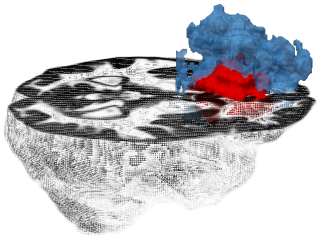
Fakultät für Informatik
Technische Universität München

Dissertation

Multi-modal Multi-temporal Brain Tumor Segmentation, Growth Analysis and Texture-based Classification



Esther Alberts

Fakultät für Informatik
Technische Universität München

Multi-modal Multi-temporal Brain Tumor Segmentation, Growth Analysis and Texture-based Classification

Esther J. Alberts

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende:

Assistant Prof. Dr. Laura Leal-Taixé

Prüfende der Dissertation:

1. Prof. Dr. Björn Menze
2. Prof. Dr. Claus Zimmer

Die Dissertation wurde am 08.10.2018 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 26.02.2019 angenommen.

Abstract

Brain tumor analysis is an active field of research, which has received a lot of attention from both the medical and the technical communities in the past decades. The purpose of this thesis is to investigate brain tumor segmentation, growth analysis and tumor classification based on multi-modal magnetic resonance (MR) image datasets of low- and high-grade glioma making use of computer vision and machine learning methodologies.

Brain tumor segmentation involves the delineation of tumorous structures, such as edema, active tumor and necrotic tumor core, and healthy brain tissues, often categorized in gray matter, white matter and cerebro-spinal fluid. Deep learning frameworks have proven to be among the most accurate brain tumor segmentation techniques, performing particularly well when large accurately annotated image datasets are available. A first project is designed to build a more flexible model, which allows for intuitive semi-automated user-interaction, is less dependent on training data, and can handle missing MR modalities. The framework is based on a Bayesian network with hidden variables optimized by the expectation-maximization algorithm, and is tailored to handle non-Gaussian multivariate distributions using the concept of Gaussian copulas. To generate reliable priors for the generative probabilistic model and to spatially regularize the segmentation results, it is extended with an initialization and a post-processing module, both based on supervoxels classified by random forests.

Brain tumor segmentation allows to assess tumor volumetry over time, which is important to identify disease progression (tumor regrowth) after therapy. In a second project, a dataset of temporal MR sequences is analyzed. To that end, brain tumor segmentation and brain tumor growth assessment are unified within a single framework using a conditional random field (CRF). The CRF extends over the temporal patient datasets and includes directed links with infinite weight in order to incorporate growth or shrinkage constraints. The model is shown to obtain temporally coherent tumor segmentation and aids in estimating the likelihood of disease progression after therapy.

Recent studies classify brain tumors based on their genotypic parameters, which are reported to have an important impact on the prognosis and the therapy of patients. A third project is aimed to investigate whether the genetic profile of glioma can be predicted based on the MR images only, which would eliminate the need to take biopsies. A multi-modal medical image classification framework is built, classifying glioma in three genetic classes based on DNA methylation status. The framework makes use of short local image descriptors as well as deep-learned features acquired by denoising auto-encoders to generate meaningful image features. The framework is successfully validated and shown to obtain high accuracies even though the same image-based classification task is hardly possible for medical experts.

Zusammenfassung

Die Bildanalyse von Gehirntumoren hat sich in den vergangenen Jahren sowohl in der Medizin als auch in den technischen Disziplinen zu einem sehr aktiven Forschungsfeld entwickelt. Diese Dissertation untersucht die Segmentierung, Klassifizierung und das Tumorwachstum von Gliomen auf Basis von multispektralen Magnetresonanztomographiebildern (MRT) mittels Computer Vision und maschinellem Lernen.

Die Segmentierung von Hirntumoren besteht aus der Abgrenzung verschiedener Tumorgewebestrukturen (kontrastmittelanreichernder Anteil, Nekrose und Ödem), sowie Unterscheidung des Tumors von der gesunden Hirnsubstanz (unterteilt in graue -, weiße Substanz und Liquor). Ein zunehmende Zahl von Studien legt nahe, dass neuronale Netze in der Lage sind, Gehirntumore mit sehr hoher Genauigkeit zu segmentieren, insbesondere wenn genügend segmentierte Bildern in der Trainingsphase vorliegen. Im ersten Projekt wird ein alternatives, flexibleres Modell erstellt, welches erlaubt die modellierte Segmentierung durch einfache Benutzerinteraktion zu korrigieren. Es ist zudem weniger von Trainingsdaten abhängig und erfordert nicht dass die MRT-Kanäle in der Applikationsphase vollständig sind. Das Modell basiert auf einem Bayesschen Netz mit versteckten Variablen, welche durch den Expectation-Maximization-Algorithmus gelernt werden. Es nutzt Gauß-Copulas um Abhängigkeiten in einer nicht-normalverteilten, multivariaten Verteilung zu modellieren. Um den räumlichen Zusammenhang der finalen Segmentierung zu gewährleisten, sowie robuste a-priori Wahrscheinlichkeiten für das probabilistische Modell zu erstellen werden zwei Module basierend auf durch Random Forests klassifizierten Superpixeln integriert.

Gehirntumorsegmentierungen erlauben es Tumorvolumen zu messen, was von zentraler Bedeutung für die Kontrolle der Wirksamkeit der Therapie ist. Im zweiten Projekt werden zeitliche MRT-Bildserien analysiert. Gehirntumorsegmentierung und Tumorwachstumsanalyse werden mittels eines Conditional Random Field (CRF) vereint. Dieses CRF enthält gerichtete Verbindungen, welche Wachstums- oder Schrumpfungsbeschränkungen entsprechen. Das Modell generiert zeitlich kohärente Segmentierungen und schätzt die Wahrscheinlichkeit von Tumorwachstum nach einer Behandlung.

Neuere Studien identifizieren genetische Parameter als gute prognostische Biomarker für Gehirntumore. Darauf aufbauend wird in einem dritten Projekt untersucht, ob die relevanten genetischen Parameter auf Basis von MRT-Bildern vorhergesagt und auf diese Weise risikoreiche Biopsien vermieden werden können. Hierzu wird ein multispektrales Bildklassifizierungsmodell kreiert, welches Gliome in drei durch DNA Methylierung definierten genetischen Klassen zuweist. Dieses Modell berechnet charakteristische Bildmerkmale automatisiert mittels sog. “short local image” Deskriptoren und neuronaler Netze. In der Validierung erreicht die Methodik eine hohe Genauigkeit, obwohl die Aufgabe selbst für medizinische Experten kaum lösbar ist.

Acknowledgments

Foremost, I'd like to thank both my supervisors, prof. B. Menze from the Computer Science department and prof. C. Zimmer from the neuroradiology department of the university hospital Klinikum Rechts der Isar, for giving me the opportunity to conduct research on a fascinating research topic, so relevant for clinical practice, in such a stimulating environment. Working in close cooperation with the neuroradiology department, it was an absolute privilege to regularly interact with medical doctors, amongst which Thomas Huber, Jan Kirschke and Benedikt Wiestler. I highly appreciate that, even amidst busy schedules, they always found time slots to meet, discuss, exchange ideas and give constructive feedback. Special gratitude goes out to Benedikt, for our fruitful collaboration and for giving me the opportunity to extend my research topic towards radiogenomics and genetical phenotyping.

In addition, I would like to thank all my lab mates and office colleagues, for inspiring and supporting me and for the many lunch and coffee breaks we have shared. Special thanks to Marie, Markus and Jana, for emotional, motivational and research-related support throughout my PhD years, and to Anjany, Giles and Florian, for the many interesting technical discussions and the collaboration we established towards the end of my PhD.

With a special mention to my master and bachelor students, Alexandra, Hang, Tolu, Santiago, Christoph, Stevica, Yusuf and Enes. Thank you for your thrust and good work, I very much enjoyed our collaboration and wish you all the best for the future!

To my parents, my siblings, my family, to my close friends and to Yoann, thank you for your patience, for your support, for believing in me and for helping me to achieve my goals.

Contents

List of Figures

List of Tables

1 Introduction

Brain tumor analysis covers a broad spectrum of medical and technical research topics. Section 1.1 provides a concise motivation for each research project elaborated in this thesis as well as their main scientific contributions. Section 1.2 discusses the general outline of the thesis.

1.1 Motivation and Contributions

Primary malignant brain tumors have a high impact on society and the health care system. They account for approximately 2% of the total cancer burden in adults and are associated with very poor survival rates in comparison to other types of cancer in adults [13]. Today, brain tumor analysis is an active field of research, receiving a lot of attention from both the medical and the technical communities. This thesis focuses on the following three brain tumor analysis topics: brain tumor segmentation and uncertainty assessment, brain tumor growth analysis and brain tumor classification.

Brain tumor segmentation. Brain tumor segmentation is highly valuable for clinical practice. Firstly, it can aid in brain tumor diagnosis and staging. Secondly, it can assist in treatment planning, including radiotherapy planning and computer-assisted surgery. And finally, because it allows to directly assess tumor volume, it can aid in treatment assessment and help to detect disease progression. In comparison to manual brain tumor segmentation, *automated* brain tumor segmentation techniques are particularly interesting. They can be considerably less time-consuming, less costly and more objective. In recent years, (semi-)automated brain tumor segmentation has become a major field of research within the medical image analysis community. Since 2012, the *Multi-modal Brain Tumor Image Segmentation Benchmark* (BRATS) challenge is organized each year in conjunction with the international conference on *Medical Image Computing and Computer Assisted Interventions* (MICCAI) [2]. The BRATS challenge includes a vast set of accurately annotated and neatly preprocessed training data and test data. This dataset is carefully reviewed by multiple expert neuroradiologists and grows larger and more accurate every year. In addition to the quality and size of the training and test data, the success of the challenge is strongly associated with the public availability and easy access to the training data (including ground truth segmentations) and test data (without ground truth segmentations – these are kept to the organizers for internal validation).

Building further on the unsupervised segmentation method presented in [8], this thesis presents a framework for robust, fast and accurate (semi-)automated brain tumor

segmentation. To this end, it makes use of generative probabilistic models, generating brain tumor segmentations by grouping voxels with similar statistics in an arbitrary high-dimensional feature space. In Chapter 4, the probabilistic model is extended with a Conditional Random Field (CRF), applied directly on the probabilistic segmentation maps generated by the probabilistic model. The purpose of the CRF is to incorporate spatial neighborhood dependencies, such that the final segmentations are spatially coherent. Traditionally, the likelihood model in probabilistic models implemented for image segmentation is modeled by a mixture of multivariate Gaussians. In Chapter 7, the probabilistic model is specifically tailored to cope with non-Gaussian intensity distributions, by implementing Gaussian copulas and univariate Gumbel distributions in the likelihood model. Furthermore, random forest supervoxel classifiers are presented as a new technique to generate reliable priors for the probabilistic model as well as to spatially regularize the segmentations generated by the probabilistic model (replacing the CRF proposed in Chapter 4).

Brain tumor segmentation will play a central role throughout the thesis. All other research projects explored in this thesis – brain tumor classification, uncertainty assessment and growth analysis – rely on brain tumor segmentation frameworks.

Longitudinal datasets and growth analysis. The very first research project, presented in Chapter 4), was established as a collaboration with the neuroradiology department of the university hospital[1]. In this project, a framework is presented for *longitudinal* brain tumor segmentation, in which segmentations are calculated for several time points at once. For this purpose, the neuroradiology department provided a longitudinal image dataset, containing 30 glioblastoma patients together with expert annotated ground truth for edema and active tumor [4].

Inspired by the work presented in [14], this data is used to build and test a framework in which longitudinal brain tumor segmentation and *growth assessment* are combined. In a first step, segmentations are calculated for each time point separately by means of a generative probabilistic model. In a second step, these segmentations are refined for all time points at once, using a conditional random field (CRF) spanning over the entire multi-modal, longitudinal sequences. The CRF is extended with growth and shrinkage constraints along time, which are implemented by directed links with infinite weight. The likelihood of such a constraint can be assessed by means of the energy values associated with the maximum-a-posteriori (MAP) tumor segmentation obtained under different configurations of the growth and shrinkage constraints.

Uncertainty assessment of brain tumor segmentation. Segmentation algorithms often have to make decisions in the presence of high uncertainty due to the high variability in tumor location, tumor shape and tumor volume, due to images suffering from motion artifacts or due to the inherent inhomogeneity of clinical data. Therefore, uncertainty assessment can be very valuable for the clinical interpretation of the segmentation results.

[1]Department of Neuroradiogy, Klinikum Rechts der Isar, Technische Universität München, Germany

In the second project, presented in Chapter 5, the framework in [8] is used for brain tumor segmentation and extended with a CRF for regularization purposes. Brain tumor segmentation uncertainty is assessed, both on a voxel and a supervoxel level, based on random MAP perturbation models, previously proposed in [15, 16], for the purpose of CRF inference uncertainty assessment.

Genomic brain tumor classification. In a third project, presented in Chapter 6, a new type of dataset becomes available. This dataset includes both image data as well as genomic data. Medical collaborators at the neuroradiology department of the university hospital[1] processed the dataset and selected ∼120 patients, each with multi-modal MR images together with one genetic tumor class label. These genetic tumor classes are based on the genetic tumor profile, which is determined via a biopsy. Medical research studies suggest that some of the parameters in the genetic tumor profile have an important influence on the choice of treatment and disease outcome [17], [18]. In this challenging project, it is investigated whether the genetic tumor profile can be predicted solely based on image data, omitting the need for a biopsy. From a medical perspective, this third project belongs to a new type of research field – referred to as *radiogenomics* – in which genetic parameters are studied in relation to imaging features. From a technical perspective, the computer vision task is now shifted from image segmentation towards image classification. *Medical image classification* is an entirely different field, with its own particular challenges. In particular, the feature space, consisting of multiple 3D images, has an extremely high dimensionality. These kind of classification tasks, in which the dimensionality of the feature space largely exceeds the number of samples, is often symbolically written as $p \gg n$. In these problems, *feature engineering*, including feature extraction and feature selection, plays a more prominent role. Deep-learning frameworks offer an elegant solution to these kind of high-dimensional image classification tasks, because they optimize feature generation and classification at once, using only one error function.

In Chapter 6, several types of image features are extracted. They include 3D linear binary patterns (LBPs), efficient 3D local binary descriptors known as BRIEFs and low-dimensional image representations generated by deep convolutional denoising auto-encoders. Furthermore, several classifiers are validated on subsets of the generated features as well as on the complete set of features, both with and without a preceeding principal component analysis (PCA) on the feature space. Investigated classifiers include random forests, support vector machines, multi-layer perceptrons, k-nearest neighbours and logistic regression. Even though this classification task is hardly possible for neuroradiology experts, the proposed framework is shown to obtain promising classification accuracies, performing far better than a majority-voter.

Semi-automated brain tumor segmentation. In a last project, a brain tumor segmentation framework is presented with semi-automated extensions in order to provide a useful tool for medical doctors or other users interested in segmenting brain tumors in a quick and transparent way. The framework makes use of the probabilistic network

presented in [8] and additionally adopts supervoxel classifiers, which allow to make abstraction of the fine voxel level. As previously described, these supervoxel segmenters first serve to generate label prior maps, adopted in the probabilistic model. After the probabilistic model, a second supervoxel segmenter serves to spatially regularize the segmentations. An important contribution is that the interactive framework is implemented on the supervoxel level instead of on a fine voxel level. This allows for extremely fast and efficient user interaction.

1.2 Outline

In this thesis, various topics of brain tumor analysis are investigated. The thesis is divided in two main parts. The first part presents the medical and technical background for brain tumor analysis. The medical background is presented in Chapter 2 and provides medical insights on the types of malignant brain tumors, more in particular for glioma and glioblastoma, as well as their treatment and prognosis (Section 2.1). Image acquisition plays an important role in brain tumor diagnosis and follow-up. Chapter 2 therefore also provides a concise background on the physics of MRI imaging (Section 2.3) and illustrates the MRI imaging protocol for brain tumor patients (Section 2.4). The technical background, addressed in Chapter 3, first discusses image preprocessing techniques typically conducted before performing subsequent image analysis (Section 3.3). Next, methods for image segmentation (Section 3.4) and image classification (Section 3.5) are presented in more technical detail.

In the second part of this thesis, the different research projects are presented (Chapters 4-7). In Chapter 4, a framework conducting brain tumor segmentation and tumor growth analysis in parallel is presented. The framework is based on a Conditional Random Field (CRF) that acts on longitudinal datasets, studying in retrospect whether the tumor has grown or rather shrunken between two time points in the given sequence. Chapter 5 presents a project which aims to assess the uncertainty of a segmentation calculated by a CRF, by means of random perturbation models. Chapter 6 presents an image classification project. In this project, the genotype of a tumor is automatically predicted based on a set of multi-modal MR images. The pipeline for image classification is based on feature extraction and standard machine learning classifiers, and is quite different from the segmentation pipelines discussed in the other chapters of the thesis. Lastly, in Chapter 7, a multi-scale brain tumor segmentation approach is presented, which aims to make brain tumor segmentation more robust compared to the segmentation techniques used in Chapter 4 and 5. The thesis is concluded with a discussion of the obtained project goals and an outlook towards follow-up research projects.

Part I

Background

2 Medical Background

This chapter provides the medical background of brain tumor types, histology, diagnosis and in particular MRI physics and image modalities. The projects presented in Chapters 4-7 cover various topics of brain tumor analysis. Every project uses a dataset of brain MRI images of patients with brain tumors. The brain tumors in these datasets will always be either low-grade or high-grade glioma, and the majority of them are glioblastoma (high-grade glioma). Section 2.1 presents an overview of brain tumor types and incidence rates and explains the classification, grading and prognosis of glioma and glioblastoma in particular. Section 2.2 provides a general overview on medical imaging techniques and Section 2.3 discusses MR imaging in more detail. Finally, Section 2.4 explains and illustrates the standard MR imaging protocol for brain tumor diagnosis, illustrating different tumor structures and healthy brain tissues.

2.1 Malignant Brain Tumors: Glioma and Glioblastoma

Many types of brain tumors exist. They can behave in two ways: either they are cancerous (malignant), or they are not cancerous (benign). Brain tumors are often also classified by the primary tumor site: primary brain tumors originate in the brain, and secondary, or metastatic, brain tumors are spread from tumors that originated in other parts of the body (most commonly the lungs, breasts, melanomas and renal or colorectal sites [19]).

2.1.1 Brain Tumor Types and Incidence

Primary malignant brain tumors account for approximately 2% of the total cancer burden in adults [13]. Though this incidence rate is considerably lower than those of other frequent cancers in adults (i.e. breast cancer at 15% or lung cancer at 14% [13]), its impact is relatively high, because it is associated with poor survival rates (cf. Section 2.1.3).

A statistical study in the US in between 2008-2012 [1] states that brain tumors have a yearly incidence rate of 28.57 in a population of 100 000 adults (age 20+ years) and that brain tumors are the most common cancer reported in children (age 0-19 years), being the second most common cancer leading to child death, after leukemia. The same study reports that the median age of all primary brain tumors is 59 and that 55% of all malignant brain tumors are diagnosed in men.

