Title

Development of Segmentation Variability Maps to Improve Brain Tumor Quantitative Assessment Using Multimodal Magnetic Resonance Imaging

Author

Rios Piedra, Edgar Anselmo

Publication Date

2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Development of Segmentation Variability Maps to Improve Brain Tumor Quantitative Assessment Using Multimodal Magnetic Resonance Imaging

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

by

Edgar Anselmo Rios Piedra

2018

ABSTRACT OF THE DISSERTATION

Development of Segmentation Variability Maps

to Improve Brain Tumor Quantitative Assessment

Using Magnetic Resonance Imaging

by

Edgar Anselmo Rios Piedra

Doctor of Philosophy in Bioengineering

University of California, Los Angeles, 2018

Professor Alex Anh-Tuan Bui, Co-Chair

Professor William Hsu, Co-Chair

Glioblastoma multiforme (GBM) is the most common type of primary brain tumor, characterized by a short survival period after diagnosis. As with most other cancers, treatment and follow-up decisions are made largely based on observed changes in tumor size and appearance during imaging studies.

The quantification of tumor measurements is problematic due to the systematic variability introduced while attempting to quantify tumor characteristics in uncertain

regions. This issue is primarily observed around the tumor boundary, where it is often hard to differentiate whether a given region is part of the tumor (*e.g.*, active, necrotic, edema, etc.) or part of normal brain tissue (*e.g.*, grey matter, white matter). This problem has significant implications because this uncertainty can affect ensuing quantitative/computational analyses. Current approaches for the segmentation of glioblastoma multiforme still face multiple challenges, often failing to consistently identify the tumor region so as to be clinically useful and reliable; moreover, these different techniques tend to produce results that differ significantly from each other (*i.e.*, measurement variability).

To address these problems, this dissertation describes a framework to help characterize factors that influence variability in brain tumor boundaries and to optimize their performance through methods that calculate an estimate of expected variability arising from different automated segmentation approaches, setting the bases for the development of better knowledge-based methods. Additionally, a novel automated method was developed to generate more robust brain tumor segmentations by taking into consideration the inherent variability of brain tumors and statistical priors that provide context-relevant information about the different brain and tumor tissues.

Altogether, this dissertation project provides further understanding of the sources of variability that arise in GBM across different image analysis methodologies and the integration of these insights into the development of tumor variability maps that can provide a better characterization of tumors.

The dissertation of Edgar Anselmo Rios Piedra is approved.

Ricky Kiyotaka Taira

Suzie May El-Saden

Benjamin M. Ellingson

Mark S. Cohen

Alex Anh-Tuan Bui, Committee Co-Chair

William Hsu, Committee Co-Chair

University of California, Los Angeles

2018

Dedicado a mis papás, abuelitos, hermano, toda mi familia y a mis amigos

¡Por fin terminamos!

TABLE OF CONTENTS

Acknowledgments

I would like to acknowledge a multitude of people for helping me so much and without whom I would not have been able to make through all this process. My utmost gratitude goes to:

Dr. William Hsu and Dr. Alex Bui, the co-chairs of my committee and big influences through all this time. Thanks for all your guidance, patience, the opportunity for being part of this awesome group (medical imaging and informatics lab) and helping me become a better person and better scientist. Particularly, thanks to Dr. Hsu for the countless hours of advice, always showing his inner kindness and good spirit, and Dr. Bui for sharing his genius and infectious good personality every time he was in the room.

Dr. Ricky Taira for all his teachings and always being willing to illuminate me with his wisdom and experience. Dr. Suzie El-Saden for her friendship, kindness, and extremely helpful medical insights. Dr. Benjamin Ellingson for all his contributions to my overall training, advice, and for allowing me to be part of his research group (brain tumor imaging lab). Dr. Mark Cohen for all his teachings, research insights and showing me how great people think.

All my other professors that I had the privilege of knowing and learn a lot from: Dr. Danny J.J. Wang, Dr. Corey Arnold, Dr. James Sayre, Dr. Grace Kim, Dr. Demetri Terzopoulos, Dr. Allan Yuille, Dr. Kyung Sung, Dr. Holden Wu, Dr. Daniel Ennis, Dr. Anand Santhanam, Dr. Denise Aberle, Dr. Craig Morioka, Dr. James Bisley, Dr. Robert Woods, Dr. Scott Chandler, Dr. Stephanie White, Dr. Mayumi Prins, Dr. Dan Ruan, Dr. Donatello Telesca, Dr. Noriko Salamon, Dr. Timothy Cloughesy and some others.

My friends, colleagues and huge positive help in many different aspects: Johnny Ho, Panayiotis Petousis, Nick Matiasz, Shiwen Shen, Simon Han, Nova Smedley, Karthik Sarma, Bobby Harris, Davis Woodworth, Anthony Hardy, Kevin Leu, Ararat Chakoyan, Kay Jann, Mayank Jog, Thomas Martin, Samantha Ma, Lirong Yan, Yi Wang, Robert Smith, William Speier, Kyle Singleton, Mary Zide, Jean Garcia-Gathright, Anna Wu, Maurine Tong, Iren Orosz, Emiko Morimoto, Akira Yogi, Hajime Yokota, Greg Fleishman, Jim Garritano, Yiajun Li, Tianran Zhang, Wenyuan Li, Justin Wood, Daniel Johnson, Audrey Winter, Catalina Raymond, Ruben Nechifor, Stanislas Rapacchi, Nathan Pham, Jessica Martinez, Harsha Kittur, Neil Rath, Zixin Deng, Sudarsan Ranganathan and many others.

Staff members that were always there willing to help and talk: Isabel Rippy, Eloisa Rodriguez, Denise Luna, Jenny Lee, Megan Lebre, Lew Andrada, Patrick Langdon, Shawn Chen, Audrey Gallego, Susana Hidalgo, Veronica Sandoval, and many others.

And finally, the institutions that were of huge (financial) help throughout these years. The University of California Institute for Mexico and the United States (UC-MEXUS), funding institutions from the Mexican and American governments, and the radiology, neurology, and neuroscience departments at UCLA.

Material from Chapter 3 contains material adapted from Rios Piedra, E., Taira, R., Cohen M.S., El-Saden, S., Ellingson, B., Bui, A., Hsu, W. "A Framework for Evaluating the Robustness of Segmentation Methodologies: Application to Brain Tumors". International Journal for Computer Assisted Radiology and Surgery (IJCARS), in revision, 2018. Rios Piedra, E., Taira, R., El-Saden, S., Ellingson, B., Bui, A., Hsu, W. "Assessing Variability in Brain Tumor Segmentation to Improve Volumetric Accuracy and Characterization of Change". Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on. IEEE, oral presentation, 2016.

Material from Chapter 4 contains material adapted from Rios Piedra, Edgar A., et al. "Brain Tumor Segmentation by Variability Characterization of Tumor Boundaries." International Workshop on Brain lesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer, Cham, 2016. Rios Piedra, E., Ellingson, B., El-Saden, S., Taira, R., Bui, A., Hsu, W. "Measuring Tumor Boundary Variability to Improve Automated Segmentation of Brain Tumors Using Multimodal MRI". International Society for Magnetic Resonance in Medicine (ISMRM), 2017.

Material from Chapter 5 contains material adapted from Rios Piedra, E., Ellingson, B., Cloughesy, T., Taira, R., El-Saden, S., Bui, A., Hsu, W. "Evaluating the Applicability of Tumor Probability Maps as a Resource for Improved Brain Tumor Segmentation ". Society of Neuro Oncology Annual Meeting, 2017. Rios Piedra, E., Orosz, I., Zide, M., El-Saden, S., Taira, R., Bui, A., Hsu, W. "A Usability Study to Evaluate the Impact of a Novel Automated Brain Tumor Assessment Application". American Medical Informatics Association, 2017. Rios Piedra, E., Taira, R., El-Saden, S., Ellingson, B., Bui, A., Hsu, W. "GlioView: An Application to Visualize Variability in Brain Tumor Segmentation to Inform the Clinical Assessment of Change". Radiological Society of North America (RSNA). Applied Science Presentation, 2016.

Vita/Biographical sketch

Education

2007 - 2011: B.Sc. Biomedical Engineering
Monterrey Institute of Technology, Mexico City.
2012 - 2014: M.Sc. Biomedical Engineering / Bioengineering
University of California Los Angeles, USA.

Professional Experience

2012 - 2018: Graduate student researcher (radiology, neurology)
2017 - 2018: Teaching assistant (biomedical engineering, neuroscience)
2011 - 2012: Research assistant (MRI department, Medical Sur hospital)

Publications and Conferences

Rios Piedra, E., Taira, R., Cohen M.S., El-Saden, S., Ellingson, B., Bui, A., Hsu, W. "A Framework for Evaluating the Robustness of Segmentation Methodologies: Application to Brain Tumors". International Journal for Computer Assisted Radiology and Surgery (IJCARS), in review, 2018.

Rios Piedra, E., Ellingson, B., Cloughesy, T., Taira, R., El-Saden, S., Bui, A., Hsu, W. "Evaluating the Applicability of Tumor Probability Maps as a Resource for Improved Brain Tumor Segmentation ". Society of Neuro Oncology Annual Meeting, e-poster presentation, 2017.

Rios Piedra, E., Orosz, I., Zide, M., El-Saden, S., Taira, R., Bui, A., Hsu, W. "A Usability Study to Evaluate the Impact of a Novel Automated Brain Tumor Assessment Application". American Medical Informatics Association, poster presentation, 2017.

Rios Piedra, E., Ellingson, B., El-Saden, S., Taira, R., Bui, A., Hsu, W. "Measuring Tumor Boundary Variability to Improve Automated Segmentation of Brain Tumors Using Multimodal MRI". International Society for Magnetic Resonance in Medicine (ISMRM), e-poster presentation, 2017.

Rios Piedra, Edgar A., et al. "Brain Tumor Segmentation by Variability Characterization of Tumor Boundaries." International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer, 2016.

Rios Piedra, E., Taira, R., El-Saden, S., Ellingson, B., Bui, A., Hsu, W. "GlioView: An Application to Visualize Variability in Brain Tumor Segmentation to Inform the Clinical Assessment of Change". Radiological Society of North America (RSNA). Applied Science Presentation, 2016.

Rios Piedra, E., Ho, King Chung, Taira, R., El-Saden, S., Ellingson, B., Bui, A., Hsu, W. "Glioblastoma Multiforme Segmentation by Variability Characterization of Tumor Boundaries" Medical Image Computing and Computer Assisted Intervention Society (MICCAI), MICCAI-BRATS Conference Workshop, 2016.

Jann, K., Smith, R., Rios Piedra, E., Dapretto, M., Wang D. "Noise reduction in Arterial Spin Labeling based Functional Connectivity Using Nuisance Variables" Frontiers in Neuroscience (Advanced Neuroimaging Methods for Studying Autism Disorder), 2016.

Rios Piedra, E., Taira, R., El-Saden, S., Ellingson, B., Bui, A., Hsu, W. "Assessing Variability in Brain Tumor Segmentation to Improve Volumetric Accuracy and Characterization of Change". Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference. IEEE, oral presentation, 2016.

Jann, K., Rios Piedra, E., Smith, R., Wang, D. "Effect of noise regression on Arterial Spin Labeling based functional connectivity". International Society for Magnetic Resonance in Medicine (ISMRM), e-poster presentation, 2015.

Shie, A., Hollada, J., Rios Piedra, E., Sung, K., Adams, J. "A comparison of trabecular microarchitecture and bone marrow fat fraction in Caucasian versus Asian American young adults." American Society for Bone and Mineral Research (ASBMR), poster, 2014.

Roldan-Valadez, E., Rios Piedra, E., Favila, R., Alcauter, S., & Rios, C. "Diffusion tensor imaging-derived measures of fractional anisotropy across the pyramidal tract are influenced by the cerebral hemisphere but not by gender in young healthy volunteers: a split plot factorial analysis of variance (ANOVA)." Chin Med J 125 (2012): 2180-2187.

Chapter 1: Introduction

1.1 Problem definition

Primary brain tumors are one of the most aggressive human cancers, making improvements in treatment and monitoring techniques one of the current top challenges in neuro-oncology. The most aggressive type (highest grade) of brain tumors is known as Glioblastoma Multiforme (GBM).

Despite current advances in therapy, the outcome for patients with high-grade gliomas remains poor, with a median survival of 14–18 months [1-2] following diagnosis. Ninety percent of patients die after 18 months and only 3.3% live longer than 5 years after diagnosis [2]. This dire course is due in part to the lack of characteristic symptoms during early stages of tumor development, and a paucity of effective non-invasive detection methods [3-4].

In GBM and other types of cancers, treatment and follow-up decisions are largely made based on observed changes in size and appearance following imaging, including magnetic resonance imaging (MRI) studies. Currently, for GBM this characterization includes different MRI sequences (*e.g.* T1 pre/post-contrast, FLAIR, *etc.)* to identify and characterize the tumor and its different subparts (*e.g.,* enhancement, necrosis, edema); however, such characterization is affected by variability in the subsequent analysis process. This problem becomes evident when different raters (*e.g.* expert neuro-radiologists,

automated or semi-automated algorithms.) characterize the same lesion and produce different results that can vary significantly from one another. More importantly, these differences can have a measurable effect on downstream analyses and clinical decisions.

Measurement variability, which, for the purposes of this dissertation, is defined as the uncertainty associated with the object's (*e.g.* tumor) 3D boundary. Such variations can be attributed to: 1) poor boundary distinction due to biological tissue heterogeneity; 2) inaccuracies in the characterization process (manual or automated); and/or 3) variability in image acquisition (*e.g.* signal levels, cross-sectional views, and noise). Although different manual and semi-automated methods exist to help the neuroradiologist characterize a GBM (*e.g.,* computing tumor volume and change over time), they do not consider the error bounds of the tumor boundary in the final measurement, and do not produce a metric that informs the clinician about significant characteristics of brain tumors (*e.g.,* variability, change).

The problem of inconsistent and inaccurate measurements in GBM segmentation originates in part because such tumors tend to have irregular shapes, discontinuous borders, and a high degree of border variability. These problems contribute to the challenging task of developing discriminative classification methods that can measure and characterize variability in high-grade brain tumors. Given the importance of imaging in diagnosis and treatment of this disease, developing methods that can consistently quantify

and subsequently characterize GBMs has the potential to enable new insights regarding disease progression, and ultimately improve patient outcomes.

Understanding, characterizing, and reducing measurement variability can lead to more accurate and consistent segmentations, resulting in improved quantitative results and ultimately increased assurance in the clinical MRI assessments and definition of changes over time.

Additionally, the investigation of this problem can characterize the root causes of tumor heterogeneity (defined as the diversity of biological entities that can be found while analyzing an image) observed during the brain tumor characterization so that this work in measurement of image variability can be considered and integrated into future processing frameworks and algorithms.

1.2 Contributions

This dissertation presents a framework to characterize the factors that influence variability in brain tumor boundaries, developing methods to optimally calculate and define the expected variability that arises from different segmentation approaches. This provides the bases for the development of better statistical- and knowledge-based methods that use prior information about cerebral tissue properties. This work is structured around three fundamental research contributions:

1. Understanding the sources of variability that arise in GBM tumor segmentation. Characterization of the tumor boundary using different segmentation approaches (algorithms) leads to the identification of factors (*i.e.,* parameters affecting algorithm performance); and their optimization can lead to a reduction of the differences observed between methods.
2. Development of a new automated tumor segmentation algorithm to determine overall tumor boundary as well as its different sub-regions. A combination of statistical- and knowledge-based methodologies can help obtain a more accurate characterization of the tumor boundary at the pixel-level using Tumor Variability Maps (TVMs).
3. Evaluating the impact of segmentation variability metrics in the assessment of medical images. Hypothesizing that these measurements contribute towards a reduced uncertainty during radiological evaluation. This evaluation also includes the development of a novel end-to-end imaging pipeline with a user interface that generates the tumor variability metrics.

These insights are integrated into the development of segmentation algorithms that produce variability maps (confidence maps), which can provide improved characterization and understanding of medical images.

1.3 Organization of the dissertation

This dissertation is organized into six chapters, including this introductory chapter. Figure 1.1 presents a diagram with the sequence and connections of the information presented in the different chapters.

Chapter 2 provides background about the different aspects that affect the characterization of brain tumors, including biological and algorithmic aspects, to bring context and relevance to the work subsequently described. This chapter also includes current clinical standards available for brain tumor characterization, their relevance to patient assessment, as well as their disadvantages. Also, the imaging techniques in terms of acquisition and current methods for automated tumor segmentation are covered in this chapter.

Chapter 3 covers the majority of the Research Contribution 1, expanding on the calculation of tumor variability in different segmentation approaches, leading to the generation of Tumor-measurement Variability Maps (confidence maps). This chapter also includes the optimization experiments for several tumor segmentation approaches, and the framework to automatically evaluate and reduce sources of variability.

Chapter 4 centers on Research Contribution 2, including the design and development of a statistical and knowledge-based algorithm to segment brain tumors. The chapter covers the development of the method, the addition of prior information into the

segmentation process to automate the definition of the tumor location, and a metric representing the degree of variability found for each case. The method is evaluated using two large databases of brain tumors with reference annotations obtained from UCLA and TCGA/MICCAI (The Cancer Genome Atlas and the Medical Image Computing and Computer Assisted Intervention society, respectively).

Chapter 5 explores the clinical utility of variability metrics (Research Contribution 3), exploring the interpretation of variability metrics when assessing tumors at a determined time (single time point), or tumor volume change in cases with multiple follow-ups. A utility study and a retrospective analysis were conducted on a subset of patients to determine the effect of these newly developed metrics on the assessment by neuro-radiologists.

Finally, Chapter 6 summarizes the findings, the technical problems and limitations identified during this project, and possible future directions that could follow the results presented in this dissertation.

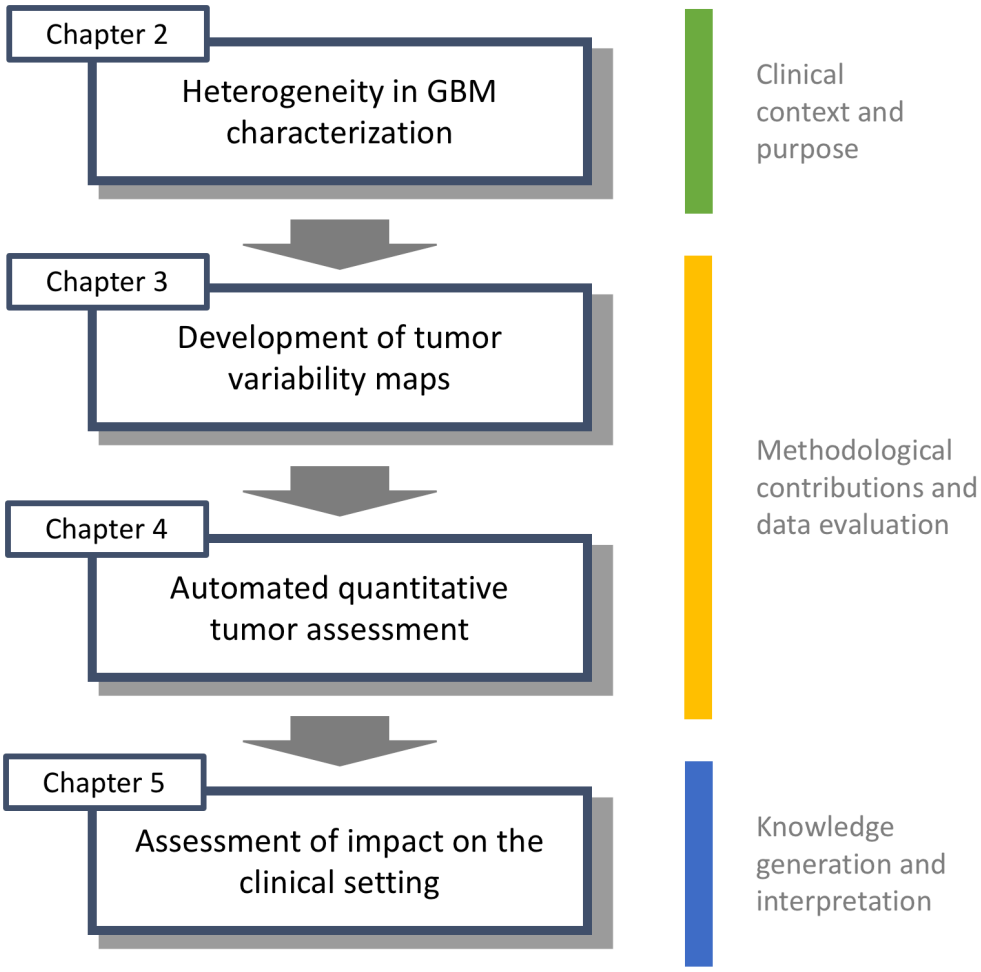


Figure1.1: Organization of the dissertation.

Chapter 2: Understanding Sources of Variability

This chapter discusses the potential sources of variability in tumor segmentation and the need for measurements of quantification error (*i.e.,* variability) in order to guide reliable clinical assessments. The chapter is organized into three parts: 1) an overview of how treatment assessment of GBM patients is performed today; 2) the sources of biological heterogeneity that influence the appearance of tumors, and; 3) the sources of image variability due to acquisition parameters and post-processing methods. A literature review is then provided of current methods for tumor segmentation and variability characterization, followed by metrics for measuring agreement between segmentations and a reference standard.

2.1 Characterization of treatment response using imaging

Various multi-institutional efforts have created clinical guidelines or standards to aid the neuroradiologist in the task of characterizing brain tumor images [5]. Some of the most used guidelines for tumor structural characterization are:

- Visually Accessible Rembrandt Images (VASARI): This controlled terminology for describing the magnetic resonance (MR) features of human gliomas was devised based upon prior work (REMBRANDT project). It consists of a comprehensive feature-set of 24 observations familiar to neuroradiologists to describe the

morphology of brain tumors in routine contrast-enhanced MRI. These guidelines exclude advanced MRI techniques such as dynamic susceptibility contrast (DSC), magnetization transfer (MT), magnetic resonance spectroscopy (MRS), diffusion tensor imaging (DTI) and perfusion/permeability characteristics [6]. Some drawbacks of these guidelines include the poor inter-reader agreement when different clinicians evaluate the same studies [7]. Table 2.1 summarizes the suggested image features for tumor structural characterization.

- Response Evaluation Criteria in Solid Tumors (RECIST): RECIST is a set of published rules that assess tumor burden to provide an objective assessment of therapy response (updated to RECIST1.1 in 2009) [8]. It is based on linear measurements along the longest axis of the tumor, and registers response in four categories: full response is defined as the disappearance of all target lesions; partial response as a decrease of 30% on total lesion size; progressive disease as 20% increase (compared to baseline) on the measurable lesions or a 5mm absolute increase in total measurements (progression can be identified by visually evident progression in the number or size of non-measurable lesions); finally, if there is not enough growth or shrinkage of the lesions, then status is defined as stable disease. The main drawback of this set of rules is that it considers only one-dimensional measurements. This terminology characterizes tumors as measurable versus non-measurable and target versus non-target (Table 2.2).

- Response Assessment in Neuro-Oncology (RANO) criteria: This set of criteria is an update of the McDonald criteria, which is based on two-dimensional measurements of the enhancing component of the tumor [9]. In recognition that contrast enhancement is nonspecific and may not always be a true surrogate of tumor response, and the need to account for the non-enhancing component of the tumor, these new criteria were developed to permit accurate assessment of imaging studies. These guidelines are under constant evolution to adapt to current clinical needs and advancements [10]. A brief description of the parameters that are considered to assess tumor response to treatment is summarized in Table 2.3.

These different structural features play an important role in determining the grade of a brain tumor and treatment selection [7,11]; however, when these standards are applied in the clinic, there is a tendency to obtain metrics that are highly variable due to the heterogeneous characteristics of the tumor, making it difficult to predict future development [7].

In most clinical centers the current measuring standard is the RANO criteria, but the current recommendations for defining tumor change and the basis of a one-dimensional metric on the largest axis of the tumor is in most cases not an appropriate representation of the lesion, leading to misinterpretations and errors (Figure 2.1). The VASARI guidelines provide a more complete set of imaging markers to characterize the

lesion, but in many cases the characterization is broad and qualitative, relying on the observer to make the judgment and arising variability issues. Additionally, these metrics are sometimes skipped in clinical practice because of the time and effort it takes to make these characterizations, motivating the need for an automated way to generate these measurements.

RECIST criteria	
Feature	**Description**
Non-measurable lesions	Series of binary values, positive if the feature is identifiable (but not measurable) - Lesions smaller than 10mm. - Skeletal metastases. - Leptomeningeal disease. - Cystic or necrotic lesions. - Tumor lesions situated in previously irradiated regions.
Measurable lesions	Defined in the same way as the non-measurable lesions but the longest diameter exceeds 10mm.
Target lesions	Definition of targeted organs (*e.g.* brain, lymph nodes, etc.) to center treatment on, in case more than one is found (selected on the basis of size).
Non-target lesions	Lesions that do not need to be quantified, characterized as total disappearance, continued presence, and unequivocal progression.

Table 2.2. Summary of structural imaging components described by the RECIST guidelines towards tumor characterization. Note: Lesions that are not directly related to brain tumor characterization were excluded (*e.g.* abdominal masses) [12].

VASARI guidelines		
#	**Name**	**Description / possible values**
f 01	Tumor Location	Frontal, temporal, insular, parietal, occipital, brain-stem.
f 02	Side of lesion center	Right, center/bilateral, left.
f 03	Eloquent brain	Speech motor, speech receptive, motor, vision.
f 04	Enhancement quality	None, minimal/wild, marked/avid.
f 05	Proportion enhancing	Percent of enhancing tumor compared to total volume.
f 06	Proportion nCET	Percent of non-enhancing tumor compare to total vol-ume.
f 07	Proportion necrosis	Percent of necrotic tumor compared to total volume.
f 08	Cysts	Positive if well-defined regions of bright T2 signal and low T1 with very thin and regular enhancing walls are found.
f 09	Multifocal (multicentric)	Multifocal, multicentric, gliomatosis.
f 10	T1/FLAIR ratio	Expansive (FLAIR = T1), mixed (FLAIR >= T1), infiltrative (FLAIR > T1).
f 11	Thickness of enhancing margin	None, thin (<3mm), thick/nodular (=> 3mm), solid.
f 12	Definition of the enhancing margin	None, well-defined, poorly-defined.
f 13	Definition of the non-enhancing margin	None, well-defined, poorly-defined.
f 14	Proportion of edema	Percent of edema compared to total volume.

f 16	Hemorrhage	Positive if intrinsic hemorrhage in the tumor matrix is found, intrinsic foci have low T2 signal and high T1 signal.
f 17	Diffusion characteristics	Facilitated, restricted, mixed. Based on ADC map.
f 18	Pial invasion	Positive if enhancing of pia next to the tumor is observed.
f 19	Ependymal extension	Positive if tumor adjacent ependymal regions are invaded.
f 20	Cortical involvement	Positive if cortex near tumor is no longer distinguishable.
f 21	Deep white matter invasion	None, Internal capsule, brainstem, corpus callosum.
f 22	nCET crosses midline	Positive if nCET crosses to contralateral hemisphere through white matter commissures.
f 23	CET crosses midline	Positive if CET crosses to contralateral hemisphere through white matter commissures.
f 24	Satellites	Positive if enhancing region is observed near dominant tumor but there are no continuous connections.
f 25	Calvarial remodeling	Positive if erosion of inner side of the skull is observed.

Table 2.1. Summary of structural imaging components described by the VASARI guidelines for tumor characterization. nCET = non-contrast-enhanced Tumor, CET = contrast-enhanced tumor. Note: Feature 15 was omitted by the authors because of unknown reasons [6].

RANO criteria for disease assessment	
Response	**Criteria**
Complete response	- Complete disappearance of all enhancing measurable and non-measurable disease sustained for at least 4 weeks. - No new lesions. Stable or improved non-enhancing (T2/FLAIR) lesions. - Patients must be off corticosteroids. - Clinical status defined as stable or improved.
Partial response	- ≥ 50% decrease compared with baseline in the sum of products of perpendicular diameters of all measurable enhancing lesions sustained for at least 4 weeks. - For volumetric assessments, ≥ 50% decrease in unidimensional metrics corresponds to a ≥65% decrease in volumetric change. - No progression of non-measurable disease. - No new lesions. - Stable or improved non-enhancing (T2/FLAIR) lesions on same or lower dose of corticosteroids compared with baseline scan. - Corticosteroid dose at the time of the scan evaluation should be no greater than the dose at the time of baseline scan. - Clinical status defined as stable or improved.
Stable disease	- Does not qualify for complete response, partial response, or progression. - Stable non-enhancing (T2/FLAIR) lesions on same or lower dose of corticosteroids compared with baseline scan.
Progression	- ≥ 25% increase in the sum of the products of perpendicular diameters of enhancing lesions compared with the smallest tumor measurement obtained either at baseline (if no decrease) or best response, on stable or increasing doses of corticosteroids. - For volumetric assessments, ≥ 25% increase in unidimensional metrics corresponds to a ≥40% increase in volumetric change - Significant increase in T2/FLAIR non-enhancing lesion on stable or

	increasing doses of corticosteroids compared with baseline scan or best response after initiation of therapy not caused by comorbid events. - Any new lesion. - Clear clinical deterioration not attributable to other causes apart from the tumor (*e.g.* seizures, medication adverse effects, complications of therapy, cerebrovascular events, infection, and so on). - Changes in corticosteroid dose. - Failure to return for evaluation as a result of death or deteriorating condition. - Clear progression of non-measurable disease.

Table 2.3. Summary of different requirements to assess disease response to treatment according to the RANO criteria. [13].

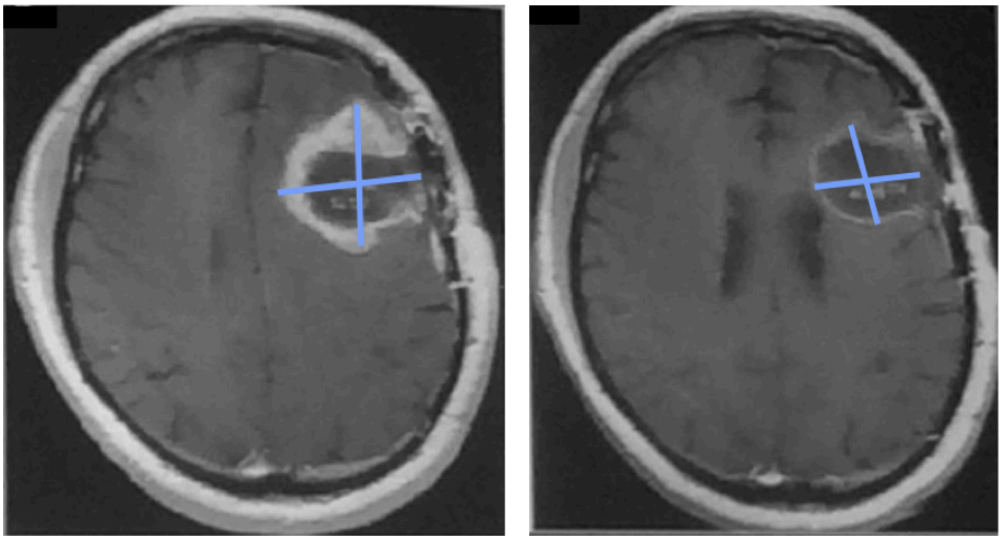


Figure 2.1. Example of where current one-dimensional measurement can fail to characterize tumors. The left side shows a tumor and resection cavity before going through therapy, the right side presents a scan after the therapy had concluded with a decrease in the enhancing margin surrounding the resection cavity; however, by the current analysis metrics, the major axis measurements do not represent the real effect of the therapy, misclassifying the case as stable disease instead of partial response.

2.2 Sources of Tumor Variability

Imaging has a critical role in understanding disease extent, progression and, making treatment decisions; However, tumors often have a mixture of diverse morphological features, such as discontinuous borders, non-uniform and unpredictable shapes, and different tissues mixed throughout the overall mass. These factors together hinder the ability of clinicians and computational approaches to determine the extent of the disease. Although this problem regarding heterogeneity was first described as an issue long ago [14], it continues to pose challenges for biological and algorithmic aspects.

Regarding biological aspects, the heterogeneous characteristics of GBMs can be observed at different levels (macroscopic, cellular, molecular), and is an important factor when trying to understand tumor progression, treatment resistance, metastatic potential, recurrence probability, and other features [7, 15]. For instance, different types of tissues can be found within the tumor (*e.g.* different types of cellular populations) and different tumors within the same organ can have independent origins and present different morphological and physiological behaviors. For example, one brain tumor can be a primary tumor and the other a result of metastatic development originating from another anatomical region.

These differences observed at the macroscopic level are likely to reflect differences in deeper aspects of the tumor including (genetic components, metabolism, proliferation

rate, and vascular activity.) that may be used to determine the most appropriate treatment. Figure 2.2 shows different examples of how this biological heterogeneity is observed in imaging studies.

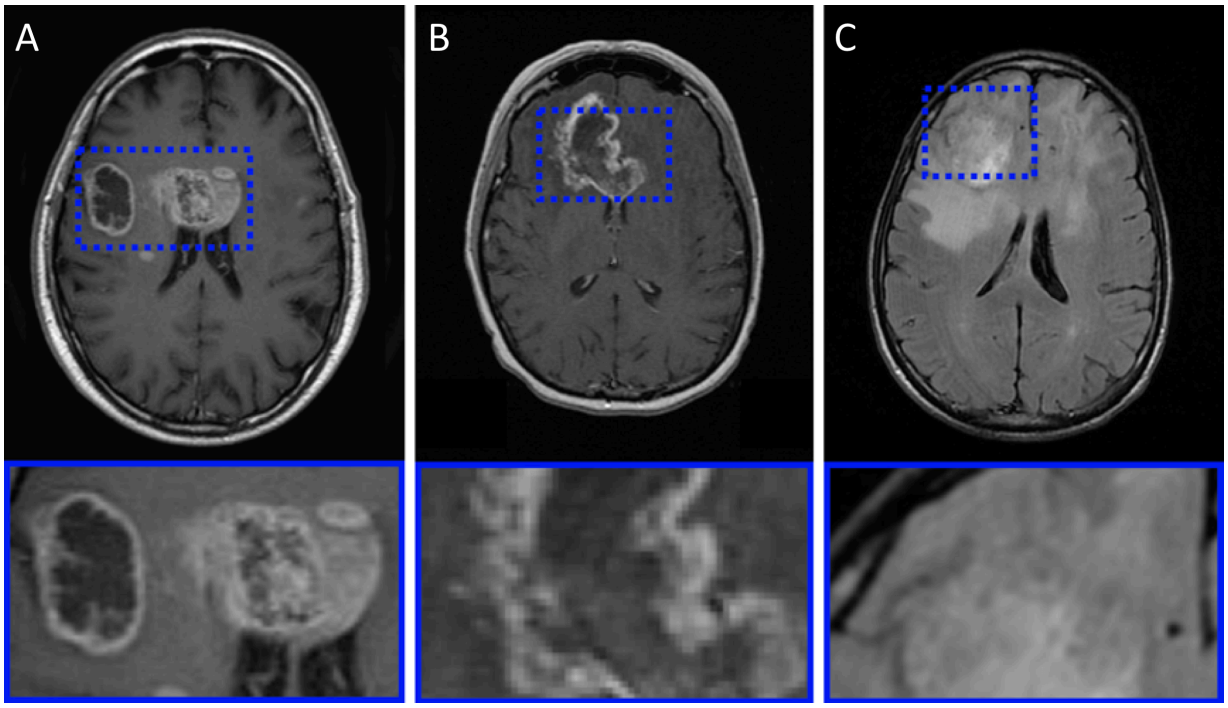


Figure 2.2. Adjacent tumors can present different characteristics between each other, and thus, different responses to treatment. Not considering these differences can lead to unsatisfactory response and an increased likelihood of relapse. Examples of tumor variability include differences in multifocal tumors with different overall appearance (A) and the different regions or tumor sub-components that are observed under different contrasts (active tumor versus necrotic tumor on B and tumor versus edema on C).

This inherent tumor variability manifests as variations in image characterization, including sharp/ill-defined borders, intensity variations between different sub-regions,

and smooth (unclear) edges observed on images that can potentially be found by applying different image analysis methods such as different mathematical approaches, different parameters of the segmentation methods, measurement assumptions and others. These methods aim to segment the tumor boundary by measuring this image variability, and determine the tumor boundary, allowing one to differentiate between normal and tumor tissue as well as hyper-intense areas that appear when contrast agents are used (in comparison to non-enhancing regions); zones of decreased intensity inside the overall tumor mass (necrosis); or morphological distortions along the white matter tracts that might correspond to edema surrounding the solid tumor.

Given that identifying different sub-regions of a tumor is a clinically significant task and is often a source of quantification error [16], this work focuses on the characterization of the different sub-regions that compose the total tumor mass.

Undoubtedly, variability and error in tumor characterization arise from different sources throughout the entire imaging process and subsequent analysis, ranging from sources of error during acquisition to quantitative feature extraction and analysis.

2.3 Variability in the imaging process

Variability in measurements affects different aspects of the medical image analysis process. In the imaging aspect, it is of critical importance to know that many subsequent *radiomic* features (*e.g.,* volume, change over time, texture) are segmentation-dependent.

Additionally, the interpretation of these imaging features has been observed to improve when precision metrics are provided with them [17].

Furthermore, having knowledge of the distribution of a feature obtained through separate measurements can increase the probability of a correct classification when there is only a single measurement, it is easier to make incorrect judgments as there is no information about the distribution or variability associated with that measurement.

Moreover, pixel distribution for different structures (*e.g.*, among normal tissues or between normal and pathologic tissues) can be complicated when it comes to grouping them, since there is overlap in pixel distribution, and the same pixel can belong to different possible classifications (tissues) (Figure 2.3).

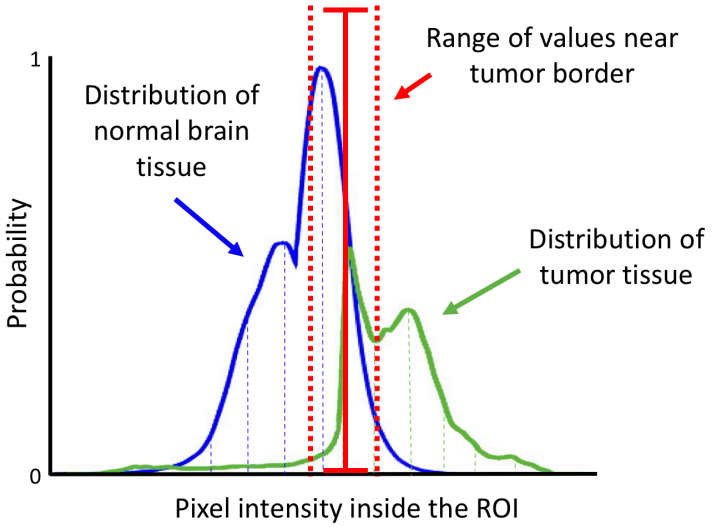


Figure 2.3. The curves exemplify the probability distributions of the intensity value for normal and tumor tissues and the tumor boundary in the intersection between them. Knowledge of the distribution of possible places where the tumor boundary is located (shown in red) obtained through separate measurements can increase the probability of a correct classification into one appropriate category -normal or tumor tissue.

Uncertainty and measurement errors are a problem that is not only present in the image analysis process or the tumor segmentation stage. Measurement variability can be found in many other aspects of the imaging process, such as acquisition (machine characteristics, patient-induced artifacts) and other processing that happens before the study can be interpreted by the clinician (image reconstruction, registration, homogenization of image parameters, etc.).

When different measurements are available, it is possible to add a confidence interval, allowing for a more informed decision. The sources of inhomogeneity throughout the imaging process, and the way these the different sources of variability influence the imaging process are presented in Table 2.8.

Sources of error in image analysis		
	Process	**Examples of the source of error**
Object to image	Scanner physical properties	Field inhomogeneity, maximum gradient amplitude, Modulation Transfer Function (MTF), quality and calibration of detectors.
	Spatial resolution	One voxel represents thousands of cells and different processes, depending on scanner matrix size, the field of view and slice thickness.
	Imaging artifacts	Motion artifacts, sampling artifacts, reconstruction inaccuracies, RF noise, chemical shift artifacts.
	Signal quality	Signal to noise ratio, optimized tissue contras t, imaging sequence parameter standardization.
Image to feature	Noise and motion correction	SNR and CNR of the images, motion correction algorithms, bias field correction, etc.
	Image registration	Accuracy of alignment, data interpolation, temporal registration, algorithm parameters.
	Tumor segmentation	Algorithm used for total tumor sub-compartments, algorithm parameters, accuracy of the output ROI, reproducibility and robustness of measurements.
	Temporal change tracking	Temporal uniformity, significant estimation of change, parameter standardization, missing information.