Figures 2.1, 2.2 and 2.3 represent the distribution of malignant and benign brain and CNS tumors, of different types of glioma and of all malignant brain and CNS tumors. These distributions have also been calculated by the CBTRUS Statistical Report 2015

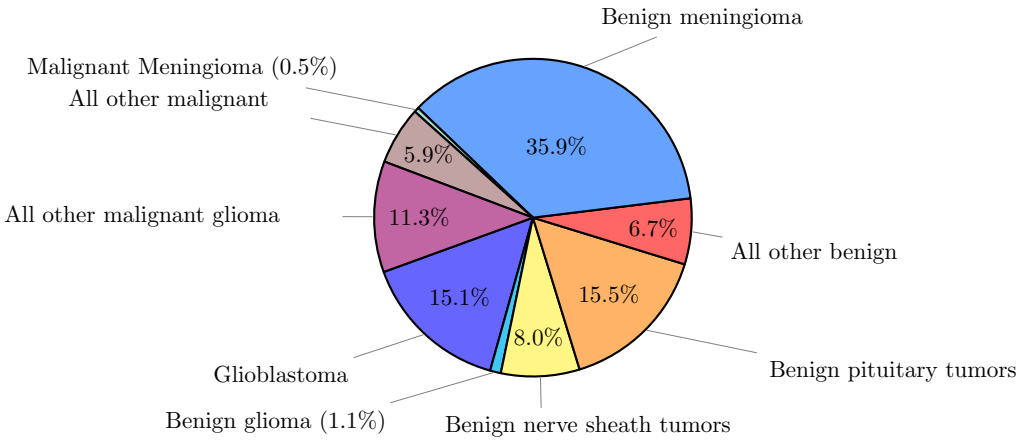


Figure 2.1: Distribution of primary brain and CNS tumors by behavior (malignant or benign). $N = 356858$, among which $N = 239835$ (67.2%) benign and $N = 117023$ (32.8%) malignant. Figure reconstructed from "CBTRUS Statistical Report" [1]

[1], which reports incidence and survival rates of brain and CNS tumors reported in the US in between 2008-2012.

Glioma is an important type of primary brain tumors. They account for ca. 30% of all brain and central nervous system (CNS) tumors and ca. 80% of all malignant brain tumors [20]. Other brain and CNS tumor types include meningioma, nerve sheath tumors and pituitary tumors. Figure 2.1 shows the distribution of primary brain and CNS tumors, distinguishing between malignant and benign behavior. Meningioma, nerve sheath tumors and pituitary tumors are very common and are almost always benign. Meningioma, for example, make up for 36.4% of all primary brain and CNS tumors and are benign in ca. 98.6% of the cases. Glioma are also very common, accounting for 27.5% of all primary brain and CNS tumors. However, gliomas are almost always malignant (in ca. 96.0% of the cases). From Figure 2.1, it is also clear that glioma is indeed the most dominant malignant brain tumor (accounting for **80% of all primary malignant** brain and CNS tumors).

Several types of glioma exist, including astrocytomas, oligodendrogliomas, oligoastrocytomas and ependymomas. The distribution of gliomas is visualized in a pie diagram in Figure 2.2. **Glioblastoma**, also called Glioblastoma Multiforme (GBM), is a glioma type, more in particular a malignant astrocytoma, which is particularly aggressive and very common. They account for 55.1% of all glioma (Figure 2.2) and for 15.1% of all primary brain and CNS tumors (Figure 2.1). Figure 2.3 illustrates the distribution of all primary malignant brain and CNS tumors. Indeed, the vast majority consists of glioma types such as astrocytomas, oligodendrogliomas, oligoastrocytomas and ependymomas. Glioblastoma, in particular, is by far the most common type of primary malignant brain tumors, accounting for **46.1% of all primary malignant** brain and CNS tumors.

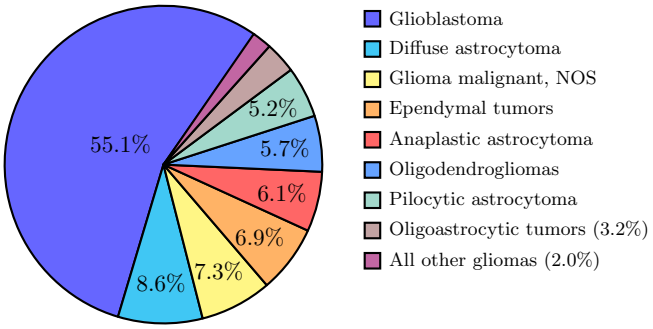


Figure 2.2: Distribution of primary brain and CNS gliomas by histology subtypes (N = 97910). Figure reconstructed from "CBTRUS Statistical Report" [1]

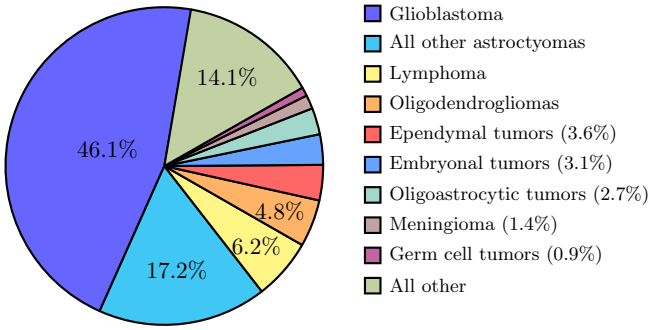
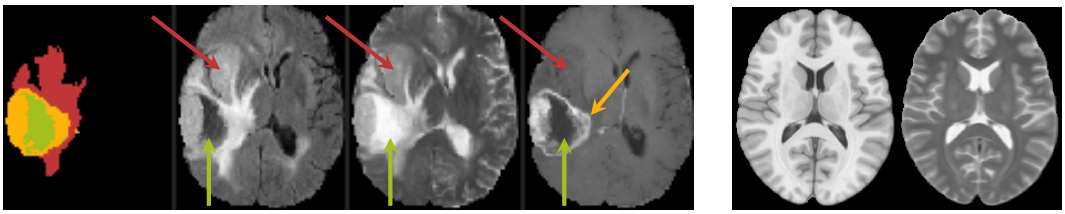


Figure 2.3: Distribution of malignant primary brain and CNS tumors by histology (N = 117023). Figure reconstructed from "CBTRUS Statistical Report" [1]



(a) Tumor structures in a glioblastoma patient **(b)** Healthy population atlas

Figure 2.4: Visualization of a brain tumor. (a) Tumor segmentation containing active tumor (●), edema (●) and necrotic tumor core (●) followed by three MR modalities: FLAIR, T2 and T1c (BRATS 2016 training case [2]). (b) T1 and T2 axial slices from the SRI24 [3] healthy population atlas.

2.1.2 Glioma: Histology, Classification and WHO Grading

Figure 2.4 illustrates a brain tumor for a glioblastoma patient. Figure 2.4a illustrates the delineation of three different brain tumor structures and, for comparison, a healthy population atlas is shown in Figure 2.4b, reflecting the brain anatomy of a healthy patient. Gliomas usually consist of several tumor substructures, which will often be referred to in this thesis. In medical image segmentation challenges, such as BRATS [2], it is common to differentiate between

1. an active tumor core, consisting of invasive proliferating tumor cells,
2. a necrotic tumor core, consisting of damaged or dying cells,
3. perifocal edema, a swelling surrounding the tumor core. Edema is associated with an abnormal accumulation of water in the affected tissue, often caused by glioma cells, and is known to facilitate glioma cell invasion [21].

Other abnormal tissues which can be caused by glioma, are brain cysts and hemorrhages. Hemorrhages are internal bleedings, which can for example occur when the intratumor vascularization is well developed and consists of dilated, thin-walled vessels as well as tumor necrosis [22]. Examples of brain hemorrhages in MR images are further discussed in Section 2.4 (cf. Figure 2.13a). A brain cyst is a fluid-filled sac which can contain blood, pus or CSF. They can develop in low-grade gliomas, but are rather rare in GBMs [23].

Gliomas can be classified into histological subtypes, such as astrocytomas, oligodendrogliomas, oligoastrocytomas, ependymomas and subgroups thereof, as seen in Figure 2.2. A more extensive histological classification of glioma is listed in the *international classification of diseases for oncology* (ICD-O) [19], which also associates a histological ICD code with each human tumor type. Similarly, gliomas have been classified according to their malignancy scale. In particular, the WHO grading of brain and CNS tumors [9], differentiates between four grades (grade I, II, III and IV) based on histological behavior and disease outcome (Table 2.1). Lower grades are associated with less aggressive tumors and often have more favorable disease outcomes. Table 2.2 gives an overview of the 2007 WHO grading for the most common glioma types. All anaplastic astrocytomas, olgiodendrogliomas and oligoastrocytomas are classified as WHO grade III. Glioblastomas are the most aggressive glioma type, classified as WHO grade IV.

In order to better differentiate between different tumor entities, the WHO 2016 classification and grading of brain and CNS tumors includes molecular genetic parameters, in addition to histology [24]. In the past years, several studies reported that the genotypic parameters of glioma have an impact on the prognosis and the therapy of the patients. In particular, IDH mutations and 1p/19q codeletion seem to play a key-role in patient outcome [17], [18]. To determine these genetic parameters, genotyping is required. In chapter 6, an algorithm is developed to automatically detect IDH mutations (IDH wildtype vs. IDH mutant) and 1p/19p codeletion, based on MRI images.

Table 2.1: WHO grading of brain and CNS tumors [9]

	Histological behaviour	**Treatment/disease progression**
I	Low proliferation	Can be cured with surgical resection alone
II	Low proliferation, but generally infiltrative	Often recur
III	Histological evidence of malignancy: nuclear atypia and brisk mitotic activity	Patients often receive adjuvant radiation and/or chemotherapy
IV	Cytologically malignant, mitotically active, necrosis-prone tumors	Often rapid pre- and postoperative disease evolution and a fatal outcome

Table 2.2: 2007 WHO grading of glioma types, extracted from [9]

	I	II	III	IV
Astrocytic tumors				
Subendymal giant cell astrocytoma	•			
Pilocytic astrocytoma	•			
Pilomyxoid astrocytoma		•		
Diffuse astrocytoma		•		
Pleomorphic xanthoastrocytoma		•		
Anaplastic astrocytoma			•	
Glioblastoma				•
Giant cell glioblastoma				•
Gliosarcoma				•
Oligodendroglial tumours				
Oligodendroglioma		•		
Anaplastic oligodendroglioma			•	
Oligoastrocytic tumours				
Oligoastrocytoma		•		
Anaplastic oligoastrocytoma			•	

2.1.3 Prognosis and treatment

The average estimated 5-year survival rate is 34.4% for all malignant brain tumors [1]. It is important to note however, that the survival rates for malignant brain tumors

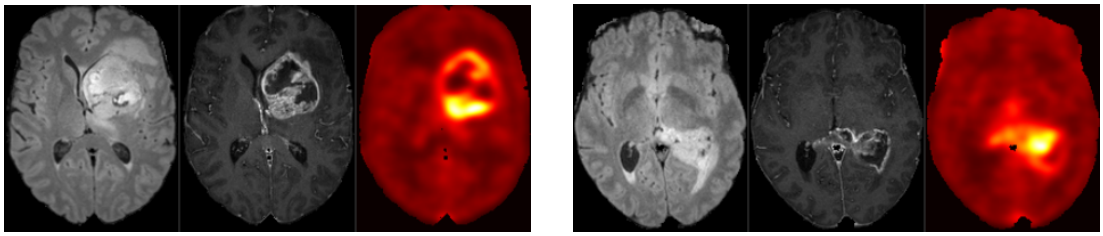


Figure 2.5: MR and PET image acquisition for two glioblastoma patients. *From left to right:* two MR images (FLAIR and T1c) and a 18F-FET PET scan.

Table 2.3: Medical image acquisition: electro-magnetic (EM) wave and radiation types. Based on Chapter 2-5 in [10].

	EM waves	wavelength (m)	radiation type	first image
MRI	Radio waves	$\sim 10^{2}$	non-ionizing	∼1973
Endoscopy	Visible light	$\sim 10^{-7}$	non-ionizing	∼1805
Radiography	X-rays	$\sim 10^{-10}$	ionizing	∼1895
CT	X-rays	$\sim 10^{-10}$	ionizing	1972
Nuclear imaging	γ-rays	$\sim 10^{-12}$	ionizing	1950s

strongly vary with the brain tumor type and its WHO classification. The estimated 5-year survival rate for glioblastoma, for example, is only 5.1% [1].

In general patients with low-grade glioma are associated with longer survival times: patients with WHO grade II glioma usually survive more than 5 years and patients with grade III glioma have survival times of 2-3 years [9]. In contrast, Glioblastomas – WHO grade IV – have a very poor prognosis: the majority of patients have survival times of about 12 to 15 months, with only 3%-5% surviving more than five years [25]. It should be noted that, apart from the WHO grading, multiple factors play an important role on patient survival time, including patient age, extent of surgical resection, contrast enhancement and tumor location [9].

In the WHO 2016 grading of glioma, IDH mutation and 1p/19p codeletion are included [24]. In general, IDH mutant glioma are associated with longer patient survival time compared to IDH wildtype [17]. Similarly, within the IDH mutant cohort, 1p/19p codeletion is also associated with a better outcome [18].

2.2 Medical Image Acquisition

Brain tumors are almost always diagnosed with magnetic resonance (MR) images. In addition, computer tomography (CT) or nuclear medicine images (such as SPECT or PET) might also be acquired in certain institutes or for particular cases. This section provides a general overview of all medical imaging techniques relevant to brain tumor diagnosis.

The most common diagnostic medical imaging techniques include radiography, computed tomography (CT), single-photon emission computed tomography (SPECT), positron emission tomography (PET) and ultrasound (US) imaging. All these imaging techniques are based on the emission or detection of electromagnetic waves, except from US imaging, which is based on the propagation of sound waves. Table 2.3 summarizes different medical imaging techniques, as well as the type and wavelength of the electro-magnetic waves emitted during image acquisition.

Both in radiography and CT, X-rays are emitted by an X-ray tube, pass through human body tissues and are detected by an X-ray detector. SPECT and PET are types of nuclear medical imaging, in which a tracer molecule, carrying radionuclides, is administered to the patient (usually intravenous). These radionuclides are unstable isotopes that emit γ-rays. For PET, a proton in the nucleus of the radionuclide is transformed into a neutron and a positron. This positron then annihilates shortly after with an electron, emitting two γ-photons in opposite directions. That is, in nuclear medical imaging, γ-rays typically originate within the patient body and are detected by γ-ray detectors closely located to the patient. Some studies show that PET scans, in particular (18)F-fluoro-ethyl-tyrosine (18F-FET) PET scans, can have a high diagnostic value for primary brain tumor patients [26]. In Figure 2.5, 18F-FET PET scans are shown for two glioblastoma patients.

In radiography, CT, PET and SPECT, patients are exposed to X-rays or γ-rays. Due to their high frequency (or short wavelength), X-rays and γ-rays are ionizing. Ionizing EM waves are able to eject electrons from their atoms, which means they can damage or destroy cells in the body. Several studies have shown that, depending on the radiation dose received by particular cells, radiation can potentially lead to genetic changes or even malignant tumors [27]. For medical image acquisition, the radiation exposure is therefore kept as low as possible, following the "As low as reasonably achievable (ALARA)" principle [10]. Ionizing radiation is measured in Gray (Gy), for the total absorbed dose, or in Sievert (Sv), for the effective dose. Yearly, we are typically exposed to 2-3 mSv due to natural sources. In comparison, the effective dose of a skull CT is 1-2 mSv [10]. Eventually, examinations involving ionizing radiation, especially CT and PET examinations, should be kept at a minimum and every individual examination should be medically justified.

Magnetic resonance imaging (MR or MRI) is one of the most frequently used modern medical image acquisition techniques. In contrast to radiography, CT, PET and SPECT, MRI uses a strong external magnetic field and radio-frequency (RF) waves, rather than ionizing radiation, in order to generate medical images (cf. Table 2.3). The next section is devoted to the physics and acquisition of MR images.

2.3 Magnetic Resonance Imaging

Radiography and CT visualize the absorption of x-rays in human tissue, which depends on the linear attenuation coefficient of the atoms or molecules contained in the body tissues. PET and SPECT visualize the spatial and temporal distribution of a tracer in

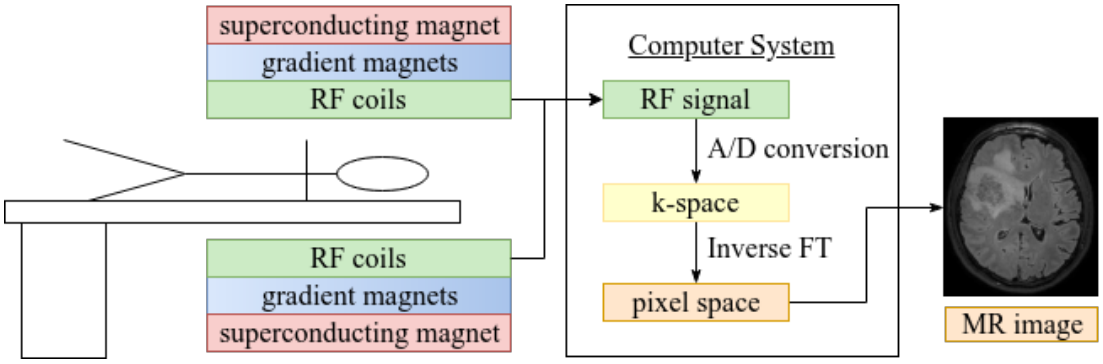


Figure 2.6: Setup of an MRI scanner. A patient is positioned in a tunnel or a C-shaped system which contains superconducting magnets, gradient magnets and RF coils. A computer system generates MR images by applying the inverse Fourier transform on the k-space presentation.

the human body based on the γ-rays that originate closely to the tracer, and are hence typically used to measure metabolism, perfusion, tumors or certain organ functions. MRI on the other hand, relies on the magnetic properties of hydrogen nuclei to produce images. These magnetic properties most often rely on the spin-spin relaxation (T_2) as well as the spin-lattice relaxation (T_1) of hydrogen protons after (a sequence of) radio-frequency (RF) pulses. These T_1 and T_2 relaxation times strongly depend on the tissue type in which the hydrogen protons are located. Hence, MRI is able to obtain anatomical images of all parts of the human body that contain hydrogen. Depending on the MRI acquisition scheme, MRI can even reach better contrast between different soft tissues than with CT [10]. The next sessions elaborate on the setup of an MRI scanner, on the magnetic properties of hydrogen nuclei in an external magnetic field, their behavior after RF pulses and how this is used to obtain MR images.

2.3.1 MRI setup

The first MR image was taken in 1973, by Paul C. Lauterbur, who later shared the Nobel Prize in Medicine or Pyhysiology in 2003 with Peter Mansfield [10]. Nowadays, an MRI system exists of 1) strong magnets establishing a constant external magnetic field B_0, often 1.5 T or 3 T, 2) gradient magnets which allow for spatial encoding along the x, y and z-axis, 3) radio-frequency (RF) coils which can emit RF pulses and detects RF signals that arise due to the magnetization vectors established by the hydrogen nuclei and 4) a computer system that generates anatomical images, as illustrated in Figure 2.6.

2.3.2 Hydrogen nuclei in an external magnetic field

MRI depends on the magnetic properties of hydrogen nuclei. A hydrogen nucleus, ${}^1_1\mathrm{H}$, is equal to one proton. Protons have a positive charge and spin about their axis, hence establishing a magnetic moment μ. When a hydrogen nucleus is placed in an

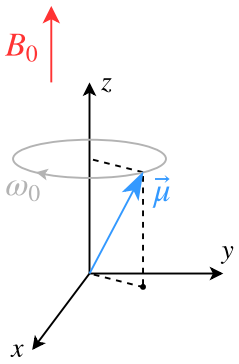


Figure 2.7: Illustration of magnetic moment μ of a hydrogen nucleus under an external magnetic field B_0 along the z-axis. The particle precesses about the z-axis with the angular frequency ω_0, which is equal to the Larmor angular frequency.

external magnetic field B_0, the particle starts precessing about the z-axis with an angular frequency ω_0, which depends on the applied external magnetic field:

$$\omega_0 = \gamma B_0 \ , \tag{2.1}$$

where γ is the gyromagnetic ratio ($\gamma/2\pi = 42.57\,\text{MHz/T}$ for ^1_1H). This is illustrated in Figure 2.7.

Because of the precession of the magnetic moment μ about the z-axis, the proton has a potential energy E. From quantum mechanics, we know that a hydrogen nucleus can only have two energy values under an external magnetic field B_0:

$$E_\uparrow = -0.5\frac{h}{2\pi}\gamma B_0 \quad \text{and} \quad E_\downarrow = +0.5\frac{h}{2\pi}\gamma B_0 \ , \tag{2.2}$$

with h the Planck constant. Most protons are in the lower energy state, also called "spin up" ($\uparrow$). A proton can switch to the higher energy state "spin down" ($\downarrow$), by absorbing a photon with energy $E_{\text{RF}} = E_\downarrow - E_\uparrow = (h/2\pi)\gamma B_0$. Because the energy of a photon can be written as a function of its EM wave frequency, $E_{\text{RF}} = (h/2\pi)\omega_{\text{RF}}$, we see that:

$$\omega_{\text{RF}} = \gamma B_0 \ , \tag{2.3}$$

where ω_{RF} is called the Larmor angular frequency. From (2.1), we see that ω_0 is in fact equal to ω_{RF}, which means that the precession frequency of a hydrogen nucleus also equals the Larmor frequency. In order to switch from a spin up to a spin down state, a proton must absorb a photon associated with an EM wave at the Larmor frequency. For a hydrogen nucleus under an external field of 1 T, this frequency is about 42.6 MHz, which falls in the spectrum of the radio-frequency (RF) waves.

2.3.3 RF pulses and relaxation

MRI systems use RF pulses, to flip hydrogen nuclei from spin up to spin down state. In dynamic equilibrium, the protons exhibit a net magnetization vector in the direction of the external magnetic field, $\mathbf{M_{tot}} = (0, 0, M_0)$, because more protons are in spin up (lower energy state) then there are in spin down (cf. Figure 2.8 (1)). The net

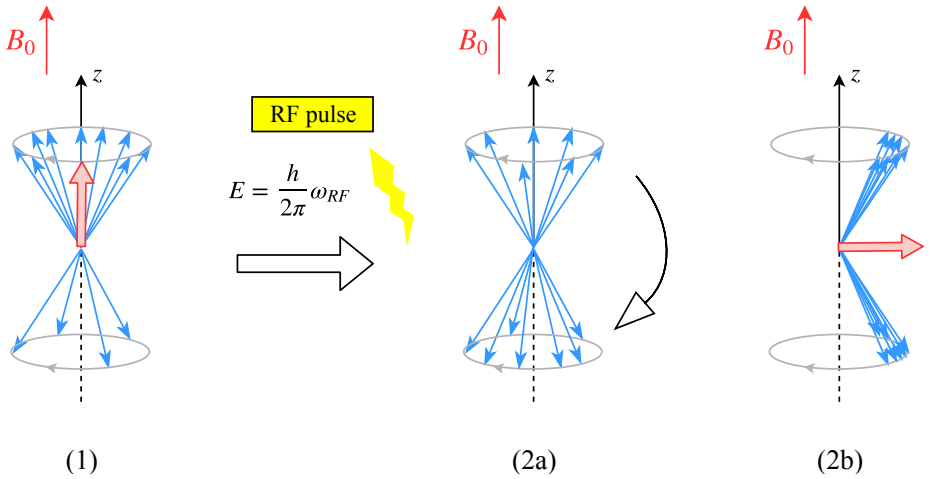


Figure 2.8: Illustration of disturbance of dynamic equilibrium of hydrogen nuclei by a 90° RF pulse: (1) dynamic equilibrium, (2a) protons flip to the higher energy state, from spin up (↑) to spin down (↓) and (3a) protons start precessing in phase. (2a) and (2b) happen simultaneously.

magnetization vector has no transverse component, because the xy-components of the individual magnetic moment of the protons cancel themselves out. During a 90° RF pulse, protons are flipped such that 50% of the protons is in spin up, and the other 50% in spin down, which reduces the longitudinal component of the net magnetization vector to zero (cf. Figure 2.8 (2a)). In addition, and simultaneously, the protons start precessing in phase, hence establishing a net magnetization vector in the transverse plane, $\mathbf{M_{tot}} = (M_0 \cos\theta, M_0 \sin\theta, 0)$ (cf. Figure 2.8 (2b)).

After the RF pulse, the hydrogen nuclei "relax" back to their dynamic equilibrium. Two types of relaxation exist: T2 relaxation and T1 relaxation, both illustrated in Figure 2.9. During T2 relaxation, or spin-spin relaxation, the spinning hydrogen nuclei loose their phase coherence. This results in a decrease of the transverse component of the net magnetization vector $\mathbf{M_{xy}}$. In fact, after a 90° RF pulse, $\mathbf{M_{xy}}$ decreases exponentially in time with a time constant T_2:

$$|M_{xy}| = M_0 \exp -\frac{t}{T_2} \ . \tag{2.4}$$

During T1 relaxation, or spin-lattice relaxation, the protons flip back to their lower energy state, releasing the excess energy as heat into the tissue. Hereby, the longitudinal component of the net magnetization vector, M_z is re-established. Similarly, this process can be described in time with a time constant T_1. After a 90° RF pulse:

$$|M_z| = M_0(1 - \exp -\frac{t}{T_1}) \ . \tag{2.5}$$

T_1 and T_2 are tissue properties and differ based on the extent and the motion of the molecules the protons are bound to. For CSF, T_2 is 20 times larger than for fatty

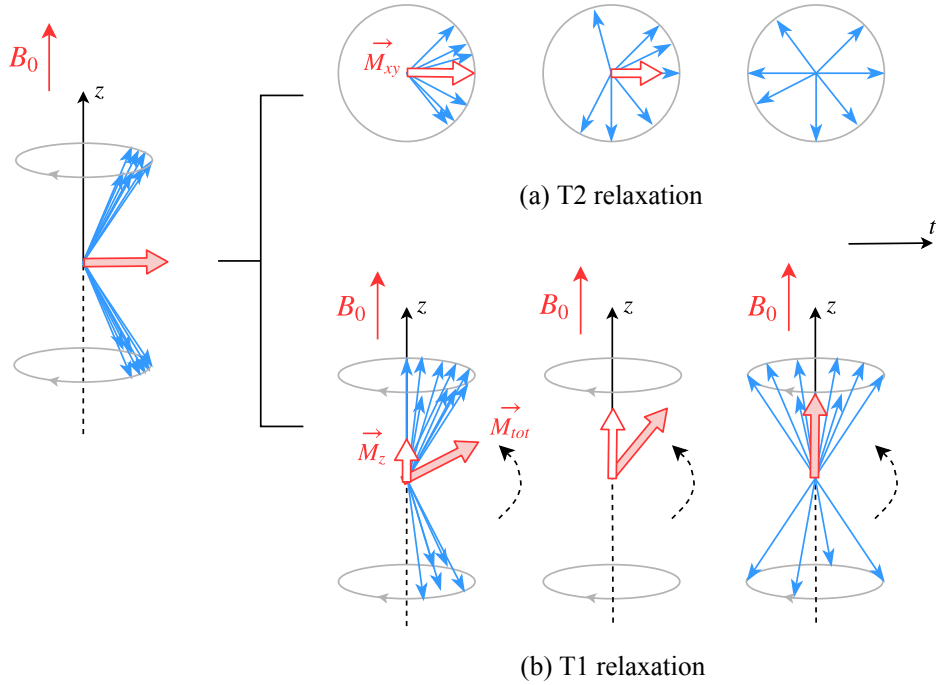


Figure 2.9: Illustration of relaxation after an 90° RF pulse. During T2 relaxation, or spin-spin relaxation, the protons dephase, resulting in the annulment of the transverse magnetic moment $\mathbf{M_{xy}}$ (a). During T1 relaxation, or spin-lattice relaxation, the "spin down" (↓) protons flip back towards their lower energy state "spin up" (↑), reestablishing the longitudinal magnetic moment $\mathbf{M_z}$ from the dynamic equilibrium (b).

tissues such as WM ($T_2 = 2000\,\text{ms}$ for CSF vs. $T_2 = 100\,\text{ms}$ for fat [10]). Similarly, T_1 is 15 times as large for CSF than it is for fat ($T_1 = 3000\,\text{ms}$ for CSF vs. $T_1 = 200\,\text{ms}$ for fat [10]).

2.3.4 From tissue-specific relaxation to measuring positional image data

In MRI systems, the transverse component of the net magnetization vector, M_{xy}, can be measured by means of the RF coils which are visualized in Figure 2.6. The RF coils detect a RF signal proportional to the transverse component:

$$s(t) = M_{xy} \exp -i\omega_0 t \ , \tag{2.6}$$

where the complex exponential function denotes clockwise rotation in the xy-plane at the precession frequency ω_0. Remember that ω_0 is in fact the Larmor frequency from (2.3), which is proportional to the external magnetic field B_0. In order to encode positional information, which is for now absent in (2.6), linear magnetic field gradients along the x-, y- and z-axis are superimposed at different time points on the external magnetic field, causing the net magnetization vector to rotate at different precession

frequencies at different positions and at different time points in the (x, y, z) coordinate system. This allows us to measure the transverse component of the net magnetization vector at different positions. In fact, this is done with the help of the k-theorem, which states that the measured RF signal $s(t)$ is equivalent to the Fourier transform of the MR image to be acquired [10]. For more mathematical and technical details on how the whole Fourrier space is sampled using different gradients at different time points, the interested reader is referred to [10].

2.3.5 MR modalities

We thus-far explained the basic principles behind MR imaging, elaborating on proton spins, magnetization vectors and RF pulses, how T1 and T2 relaxation depend on tissue types and how positional information can be encoded based on different precession frequencies. In MR acquisition, various kinds of images can be acquired, each enhancing particular tissues or structures in the brain. These different scans are sometimes called MR sequences, MR channels or MR modalities. In the following, we will refer to them as **MR modalities**.

Firstly, to acquire different **anatomical MR modalities**, different acquisition schemes can be used, based on different scanning protocols (cf. top row in Figure 2.10). These acquisition schemes typically encode 1) when and what kind of RF pulses are imposed (e.g. 90° or 180°), 2) when the magnetic gradients along the x-, y- or z-axis are activated and 3) when the RF signal is measured. Different pulse sequences are possible. Spin echo (SE), for example, first applies a 90° RF pulse, then a 180° RF pulse and then measures the signal. Gradient echo (GE), on the other hand, only imposes one RF pulse, typically smaller than 90°. TurboSE or TurboGE are similar to SE and GE, but measure several echoes per excitation. Two important parameters are the *repetition time*, TR, and the *echo time*, TE. TR is the time between two RF pulses. If one excitation consists of multiple RF pulses (e.g. as in spin echo), TR is the time between the first RF pulses of two consecutive excitations. TE is the time between the first RF pulse and when the signal is measured. T1-weighted images typically have a short TR and a short TE, while T2-weighted images typically have a long TR and a long TE. Another relevant parameter is the *inversion time* (TI). After a 180° RF pulse, the net magnetization vector is inversed (flipped 180° along the negative z-axis). During relaxation, after the inversion time (which equals $0.69T_1$), the longitudinal component of the net magnetization vector switches from the negative to the positive z-axis, and thus becomes temporally zero. Hence, by combining an inversion pulse and a tissue-dependent inversion time, the signal from a certain tissue type can be suppressed. These are called *inversion recovery pulse sequences*. Details on the purpose of different pulse sequences and different acquisition schemes are out of the scope of this thesis.

Secondly, MR can make use of **contrast agents** which can be paramagnetic, superparamagnetic or ferromagnetic substances. Contrast agents are applied intravenously and follow blood flow patterns. In particular, Gadolinium-based (Gd) contrast agents are used most often. They decrease the T_1 and T_2 time constants of surrounding hydro-

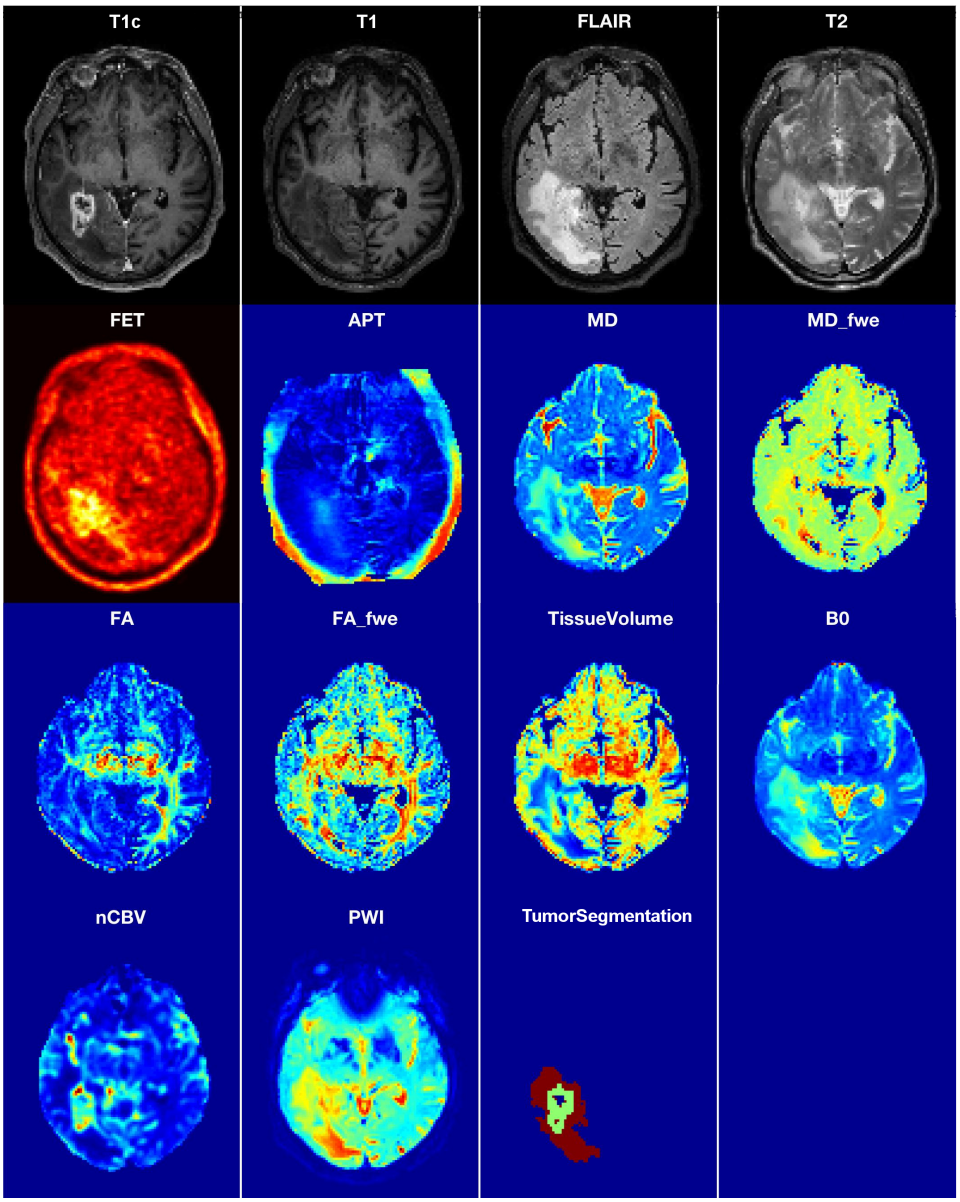


Figure 2.10: A set of image modalities acquired for a glioblastoma patient. The four standard modalities are shown at the top row (T1c, T1, FLAIR and T2). Other modalities include FET (PET scan), APT MRI (amide proton transfer), diffusion-weighted MRI (MD: mean diffusivity, FA: fractional anisotropy), B0 (T2-weighted images), functional MRI (CBV: cerebral blood volume) and perfusion-weighted MRI (PWI). The tumor segmentation contains edema (●) and active tumor (●).

gen protons, and thus appear either hypo or hyper-intense depending on the imaging modality.

Lastly, non-anatomical MRI images such as **diffusion and perfusion MRI** are often acquired by neuroradiologists in clinical practise [28]. Diffusion MRI is based on the diffusion of water molecules, which can be visualized with adapted pulse sequences. Common examples of diffusion MRI are ADC (appearant diffusion coefficient) images and diffusion tensor imaging (DTI). Typical examples of DTI include fractional anisotropy (FA) images and white matter tractography in the brain. Perfusion MRI can visualize cerebral blood flow and volume as well as vasculature. Figure 2.10 visualizes an extended set of image modalities acquired for a brain tumor patient, including anatomical MRI images as well as diffusion and perfusion MRI modalities.

2.4 MRI Protocol for brain tumor patients: T1, T1c, T2 and FLAIR

In a clinical MR study, several modalities are usually acquired. For brain tumor patients, it is common in clinical practice to acquire the following MR modalities: 1) a T1 image, 2) a T1 image with contrast agent (T1c), 3) a fluid-attenuated inversion recovery (FLAIR) image and 4) a T2 image. These four modalities (T1, T1c, T2 and FLAIR) are a world-wide accepted imaging protocol for brain tumor MRI acquisition, and are especially common in the field of brain tumor segmentation [2]. It should be noted that in some institutions another set of MR modalities might be acquired and further studied for brain tumor patients, for example in the context of medical research projects or to gain further insight in tumor metabolism or diffusion properties (cf. Figure 2.10). The datasets adopted in the projects presented in Chapters 4-7 consist in all of the four aforementioned MR modalities (T1, T1c, T2 and FLAIR), or a subset thereof (T1, T1c and FLAIR). These MR modalities are visualized in Figure 2.11 across several axial slices acquired for a glioblastoma patient. In the next paragraphs, each of the standard MR modalities are further explained and the appearance and distinguishability of healthy brain tissues (WM, GM and CSF) as well as tumor structures (active tumor, necrotic tumor and edema) is discussed. The appearance of each of these tissues in each of the four standard modalities can also be observed in parallel in Figure 2.11, which, in addition to the four standard modalities, also illustrates the tumor segmentations and the WM, GM and CSF maps.