Table 2.8. Summary of the different sources of image uncertainty and error during the imaging process. The variability sources that are going to be explored as part of the project mainly belong to the "image to feature" process. SNR = Signal to noise ratio, CNR = Contrast to noise ratio, ROI = Region of interest.

2.4 Current approaches in GBM segmentation

Given the necessity of obtaining an improved characterization of tumor and response to treatment, there have been numerous efforts in the domain of imaging informatics and computer science to try to overcome this problem. Some of the most recent approaches are briefly described in Table 2.4 [18]. These efforts are driven by the desire to generate accurate segmentations automatically in real-time.

Nevertheless, the segmentation of brain tumors remains a challenging task for several reasons, including that GBMs tend to exhibit unclear and irregular boundaries with discontinuities (biological tissue heterogeneity); automated systems are hard to train given the distinct morphological appearance of tumors; and training sets of annotated data are hard to develop and subject to annotation variability.

There are other difficulties that pertain to the image acquisition process. For instance, time differences between the uptake of contrast agent and actual image acquisition can influence the appearance of the enhancing tumor. The images frequently contain regions that in reality correspond to tumor tissue but are not different from the other structures [16].

Author	Approach	Segmentation accuracy (Dice coefficient)		
		Whole	Core	Active
Bauer [19]	Integrated hierarchical random forest classification and CRF regularization.	68	48	57
Buendia [20]	A grouping artificial immune network for tumors.	57	42	45
Cordier [21]	Patch-based tissue segmentation approach.	68	51	39
Doyle [22]	Hidden Markov fields and variational EM in a generative model.	74	44	42
Festa [23]	Random forest classifier using neighborhood and local context features.	62	50	61
Geremia [24]	Spatial decision forests with intrinsic hierarchy.	62	32	42
Guo [25]	Semi-automatic segmentation using active contours.	74	65	49
Hamamci [26]	"Tumorcut" method.	72	57	59
Meier [27]	Appearance- and context-sensitive features with a random forest and CRF.	69	50	57
Menze (II) [28]	Generative-discriminative lesion segmentation model. This method consists of an improvement of the Menze (1) segmentation method.	78	58	54
Menze (I) [29]	A generative lesion segmentation model for multimodal brain tumor segmentation.	69	33	53
Reza [30]	Texture features and random forests	70	47	55
Riklin-Raviv [31]	Generative model with latent atlases and level sets.	74	50	58
Shin [32]	Hybrid clustering and classification by logistic regression.	30	17	05
Shubbanna [33]	Hierarchical MRF approach with Gabor	75	70	59

	features.			
Taylor [34]	"Map-Reduce Enabled" hidden Markov models.	44	28	41
Tustison [35]	Random forest classifier using the open source ANTs/ANTsR packages.	75	55	52
Zhao (I) [36]	Learned MRF on supervoxels clusters.	<u>82</u>	66	49
Zhao (II) [37]	Learned MRF on supervoxels clusters with updated unary potential.	76	51	52
Zikic [38]	Context-sensitive features with a decision tree ensemble.	75	47	56
	Average result of all methods.	67	48	49

Table 2.4. A survey of different approaches for automated brain tumor segmentation submitted to the MICCAI brain tumor segmentation challenge [18]. Note the range of Dice similarity coefficient across methods and tumor regions. The best performing method is underlined.

One way to address this issue has been by using a multi-modal approach, which refers to the use of different imaging "contrasts" that together allow a better visualization of differences and boundaries between healthy and abnormal tissue (Figure 2.4). Tumor sub-regions (Table 2.5) can be better identified more accurately when several modalities or imaging sequences are combined. Also, the combination of these different imaging sequences entails the use of image pre-processing techniques (such as skull-stripping, registration, etc.).

Tumor region	MRI appearance	Clinical Interpretation
Enhancing tumor	Hyper-intense region in T1 weighted image with contrast agent.	Active viable tumor with micro-vascular proliferation, poor blood-brain barrier.
Necrotic tumor	Hypo-intense regions in T1 weighted images with or without contrast agent.	Pool of dead cells that are no longer active due to insufficient blood supply.
Non-enhancing tumor	Regions that are visible as abnormalities in T1 weighted images but do enhance when the contrast agent is injected.	Tumor suggestive of a lesion that may progress to a higher grade enhancing lesion. May also be indicative of lower-grade glioma, cortical dysplasia or part of an inflammatory process.
Edema region	Hyper-intense region in images with T2 contrast such as T2 maps, FLAIR images or T2 weighted images.	Extension of edematous regions outside the boundaries of the tumor. Likely to contain migrating tumor cells.

Table 2.5. Different areas of interest related to brain tumor image characterization. These parts combined form the total tumor mass (total lesion).

Some of the recently developed algorithms that have been tested as part of the MICCAI [18] tumor segmentation competition use multimodal approaches. In this competition, each participant generates a segmentation of the tumor area by using a specific algorithm developed by each team. Then, the results drawn from the different models are evaluated to determine which has the most similarity with the manually labeled

dataset and is the most consistent (repeatable). In the most recent results published from this competition [18-39], it can be observed that the algorithm based on Markov Random Fields (MRF) developed by Zhao [38] performed best for segmentation of total tumor, but results were not consistent when evaluating sub-regions of the tumor such as core and enhancing tumor regions—a task on which Shubbanna (MRF with Gabor Features) [33] and Festa (Random Forest) [23] performed the best. Additionally, the agreement with the manual reference standard measured using the Dice coefficient reached an average of 49% on enhancing tissues, demonstrating that further improvement in this domain is needed. A complete description of these approaches can be found in Table 2.4 and Figure 2.5. Table 2.6 summarizes the methodology (*i.e.,* algorithm) that was utilized for the development of these segmentation approaches. Finally, Table 2.7 shows the distinct kinds of imaging features that are considered by these approaches in modeling the variable boundaries of brain tumors.

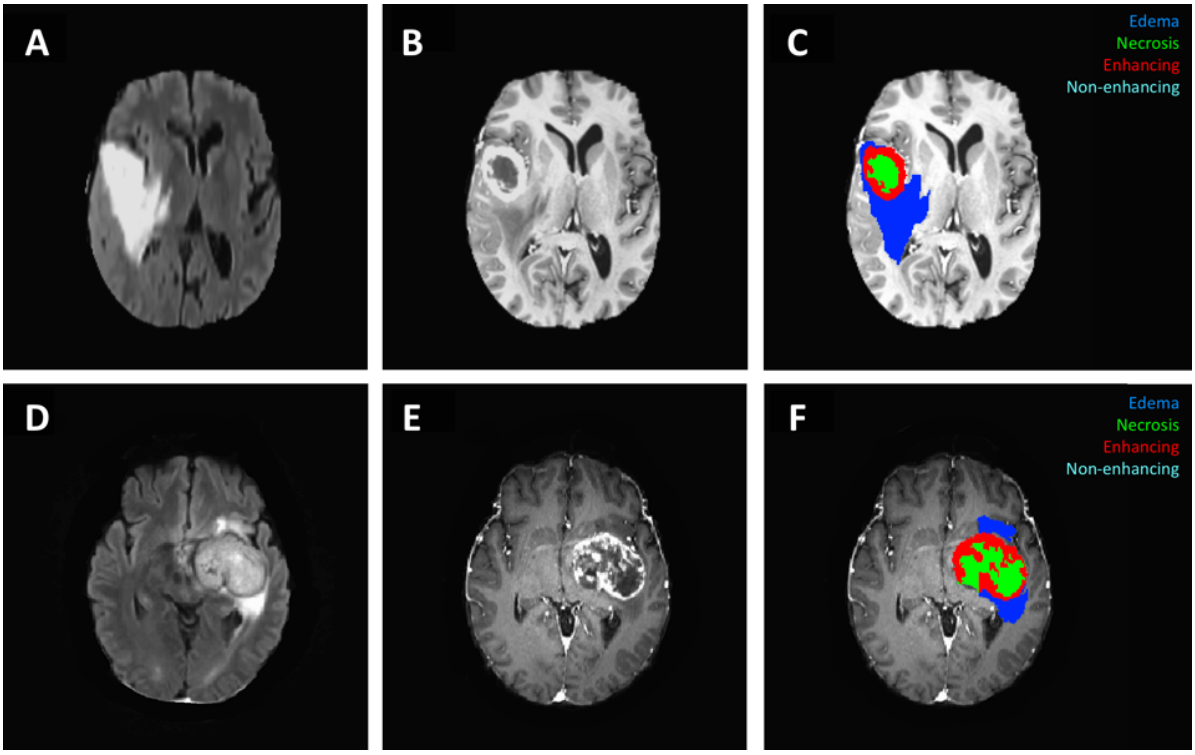


Figure 2.4. Two cases that illustrate the advantages of multimodal imaging for the segmentation of tumor subcomponents. Each row shows a FLAIR and post-contrast T1 scans (A, B and D, E respectively) and the output segmentation of the different tumor regions (C).

A brief description of the properties (features) used by the algorithms that have been published as part of the mentioned tumor segmentation competition is listed below, categorized by the types of features that are used. A more detailed description of the algorithms utilized on the segmentation challenge is found in the supplementary materials for this chapter found at the end of the dissertation.

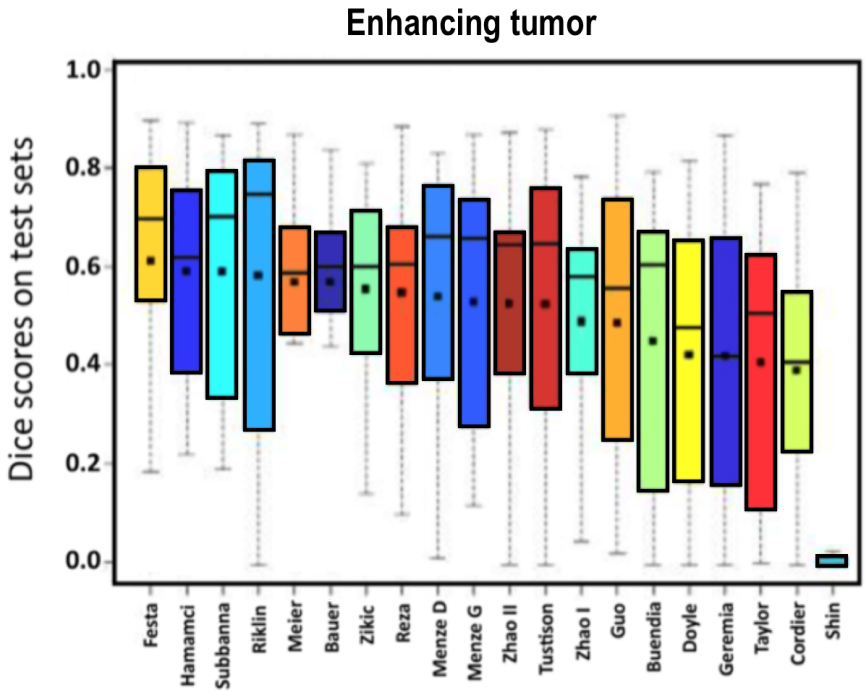


Figure 2.5. The box plot shows the Dice coefficients for different methods on the segmentation of active tumors. Results were obtained from a sample of 100 datasets from the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) brain tumor segmentation competition for the last two years [18].

- Pixel intensity: Pixel intensity is used in most (if not all) image processing approaches as one of the most important means to classify objects or structures of interest. In MR images, the intensity captured in each pixel has a direct correlation to the magnetization properties of different tissues. Thus, this property can be used to identify or group different structures with similar intensities. Pixel intensity can be considered on a pixel-wise basis or in conjunction with the local neighborhood as image patches. Although pixel intensity is one of the default features to use, some disadvantages include the sensitivity to image artifacts and noise,

partial volume effects (different structures being represented in a single pixel), boundary definition in smooth images, and problems differentiating different objects with similar intensities.

- Pixel neighborhood: Information about the pixel surroundings can help contextualize the image intensity observed for a given pixel, so considering the adjacent regions to it has become general practice in image analysis methods. This aims to reduce the effect on fine intensity fluctuations originated from noise or image artifacts, often performing the pixel classification by considering the normalized (average) value of the pixel with its neighbors instead of the single pixel value. This can be performed on a pixel-wise basis or on image patches using, uniform grids of a defined size or image superpixels. Finally, pixel neighborhood can be considered in two dimensions by looking at the 8-pixel neighborhood around the pixel or in three dimensions by considering the adjacent slices similar to a region-growing approach). Disadvantages include the high sensitivity to preprocessing techniques such as noise reduction filters and intensity normalization, a decreased capability of finding fine structures similar to an erosion, and the increase of the complexity while implementing multiple neighbor comparisons.
- Textural patterns: Textural metrics are used to identify patterns or specific arrangements in the images that can lead to a better definition of the boundary observed between two objects. Common textural features include mean, variance,

co-occurrence matrices, kurtosis, uniformity, compactness and many others. These features are often used in learning-based methods to significantly increase the information space to characterize the images during the training phase. Additional advantages include a rather straightforward implementation, and the capacity to detect semi-repetitive patterns in the images, which are otherwise difficult to find. Disadvantages include the difficulty in understanding the biological meaning of many of those features (*e.g.* the role of the inverse kurtosis for the boundary detection), the fact that the classification performance changes depending on the scale at which the features are obtained, and the possible introduction of incorrect correlations or false-positives (type 1 errors) given the non-specificity of these features.

- Edge detection: Edge detection techniques include the calculation of the image gradient. Magnitude and phase maps show the direction on which the image intensity varies, and different kinds of filters can be used to enhance certain image properties such as high-low-band pass filters and Gabor filters. Edge detection methods are made to detect regions where sharp changes in intensity occur, improving classification capacity by finding the optimal boundary location and reducing the effect of random noise. Some difficulties are manifested in image regions that are highly variable or have blurry edges. Their performance is also

significantly dependent on the filter parameter selection including kernel size and shape.

- Image symmetry: A normal brain has approximate bilateral symmetry, allowing for comparisons with the contralateral hemisphere to locate potential abnormal regions. This is one of the most discriminating features radiologists use when assessing brain images. The presence of a tumor and other conditions that influence brain anatomy such as stroke or traumatic brain injury often distort brain anatomy, making morphological symmetry comparisons useful when trying to detect these anomalies. This is based on the assumption that normal areas have a high correlation with the contralateral elements [97]. Some challenges arise when the lesion is located on the symmetry plane used to make the comparison, or when there is a large abnormality that introduces morphological distortions on both sides of the image. Image realignment or registration is often performed to correct for rotations or translations which occurs during acquisition, potentially making symmetry analysis difficult.
- Atlas-based labels: These features are based on the use of images previously annotated by an expert, generating labels that correspond to different brain structures such as different tissues, pathways or functional organization. This information is then extrapolated to a test image to provide knowledge and context

about the location of the different structures. This helps to identify abnormal regions in cases when there is low contrast, noise, or situations that can be ambiguous or difficult to classify using intensity analysis. A similar approach can be used to create custom databases (dictionaries) for a specific situation. For example, the creation of labels that contain examples of healthy and pathological tissues can be compared with a test set of images. Challenges include the need for image registration to match the atlas. In some cases, large deformations (big tumors) make it difficult to match the structures (labels) with the atlas (usually normal brains), and structure correspondence is barely achieved.

Another way to reduce the variability in segmentation includes the measurement of several ROIs for each case and generation of a "consensus" result that is obtained from these different approximations to better represent the real tumor boundary. Figure 2.6 shows an example of how different methodologies can derive different results for the same case. Figure 2.7 shows an example of how these different boundary approximations can derive variability maps that represent tumor likelihood (confidence maps). These ideas of multimodal image analysis and determination of tumor boundary are explored in subsequent chapters (Chapter 3 and Chapter 4).

Author	Method used						
	Random Forest	CRF	Dictionary match	Label propagation	HMRF	Log Regression	Graph-cut method
Bauer	X	X					
Cordier			X	X			
Doyle					X		
(Cont..)	Random Forest	CRF	Dictionary match	Label propagation	HMRF	Log Regression	Graph-cut method
Geremia	X						
Meier	X	X					
Shin			X			X	
Shubbanna					X		
Taylor					X		
Zhao(I)					X		X
Zhao (II)					X		X
	Random Forest	Region-based	Edge-based	Cell Automata	EM algorithm	AIN	ICA
Buendia						X	
Festa	X						
Guo		X	X				
Hamamci				X			
Menze (I)	X		X				
Menze (II)					X		
Reza	X						

Riklin-Rav.			X				
Tustison	X						X
Zikic	X						

Table 2.6. Different methods that each approach use to find tumor boundaries. Abbreviations: CRF = Conditional Random Field, HMRF = Hidden Markov Random Field, EM = Expectation Maximization, AIN = Artificial Immune Network and ICA = Independent Component Analysis. References to these publications can be found in Table 2.4.

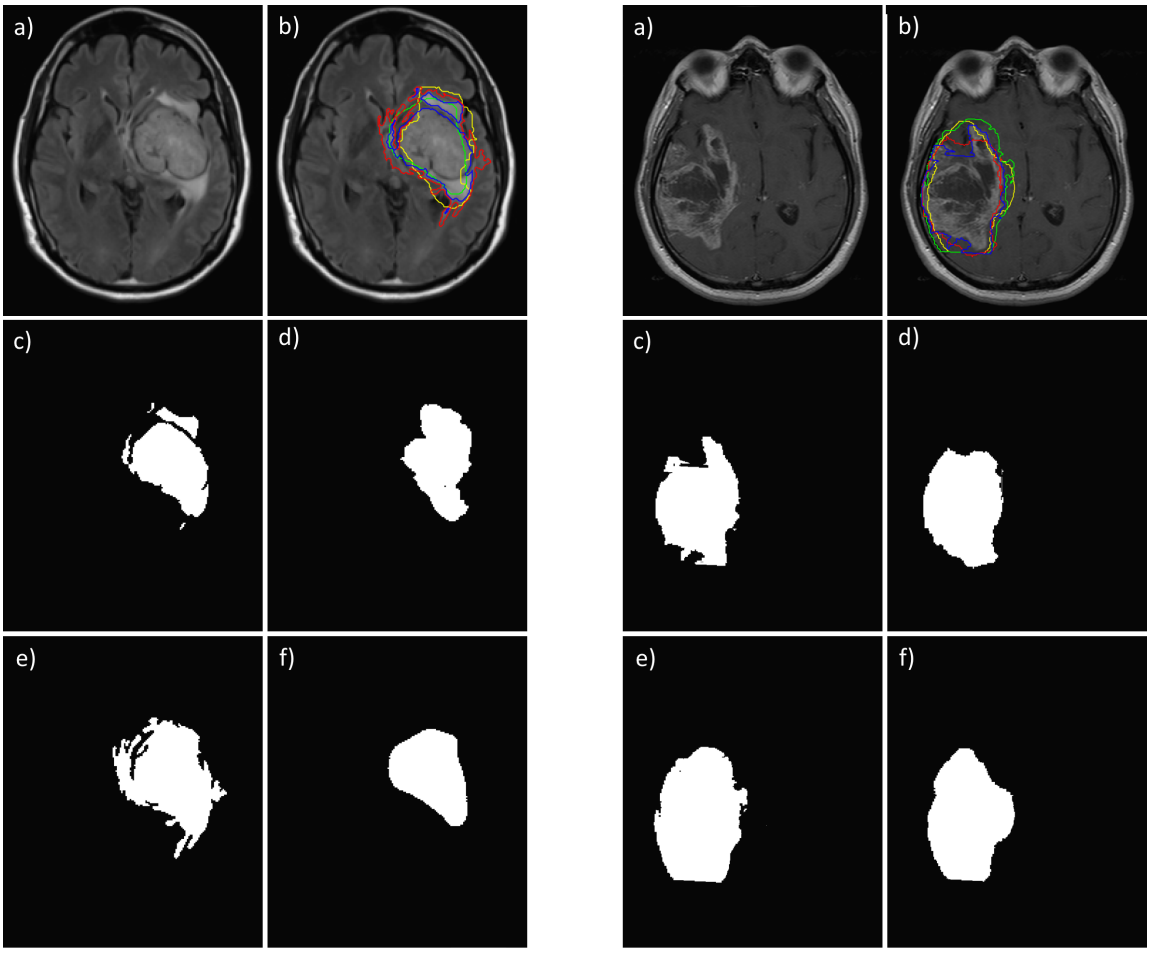


Figure 2.6. Example of different ROIs generated by different methodologies for the same object. Columns on the left show a T2-weighted FLAIR study and columns on the right show a T1 with contrasts study. Images a) and b) show the input images with and without the segmentation results, respectively. The binary images c)-f) show the individual results of the segmentation methods.

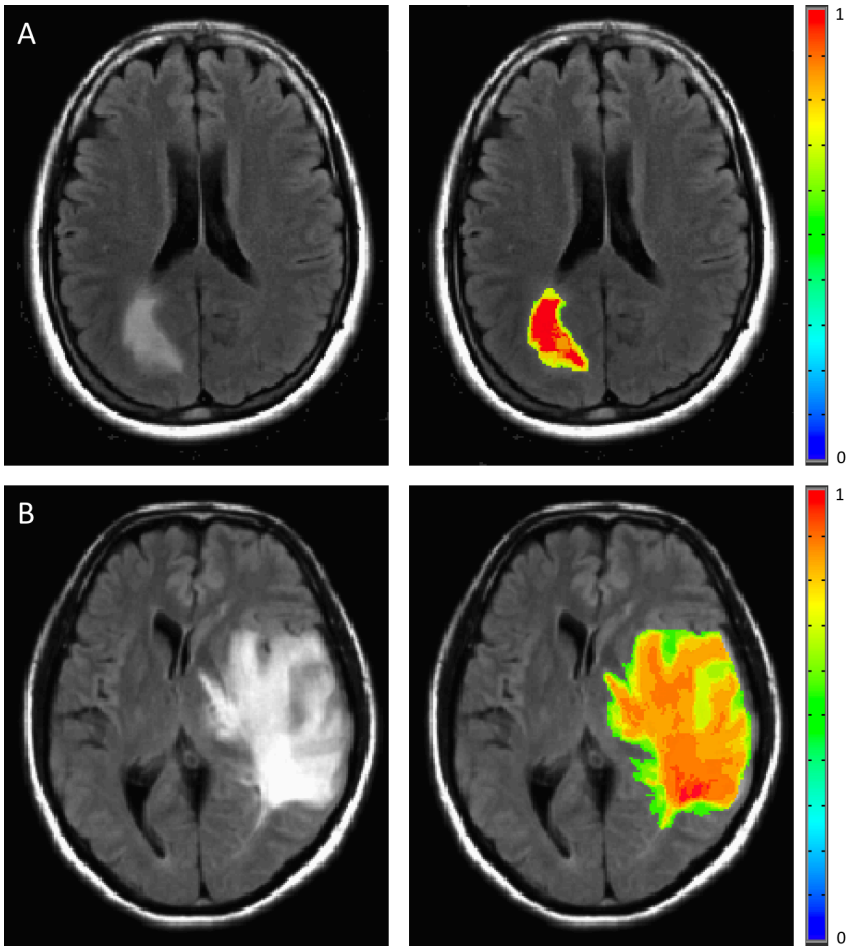


Figure 2.7. Example of a possible tumor likelihood distribution map generated using different segmentation methods to represent variability on a FLAIR MRI. Row A shows a small tumor with seemingly clear margins (left) and the different margins found by. Row B shows a bigger tumor with more complex margins that result in less agreement among the segmentation approaches due to increased tumor variability in hyper-intensity ranges and shape. The color bar indicates the agreement among segmentation results as a probability estimate.

Author	Input imaging features					
	Pixel Intensity	Pixel neighbor-hood	Texture patterns	Edge detec-tion	Image sym-metry	Atlas-based la-bels
Bauer	X	X	X	X		
Buendia	X		X			
Cordier	X	X				X
Doyle	X	X	X			
Festa	X	X	X		X	
Geremia	X					
Guo	X			X		
Hamamci	X	X				
Meier	X		X	X	X	X
Menze (II)	X		X			
Menze (I)	X					X
Reza	X		X			
Riklin-Raviv	X			X		
Shin	X					X
Shubbanna	X	X		X		
Taylor	X	X				
Tustison	X	X			X	X
Zhao (I)	X	X				
Zhao (II)	X	X				
Zikic	X		X			X

Table 2.7. Summary of the types of imaging features that are used by the tumor segmentation approaches described by the tumor MICCAI tumor segmentation competition [18]. For this table: Pixel intensity refers to features dependent on the intensity properties observed by a pixel or group of pixels. It is directly correlated with the type of tissue and with the acquisition parameters used while on the scanner. Pixel neighborhood information refers to distance metrics from a given pixel or image patch to its neighbors or a reference area. Textural patterns refer to first order image features such as mean, variance, and co-occurrence matrices. Edge detection refers to the use of filters and image gradients to find the image boundary. Image symmetry refers to comparisons on different axes to identify local differences under the assumption that both sides should be similar in anatomically. Atlas-based labels refer to the use of previously developed maps that integrate knowledge about a standardized location of cerebral structures. References to these publications can be found in Table 2.4.

2.5 Statistical metrics to evaluate segmentation performance

To obtain objective and accurate measurements of how good a given method performs compared to current standards (*i.e.,* reference standard), different statistical evaluations are often utilized. These metrics that can help verify if the results from a proposed segmentation method are close to the measurements obtained manually by trained radiologists, allowing a quantitative comparison, and helping identify cases or situations where improvements need to be made. Some of the most common metrics include:

- The Dice coefficient: A measure frequently used in image processing to compare binary images, and is the most used metric for validation of segmentation methods [40]. It is calculated by the intersection ("AND" operator) of the input images

divided by the total number of pixels contained inside each mask according to the following expression:

$$Dice = \frac{2\,|A \cap B|}{\lceil A \rceil + \lceil B \rceil} = \frac{2\,TP}{2TP + FP + FN} = \frac{2 * Precision * Recall}{Precision + Recall}$$

where A and B represent the two input binary images (in this case the reference segmentation and the automated result generated by an algorithm). TP refers to the true positives. FP refers to false positives, and FN to false negatives.

- Hausdorff distance: Refers to the maximum spatial distance (in voxels) between two finite points between input volumes A and B. It is defined as follows [40]:

$$HD(A,B) = \max\big(h(A,B), h(B,A)\big)$$

where h(A,B) is called the directed Hausdorff distance and is defined as:

$$h(A,B) = \max_{a \in A} \min_{b \in B} \|a - b\|$$

and $||\ a - b\ ||$ is the norm between these two points (*e.g.,* Euclidean distance).

- Jaccard index: Is an overlap metric defined as the intersection between the two images, divided by their union. It is defined as [40]:

$$Jaccard = \frac{|A \cap B|}{\lceil A\ \cup\ B \rceil} = \frac{TP}{TP + FP + FN}$$

The Jaccard index is similar to the Dice coefficient, one can be transformed into the other as follows:

$$Jaccard = \frac{DICE}{2 - DICE}$$

- Precision: This measure refers to the fraction of selected pixels that are relevant. It is also called positive predictive value and is defined as:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- Recall: This measure refers to the fraction of relevant pixels that are retrieved with respect to the total number of relevant pixels. It is also called sensitivity and is defined as:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Although these metrics have the goal of evaluating the performance of a test segmentation against a reference standard, they have different means to achieve this purpose. The Dice coefficient is, by general consensus [40], the most used metric in research papers on image segmentation. The Jaccard coefficient is a similar metric to the Dice coefficient in that it also measures overlap and thus can be transformed into the Dice coefficient. The Hausdorff distance measures the maximum distance between the pixels from the two images and thus is different from the overlap metrics (Dice and Jaccard). It is useful to determine the outliers of the results produced by the system. This distance will be significantly large when a region far away from the target is identified as a lesion (false positive). Finally, precision and recall are common statistics used broadly in informatics and computer science. As defined above, these metrics combine to give information of

how much relevant information is left out from the final result, and how much irrelevant information was included as part of the results (false positives and false negatives). This provides insights into how well a system is trained (loose-fit vs over-fit) and suggests the next steps needed to improve the performance of the system.

Although the Dice coefficient is the standard metric for evaluation of segmentation methods and many other applications in image analysis and is the standard metric for replicability studies, it also has some limitations. Some of them include the limited insights about outliers that can be detected by distance-based metrics such as the Hausdorff metric, the consideration of similarity that occurs by chance (Kappa coefficient), and that this score does not capture other sources of variability between the regions other than overlap (e.g., variation of intensity) given that it only considers binary inputs.

Chapter 3: Development of tumor measurement variability maps

3.1 Overview

This elaborates on the relevance of accurate segmentation of diagnostic medical images for the delineation of ROIs and the generation of meaningful quantitative features. Problems in image segmentation that arise from the non-uniform appearance of tissues and other factors such as differences in image acquisition hardware and protocols are discussed. This chapter explores the methods for determination of variability metrics in regions of interest, and how differences in algorithms and input parameters impact segmentation performance in brain imaging studies. Different approaches that exist for the determination of the tumor boundary can be optimized at the algorithm level to improve results that can be generalized to brain tumor segmentation.

We developed a framework that evaluates the robustness and reliability of different segmentation approaches by systematically evaluating the performance of each method through a range of input parameters. This process adaptively determined the range of parameters to test, and then optimizes the selection of parameters.

The following subsections include the characterization of the datasets utilized for this work, followed in Section 3.5 by a review of the different segmentation methods to

provide intuition about the influence of each parameter. Section 3.6 describes the optimization framework and parameter-selection process as well as the results obtained for the different algorithms. Finally, a discussion of key findings is provided in Section 3.7.

3.2 Input datasets

To achieve a robust characterization of and quantification of variability and segmentation error, a large dataset of GBM imaging studies was utilized in the experiments described in this chapter and the rest of the dissertation. This multimodal imaging data was provided from two different sources: a set of 300 subjects from the UCLA Neuro-Oncology Program (through the UCLA Brain Tumor Imaging Laboratory), and a set of 220 subjects from The Cancer Genome Atlas (TCGA) [41] curated by the Medical Imaging Computing and Computer Assisted Interventions society (MICCAI). This data from the TCGA/MICCAI is a subset of the larger TCGA dataset that was manually annotated as part of the MICCAI 2016 tumor segmentation challenge [18, 42].

The dataset characteristics for the input datasets are provided in Table 3.1, with a more specific characterization of sequence parameters shown in Table 3.2. Standard acquisition parameters are presented in Table 3.3 for reference.

The other dataset used as a training set for the system is the publicly available data from the TCGA/MICCAI brain tumor segmentation challenge [18]. This dataset

consists of 220 multimodal MRI scans (resolution of 240 x 240 x 154 pixels) as well as manually generated contours of the brain tumor region (whole tumor and sub-components). The available imaging sequences include T1-weighted, T1-weighted with contrast enhancement, T2-weighted, and FLAIR and were acquired according to standard parameters described in Table 3.3.

Element	Description	
	UCLA	TCGA/MICCAI
Total number of cases	300	220
Number of follow-ups	18 ± 13 (minimum 2, maximum 17)	Only baseline scan was available
Sequence availability	T1+C: 90% T1: 90% FLAIR: 70% T2: 90%	T1+C: 100% T1: 100% FLAIR: 100% T2: 100%
Field Strength	1.5T = 88% 3.0T = 12%	No information available
Reference standard	Available for total tumor	Available for total tumor and the different sub-components

Table 3.1. General information about UCLA and TCGA datasets.

	T1+C	FLAIR	T2	T1
Slice thickness [mm]	3.24 ± .92	3.99 ± .90	3.42 ± .78	3.36 ± .88
Echo time [ms]	12 ± 04	111 ± 16	82 ± 25	12 ± 6
Repetition time [ms]	638 ± 298	8345 ± 853	4202 ± 768	630 ± 341
Inversion time [ms]	948 ± 245	2228 ± 207	N/A	961 ± 242
Frequent sequence example	MPRAGE	FLAIR	TSE	MPRAGE

Table 3.2. Parameter characterization of input MRI data (UCLA dataset). It was not possible to obtain acquisition parameters information for the TCGA/MICCAI dataset (as DICOM header was removed from the data) but was assumed to be similar to the standard parameters for brain tumors.

	T1+C	FLAIR	T2	T1
Slice thickness [mm]	1	3	3	1
Echo time [ms]	Minimum	100 – 140	80 – 120	Minimum
Repetition time [ms]	< 800	> 6000	> 2500	< 800
Inversion time [ms]	1100	2500	N/A	1100
Frequent sequence example	MPRAGE	TSE	TSE	MPRAGE

Table 3.3. References for the standard parameters utilized in brain tumor MRI structural sequences [43].

3.2.1 Creation of a reference standard

Two sources of reference segmentations were used for two sources of data utilized in this work.

3.2.1.1 MICCAI/TCGA

The MICCAI/TCGA dataset consists of a set of newly diagnosed preoperative brain tumor cases with multi-parametric MRI, including T1-weighted (axial 2D acquisition with 1-6 mm slice thickness), T2weighted (axial 2D acquisition with 2–6 mm slice thickness), Fluid Attenuated Inversion Recovery [FLAIR (axial 2D acquisition with 2–6 mm slice thickness) and post-contrast T1-weighted images (3D acquisition using gadolinium and 1mm isotropic voxel size) for all subjects. The reference standard was defined based on the manual tumor segmentations generated for each case by up to seven raters. The final segmentation was assigned by consensus [18,42].

3.2.1.2 UCLA

Reference standards for tumor masks are generated manually by trained experts using a semi-automated approach that included intensity thresholding. The overall process included generation of a manual region of interest, followed by a user-defined threshold applied on the three-dimensional volume, and a final manual correction to ensure the resulting ROI was correct. This process was performed using AFNI software [44]. Although these annotations were generated by trained professionals, there is vari-

ability associated with the different masks that are produced, which impacts the assessment of the automated approaches and evaluation of the tumor-measurement variability maps (confidence maps).

To measure the variability associated to these annotations, a sample of 50 cases (sample determined as detailed in Table 3.5) was contoured by three different experts, and each rater segmented the same tumor twice to generate independent measurements. The 50 cases included 15 small tumors (≈3k to 9k μL), 18 medium-sized tumors (≈25k to 33k μL) and 18 large tumors (≈90k to 150k μL). Subsequently, an inter- and intra-rater correlation using a kappa statistic (with a probability of chance agreement of zero) obtained using the Dice similarity coefficient was performed for all the combinations between raters as well as against the same rater.

The intra-rater agreement refers to how well the same person agrees with his/her own results when doing repeated measures of the same object, while inter-rater agreement refers to the comparison between measures of the same object obtained by different raters. The values of these two metrics will be high (close to one hundred percent) when there is little variation between the results of each measure and will decrease as the difference between the two inputs increases.

Also, a similar experiment was performed to evaluate the variability in the regions of the tumor mass where the most variability is found (*e.g.*, the edges). To evaluate this, the same algorithm to evaluate variability was performed on the manual tumor masks

with the difference that the region of the tumor where all raters agreed was removed (a logical AND operation between the tumor masks). The result of both experiments is presented in Table 3.4.

Metric	Rater agreement
Inter-rater agreement (overall)	91% ± 2%
Intra-rater agreement (overall)	93% ± 1%
Intra-rater agreement (rater 1)	93% ± 3%
Intra-rater agreement (rater 2)	95% ± 1%
Intra-rater agreement (rater 3)	93% ± 3%
Inter-rater agreement (edges only)	55% ± 8%
Intra-rater agreement (edges-only)	60% ± 7%

Table 3.4. Intra- and inter-rater agreement on a sample set of brain tumor manual segmentations (complete tumor mask and only-edges tumor masks).

3.2.2 Sample size calculations

Throughout the different experiments and trials made in the subsequent subsections, a minimum sample size was defined to ensure the results have sufficient statistical relevance (power). These calculations were performed using a power analysis of one proportion as it helps to assess whether a population proportion is significantly different from hypothesized value. For the purposes of the experiments performed in this dissertation and under the general hypothesis that attempts to compare where a given tumor

segmentation is similar or not to a manual standard generated by the clinician, we aimed to have less than 15% error rate (80% power) and significance level of 0.05, since this margin of error is based on clinical standards for GBM evaluation [6,45]. Finally, the statistical tests that aid calculation of the appropriate sample size includes tests of non-inferiority and Receiver Operating Characteristic (ROC) curve power analysis. The former is used to determine if automated segmentation can produce results that are not inferior to the ones produced by the reference method (manual expert segmentation). The latter is performed to evaluate the performance of a diagnostic test -clinical assessment by a neuroradiologist - in terms of its false-positive and true-positive ratios with respect to the reference clinical assessment which was assumed to be correct [46]. The specific results obtained for the different experiments and datasets are summarized in Table 3.5.

Source	Total size (# of cases)	Subset (# of cases)	Purpose
UCLA	300	≥ 233[a]	- Validation of knowledge-based segmentation method (Chapter 5). The reported results include a test on 260 subjects. - A secondary sample of 31[c] cases was utilized for clinical evaluation (Chapter 6)
		≥ 50[b]	- Used for validation of optimization results using segmentation algorithms (Chapter 3)
TCGA /MICCAI	220	≥ 50[b]	- Used for the main optimization experiment using segmentation algorithms (Chapter 3) - This sample was expanded to 84 while attempting to balance the dataset in terms of tumor sizes and border properties (Chapter 3)
		136	- Remaining cases that were used for training the multimodal segmentation approach developed in Chapter 5. Considering that only the TCGA/MICCAI dataset contained labels for the different tumor components.

Table 3.5. Sub-divisions of input datasets and statistical considerations regarding sample size for the different studies and experiments described in Chapters 3, 4 and 5. Superscript "a" indicates a power analysis test of non-inferiority of one mean (t-test). Superscript "b" indicates a power analysis of one proportion. Superscript "c" indicates a ROC curve power analysis test. All tests were performed aiming to obtain 80% power, with a 15% error range and a significance level of .05.

3.3 Data preprocessing

Before the optimization and segmentation processes, a series of pre-processing steps encompassing image registration, intensity normalization, image denoising, and removal of non-cerebral tissues (skull-stripping) took place. The following steps were implemented to reduce possible sources of error such as motion, intensity anomalies, and removal of unwanted skull tissue:

- Image registration. Image registration of all selected sequences to a normalized space (*i.e.,* MNI_152_1mm atlas) was accomplished using the FMRIB Linear Image Registration Tool (FLIRT) [46-48], a fully automated, robust, and accurate tool for linear (affine) intra- and inter-modal brain image registration. Registration helps to correct variations in the different scans (rotations, translations, slight deformations, etc.) and allows the visualization and measurement of tumor subcomponents in all MRI sequences, for instance visualizing areas of edema in post-contrast T1-weighted images.
- Image denoising and bias correction. Image correction was done using a noise reduction filter and the bias correction tool implemented in the Statistical Parametric Mapping (SPM) toolbox [49]. MR images often are corrupted by a smooth, spatially varying intensity modulation (bias). These artifacts, although not usually a problem for manual visual inspection, can impede automated image processing,

affecting image intensity distribution and textural patterns. The SPM algorithm models the bias field as a multiplicative n-dimensional random vector with zero mean Gaussian prior probability density, allowing the bias field to be treated as an additive artifact, calculated from the maximum intensity projection [50].

- Intensity normalization. The pixel values of the input files are adjusted to modify the original histogram distribution to a normalized scale. Normalization is used often when comparing across subjects (or time), and when the pixel intensity itself is not as relevant for the analysis compared to the interactions with the other elements or the location on image grid [49]. This adjustment in the intensity distribution was performed using the non-parametric non-uniform intensity normalization (N3) algorithm, a popular method for MRI processing [51].
- Skull stripping. Intracranial segmentation (or skull-stripping) removes non-cerebral tissues such as skull, eyeballs, and skin. Out of the several options tested (*e.g.*, BET, SPM, 3DSkullStrip), the 3DSkullStrip approach was selected. This automatic tool is part of the AFNI package [44] and is a modified version of BET [52]. It uses the spherical surface expansion paradigm and includes adjustments to avoid eyes and ventricles, reduce leakage into the skull and use data outside the surface to guide the surface evolution. Even though the registration step partially allows for the use of a default brain mask, this step helps to obtain better

overall results including better-defined brain margins. and use cases with significant deformities that are harder to fit with a general brain mask such as tumor recurrence cases with resection cavities.

3.3.1 Evaluation of Registration Error

Although the main objective of this project is to evaluate the variability observed when using different approaches to measure the tumor boundary and evaluate the clinical implications of tumor segmentation variability, the processes that take place before the segmentation (*e.g.,* image registration) are also susceptible to variation and error (Table 2.8).

An experiment was performed to evaluate registration error influences on the tumor boundary obtained from the segmentation process. The overall process is illustrated in Figure 3.1. An exploratory random sample of ten subjects from the UCLA dataset (five FLAIR and five T1 with contrast enhancement volumes) were selected to undergo image registration [47] using different parameter combinations, obtaining a total of 80 different combinations per subject (Table 3.6). After all the registration results were obtained (a total of 800 different results), a tumor mask for each of these volumes was generated and, by using the inverse transformation matrix obtained at the registration step, the tumor masks were transformed back into the original space before the registration occurred.