T1. In T1 images, the images mostly reflect differences in spin-lattice relaxation, using a short TR and a short TE. The magnetization of the longitudinal and transverse magnetization vectors under a short TR and a short TE are shown in Figure 2.12a. It can be observed that the transverse magnetization vector for fat is much higher than it is for CSF after a short TR and a short TE. As a consequence, WM (very fatty tissue) appears brightest, GM appears grayish and CSF appears dark in T1 images (cf. Figure 2.11). The tumor structures appear rather darkish on T1 images. They are

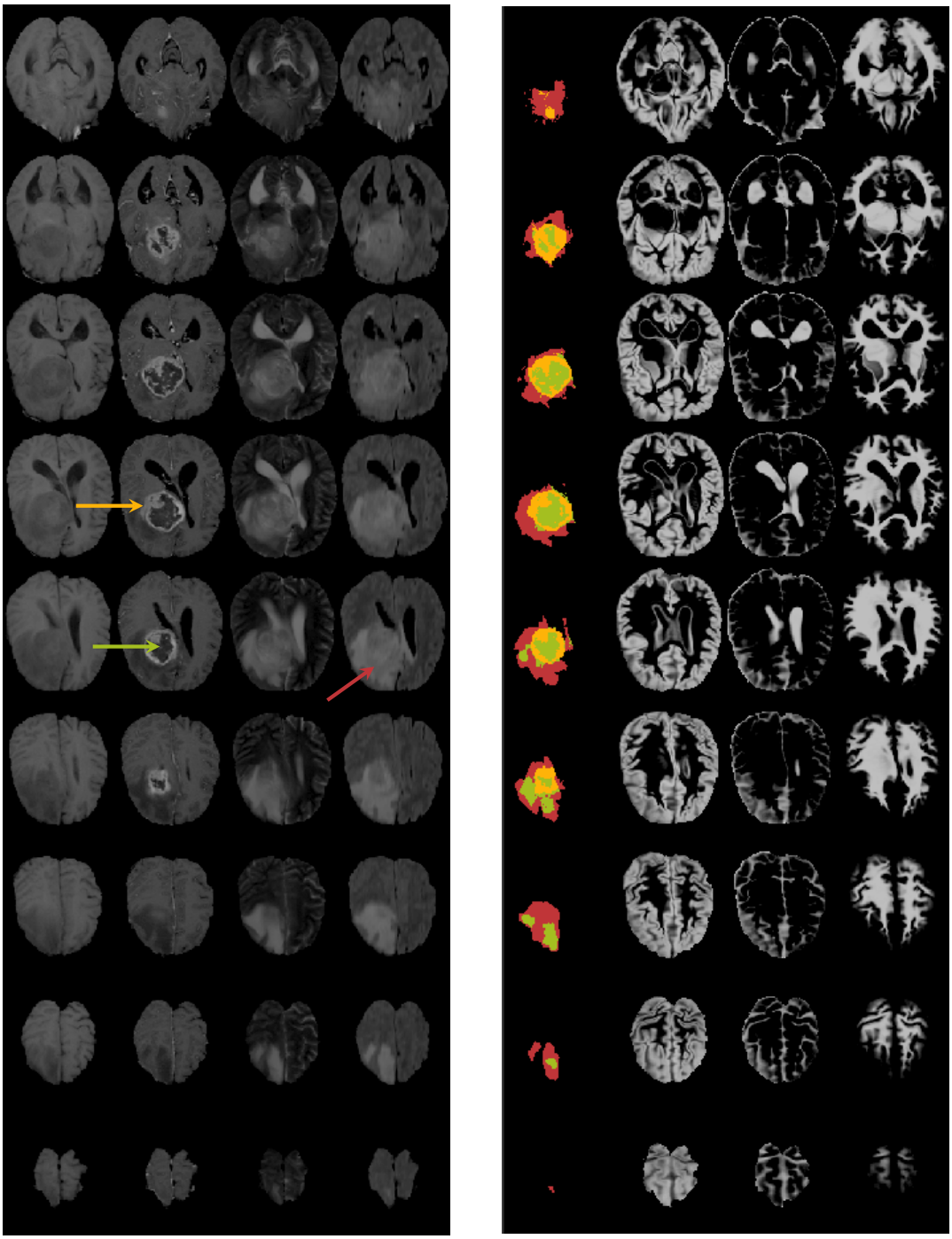
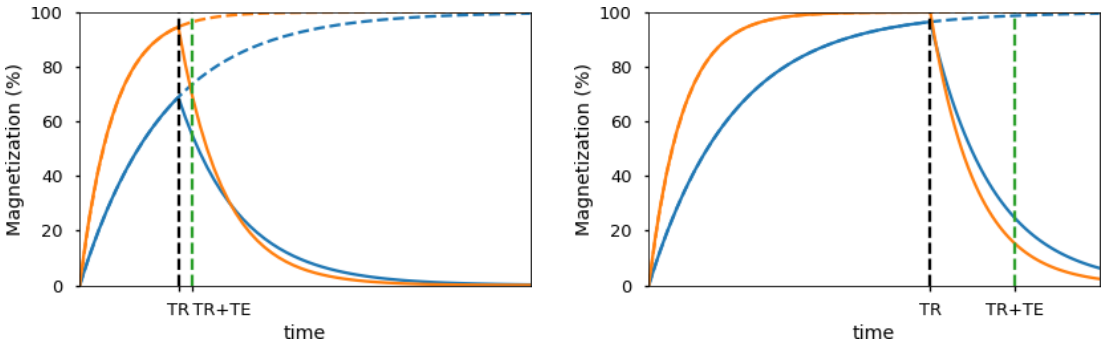


Figure 2.11: Left: axial slices of four standard MR modalities (T1, T1c, T2 and FLAIR) acquired for a glioblastoma patient from BRATS 2015 [2]. Right: Colored tumor segmentation containing active tumor (●), edema (●) and necrotic tumor core (●) and probabilistic tissue segmentations for GM, CSF and WM.



(a) T1-weighted image acquisition is associated with short TR and short TE. Magnetization of fat is stronger than liquid at acquisition time.

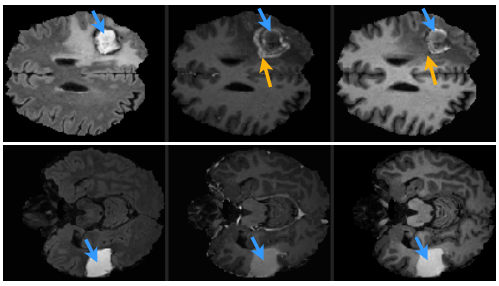
(b) T2-weighted image acquisition is associated with long TR and long TE. Magnetization of liquid is stronger than fat at acquisition time.

Figure 2.12: Approximation of relaxation processes for T1 and T2 weighted image acquisition for fatty tissues (–) and for CSF (–). T1 relaxation is shown up to TR (magnetization of the longitudinal component). At TR, a 90° RF pulse is applied. After TR, T2 relaxation is shown (magnetization of the transverse component). At TR+TE, the signal is measured to acquire the T1- or T2-weighted image.

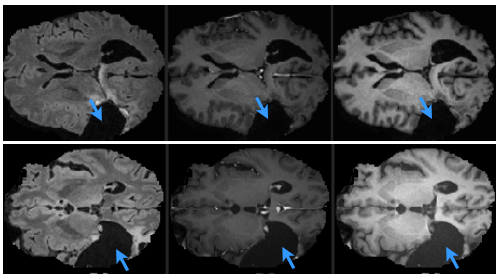
quite similar to the GM intensities, and are therefore not clearly distinguishable (cf. first column in the left image in Figure 2.11).

T1c. T1 contrast-enhanced images are very similar to T1 images, except that a contrast agent has been administered to the patient. The contrast agent follows the blood flow of the patient, and accumulates in the active part of the tumor. MRI Gadolinium-based contrast agents are known to decrease the T_1 and T_2 time constants of surrounding hydrogen protons [10]. Consequently, active tumor appears very bright (hyperintense) on T1c images (cf. yellow arrow in Figure 2.11). Active tumor is typically very clearly distinguishable in T1c images. Hemorrhages can also appear hyperintense in T1c, but will also appear hyperintense in T1, whereas active tumor will generally appear hypointense in T1 (cf. blue and yellow arrows in Figure 2.13a).

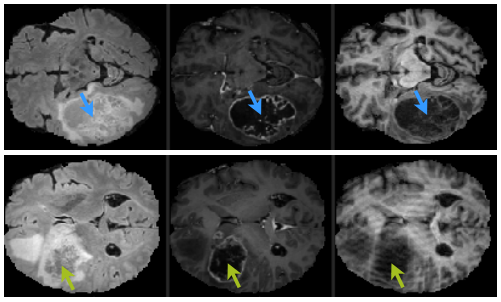
T2. In T2 images, the images mostly reflect differences in spin-spin relaxation, using a long TR and a long TE. The magnetization of the longitudinal and transverse magnetization vectors under a long TR and a long TE are shown in Figure 2.12b. It can now be observed that the transverse magnetization vector for CSF is much higher than it is for fat. As a consequence, CSF now appears brightest, GM still appears grayish and WM (very fatty tissue) now appears dark (cf. Figure 2.11). The tumor structures appear rather bright in T2 images (cf. third column in the left image in Figure 2.11). Sometimes the spherical tumor core, containing active and necrotic tumor, is clearly distinguishable from the surrounding edema.



(a) internal bleedings (hemorrhages) in FLAIR, T1c and T1).



(b) Resection cavities in FLAIR, T1c and T1.



(c) Necrotic tumor cores in FLAIR, T1c and T1.

Figure 2.13: Particular brain tumor structures in FLAIR, T1 and T1c (from left to right), for cases taken from the dataset reported in [4]. (a) internal bleedings (hemorrhages) (blue arrows), appearing hyperintense in T1c and T1; note that active tumor (yellow arrow) is only hyperintense in T1c and must be hypointense in T1, (b) resection cavities (blue arrows) and (c) necrotic tumor cores, which can appear hypointense in FLAIR (green arrows) or not (blue arrows).

FLAIR. Fluid-attenuated inversion recovery (FLAIR) images are similar to T2 images, but suppress the CSF signal. As explained in Sect. 2.3.5, inversion recovery pulse sequences can suppress the signal from a certain tissue type by combining an inversion pulse and a tissue-dependent inversion time. In FLAIR, the inversion time of fluid is adopted, hence suppressing the CSF signal, which now appears dark (cf. Figure 2.11). In FLAIR the tumor structures appear bright and are very clearly distinguishable from the healthy tissues (cf. last column in the left image in Figure 2.11). Necrotic tumor core (cf. green and blue arrows in Figure 2.13c) and resection cavities (cf. blue arrows in Figure 2.13b) can appear hypointense in FLAIR.

2.5 Conclusion

Gliomas, and glioblastomas in particular, are the most common malignant primary brain tumors. They are often diagnosed by means of MRI scans. In MRI scans, one can often discriminate between different tumor substructures. In this thesis, the following

tumor substructures will be considered in the technical projects presented in Chapter 4-7: active tumor, non-active tumor, necrotic tumor and edema (according to [2]). The adopted MRI scans will consist of T1, T1c, T2 and FLAIR, or a subset thereof. They each enhance different tumor structures: active tumor appears hyperintense in T1c and hypointense in T1, edema appears hyperintense in FLAIR and T2 and necrotic tumor core appears hyperintense in T2 and hypointense in T1c and often also in FLAIR. In addition to MRI scans, a biopsy can be taken for diagnosis purposes, to identify the genetic profile, which has an important influence on the choice of treatment as well as disease outcome.

3 Technical Background

This chapter aims to provide a technical background for the projects presented in Chapters 4-7. Section 3.1 introduces a few important technical concepts, such as machine learning, computer vision and data science concepts, that are at the basis of this thesis. A general problem statement is concisely presented in Section 3.2. Section 3.3 elaborates on image preprocessing, focusing on how the images are prepared before they are subject to further analysis such as segmentation, classification or growth estimation. Section 3.4 gives a general overview of medical image segmentation techniques and elaborates a bit further on the techniques adopted in the projects described in Chapter 4, 5 and 7. Similarly, Section 3.5 provides a general overview of medical image classification, presenting common texture features and classifiers used for image classification, and focuses on the techniques used in Chapter 6.

3.1 Introduction: Background on Related Fields

Medical image analysis makes use of several fields related to data science, such as machine learning, optimization, computer vision, image processing, pattern recognition and statistical modeling. In Chapter 4-7 several paradigms, algorithms and techniques from these fields will be adopted. This section provides a short introduction as well as general concepts and definitions related to machine learning, computer vision and statistical modeling, including examples of the techniques adopted in this thesis.

3.1.1 Machine learning

Machine learning is a field of *artificial intelligence*, concerned with "getting computers to act without being explicitly programmed" (Andrew Ng, Stanford). Instead of relying on rules-based programming, computers automatically learn themselves, through data and observations, to correctly *generalize* to unseen data samples. Several machine learning algorithms exist, each with their own strengths. Most algorithms are *parametric models*, which means that the algorithm needs to learn its own parameters during a *training phase*, before it can process unseen samples in a testing phase. The *parameters* θ, often also referred to as *weights* $\mathbf{w}$, are commonly learned by minimizing some *error function* E: $\theta = \arg\max_\theta E(\theta)$. Hence, *optimization* is an important aspect of machine learning. Machine learning is often subdivided into unsupervised, supervised and semi-supervised machine learning, as well as reinforcement learning. In *supervised machine learning*, training data, containing samples in input space together with their output label(s), is used to learn a machine learning function f mapping input samples $\mathbf{x}$ to (an) output label(s) y: $y = f(\mathbf{x})$. In *classification*, the output labels

are discrete. In *regression*, they are continuous. In *unsupervised machine learning*, only input samples are available, without their corresponding output labels. The set of labels might be known but unavailable for the training set, for example if you know there are k tumor types, but you don't know the tumor types of your training set samples, or they might be unknown and something to be learned, for example when you try to identify patterns in your data. Clustering, density estimation, anomaly detection, generative modeling and dimensionality reduction are important pillars of unsupervised machine learning. This dissertation covers several supervised machine learning algorithms (random forests (RFs), k-nearest neighbours, support vector machines (SVMs), logistic regression, multi-layer perceptrons (MLPs) and conditional random fields (CRFs)) as well as a few unsupervised techniques (Bayesian networks, probabilistic graphical models, Expectation-Maximization (EM) clustering and deep noisy auto-encoders). In semi-supervised machine learning, the output labels are only known for a part of the training data. Although semi-supervised learning is a field that can be tremendously valuable for medical image analysis (because it is expensive to acquire large training datasets with accurate output labels), it isn't explicitly adopted in this dissertation. In reinforcement learning, algorithms aren't tweaked to optimize some output y, but instead they learn to take actions, based on current and previous states, which optimize some kind of reward. Although it has recently found applications in medical image analysis, especially since the rise of deep reinforcement learning frameworks [29], it will also not be further discussed in this dissertation.

3.1.2 Computer vision

Computer vision is another important theme throughout this thesis. According to [30], computer vision is a field in computer science and engineering, aiming (1) to build computational models of the human visual system and (2) to build autonomous systems that can (out)perform tasks of the human visual system, such as recognition, object detection, scene segmentation, object tracking along time, ... The authors of [31] give a technical overview of computer vision algorithms and applications, covering various topics that mostly relate to the automation of tasks related to the human visual system. For this thesis, the most relevant topics include 1) geometric image transformations and deformations (i.e. image registration), 2) image processing (local and morphological image filters, Markov random fields (MRFs)), 3) feature detection and matching (e.g. feature descriptors and edge detection), 4) image segmentation and 5) recognition (object detection). Other topics are related to robotics and augmented reality (image motion, object tracking, image rendering, shape and appearance modeling, ...), and will not be treated in this thesis.

Several computer vision topics are further discussed in the following sections. Techniques related to image processing appear in Section 3.3. In Section 3.4 and Section 3.5, image segmentation and classification techniques are discussed.

3.1.3 Statistical modeling using probability distributions

A statistical model consists of a set of probability distributions $\mathcal{P}$ calculated based on a set of samples in a sample space $\mathcal{S}$, attempting to generalize to a larger population. In short, according to [32], a statistical model is "a set of probability distributions on the sample space $\mathcal{S}$". Statistical models can for example be used to verify statistical hypotheses such as "are the results with technique B significantly better compared to using technique A?".

In this dissertation, statistical models are used to estimate population densities using a variety of probability distributions. Several parametric probability distributions are used throughout this thesis. The normal distribution (*Gaussian distribution*) is used extensively throughout this thesis: Gaussian functions are used in the pairwise potentials in the energy function of the conditional random field in Chapter 4, Gaussian filters are used for intensity image smoothing, multivariate Gaussians are used to model multi-modal intensity distributions in the Expectation-Maximization algorithm in Chapter 4 and Gaussian copulas are used to model the dependencies in a multivariate probability function in Chapter 7. Gaussian distributions are bell-shaped symmetrical distributions, widely used to represent the distribution of continuous variables. Their common use is related to the *central limit theorem*, stating that the sum of a set of independent and identically distributed (i.i.d) random variables with finite variance, has a distribution that becomes increasingly Gaussian as the number of samples increase [33]. As the number of samples becomes infinite, the distribution becomes normal. A univariate Gaussian distribution of a continuous variable x is parametrized by two scalars: a mean μ and a standard deviation σ (or a variance σ^2):

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp -\frac{(x-\mu)^2}{2\sigma^2} \quad . \tag{3.1}$$

A multivariate Gaussian distribution of a n-dimensional vector $\mathbf{x} \in \mathbb{R}^n$ is parametrized by a mean vector $\mu \in \mathbb{R}^n$ and a covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$:

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Sigma}|}} \exp -\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) \quad . \tag{3.2}$$

If the variables in $\mathbf{x}$ are statistically independent, the covariance matrix is diagonal. A more detailed discussion of the Gaussian distributions and its peculiarities, the interested reader is referred to [34] (Section 2.3).

Other distributions adopted in this thesis include the *Gibbs or Boltzmann distribution* used in the context of a conditional random field (CRF). The distribution is formulated as a probability distributions for a certain state x out of a set of N possible states. Given state x with energy E_x, the probability distribution for state x is then given as:

$$p(x) = \frac{1}{Z} \exp -\frac{E_x}{\alpha} \quad \text{with} \quad Z = \sum_{j=0}^{N} \exp -\frac{E_j}{\alpha} \quad , \tag{3.3}$$

with Z the partition function, which serves as a normalization factor, and α a constant of the distribution ($\alpha = kT$ – with k the Bolzmann constant and T the temperature

– in the context of statistical mechanics [35]). The partition function is usually very expensive to calculate, as it requires a summation over all possible states.

Lastly, the *Gumbel distribution* is used for several purposes in this thesis. In Chapter 5, Gumbel samples are used to perturb the energy function of the conditional random field in order to access uncertainty. In Chapter 7, The Gumbel distribution is used to model the *asymmetrical* marginal intensity distributions in each channel for each class. The Gumbel distribution is an asymmetrical distribution, often used to model the distribution of maximum values, such as monthly maximum values of daily temperature or rainfall [36]. It is parametrized by a location parameter, μ, and a scale parameter, β:

$$p(x) = \frac{1}{\beta} \exp -\Big(\frac{x-\mu}{\beta} + \exp -\big(\frac{x-\mu}{\beta} \big) \Big) \ . \tag{3.4}$$

In this form, the Gumbel distribution is positively skewed (right-skewed). The amount of skewness is constant – it doesn't depend on the scale or location parameter (unlike the asymmetrical Gamma distribution). The distribution of a negatively skewed (left-skewed) Gumbel distribution can be written as:

$$p(x) = \frac{1}{\beta} \exp \Big(\frac{x-\mu}{\beta} - \exp \frac{x-\mu}{\beta} \Big) \ . \tag{3.5}$$

3.2 Problem Statement

This section aims to give a short overview of the types of data studied in this thesis and the type of tasks associated with them.

Medical Image Data. In this thesis, the most prominent type of medical image data are the T1, T1c, T2 and FLAIR MR images of brain tumor patients, as discussed in Section 2.4. In Chapter 4 and Chapter 5 these *multi-modal* MR sequences are available for several time points. This type of multi-temporal image data sets are referred to as *longitudinal datasets*. Figure 3.1 shows a multi-modal longitudinal dataset for one patient. In such a dataset, a 3-dimensional volume is present for each time point and each modality type, leading to 5-dimensional image datasets. In Chapter 6 and Chapter 7, the datasets are treated for individual time points, leading to (4-dimensional) multi-modal (unitemporal) datasets. The image datasets adopted in these thesis are either publicly available (e.g. the BRATS dataset [2], the REMBRANDT dataset [5] or the TCIA dataset [37]) and others have been shared by the "Klinikum Rechts der Isar – department of neuroradiology" (cf. works published in [38], [39] and [12]). The image data itself can come in different formats and can be partly preprocessed or can come as raw image data directly from an MR scanner. The image data in Figure 3.1 is partly preprocessed: the images are co-registered (they are aligned) but haven't been skull-stripped. Section 3.3 elaborates on image data preprocessing steps.

Genetic Data. In Chapter 6 (cf. [11] and [12]) genetic data is used in addition to the image data. The field of studies involving both imaging and genetic data is often

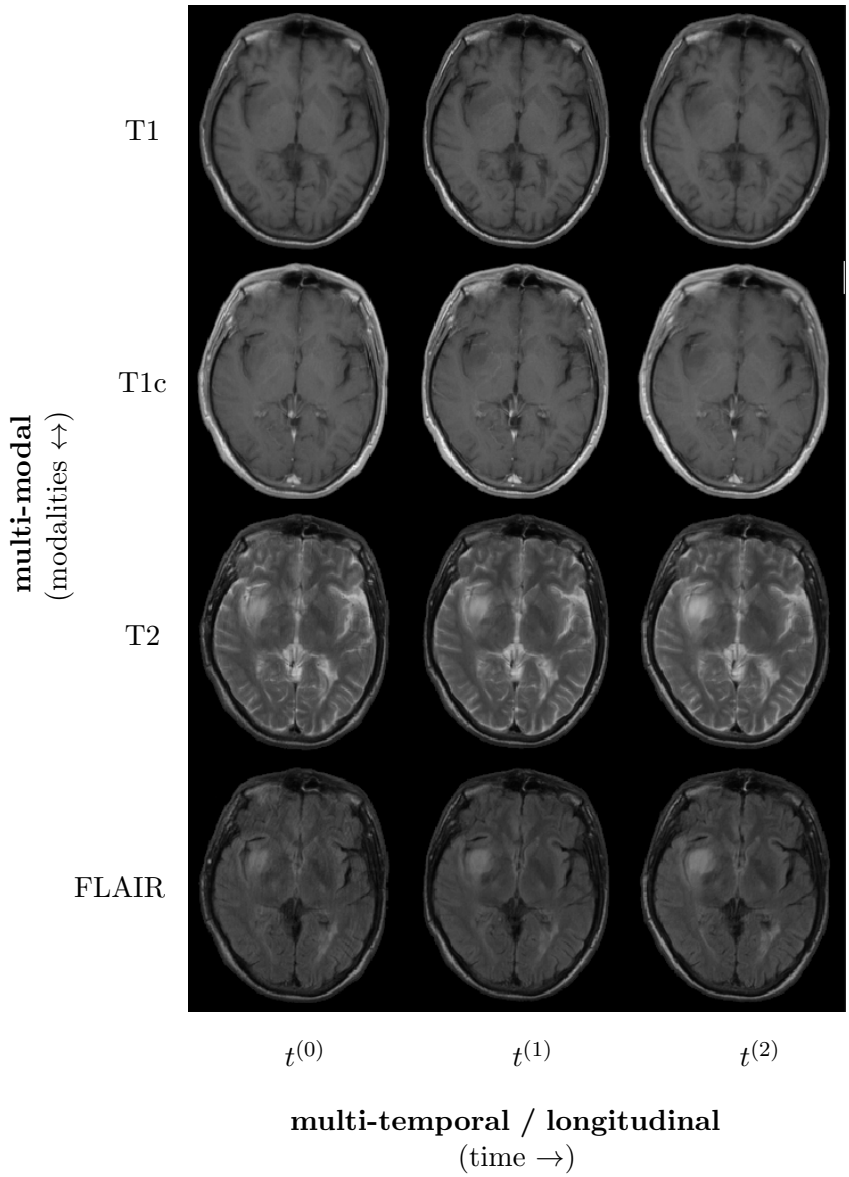


Figure 3.1: Patient-specific multi-modal multi-temporal MRI image data. Axial slices are shown for each 3-dimensional image volume, available for several modalities (T1, T1c, T2 and FLAIR) along several time points $t^{(i)}$.

Table 3.1: Derivation of tumor classes based on genetic parameters. Tumor classes are adopted for the multi-modal image classification tasks in [11] and [12].

Tumor class	IDH mutant	1p/19q co-deletion
0: CIMP Negative (IDH wild type)	×	×
1: CIMP Non-Codel	√	×
2: CIMP Codel	√	√

referred to as *radiogenomics*. The genetic data consists of genetic parameters of the tumor's genetic profile, identified by means of a biopsy. Recent studies suggest that two genetic parameters – the IDH mutation status and the presence of 1p/19q co-deletion – play a key-role in disease outcome [17], [18]. Based on these two genetic parameters, medical doctors at the "Klinikum Rechts der Isar – department of neuroradiology", categorized the genetic tumor profile of the patients into three classes: CIMP Codel, CIMP Non-Codel and CIMP Neg (cf. Table 3.1).

Image Analysis Tasks. Based on the multi-modal, multi-temporal image datasets and the genetic parameters, several tasks are investigated. In this thesis, brain tumor segmentation, brain tumor growth analysis and brain tumor classification are studied. Each of these tasks can be described as a set of inputs and outputs. In brain tumor segmentation, a 3-dimensional label image (also referred to as a segmentation image) is calculated based on a set of multi-modal images. Tumor growth analysis is studied in this thesis as an extension of the image segmentation pipeline. For the study of tumor growth analysis, multi-modal multi-temporal image datasets are available and a 3-dimensional label image is generated for every time point. Furthermore, tumor volumetry can be assessed along time and the probability of growth is evaluated between every two consecutive time points, such that the time point of disease progression can be identified. Tumor segmentation topics and implementations are presented in Chapters 4,5 and 7. The last image analysis task involves brain tumor classification. During image classification, a genetic tumor class is predicted based on a set of multi-modal images. A proposed pipeline for brain tumor image classification is presented in Chapter 6.

In this chapter, a background on image segmentation techniques is presented in Section 3.4 and a background on image classification techniques is presented in Section 3.5.

3.3 Preprocessing

In this thesis, *preprocessing* is referred to any method or technique applied to the image data before further analysis, such as image segmentation, image classification or growth estimation, takes place. Preprocessing is important, because methods implemented for further analysis make certain assumptions about the input image data. It might for example be necessary that images have a certain image dimension or it might be

required that the image dimensions for the images within a multi-modal image dataset are the same. Preprocessing typically consists of 1) image format conversion, 2) image registration or image resampling, 3) biasfield correction, 4) brain extraction or skull stripping and finally 5) intensity normalization. Each of these steps are elaborated in the next paragraphs.

It is important to note that this section attempts to give an overview of what a brain MR preprocessing pipeline might look like. Depending on the application, some of these preprocessing steps might be obsolete. Furthermore, publicly available datasets are typically already (partially) preprocessed. In such cases, a good approach is to check to what extend the images are preprocessed and what steps should still be conducted before proceeding with further image analysis tasks.

3.3.1 Image Format Conversion

The first important property of newly acquired image data, is its imaging format. Imaging data consists of raw image data (i.e. gray values) as well as some image metadata (patient name, patient id, study time, modality type, image dimensions, slice dimensions, raw positional coordinates, ...). Because of privacy protection, it is extremely important that medical staff (pseudo-)anonymizes patient data before it is transferred to a portable device that can leave the hospital territory. For this purpose, several *(pseudo-)anonymization* functions can be used.

Reasons for image format conversion include: 1) transparency and portability, 2) compatibility with image visualization tools, programming language or image analysis libraries and 3) memory efficiency and read/write speed.

Transparency and portability. Raw image data from an MR scanner typically comes in a *Digital Imaging and Communications in Medicine* (DICOM) format. This format consists of DICOM objects, for example assembled in XML files, containing several attributes, such as patient ID, imaging modality, date, time as well as references to pixel data frames. The pixel data frames are typically stored in separate files for each 2D slice within the 3D volumes [40]. The DICOM format is the standard imaging format used for radiological images in hospitals worldwide, and is compatible with the *Picture Archiving and Communication System* (PACS) systems used in radiology departments. Because of its complexity (DICOM objects consist of several files), the DICOM format can be quite tedious to work with outside PACS systems. NIfTY [1] (*.nii*, *.nii.gz*) and MHA[2] (*.mha* or *.mhd* & *.raw*) are popular imaging formats that are more transparent: they contain imaging metadata and 3D imaging data in one file.

Visualization and Imaging Toolkits. The image format should be compatible with an image visualization tool that is easy accessible and quick in use. Some typical 3D visualization tools include ITKsnap [41] and 3D Slicer [42]. They both support a

[1] Neuroimaging Informatics Technology Initiative: `https://nifti.nimh.nih.gov/`
[2] ITK wiki: `https://itk.org/Wiki/MetaIO/Documentation`

different set of image formats. Visualization of DICOM datasets can be cumbersome, because the slices can get mixed up, leading to awkward 3D representations. A common problem can be that the DICOM format of your image data might not be compatible with the DICOM formats supported by the image visualization tool. Popular computer vision packages for image processing include ITK, OpenCV and NiBabel, which all allow to read and write NIfTY or MHA images, and to a less transparent extent DICOM datasets.

Memory efficiency. Some image formats are more memory-efficient than others. If entire datasets are converted towards heavy image formats, they can quickly take up large extends of data volume. As an example, consider a NIfTY image with *image dimensions* $155 \times 240 \times 240$, *pixel data type* signed 16-bit integer (i.e. 2 Bytes per pixel) and intensity values in $[0, 1646]$. The image data alone sums up to $(155 \times 240 \times 240 \times 2\,\mathrm{B} = 17856000$ bytes$) \approx 17,86\,\mathrm{MB}$. In addition, a bit of extra memory is needed for the image metadata (pixel data type, pixel dimensions, image dimensions, string description, ...). Table 3.2 depicts the memory used to store this image when it is saved using different image formats and different pixel data types. When this image is saved to a *.nii* image, it occupies 17,9 MB (17.856.352 bytes). However, when saved to the NIfTY compressed format, *.nii.gz*, the image contains only 1,7 MB (1.670.612 bytes). That is 10 times less memory needed to store the same imaging data! This lower memory space of compressed NIfTY formats, comes of course at a cost of slower reading and writing access.

Furthermore, for most image formats (including NIfTY and MHA files), casting the pixel data towards more space-efficient integer or floating point data types can considerably reduce storage use. Caution should of course be taken that the image data remains unaltered under the casting operation. As an example, a label image with only 5 segmentation labels is best saved using signed 8-bit integers (covering values in $[-128, 127]$) instead of signed 16-bit integers (covering values in $[-32768, 32767]$). The effect of saving an image with different pixel data types is also listed in Table 3.2.

Finally, care must always be taken that the image format is compatible with the image data. For example, image data with pixel data type signed or unsigned 64-bit integer ('uint64' or 'int64' in NumPy) are not compatible with the NIfTY-1 format.

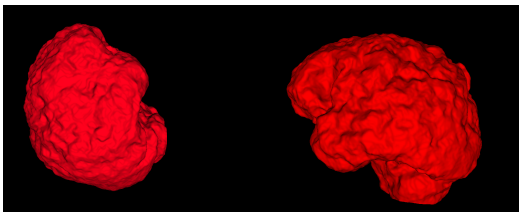
3.3.2 Image Registration

In order to extract information out of a set of several images taken from the same object, the images first need to be aligned to each other. This is accomplished during image **co-registration**. Image co-registration is the process during which the images are resampled to the same image space, resulting in a set of aligned images. Figure 3.2 illustrates two pairs of identical 3-dimensional brains. The pair to the left is not properly co-registered: the brain in the right image is translated, rotated and scaled compared to the brain in the left image. The pair to the right shows the same two brains, co-registered to an atlas reference space.

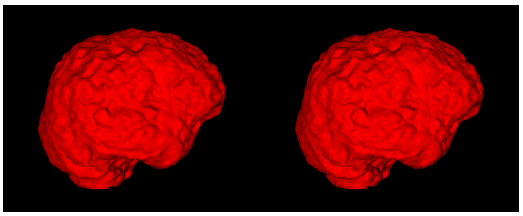
Table 3.2: Memory efficiency for different imaging formats, calculated for a 3-dimensional gray-scale image with image dimensions $155 \times 240 \times 240\times$, pixel values in between $[0, 1646]$ and some small amount of imaging metadata. Image conversions are performed using SimpleITK and NumPy in Python. Compressed NIfTY images seem to be much smaller and casting towards less heavy pixel data types is much more memory efficient.

Image format	pixel data type	size (MB)
NIfTY-1 (.nii)	unsigned integer 16 ('uint16')	17.9
	unsigned integer 32 ('uint32')	35.7
	64-bit double-precision floating point ('float')	71.4
Compressed NIfTY-1 (.nii.gz)	unsigned integer 16 ('uint16')	1.7
	unsigned integer 32 ('uint32')	1.9
	64-bit double-precision floating point ('float')	2.1

Registration types. During registration, a *floating image* is registered to a *reference image*. In a *rigid registration*, the floating image can only be translated or rotated to better align the reference image. In an *affine registration*, the floating image can be scaled and sheared in addition. Rigid and affine transformations are both linear transformations defined by a transformation matrix. A 3-dimensional affine transformation for example, is defined by a transformation matrix consisting of 12 degrees of freedom: $T_{\text{aff}} \in \mathbb{R}^{4\times3}$. *Non-rigid registrations*, also called deformable or elastic registrations, calculate non-linear transformations. In contrast to rigid or affine registration, they are able to capture local deformations. Non-rigid transformations are defined by a deformation field, which implicitly encode a displacement for every voxel in the voxel grid. They are typically more expensive to calculate in comparison to rigid or affine registrations.



Two brains in different image spaces.



After co-registration to reference space.

Figure 3.2: Illustration of co-registration. To the left, two non-registered brain masks are visualized: the images are rotated, translated and scaled compared to each other. To the right, the same two brain masks have been co-registered to a reference space.

Multi-modal image co-registration. To co-register the T1, T1c, T2 and FLAIR modalities, affine registrations are most appropriate. Performing non-rigid registrations on input images as part of a preprocessing pipeline is highly discouraged: they allow for local deformations, which can alter the tumor shape and the brain tissue structures in the input images.

During multi-modal image co-registration, the four modalities can be registered among themselves (for example by registering all modalities to the T1 image), or they can all be registered to a *reference space* (for example the BRATS reference space [2] or the SRI24 atlas space [3]). In case the images are to be passed to an algorithm that has been trained on a training set, it is generally a good idea to register all modalities to the reference space adopted in the training set.

Figure 3.3 illustrates an example of inter-modality co-registration, where T1c, FLAIR and T2 are registered towards the T1 image modality. The modalities are usually acquired one after the other with the same scanner, causing the inter-modality transformations to be rather small. Figure 3.3 also illustrates why careful visual inspection is usually required in order to identify whether the images have already been co-registered.

Implementation of affine registration. Affine registrations can be calculated using the `reg_aladin` function from the NiftyReg library [3]. This implementation makes use of a multi-scale iterative scheme, in which each iteration consists of [43]:

1. a **block-matching step**, during which corresponding blocks in the floating image and in the reference image are identified using a correlation coefficient as a similarity metric, and

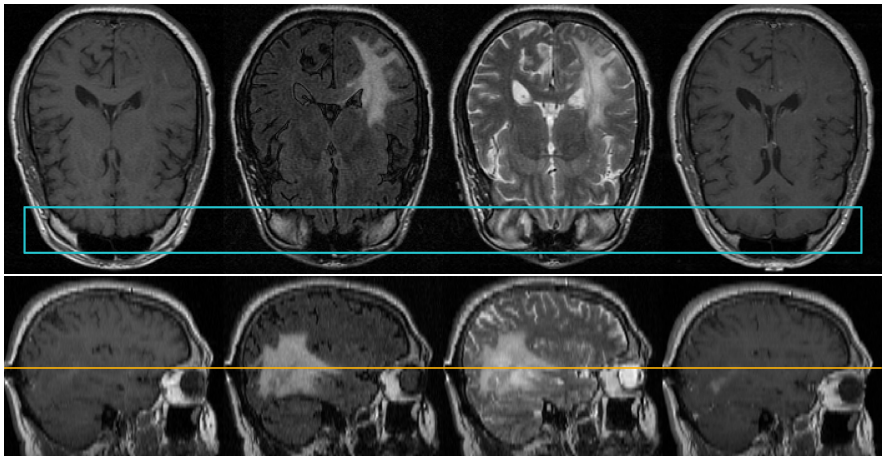
2. an **estimation of the rigid or affine transformation**, during which the distances between the matched blocks in the reference image and the transformed image are minimized based on the Least Trimmed Squares criterion.

The algorithm starts at a coarse scale (subsampling the original images) and refines the scale through the iterations [44]. Another popular alternative for affine registrations is the `antsRegistration` function from the Advanced Normalization Tools (ANTS) library [45].

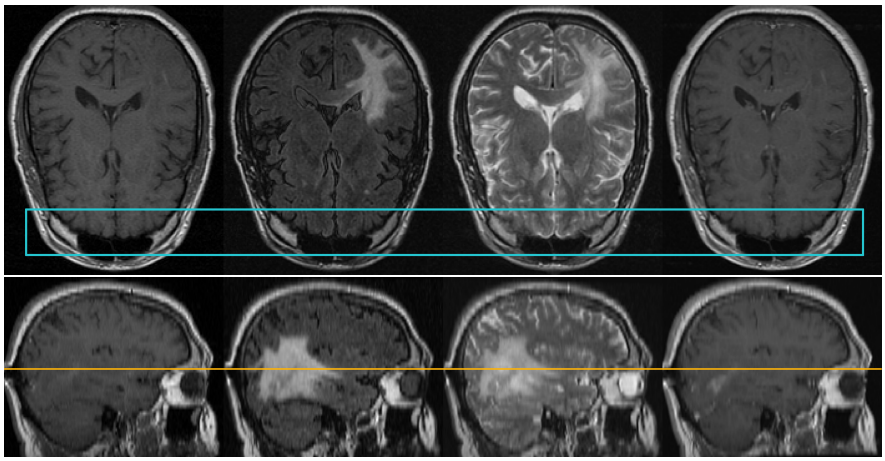
Adoption of non-rigid registrations. Non-rigid registrations allow for local deformations and are often used to model growth, motion or tissue deformations. In this thesis, they are used to align population atlases with a patient brain, such that the population atlases can better resemble patient-specific brain structures (cf. Section 3.3.6).

Image resampling. Image resampling is the process during which the image dimensions of an image are altered. In contrast to image registration, the field-of-view of the image is not altered and the objects in the image do not undergo any transformation of any kind. Image resampling simply consist of either upsampling or subsampling the

[3] `https://sourceforge.net/projects/niftyreg/`



Before co-registration (T1, FLAIR, T2, T1c).



After co-registration (T1, FLAIR, T2, T1c).

Figure 3.3: Illustration of image registration: modalities are co-registered towards the T1 image. Difference in alignment are apparent at the eye level (indicated in blue and yellow).

image grid. This is typically done by linear interpolation, nearest-neighbor interpolation or *b*-spline interpolation, where the order of the *b*-splines can be chosen by the user (cf. the `scipy.ndimage.zoom` function as part of the Scipy package in Python or the `itk::ResampleImageFilter` of the ITK toolkit). For brain MR scans, linear interpolation is usually sufficient. When higher-order splines are used, the user should be aware the range of pixel intensities can be altered and that negative pixel values might suddenly pop up. Image resampling might be needed if the preprocessed data is to be *isotropic*, meaning that the pixel spacing should be identical in every direction (along the x-, y- and z-axes).