After the masks were warped to the original space, the similarity between those results and an initial reference was measured using the Dice coefficient. Given that the segmentations were verified to be accurate in each case, the differences observed in the similarity evaluation were assumed to be due to the different registration parameters used for each case.

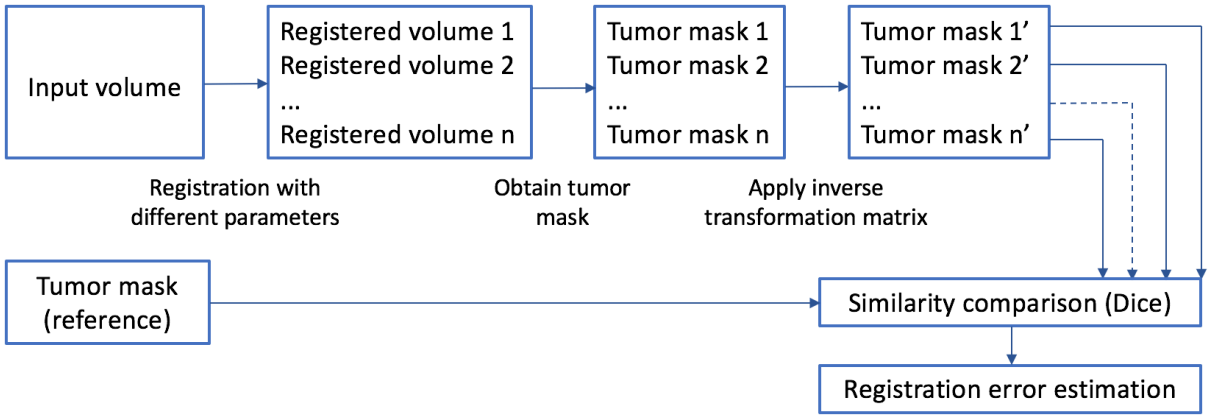


Figure 3.1. The overall process followed for the evaluation of registration variability. The input volumes are registered using different registration parameters, then a tumor mask is obtained for each case and transformed to the original image space (tumor mask'). Finally, these tumor regions were compared with a reference standard to obtain variation metrics. The differences observed from each tumor mask with respect to the reference indicates the error introduced by the registration process.

Parameter	Parameter options
Degrees of freedom (model)	12 (affine), 9 (traditional), 7 (global rescale), 6 (rigid body).
Cost function	Correlation ratio, mutual information, normalized mutual information, normalized correlation and least squares.
Interpolation	Tri-linear, nearest neighbor, spline, sinc.

Table 3.6. Different parameters for image registration. These parameters control how the registration algorithm evaluated the goodness of fit for a certain solution, the degree of deviation the different images can have between each other (similarity), and the method used to produce the missing locations on the image volume when resampling to the new image space.

This evaluation using Dice coefficients helped characterize the in-plane variation (observed as displacement on the x- or y-axis) measured as the overlap of the ROIs on each slice, and the through-plane variation (slight shifts on the z-axis) as the similarity is evaluated in three-dimensions.

The average similarity observed overall in these different experiments was 94.9%; therefore, the measured variability due to image registration is 5 ± 1%. The variability observed only for the FLAIR sequences was 5 ± 3% and for post-contrast T1 6 ± 2%. Although there is a measurable amount of error added from the registration process of the imaging volumes, the amount of variation observed in these experiments was less than 15%, defined according to clinical guidelines as discussed in Section 3.2.2. Additionally, the registration step is not a component of the segmentation methodologies as

they do not require spatially aligned images as input. Registration error was not considered for the remainder of this work, although other considerations about the role of registration in the imaging pipeline are further discussed in Chapter 6.

3.4 Experiment set-up

Distinguishing tumor from surrounding normal tissue is frequently challenging in imaging studies given tumor variability, discontinuous borders and, fuzzy tissue boundaries. As a result, both computational approaches and even experienced radiologists can have difficulty achieving consistent and accurate segmentation and classification results [7, 53]. As a consequence, if automated segmentation is to yield reliable boundaries from which meaningful quantitative features can be extracted, the choice of algorithm and input parameters needs to be adapted for specific types of images and segmentation tasks [16-54].

This study presents a framework to discover the optimal segmentation method and parameters to use given a set of input images with objects of interest with non-uniform boundaries. This evaluation was accomplished by performing a sensitivity analysis for each algorithm, finding the parameter ranges that achieve the most accurate and consistent segmentation result using the Dice similarity coefficient compared to the manual labeling of the input data.

The process described in this section is illustrated in Figure 3.2. The development of this framework included a sample of 84 subjects. This subset of the TCGA/MICCAI dataset was selected based on the statistical power calculations shown in Table 3.5. This dataset was utilized given that a manual reference standard for all different tumor sub-components was available. This sample reflects a range of tumor sizes (from 9.3 cm^3 to 388.9 cm^3 total tumor volume, including edema, and morphological characteristics (smooth, irregular margins). Cases were split into 80% for training and 20% for testing.

The input images are preprocessed as described in the previous section and segmented using different existing algorithms. Each of the segmentation methods was evaluated and optimized on the annotated training examples, and compared based on the degree of variability observed in the segmentation results given the utilized parameters.

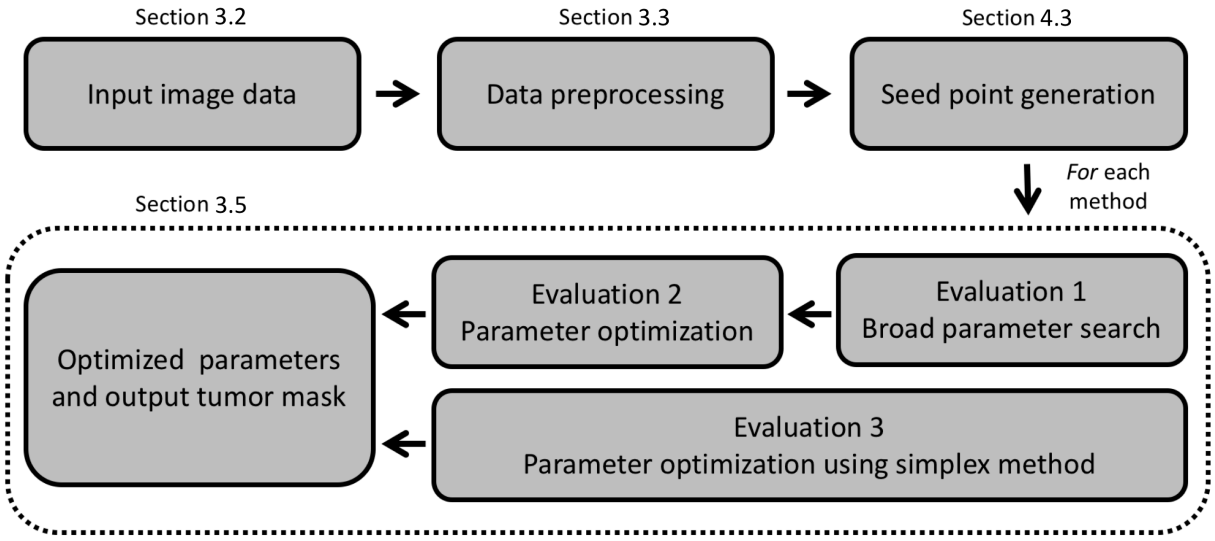


Figure 3.2 Overall steps in the pipeline processing scheme from the data input to the generation of the final tumor mask, and the optimal combination of segmentation parameters. The dotted line indicates a repeated process (*i.e.,* a *for* loop).

Evaluations 1 and 2 are iterative steps in which different parameter values are fully evaluated. Evaluation 3 uses a popular optimization algorithm to provide an alternative to the complete parameter search. Comparison of results and processing time is commented on in the discussion section. Finally, a five-fold cross-validation was used to obtain the optimized parameters for each method.

The contribution of this section is a generalizable process for identifying the segmentation approach that produces accurate and consistent tumor boundaries across the input dataset. As with all supervised learning methods, this framework requires an initial set of reference annotations from which the optimal algorithm and set of parameters can be determined; nevertheless, the process of discovering the optimal approach is fully automated, including the definition of a seed point/region [55] and the selection of which parameter values to test.

3.5 Evaluation of variability in different segmentation approaches

For this section, four different types of segmentation methods were evaluated: region-based, edge-based, level-set, and multilevel segmentation approaches. These methods have been applied in a variety of medical image-analysis tasks including brain tumor segmentation [56-62]. Besides their widespread use, these methods were selected

because each has a set of parameters that employ different imaging aspects to modify the algorithm behavior and influence the output ROI. In the following sub-sections, we briefly describe these selected methods.

3.5.1 Region-based Segmentation

This method detects the tumor boundaries based on energy, area, and length measurements to maximize the difference between the inside and outside of an object [63]. During the original development of this algorithm, an initial ROI defined as a bounding box is required. The algorithm then evolves the ROI guided by calculating an energy metric (minimal partition problem) [64] based on the gradient of the objects inside the initial ROI. This process was repeated several times (curve evolution) using the "mean-curvature flow evolution method."

This edge detection method relies on minimizing the internal and external energy found inside the polygon, which is seen in practice as the ROI moving inward or outward according to the algorithm parameters α, β, and λ (smoothness, rigidness, and object-attracting force, respectively) [65]. The energy minimization considers the image (u_0) average intensity values on the regions contained inside and outside the contour C, as defined in the equation below. The algorithm converges when the average intensity between both sides is the most different and the individual regions are homogeneous. Then

the following fitting term is applied [64]:

$$E(C, c_1, c_2) = \mu \cdot Length(C) + \tau \cdot Area\big(inside(C)\big) + \lambda_1 \int_{inside(C)} |u_0(x,y) - c_1|^2 dx\, dy + \lambda_2 \int_{outside(C)} |u_0(x,y) - c_2|^2 dx\, dy$$

where the energy functional E is the calculated energy value for any C curve at any iteration, c_1 and c_2 are the average intensity values located inside and outside of the current contour, and λ_1 and λ_2 are regularization terms. Also, μ and τ are regularization terms length and area respectively. If the average intensity of the object is very similar to the background in the object region, the method will have more trouble finding the correct boundary.

3.5.2 Edge-based segmentation

This technique is based on the minimization of the distance measured in geodesic maps, formed according to intrinsic geometric image measurements [66-68]. It also considers a measure of image energy as the contour evolves that is calculated in a distinct manner compared to the previous approach. In general, this approach finds the curve that minimizes the measured geodesic distance (the local shortest paths between points in space) as well as the curve energy. This energy (active contour component) helps to locate the boundaries that are defined on the gradient originating from the image, and the geometric aspect allows for stable boundary detection when their gradients suffer

from large variations, also allowing it to divide or merge as needed. The model for boundary detection is defined in the following way [68-69]:

$$\frac{\partial \alpha}{\partial t} = g(I)|\nabla \alpha| \text{div}\left(\frac{\nabla \alpha}{|\nabla \alpha|}\right) + \varepsilon g(I)|\nabla \alpha| = g(I)(\varepsilon + \kappa)|\nabla \alpha|$$

where α is defined as $\alpha = (\varepsilon+\kappa)|\nabla \alpha|$, and ε changes in function of $\varepsilon = (E(t) / N) - \kappa$, where $E(t)$ is known as the Euclidean heat flow, N is the normal inward to the curve and κ is known as the Euclidean curvature; $\nabla \alpha$ attracts the curve to the boundaries of the object, and $g(I)$ is the stopping function defined by the force of the external regions of the image. The main goal of $g(I)$ is to stop evolving the curve when it arrives at the object's boundaries. The curve stops evolving when the measured distance equals zero or when the maximum number of iterations is reached.

The curve evolution in this method is independent of the image topology, meaning that there is no need to know the structure of the solution to be able to run this method. It also allows detection of objects in the image without requiring information about their number or shape characteristics. It can also handle situations where the boundary has high gradient variations or small gaps (frequent in real images) [70-71].

3.5.3 Distance Regularized Level-Set Evolution

The distance regularized level-set evolution method (DRLSE) is a contour evolving algorithm that allows for cusps, corners, and automatic topological changes. It starts with an initial ROI in the input image, and it changes automatically to adjust to the optimal boundary of the object of interest. Different from the standard level-set method, the DRLSE does not need a reinitialization step throughout the process [72-74]. This method was formulated as a gradient flow that minimizes an energy measurement with the addition of a distance regularization term, and an external energy term that drives the motion of the initial level set toward the locations of the edges. The distance regularization term has a potential function (a signed distance profile near its zero-level set) that aides the level-set function to reach its minimum points, stabilizing the shape of the level-set function.

The mechanism used by the DRLSE method to maintain numerical stability without reinitialization is described as an energy minimization with distance regularization, $\xi(\phi)$, as follows [74]:

$$\xi(\phi) = \sigma R_p(\phi) + \xi_{ext}(\phi)$$

where $R_p(\phi)$ is the level set regularization term; σ is a positive constant (similar to a weighting factor of the regularization term), and ξ_{ext} is the external energy that reaches

a minimum when the level set is located on the object boundary or edge. The overall role of $R_p(\phi)$ is to maintain the distance property of the level-set function, defined as:

$$R_p(\phi) = \int_{\Omega} p(|\nabla\phi|)dx$$

where p represents the energy density function or potential function (p>0), this function often has a smoothing effect but also maintains the signed distance property ($\nabla\phi$) that ensures an accurate computation during the curve evolution.

3.5.4 Multilevel Segmentation

This approach is based on segmentation by the weighted aggregation (SWA) algorithm [75], which is a fast and effective approach that uses the concept normalized cuts [76] to achieve image segmentation. It also includes the use of a class of models using a Bayesian approach to represent the different characteristics of the object of interest [77].

The SWA algorithm aims to extract multi-scale segments that are classified using a decision tree algorithm based on the calculation of affinities between regions of the image. The incorporation of a model for classification develops a modified SWA algorithm that can use class-based probabilities that integrates the classes of the model with the affinity measurements

To define the weighted aggregation segmentation, a graph with different levels (pyramidal approach) is used to represent a multimodal intensity vector for every voxel in the image. In this pyramidal model, the finest layer was produced by taking every pixel in the lattice individually, and then the next levels are calculated by evaluating the affinity of each pixel using a six-neighbor connectivity scheme. The node affinities are calculated as a saliency coefficient [78], which is defined as:

$$\Gamma(R) = \frac{\sum_{u \in R} v \notin R^{w_{uv}}}{\sum_{u,v \in R} R^{w_{uv}}}$$

where Γ is the saliency coefficient, u and v are the nodes defined in the image R and w is the weight that represents the affinity between two nodes. This affinity w is normally defined as $w = exp(-D(s_u, s_v, \theta))$, where D is a non-negative distance and theta is a predetermined parameter that controls the saliency parameter, and s_u and s_v denote the current property or statistic being measured at graph nodes u and v.

Subsequently, the SWA algorithm proceeds by iteratively producing the different levels by reducing the complexity of the image using the following approach:

Pseudocode 3.1. Multilevel segmentation of brain images

1. **Initialize** at time t_0 preliminary grid on finest level
2. **Define** nodes defined by the voxel neighbors
3. **Choose** a representative set of nodes such that $\sum_{v \in R} w_{uv} \geq \beta \sum_{v \in V} w_{uv}$
4. **Define** graph $G^{t+1} = (V^{t+1}, \varepsilon^{t+1})$:
    a. **Compute** interpolation weights $P = {}^{w_{uU}}/_{\sum_{v \in v^t} w_{uV}}$
    b. **Accumulate** statistics to current coarse level and interpolate affinity from finer level
    c. **Use** coarse affinity to modulate the interpolated affinity
    $w = \exp(-D(su, sv, \theta))$
    d. **Create** an edge ε^{t+1} between $U \neq V \in V^{t+1}$ when $W_{uv} \neq 0$.
5. **Increase** t to t+1
6. **Repeat** steps 2 through 4 until V =1 or ε=0

The parameter β in step two governs the amount of coarsening that occurs at each layer in the graph (empirically defined as 0.2). During step three, the affinity parameter *w* includes the finer level (scale) affinities which are interpolated to the more general level; a similar process is followed for all the regions of the image.

Finally, to extract the object of interest from the pyramid, a class likelihood function is computed using the Bayesian affinity model. For each voxel, the most likely class is computed at each level in the pyramid. Node memberships for the more general pyramid levels are calculated by the SWA interpolation weights so that at the end every voxel is

labeled as the class for which it is part of at the biggest number of levels along the pyramid. Examples of results generated by this algorithm can be observed in Figure 3.8.

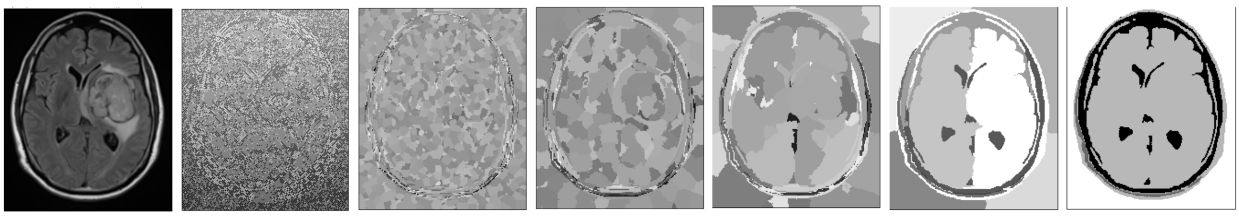


Figure 3.8. Example of the output of the multilevel segmentation algorithm. The different levels and clusters in each of them are dependent on the affinities that are calculated among pixels and the input parameters to the algorithm

3.6 Methods and results

3.6.1 Parameter Evaluation

The robustness of each segmentation method was assessed by systematically evaluating the different parameter combinations for each algorithm, and then finding the optimal parameter values. The optimization process proceeded in three parts. First, a broad parameter search to identify the plausible range for each parameter, considering the influence of the parameter on the final ROI (Evaluation 1). Second, a more granular search of the parameter space was conducted to determine the optimal value for each of the test parameters (Evaluation 2). Third, in parallel, we also explored an optimization algorithm as a faster alternative to find the optimal parameter values. We compared the

results and parameters identified using this method versus doing a full search (Evaluation 3).

These evaluations were performed separately in the T2-weighted input images (FLAIR and T2) and T1-weighted images (T1 with contrast enhancement) to assess differences in parameter optimization findings given the imaging contrast parameters. T2-weighted sequences are used often to measure total tumor extent with edema regions, while T1 with contrast is used to delineate enhancing and necrotic tumor regions (also known as core-tumor) [2,17].

To begin the optimization, the initial parameter values were set either to their default values defined in reference publications [64, 67, 68, 73, 74, 77] or to neutral non-contributing values according to the mathematical definition. Table 3.7 shows the parameters that were identified within each algorithm and were used in the described optimization experiments. It is worth noting that for the multilevel segmentation the number of levels was not a parameter to optimize. Since all the output levels were generated and tested when defining the output boundary, it is possible to check all levels for the cases as it is computationally efficient.

The resulting ROI from each segmentation test was compared to the reference standard using the Dice similarity coefficient. As previously discussed, the Dice coefficient was used as it provides a direct comparison between the automatic and ground

truth segmentations, and normalizes the number of true positives to the average size of the two segmented areas.

Finally, the input dataset of 84 subjects was stratified based on tumor volume and tumor boundary characteristics. The categorization by size included a division of input cases in three equally-sized groups based on the total tumor volume, including the region of edema. The mean tumor volume was 49.2 ± 20.1 cm^3 for small tumors, 98 ± 14.3 cm^3 for medium-size tumors, and 187.6 ± 63.7 cm^3 for large tumors. For the classification of tumor boundaries, two groups were defined: smooth and irregular. The groups were found by splitting the total number of subjects into two groups according to a surface-to-volume ratio calculation used as a proxy for surface smoothness (also known as sphericity) [79]. A threshold value of 0.63 was defined in order to have equally sized groups (for both T1- and T2-weighted sequences). To evaluate for statistically significant differences in segmentation performance between the different methods and the different subgroups, we used a one-way analysis of variance (ANOVA) test.

Algorithm	Parameter	Description
Region-based segmentation	Foreground weight	Weight inside contour currently being evolved.
	Background weight	Weight outside contour currently being evolved.
	Smooth factor	Gaussian filter (smoothing filter) that affects the details observed on edges.
	Contraction bias	A coefficient that biases the evolving contour towards a shrinking or expanding behavior.
Geodesic segmentation	Advection weight	Controls the sensitivity to the local gradient direction. It is a component of the calculation of the speed or rate of change parameter.
	Exponent	Exponent used in the calculation of the stopping function.
	Smooth factor	Gaussian filter (smoothing filter) that affects the details observed on edges.
	Contraction bias	A coefficient that biases the evolving contour towards a shrinking or expanding behavior.
Distance regularized level set	Weighted length coefficient	Coefficient that weights the length of current contour to control curve evolution.
	Weighted area coefficient	Coefficient that weights the total area inside current contour to control curve evolution.
	Width of the Delta function	Refers to the alpha coefficient on delta function used while updating contour.
	Scale of Gaussian kernel	Controls the lambda coefficient on Gaussian kernel, used as smoothing factor.
	Cut salient segment coefficient (CSSC)	Binary indicator that controls the method to follow when dealing with the edges between the different

Multilevel segmentation		sections of the image (to cut-off the edges or to keep them).
	Fine affinity coefficient (FAC)	This coefficient has a parameter (defined as Theta on the affinity formula $w= \theta\|i_u\text{-}i_v\|$). Works as a weight and it usually takes a value of 10.
	Strongly connected coefficient (SCC)	Parameter referred as beta on the weighted aggregation segmentation. It is usually set as two.
	Interpolant cut coefficient (ICC)	The percentage of total volume an edge needs to have from a fine node to be connected to the coarser node. The higher this value, the fewer connections in the final result.

Table 3.7 A summary of the parameters that can be varied for each algorithm and a short description of the role of that variable on the behavior of the segmentation approach.

3.6.1.1 Evaluation 1: Broad parameter search

The initial evaluation was aimed to determine the range of permissible values for each parameter from which the optimal parameter can be found. The goal is to discover the range of input values that yield meaningful results that are not grossly over- or under-segmented. Initially, a coarse sampling of the parameter combinations was performed to establish a more constrained interval on which a more granular investigation of optimal parameter combinations could be performed. The initial interval was defined by iteratively testing different parameter values and constraining the interval according to the similarity values obtained using the Dice coefficient. If the Dice coefficient was found to

be 0 (*i.e.,* no overlap between the segmentation and the reference standard), that parameter value was assumed to be invalid.

Two examples are provided to illustrate the results of a broad parameter search. The contraction bias parameter in the region-based approach helps modulate the ROI evolution to shrink or expand more rapidly. While the parameter may be any value between -1 to 1 (*i.e.,* a positive or negative proportion), the broad parameter search further narrows this range from -0.5 to 0.35. Similarly, in the DSLRE segmentation method, the kernel width parameter can have a wide range of real values and its tuning directly affects how the ROI edges are defined. For this dataset, the broad parameter search narrowed this range from any integer to values ranging from 1 to 11.

3.6.1.2 Evaluation 2: Refined parameter search

After the broad parameter search identified the minimum and maximum values for each parameter to test, a refined search was performed using more granular steps. The selection of what additional parameters to test was made adaptively based on the magnitude of change in Dice similarity coefficients. Based on the prior results of tested parameters, additional values were selected according to the trend (gradient) observed.

The different parameter combinations were tested using permutations without repetition. Each parameter was tested and optimized one at a time, varying and evaluating the performance of one parameter at a time while leaving the others constant, but at the end testing all possible values. If the best performing parameter combination (the combination that achieved the highest Dice coefficient) included a parameter value that existed on the boundary, the initial interval was expanded to cover additional parameter values in that direction until the global optimum was found. The parameter value that achieved the optimal segmentation result varied across cases; hence, the selected parameter was determined to be the most frequent value of the different optimum values obtained across all training subjects.

3.6.1.3 Evaluation 3: Optimized parameter search

The simplex method [80] is a popular optimization method that searches through possible different solutions (*i.e.,* combinations) until the optimal feasible solution is obtained based on comparison with an objective function, in this case, the Dice coefficient as described in [81]. This method has been used on a variety of image processing approaches, with good performance and reduced computation time for each iteration [82]. Specifically, we used the dual-simplex approach, which represents an improvement of the original algorithm that produces more robust results. It adds more constraints to the

possible solution set and avoids possible infinite cycles that can occur in some cases. The objective function was to minimize the error observed in the segmentation methods (*i.e.,* similarity coefficient expressed in terms of true positives and true negatives). The input conditions were given by the different parameter combinations. The optimal set of parameters was discovered based on the optimal performance (minimum error) provided as output from the algorithm. Table 3.8 shows the results obtained for this evaluation. We also evaluated whether the simplex approach achieved the same results as the exhaustive search (Evaluations 1 and 2).

It is important to note that the Simplex algorithm was utilized as an example of one of the many optimization algorithms that have been developed. It was selected because of its interpretability, speed of calculation and also because of its good performance on other medical image analysis applications. Other approaches like gradient descent or other stochastic optimization methods can be utilized in this framework to obtain the same results.

	Parameter	FLAIR (M±SD)	T2 (M±SD)	T1+C (M±SD)	Simplex (M±SD)	Default (M±SD)
Region-based segmentation	Smooth factor	0 ± 0	0.6 ± 0	0 ± 0	0 ± 0	0 ± 0
	Contraction bias	0 ± 0	-.25 ± .1	0 ± 0	0 ± 0	0 ± 0
	Foreground weight	8 ± 0	8 ± 0	8 ± 1	6 ± 2	1 ± 0
	Background weight	4 ± 0	8 ± 1.7	4 ± 0	4 ± 0	1 ± 0
	Dice coefficient across all test sets	*.75 ±.05*	*.73 ±.02*	*.61 ±.08*	*.72 ±.03*	*.70 ±.01*
Edge-based segmentation	Smooth factor	1 ± 0	5 ± 0	1 ± 0	3 ± 0	1 ± 0
	Contraction bias	-0.25± .1	-0.25± .1	0.5 ± 0	-0.25 ± 0	.3 ± 0
	Advection weight	8 ± 1.7	8 ± 0	8 ± 0	5 ± 2.3	1 ± 0
	Exponent	1 ± 0	1 ± 0	1 ± 0	1 ± 0	1 ± 0
	Dice coefficient across all test sets	*.80 ±.01*	*.78 ±.02*	*.72 ±.05*	*.73 ±.07*	*.52 ±.02*
Distance-regularized level-set	Length weight	8 ± 0	8 ± 0	8 ± 0	8 ± 0	5 ± 0
	Area weight	-2 ± 0	-4 ± 1.6	-8 ± 1.6	-4 ± 0	-3 ± 0
	Delta width	1 ± 0	1 ± 0	5 ± 0	1.2 ± 0	1.5 ± 0
	Kernel width	1 ± 0	1 ± 0	1 ± 0	1 ± 0	8 ± 0
	Dice coefficient across all test sets	*.74 ±.27*	*.67 ±.18*	*.78 ±.30*	*.68 ±.32*	*.35 ±.16*

Multi-level segmentation	Fine affinity coeff. (FAC)	1 ± 0	1 ± 0	1 ± 0	1 ± 0	10 ± 0
	Strongly conn. coeff. (SCC)	0.3 ± 0.1	0.2 ± 0.1	0.5 ± 0.3	0.2 ± 0.2	0.2 ± 0
	Interpolant cut coeff. (ICC)	0.6 ± 0.1	0.4 ± 0.2	0.7 ± 0.1	0.6 ± 0.3	0.4 ± 0
	Cut salient coeff. (CSSC)	1e-6 ± 0	1e-6 ± 0	1e-6 ± 0	1e-6 ± 0	1e-6 ± 0
	Dice coefficient across all test sets	*.82 ±.07*	*.78 ±.16*	*.76 ±.21*	*.68 ±.17*	*.71 ±.23*

Table 3.8 Results from optimization and cross-validation experiments. Showing parameter ranges for each method on first four rows (mode ± standard deviation across all folds) and the associated Dice coefficient for each case on the fifth row. Results for FLAIR, T2, simplex, and default values are measured with respect to the total tumor (core tumor and edema region) and the results for T1+C only include the core tumor (enhancing and necrotic regions)

3.6.2 Performance across different MRI sequences

Evaluations and cross-validation were performed on all subjects, automatically testing 108,948 different combinations for each segmentation approach (region, edge, and level-set) and sequence (T1+C, FLAIR, T2).

The results in test cases represent an improvement compared to the results obtained using the default parameters, showing the largest improvement on the level-set-based algorithm and demonstrating that the optimization on methods parameters translated to a sensible improvement in segmentation performance (Table 3.8). The optimized

parameters found on the T2-weighted sequences (T2 and FLAIR) presented some similarity but were not the same, suggesting that other factors besides image intensity—such as overall tumor size or tumor border variability— may be influential to define whether a segmentation method is able to delineate the tumor boundary reliably.

The number of high-performing combinations defined as a Dice similarity coefficient > 0.9 across all methods ranges from 56 (level-set based) to 412 (edge-based), representing only 0.05% and 0.38% respectively of the parameter combinations that were tested. The mean Dice coefficient for region-, edge-, level-set-based and multilevel segmentation methods on the T2-weighted sequences were 0.772 (0.743–0.801; 95% CI), 0.809 (0.788–0.830), 0.803 (0.782–0.823), and 0.762 (0.722–0.834), respectively. In the same way, the mean Dice coefficient on the T1-weighted sequences were 0.538 (0.485–0.591), 0.619 (0.570–0.668), 0.630 (0.582–0.678), and 0.659 (0.614–0.704), respectively. These lower values on T1 sequences are believed to have occurred because the combination of hyper- and hypo-intense tumor components (enhancing and necrosis) contribute towards increased boundary variability (more information about it can be found in the discussion section for this chapter). The edge-based approach had the largest proportion of combinations (n=412) with a high performance, suggesting that the edge-based approach is more robust (*i.e.,* less sensitive) to parameter perturbations, and is, therefore, likely to provide better results in datasets with highly variable appearance. Figures 3.3a and b show some examples of the output segmentations for each approach.

Table 3.9 summarizes the parameters that appeared most frequently in the highest-performing cases or the lowest-performing cases. The parameters that produced the lowest overall performance may be related to differences in the imaging aspects considered by the segmentation method. For example, although the mathematical definition of the edge-based approach indicates that the "exponent" parameter can take a value of 1 or 2, this second value does not perform better than a value of 1 in most cases. This observation could be because more complex terms might trigger bigger changes that affect the curve evolution process. Table 3.10 shows the correlation coefficients calculated for the different pairs of parameters for each segmentation method. No highly correlated parameters were discovered.

3.6.3 Performance comparison between default and optimized parameters

The results presented in Figure 3.4 show the difference in performance using a five-fold cross-validation when comparing each segmentation based on their default versus optimized parameters stratified by MRI sequence used. For the T2-weighted images, we found a statistically significant improvement on the edge-based (0.480–0.553 vs 0.788–0.830, 95% CI; $p = 0.012$), level-set-based approaches (0.202–0.320 vs 0.782–0.823, 95% CI; $p = 0.034$), and multilevel segmentation (0.492–0.631 vs 0.714–0.823, 95% CI; $p = 0.011$). We observed that the region-based approach produced the best

mean segmentation accuracy when using the default parameters. Results improved only marginally after the optimal parameters where utilized (0.711–0.771 vs 0.743–0.801, 95% CI; p = 0.829). Although we found differences between the optimal and default parameter values, especially in the foreground and background weights, they were not influential in defining the final tumor boundary.

The same trends were observed in the T1-weighted images. The Dice coefficients obtained with the T1-weighted images are lower overall than in the T2-weighted case for the region (0.454–0.553 vs 0.743–0.801, 95% CI), edge (0.385–0.479 vs 0.788–0.830, 95% CI), level-set methods (.387-.478 vs .782-.823, 95% CI), and multilevel segmentation (0.540–0.682 vs 0.573–0.724, 95% CI). The lower performance in the post-contrast T1-weighted scans may be due to the attempt to quantify enhancing and necrotic tumor components (and non-enhancing tumor in-between) simultaneously. The variability between these different tumor parts could cause the algorithms to struggle to find the best boundary. The degree of variability of the enhancing regions is more challenging for segmentation methods to delineate than the smoother, lobulated edges that are observed on T2-weighted images representing regions of edema (Figure 3.4). This trend on lower performance in post-contrast T1 was observed in other previous studies [18].

Method	Criteria	Name	Interval
Region-based segmentation	Recurrent parameters in high-performance cases	Smooth factor Contraction bias Foreground weight Background weight	[0,0.6] [-.1, .1] [7,8] 4, 8
	Recurrent parameters in low-performance cases	Background weight	0
Edge-based segmentation	Recurrent parameters in high-performance cases	Smooth factor Contraction bias Advection weight Exponent	[1,5] [-.5, 0] [4,8] 1
	Recurrent parameters in low performance cases	Contraction bias Exponent	[-.6, -.2] 2
Distance-regularized Level-set	Recurrent parameters in high-performance cases	Length weight Area weight Delta width Kernel width	[7, 8] [-10, -2] [1, 5] [1, 3]
	Recurrent parameters in low-performance cases	Delta width	[7,9]
Multilevel segmentation	Recurrent parameters in high-performance cases	Fine affinity coeff. Cut salient coeff.	[1] [1e-6]
	Recurrent parameters in low-performance cases	Fine affinity coeff.	[15, 20]

Table 3.9 Parameters and parameter values that were part of the combinations that produced the best performing results and also the ones that produced the lowest similarity coefficients

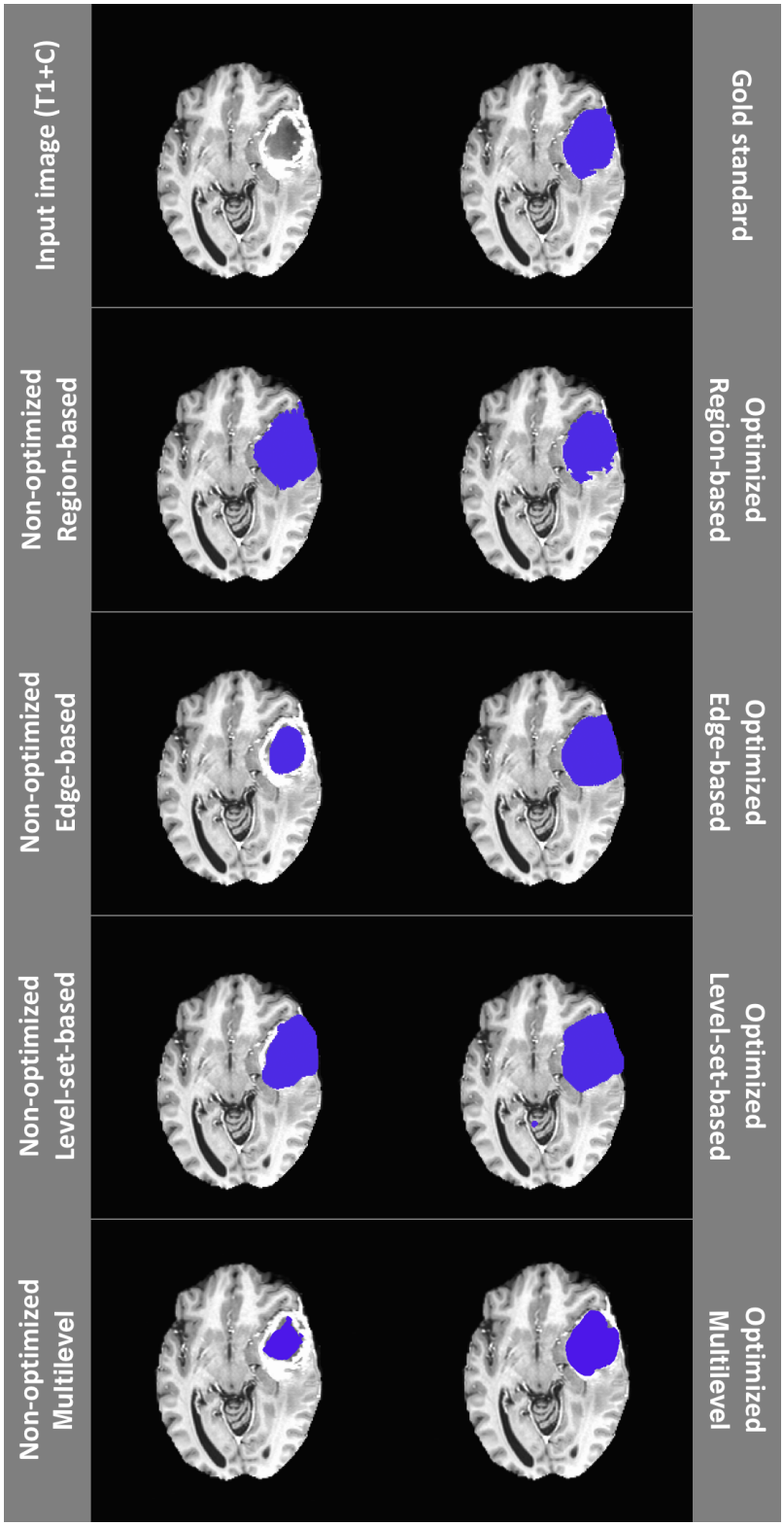


Figure 3.3a Comparative example (T1+C) of resulting ROIs obtained for one subject, comparing the default (left) and optimized (right) parameters.

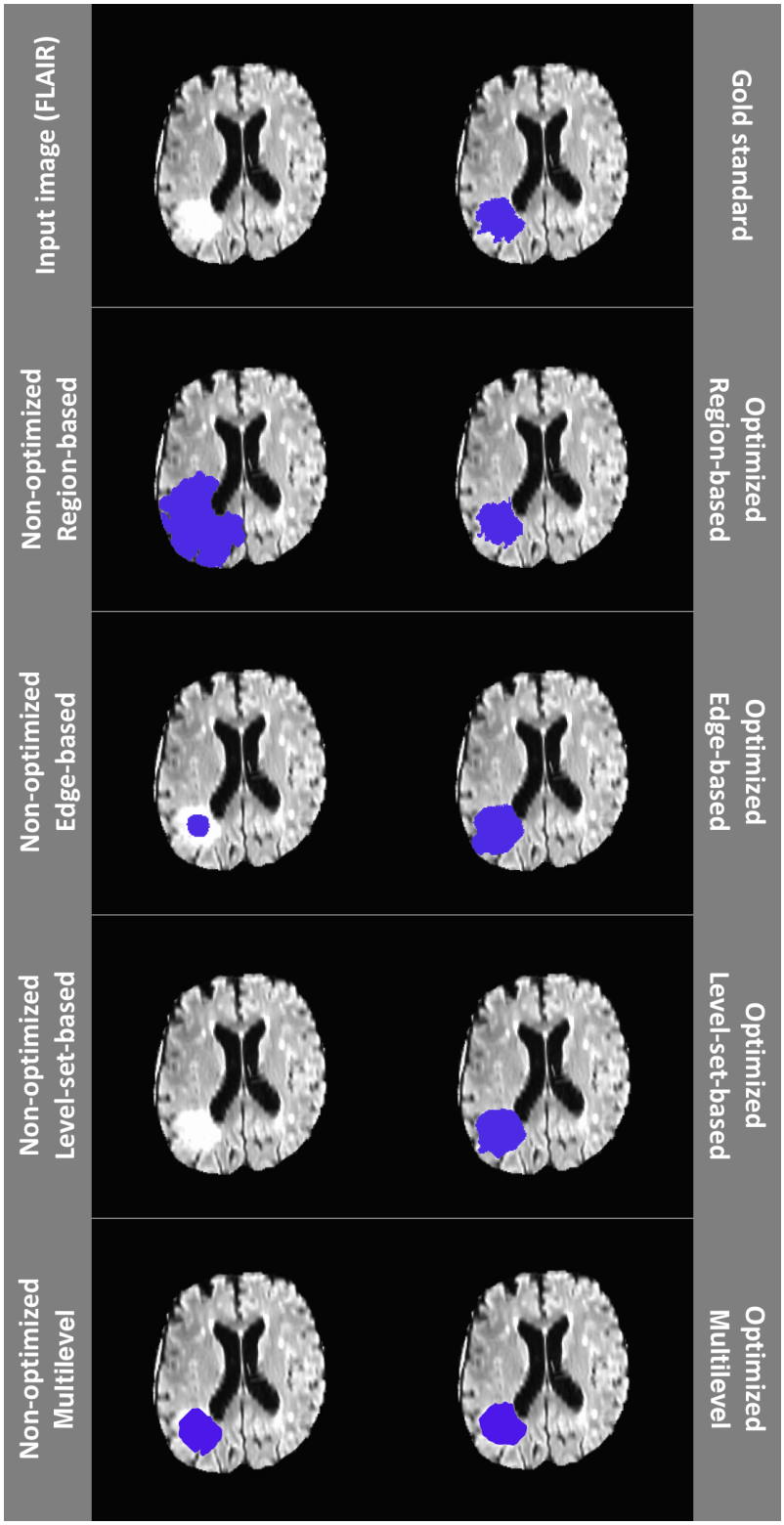


Figure 3.3b Comparative example (FLAIR) of resulting ROIs obtained for one subject, comparing the default (left) and optimized (right) parameters.