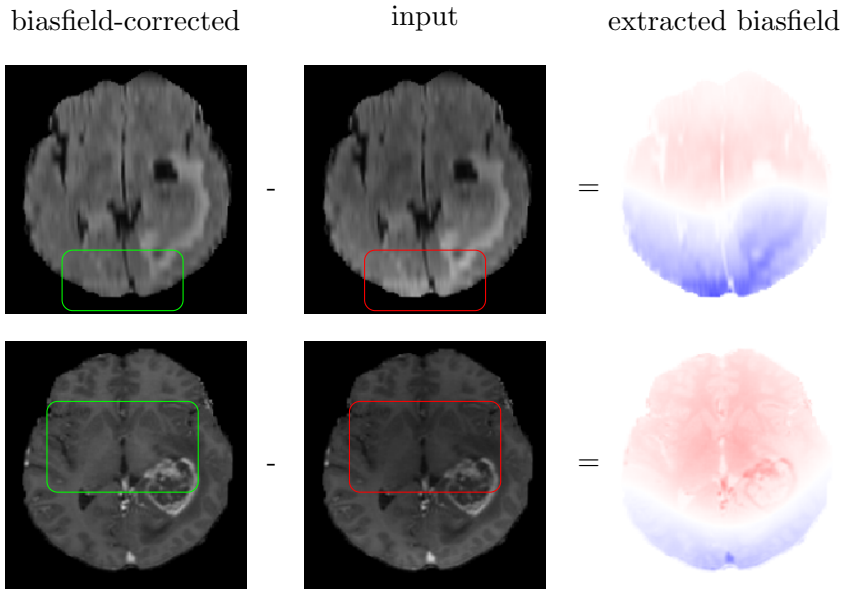


Figure 3.4: Biasfield correction of a flair (top row) and a t1c image (bottom row), calculated using ANTs N4 bias field correction. Additive extracted biasfields are shown to the right. Note how the red encircled regions in the input images have improved in the biasfield-corrected images.

3.3.3 Biasfield Correction

Due to the magnetic field inhomogeneities of MRI scanners, MR images can suffer from biasfield perturbations. A biasfield is a low-frequency noise signal, which can perturb the high-frequency contents in the MR image, making edges or contours less pronounced [46]. For a long time the nonparametric nonuniform intensity normalization (N3) algorithm has been the most popular library for MR biasfield correction, both because it works well for a large range of applications and because of its public availability [47]. The **N3 biasfield correction** algorithm follows an iterative scheme in which the biasfield is modeled based on a B-spline least-squares fitting. The authors in [48] developed a **N4 biasfield correction**, publicly available for example via the `itk:N4BiasFieldCorrectionImageFilter` in ITK 5.0.0, as an improvement over the N3 biasfield correction. Specifically, the B-spline approximation strategy is improved (by estimating residual bias fields in each iteration instead of the total bias field), the execution times are faster (parallelization of the B-spline approximation algorithm) and a multiresolution approximation strategy is employed [48]. Figure 3.4 illustrates the N4 biasfield correction of a FLAIR and a T1c image, together with their extracted additive biasfields.

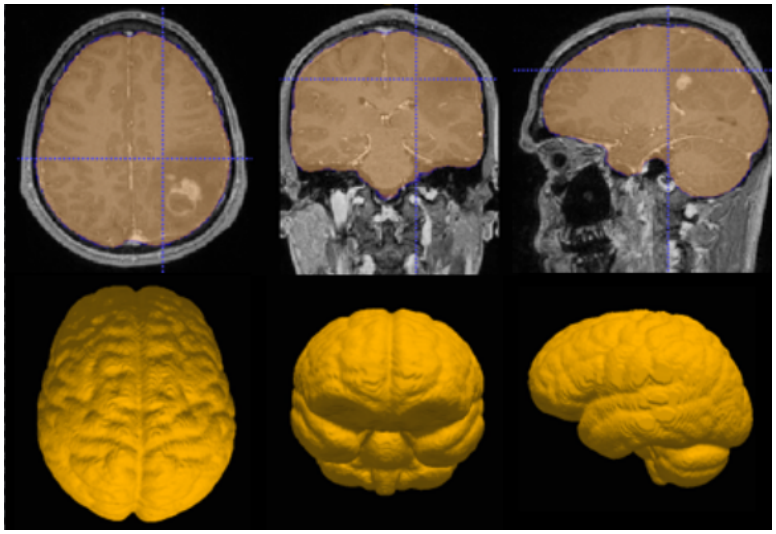


Figure 3.5: Illustration of brain extraction for a brain tumor patient from the REMBRANDT dataset [5]. The binary brain extraction mask is visualized in yellow and delineates the soft brain tissues, effectively extracting irrelevant parts such as bone (skull), the eyes and the face. By courtesy of Jana Lipkova for visualization.

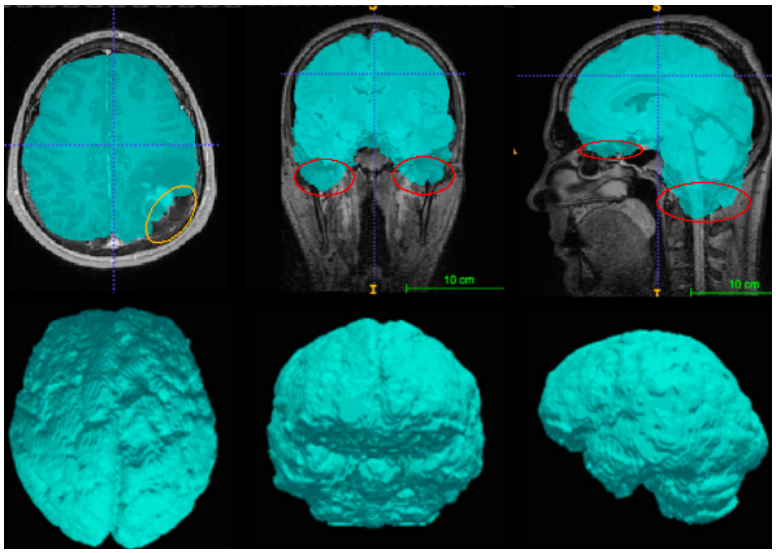


Figure 3.6: Illustration of brain extraction for a brain tumor patient from the REMBRANDT dataset [5], calculated using BET [6]. Erroneous delineations are indicated in yellow (due to the presence of a tumor) and in red (casual oversegmentations). By courtesy of Jana Lipkova for visualization.

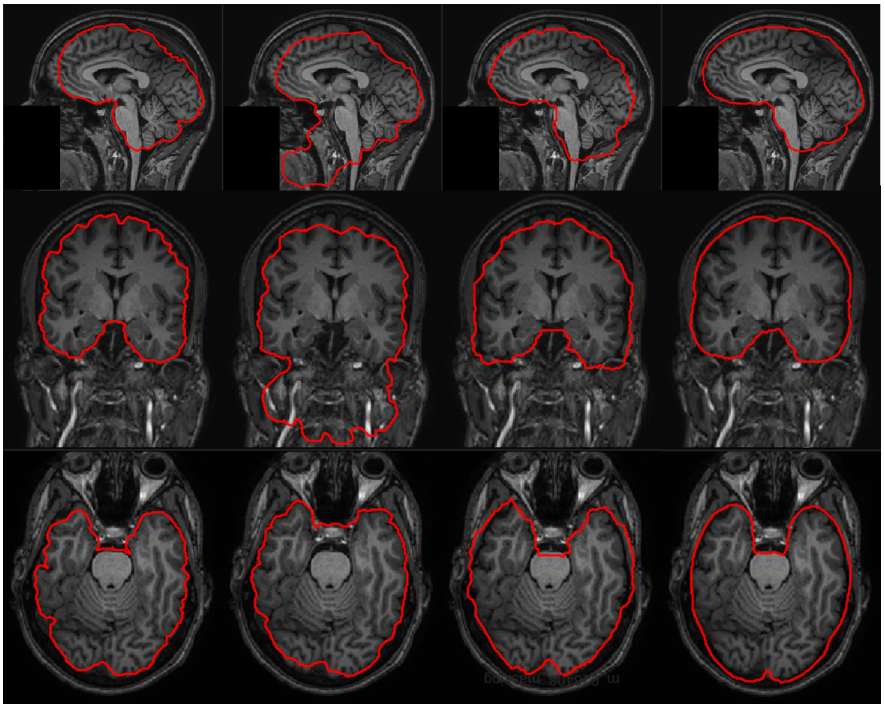


Figure 3.7: Illustration of brain extraction calculated with several methods. From left to right: BET [6] (`reduce_bias = True`), BET [6] (`reduce_bias = False`), ITK SkullStrip [7], ground truth.

3.3.4 Brain Extraction

Brain extraction or skull stripping involves the removal of all structures and tissues in the MR scans that surround the brain tissues GM, WM and CSF. These include the eyes, the face, the skin and the skull. Figure 3.5 illustrates a brain extraction mask for a patient with a brain tumor. There are two reasons brain extraction is usually done. Firstly, it ensures further patient anonymization, as the face is now explicitly removed from the image data. The DICOM pseudo-anonymization functions mentioned in Section 3.3.1 only remove patient names or patient ids. They do not change the raw image data. Secondly, it allows to simplify the tasks assessed in further analysis. For tumor segmentation or tumor type classification, the skull, eyes and face are namely irrelevant. The vast majority of brain images shown in this thesis have been skull-stripped (cf. Figure 3.4).

Several brain extraction tools are available: The Brain Extraction Tool (BET) [6], Robust Brain Extraction (ROBEX) [49] and the Insight Segmentation and Registration Toolkit (ITK) also provides a skull-stripping module `SkullStrip` containing a skull-stripping filter `itk::StripTsImageFilter` [7]. For images containing brain tumors, results acquired by these methods can be a bit off (cf. Figure 3.6). In such cases, visual inspection might be necessary and sometimes a manual adaption is required. Figure 3.7 visualizes brain extraction masks calculated with several publicly available methods.

3.3.5 Intensity Normalization

Intensity normalization plays an important role in the preprocessing of MRI images. In contrast to radiography or CT, the absolute voxel intensities in MR images do not reflect physical quantities and don't have a unity. As such, the voxel intensities in an MR image are only meaningful in relation to the intensity distribution of the MR image. The intensity distributions in MR images depend on the scanning protocol and various scanner-specific parameters (cf. Section 2.3.5). As a result, MR images taken from similar objects with different scanners or with different scanning parameters often exhibit substantial variations in intensity distributions.

Most image segmentation and classification methods make decisions based on voxel intensities and make assumptions on the image intensity distribution of the MR images, for example requiring the images to have a standardized intensity range [50]. As such, intensity normalization is a crucial part of the preprocessing pipeline. During intensity normalization, the intensity distributions in the different MR images are normalized, scaled or aligned, typically by means of gray level transformations or histogram-based normalizations [50], [51], [52]. A few common intensity normalization techniques are described in the next paragraphs. Figure 3.8 illustrates the histograms and image slices of the normalized images accordingly. Note that, for brain tumor analysis, only the voxels within the brain mask are relevant. It is therefore common to first apply brain tumor extraction (as described in Section 3.3.4) and only consider the voxels within the brain mask during intensity normalization.

Intensity rescaling. Rescaling is the most straight-forward intensity normalization technique. The intensities in image I are linearly scaled in order to cover a range $[a, b]$:

$$X' = (b-a)\frac{X - \min_X}{\max_X - \min_X} + a \ ,$$

where X and X' are the original and rescaled images resp. and $\max_X$ and $\min_X$ are the maximal and minimal intensities in the original image I.

By rescaling a set of images, it can be ensured that all intensities will be in a fixed and known range $[a, b]$. However, rescaling is very sensitive to outliers: the intensity distribution of a rescaled image will horizontally shift ans shrink as $\max_X$ increases or $\min_X$ decreases.

z-transformation. When the pixel intensities are z-transformed, they are linearly transformed such that the transformed intensities exhibit zero mean and unit standard deviation:

$$Z = \frac{X - \mu_X}{\sigma_X} \ ,$$

where X and Z are the original and z-transformed images resp. and μ_X and σ_X are the mean and standard deviation in the original image X.

By z-transforming a set of images, the intensity distributions across the different images are well aligned, they have all zero mean and unit standard deviation. However, the range of intensities is not fixed and varies from one image to another.

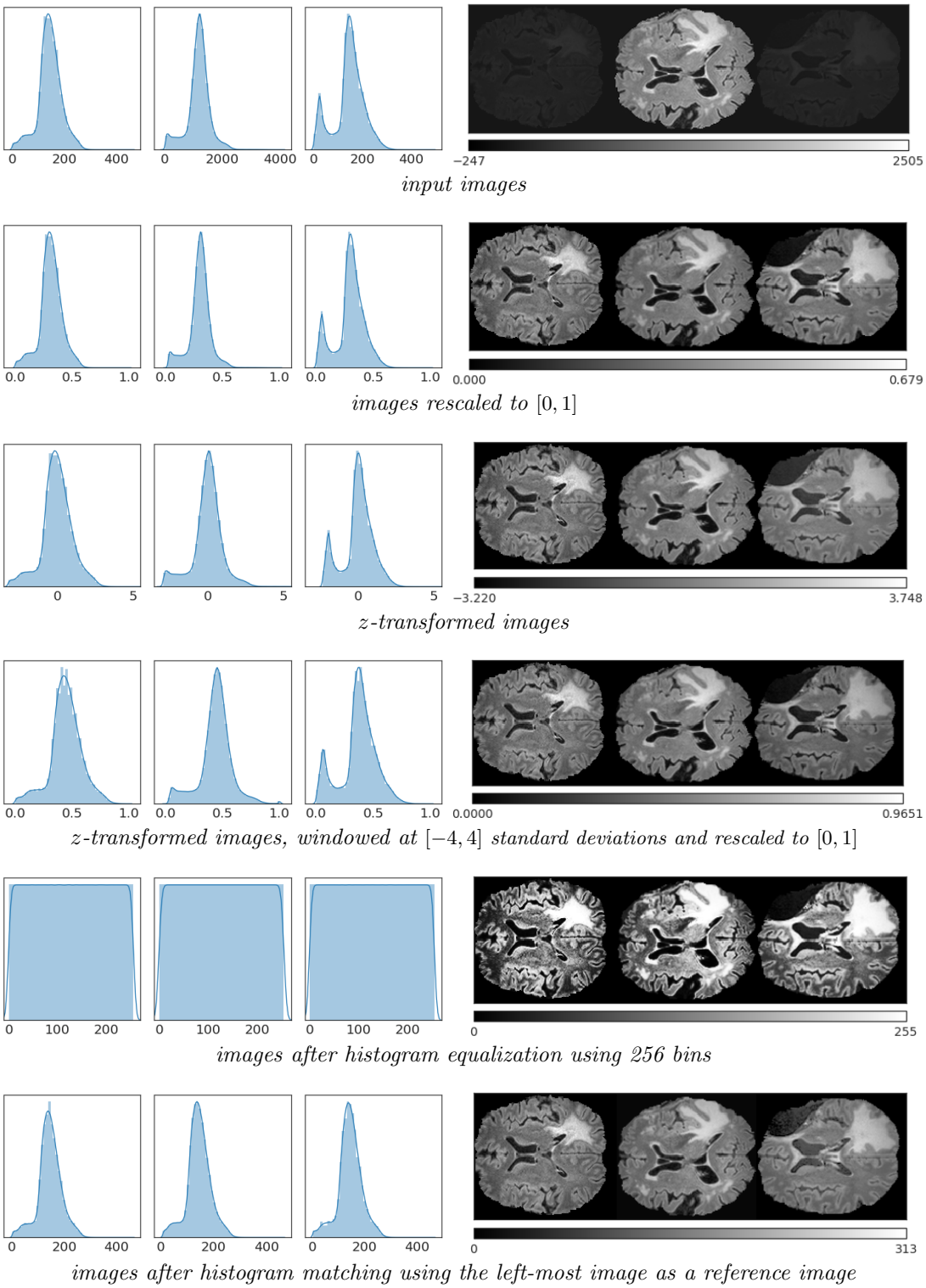


Figure 3.8: Illustration of different intensity normalizations. To the left, the histograms of the normalized images are shown. Only the voxels within the brain mask are considered in the histograms. To the right, axial slices of the normalized images are shown together with their colorbar mappings.

Windowing. During windowing, outliers are set to a lower and upper threshold, a and b, such that the intensities cover a range $[a, b]$:

$$x_i' = \begin{cases} a & \text{if} \quad x_i < a, \\ b & \text{if} \quad x_i > b, \\ x_i & \text{else.} \end{cases} ,$$

where x_i and x_i' are the original and windowed image intensities in voxel i.

By windowing a set of images, the intensity range will be fixed and known. It can be difficult however, to identify appropriate values for the lower and upper thresholds. A good technique, especially for visualization purposes, can be to first z-transform the images, then window the images at n standard deviations and then rescale the images in $[0, 1]$. Doing so, the intensity histograms are nicely spread over the intensity range and are not as much sensitive to outliers (cf. Figure 3.8).

Histogram equalization and histogram matching. During histogram equalization, the intensity histogram is quantized into n bins, and is flattened such that each bin is equally frequented. For histogram matching, a reference image is needed. During histogram matching, intensities are altered such that the resulting histogram is equal to the histogram of the reference image.

For brain MR images in particular, clever image normalization techniques have been proposed that rely on GM, WM or CSF segmentations and align the intensities accordingly [52]. A large variety of other image normalization techniques have been investigated, but fall out of the scope of this thesis [50].

3.3.6 Generating tissue segmentations using atlas registration

Some image analysis tasks require prior GM, WM and CSF probability maps for each patient. In this thesis, these prior tissue probability maps are used to initialize the models presented in Chapter 7. WM, GM and CSF segmentation methods (such as the one presented in [53]) also require prior GM, WM and CSF probability maps to initialize the model. These tissue priors are commonly generated by registering a population atlas to patient space via an affine transformation ([53], [8], [54]). In this thesis, we make use of affine and non-rigid registrations between patient and atlas space, in order to generate more reliable patient-specific tissue priors.