3.6.3 Performance across different tumor morphology

Figure 3.5 and Table 3.11 depict the results obtained for the test datasets stratified by tumor size (small, medium, and large as described in Section 3.2). When default parameters are utilized, the region-based approach produced results that were significantly better in all three tumor sizes than the other two methods shown in Figure 3.5 (For large tumor sizes: 0.711-0.771 vs 0.480-0.553, 95% CI, p = 0.016 and 0.711-0.771 vs 0.202-0.320, 95% CI; p = 0.003). When optimized parameters were used, an improvement was observed across all four methods and was especially noticeable in large tumor sizes for the region (0.816-0.853, 95% CI, p = 0.347), edge (0.815-0.858, 95% CI, p = 0.002), level-set (0.800-0.845, 95% CI, p = 0.001), and multilevel segmentation (0787.-0.834, 95% CI, p = 0.116) methods. Notably, the most marked improvement was seen for the level-set and the edge-based approaches.

Additionally, Table 3.12, with Figures 3.6a and 3.6b (T1-weighted and T2-weighted contrasts, respectively), show the results obtained while analyzing the results stratified by tumor boundary. The cases with smoother margins pointed toward less overall variability in segmentation results. The use of the optimized parameters improved performance in both smooth and irregular boundaries. We found that the region-based algorithm had the best performance when the standard parameters were used (0.723-0.805 for smooth borders and 0.673-0.763 for irregular borders, 95% CI) while

the edge-based approach performed the best when the optimized parameters where utilized (0.819-0.869 for smooth borders and 0.742-0.806 for irregular borders, 95% CI). Statistical significance between the results obtained using the default parameters and the optimized set was found for the edge-based (0.479-0.599 vs 0.819-0.869, 95% CI; $p = 0.046$), level-set-based (0.194-0.376, 95% CI vs 0.816-0.865, 95% CI; $p = 0.032$), and multilevel segmentation algorithms (0.371-0.532, 95% CI vs 0.691-0.785, 95% CI; $p = 0.022$).

In summary, for the evaluated dataset, the segmentation algorithms achieved a higher accuracy in tumors that were larger and more spherical in shape. The cases with the lowest Dice coefficients were those that were characterized as small tumors, having multiple tumors, having varying pixel intensities across the tumor border, or having higher necrosis proportions (greater hypo-intense regions on post-contrast T1).

		Smooth factor	Contraction bias	Foreground weight	Background weight
Region-based segmentation	**Smooth factor**	1	0.041	0.302	0.021
	Contraction bias	0.151	1	-0.36	-0.022
	Foreground weight	0.200	0.167	1	0.299
	Background weight	0.030	0.190	0.170	1
		Smooth factor	**Contraction bias**	**Advection weight**	**Exponent**
Edge-based segmentation	**Smooth factor**	1	0.195	0.272	-0.059
	Contraction bias	0.131	1	-0.026	0.230
	Advection weight	0.174	0.039	1	-0.189
	Exponent	0.025	0.219	-0.147	1
		Length weight	**Area weight**	**Delta width**	**Kernel width**
Distance-regularized Level-set	**Length weight**	1	0.292	0.233	0.115
	Area weight	0.344	1	0.228	0.279
	Delta width	-0.014	-0.048	1	0.198
	Kernel width	-0.032	-0.162	0.242	1

Table 3.10 Pearson correlation results to identify dependencies between parameters. Underlined results are statistically significant. Values on lower diagonal correspond to results in T2-weighted sequences (FLAIR and T2) and values on the upper diagonal to post-contrast T1-weighted. This test was not performed for the Multilevel Segmentation.

Segmentation method	Parameters utilized	Dice coefficient							
		All		Small		Medium		Large	
		T1w	T2w	T1w	T2w	T1w	T2w	T1w	T2w
Region-based segmentation	Default	.503	.741	.347	.615	.489	.786	.675	.822
	Optimized	.538	.772	.377	.667	.528	.816	.708	.835
Edge-based segmentation	Default	.432	.516	.311	.449	.436	.535	.549	.565
	Optimized	.619	.809	.489	.739	.672	.851	.697	.836
Distance regularized level-set	Default	.417	.261	.333	.319	.370	.239	.549	.222
	Optimized	.630	.803	.506	.742	.687	.844	.697	.822
Multilevel segmentation	Default	.622	.561	.468	.729	.553	.645	.472	.589
	Optimized	.659	.762	.691	.802	.781	.817	.676	.793

Table 3.11 Results showing the Dice coefficient for each segmentation method for all cases, cases separated by lesion size groups (as defined in the reference segmentation) and type of MRI sequence, T2-weighted (FLAIR and T2 obtained similar results) and T1-weighted (T1 with contrast enhancement).

Segmentation method	Parameters utilized	Dice coefficient					
		All		Smooth margins		Irregular margins	
		T1w	T2w	T1w	T2w	T1w	T2w
Region-based segmentation	Default	.503	.741	.497	.764	.510	.718
	Optimized	.538	.772	.533	.800	.542	.744
Edge-based segmentation	Default	.432	.516	.449	.539	.415	.493
	Optimized	.619	.809	.660	.844	.578	.774
Distance regularized level-set	Default	.417	.261	.357	.285	.477	.237
	Optimized	.630	.803	.675	.841	.584	.765
Multilevel segmentation	Default	.622	.561	.430	.561	.507	.432
	Optimized	.659	.762	.783	.812	.539	.744

Table 3.12 Results showing the Dice coefficient for each segmentation method for all cases, cases separated by margin type (defined by a measure of tumor sphericity) and type of MRI contrast, T2-weighted (FLAIR and T2 obtained similar results) and T1-weighted (T1 with contrast enhancement).

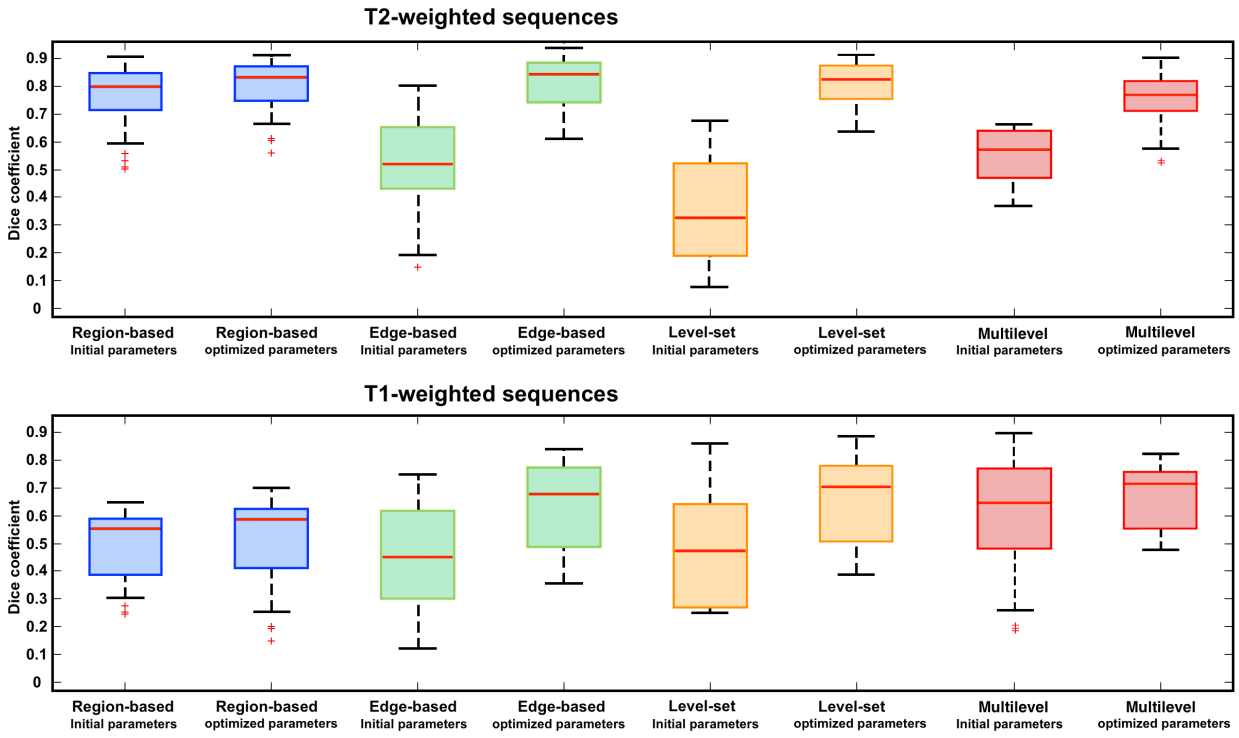


Figure 3.4 Boxplot that compares the performance distribution of the test cases (five-fold cross-validation) before and after the parameter optimization. The boxplot on top shows results for the T2-weighted sequences (FLAIR and T2) and the one on the bottom the results for the T1-weighted sequence (post-contrast T1). A paired t-test shows a significant improvement ($p<0.05$) for the edge-based and level-set-based algorithms.

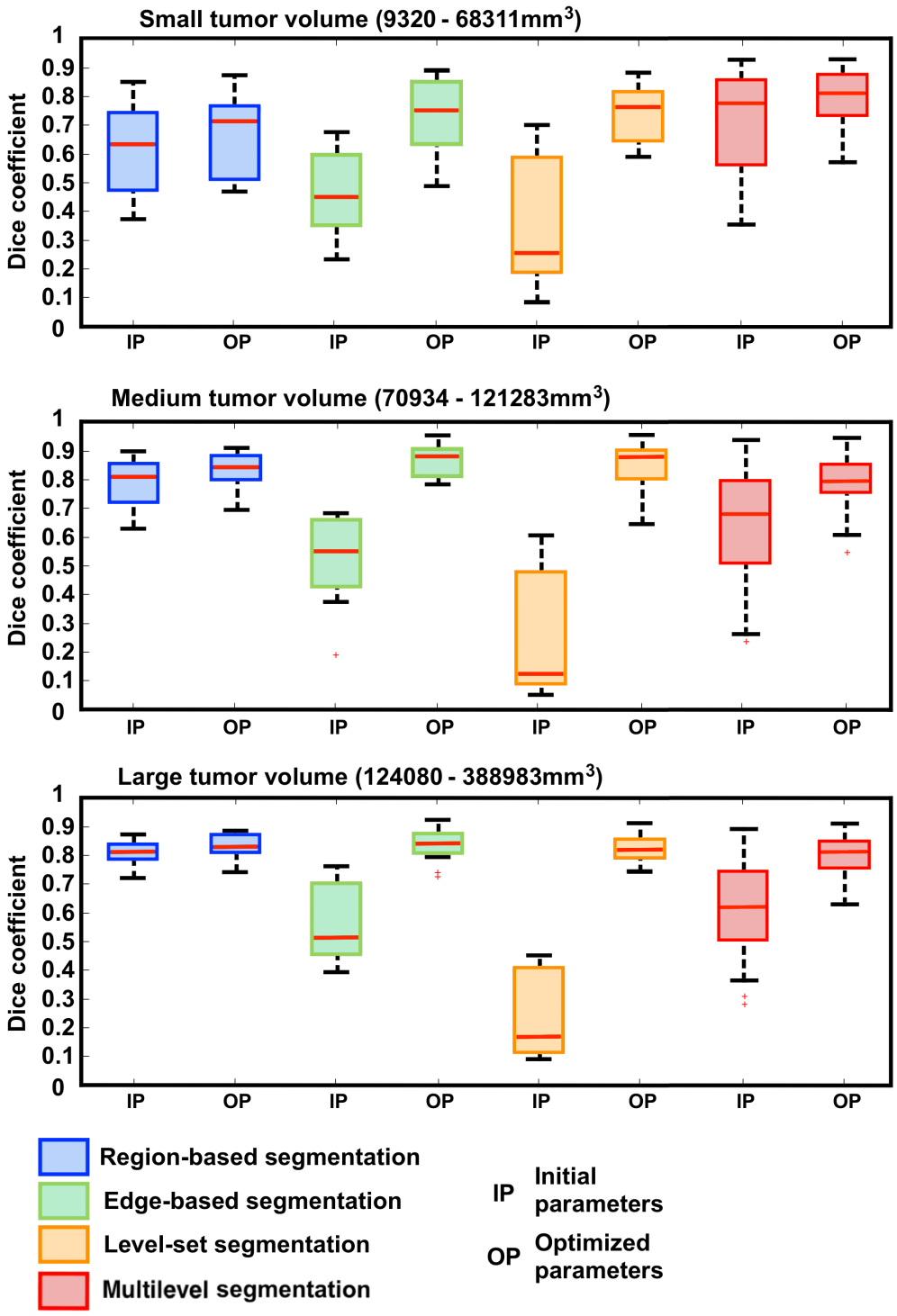


Figure 3.5 Boxplot that compares the performance distribution by tumor size (small, medium and large tumors). Improvement noticed on all three categories, being specially marked on the case where large tumors were analyzed. A paired t-test shows a significant difference ($p<0.05$) for the edge-based and level-set-based algorithms in all cases

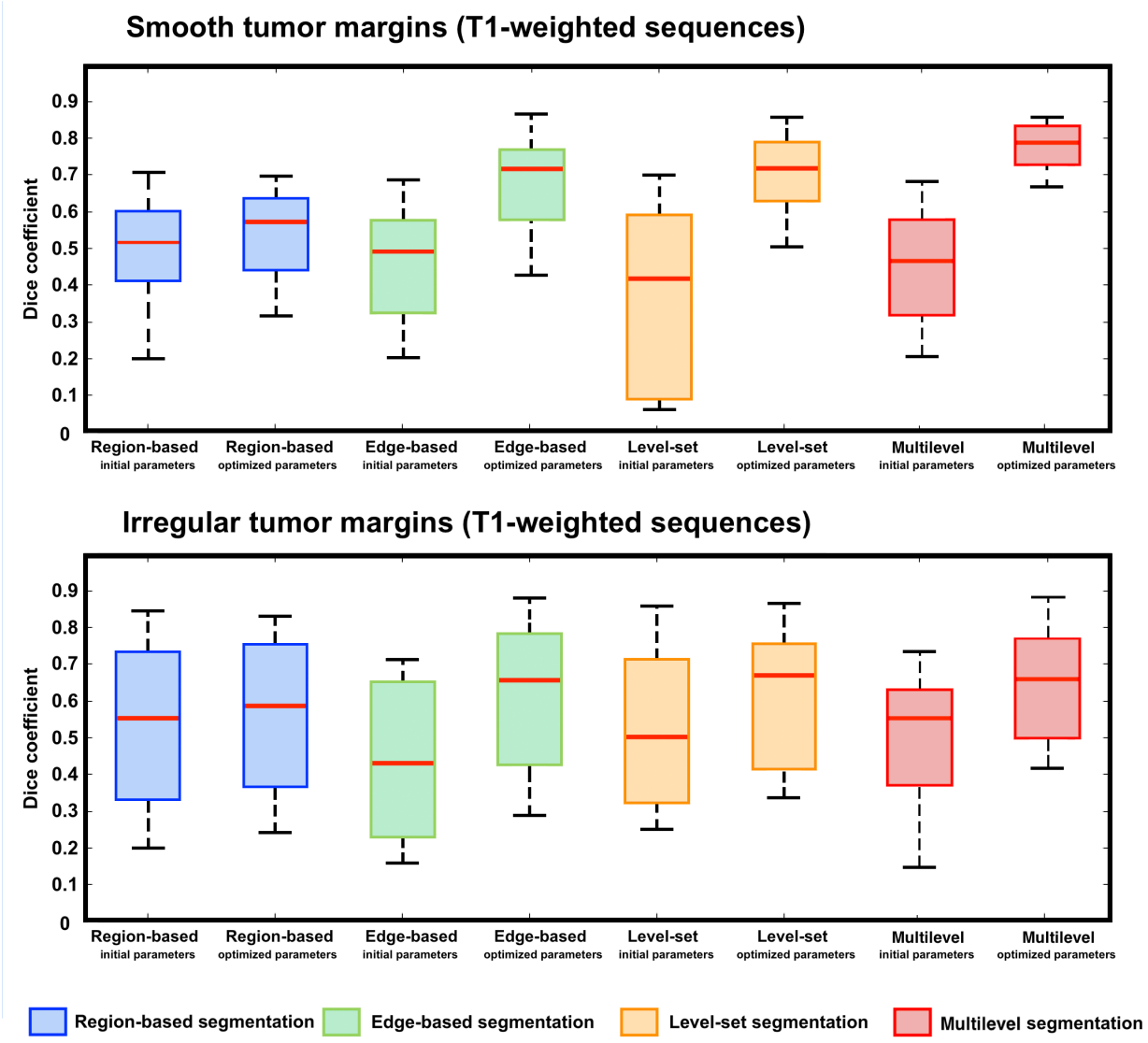


Figure 3.6a Boxplot that compares the performance distribution in the T1-weighted images (Post-contrast T1 sequence) by tumor boundary characteristics (smooth vs irregular edges). Cases with better-defined margins tend to present less variability in results, compared to the ones with diffuse margins. We observed that all methods improved when testing the cases using the optimized parameters (being statistically significant for the edge-based and level-set-based methods).

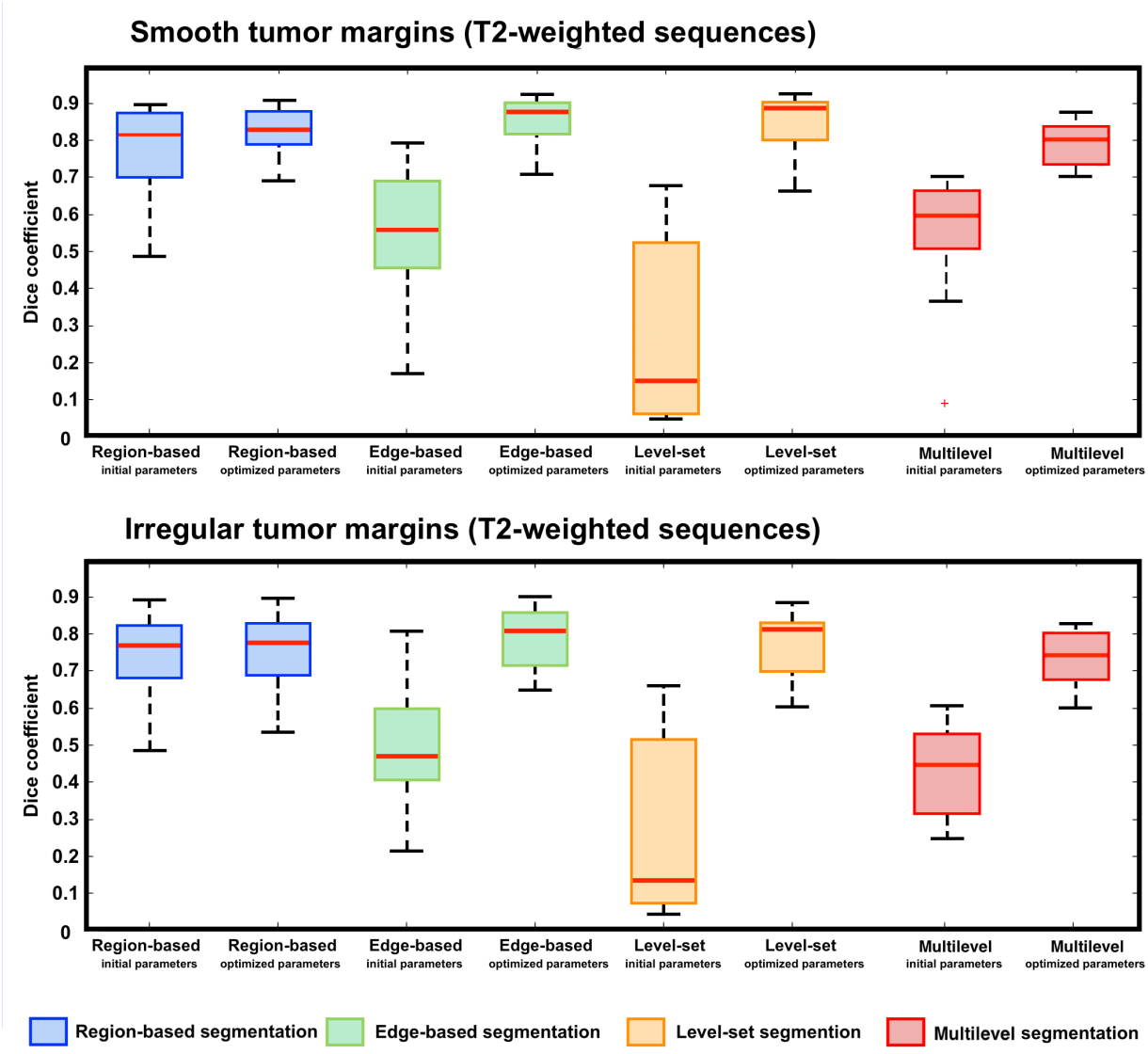


Figure 3.6b Boxplot that compares the performance distribution in the T2-weighted images (FLAIR and T2) by tumor boundary characteristics (smooth vs irregular edges). Cases with better-defined margins tend to present less variability in results compared to the ones with diffuse margins. We can observe that all methods improved when testing the cases using the optimized parameters (being statistically significant for the edge-based and level-set-based methods).

3.6.4 Influence of seed point selection in resulting ROI

We evaluated the effect of perturbations of seed point placement in the final ROI in the region-, edge- and level-set based algorithms. The multilevel segmentation was not subject to this source of variability. The experiment included varying the initial coordinates by 3–6% in each direction. The outputted ROI had high similarity (Dice coefficient of 0.93 to 0.98) with respect to the results obtained using the original input ROI, concluding that small perturbations on the initialization points did not have a major effect in the resulting segmentation.

The variability observed when introducing small perturbations (difference in Dice coefficient ranging from 3% to 6%) to the input seed regions is presented in Table 3.13. The resulting tumor segmentations obtained high Dice similarity scores, obtaining results that are 93% to 98% and are similar to the results obtained without the variation in the seed region

	T2-weighted images	T1-weighted images
Region-based segmentation	0.95 ± 0.01	0.95 ± 0.02
Edge-based segmentation	0.96 ± 0.01	0.94 ± 0.00
Distance regularized level-set	0.98 ± 0.02	0.93 ± 0.02

Table 3.13 Seed point variability observed when introducing small perturbations on the initial region of interest. This experiment was not performed on the Multilevel Segmentation algorithm.

3.6.5 Validation of optimization framework using a different dataset

To assess the validity of the proposed framework, a subset of the UCLA dataset was utilized to determine whether the results remained consistent when applied to a different dataset. A sample of 50 cases was selected to achieve a statistical power greater than 85% with a significance level of 0.05 (Table 3.5). The same parameter optimization was performed using the four methods previously described in Section 3.6 (edge based, region, level-set-based and multilevel segmentation). The optimization occurred only in post-contrast T1 cases given that for this dataset, manual segmentation (reference standard) only exists for tumor core (enhancing plus necrotic).

As shown in Table 3.14, the results obtained for this experiment are different in only some instances from to the ones obtained previously. Except for two parameters (contraction bias for the edge-based segmentation and delta width for the level-set method), these new results for optimal parameters are equal or within the intervals for optimal performance previously reported (median ± standard deviation). The most difference was the delta-width parameter (level-set-based segmentation). It is hypothesized that this is because it is part of a regularization term because it defines the size of the kernel to be applied during the segmentation process, it is possible that the average large size of tumors may play role in the increase of the value of this parameter with respect to its original value.

	Parameter	Default parameters	Optimization 1 (Section 3.6)	Experiment 2 (new results)
Region-based segmentation	Smooth factor	0 ± 0	0 ± 0	0 ± 0
	Contraction bias	0 ± 0	0 ± 0	0 ± 0
	Foreground weight	1 ± 0	8 ± 1	8 ± 0
	Background weight	1 ± 0	4 ± 0	4 ± 0
	Dice coefficient across all test sets	*.70 ±.01*	*.61 ±.08*	*.64 ±.26*
Edge-based segmentation	Smooth factor	1 ± 0	1 ± 0	1 ± 0
	Contraction bias	.3 ± 0	0.5 ± 0	0.45 ± .11
	Advection weight	1 ± 0	8 ± 0	8 ± 0
	Exponent	1 ± 0	1 ± 0	1 ± 0
	Dice coefficient across all test sets	*.52 ±.02*	*.72 ±.05*	*.68 ±.02*
Distance-regularized level-set	Length weight	5 ± 0	8 ± 0	8 ± 0
	Area weight	-3 ± 0	-8 ± 1.67	-7 ± .32
	Delta width	1.5 ± 0	5 ± 0	4 ± .44
	Kernel width	8 ± 0	1 ± 0	1 ± 0
	Dice coefficient across all test sets	*.35 ±.16*	*.78 ±.30*	*.73 ±.41*
Multilevel segmentation	Fine affinity coeff	10 ± 0	1 ± 0	1 ± 0
	Strongly conn coeff.	0.2 ± 0	0.5 ± 0.3	0.5 ± 0.10
	Interpolant cut coeff	0.4 ± 0	0.7 ± 0.18	0.6 ± 0.14
	Cut salient coeff	1e-6 ± 0	1e-6 ± 0	1e-6 ± 0
	Dice coefficient across all test sets	*.71 ±.23*	*.76 ±.21*	*.63 ±.35*

Table 3.14 Table that shows the parameter optimization results (tumor core measured in post-contrast T1 in the UCLA dataset) compared to the results found in previous experiments (TCGA/MICCAI dataset). Results show that the framework for parameter optimization can achieve consistent results when using different datasets.

This additional experiment demonstrated that although there is some expected variation in the results that were encountered, this framework can consistently find the best performing parameters for different algorithms in a specific domain using different datasets and different across domains (stroke, lung nodules).

3.6.6 Application of multiple algorithms to represent tumor boundary variability

One of the practical applications of having diverse optimized algorithms for tumor segmentation is the potential to automatically characterize the tumor boundary and obtain multiple measurements about the likely location of the tumor boundary. Having different measurements is statistically useful to determine differences between groups (as performed in previous sections) and can be carried over into the measurement of different tumor biomarkers until the clinical decision-making process.

This section covers the creation of tumor measurement variability maps at the algorithm-level using different approaches. The next chapter goes more in-depth to examine the pixel-level variability that arises from image intensity distribution.

Figures 3.7 a and b show examples of tumor measurement variability maps generated using five different segmentation algorithms. The first four are the segmentation approaches covered in this chapter (region-based, edge-based, level-set-based, and mul-

tilevel segmentation). The fifth algorithm is a statistical classifier (Support Vector Machine) that has been previously trained and independently optimized [19] for brain segmentation of high-grade gliomas. Section 3.6.6.1 has a brief description of this methodology.

3.6.6.1 Segmentation using a statistical classifier

This segmentation combines a Support Vector Machine (SVM) [83] classifier using intensity-based and texture-based image features followed by a hierarchical regularization based on conditional random fields (CRFs). This regularization using CRFs introduces spatial constraints to the classification by making the assumption that voxels are independent of their neighbors [84]. The hierarchical approach works by applying a regularization step, assuming that the data is evenly distributed, at different stages of the classification [19,85], improving the robustness and speed of the model and its accuracy by considering neighbor relationships. This method was tested in the brain tumor data during the MICCAI segmentation competition [18].

During the SVM training step, imaging and textural features are extracted from the input imaging data. Among those features, the most important ones are the ones derived from image intensity in the different image contrasts. A group of first-order tex-

ture features (mean, variance, skewness, kurtosis, energy, entropy) are extracted to support the classification. In the end, this model is trained using a 28-dimensional feature vector which is formed by the voxel-wise concatenation of the intensity and texture features obtained for each voxel.

When the features from the training phase have been extracted, this discriminative classifier attempts to find a boundary that can separate the feature space into different categories. This separation of the different features is obtained by a minimization of a hyperplane that divides the feature space as boundary curves in a 2D feature map. The optimization process is defined by:

$$\min \frac{1}{2} w^T w + C \sum_{i=1}^{l} \varepsilon_i$$

where w is the normal vector of the separating hyperplane, C is a penalty parameter for the classification variable ε found by solving the expression $1\text{-}\varepsilon < y_i(w^T\phi(x_i))$, where x and y are the instance-label pairs of the dataset. The regularization is done in two different stages using a conditional random field method obtained by using the following energy minimization model:

$$E = \sum_i V(y_i) + \sum_{ij} w(y_i, y_j, x_i, x_j) D_{pq}(y_i, y_j)$$

The expression is based on graph-cuts and primal-dual decomposition [86]. E is

the data energy at the current point, *V* is the unary potential function, *w* describes the neighborhood relationships and D_{pq} is a distance function for the pairwise potentials calculated directly from the voxel-wise grid. The potential V is defined as $V(y_i) = c(1-d(\ddot{y}_1,y_1))$ where c is a constant to adjust the weight on the pairwise potentials, d is the Kronecker function and $\ddot{y}_i$ is the new label after the regularization.

As the last step in the algorithm, after image pixels have been classified, a secondary 2D neighborhood regularization is applied by considering the neighborhood structure (8 neighbors in 2 dimensions). This step allows for overall smoother and more accurate results.

3.6.6.2 Evaluation of variability maps derived from multiple algorithms

In summary, this section puts together a set of different algorithms (region-based, edge-based, level set based, multilevel segmentation, and statistical classifier) to generate Tumor-measurement Variability Maps (confidence maps) (Figures 3.7 a and b). The determination of which algorithm was more likely to produce the most accurate boundary was determined by assigning a weight that is proportional to the algorithm's performance observed during the optimization/cross-validation experiments described in section 3.6.1.

Corresponding weights were defined proportionally according to the performance observed when working individually in the full dataset (84 subjects). Results are shown in Table 3.15.

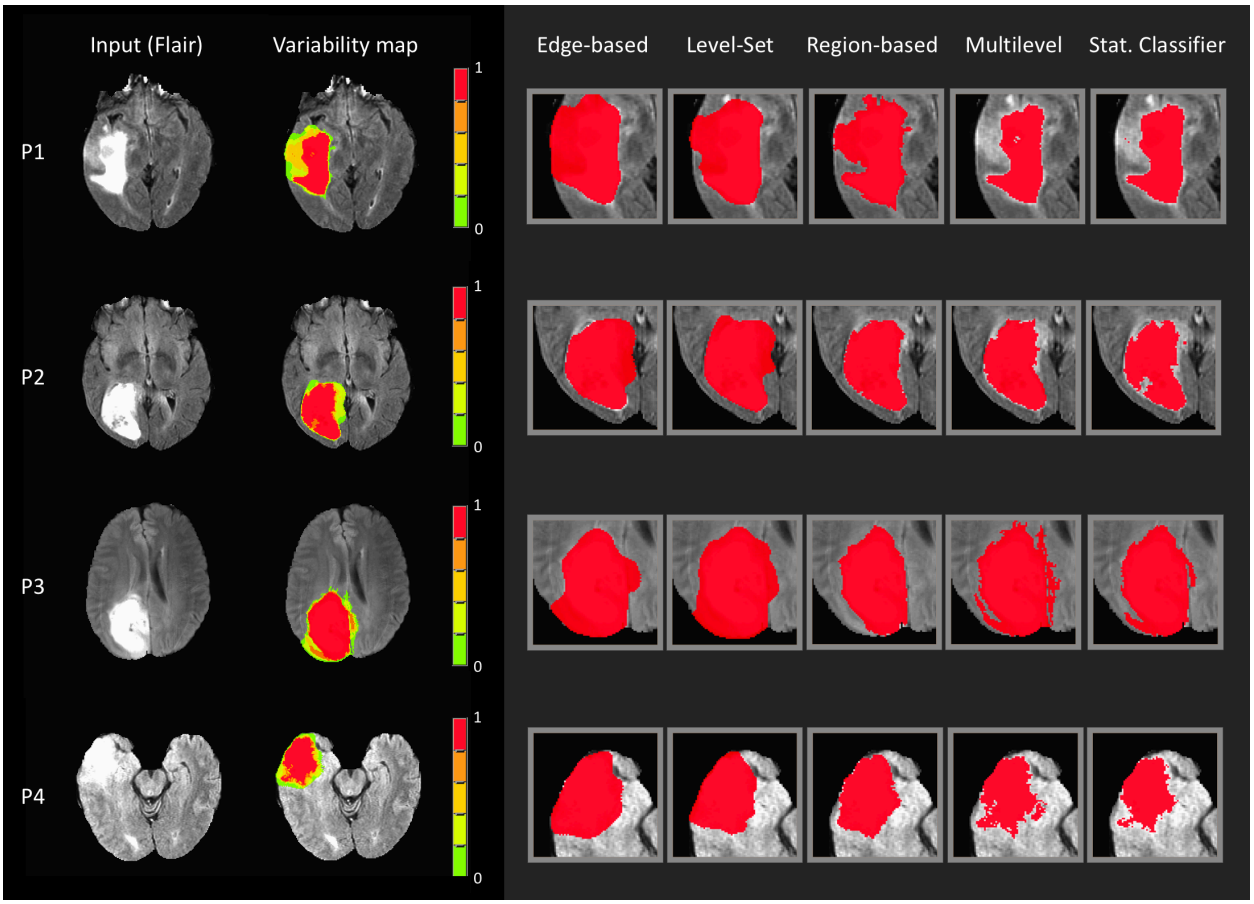


Figure 3.7a. Examples (FLAIR) of tumor measurement variability maps generated using different algorithms. The first two columns on the left show the input image and resulting tumor variability map. The other columns on the right show the individual components of the variability map provided by the different algorithms.

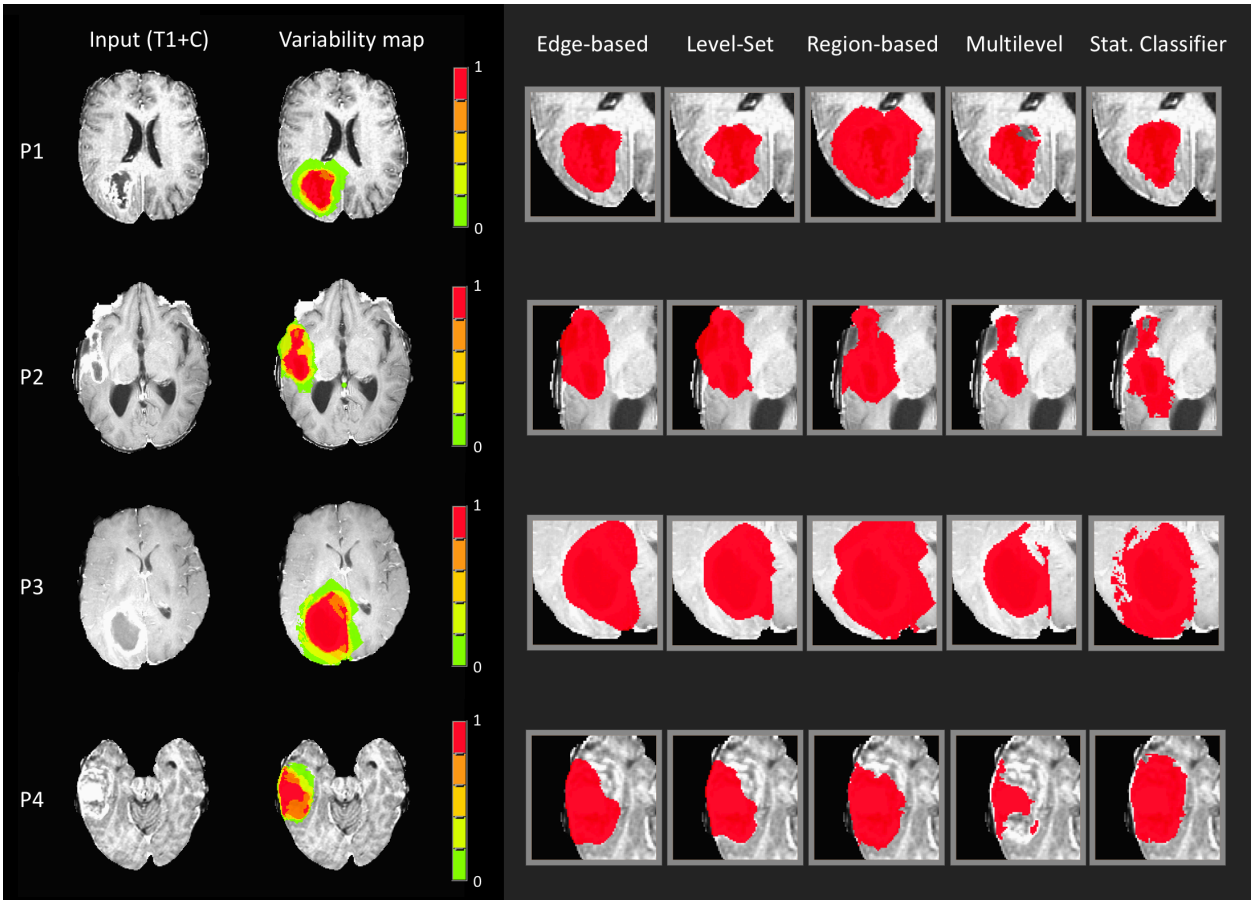


Figure 3.7b. Examples (T1+C) of tumor measurement variability maps generated using different algorithms. The first two columns on the left show the input image and resulting tumor variability map. The other columns on the right show the individual components of the variability map provided by the different algorithms.

	Region-based	Edge-based	Level-Set based	Multi-level segment.	Statistical classifier	Dice coefficient (on test set)
Weight on T2 sequences	20.13%	21.09%	20.93%	19.86	19.96%	0.782
Weight on T1 sequences	17.83%	20.52%	20.88%	21.85%	18.89%	0.659

Table 3.15. Different weight for the ensemble of algorithms and their performance when evaluating performance on full dataset (84 subjects).

3.7 Discussion

We presented a generalizable framework to define and evaluate segmentation algorithms and their input parameters when attempting to analyze imaging datasets with objects of interest of non-uniform boundary characteristics. The framework assesses the performance of each approach using a parameter search that automatically defines the input seed region and the testing interval for each of the algorithm parameters. We identified the set of optimized parameters that produce the best results for each algorithm based on the test dataset. We additionally explored the variability in results by controlling for tumor size and margin. Improvements in segmentation accuracy ranged from 7-51% when comparing between default and optimized parameter values. While demonstrated in the domain of brain cancer, our approach can be adapted to other areas as well.

The edge-based segmentation algorithm achieved consistently higher Dice coefficient compared to the other algorithms. The region-based approach usually had the best performance when using the default parameters. Although performance increased using the optimized parameter set, results did not improve significantly. The level-set-based segmentation was observed to have the largest differences before and after optimization, showing large ranges of variability given the input parameters, and making a marked improvement when the new optimized parameters were utilized. Finally, the multilevel approach showed a particular reduction in the error intervals (Figure 3.6 and Figure 3.5)

when the optimized parameters are used compared with the default parameters, demonstrating that segmentation parameters can be a significant source of segmentation variability at the algorithm level.

Algorithm performance was lower in tumors with smaller volumes and was lower in tumors with variable borders. Smaller tumors were more influenced by the selection of the seed region. In addition, the intensity ranges within the region are smaller, resulting in a narrower range of parameter values that achieve high Dice similarity coefficients. Variable borders tend to represent an overlap between normal and abnormal tissue, reducing the contrast between the two regions and creating "gray areas" that make it difficult for the algorithm to determine if they are part of the intended object to segment or part of what is being classified as the background. On brain tumor images. these areas can correspond to a non-enhancing tumor or transition areas between tissues with edema and normal brain tissue).

Limitations of this work include the computational power and time that are required to evaluate all possible parameter combinations. The data processing was distributed across three workstations with 2.5-3.4GHz Intel Core i7 processors and 16–64 GB of memory. An individual iteration of each segmentation approach took at least 1.5 minutes, but thousands of iterations were needed to fully evaluate the parameter space. Although the simplex method reduces the amount of time and parameter combinations needed to evaluate, as noted in Table 3.8, the approach does not always find the set of

parameter values that achieve the highest accuracy. Furthermore, the framework is most effective when trained in a large dataset that represents the level of variability of the domain of interest. We selected a representative sample of patients from the MICCAI/TCGA dataset that contained tumors with different characteristics, including cases with different tumor sizes, smooth and irregular margins, simple and complex shapes, different degrees of tissue hyper-intensity, and some cases with imaging artifacts. Additionally, because of the complexity of the segmentation task, inherent variability exists in the reference standard as well. The reported intra- and inter-rater agreement for these tumor segmentation methods ranges between 74% and 85% [18].

Future work can include the development of a framework that is not dependent on an annotated dataset, generating the comparison standard by inter-method consensus, or by using an additional statistical or knowledge-based prior. Additionally, In these experiments for the generation of the ensemble method we selected the individual weight of each algorithm as well as the optimal parameters by training the algorithms separately. Although we produced results that far outperform the base parameters for each algorithm, a version of this ensemble approach that optimizes each method based on the performance with respect to the other methods can be done by using a multivariate approach. An example can include training one algorithm to work better in cases

where the other approaches are known to perform worse by using the principal component analysis (PCA). This is part of our future work to potentially improve and further advance the ensemble segmentation maps.

The development of methods to increase precision and reduce the variability of measurements on imaging data is of high relevance to distinguish between the biological heterogeneity of the tumor. Suboptimal selection of algorithms and parameters can lead to inconsistent delineations of ROIs, which can lead to incorrect conclusions whether interpreted by human readers or machine learning algorithms. This problem is compounded when analyzing groups of subjects or individuals at different time points since each measurement can vary the algorithm performance.

While our framework was optimized to produce segmentations that mimic ones generated by trained radiologists using the Dice method as an objective function, it might be desirable to optimize based on some other criteria such as which segmentation method results in the most consistent and precise prognosis. Arguably, for automated methods to reach their full potential, they should be able to provide information beyond what is achieved by traditional manual analysis, not simply replicate existing findings.

Notably, while our results depend on these choices of objective functions, our overall approach to optimization does not. When a larger scale of relevant patient-oriented measures such as outcome or survival measurements become available in the patient records, it could be possible to use them as part of the optimization criteria and

develop a global optimization from which trends and meaningful observations can be extracted: from the choice of imaging parameters through to the segmentation software.

Other notable prior efforts have explored the use of different segmentation methods to obtain an improved ROI. Huo *et al*. [87] tested an ensemble approach on a series of postoperative post-contrast T1 magnetic resonance (MR) imaging scans. The authors concluded that collective segmentation produces a better result compared to the one obtained by an individual method. Our work differs in that we focus on reducing the segmentation variability of the algorithm by performing a comprehensive evaluation of its segmentation parameters and testing each method on a larger sample of multi-channel brain tumor data. Another paper by Warfield et al. [88] demonstrated how using multiple segmentation methods can generate an estimate of the true segmentation. While they also explored variability in segmentation, their method was aimed at finding an optimized ROI to be used as a reference when manual annotations were unavailable. While they did not attempt to improve the individual performance of each algorithm, this framework systematically evaluates and optimizes segmentation methods to maximize performance and achieve more consistent, accurate and robust results.

3.7.1 Summary of findings

In this chapter, we attempted to determine a measurement of tumor boundary variability by looking at the different border calculations obtained by using multiple segmentation approaches. We demonstrated that these methods can be optimized for the purpose of brain tumor segmentation by training and updating parameter values on a set of training data, and provided an automated framework that obtains the optimal parameter values. Finally, the developed framework is able to determine which segmentation can have the most stable (*i.e.,* robust) results across a set of data, and which ones tend to present more variability in results depending on the type of data that is presented to them (homogeneous borders vs. heterogeneous border, etc.).

The following chapter elaborates on the development of a new approach that aims to determine the variability of not only the overall tumor boundary, but from the different tumor components (edema, enhancing, etc.) and uses prior information on cerebral tissue sample to determine the most likely intensity ranges on which the tumor boundary can be located. A clinical evaluation is performed to evaluate the utility of the variability metrics that were discussed throughout this chapter.

Chapter 4: Automated quantitative tumor evaluation

4.1 Overview

This chapter describes a new approach for automated tumor segmentation using multimodal MR images. It is particularly burdensome for clinicians to manually review imaging studies. In addition to increased labor and expense, manual measurements can have a high degree of measurement variability [17] due to the inconsistency and diversity of MRI acquisition parameters (*e.g.,* echo time, repetition time, etc.) and strategies (2D vs. 3D), along with hardware variations (*e.g.,* field strength, gradient performance, etc.) that change the appearance of the tumor [7].

This chapter addresses the hypothesis that the inherent variability in tumor volume measurements can be leveraged to provide a more accurate assessment of tumor burden and produce an estimate of tumor segmentation variability in the form of Tumor-measurement Variability Maps (TVMs). These represent a measure of uncertainty along the tumor boundary (image variability). Different from the other methods that were described in previous sections, this new algorithm considers the variability that arises from the intrinsic tumor variability (*e.g.,* different tumor components) and segmentation error (*e.g.,* poorly detected tumor edges) to derive the tumor boundary and produce an estimate of segmentation error [87]. Additionally, it incorporates the calculation of a prior probability distribution of the different brain tissues, characterization of morphological

features using a super-pixel-based approach, and localization of the tumor ROI using a learning approach that captures morphological characteristics of previously false-positive classifications.

The relevance of this work lies in the calculation of segmentation error across the tumor boundary, a type of metric that is largely unexplored [19], and that is increasingly obtaining more attention [10, 87, 89] given that variability in measurements can directly affect assessment of medical images, treatment planning, and clinical monitoring that may potentially result in erroneous treatment decisions or conclusions about potential therapeutic benefits.

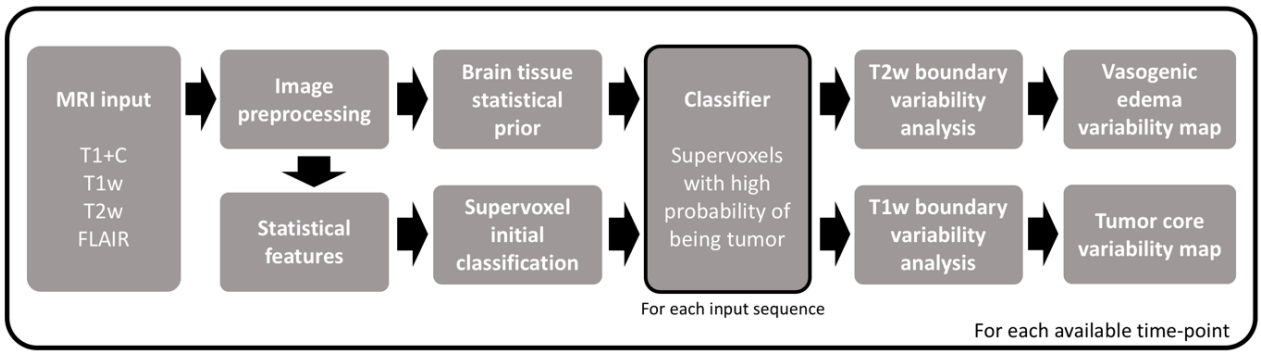
4.2 Multimodal knowledge-based tumor segmentation

We developed a processing pipeline to automate the segmentation from the raw MRI images to create tumor variability maps (confidence maps) that indicate tumor extent. This pipeline is illustrated in Flowchart 4.1 and Pseudo Code 4.1. As the first step, the algorithm selects all MR modalities of interest, including pre-contrast T1-weighted images, post-contrast T1-weighted images (T1+C), T2-weighted images, and T2-weighted fluid-attenuated inversion recovery (FLAIR). Preprocessing steps performed to normalize and register the studies were discussed in Section 3.3.

Pseudo Code 4.1: Brain Tumor segmentation through estimation of variability

for each follow up: d = {d_1,… d_{t-1}, d_t}
    // **Data preprocessing**
    Read directory and load MRI Sequences → FLAIR, T1, T1+C, T2
    Register and skull strip all volumes;
    Normalize and denoise all volumes;
    Calculate tissue probability masks for white and gray matter → WM, GM
    Obtain Subtraction map T1+C - T1 → DeltaMap
    // **Preliminary ROI identification**
    for sequences: m = {FLAIR, T2, DeltaMap}
        for orientations x = {axial, coronal, sagittal}
            Obtain brain tissue distributions(WM,GM)
            Cluster volume using SLIC (input volume(m,x), k=10) → Vol. superpixels
            Extract image features(m,x) → Histogram, Symmetry, Inhomogeneity
            Find tumor preliminary ROI by using extracted features → 3DtumorROI
        end
    end
    // **Tumor segmentation**
    for tumor regions: z = {Edema, Enhancing, Necrosis}
        Get tissue distribution inside ROI (3DtumorROI, WM, GM)
        Evaluate rate of change in region z → top regions of intensity variation (t(i))
        Apply thresholds at intensity t(i) in ROI(z) → Tumor subregion variability map
    end
    Obtain binary tumor masks by obtaining mask consensus in variability maps
    Locate unclassifed regions inside tumor area → Non-enhancing tumor
    Warp back all volumes to original scan space
end

The system is divided into a series of preprocessing strategies, followed by the proposed tumor segmentation algorithm. This approach finds an approximate tumor ROI by using a knowledge-based approach that considers the prior distribution of brain tissues described in the next section. The intensity variation observed in the approximate tumor ROI is then analyzed to find the possible tumor boundaries for the TVM. This approach was evaluated using the 2015 Multimodal Brain Tumor Image Segmentation Benchmark (BRATS) dataset [18].



Flowchart 4.1. Overall segmentation process from input multimodal MRI, feature extraction, preliminary tumor ROI calculation by supervoxel classification, and a tumor histogram variability analysis to generate segmentation error estimates for the overall tumor boundary and different tumor components. The process is repeated for all time-points available for an input subject. The output variability maps are a graphical representation that reflects the likely location of a variable tumor boundary.

4.3 Definition of the tumor ROI using statistical priors

The selection of an initial tumor ROI to initiate the segmentation was performed by obtaining a statistical prior distribution of the normal cerebral tissues to determine

an approximate location of the tumor area (Figure 4.1). This process was performed for the T1- and T2-weighted images. We adapted the method described in [90], which models image intensities as a mixture of k Gaussian distributions. The value of k is selected based on the number of tissues, such as gray matter, white matter and cerebrospinal fluid where each Gaussian cluster is modeled by its mean (μ_k), variance (σ_k) and a mixing proportion. Bayes rule is employed to produce the posterior probability of each tissue class. The probability of observing an element with intensity y_i on the k^{th} Gaussian is given by:

$$P(y_i|k = k, \mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y_i - \mu_k)^2}{2\sigma_k^2}\right)$$

Finally, the probability is maximized by minimizing the cost function:

$$\varepsilon = -\log P(y|\mu, \sigma, \gamma) = -\sum_{i=1}^{I} \log\left(\sum_{k=1}^{K} \frac{\gamma_k}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y_i - \mu_k)^2}{2\sigma_k^2}\right)\right)$$

where K is the total number of Gaussian distributions (one for each tissue), and I is the total number of image elements. The update of the mixture proportion (γ_k) is performed by the expectation maximization (EM) algorithm and generates pixel-wise probability maps for cerebral gray matter (GM), white matter (WM), cerebrospinal fluid (CSF), and soft tissue [90-91]. Finally, these maps are used in our framework to generate an initial ROI to initialize the segmentation methods by selecting the intensity regions of high

probability for the different normal brain tissues. Main areas of low probability are defined as likely to be part of tumor tissue, allowing for detection of multiple distinct lesions in the image volumes. Implementation of this approach permitted us to evaluate the influence of seed point selection on the resulting ROI.

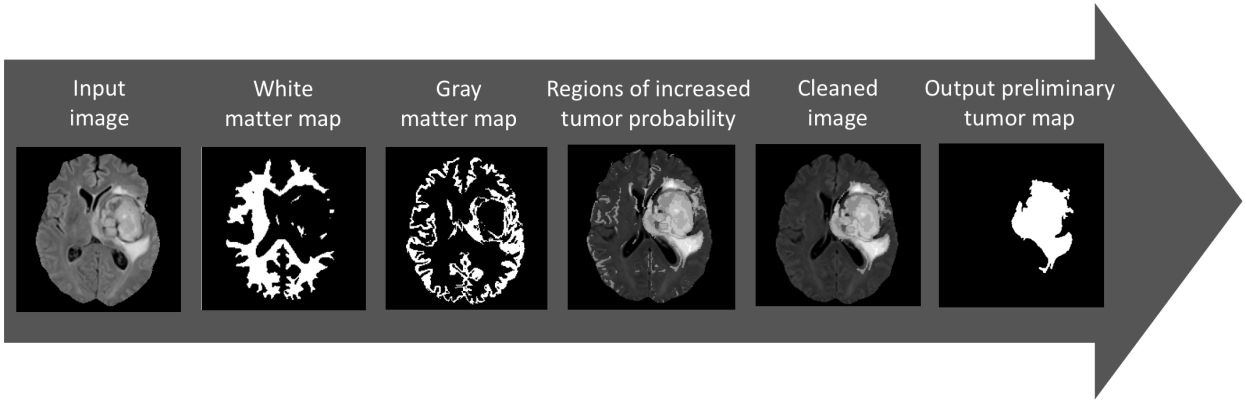


Figure 4.1. The overall process followed to obtain the preliminary tumor ROI using prior information encoded in brain tissue pixel probability maps.

4.4 Tumor segmentation

After this step is completed, using the information provided by the tissue distribution probability maps as well as imaging features from each MR volume, an initial tumor ROI is obtained. This process involves the partition of the images into superpixels by using the SLIC algorithm [92]. This method implements an adaptation of the k-means clustering approach that provides an efficient and fast segmentation of an input image while combining color and spatial proximity to generate the superpixels. By default, in a simple implementation of the SLIC algorithm, only one parameter has to be set (the

number of superpixels to be extracted) before being able to use it. We modified the approach to automatically select the number of clusters based on the histogram distribution of the input imaging volume (3D) [93], setting an initial histogram partition parameter k at ten. This was based on the type of normal and tumor tissues present on the input images -gray matter, white matter, cerebrospinal tissue, soft normal tissue, bone tissue, enhancing tumor, necrotic tumor, non-enhancing tumor, edema and empty spaces/air. It is worth noting that this parameter represents the starting point for partitioning the input image. Its value does not influence the output ROI as it can contain as many of the different clusters of pixels generated in this step depending on the output of the previous step (ROI generation using prior tissue information).

A preliminary tumor ROI is obtained by selecting the superpixels that represent the regions with the lowest probability of being normal brain tissue according to the information provided by the described distribution probability maps for normal cerebral tissues (including gray matter, white matter, and cerebrospinal fluid) [90]. This process is iterated under different orientations (axial, coronal, sagittal) to increase the accuracy of the initial ROI, hypothesizing that different tumor shapes may be easier for the algorithm to identify if visualized under different perspectives. For instance, a u-shaped tumor might be visualized as two different small structures in the axial view, but as a continuous and more defined mass in the coronal view, as shown in Figure 4.2. The

preliminary tumor ROI is then obtained by taking the union of all regions generated across different perspectives, resulting in a single volumetric ROI.

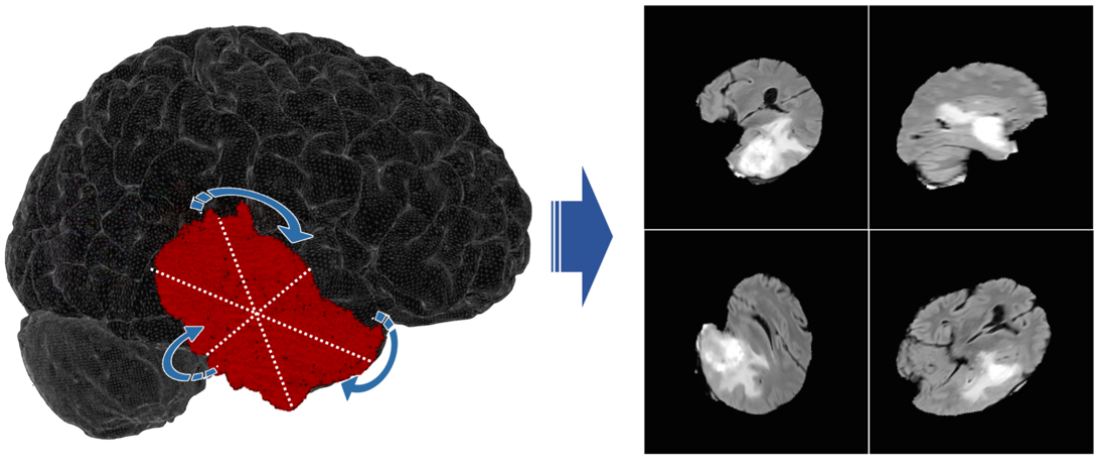


Figure 4.2. To have multiple estimates of the tumor boundaries, multiple rotations are found along the tumor major axes so at the end of this process these measurements can be aggregated into the measurement variability map. In this work, the segmentation occurs on the three main tumor axes and ten random rotations on each axis.

4.5 Multimodal tumor boundary selection

The next step involves identifying a set of tumor boundaries for the total tumor mass as well as for the tumor sub-regions (*i.e.,* edema, necrosis, enhancing and non-enhancing tumor) using the preliminary tumor ROI defined in the previous step. This approach represents the tumor variable boundary by performing multiple measurements and then combining them into a TVM to quantify the uncertainty associated with segmentation boundaries.

The specific tumor boundaries are obtained as follows: A single definition for T2 abnormality was used to define a "T2 abnormal ROI" using the preliminary ROIs found in the FLAIR and T2 contrast images. Regions of edema were extracted by ranking the intensity rate of change on the preliminary tumor ROI histogram, defining boundaries as the locations where the highest total variation across the histogram corresponding to the tumor region are found. The tumor variability map is produced by aggregating the different binary ROIs obtained at each of these identified values on the tumor histogram.

The same process is followed to locate the enhancing and necrotic regions, but using the post-contrast T1 sequence or a subtraction map, (defined as T1+c - T1 volumes) [94]. Similarly, the TVM for these regions was then obtained by aggregating these different approximations of the tumor boundary (Figure 4.8). A binary representation of the tumor mask was obtained by using the following expression for majority agreement using the TVM as input:

$$I = \begin{cases} P_{i,j} \geq \dfrac{n}{2} & \therefore \quad I_{i,j} = 1 \\ P_{i,j} < \dfrac{n}{2} & \therefore \quad I_{i,j} = 0 \end{cases}$$

where I is the output binary image, $P_{i,j}$ is the intensity at pixel location i, j of the TVM P and n is the number of discrete probability levels defined in the variability map. The output is a set of masks that represent the tumor extent and the different sub-regions with the possibility to calculate variability metrics (*e.g.*, agreement ratio, standard deviation, statistical change measurement, and others).

4.5.1 False-positives identification and correction

Although tumor variability can be characterized from different points of view including different segmentation methods, variability from parameters and, variability in tumor boundary, it is still a possibility to obtain results where an incorrect region is selected as tumor region (*i.e.,* false positives).

To increase the sensitivity (true positive rate) of this approach, a database dictionary was developed that accumulate sets of previously identified examples and information about their location (coordinates in two and three dimensions), mean intensity, and shape characteristics. For example, thin, elongated shapes may correspond to vascular structures that are indicative of false positives. When analyzing a new case, a comparison against this database is performed (comparing intensity, morphology/shape, and location), and the regions with high correlation with respect to the previously observed example are eliminated. Examples of cases when this approach is particularly useful include cases when regions of the skull failed to be eliminated and are close to the tumor. Enhancing vessels close to the tumor, and normal brain structures, especially in the cerebellum, that usually appears brighter than other brain regions but is not a tumor.

The construction of the database of false positives was created using a set of 50 training subjects in three MR sequences (post-contrast T1, FLAIR and T2) that were a

representative sample of tumor of different sizes and boundary characteristics as described in Chapter 3. Annotations were corroborated by a trained neuro-radiologist to verify that cases represented false positives.

4.6 Evaluation, validation and results

4.6.1 Performance of tumor ROI detection

During the first step of the multimodal segmentation, the initial tumor mask was obtained using the supervised learning approach defined above. We wanted to evaluate if this method was a feasible and reliable choice for the generation of the initial ROI.

For this purpose, the input preliminary ROIs from the UCLA dataset (n = 260) were analyzed and compared to the reference standard (core tumor measured in post-contrast T1 sequences), displaying a precision of 53% and recall of 73%. The precision metric helps determine what proportion of the pixels included into the ROI are indeed part of the tumor, while recall determines the proportion of pixels that were accurately included into the ROI, with respect to the total number of pixels that compose the tumor region. These results indicate that the preliminary tumor ROI usually generated a loose region that tended to contain the majority of the tumor but also tended to overextend to regions that are no longer tumor, consisting of only normal tissue. The large recall can be interpreted as a large ROI that contained most of the interesting pixels. The small

precision value indicated that regions selected preliminarily as tumor also include many other regions that are not a tumor.

A discussion of the instances where this method failed to find the overall tumor region (7 cases) appears below in the discussion session for this chapter (Table 4.7). A discussion of the results obtained after the segmentation algorithm was completed can be found in Section 4.6.4.

4.6.2 MRI phantom experiments

This validation experiment evaluated to what degree the results generated by the image segmentation method match the real dimensions observed on the real object that is being scanned (*i.e.,* the measured area/volume match). This experiment was performed by scanning an imaging phantom (a physical object with known characteristics) and evaluating the segmentation error found on the measurements performed on the images and the phantom's real dimensions.

The phantom experiments help quantify how much variation captured in the segmentation results correspond to other factors (*e.g.,* noise, scanner variations or contrast artifacts). By measuring the boundary variation when an object of stable characteristics (no variability associated) is segmented, the error associated with other factors besides tumor variability can be quantified (these variations are thought to be small).

The overall procedure includes using the MRI machine to characterize different image phantoms, including a commercially available phantom as well as other tailored phantoms that serve the purpose of evaluating variability in segmentation. The experiments presented in the following subsections used the standard image acquisition sequence used in standard brain tumor scans including spin echo, gradient echo, and inversion recovery MRI sequences. The segmentation error evaluation using the Dice coefficient was performed using the Dice similarity coefficient, trying to prove the hypothesis that calculated volume does not significantly differ from the real physical volume of the object.

4.6.2.1 ADNI phantom

The Alzheimer's Disease Neuroimaging Initiative (ADNI) [95] phantom is a sphere that contains spheres of different sizes and locations. Some of these spheres were segmented and compared to the volume reported by the manufacturer. Table 4.1 summarizes results from this experiment. Figure 4.3 and Figure 4.4 present the segmentation results in the different sections of the phantom.

ADNI phantom	Real volume [mL]	Calculated volume [mL] Mean ± SD
Large (6cm diameter) sphere (n = 1)	113.1	112.77 ± 0
Medium (3cm diameter) spheres (n = 4)	14.14	13.51 ± 0.42
Small (1 cm diameter) spheres (n = 108)	0.52	0.48 ± 0.03

Table 4.1 Segmentation results using the ADNI phantom. The variability in the medium-sized sphere comes from different results obtained for each of the four spheres. For the small spheres, five different measurements were made. There was no variability associated with the single large sphere.

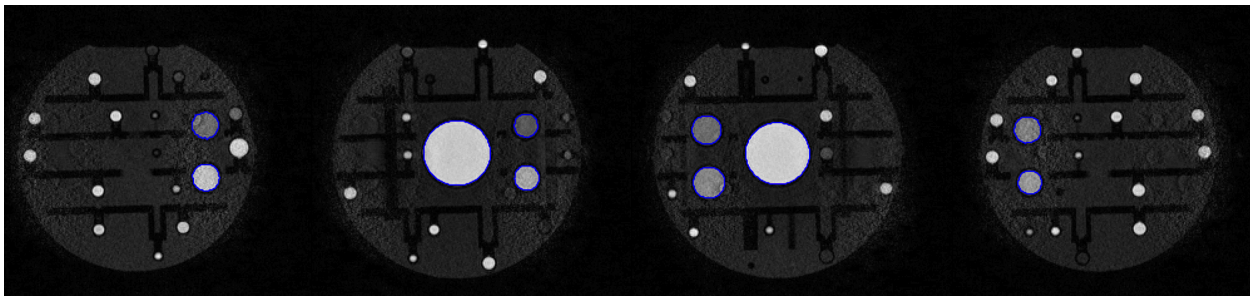


Figure 4.3 Example of the segmentation produced on the different spheres in the ADNI phantom.

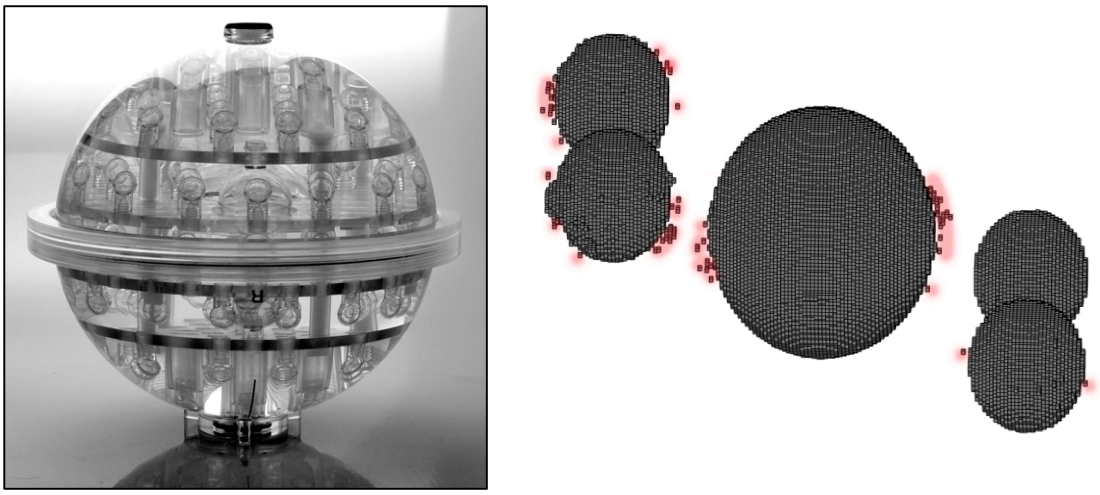


Figure 4.4. The ADNI phantom and 3D visualization of output segmentation that shows some of the false positives and false negatives that influenced the volume calculation. The false positive voxels are indicated with a red shadow. These are small pixel regions that do not correspond to the sphere.). Image obtained from www.phantom-lab.com/magphan-adni/

4.6.2.2 Latex phantom

The purpose of this latex phantom was to determine the performance of the algorithm in cases with irregular shapes and contours. A shape of a glove was utilized for its complicated shape and slight similarity to a common morphological "finger-like" characteristic of infiltrative edema in brain tissues. To fix the sample into the container, a high-density colloidal suspension that changes its molecular configuration with changes in temperature was utilized (*i.e.,* Jell-O). A much higher concentration was used inside the glove to observe different contrasts in the two different regions of the phantom. Table 4.3 summarizes results from this experiment. Figure 4.5 displays the phantom and the segmented result.

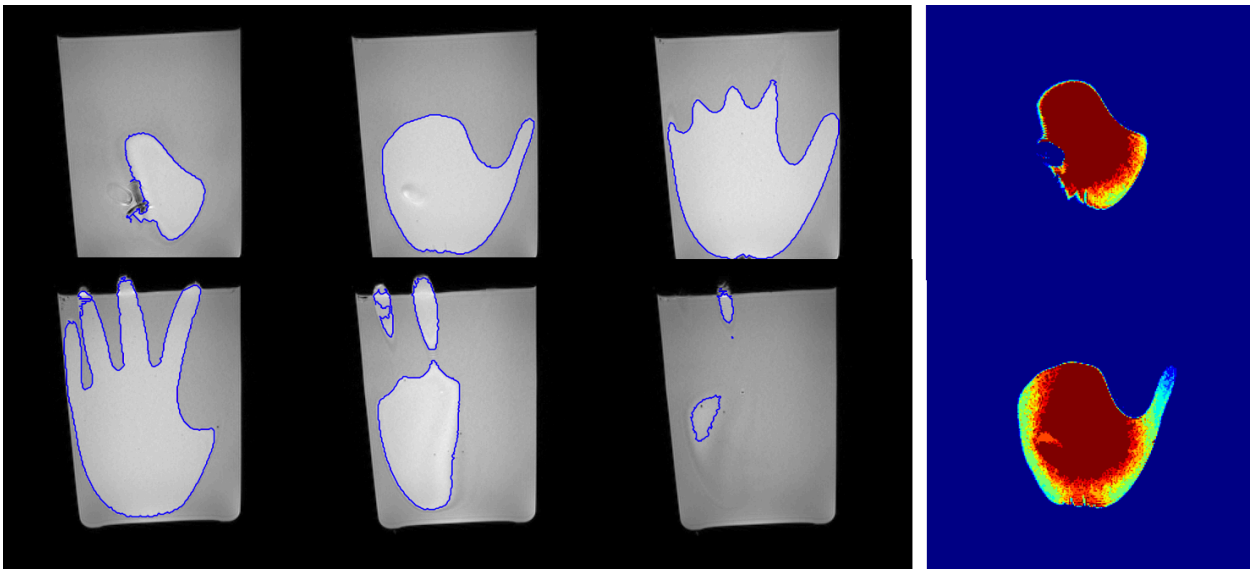


Figure 4.5. Output segmentation of the Latex phantom. Images on the left show the boundary selected for the phantom, and the images on the right show the different estimates calculated based on intensity differences through the image volume.

4.6.2.3 Organic tissues

Similar to the previous phantom, a final experiment was performed with biological tissues (*i.e.,* raw meat) to simulate the heterogeneity of different organic tissues, and their different relaxation characteristics when in the presence of a magnetic field. Two different phantoms were created for this purpose, one with meat in a spherical shape and another with flattened meat. Table 4.2 summarizes results from this experiment. Figure 4.6 displays the phantom and the segmented result. From these results, we can observe that in general, the segmentation on T1 produced results that in general overestimated (meatball) or underestimated (burger) the real volume while the results segmented on FLAIR produced the most accurate results. Reasons for this difference in performance might be related to the different physical properties of the protein/fat/water content of the sample that affect the contrast observed under the different sequences, being more easily processed on FL and T2 sequences in this case.

Phantom	Real volume [mL]	FLAIR [mL]	T2 [mL]	T1 [mL]
Latex phantom	Total = 500 mL	495.1±5.1	**492.5±4.4**	**488.7±6.2**
Organic tissue phantom 1 (meatball)	Total = 324 mL	327.8±9.3	**317.9±1.2**	**367.3±7**
	Large = 165 mL	169.1±7.4	**153.4±6.1**	**175.5±8.0**
	Medium = 105 mL	100.4±2.8	98.6±3.5	**116.1±4.6**
	Small = 54 mL	58.2±2.5	55.8±1.4	**75.7±2.5**

Organic tissue phantom 2 (hamburger)	Total = 217 mL	214.1±2.1	**219.1±2.7**	**200.3±4.7**
	Piece 1 = 117 mL	**112.7±1.5**	116.8±2.1	**105.8±4.3**
	Piece 2 = 100 mL	101.3±2.5	102.2±3.3	**94.5±5.2**

Table 4.2. Assessment of the different tested phantoms, comparing real volume and measured volume on the image scans using the proposed segmentation algorithm. The estimates of variability were obtained by performing additional segmentation tests on the objects. The result shown in bold indicate the results that are significantly different (using a Student's t-test) from the real volume measurement.

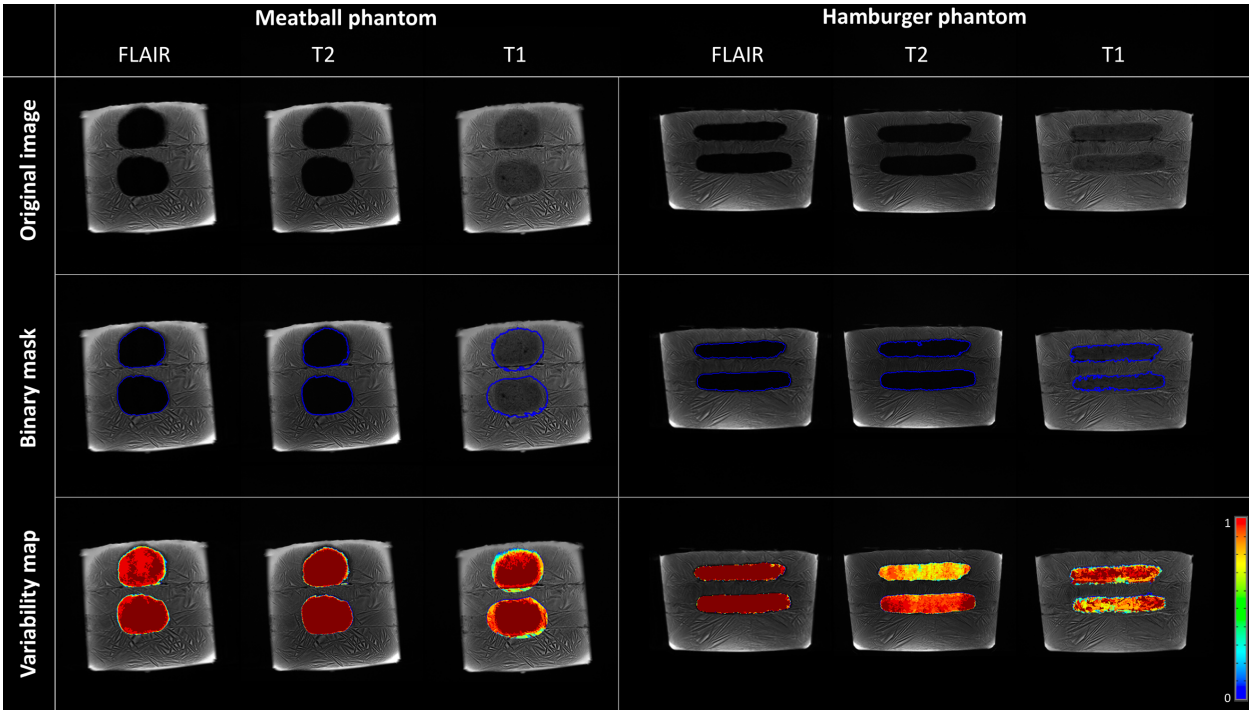


Figure 4.6 Output segmentation of the organic tissue phantoms. The phantom with spherical shapes appears on the left, and phantoms with a flattened appearance on the right. There is a distinctive tissue heterogeneity in these phantoms, different from the homogeneous characteristics observed on the ADNI or glove phantoms.

4.6.3 Evaluation of MICCAI 2015 data

This approach was validated in all 220 cases of the BRATS 2015 dataset, segmenting three components: whole tumor, tumor core (enhancing and necrotic components), and active tumor (enhancing component).

The Dice coefficient for total tumor mass of 0.74 (median: 0.77, 1st quartile: 0.66, 3rd quartile: 0.84), 0.54 for the tumor core (median: 0.57, 1st quartile: 0.37, 3rd quartile: 0.75), and 0.54 for the active tumor (median: 0.60, 1st quartile: 0.29, 3rd quartile: 0.76). Figure 4.7 displays the Dice coefficient as well as the Hausdorff distance metric for this dataset.

Figure 4.8 displays examples of input images and the output tumor variability maps (confidence maps) for edema, enhancing and necrotic regions of the tumor, as well as the representation of these tumor compartments overlaid on the image as binary masks. This binary representation of the TVMs (as previously described) is performed with the purpose of similarity computations against the binary reference. This facilitates easy integration with other processing pipelines (evaluation of clinical variables, genetic algorithms, etc.) or statistical approaches that require a binary input (Figure 4.9).

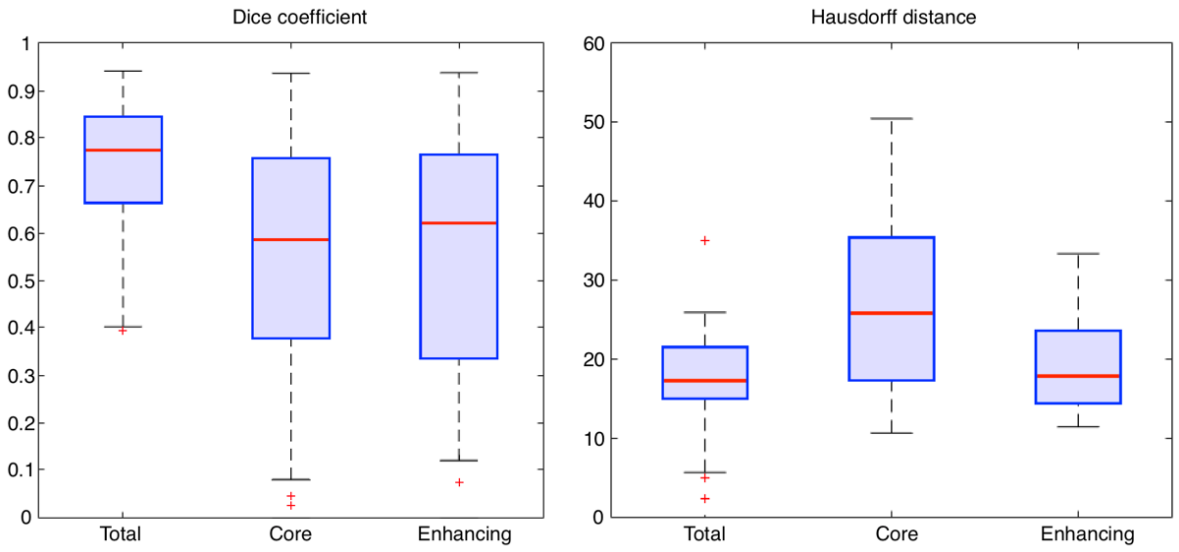


Figure 4.7. Box-plots showing the proposed method's performance for the BRATS 2015 dataset. It plots the Dice similarity coefficient on the left (higher is better) and Hausdorff distance on the right (lower is better) when comparing a binarized tumor mask generated by the proposed approach and a segmentation reference standard. Label "Total" refers to all tumor components (edema, enhancing, necrosis and non-enhancing tumor), "Core" refers to the tumor core (excluding regions of vasogenic edema), and 'Enhancing' refers to active tumor cells with microvascular proliferation.

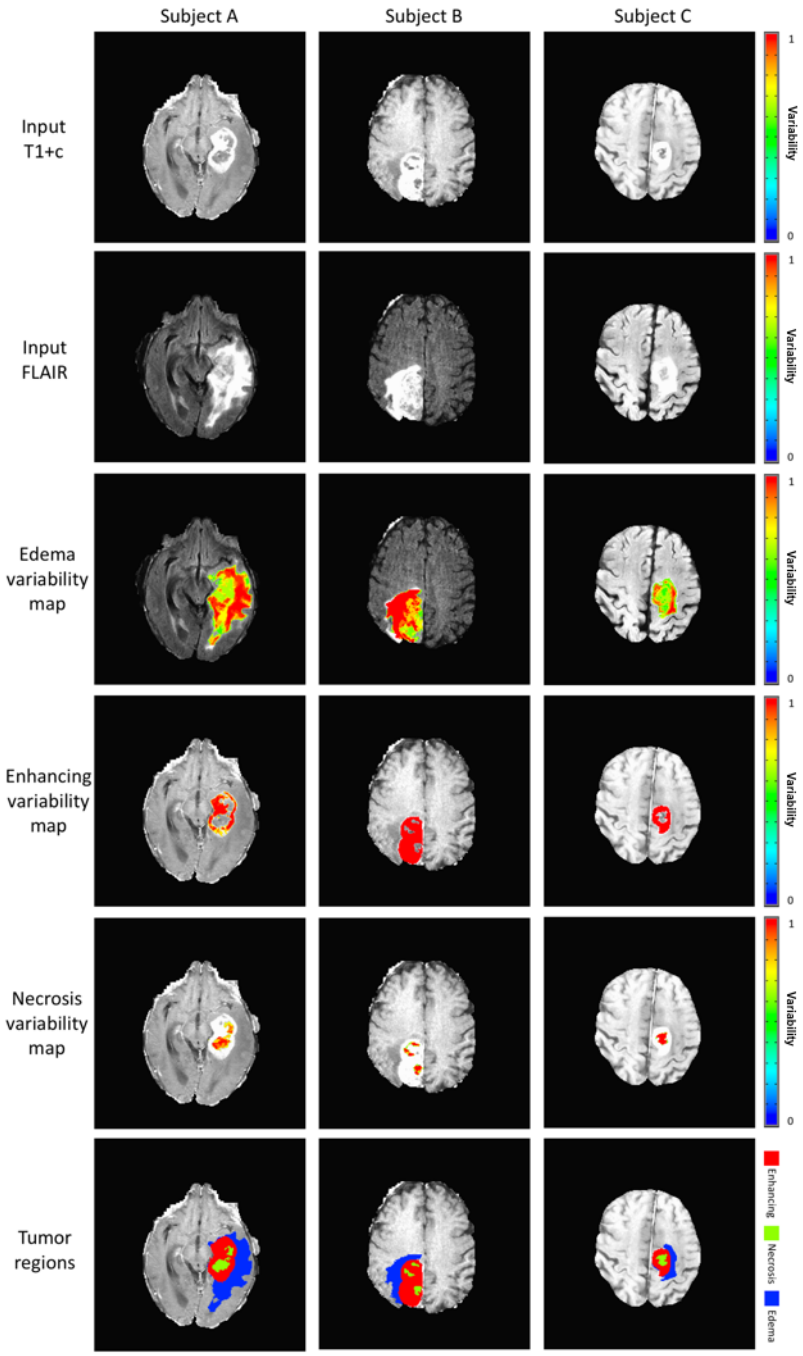


Figure 4.8. Segmentation results for different subjects. First and second rows represent the input post-contrast T1 and FLAIR scans for each subject (columns). Rows three to five show different boundary estimations represented as variability maps in each sub-component (edema, enhancing, and necrosis respectively). The color bar represents the pixel-wise probability for each case. The bottom row displays a color-coded binary mask for tumor sub-components: enhancing (red), edema (blue) and necrosis (green).