Registration flow. The affine and non-rigid registrations are calculated between a patient T1 image and an atlas T1 image, and the resulting registration transformations are applied on atlas GM, WM and CSF probability maps. All atlas images (T1 image as well as GM, WM and CSF probability maps) are taken from the SRI24 atlas [3]. The registration flow is schematically depicted in Figure 3.9 and goes as follows:

1. A T1 image in atlas space is affinely registered to patient space, generating a affinely registered T1 atlas image. During this affine registration, an affine trans-

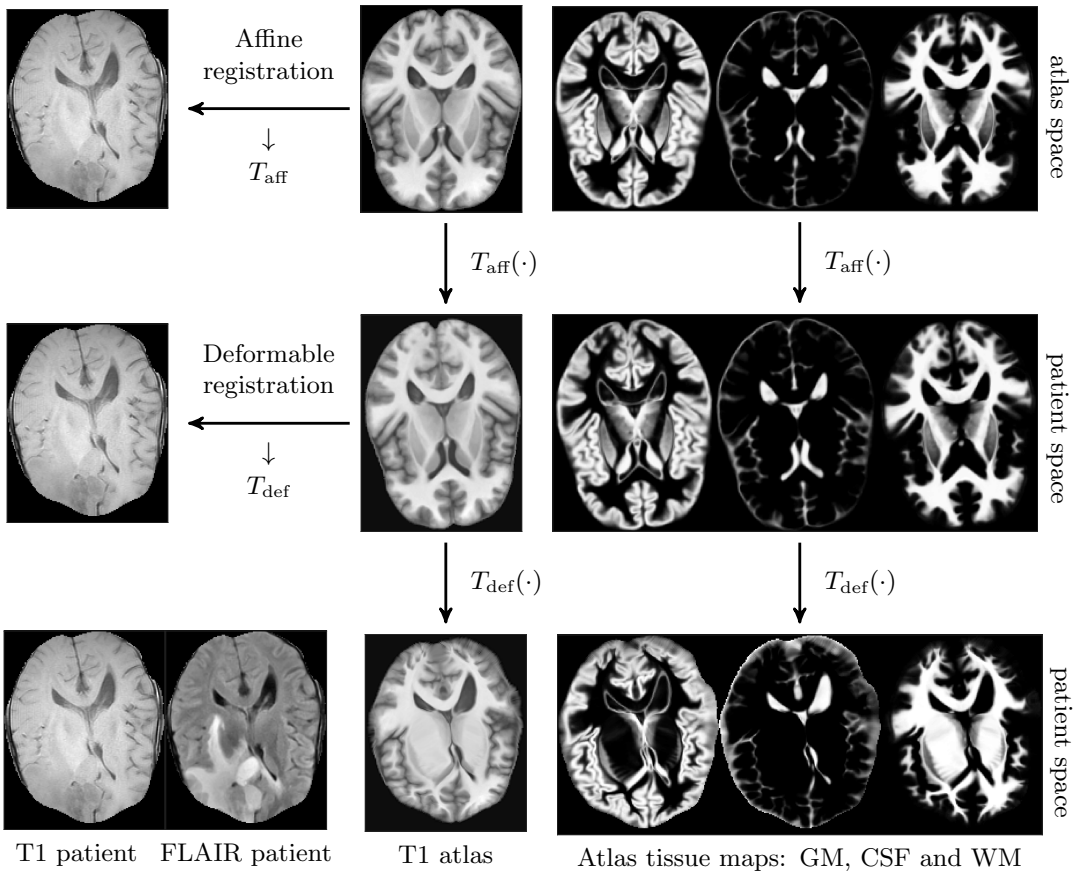


Figure 3.9: Registration flow adopted to create patient-specific tissue segmentations for WM, GM and CSF based on affine and deformable registrations T_{aff} and T_{def} between subject space (T1 patient and FLAIR patient) and atlas space (T1 atlas and WM, GM and CSF atlas).

formation matrix $T_{\text{aff}} \in \mathbb{R}^{4\times4}$ is calculated, mapping images in atlas space to patient space.

2. The GM, WM and CSF atlas maps can be registered to patient space, using T_{aff} calculated in the previous step. This generates affinely registered GM, WM and CSF atlas images.

3. A non-rigid deformation is calculated between the affinely registered T1 atlas image and the patient T1 image. During non-rigid registration, a control point grid is calculated from which a displacement for each voxel in the voxel grid can be derived. These voxel displacements are contained in a deformation field T_{def}.

4. The affinely registered GM, WM and CSF atlas images can be non-rigidly transformed to the patient T1 image using the deformation field T_{def}.

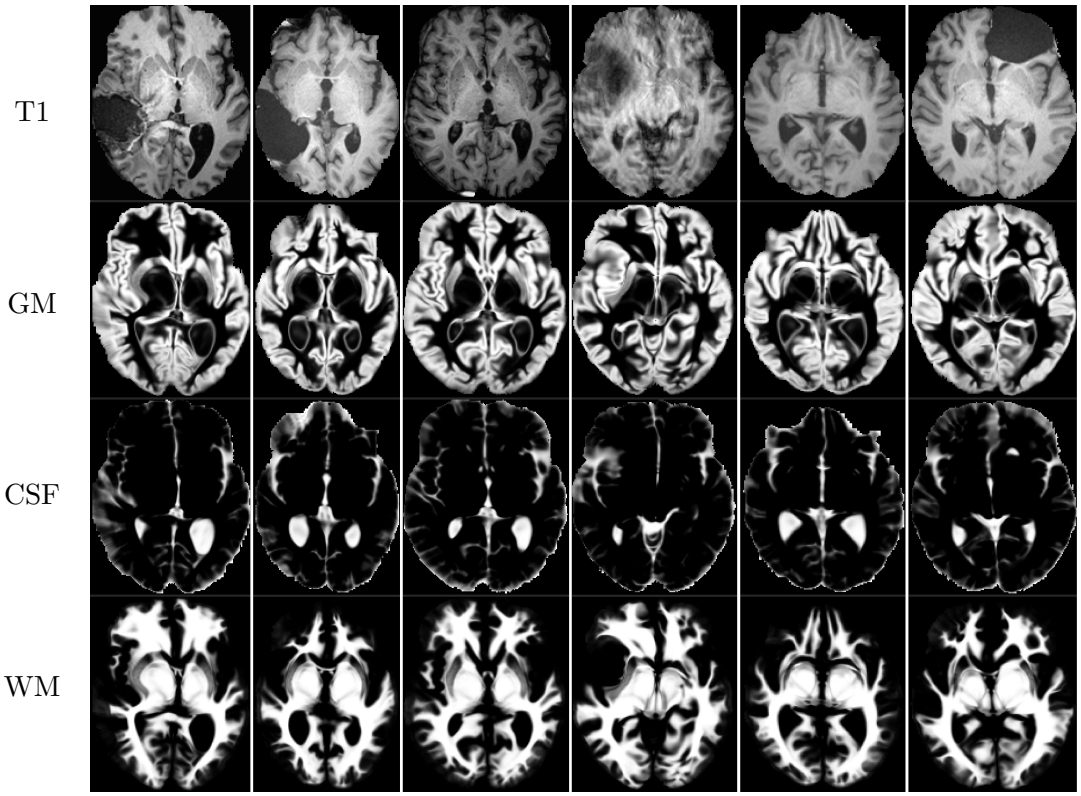


Figure 3.10: Examples of patient-specific tissue segmentations for WM, GM and CSF using the registration flow depicted in Figure 3.9. Note that the tissue segmentations are even accurate in the presence of resection cavities (cf. columns 1, 2 and 6) or if the patient T1 image is very blurry (cf. column 4). T1 patient images are taken from the dataset reported in [4].

Figure 3.10 illustrates more examples of tissue segmentations generated by means of the registration flow described above. The advantage of using non-rigid registrations, in addition to affine registrations, is that local deformations can be captured. As can be seen in Figure 3.10, local deformations are very useful when working with brain tumor images, as the GM and WM maps can be locally distorted in comparison to healthy brain images (cf. upper right image and lower right image in Figure 3.9).

Implementation of the affine and non-rigid registrations. Registrations are calculated using the `reg_aladin` (for affine transformations) and `reg_f3d` (for non-rigid registrations) functions implemented in the NiftyReg library. The affine registrations are calculated based on a multi-scale iterative block-matching scheme (cf. Section 3.3.2). The non-rigid registrations are calculated based on fast free form deformations using cubic B-splines calculated on a control point grid. The fast free form deformation algorithm also makes use of a multi-scale scheme, during which the resolution of the control point grid increases with the iterations [55].

3.4 Image Segmentation

Image segmentation involves the delineation of specific objects or structures present in an image. In medical image segmentation several segmentation tasks have been studied, ranging from the delineation of cells, tissues, organs or pathological structures. One of the major contributions of this thesis involves the segmentation of healthy brain tissues (such as GM, WM and CSF) and tumorous structures (such as edema, active tumor core and necrotic tumor core). In Chapter 4, a segmentation framework is presented that combines the segmentation of images available for several time points.

Segmentation algorithms make use of tools from machine learning, pattern recognition, statistics and computer vision. There exist a wide variety of image segmentation algorithms. This section only aims to provide a background on a small subset of them, focusing on those that are very relevant (such as convolutional neural networks) and those that have been adopted in this thesis. First, in Section 3.4.1, general image segmentation concepts are discussed, as well as some validation metrics. Then, a few selected segmentation algorithms are presented. Probabilistic graphical models, discussed in Section 3.4.2, are abundantly adopted in this thesis in the context of the Expectation-Maximization (EM) algorithm and conditional random fields (CRFs). Furthermore, random forests (RFs) are adopted in the context of supervoxel classifiers (cf. Chapter 7) and many medical image segmentation studies in the last decade have been conducted based on them. They will be presented in Section 3.4.3. In recent years, deep neural network implementations, such as convolutional neural networks, have become particularly successful in the context of medical image segmentation challenges. Therefore, they will be shortly addressed in Section 3.4.4.

3.4.1 Segmentation: Validation and Evaluation

Ground truth annotations. Image segmentation can be described as a classification task, more in particular, a *voxel* classification task. Therefore, both supervised or unsupervised (or even semi-supervised) techniques can be used. Supervised techniques can be problematic if no training sets with ground truth are available. *Ground truth* segmentations – also referred to as *annotations* – are image segmentations that reflect the perfect image segmentation. They are often acquired by manual segmentations done by medical experts or by working students. Many institutions have invested in creating reliable ground truth segmentations in order to be able to train and test segmentation algorithms. For brain tumor segmentation, the BRATS dataset has been created in 2012. The BRATS 2016 training dataset contains 244 segmentation cases (each with four tumor labels), all carefully reviewed by expert neuroradiologists. Unsupervised techniques do not necessarily need these big, accurately annotated, training datasets. However, just like supervised methods, they also need a test set including ground truth segmentations, on which they can be validated. The segmentations of the BRATS 2016 test dataset are kept to the organizers, to be able to correctly validate all segmentation algorithms.

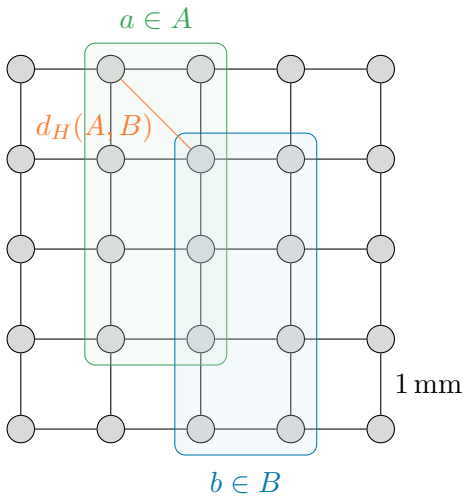


Figure 3.11: Illustration of two segmentations and their Hausdorff distance. A is the set of predicted tumor voxels (in blue) and B is the set of tumor voxels in the ground truth (in green). The Hausdorff distance is indicated in orange ($d_H(A, B) = \sqrt{2}\,\text{mm}$). Note that the Dice score is very low ($\frac{6}{16} = 37.5\%$), although the segmentations have the exact shame shape and are only shifted by $\sqrt{2}\,\text{mm}$.

Validation scores. In this thesis, the most commonly discussed segmentation validation score is the F1 score, also called the Dice score. It is a measure for voxel-wise overlap between the predicted segmentation and the ground truth segmentation. In this thesis, it is every time evaluated for each tumor label T separately against all other labels. It is defined as:

$$\text{F1 score / Dice score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{3.6}$$

$$= \frac{2 \ \text{precision} \ \text{recall}}{\text{precision} + \text{recall}} \tag{3.7}$$

$$= \frac{2 \mid A \cup B \mid}{\mid A \mid + \mid B \mid} \ , \tag{3.8}$$

where TP are the true positives, FP are the false positives, FN are the false negatives, the *precision* is the ratio of true positives over the number of true positives and false positives and *recall* the ratio of the number of true positives over the number of true positives and false negatives and A is the set of voxels which have been assigned a tumor label and B the set of voxels that have a tumor label in the ground truth.

Because the F1 score is a measure of voxel-wise overlap, the scores can be low for small structures, even if they resemble the ground truth very well. This is illustrated in Figure 3.11. Therefore, it might be useful to evaluate the Hausdorff distances in addition to the F1 scores. *Hausdorff distances* quantify surface distance rather than voxel-wise overlap. They are defined as the maximum distance obtainable between two elements a and b, with $a \in A$ and $b \in B$, b taken such that no other element in B is closer to a:

$$d_H(A, B) = \max\{\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b)\} \ , \tag{3.9}$$

where sup is the supremum and inf is the infimum. The Hausdorff distance is illustrated for two simplified segmentations in Figure 3.11.

In order to quantify the accuracy of the segmentation borders, *TRIMAP scores* are also reported in this thesis. TRIMAP scores quantify the F1 score only for the voxels that are close to the ground truth segmentation border. Therefore, they are usually reported in function of a radius r, representing the maximal distance from the considered voxels to the segmentation border.

3.4.2 Probabilistic Graphical Models

Probabilistic graphical models are diagrammatic representations of probability distributions [34]. They are typically represented by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ comprising vertices $\mathcal{V}$, representing random or deterministic variables, and edges $\mathcal{E}$ connecting vertices. Two types of graphical models exist: directed graphical models and undirected graphical models. Directed graphical models (edges are illustrated by arrows) are often called *Bayesian networks* and are useful to express causal relationships. Undirected graphical models are Markov random fields (MRFs) and are useful to express soft constraints between variables. In the research topics elaborated in this thesis, both Bayesian networks and Conditional random fields (CRFs) – a type of Markov random fields – are adopted.

3.4.2.1 Directed Graphical Models: Generative Probabilistic Models

Generative probabilistic models have been used for several medical segmentation applications, with a strong prevalence in brain analysis ([53],[8], [56]). The models are optimized by the Expectation-Maximization algorithm and are sometimes also referred to *atlas-based* methods, as they use population atlases to initialize healthy tissue classes.

The EM algorithm is a technique to minimize the negative log-likelihood of an observed dataset $\mathbf{x}$ in the presence of unknown parameters θ and latent variables $\mathbf{z}$. The log-likelihood can be written as:

$$\mathcal{L}(\theta; \mathbf{x}) = \log p(\mathbf{x}|\theta) = \sum_{v \in \mathcal{V}} \log p(\mathbf{x}_v|\theta) \tag{3.10}$$

$$= \sum_{v \in \mathcal{V}} \sum_{\mathbf{z}_v} p(\mathbf{x}_v, \mathbf{z}_v|\theta) \ , \tag{3.11}$$

with $v \in \mathcal{V}$ the voxels in the voxel grid. Note that, by writing the data log-likelihood as a sum over the voxels, it is assumed that voxels are statistically independent. However, many works include dependencies on neighboring voxels by using smart label priors $p(\mathbf{z}_v)$ in the joint distribution $p(\mathbf{x}_v, \mathbf{z}_v|\theta)$ (for example smoothened segmentation maps). Due to the presence of the hidden variables, setting the gradient of the data log-likelihood in (3.10) to zero will not leave us with a closed-form solution for the parameters **theta**. Therefore, the EM uses a technique called *iterative decoupling*: in each iteration it fixes the parameters θ and calculates the probability distributions over the hidden variables based on the current values for θ. Then, it fixes the probability distributions over the hidden variables, and updates the parameters θ accordingly. These

two steps in each iteration are described as the E(xpectation) and the M(aximization) step:

$$\text{E-step:} \quad Q_v(\mathbf{z}_v) \leftarrow p(\mathbf{z}_v|\mathbf{x}_v, \theta^{(i)}) \tag{3.12}$$

$$\text{M-step:} \quad \theta^{(i+1)} = \arg\max_{\theta} \sum_{v \in \mathcal{V}} E_{Q_v} \left[\log p(\mathbf{x}_v, \mathbf{z}_v|\theta)\right] \tag{3.13}$$

The main model choices to be made are then

1) how to model $p(\mathbf{x}_v, \mathbf{z}_v|\theta)$, such that $p(\mathbf{z}_v|\mathbf{x}_v, \theta^{(i)})$ can be calculated in the Expectation step:

$$p(z_v{=}l|\mathbf{x}_v, \theta_l) = \frac{p(\mathbf{x}_v, z_v{=}l|\theta_l)}{\sum_{l=1}^{k} p(\mathbf{x}_v, z_v{=}l|\theta_l)} \ , \tag{3.14}$$

2) how to initialize the algorithm (either by setting $\theta^{(0)}$ or by starting with some initial estimation for $p(\mathbf{z}_v|\mathbf{x}_v, \theta^{(i)})$) and

3) how many hidden variables to take up in $\mathbf{z}_v$ and how many labels each hidden variable $\mathbf{z}_v$ should contain.

The research projects in Chapter 4-5 adopt the EM-framework presented in [8] to create initial segmentation maps that are fed to CRFs. In Chapter 7 a new EM-segmentation framework is developed and explained in detail. The main differences between the EM-framework from [8] and the one presented in Chapter 7 is how the hidden variables are implemented and how $p(\mathbf{x}_v, \mathbf{z}_v|\theta)$ is modeled.

In [8], the hidden variables $\mathbf{z}_v$ are modeled as a sequence of variables: $\mathbf{z}_v = [k_v, t_v^1 \dots t_v^m]$. Here, k_v is the label of the healthy tumor class: $k_v \in \{0, 1, 2\}$ for WM, GM and CSF and t_v^i indicates the tumor presence in modality i: $t_v^i \in \{0, 1\}$. The full state table for $\mathbf{z}_v$, as calculated in [8], is given in Table 3.3. Due to *inclusion constraints* some of the possible states are considered invalid and ignored. Note that the posterior probabilities $p(\mathbf{z}_v|\mathbf{x}_v, \theta^{(i)})$ need to be estimated for all possible states of $\mathbf{z}_v$, in the E-step. Tumor segmentations as well as WM, GM and CSF segmentations can be calculated by summing up $p(\mathbf{z}_v|\mathbf{x}_v, \theta^{(i)})$ for different states of $\mathbf{z}_v$ (as explained in the caption in Table 3.3). Solving for 12 different states is computationally expensive, especially since, in the end, only the tumor segmentations are relevant. Moreover, the probability that the EM converges to a global minimum is smaller in the presence of several hidden variables [57]. Therefore, the EM-framework presented in Chapter 7 only models one hidden variable: $\mathbf{z}_v = z_v \in \{1, \dots l\}$ with l possible label values: GM, WM, CSF, active tumor, non-active tumor, necrotic core and edema.

In [8], $p(\mathbf{x}_v, \mathbf{z}_v|\theta)$ is modeled using tissue prior maps based on population atlases. Tumor priors are initialized using flat priors and updated with the tumor segmentations of the previous iteration. Furthermore, tumor and healthy tissue is modeled in each channel separately, using univariate Gaussians. In Chapter 7, label prior maps are generated by another algorithm (a supervoxel random forest classifier). Furthermore, each

Table 3.3: Full state table for the hidden variable sequence modeled in each voxel v along four standard MR modalities (T1, T1c, T2 and FLAIR), according to [8]. Due to the *inclusion constraints*, tumor is assumed to be present in T2 if it is present in T1c or T1 and in FLAIR if it is present in T2. No tumor is modeled in CSF. To get a tumor segmentation for active tumor, the probabilities for state 4, 5, 10 and 11 are summed up, for tumor core, the probabilities for state 2-5 and 8-11 are summed up and for whole tumor, the probabilities for state 1-5 and 7-11 are summed up. To get a segmentation for WM, the probabilities for state 0-5 are summed up, for GM, the probabilities for state 6-10 are summed up and for CSF, the probabilities of state 12 are taken.

possible states for $\mathbf{z}_v$	k_v	t_v^1 (T1)	t_v^2 (T1c)	t_v^3 (T2)	t_v^4 (FLAIR)
state 0: WM, no tumor	0 (WM)	0	0	0	0
state 1: WM, tumor in FLAIR	0	0	0	0	1
state 2	0	0	0	1	1
state 3	0	1	0	1	1
state 4	0	0	1	1	1
state 5: WM, tumor in all	0	1	1	1	1
state 6: GM, no tumor	1 (GM)	0	0	0	0
state 7: GM, tumor in FLAIR	1	0	0	0	1
state 8	1	0	0	1	1
state 9	1	1	0	1	1
state 10	1	0	1	1	1
state 11: GM, tumor in all	1	1	1	1	1
state 12: CSF, no tumor	2 (CSF)	0	0	0	0

label is modeled in a multivariate fashion over all channels simultaneously. It therefore uses univariate Gumbel distributions (defined for each channel separately) and multiplies it with multivariate Gaussian copulas (defined over all channels simultaneously) to efficiently model inter-channel dependencies. Doing so, non-Gaussian multivariate distributions can be modeled, which is shown to be very relevant when modeling tumor intensity distributions in the standard MR modalities.