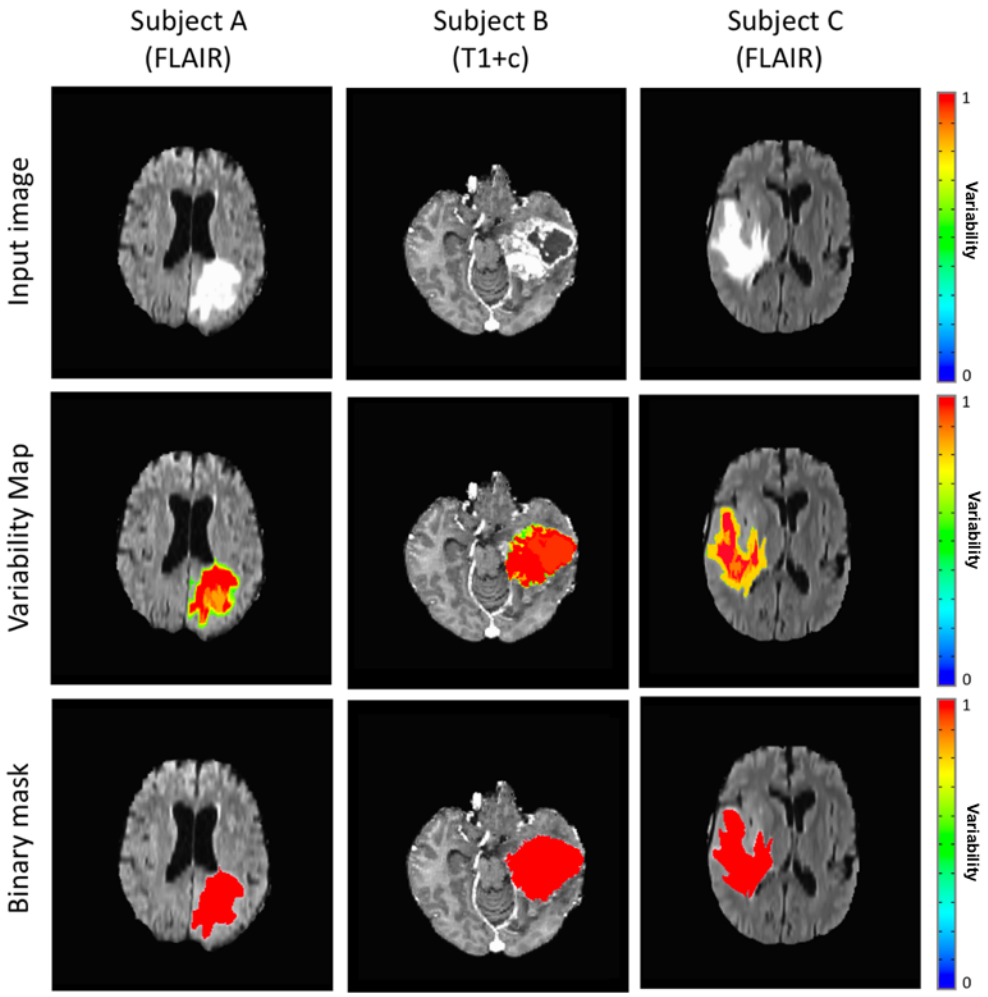


Figure 4.9. Examples of brain tumor segmentation results for total tumor mass in three different cases. The first row shows one of the four imaging sequences that are used to perform the tumor segmentation. FLAIR is used in this example for the examples on the left and right and T1+c is used for the example on the center. The tumor variability map is displayed for a total tumor in the second row, highlighting in red the regions where the algorithm is almost certain that there is an abnormality in that pixel region. Other color intensities indicate decreased belief in tumor abnormality according to the color bar located on the right. The third row shows the binarization of the variability map, according to a majority agreement of the different estimates using a threshold of 0.5. This enabled comparison with binary reference standards to evaluate the accuracy and also allowed for other more standard metrics and use of results generated by other processing pipelines.

4.6.4 MICCAI BRATS2016 segmentation competition

The Multimodal Brain Tumor Image Segmentation (BRATS) benchmark was initiated in 2012 as part of the Medical Image Computing and Computer Assisted Interventions (MICCAI) conference. The purpose of this competition included the analysis of several segmentation methods with a common dataset to objectively evaluate the performance of each. a second objective was to define which quantitative evaluation could better adjust to a reference standard annotated and agreed to by seven different raters. The BRATS image data and manual annotations are publicly available through an online evaluation system as an ongoing benchmarking resource. Advantages of the participation in this competition include having the algorithm evaluated by an external evaluation group, and having access to a large set of annotated data. This dataset contained different examples of non-enhancing tumor cases, which were not common in other datasets utilized (UCLA), providing a comprehensive set of training data to improve the algorithm classification of non-enhancing tumor tissues. Figure 4.10 displays some of these cases.

As an external validation of our method, we participated in the 2016 segmentation challenge using our proposed method. Similar to other years, the competition centered on the accuracy in identification and delineation of the overall tumor boundaries and different subcomponents -binary boundaries without consideration of segmentation variability. New cases added for 2016 included cases with more than one scan, adding a

temporal component into the challenge. The extra task determined if the tumor had progression, remission, or was unchanged based on the tumor volume change over time.

The dataset comprised 220 training cases with a reference standard provided for each case, and 191 test cases. Similar to the dataset described in Chapter 3 (Table 3.1), the set included all four image sequences, and input images had a resolution of 240 x 240 x 154 pixels.

Results for this competition had not been published at the time of the publication of this dissertation. Preliminary results indicated that our approach had a mean Dice coefficient of 0.75 (total tumor), with an approximate variability range of 0.60 to 0.85.

Preliminary results are shown in Figure 4.11. It was observed that as in previous years (Figure 2.3) there was a wide variability in results achieved by the different approaches, ranging from less than 0.40 (worst performance) to more than 0.80 (best performance). Figure 4.12 displays some of the results obtained for this dataset.

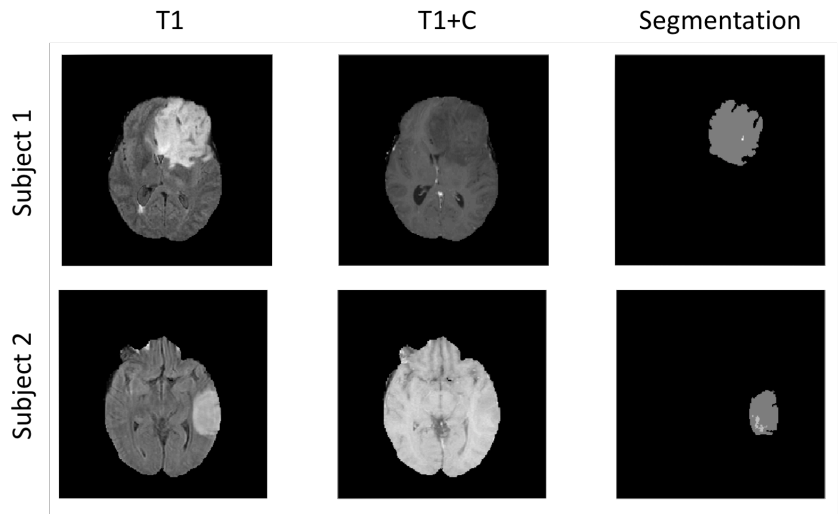


Figure 4.10. Examples of non-enhancing tumor tissues in the 2016 MICCAI dataset.

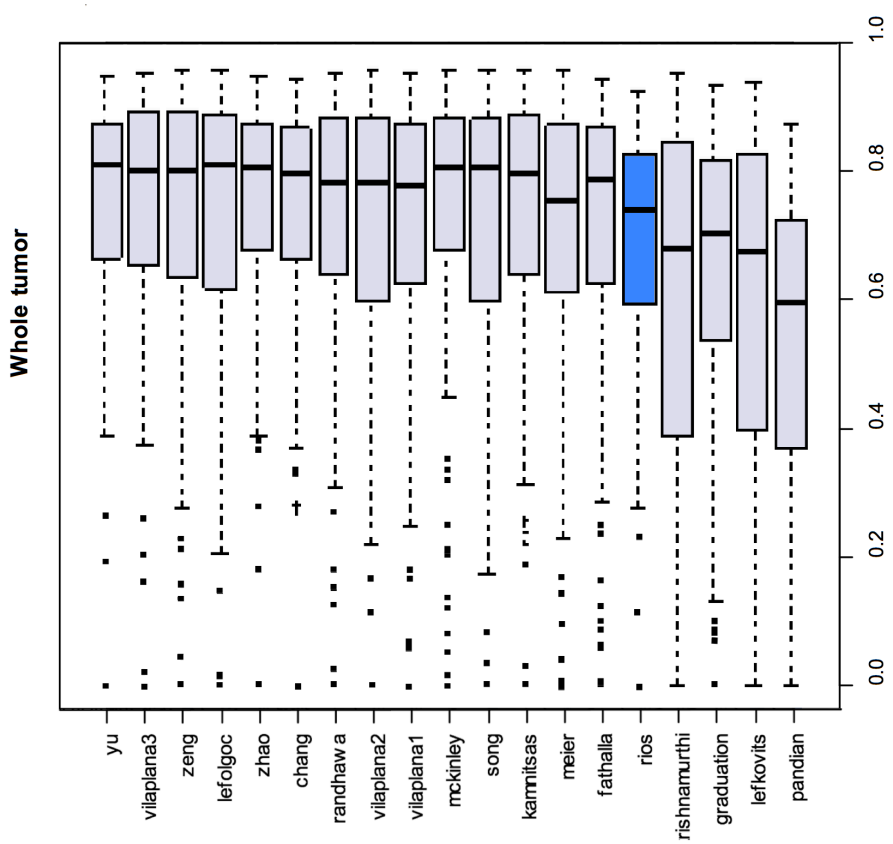


Figure 4.11. Results for the BRATS tumor segmentation competition 2016. Adapted from Menze *et al.* from the Neural Information Processing Systems (NIPS) conference 2017 and MICCAI tumor segmentation competition 2016 [96]. Full results of this competition, including the temporal assessment, had not been released at the time this dissertation was completed. The submission shown in solid blue indicates the result achieved by our team in the test data with respect to other participants.

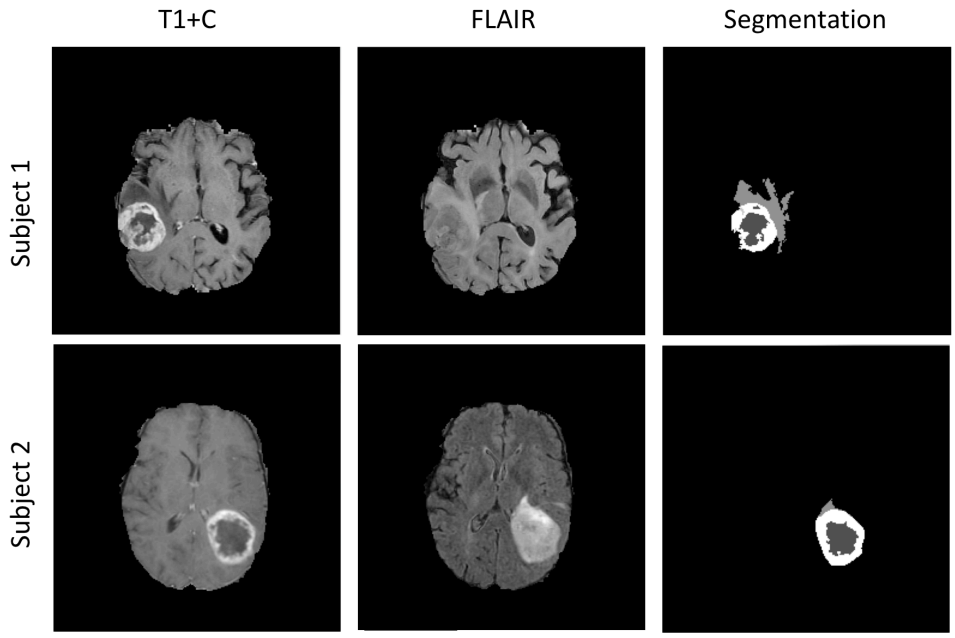


Figure 4.12. Examples of results obtained in the test set for the 2016 MICCAI dataset. Note that the input images were preprocessed, and the reference segmentation was not available for these examples.

4.6.5 Evaluation of UCLA dataset

The knowledge-based algorithm was evaluated using a large set of tumor cases provided by the neuro-oncology program at UCLA. The dataset consisted of 300 different cases. A complete characterization of this dataset was described in Section 3.2.

A unique feature about this dataset was that several of the cases had a temporal component (*i.e.,* more than one follow-up), thus allowing analysis of progression and clinical assessment for these cases. Section 5.3 elaborates on the clinical applicability of the variability maps using data obtained from this dataset. The segmentation reference

standard from this dataset was manually generated by expert annotators, as previously described in Section 3.2.1.

The total sample included 260 cases drawn from the 300 cases from this dataset. Section 3.2.2 presented detailed characteristics of this dataset and the sample size calculations. Table 4.3 summarizes cases that were successfully processed, as well as the causes for cases that were not able to go through the complete pipeline.

Table 4.4 presents the performance of the proposed knowledge-based segmentation approach compared with the multiple segmentation methods tested in Chapter 3. It is evident that using the UCLA dataset (which was used mainly for testing purposes) our proposed method had a better overall performance with respect to other approaches. However, when measuring the performance on the TCGA/MICCAI dataset, the edge-based and multilevel algorithms had a higher Dice similarity coefficients. Reasons for this can include that the output ROIs produced by these methods were able to be tailored more specifically for the current dataset and that the output ROI was more in agreement compared to the other methods (number of false positives vs number of false negatives inside the output mask). This highlights the utility of having an ensemble of models that attempt to delineate tumor boundaries.

Lastly, the knowledge-based method outperformed (on both datasets) the ensemble output generated by aggregating other five methods (section 3.6.6.2), which can be

an indication that pixel-level variability is better suited to characterize the variability observed in brain tumors.

Similarly, Figure 4.13 displays the overall performance of cases from the UCLA cohort compared with the previous cases that were obtained from the MICCAI/TCGA dataset Chapter 3), and Figures 4.14 a-c display the tumor variability maps as well as the binary masks obtained for some of the cases. Examples of the temporal component of this dataset are shown in Section 5.3.

4.6.6 Evaluation of the utility of T2-weighted sequences for brain tumor segmentation

The MRI sequences routinely used for brain tumor segmentation include two T2-weighted sequences, FLAIR, and T2. These two sequences provide similar insights about tumor extent, mainly the infiltrative edema.

Given that it is important to reduce scan time and optimize the time under which the patient is examined, this experiment aims to evaluate the contribution of these sequences towards accurate segmentation results (together and individually). Consequently, an assessment of a random sample of 30 different cases (10 small tumors, 10 medium sized tumors, and 10 large tumors) was drawn from the TCGA/MICCAI dataset (because all sequences where available) and was reanalyzed to observe differences when one sequence was removed from the analysis.

Results are presented in Table 4.5. The combination of both sequences proved to be ideal for improved performance; but in case only one was utilized, the FLAIR sequence outperformed T2. A major reason for this change is the characteristic of such sequence to null the signal from fluids such as cerebrospinal fluid and water that can be classified as the tumor (infiltrative edema). Also, the increased contrast observed in FLAIR between normal cerebral tissue and hyper-intense edema with respect to the T2 can play a role in the improved performance of FLAIR over T2.

4.6.3 Concluding remarks

In this chapter, we presented a novel algorithm to determine the tumor boundary as well as an estimate of its variability evaluated from the pixel-level perspective. This was accomplished by using a combination of statistical approximations to the normal brain tissues, the characterization of the intensity differences observed near the boundaries of the different tumor components, and a system that provides a way to reduce the rate of false-positives by comparing the current cases with previously observed examples.

This method was tested using phantoms with different physical characteristics and diverse ranges of variability to determine how accurate the results can approach the real object. Finally, we evaluated this method in both the UCLA and TCGA/MICCAI datasets, including an external evaluation that occurred during the BRATS segmentation competition held in 2016.

Number of cases	Processing outcome	Comment
260 (86.66%)	Completed successfully	Tumor variability maps and binary maps were successfully generated. All steps completed without error.
6 (2%)	Error while generating prior brain tissue probability maps	Two different causes were identified for this problem. One includes abnormalities in the intensity distribution (due to artifacts, inappropriate sequence that causes images overly bright or dark) of input images (since acquisition). Another problem is the missing point of origin in the three- dimensional volume (usually the anterior commissure). The proposed pipeline has a built-in automated method to assign the origin to this cerebral structure that worked in several of cases successfully processed.
23 (7.66%)	Image registration failed to complete.	Due to extreme anatomical deformation that brain tumors can cause, sometimes the registration can produce images that are empty or that contain erroneous shifts, rotations, and deformations on the output images. Although the registration of the images is critical for the adequate identification of tumor sub-regions and visualization of tumor progression through time), a downside is that this process has to work for all input sequences, which is not always the case. Ways to fix it include the implementation of the processing framework with an additional registration algorithm done one time.

7 (2.33%)	Error while obtaining tumor initial region of interest	When the initial tumor ROI from the extracted imaging features and prior tissue distribution maps is not found (empty image), the subsequent process cannot be completed. Methods to solve this issue include the segmentation under different planes and orientations built-in into the program too.
4 (1.33%)	Tumor reference and automated result cannot be compared properly.	This error occurred when the segmentation process completed successfully but the transformation to the original space (inverse registration) in order to compare it to the reference was not completed. Reasons include the automated and tumor ROIs do not match (different sizes) due to an incorrect inverse registration, or the manual reference could not be found.
Total = 300		

Table 4.3. Summary of results for the UCLA dataset, showing the number of cases that were successfully processed and the cases that did not complete the process.

	Knowledge-based	Methods ensemble	Region-based	Edge-based	Level-set segment.	Multilevel segment.	Statistical classifier
UCLA dataset (n = 260)	0.636	0.572	0.528	0.585	0.609	0.594	0.522
TCGA/MICCAI dataset (n = 85)	0.738	0.711	0.614	0.721	0.789	0.760	0.631

Table 4.4. Performance of segmentation algorithms in different datasets. The reported Dice coefficients correspond to core tumor (enhancing + necrosis components). Results for TCGA/MICCAI row are the same as the ones reported in Chapter 3. The sample size was determined according to results presented in Table 3.5

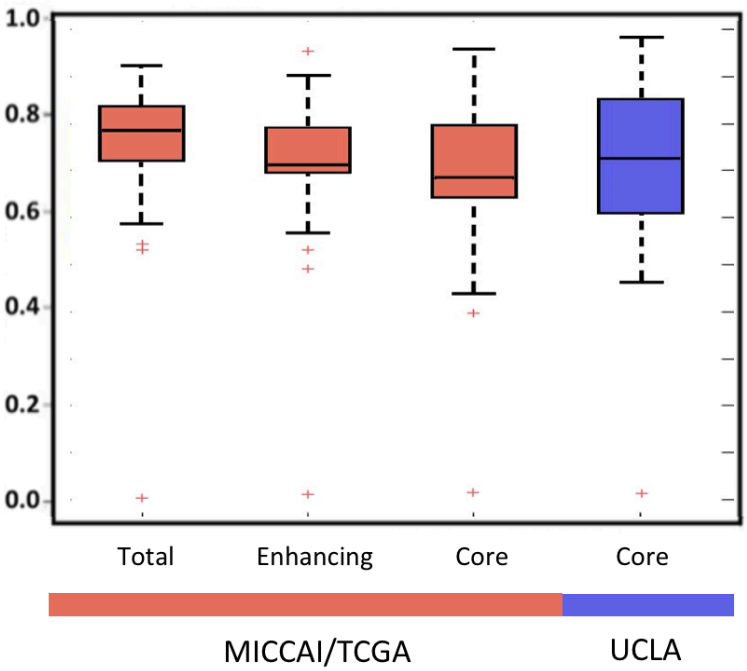


Figure 4.13. The overall performance of the knowledge-based segmentation method in the UCLA dataset (260 cases) compared to the TCGA/MICCAI 2015 cases (131 cases). Note that the only tumor core was assessed for this analysis because a manual reference was not available for the other tumor subcomponents. It can be seen that performance of the segmentation approach for core tumors was better in the UCLA set. Some possible reasons include a bigger sample size in the UCLA set and higher proportion of non-enhancing tumors difficult to segment for all state-of-the-art methods in the TCGA/MICCAI datasets.

	Both inputs (FLAIR & T2)	FLAIR only	T2 only
Small tumor sizes	.695	.678	.633
Medium tumor size	.714	.708	.671
Large tumor size	.742	.712	.704

Table 4.5. A pilot study to evaluate the value of information provided by T2 and/or FLAIR to the segmentation of total tumor. It can be concluded by the results that the contribution of the FLAIR sequence is greater than that provided by T2. Tumor sizes were considered based on total tumor volume according to the information found on table 3.6.3.

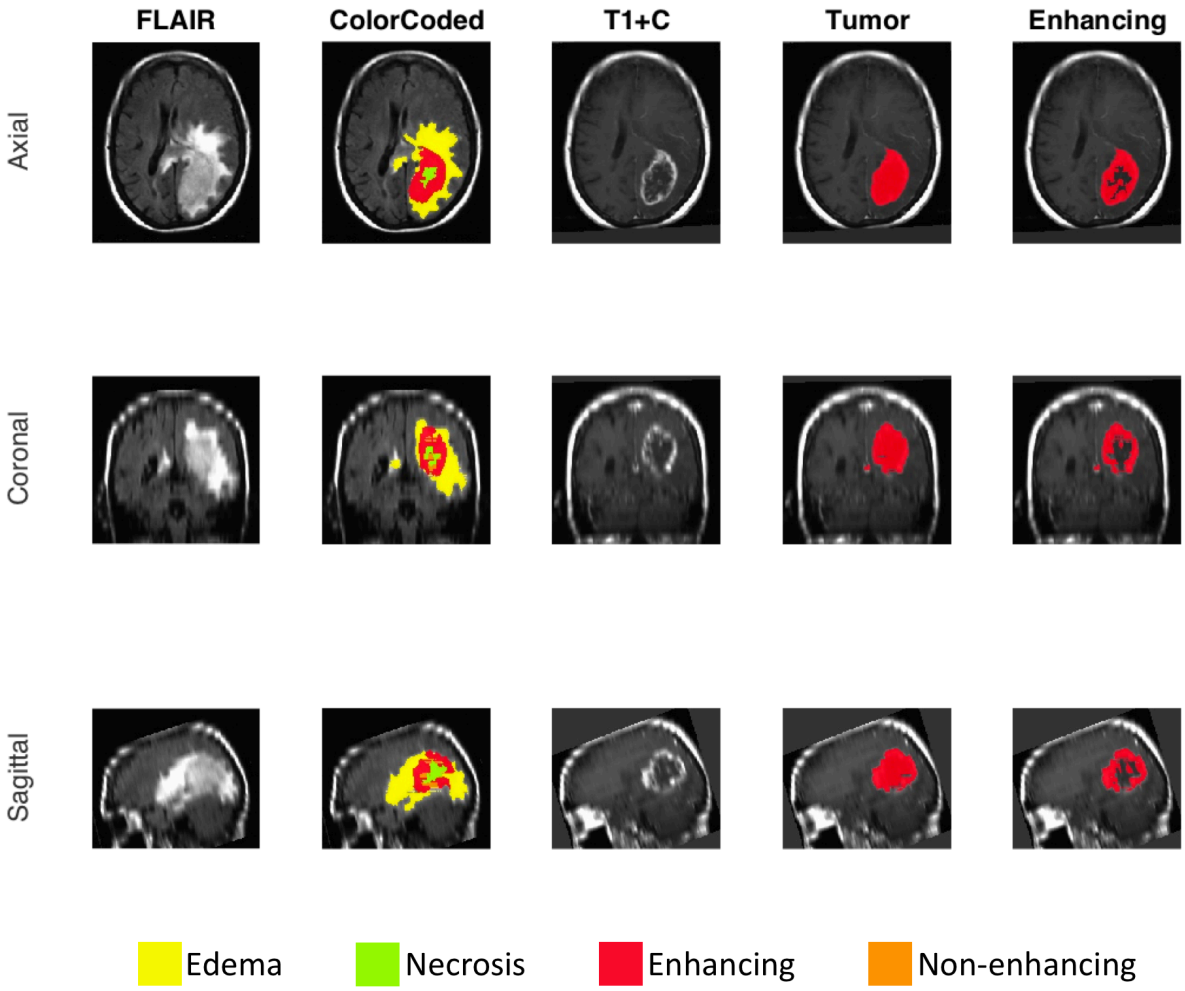


Figure 4.14a. Example 1 of the output segmentation from the UCLA dataset showing three different views of the same patient to visualize the three-dimensional result. The different tumor components displayed using different colors. The "Tumor" and "Enhancing" masks are the binary results derived from the Tumor Variability Maps.

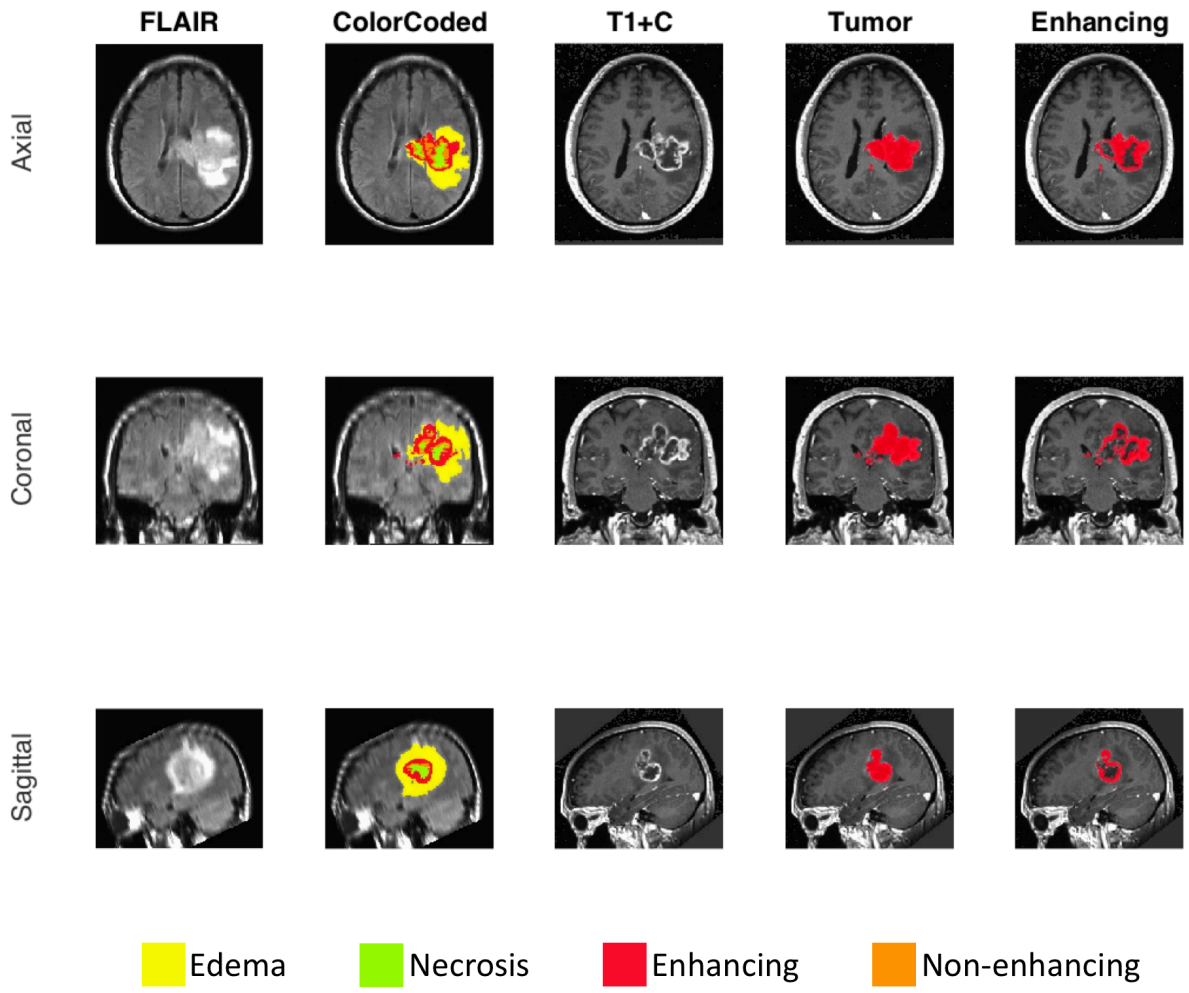


Figure 4.14a. Example 2 of the output segmentation from the UCLA dataset showing three different views of the same patient to visualize the three-dimensional result. The different tumor components displayed using different colors. The "Tumor" and "Enhancing" masks are the binary results derived from the Tumor Variability Maps.

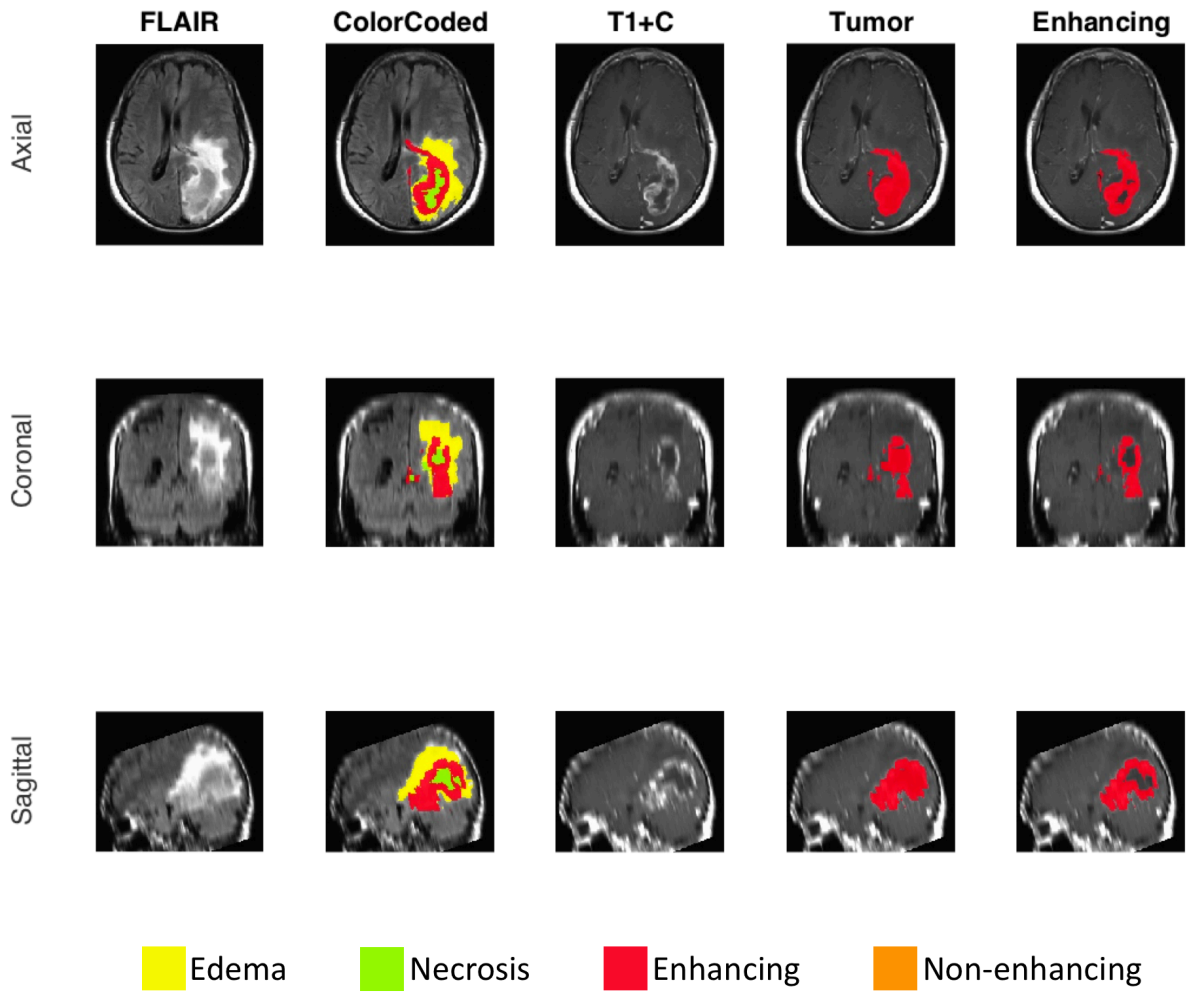


Figure 4.14a. Example 3 of the output segmentation from the UCLA dataset showing three different views of the same patient to visualize the three-dimensional result. The different tumor components displayed using different colors. The "Tumor" and "Enhancing" masks are the binary results derived from the Tumor Variability Maps.

Chapter 5: Impact in the Clinical Setting

5.1 Overview

In this chapter, we describe our efforts to incorporate the segmentation methods described in prior chapters into a prototype neuro-oncology workstation with the goal of providing insights about the utility and applicability of the developed tumor variability metrics into clinical practice. We describe the workstation (Glioview) followed by subsequent (prospective) evaluations of the workstation.

First, we introduce the user interface that implements the segmentation algorithm introduced in the previous chapter, explaining how it is structured, and detailing its overall contribution of displaying the segmentation variability estimate as part of the visualization and clinical assessment.

Then we perform two evaluations, the first being a study that aims to determine the overall utility and understanding of the different visualizations to represent boundary variability. A second more clinical study is then presented in which a series of retrospective cases are analyzed to determine if a more quantitative determination of change can lead to more informed decision making.

We conclude the chapter detailing the results found in both evaluations, with emphasis on the opinions generated by the neuroradiologists with respect to its potential in a clinical setting.

5.2 Glioview: An application to visualize boundary variability.

An important component of the experiment covered in this chapter includes the use of an appropriate method to display the temporal progression, variability estimates, and segmentation results to the user (*i.e.,* neuroradiologist). For this purpose, a new visualization interface (Glioview) was developed, on which the different imaging studies, as well as the visualization of variability estimates are displayed.

As shown in Figure 5.1 and Figure 5.2, Glioview is able to load the desired images and run the automated segmentation approach described in Chapter 4. Additionally, it has diverse tools that aid the neuroradiologist to interpret and understand the variability metrics generated after the computations have been completed. Some of these visual aids include a boxplot with variability estimates, a timeline with variability estimates, a bar plot with the calculated volume for each of the tumor components and, most importantly, the tumor variability maps that display the degree of boundary variability across the tumor volume.

A demonstration of the functionality of this application can be found on the following YouTube permanent link (*https://www.youtube.com/watch?v=MhhqCOkOlHM*).

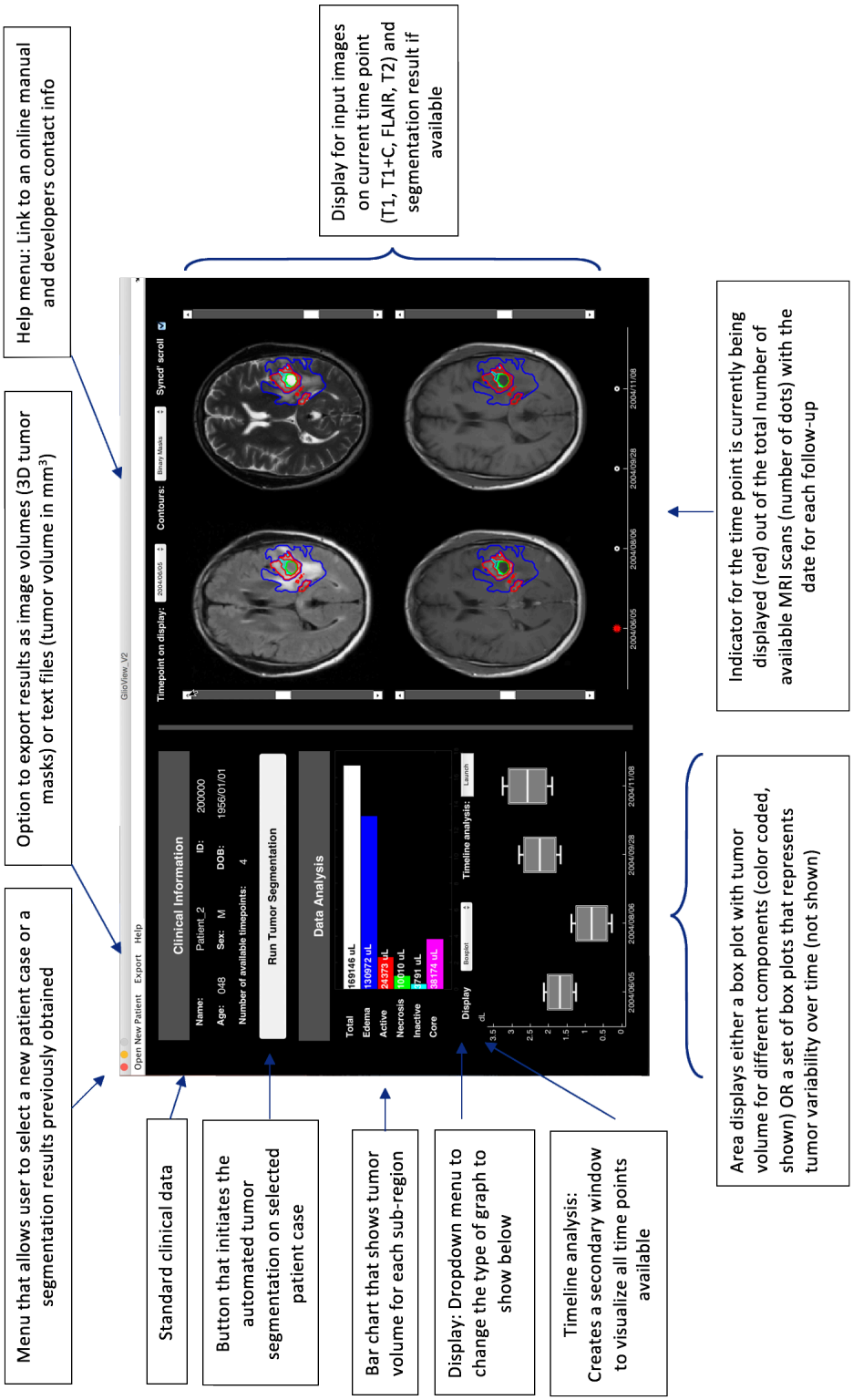


Figure 5.1. Glioview user interface. This is the main interface where the user can load, process, and visualize the different contrasts for a given time point. A description of the overall functionality is provided for each component of the interface.

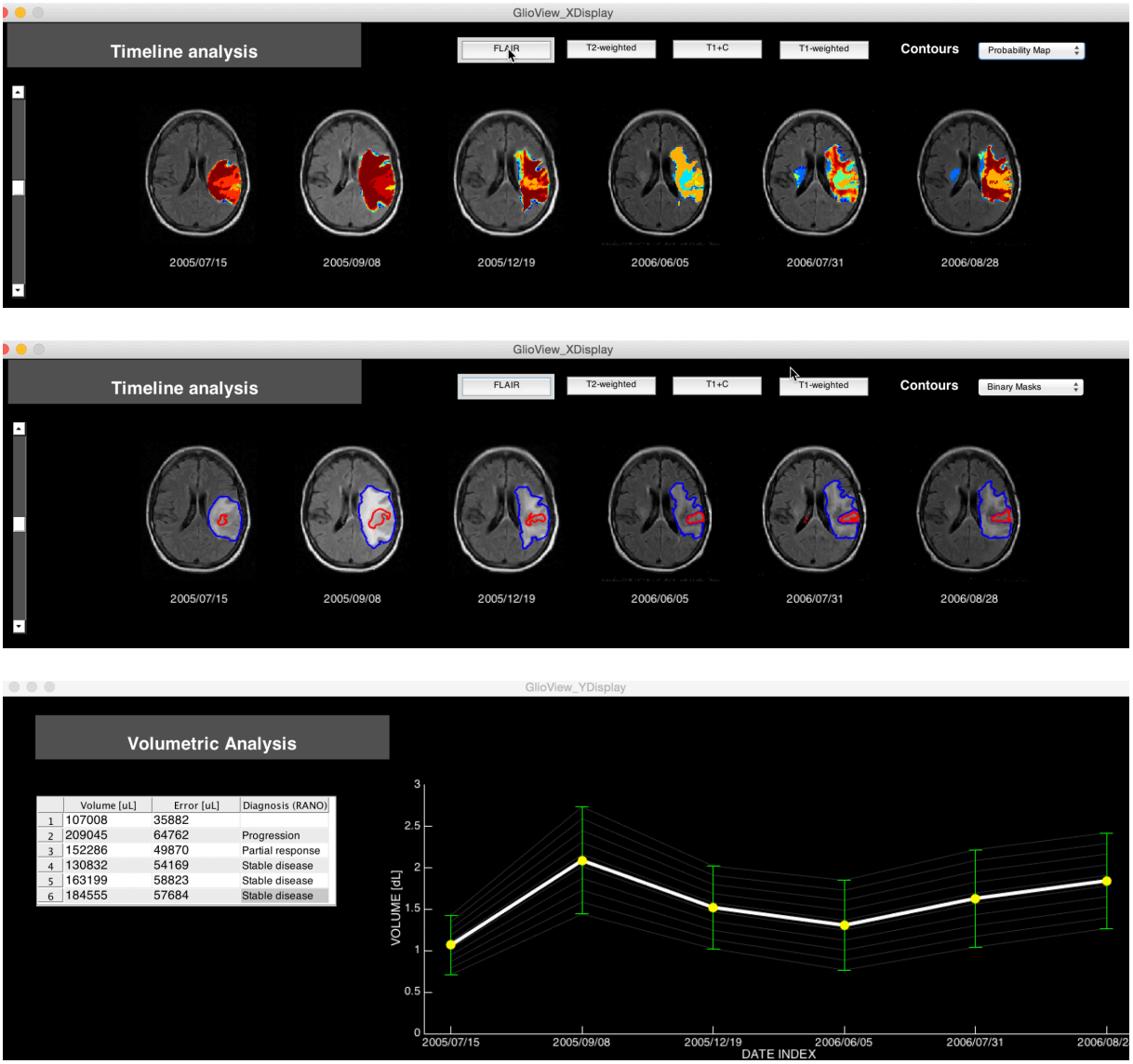


Figure 5.2. Additional windows generated by Glioview for temporal evaluation and assessment. The top image displays the tumor variability maps as discussed in Chapter 4 for each of the follow ups available for this case. The figure in the middle display the derived binary maps (contours over the image volumes). The bottom image shows a timeline with the estimated volume and associated variability for each of the cases, together with an automated assessment for each transition.

5.2.1 Radiological utility study

To assess the utility of Glioview for the practicing clinician, a utility study was conducted among neuroradiologists to evaluate whether our computer-aided system for automated image analysis can alleviate the challenges of image interpretation in a subset of GBM patients.

In this exploratory study, eight experienced neuroradiologists evaluated the automated tumor segmentation application on a set of six different patients with multiple follow-ups, and subsequently, answered a survey that consisted of ten questions. This survey was created with the help of Dr. William Hsu and Dr. Mary Zide [98] specifically to evaluate this new interface. The questionnaire included an evaluation of the user interface (intuitiveness of single and multiple data points), display (data display and arrangement), understanding of data (understanding probability maps and graphs), and program utility (performance, work facilitation). The questions utilized in the questionnaire are shown in Table 5.1; most of the responses were recorded using a Likert scale (1 to 5 ranking) [99,100].

Question	Possible answer
Overall, I think the application was intuitive (*i.e.,* easy) to use.	Likert scale (1-5)
Overall, I was satisfied with the way the data is displayed.	Likert scale (1-5)
I completely understood the meaning of the tumor variability maps generated by the application.	Likert scale (1-5)
I am overall satisfied with the performance of the program.	Likert scale (1-5)
I think this application can facilitate my work (radiology).	Likert scale (1-5)
Have you used/seen another application for automated brain tumor segmentation before? (Yes/No answer)	Likert scale (1-5)
Have you used/seen another application that incorporates analysis of different patient follow-ups (temporal analysis) in the results?	Yes / No
Which set of results did you find to be the most informative (most added value) for clinical purposes?	Bar graph with tumor regions Patient timeline Variability maps Binary masks

Table 5.1. Set of questions the clinician answered after using the user interface for tumor variability estimation.

5.2.1.1 Results for the utility study

Results were gauged using a Likert-type scale ranging from 0 (no agreement) to 5 (complete agreement) (bar chart below) and found high agreement (4.62 ± 0.51) regarding the user interface being helpful to better evaluate data, while a smaller proportion indicated it would improve their diagnostic workflow (4.12 ± .99). Additionally,

visual aids that can display error-range in tumor measurements (*e.g.,* timeline) were identified by the users (six out of eight) to be significantly helpful during the decision-making process. Figure 5.3 and Figure 5.4 summarize the results of the questions in the utility study.

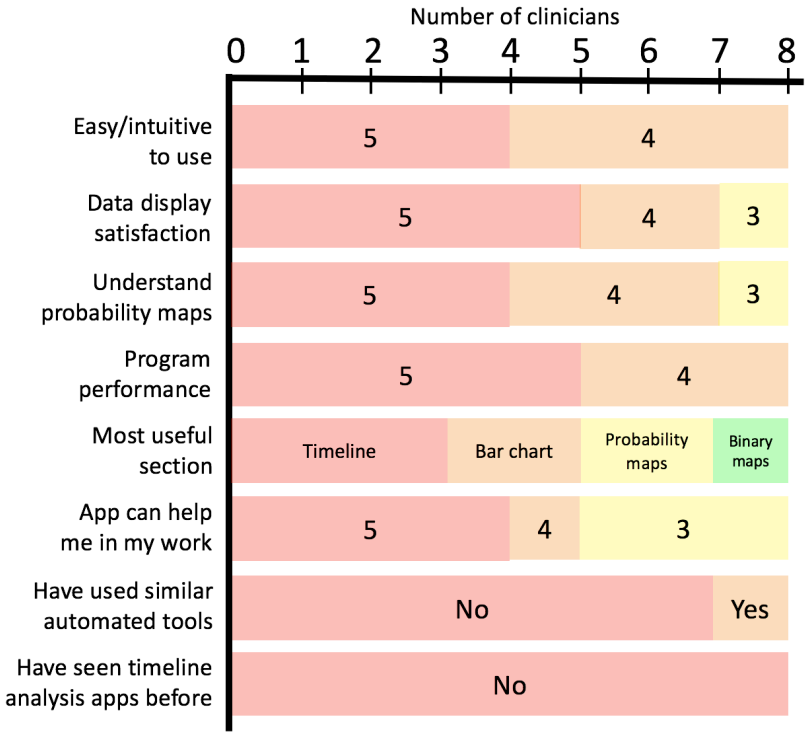


Figure 5.3. Distribution of answers for each question divided by the number of clinicians that answered the questions

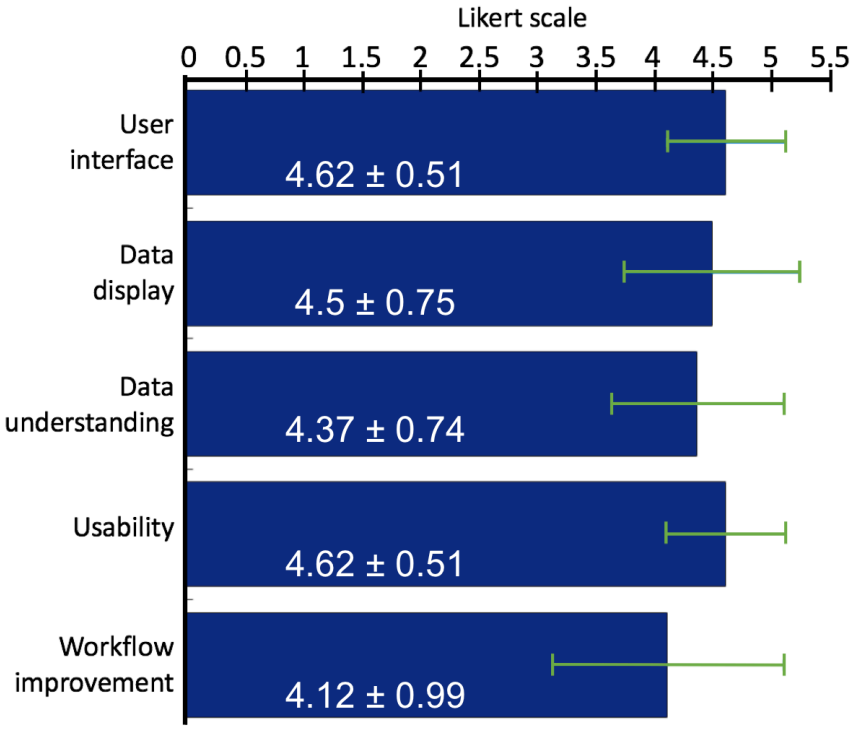


Figure 5.4. Summary of the answers by category of the questions in the utility study. The actual numbers are shown inside each of the bars.

5.2.1.2 Discussion for the utility study

With this exploratory utility study, we demonstrated that computer-aided approaches have a positive effect on data understanding and effective work of neuroradiologists. Volumetric analysis proved to be helpful in analyzing brain tumor cases. With the feedback and comments received during this evaluation, we found possible ways to improve the visualization. These included the display of variability estimates for T1-weighted and T2-weighted estimates at the same time and display of the variability range in the bar-chart.

5.3 Retrospective analysis

Given that the clinical evaluation using the RANO criteria [9,45] considers tumor characterization from the perspective on tumor volume changes through time, this experiment centered on the analysis of a sample of patients from the UCLA dataset (section 3.2) for which temporal data was available. A retrospective analysis comparing the assessment from the neuroradiologist at the time of the scan and a new assessment by expert neuroradiologists (Dr. El-Saden and Dr. Salamon) was performed. Given that the imaging scans were the same and the skill set of both clinicians was assumed to be equal, the critical comparison between both assessments was the availability of quantitative metrics for tumor boundary variability. This was provided by the segmentation algorithm presented in the previous chapter. The overall process to perform this experiment is presented in Figure 5.5.

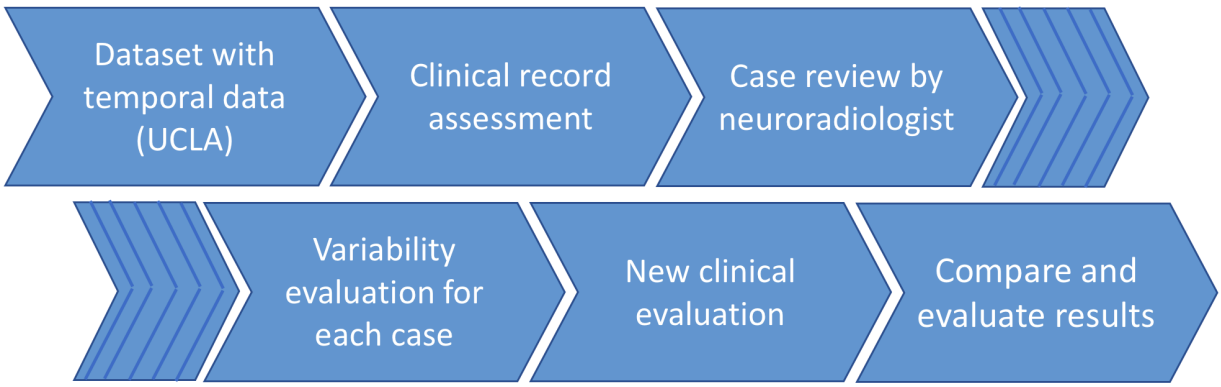


Figure 5.5. Flowchart of the process followed for the assessment in clinical practice. Starting from the selection of temporal cases and clinical information until the new assessment was reached, and compared with the previous assessment when the variability for each case was available.

The driving hypothesis that was followed for this experiment is that the variability metrics generated by the approach proposed in this dissertation (Chapters 3 and 4) can be leveraged by the neuroradiologist towards a better understanding of cases with high variability over several follow-ups. The primary aim was to achieve a more objective, specific and accurate assessment of the case during review.

These new assessments were compared with the previous clinical analyses, and with the conclusions achieved by the automated software. For the purposes of this experiment, the reference standard was considered to be the automated evaluation obtained by measuring the volumetric change using RANO (volumetric) criteria, taking into consideration the tumor volumes generated by the automated approach that were verified to be accurate.

5.3.1 Study cohort

A subset of 31 patients from the UCLA dataset was selected to perform this evaluation following the sample size calculation in Table 3.5 (80% power using a ROC curve power analysis calculation). These cases had a temporal component, having a minimum number of follow-up scans of two and a maximum number of follow-ups of eight. The total number of cases analyzed for this analysis included 93 different follow-ups. The inclusion criteria included randomly selected cases with multiple follow-ups, without resection (but including biopsy) and with previous clinical notes available (Figure 5.6)

For all cases, the clinical information was also available to obtain the previous diagnosis. The clinical notes were accessed and manually read to generate a correlation between the clinical notes (free-text) with the volumetric RANO criteria.

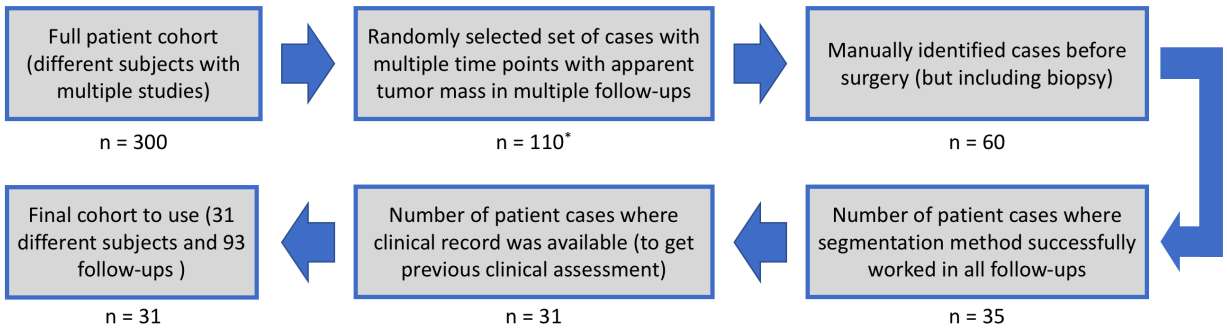


Figure 5.6. Inclusion criteria for the retrospective pilot clinical study. *Number of cases limited to 110 because of time constraints and the intensive manual nature of image examination.

As discussed in previous chapters, the proposed volumetric assessment using the RANO criteria considers a volumetric measurement on which a decrease in overall volume greater or equal to 65% corresponds to treatment response, and an increase in overall volume greater or equal than 40% corresponds to progressive disease. The interval between these two states is classified as stable disease. From the selected cases, 55 (59.1%) were found to be cases of progressive disease, 35 (37.6%) cases of stable disease, and 3 (3.3%) cases of treatment response. Table 5.2 presents some examples of how the free text of medical notes was converted into assessments using the RANO criteria.

After the imaging studies were successfully processed, and clinical data were obtained for each of the subjects, the proposed retrospective studies were performed using Glioview. All 93 cases were manually assessed to be cases in which the tumor was still present in the different follow-ups, no resection occurred during the process (biopsy was accepted). Cases where resection happened immediately (a frequent scenario) were not included in the study since the objective was to evaluate the impact of variability metrics on tumor volume change over time.

As mentioned, the initial (retrospective) clinical assessment was extracted from clinical notes and manually correlated to either progressive disease, stable disease, or treatment response according to the RANO criteria, this assessment was supervised by an expert neuroradiologist (Dr. El-Saden). Automated diagnosis with estimates of variability was then performed, and the volume change over time was assessed according to the same criteria. The expert neuroradiologist then performed an assessment of the cases that were found to have a disagreement between the automated volumetric evaluation and the previous analysis at the time of the study. In order to not bias the judgment of the new assessment, at the time of analysis the neuroradiologist did not have access to either the previous clinical notes or the diagnosis from the automated framework. The new clinical assessment was performed by looking at the variability estimated for each of the time points such as boxplots, timeline and variability maps overlapped in the imaging studies, which were not available in the previous evaluation, as well as observing

the different time points and different sequences side by side to evaluate change over time. This was not always available during the first evaluation done at the time of the scan (Figure 5.3).

Case example	Clinical text
Example 1	A ring enhancing mass centered in the left superior temporal gyrus has mildly increased in size. It measures *5.0 x 3.9 x 4.3 cm*, (AP x TV x CC), previously *3.3 x 2.7 x 2.9 cm* at the same level. Areas of associated intra-tumoral hemorrhage are unchanged. Mild interval growth of enhancing tumor in the left temporal lobe.
Example 2	There has been *no interval change* in the size or enhancement surrounding the left parietal resection cavity and no new lesions are identified. There has, however, been mild interval increase in the amount of perilesional T2 hyperintensity, likely in part related to vasogenic edema.
Example 3	The current scan demonstrates two dominant areas of heterogeneous enhancement with central non-enhancing regions. The more posteriorly-directed rim-enhancing mass currently has a *maximal diameter of 21 mm*. *Previously, this lesion measured 19 mm and showed less central necrosis*. The more anteriorly and superiorly-enhancing process is more amorphous. Previously, *this measured 9 mm in maximal diameter*. *Currently, it measures 18 mm in maximal diameter and has developed additional central necrosis*. A more amorphous area of enhancement is noted inferior and medial to this. When compared to the prior scan, therefore, there is *significantly more new enhancing tissue*.

Table 5.2 Examples of clinical notes analyzed for this experiment that was correlated with a RANO diagnosis. This process was validated by the expert neuroradiologist to be accurate. Note that even one-dimensional measurements were not always measured by the clinician, and sometimes only a qualitative assessment was conducted. Key sentences used to identify tumor status appear in italics.

5.4 Evaluation and results

5.4.2 Results of a retrospective study

Of the 31 cases examined in the retrospective analysis, 26 (83%) of them were discovered to have the same agreement between the new analysis and the previous conclusion. Of the 5 cases (17%) that were discovered to be in disagreement (initial diagnosis and automated result), each was consequently analyzed by the expert neuroradiologist to determine which analysis was found to be more accurate, and the reasons for the differences. The cases in which the neuroradiologist was in agreement with the initial diagnosis found on the clinical record (and not the automated analysis) include:

- Case in which a small resection cavity with hyper-intense blood products was classified by the algorithm as part of the tumor and not considered in the clinical analysis. It is known that in cases with resection cavities, these empty spaces sometimes are filled with blood products, that can appear hyper-intense, even without the use of a contrast agent. This is problematic for segmentation algorithms as it may be identified as part of the enhancing tumor. Tumor T1 subtraction maps can help to discriminate between hyperintense blood and enhancing tissue, but in this case, the algorithm still included it in the boundary since the blood pool was very close to the other tumor components.

- A subject that had a disagreement because other clinical comments not related to change in tumor volume were considered to reach a final diagnosis (*e.g.,* significant mass-effect or morphological changes in cerebral structures between follow-ups). Although volume change and three-dimensional characterization is the main biomarker used to assess tumor progression, in complex cases the neuroradiologist may aid his/her decision using additional factors to reach a conclusion. These additional features should be considered to be added to future modifications of the algorithm.
- Case in which the clinical notes were non-specific, and were incorrectly classified when the initial review occurred. It was observed that as the evaluations were not quantitative, the assessment can be non-specific and difficult to interpret at a later time.

The cases in which the neuroradiologist was in disagreement with the automated assessment with variability estimate include (Figure 5.7):

- A case of stable disease classified as progression. After the new analysis was concluded it was found that the three-dimensional measurement with the estimate of variability showed an increase of volume that was not initially detected with one-dimensional metrics. There was a reported decrease in enhancing margins, and disagreement with a previous note on perilesional T2 signal in the case. Both

neuroradiologists were in agreement with this assessment. Figures 5.6 a and b present an example of this case.

- A case of progressive disease classified as stable disease. In this case, the new analysis found changes in subsequent edema that were not noted in past clinical notes. Additionally, there was an indication of nodular enhancement. These two differences are an indication of disease progression. Both neuroradiologists were in agreement and concluded that the previous assessment did not consider the critical aspects noted above, possibly due to lack of a three-dimensional segmentation and appropriate visualization. Figures 5.6 c and d present an example of this case.

5.5 Conclusions

The variability associated with imaging features (*e.g.,* tumor volume) is a fundamental characteristic that must be known to make an informed tumor assessment [53,101]. The experiments performed in this chapter aimed to assess whether our proposed methods can increase the confidence of the neuroradiologist through a more quantitative and better-informed image characterization of brain tumors. The ability to evaluate the implications of radiological interpretations and determine if volumetric assessment and error measurement provide an additional benefit with respect to the current guidelines for tumor characterization.

The retrospective study that was performed, and the novel user interface that allows the neuroradiologist to analyze volumetric changes and observe the variability estimates for each time point were determined to be effective, and have a potential impact in the way neuroradiologists analyze the images. This approach promotes the movement from one-dimensional measurements to volumetric measurements on which statistical analysis can be performed.

Equivalence tests were used to determine whether the means for product measurements or process measurements are close enough to be considered equivalent. Given that small differences between products were not always functionally or practically important (*e.g.,* thresholds determined by clinical guidelines can determine if the tumor was different from the previous observations). These differences among the group mean can be tested by doing analyzes of the variance of the means of the different groups (analysis of variance for many samples or t-test for a paired groups).

Several challenges were noted. During the analysis of clinical notes, it was observed that clinicians often did not make explicit references to the RANO criteria while reading the imaging studies. Also, clinical notes tended to be conservative when assessing change (*e.g.,* when assigning progressive disease) and considered different aspects of brain anatomy (*e.g.,* cerebral mass-effect or brain deformation) and not just tumor volume as criteria to define the status of tumor progression over time (Table 5.2). Finally, another challenge to this study is that we did not have access to information

about the treatment that was being administered to the patient, possibly limiting the assessment that the current doctor made with respect to the previous diagnosis.

In summary, we demonstrated that in cases where differences in tumor volume were not visibly different from each other, and in which the tumor boundary can be fuzzy and unclear, our approach provided a useful assessment of boundary uncertainty. Such a metric can help the neuroradiologist determine which case requires a more careful analysis as it can be a signal of biological heterogeneity or other important clinical factors.

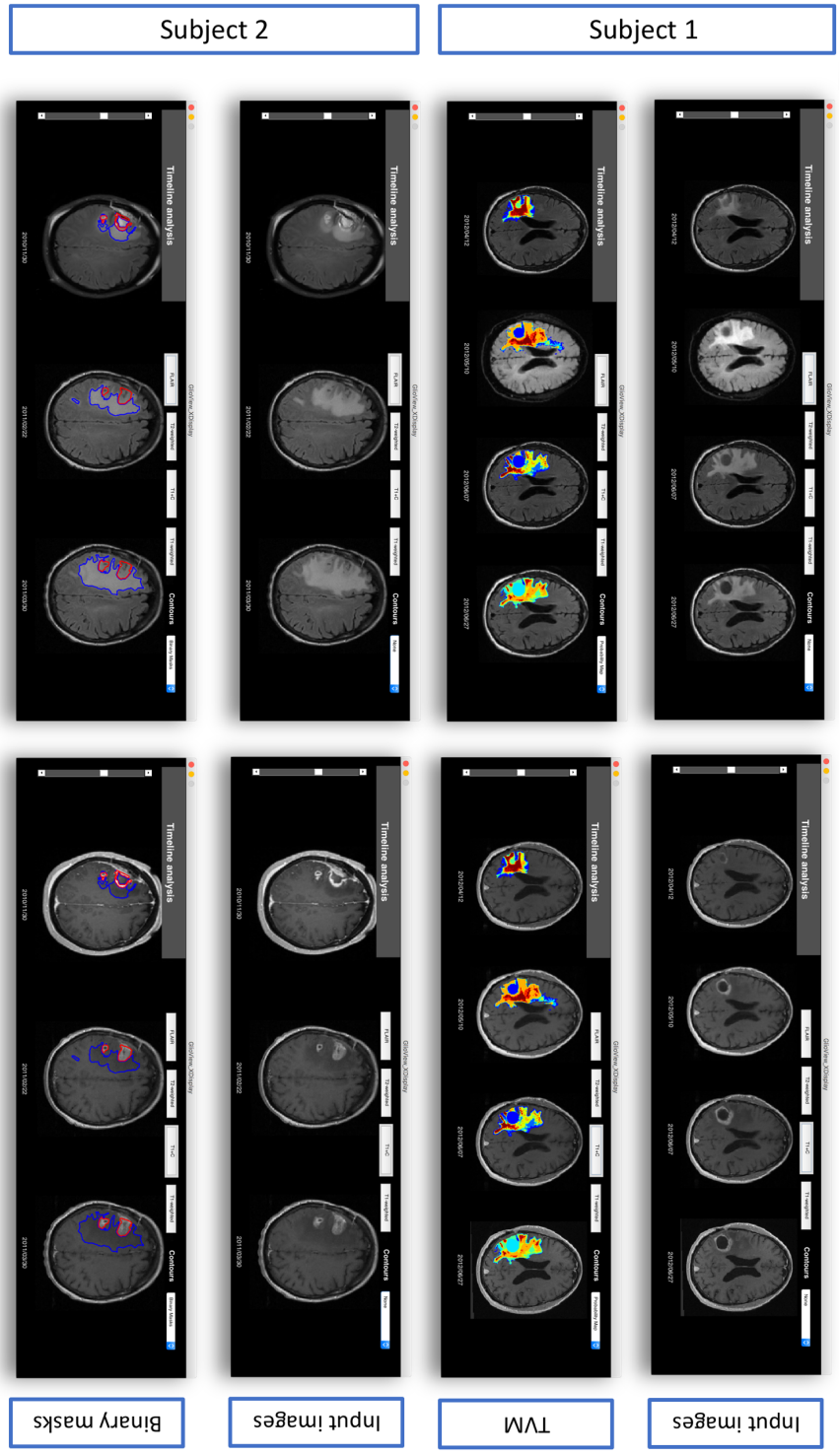


Figure 5.7 The two different cases that were found to be in disagreement with the original clinical assessment after a volumetric analysis that included an estimate of variability. First two rows show Subject 1 with and without variability maps, bottom rows show Subject 2 with and without binary contours overlapped.

Chapter 6: Conclusion

6.1 Summary of findings

This dissertation centered on the understanding, characterization, optimization, and representation of tumor variability in medical imaging, specifically in the domain of GBM using magnetic resonance imaging (MRI). The contributions of this work are outlined in the following paragraphs, with comments on the specific theoretical or practical contribution that is provided in this dissertation:

- Understanding the sources of variability that arise in GBM tumor segmentation: The work presented in Chapter 3 contains most of the work relevant to this point. Understanding the sources of variability was possible by analyzing different segmentation algorithms and their intrinsic parameters to determine which and to what degree these parameters can determine the final tumor boundary that is produced by each of the algorithms. A general optimization framework was created to provide an automated and reliable tool to test for the different parameter combinations that can exist for each algorithm. These different methodologies were applied to define tumor variability maps that represent boundary variability at the algorithm-level.
- Development of a novel and automated tumor segmentation algorithm that characterizes tumor variability in its different subcomponents. Towards fulfillment of this

objective mainly presented in Chapter 4), a complete and automated image processing pipeline that combines statistical methods and knowledge-based methodologies were created. This framework characterizes the variability of the tumor boundary at the pixel-level by evaluating the differences in tumor and normal tissue intensities near the tumor boundary. Additionally, this new algorithm is capable of calculating a variability map for the different tumor sub-regions (edema, enhancing, necrosis) as well as total tumor and core tumor, all made possible by the use of a multimodal approach that uses different MRI contrasts. The algorithm that was developed towards the fulfillment of this contribution was tested using different independent datasets, including a large dataset provided by UCLA, and a public repository of GBM images from TCGA. This was evaluated by a team of independent researchers as part of the MICCAI tumor segmentation competition during 2016. This method was also observed to work using a set of different MRI phantoms, ranging from highly standardized phantoms to other personalized phantoms.

- Evaluation of the impact of segmentation variability metrics in clinical image assessment. For this last contribution, primarily covered in Chapter 5, an evaluation of the potential practical use of the tumor variability maps in a clinical setting and the work of the neuroradiologist was completed. To accomplish this, a retrospective clinical study was performed, using a set of previously analyzed tumor cases, and contrasting them with a new assessment by an experienced clinician when he/she has access to

variability metrics. As part of this study, a new user interface was developed allowing the clinician to visualize the images and the segmentation results, along with the estimation of variability and change over time represent in different ways. Results of this study included some cases in which a combination of variability metrics, three-dimensional measurements and an overall better visualization (especially a progression timeline) caused the new assessment to be different from the previous assessment.

6.2 Limitations

The work presented in this dissertation attempted to be comprehensive and complete; nevertheless, there are always possible areas of improvement that for diverse reasons were not considered at the time when this document was written.

- Sources of variability. In this work, we attempted to characterize variability in different aspects of the imaging process, such as the differences found when using different algorithms and parameter combinations as well as the differences that could be found while analyzing the tumor boundary in each imaging volume at the pixel-level. Although these two sources of variability were demonstrated to be of significant importance for variability characterization, there are other sources of variability that likely have important effects on the imaging process. One example of this is the acquisition and initial processing (*i.e.,* image reconstruction) of the image. Different

artifacts and numerical complications can arise and lead to representations of the object of interest (*e.g.,* tumor) that differ from the real object. Other sources of variability include acquisition sequence standardization, noise and motion correction, and a better determination of temporal change (tumor evolution over time).

- Dataset size and annotations. UCLA currently has one of the biggest datasets of brain tumors in the country. Nevertheless, different issues such as data inconsistency, missing information, and availability of annotated data limit the use of other algorithms that require very large amounts of data to be appropriately trained (*e.g.,* deep learning). Additionally, although there have been efforts to standardize image acquisition, there are still some variations when the data is acquired (changes in parameters to reduce scan time, incomplete sets of imaging sequences). Considering that many of the utilized cases for this work were originally obtained up to a decade ago, some even before the FLAIR sequence became a routine sequence in brain tumors, which occurred approximately in 2006.
- The accuracy of segmentations. As noted in Chapter 2, brain tumor segmentation is a very challenging task, and different research groups in diverse locations have investigated the best possible way to find the most accurate tumor boundary. Despite these efforts, these approaches are still not ready to be effectively used in the clinical setting. In difficult cases with heterogeneous biological characteristics, segmentation methods may be able to include only a part of the total abnormality or even miss

it completely. As noted before, the approach that was described was tested in different datasets (UCLA, TCGA), internally and externally (MICCAI). The results obtained from these different evaluations were comparable to currently existing approaches, and were quite useful during the task of training and testing the system.

- Working on pre-resection cases. Linked to the previous point, this work centered on the characterization of cases before resection occurred. Although there are many cases in which tumor resection does not occur (not advisable based on location, affected areas or size of the tumor), many other cases do undergo resection, and an appropriate characterization of these cases surely can lead to meaningful clinical conclusions. Unfortunately, there are several hurdles to surpass including the handling of resection cavities by different algorithms (skull stripping, registration) that normally cannot handle cases with such extreme brain deformations and related biological problems. These include the timely identification of tumor pseudo-progression or other effects that are visible in the images that are related to treatment (*e.g.*, radiotherapy, immunotherapy), but not directly related to the tumor itself.

6.3 Future directions

Although the work presented in this dissertation has advanced our knowledge in tumor characterization using MRI, there is work that remains to be done. There are many other brain and non-brain diseases in which this work can be applied, some of them

including stroke, lung nodules, prostate cancer, and breast cancer. Characterization of the variability in measurements and the use of different metrics to find the true characteristics of an object is a general idea that is applied in science. It often does not reach medical applications due to practical and methodological constraints related to the reduced speed and high accuracy required for clinicians and other such as biological heterogeneity and characterization variability.

Future work can center on translating these methods into other domains and expand applicability into the clinical setting. Some ongoing efforts in this direction include the brain tumor division of the MICCAI society that runs the brain tumor segmentation competition. One outcome of this competition includes finding a combination of the best algorithms that can generate the most accurate segmentation and the identification of factors that account for the variability between them. Such aims are well aligned with the contributions of this work.

Another area of opportunity includes going a step closer to the clinical setting to investigate the utility and impact of translating the variability estimates generated for tumor segmentation and apply them to the generation imaging biomarkers and *radiomics* features. This can inform neuroradiologists how precise a measurement is, and possibly how much importance to give to a certain biomarker depending on the ranges of variability that are being provided.

Additionally, in order for the algorithm to work in post-resection cases, it would be ideal to have labeled cases on which resection has occurred (in multiple locations and when the resection cavity is hypo-intense due to the presence of CSF and/or hyper-intense due to the presence of blood), utilize a resection-robust method that is able to skull-strip the brain with a resection cavity and, possibly complimented using another data source (e.g., clinical information) that s information about the surgery (resection, biopsy).

Instead of focusing on clinical applications, the frameworks and ideas presented in this work can be applied to the image acquisition process. Efforts to identify what are the causes of variability in the imaging process (motion, acquisition parameters, reconstruction process), and optimizing the process so output images have better quality and can be adapted to different needs (children vs. adults, calcifications vs. soft tissue lesions are promising areas of research.

6.3.1 Applicability of tumor variability maps

Using the proposed multimodal framework for automated, probabilistic brain tumor segmentation by using variability in estimates of the tumor boundary, and by exploiting tumor variability from different imaging sources, this algorithm is able to automatically generate tumor probability maps. Alternatively, it adds a measure of uncertainty to binary tumor segmentations. By explicitly quantifying the error associated with

any given segmentation, we believe that this added information is critical to understand and judge the actual tumor extent by a radiologist or neuro-oncologist when interpreting follow-up imaging data in a clinical setting.

When comparing the results with other approaches proposed in previous years, our results are comparable to or surpass the mean performance of other algorithms [18] (Reza, Meier, Cordier, Bauer, Festa, Geremia, Buendia, Taylor, Shin) but progress still has to be made. One approach to improve our results exploits a classifier-based on convolutional neural networks (CNN) [102-103] to help in the definition of the preliminary tumor ROI. It can reduce the number of false positives during the tumor boundary selection. Combining the result of our knowledge-based approach and the result of the CNN, trained to classify whether an individual voxel is part of brain tumors using an independent dataset, might contribute towards better results in the different tumor contours. Examples where the proposed approach did not segment the tumor accurately are shown in Figure 6.1).

The inclusion of variability calculations into segmentation methodologies can lead to improved results, and ultimately provide more meaningful data to clinicians since the knowledge of measurement variation is fundamental to make more objective decisions. Finally, additional work can include the addition of extra tumor biomarkers (tumor volume, thickness of enhancing margin, necrosis proportion), and evaluate their variability in brain tumor cases. Investigations should renew determination of statistical analyses,

together with visualizations similar to the what was presented in Chapter 5 with Glioview (Figure 6.2), can further aid in the medical decision-making, increasing the radiologist's efficiency, accuracy, and eventual outcome at evaluating imaging studies.

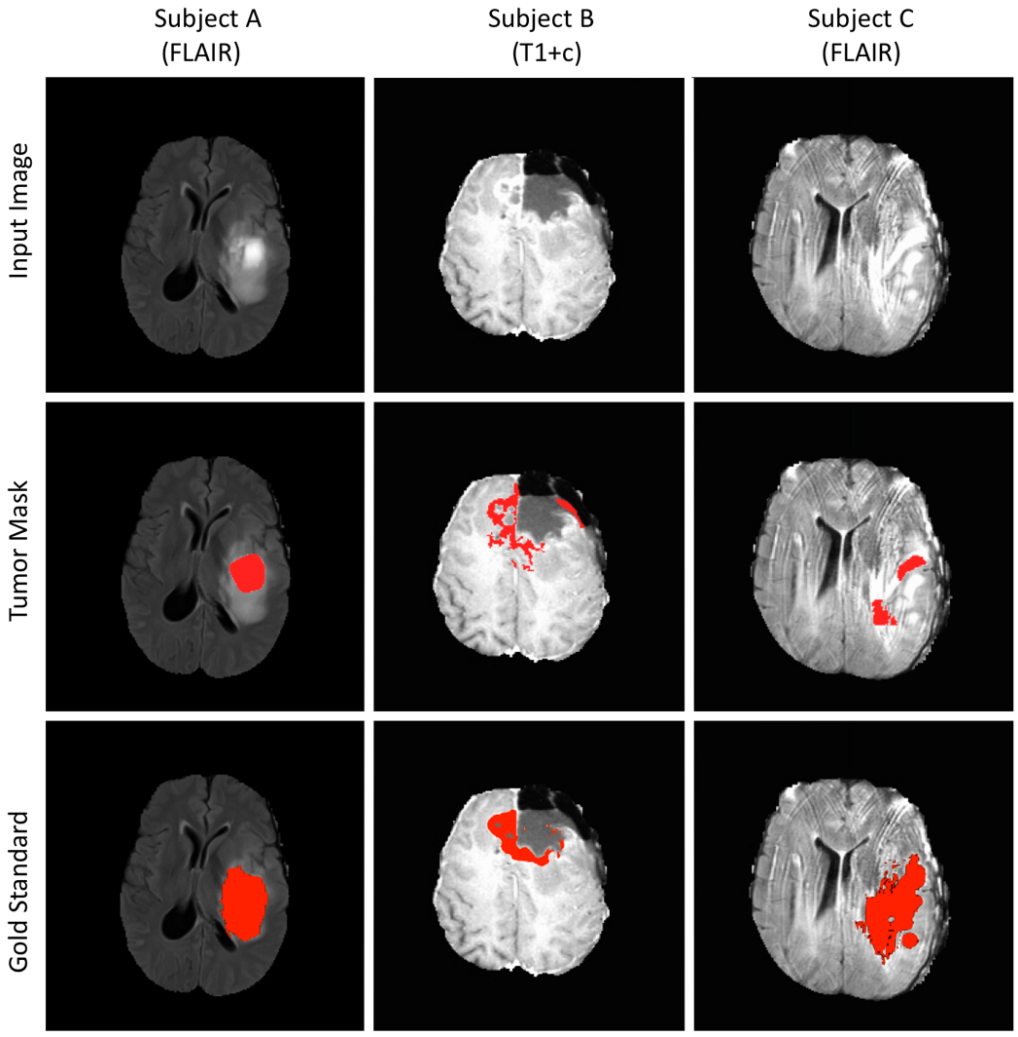


Figure 6.1. Examples where our approach did not segment the tumor appropriately. The first column shows an example where the tumor is under-segmented. We believe this is because of a miscalculation in the prior information that was obtained as a preprocessing step. Only the hyper-intense edema was selected as part of the tumor, leaving the darker parts unselected. The second column shows a resection cavity on which the full area of enhancement was not captured, and the third column shows how image artifacts can also cause problems while trying to segment the tumor.

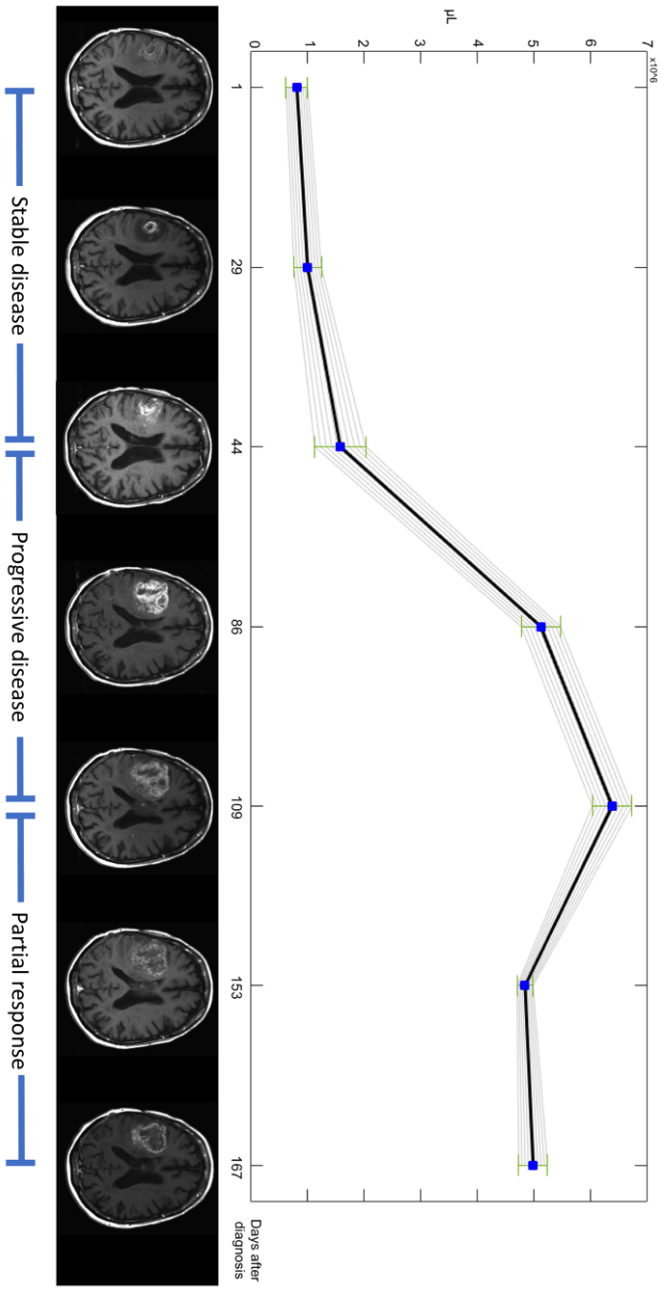


Figure 6.2. Tumor progression over time. Showing volume measurement for each time-point with its respective error estimate. As time passes it becomes critical to evaluate if the action was taken at a given point in time (*e.g.,* chemotherapy, radiotherapy) had a significant effect on the tumor characteristics observed in follow-up MRI scans. This approach enables clinicians to have different estimates of tumor characteristics (*e.g.,* tumor volume, growth rate), and statistically define if significant change over time is observed, a vital task for subsequent treatment planning.

Altogether, this dissertation explored topics in medical informatics, MRI automated assessment, and the development of algorithms applied to neuroimaging. Contributing to the improvement of the radiological assessment and patient outcome.

Appendix

Supplementary material for Chapter 2

Literature review for previously published methods for brain tumor segmentation

Integrated hierarchical random forest classification and CRF regularization [19]

Bauer, Wiest, et al.

This automatic method for brain tumor segmentation is based on a statistical classification of healthy tissues and pathologic tissues. The segmentation task is modeled as an energy minimization problem in a conditional random field (CRF) formulation, and a decision forest algorithm is used to perform the final classification.

This approach uses a more than forty features as input to the classifier, that includes intensity features, first-order textural features and other features from image gradient and pixel neighborhood (a distance metric from image patches). The random forest classifier was trained using a five-fold cross-validation in the training dataset. The parameters of the algorithm were chosen empirically.

The approach was compared against a classification using a Support Vector Machine (SVM) with similar image features, obtaining comparable results but the different regularization approach that is utilized da faster and more efficient classification. The method has difficulty in classifying datasets that are very different to what was observed during the training phase of the method.

Bit-grouping artificial immune network for tumor segmentation [20]

Buendia, Taylor, *et al.*

This approach uses an Artificial Immune Network to perform an automated analysis of GBM MRI images. This system recognizes the abnormal regions as "pathogenic image patterns", and then process them to output the tumor region for the different abnormal regions observed using the different MRI imaging contrasts obtained.

The input parameters for the detection of the abnormality include a combination of intensities of 2D and 3D image blocks of varying sizes for each sequence, textural patterns (such as mean, variance, and co-occurrence matrices). The grouping approach that was used (multi-bit grouping) was chosen to be groups of 2 bits for improved computational efficiency and was made based on the location of each bit within the input image and the relevance observed when extracting the initial features.

Before running the segmentation approach, a preprocessing pipeline was put in place to remove noise and some inhomogeneities from the input images, as well as matching the intensity histogram of the input volumes. Also, several post-processing options were implemented, such as finding and extracting connected components and performing dilation and erosion on those components to improve the accuracy of the final result. The system was trained using a 20-fold cross-validation on a set of high-grade tumors from the BRATS 2013 training set. The evaluation was done for three different tumor sub-regions - whole tumor, tumor core and enhancing region - using the

Dice coefficient comparison metric.

Patch-based tissue segmentation approach [21]

Cordier, Menze, *et al.*

This approach provides an automated method that was inspired using manual brain region labeling, and is similar to a multi-atlas label propagation method. This method utilizes a database of local normal image patches and another set of pathologic image patches. For each of the input test cases, the image patches are extracted from the image and compared to the ones contained in the database, enabling a labeling of test patches afterward.

The comparison of the different image patches is done by performing a sum of squared differences comparison after a global intensity alignment. When the patches are being compared, this evaluation is also performed in the five image patched that are most similar to the patch that was retrieved using a nearest-neighbor approach. The utilized image features include intensity patches extracted from the different MRI sequences, with the addition of exponential weight to improve system performance. The most important parameter in this algorithm consists of the patch size (defined manually) and the number of training cases. The number of similar patches that were retrieved using the nearest neighbor approach had almost no effect in the output tumor region.

This process is performed between every voxel in the test case and every training

image patch. For each defined label, the results are obtained by doing a weighted voting approach on the resulting probability-like map (50% percentile). At each step of the algorithm, weighted votes are rebalanced based on label frequencies to penalize image patches that would be picked so often that it would introduce a systematic bias. The output tumor ROI is obtained by applying post-processing techniques to reduce noise in the resulting images and image smoothing.

The algorithm has some problems when trying to classify thin regions of enhancing tumor. The necrotic component is sometimes missed or misclassified. The algorithm could use shape criteria to discriminate and ignore false positive regions located mainly in regions close to the occipital lobe or cerebellum.

Hidden Markov fields and variational EM in a generative model [22]

Doyle, Vasseur, *et al.*

This approach for tumor segmentation works using multiple MRI sequences to achieve automated tumor detection. The parameters of this model that are used to distinguish the tumor components are estimated using the expectation maximization algorithm with Markov random field constraints, and therefore does not have a training phase.

Hidden Markov Random Field (MRF) models are able to consider general patterns that overall encode complex interactions. For this approach, in particular, certain tissue

combinations in the image neighborhood are penalized more than others to provide a representation of the tissue these areas represent. This is achieved by using the mean field principle to generate a variational approximation for tractability in the presence of Markov dependencies.

The data model used for this approach included five tissue classes: white matter, gray matter, ventricular CSF, extraventricular CSF, and "other". The tumor tissues were modeled by four different tissue classes: edema, non-enhancing, enhancing and necrotic. In case there is missing data and the complete interactions cannot be accurately determined, then further constraints on the model can be applied, such as consideration of the tumor as a single entity if there is not enough information about tissue types.

In order for the approach to work with multiple sequences, the multiple volumes are registered to an existing MRI atlas, then simple morphological comparison methods are performed on the tumor ROI previously obtained by the MRF to achieve a final classification for the four different tissue classes. This algorithm was trained and tested using the BRATS dataset from 2012 and 2013. All parameters of this model were estimated iteratively during the segmentation process.

Random forest classifier using neighborhood and local context features [23]

Festa, Pereira, *et al.*

Similar to other approaches. this algorithm uses different MRI sequences to train a random forest classification model to achieve automated tumor segmentation. The pre-processing steps to reduce sources of error were included into the process (bias field correction, intensity normalization, histogram marching, and in some cases cropping of the imaging volumes where no tumor is observed).

As part of the training process, each brain voxel is classified based on several features extracted from the different images. The number of trees in the forest was defined to be 50 with a possibility of maximum 25 levels on the branches of each tree. The extracted features included intensity measurements from the input images as well as from subtractions of them. These include post contrast T1 minus regular T1 image volumes; 2D neighborhood information such as mean, median, intensity range and other texture information, and three-dimensional information with each pixel and their neighbors located 3mm in each direction.

These features are used to build the decision forest. For every test case, each pixel runs down this created structure and the final pixel classification is obtained based on majority voting, where every tree provides a classification based on the branch arrangement. Finally, a post-classification step was performed to eliminate small regions (false positives). The approach was built for low-grade and high-grade gliomas.

Spatial decision forests with intrinsic hierarchy [24]

Geremia, Menze, *et al.*

This method uses a method called "Spatially Adaptive Random Forests" to automatically learn and classify MRI images with brain tumors. Different from traditional decision tree approaches, this method introduces a multiscale 3D image representation plus a structured labeling system. They significantly increase the time it takes to train the system, but also increase the classification accuracy.

To increase the classification speed, the hierarchy is built using coarser images obtained by the use of the SLIC algorithm. These simplified images are equivalent to a supervoxel representation. This results in the ability to work with semantic regions within the tree hierarchy, boosting the parsing and improving the classification accuracy.

Semi-automatic segmentation using active contours [25]

Guo, Schwartz *et al*.

This semi-automated approach takes multimodal MRI brain tumor images and a manually defined input ROI that roughly encloses the tumor region, only on one slice. This algorithm combines the region- and edge-based approaches to find the tumor boundary inside that ROI, with the advantage that it runs very fast (under 1minute).

The general idea of this approach is to take the user-defined ROI (defined on the post-contrast T1 weighted image) as a reference and propagate tumor contour through

the imaging volume using interpolation. Following this, the active contour approaches segment the total tumor mass. Enhancing and necrotic tumor components are separated by adaptive thresholding. A similar approach is followed on the FLAIR images to find the edema region, defining the total abnormality, and then subtract the tumor mass found on the T1-weighted sequences.

The implemented active contour models allow for topological changes (merging and splitting), which is a necessary property when trying to segment regions with highly variable shapes, such as brain tumors. This method can also function when the complete set of input sequences is not available, being able to segment only the enhancing tumor or total mass when there is not enough data available.

The system is built with a user interface to draw the initial ROI and show the segmentation results for easy visual inspection, with the possibility of manual refinement in case it is needed.

The Tumorcut method [26]

Hamamci and Unal

This is a semi-automated tumor algorithm aims to segment the gross tumor mass along with the sub-components of the tumor, being able to work with high-grade and low-grade gliomas. In addition to the MR input volumes, the method requires a manually drawn line through the maximum diameter observed on the tumor, as described on the

RECIST criteria, for each of the input volumes.

This method uses the input line as a seed for tumor segmentations and to identify the final tumor boundary called cellular automata. This works by optimizes the boundary using grow-cut segmentation with an added sensitivity parameter that can be adjusted to adapt for different levels of tumor variability and regions with smooth boundaries.

The biological approach on which the cellular automata is based on a dynamic system defined inside a grid, and iteratively expands by using the grow-cut algorithm. This algorithm aims to find the shortest paths by properly choosing among different transition rules to achieve tumor segmentation. These transitions rules can be adapted depending on the sequence that is being segmented or to account for other special scenarios such as the presence of resection cavities.

Tumor regions are selected by doing logical operations on the results obtained from the previous step. A smoothness constraint using level set active surfaces is imposed to improve result accuracy.

Appearance- and context-sensitive features with a random forest and CRF [27]

Meier, Bauer, *et al.*

This approach for automatic tumor segmentation is divided into three different parts: feature extraction to create a feature space for the training of the random forest

algorithm, a classification step and finally spatial regularization. The process also includes a previous step in which the MRI input images are noise-reduced, normalized and bias-field corrected. The parameters of the approach (such as deepness level and a number of branches) are defined according to the grid search algorithm.

The explored feature space is composed of 257 features extracted from the images, including intensities from the multimodality inputs, as well as first-order texture features and information from the image gradient. Unlike other approaches, this method also includes context-sensitive features such as symmetry metrics and structural coordinates derived from a normalized atlas.

The classification forest uses the maximum a posteriori rule to calculate the posterior probability output during the classification task. The regularization step includes an energy minimization of a conditional random field. The output of this step is a measure of the strength to which each pixel might be linked to each of the defined tumor classes, then the strongest link is selected as a final label for the pixel. The algorithm showed good results overall, presenting the most number of false positives in the infratentorial region (cerebellum), and when making a classification between edema and non-enhancing tumor.

Generative-discriminative lesion segmentation model [28]

Menze, Geremia *et al.*

The proposed method is based on a generative approach similar to a classification forest that is able to encode standard image intensity and textural features as well as information about expected physiology and pathophysiological changes. Then a classification forest classifies the different input imaging into the different tumor classes by considering the association strength based on the measurement of the geodesic distance of each pixel to a defined tumor class. The output of the method is a probability map for the tumor region observed in each imaging volume.

While the output of this segmentation method provides a better insight into the different pathophysiological processes that occur in the tumor. The highly specific tumor components such as edema, necrotic core, enhancing and non-enhancing regions, cannot be directly associated with the results obtained from this method. To compensate for this, several structural features are extracted from the resulting tumor map to further constrain the classification and identify tumor subcomponents.

The model consisted of no more than 2000 samples from each input volume, generating a group of 300 trees with a level depth of 10. To improve results, some standard post-processing techniques were applied as well as Markov random field smoothness constraints on the class labels.

Generative lesion segmentation model [29]

Menze, Van Leemput, *et al.*

This fully automated approach for tumor segmentation implements a protocol that combines a standard Expectation Maximization (EM) generative model and population priors (atlas-based) to determine tumor labels using multi-sequence data. This method exploits the image spatial location information to segment the normal regions in the brain. It includes a latent variable on each voxel to assess the probability of that voxel being a part of the tumor region.

The overall process includes the use of a physiological tissue atlas with a personalized lesion prior, clustering of the input images using the EM algorithm into three different regions for healthy tissue (white matter, gray matter, and CSF), and an outlier class of tumor tissue. These outliers are defined as the regions more than three standard deviations away from the average intensity value observed in normal tissue classes.

For a semantic interpretation that is in line with the class definitions of the segmentation challenge, channel-specific segmentations are returned by the algorithm and transformed to the edema and tumor core classes, which includes enhancing and necrotic regions. Finally, a region post-processing occurs to remove the likely false-positives that may appear. Small regions defined as abnormal.

Texture features and random forests [30]

Reza, Iftekharuddin *et al.*

This fully automated algorithm achieves multi-class tumor segmentation with MRI images. This algorithm starts by preprocessing the data (image normalization, skulls-stripping and image co-registration). Two types of image features can be extracted. The first group contains intensity features from input images as well as difference images, and the second group included novel textural features such as fractals (using the PTPSA approach), "textons" (vectors that describe the human perception of texture), and first and second-order textural features.

These features were used as input for a classical random forest classification algorithm. The algorithm involves a 3-fold cross validation during the training phase. The samples obtained from the images were evenly divided between tumor tissue and normal tissue. This combination performed better compared to other unbalanced proportions. The classification is done for each of the pixels on the input imaging volumes similar to the standard random forest algorithm description. The final label was assigned based on the consensus or agreement between the individual trees of the forest.

The whole process lasts approximately one hour and a half per subject, and the cross-validation takes about 15 minutes (Xeon processor with 15 Gb of RAM). Once the labels are generated for the test images, a 3D volume is outputted for further clinical evaluation. This classification algorithm was trained with low grade and high-grade

gliomas and was tested using the testing set provided by the 2013 BRATS challenge.

Generative model with latent atlases and level sets [31]

Riklin-Raviv, Van Leemput *et al.*

This joint segmentation and patient-specific approach use MRI images to find the tumor segmentation by using a statistically-driven level-set algorithm. This approach does not require a training step. It is a semi-automated approach that requires the user to provide one or more seed points near the centroid of the tumor so different level set functions can be initialized around that area, subsequently evolving to cover the entire tumor mass.

The segmentation is approached as a partitioning problem, with the tumor component seen as foreground and normal tissue as background. A joint evolution of four level-set functions evolves based on the gray level distribution observed in the input image. The individual segmentation of each scan can consider segmentation of the other available sequences by sharing common information to improve the overall result. This common information, in the form of a spatial tumor location distribution map, is created as the level sets evolve using an alternating minimization procedure.

This method was tested on a series of 25 multimodal images from the BRATS dataset. They also tested the performance of the results for longitudinal studies for tumor evolution detection. The methodology was adapted to segment the tumor regions

and different subcomponents but can be tailored for detection of other abnormalities.

Hybrid clustering and classification by logistic regression [32]

Shin *et al.*

This method is an unsupervised learning approach using a logistic regression approach to segment the tumor core as well as edema regions. The preprocessing steps of the MRI imaging volumes included image normalization, bias correction, and image noise reduction.

The algorithm works by first extracting a series of four-dimensional imaging patches, regions of 3x3x3 pixels on each of the available MRI sequences available. Then second-order polynomial features can be extracted from those patches and used as input for the classification step. A dictionary for the behavior of the image intensities in the multiple input sequences (Flair, T1, T1C, and T2) was performed by the implementation of a sparse autoencoder. Image intensities in the same pixel position on the different image modalities are convolved with each of the dictionary entries to obtain a binary label that represents the pixel final classification as part of normal tissue or tumor tissue.

This approach included two cross-validations, and no additional post-processing steps were applied to the result produced by the classification technique. The performance was assessed using real tumor data as well as synthetic data generated from normal brain scans to simulate brain tumors. The main limitation of this approach is

that it is dependent of the final pixel classification on the examples contained using the developed dictionary. The addition of post-processing steps might provide an improvement in overall performance.

Hierarchical MRF approach with Gabor features [33]

Shubbanna, Precup *et al.*

This method is a fully automated hierarchical probabilistic model adapted for brain tumor segmentation on multi-sequence MRI images. It is based on the application of multi-window Gabor filters and Markov Random Fields to identify the edges of the tumor. The preprocessing steps used during this approach included bias field correction, intra-subject registration of all the input sequences, and image volume normalization.

This algorithm consists of two different stages. The first aims to generate an approximate tumor segmentation (whole mass and its different regions) as well as approximate the distribution of normal tissue using Gabor functions. The second is a Bayesian framework implemented in such way that the first tumor masks are used to calculate probabilities of the patterns observed on the Gabor filter to be part of normal or tumor tissue. Once these probabilities are obtained, the coarse masks previously obtained are refined using a modified conditional Markov Random Field that separates the tumor subclasses from the different healthy tissues.

The implemented Markov Random Field differs from the standard implementations in that it takes intensity differences between neighboring voxels into the likelihood model and considers transition probabilities between neighboring voxel classes. As a secondary stage on the boundary refinement, local inhomogeneities are included to have a smoothing effect on the output results.

This method was trained and tested on the MICCAI 2012 and 2013 datasets (for low and high-grade gliomas) and was evaluated for all tumor subclasses (necrotic core, edema, solid tumor, and enhanced tumor), as well as overall tumor mass. The method obtained good results compared to existing approaches (30% improvement). Disadvantages include having a dependency on image normalization because the Gabor filters directly depend on that information to generate the result. The program had encountered difficulties during the identification of some groups of the tumor (high-grade gliomas in particular).

Map-Reduce Enabled hidden Markov models [34]

Taylor, John, *et al.*

This novel approach called Map-Reduce enabled the extension to hidden Markov Models. It enables training and segmentation of solid tumors and edema sub-regions in multi-modality MRI brain images. The preprocessing steps involved in this approach include removal of spatial inhomogeneities due to patient movement or machine artifacts

that may present difficulties for the subsequent segmentation process. It also included bias field correction and histogram matching.

The training step of this approach included extraction of image features (intensity metrics for each of the available sequences as well as information from the neighboring pixels). The Map-Reduce technique involves the use these vectors to create a mapper function for every single training case, and then adding these expressions case-by-case into a hidden Markov model that updates feature vectors given each of the mappers that are added into it. The final model normalizes the probabilities obtained on the hidden Markov model and then used image classification on a test case.

This program is fast (about a minute per subject) and can run parallel segmentations on all four MRI contrasts simultaneously. Shortcomings of the program include lack of atlas-based prior or advanced textural features that might improve accuracy. This method was tested using a subset of the 2013 BRATS dataset.

Limitations of the current algorithm include lack of support for spatial features, neighborhood-based textural features, and utilization of atlas-based priors, which have been shown to improve segmentation accuracy.

Random forest classifier using the open source ANTs/ANTsR packages [35]

Tustison, Wintermark*, et al.*

This automated approach for tumor segmentation uses a random forest approach with the addition of the implementation of a processing package called ANTs (Advanced Normalization Tools). This package is implemented in R and aims to extract information from complex datasets on imaging datasets. This involves features derived from statistical atlases on co-registered images and image vector feature extraction for use on spectral decomposition approaches such as ICA.

The segmentation algorithm consists of various preprocessing steps including: construction of a multivariate symmetric template using the ANTs package; image noise reduction techniques (normalization, bias field correction and image rescaling); and the generation of image features including first-order statistical features (such as mean, variance and standard deviation), and geometrical feature components (eccentricity, elongation, and symmetry features). These features are obtained through the use of the ANTs package.

The extracted features are then utilized towards the construction of the initial layer that composes the random forest. The segmentation step involves the use of a maximum a posteriori Markov random field to obtain spatial prior probabilities used to construct deeper forest levels and output labels. A final refinement step is implemented by using binary morphological alterations determined empirically. The processing time for each patient can take up to 2 hours.

Learned MRF on supervoxels clusters [36]

Zhao, Corso *et al.*

This method is based on an initial tumor over-segmentation (estimation of supervoxels), and then constrains the segmentation to refine only the image labels that are selected as probable tumor components by a Markov random field defined on the supervoxel image.

The labeling process occurs for every tumor sub-region (edema, necrosis, enhancement, and non-enhancement) as well as for normal regions (cerebral tissue and cerebrospinal fluid, CSF). The class likelihood on the Markov random field is obtained by evaluating each supervoxel with the most adjacent neighbors located on a Gaussian mixture model, which models one Gaussian distribution for each of the possible tumor labels. The tumor edge is identified by assessing the boundaries of each pair of supervoxels using a defined smoothness term, and a subsequent boundary optimization by using the graph-cuts approach.

The training phase of this approach included a two-fold validation on high- and low-grade gliomas, generating individual classifiers for each group. Similar to other supervised methods, this approach is highly dependent on the intensity standardization that is performed as a preprocessing step.

In some cases, the method fails because of the presence of image artifacts, incomplete sets of MRI sequences, or preprocessing errors that interfere with the classification step. The method was tested using the BRATS dataset for 2013.

Learned MRF on supervoxels clusters with updated unary potential [37]

Zhao, Zarikaya, *et al.*

This automated approach for tumor segmentation estimates the likelihood of pixels on a three-dimensional joint histogram obtained by normalizing and registering the different MRI input volumes, performing a voxel-by-voxel extraction. This method initially performs a tumor over-segmentation (estimation of supervoxels) and then constrains the segmentation to refine only the image labels that are selected as probable tumor components by a Markov random field defined on the supervoxel image.

The initial steps of the program, including the normalization, registration and image noise reduction are performed using an external program called SUSAN. The estimation of the image supervoxels is performed using a program called SLIC 3D which works by evaluating intensity similarity throughout the imaging volume. Image features are extracted, such as region size and intensity information, and normalized by the calculation of the standard z-score, normally used to group all the image vectors on the same scale.

Afterward, the class likelihood for every super-pixel is evaluated using a Markov random field model with the addition of Graph Cuts approach to define the optimal boundary on pairs of superpixels. The process was validated using a 2-fold cross-validation, and used a subset of the BRATS segmentation challenge to evaluate the methodology. The method has promising results, with the added benefit that over segmentations limited by the initial boundary defined by the supervoxel, which limits the inclusion of false-positive regions adjacent to the tumor region.

Context-sensitive features with a decision tree ensemble [38]

Zikic, Glocker, *et al.*

This automated method implements a random forest classification approach on multi-modality MRI images (T1, T2, post-contrast T1 and FLAIR images) to achieve tumor segmentation of high-grade gliomas, with the novelty of the introduction of context sensitivity that arises from two different places. First, the construction of the trees is based on other additional inputs such as patient-specific tissue probabilities, rather than only the standard MRI input volumes as in the conventional forest classification approach. As the second component employed to add context sensitivity is the removal of noise from the input images and other correction techniques before the classification is performed (registration, bias field correction, and normalization).

The mentioned tissue-probabilities are computed as the posterior probability of the image region that belongs to a specific Gaussian distribution (one for every tissue label). These approximations to tissue distributions are generated during the training process. For all cases the images were co-registered between each subject to the T1 volume, Noise correction strategies take place, and the feature extraction and classification occur following a standard random forest.

Other parameters of the decision forest include a total of 100 trees with a maximum branch level of 20 branches. The algorithm was tested using a two-fold cross-validation on the BRATS 2012 training data. Potential areas of improvement for these methods are to account for highly variable boundaries, and train the approach on a bigger dataset of pre-surgery brain tumor cases. as Also useful would be the addition of post-surgical cases to the training and classification steps.

References

1. Lima, Flavia RS, et al. "Glioblastoma: therapeutic challenges, what lies ahead." *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1826.2 (2012): 338-349.

2. Kimberley Mak, Gillian Lieberman. “Imaging in Glioblastoma Multiforme; diagnosis, treatment and follow-up” Harvard Medical School Radiology Clerkship BIDMC. (2008)

3. Pope, Whitney B., et al. "MR imaging correlates of survival in patients with high-grade gliomas." *American Journal of Neuroradiology* 26.10 (2005): 2466-2474.

4. da Cruz, LC Hygino, et al. "Pseudoprogression and pseudoresponse: imaging challenges in the assessment of posttreatment glioma." *American Journal of Neuroradiology* 32.11 (2011): 1978-1985.

5. Duncan, James S., and Nicholas Ayache. "Medical image analysis: Progress over two decades and the challenges ahead." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22.1 (2000): 85-106.

6. Wiki for the VASARI (Visually Accessible Rembrandt Images) feature set; The National Cancer Institute Web site. Updated 2014. *https://wiki.cancerimagingarchive.net/display/Public/VASARI+Research+Project*

7. Inda, Maria-del-Mar, Rudy Bonavia, and Joan Seoane. "Glioblastoma multiforme: A look inside its heterogeneous nature." *Cancers* 6.1 (2014): 226-239.

8. Eisenhauer, EA1, et al. "New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)." *European journal of cancer* 45.2 (2009): 228-247.

9. Wen, Patrick Y., et al. "Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group." *Journal of Clinical Oncology* 28.11 (2010): 1963-1972.

10. Sul, Joohee, and Daniel M. Krainak. "Brain tumor clinical trials imaging: a (well-standardized) picture is worth a thousand words." (2015): 1179-1180.

11. Gutman, David A., et al. "MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set." *Radiology* 267.2 (2013): 560-569.

12. Eisenhauer, EA1, et al. "New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)." *European journal of cancer* 45.2 (2009): 228-247.

13. Wen, Patrick Y., et al. "Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group." *Journal of Clinical Oncology* 28.11 (2010): 1963-1972.

14. Zacharaki, Evangelia I., et al. "Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme." *Magnetic Resonance in Medicine* 62.6 (2009): 1609-1618.

15. Athanasiou, Lambros S., et al. "Error propagation in the characterization of atheromatic plaque types based on imaging." *Computer methods and programs in biomedicine* 121.3 (2015): 161-174.

16. Bauer, Stefan, et al. "A survey of MRI-based medical image analysis for brain tumor studies." *Physics in medicine and biology* 58.13 (2013): R97.

17. Omuro, Antonio, and Lisa M. DeAngelis. "Glioblastoma and other malignant gliomas: a clinical review." *Jama* 310.17 (2013): 1842-1850.

18. Menze, Bjoern, Mauricio Reyes, and Koen Van Leemput. "The Multimodal Brain TumorImage Segmentation Benchmark (BRATS)." (2014).

19. Bauer, Stefan, Lutz-P. Nolte, and Mauricio Reyes. "Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Berlin, Heidelberg, 2011.

20. Buendia, Patricia, et al. "A grouping artificial immune network for segmentation of tumor images." *Multimodal Brain Tumor Segmentation* (2013): 1.

21. Cordier, Nicolas, et al. "Patch-based segmentation of brain tissues." *MICCAI challenge on multimodal brain tumor segmentation*. IEEE, 2013.

22. Doyle, S., et al. "Fully automatic brain tumor segmentation from multiple MR sequences using hidden Markov fields and variational EM." *Procs. NCI-MICCAI BraTS* (2013): 18-22.

23. Pereira, Sérgio, et al. "Automatic brain tissue segmentation of multi-sequence MR images using random decision forests." *Proceedings of the MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS'13)* (2013).

24. Geremia, Ezequiel, Bjoern H. Menze, and Nicholas Ayache. "Spatial decision forests for glioma segmentation in multi-channel MR images." *MICCAI Challenge on Multimodal Brain Tumor Segmentation* 34 (2012).

25. Guo, Xiaotao, Lawrence Schwartz, and Binsheng Zhao. "Semi-automatic segmentation of multimodal brain tumor using active contours." *Multimodal Brain Tumor Segmentation* (2013): 27.

26. Hamamci, Andac, and Gozde Unal. "Multimodal brain tumor segmentation using the tumor-cut method on the BRATS dataset." *Proceeding MICCAI-BRATS* (2012): 19-23.

27. Meier, Raphael, et al. "Appearance-and context-sensitive features for brain tumor segmentation." *Proceedings of MICCAI BRATS Challenge* (2014): 020-026.

28. Menze, Bjoern H., et al. "A generative model for brain tumor segmentation in multi-modal images." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Berlin, Heidelberg, 2010.

29. Menze, Bjoern H., et al. "Segmenting glioma in multi-modal images using a generative-discriminative model for brain lesion segmentation." *Proc MICCAI-BRATS (Multimodal Brain Tumor Segmentation Challenge)* 8 (2012).

30. Reza, S., and K. M. Iftekharuddin. "Multi-class Abnormal Brain Tissue Segmentation Using Texture." *Multimodal Brain Tumor Segmentation* (2013): 38.

31. Raviv, T. Riklin, K. V. Leemput, and Bjoern H. Menze. "Multi-modal brain tumor segmentation via latent atlases." *Proceeding MICCAIBRATS* (2012): 64-73.

32. Shin, Hoo-Chang. "Hybrid clustering and logistic regression for multi-modal brain tumor segmentation." *Proc. of Workshops and Challenges in Medical Image Computing and Computer-Assisted Intervention (MICCAI'12)*. 2012.

33. Subbanna, Nagesh K., et al. "Hierarchical probabilistic Gabor and MRF segmentation of brain tumours in MRI volumes." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Berlin, Heidelberg, 2013.

34. Taylor, Thomas, et al. "Map-reduce enabled hidden Markov models for high throughput multimodal brain tumor segmentation." *Multimodal Brain Tumor Segmentation* 43 (2013).

35. Tustison, Nicholas J., et al. "Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (simplified) with ANTsR." *Neuroinformatics* 13.2 (2015): 209-225.

36. Zhao, Liang, Wei Wu, and Jason J. Corso. "Brain tumor segmentation based on GMM and active contour method with a model-aware edge map." *BRATS MICCAI* (2012): 19-23.

37. Zhao, Liang, Duygu Sarikaya, and Jason J. Corso. "Automatic brain tumor segmentation with MRF on supervoxels." *Multimodal Brain Tumor Segmentation* 51 (2013).

38. Zikic, D., et al. "Context-sensitive classification forests for segmentation of brain tumor tissues." *Proc MICCAI-BraTS* (2012): 1-9.

39. Porz, Nicole, et al. "Multi-modal glioblastoma segmentation: man versus machine." *PloS one* 9.5 (2014): e96873.

40. Taha, Abdel Aziz, and Allan Hanbury. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool." *BMC medical imaging* 15.1 (2015): 29.

41. The Cancer Genome Atlas. TCGA Research Network: http://cancergenome.nih.gov/

42. Bakas, Spyridon, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features", Nature Scientific Data, (2017).

43. Liang, Zhi-Pei, and Paul C. Lauterbur. *Principles of magnetic resonance imaging*. SPIE Optical Engineering Press, 2000.

44. The AFNI program (Analysis of Functional NeuroImages) National Institute of Health, USA.

45. Ellingson, B. M., Patrick Y. Wen, and Cloughesy. T. F. "Modified criteria for radiographic response assessment in glioblastoma clinical trials." *Neurotherapeutics* 14.2 (2017): 307-320.

46. Hajian-Tilaki, Karimollah. "Sample size estimation in diagnostic test studies of biomedical informatics." Journal of biomedical informatics 48 (2014): 193-204.

47. M. Jenkinson, C.F. Beckmann, T.E. Behrens, M.W. Woolrich, S.M. Smith. FSL. NeuroImage, 62:782-90, 2012

48. M. Jenkinson, P.R. Bannister, J.M. Brady, and S.M. Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage, 17(2):825-841, 2002.

49. Ashburner, John, and Karl J. Friston. "Nonlinear spatial normalization using basis functions." *Human brain mapping* 7.4 (1999): 254-266.

50. W. M. Wells, E. L. Grimson, R. Kikinis, and F. A. Jolesz, "Adaptive segmentation of MRI data," *IEEE Trans. Med. Imag.*, vol. 15, pp. 429–442, Aug. 1996.

51. Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. "N4ITK: improved N3 bias correction." *IEEE transactions on medical imaging* 29.6 (2010): 1310-1320.

52. M. Jenkinson, M. Pechaud, and S. Smith. BET2: MR-based estimation of brain, skull and scalp surfaces. In Eleventh Annual Meeting of the Organization for Human Brain Mapping, 2005.

53. Kimberley Mak, Gillian Lieberman. “Imaging in Glioblastoma Multiforme; diagnosis, treatment and follow-up” Harvard Medical School Radiology Clerkship BIDMC. (2008)

54. Gordillo, Nelly, Eduard Montseny, and Pilar Sobrevilla. "State of the art survey on MRI brain tumor segmentation." *Magnetic resonance imaging* 31.8 (2013): 1426-1438.

55. Piedra, E. A. R., Ellingson, B. M., Taira, R. K., El-Saden, S., Bui, A. A., & Hsu, W. "Brain Tumor Segmentation by Variability Characterization of Tumor Boundaries." International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer, Cham, 2016.

56. Zhang, N., Ruan, S., Lebonvallet, S., Liao, Q., & Zhu, Y. "Kernel feature selection to fuse multi-spectral MRI images for brain tumor segmentation." *Computer Vision and Image Understanding* 115.2 (2011): 256-269.

57. Rangayyan, R. M., El-Faramawy, N. M., Desautels, J. L., & Alim, O. A. "Measures of acutance and shape for classification of breast tumors." *Medical Imaging, IEEE Transactions on* 16.6 (1997): 799-810.

58. O'sullivan, F., Roy, S., O'sullivan, J., Vernon, C., & Eary, J. "Incorporation of tumor shape into an assessment of spatial heterogeneity for human sarcomas imaged with FDG-PET." *Biostatistics* 6.2 (2005): 293-301.

59. Zacharaki, E. I., Wang, S., Chawla, S., Soo Yoo, D., Wolf, R., Melhem, E. R., & Davatzikos, C. "Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme." *Magnetic Resonance in Medicine* 62.6 (2009): 1609-1618.

60. Ahmed, Shaheen, Khan M. Iftekharuddin, and Arastoo Vossough. "Efficacy of texture, shape, and intensity feature fusion for posterior-fossa tumor segmentation in MRI." *Information Technology in Biomedicine, IEEE Transactions on* 15.2 (2011): 206-213.

61. Chuanyun, Xu, and Zhang Yang. "Cell Contour Irregularity Feature Extraction Methods based on Linear Geometric Heat Flow Curve Evolution." *change* 7.3 (2014).

62. Corso, Jason J., et al. "Efficient multilevel brain tumor segmentation with integrated bayesian model classification." *IEEE transactions on medical imaging* 27.5 (2008): 629-640.

63. D. Mumford and J. Shah, "Optimal approximation by piecewise smooth functions and associated variational problems," *Commun. Pure Appl. Math*, vol. 42, pp. 577–685, 1989.

64. Chan, Tony F., and Luminita A. Vese. "Active contours without edges." *Image processing, IEEE transactions on* 10.2 (2001): 266-277.

65. K. Siddiqi, Y. B. Lauziére, A. Tannenbaum, and S. W. Zucker, "Area and length minimizing flows for shape segmentation," *IEEE Trans. Image Processing*, vol. 7, pp. 433–443, Mar. 1998.

66. Olver, P.J., Sapiro, G., and Tannenbaum, A. Invariant geometric evolutions of surfaces and volumetric smoothing. *SIAM J. of Appl. Math.*, (to appear).

67. Kichenassamy, S., Kumar, A., Olver, P., Tannenbaum, A., and Yezzi, A. 1995. Gradient flows and geometric active contour models. *Proc. ICCV*, Cambridge.

68. Sapiro, G., Kimmel, R., Shaked, D., Kimia, B.B., and Bruckstein, A.M. 1993. Implementing continuous-scale morphology via curve evolution. *Pattern Recog.*, 26(9):1363–1372.

69. Caselles, Vicent, Ron Kimmel, and Guillermo Sapiro. "Geodesic active contours." *International journal of computer vision* 22.1 (1997): 61-79.

70. Shah, J. 1995. Recovery of shapes by evolution of zero-crossings. Technical Report, Math. Dept. Northeastern Univ, Boston, MA.

71. Terzopoulos, D., Witkin, A., and Kass, M. 1988. Constraints on deformable models: Recovering 3D shape and nonrigid motions. *Artificial Intelligence*, 36:91–123.

72. V. Caselles, F. Catte, T. Coll, and F. Dibos, "A geometric model for active contours in image processing," *Numer. Math.*, vol. 66, no. 1, pp. 1–31, Dec. 1993.

73. S.-C. Zhuand, A. Yuille, "Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 18, no. 9, pp. 884–900, Sep. 1996.

74. Li, C., Xu, C., Gui, C., & Fox, M. D. "Distance regularized level set evolution and its application to image segmentation." *Image Processing, IEEE Transactions on* 19.12 (2010): 3243-3254.

75. Prastawa, M., Bullitt, E., Ho, S., Gerig, G.: A brain tumor segmentation frame- work based on outlier detection. Medical Image Analysis Journal, Special issue on MICCAI 8(3) (2004) 275–283

76. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8) (2000) 888–905

77. Corso, Jason J., Eitan Sharon, and Alan Yuille. "Multilevel segmentation and integrated Bayesian model classification with an application to brain tumor segmentation." *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006*. Springer Berlin Heidelberg, 2006. 790-798.

78. Sharon, E., Brandt, A., Basri,R. Segmentation and boundary detection using multiscale intensity measurements. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Volume I. (2001) 469–476

79. Aerts, Hugo JWL, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink et al. "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach." *Nature communications* 5 (2014).

80. Lawler, Eugene L. *Combinatorial optimization: networks and matroids*. Courier Corporation, 2001.

81. Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation." *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016

82. Holia, Mehfuza, and V. K. Thakar. "Image registration for recovering affine transformation using Nelder Mead Simplex method for optimization." *International Journal of Image Processing (IJIP)* 3.5 (2009): 218.

83. Sharon, E., Brandt, A., Basri, R. Segmentation and boundary detection using multiscale intensity measurements. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Volume I. (2001) 469–476

84. Komodakis, N., Tziritas, G., Paragios, N.: gle and dynamic MRFs: Setting the state of the art with primal-dual strategies. Comput. Vis. Image Und. 112(1), 14– 29 (2008)

85. Lee, C., Wang, S., Murtha, A., Brown, M., Greiner, R.: Segmenting Brain Tumors using Pseudo–Conditional Random Fields. In: Metaxas, D., Axel, L., Fichtinger, G., Sź ekely, G. (eds.) MICCAI 2008, Part I. LNCS, vol. 5241, pp. 359–366. Springer, Heidelberg (2008)

86. Komodakis, N., Tziritas, G., Paragios, N.: gle and dynamic MRFs: Setting the state of the art with primal-dual strategies. Comput. Vis. Image Und. 112(1), 14– 29 (2008)

87. Huo, Jing, Kazunori Okada, Eva M. van Rikxoort, Hyun J. Kim, Jeffry R. Alger, Whitney B. Pope, Jonathan G. Goldin, and Matthew S. Brown. "Ensemble segmentation for GBM brain tumors on MR images using confidence-based averaging." Medical physics 40.9 (2013).

88. Warfield, Simon K., Kelly H. Zou, and William M. Wells. "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation." IEEE transactions on medical imaging 23.7 (2004): 903-921.

89. Prastawa, Marcel, et al. "A brain tumor segmentation framework based on outlier detection." *Medical image analysis* 8.3 (2004): 275-283.

90. Ashburner, John, and Karl J. Friston. "Unified segmentation." Neuroimage 26.3 (2005)

91. Ashburner, J., Friston, K.J., 1997. Multimodal image coregistration and partitioning—a unified framework. NeuroImage 6 (3), 209–217.

92. Ji, Shiyong, et al. "A new multistage medical segmentation method based on superpixel and fuzzy clustering." Computational and mathematical methods in medicine 2014 (2014). 01-15.

93. Achanta, Radhakrishna, et al. "SLIC superpixels compared to state-of-the-art superpixel methods." IEEE transactions on pattern analysis and machine intelligence 34.11 (2012): 2274-2282.

94. Nechifor, Ruben E., et.al. "Novel Magnetic Resonance Imaging Techniques in Brain Tumors". Topics in Magnetic Resonance Imaging 24.3 (2015): 137B146

95. Gunter, Jeffrey L., et al. "Measurement of MRI scanner performance with the ADNI phantom." *Medical physics* 36.6 (2009): 2193-2205.

96. Menze, Bjoern. , The Multimodal Brain Tumor Segmentation Challenge - results update. MICCAI 2016.

97. Zhang, Hejia, Xia Zhu, and Theodore L. Willke. "Segmenting Brain Tumors with Symmetry." *arXiv preprint arXiv:1711.06636* (2017).

98. Rios Piedra, E., Orosz, I., Zide, M., El-Saden, S., Taira, R., Bui, A., Hsu, W. "A Usability Study to Evaluate the Impact of a Novel Automated Brain Tumor Assessment Application". American Medical Informatics Association, poster presentation, 2017.

99. Gliem, Joseph A., and Rosemary R. Gliem. "Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales." Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, 2003.

100. Hornbaek, Kasper. "Current practice in measuring usability: Challenges to usability studies and research." *International journal of human-computer studies* 64.2 (2006): 79-102.

101. Rios Piedra, E., Taira, R., El-Saden, S., Ellingson, B., Bui, A., Hsu, W. "Assessing Variability in Brain Tumor Segmentation to Improve Volumetric Accuracy and Characterization of Change". Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference. IEEE, oral presentation, 2016.

102. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

103. Kamnitsas, Konstantinos, et al. "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation." *Medical image analysis* 36 (2017): 61-78.