3.4.2.2 Undirected Graphical Models: Conditional Random Fields

Conditional random fields (CRF) have been proposed in [58] for the purpose of segmenting and labeling sequence data. CRFs are undirected graphical models specifically tailored to model a conditional probability distribution $p(\mathbf{y} \mid \mathbf{x})$, defined on an output segmentation $\mathbf{y}$ conditioned on a certain input $\mathbf{x}$. The input typically are the features or the image data. An example of a graphical model of a CRF designed for image segmentation is illustrated in Figure 3.12.

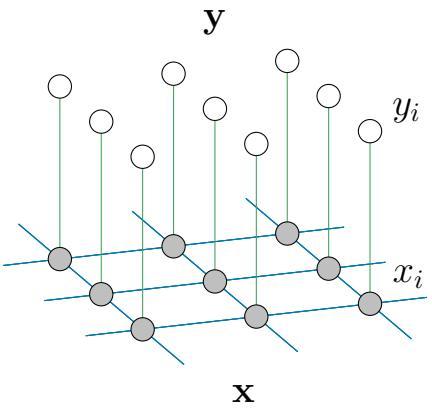


Figure 3.12: Example of a graphical model of a CRF, implemented for the purpose of image segmentation. The input image $\mathbf{x}$ is observed, whereas the output segmentation $\mathbf{y}$ is to be predicted. The edges reflect unary potentials (–), defined between voxels x_i and their segmentation labels y_i, as well as binary or pairwise potentials (–), defined between two voxels.

CRFs are characterized by an energy function, implemented as a summation over potential functions. When only considering unary and binary potentials (as in Figure 3.12), the energy function can be written as:

$$E(\mathbf{y} \mid \mathbf{x}) = \sum_{i \in \mathcal{V}} \phi_i \left(y_i, x_i\right) + \sum_{(i,j) \in \mathcal{E}} \phi_{ij} \left(y_i, y_j, x_i, x_j\right) \quad , \tag{3.15}$$

where $\phi_i \left(y_i\right)$ and $\phi_{ij} \left(y_i, y_j\right)$ are the unary and pairwise potentials, respectively.

The probability distribution over the output segmentation $\mathbf{y}$ can then be written by means of the Gibbs or Boltzmann distribution (cf. (3.3)):

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z} \exp -E(\mathbf{y} \mid \mathbf{x}) \quad . \tag{3.16}$$

3.4.2.3 Inference

In graphical models, some of the vertices have known (observed) values. During inference, the posterior distributions of (subsets of) the unobserved vertices are calculated, given the values of the observed vertices. In image segmentation, the unobserved variable typically is the full output segmentation (spanning multiple vertices), while the observed vertices can be the input images. In the CRF, the output segmentation is calculated by minimizing the energy function. Hence, during CRF inference, the energy function is minimized as a function of the output segmentation.

Various inference methods exist for graphical models, being either exact (sum-product or max-sum algorithms if no loops are present, junction tree algorithm [34]) or approximative (iterative conditional modes, loopy belief propagation based on local message passing). For image segmentation, algorithms based on graph-cut are very efficient [59]. In Chapter 4 inference is done by a graph-cut algorithm described in [60]. In Chapter 5, inference is calculated by a mean field approximation, iteratively optimized through a series of message passing steps [61].

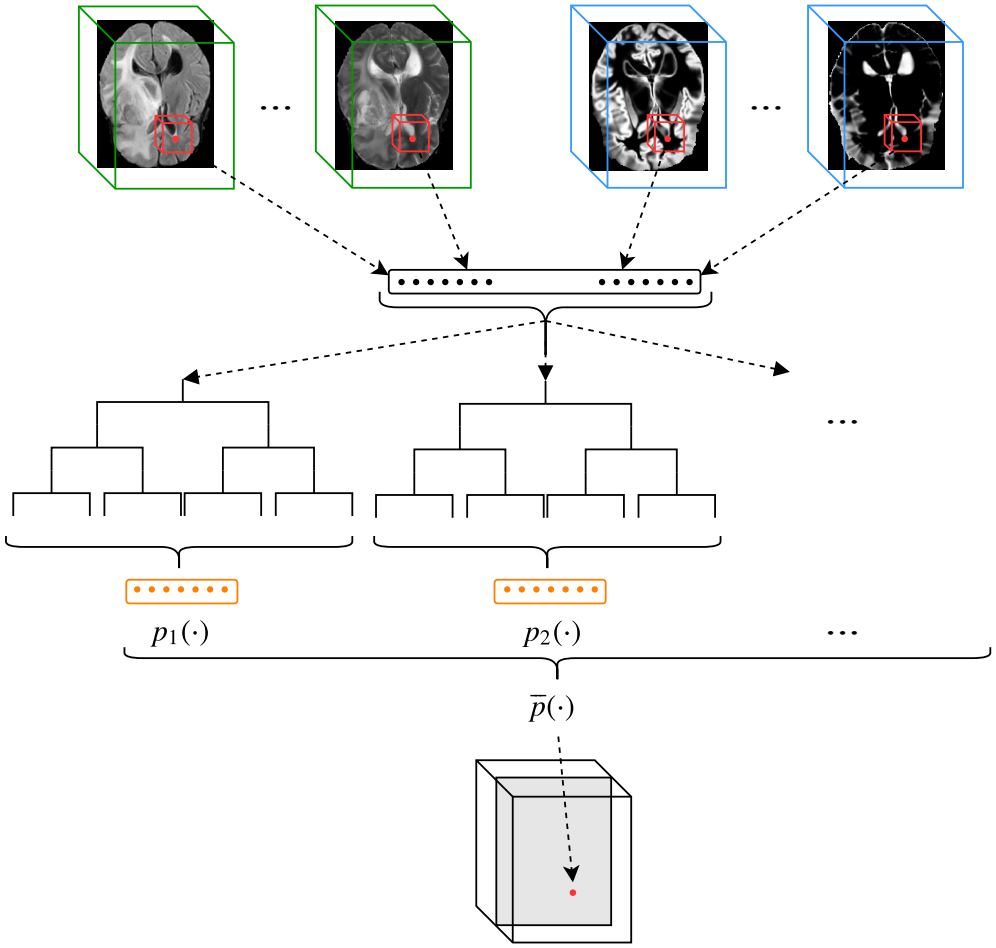


Figure 3.13: Example of how random forests can be used for image segmentation: features are extracted within a patch around the center voxel (in red) in the input images (blue cubes) or in their corresponding tissue segmentations (green cubes). The feature vector is then fed to the trees in the random forests and the final label probabilities for the red center voxel are obtained by averaging them across the trees.

3.4.3 Random Forests

Random forests (RF) is a popular algorithm used for image segmentation [62]. An overview of the most common approach is depicted in Figure 3.13. First, features are extracted for each voxel in the 3D volumes. These volumes can consist of the medical image data (depicted as blue cubes in Figure 3.13), or of additional feature maps, such as segmented brain tissues (depicted as green cubes in Figure 3.13). Then, these features are classified by a RF, which will predict label probabilities for each segmentation label. Lastly, the labels predicted for each voxel are rearranged in a 3D volume, resulting in the label image.

RF is an ensemble machine learning algorithm. Ensemble methods predict labels based on the prediction of several base estimators. In RF, the base estimators are decision trees, which are learned by *bagging*: a powerful technique to reduce the variance of machine learning algorithms. By bagging decision trees, each decision tree is trained on a set of training samples, sampled with replacement from the training set. The number of decision trees is a hyper-parameter of the RF algorithm. The decision trees consist of internal nodes, which are split into two child nodes. Child nodes can again be internal nodes and split further, or they can be leaves without any further descendants. The split in each node is made based on one feature. During training, the feature that can establish the *best split* needs to be determined for each internal node. In RF it is important to notice that the algorithm only considers m features, *randomly subsampled* from the full feature set, to determine the best split for each internal node (m is a hyper-parameter of the RF algorithm). The quality of a certain split is determined by the split criterion, often either the Gini impurity or the Entropy. Overfitting can for example be avoided by setting a fixed depth of the trees or by defining a minimum number of training samples required to split an internal node (this will stop splitting nodes that contain only a small number of training samples and make them leaves).

In order to establish accurate image segmentations, it is important to feed the relevant voxel features to the random forest. The authors in [63] propose to use 3-dimensional binary context features, called BRIEF features, and classify them later by an adapted random forest algorithm, called vantage point forests. In [64] a fully-automated brain tumor segmentation method based on RF, "BraTumIA", is presented. Both appearance-sensitive and context-sensitive features are extracted. Appearance-sensitive features are based on the MR images (they include local mean and variance in a small neighborhood around the center voxel), whereas context-sensitive features encode information related to location and spatial arrangement.

In this thesis, RFs are used for image segmentation in the context of supervoxel classifiers in Chapter 7. Here, supervoxel features are extracted for supervoxels (not for voxels), and labels are predicted via a RF for the entire supervoxels.

3.4.4 Convolutional neural networks

In recent years, convolutional neural network (CNN) implementations have outperformed the state-of-the-art in many medical image segmentation tasks, including brain tumor segmentation. CNNs are an important pillar in the sudden success of *deep learning*. Although CNNs have been around since the 80', inspired by Fukushima [65] and LeCun [66], they have only become very accurate and successful since a few years – most likely due to increasing annotated training datasets as well as recent advances made in GPU computing.

CNNs for image classification. The first CNNs were developed for **image classification**, where one label is predicted for a full input volume. The input can consist of multi-modal 2D slices (for example RGB images) or even of multi-modal 3D volumes. It consists of several *hidden layers*, each traditionally consisting of a convolutional layer,

an activation layer and a pooling layer. In the *convolutional layer*, the input images are convolved with k kernels, generating k feature maps. In the *activation layer* every pixel in the features maps is activated by a non-linear activation function (e.g. the rectified linear (ReLu) or the sigmoid activation function). And finally, in the *pooling layer*, the feature maps are down-sampled by pooling filters acting at a certain stride (e.g. taking the maximum value as in max-pooling layers). The net effect is that the feature maps in deep layers are more numerous and smaller in dimension. In other words, the layers become coarser as the network grows deeper. In the last layers, the feature maps are flattened into a feature vector, which is then fed to a number of fully-connected layers. Similar to the multi-layer perceptron, the last layer results in one activated unit for each classification label. Label probabilities are then obtained via the softmax function. During training, the kernel weights of the convolutional layers are learned via *backpropagation* of the error obtained between the predicted labels and the true labels. This error E is expressed by means of an *error function* or a cost function, for example weighted cross-entropy. During backpropagation, error gradients are calculated for each kernel weight based on one training sample. After processing a mini-batch of m training samples, each kernel weight w_i is updated based on their calculated error gradients:

$$w_i^{(t)} = w_i^{(t-1)} + \Delta w_i^(t) \tag{3.17}$$

$$\Delta w_i^(t) = \mu \Delta w_i^{(t-1)} - \alpha\eta \frac{1}{m} \sum_{j=1}^{m} \frac{\partial E_j}{\partial w_i} \ , \tag{3.18}$$

with E_j the error for training sample j, η the learning rate, $0 < \alpha \leq 1$ the weight decay factor and μ the momentum (these can be hard-coded or optimized using for example the ADAM optimizer [67]). After processing all mini-batches in the training set and updating the kernel weights accordingly, the network is said to be trained for one *epoch*. It should however be noted that, before the training set is divided in *mini-batches*, it is commonly augmented. During *data augmentation*, images in the training set are transformed (scaled, rotated or translated, or a combination thereof) and added to the original training set. This increases the training data and learns the network to recognize shifted, scaled or rotated objects. Other frequently adopted methods used in deep learning implementations include *batch normalization* [68] and *dropout layers* [69].

Patch-based CNNs for image segmentation. The first CNN implementations for medical image segmentation were patch-based. In [70], patches are extracted from input volumes and classified using the network. The adopted classification label is then assigned to the center pixel in the input volume at which the patch was extracted. Patch-based CNNs implemented for brain tumor segmentation are presented in[71], using 2D multi-scale patches, and in [72], using 3D patches. In Figure 3.14, an example of a patch-based CNN for image segmentation is given, classifying 3D patches of size $14 \times 14 \times 14$.

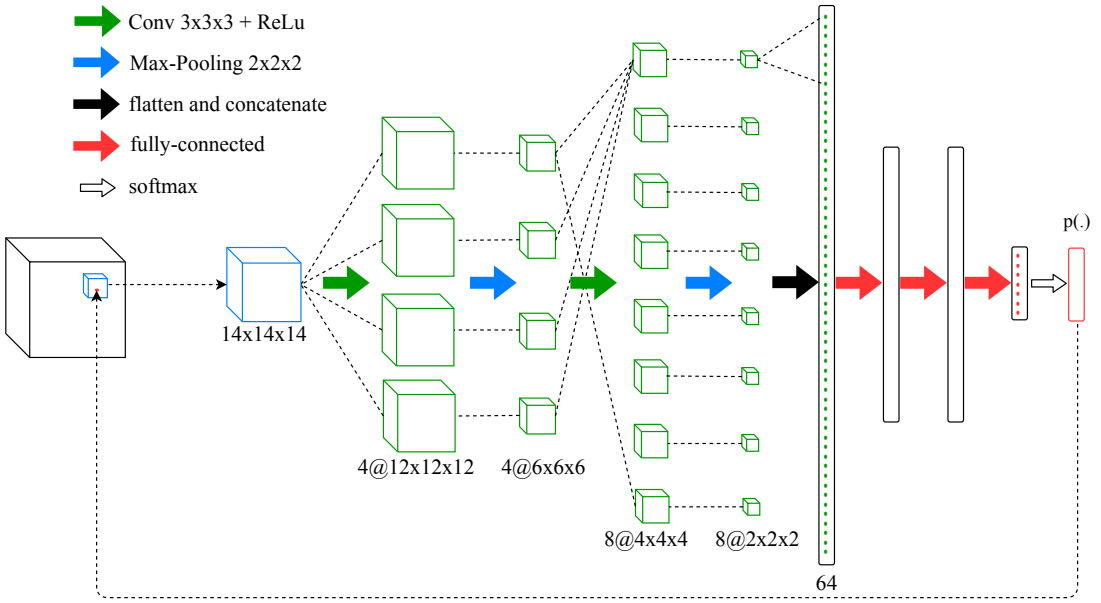


Figure 3.14: Example of a patch-based CNN for image segmentation, classifying 3D patches in an input volume. Patches (blue cubes) are extracted from the input volume (black cube) and classified by a CNN. The illustrated CNN consists of 5 hidden layers. The first two consist of a 3D convolutional layer with $3 \times 3 \times 3$ kernels, a Rectified Linear (ReLu) activation layer and a pooling layer with kernel size and stride $2 \times 2 \times 2$. Feature maps are visualized as green cubes (their dimensions are indicated below them: ¡number of feature maps¿@¡dimension of each feature map¿). The last three layers are fully-connected. Finally, the softmax function results in a label probability vector for the center pixel shown in red. (The network architecture is chosen as an example and is not taken from another source.)

Though patch-based CNNs have access to more training samples than image-based CNNs (there are more patches than image volumes), they can only include context within a certain bandwidth, limited by the size of the extracted patches.

Fully-convolutional networks for image segmentation. Patch-based CNNs for image segmentation can only include limited context information. Therefore, [73] developed a fully-convolution network (FCN) for image segmentation acting on full image volumes of arbitrary dimensions. In short, the authors remove the fully-connected layer of existing CNNs for image classification and attach deconvolution layers to upsample the deeper coarse layers back to the fine voxel grid. Doing so, the FCN consists of a contractive path, similar to the convolutional layers in a CNN classification network, followed by an expansive path, in which the feature maps are upsampled to the original dimensions of the input volume. The error function is now typically expressed by means of the Dice score. In [74], this network is further refined to the U-net. The main contribution

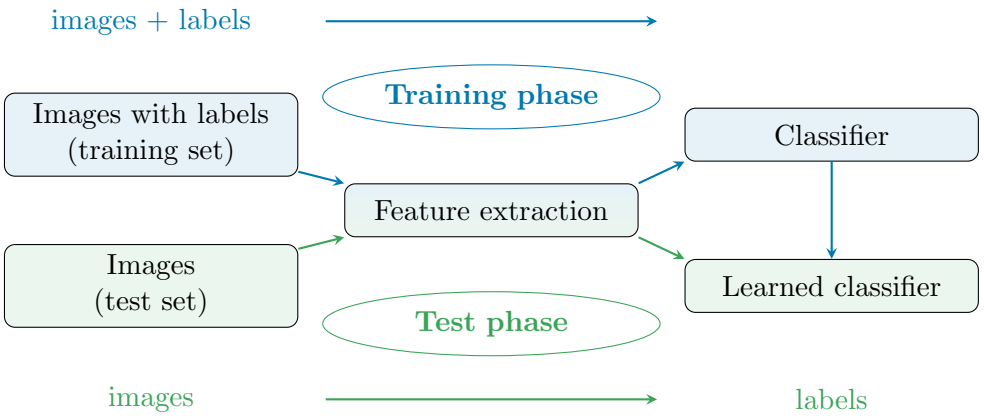


Figure 3.15: Illustration of the training and test phase in image classification.

of the U-net, is to concatenate feature maps calculated in the contractive path, to the feature maps calculated in the expansive path. This allows the network to propagate context information adopted in the contracting path towards the upsampled feature maps in the expansive path. FCNs for brain tumor segmentation have been proposed in [75],[76],[77].

3.5 Image Classification

In this thesis, image classification is discussed as a type of *image identification*: each image in the training set has one label and this label is to be predicted for the images in the test set. This is different from *object recognition*, in which a set of objects are identified in the images, together with their location in the image (such as pedestrians, cars, driving lane and road signs in the case of driving scene perception).

During supervised image classification, an algorithm is first trained based on a training set consisting of input images and their according labels. Once the algorithm has been trained, it can then be used to predict the labels of new input images, for example those present in the test set (cf. Figure 3.15).

Image classification is discussed in this thesis in the context of *radiogenomics*. In fact, in the studies conducted in this thesis, multiple images are available for each tumor. The tumor type is hence determined based on a multi-modal set of images. This is illustrated in Figure 3.16. This type of image classification is also called "multi-modal image classification". Note that the input space has an extremely high dimensionality. In Chapter 6, a multi-modal image classification framework is presented, based on [11]. This framework has also been used to successfully predict the tumors genetic profile based on diffusion-weighted MR images instead of anatomical MR images [12].

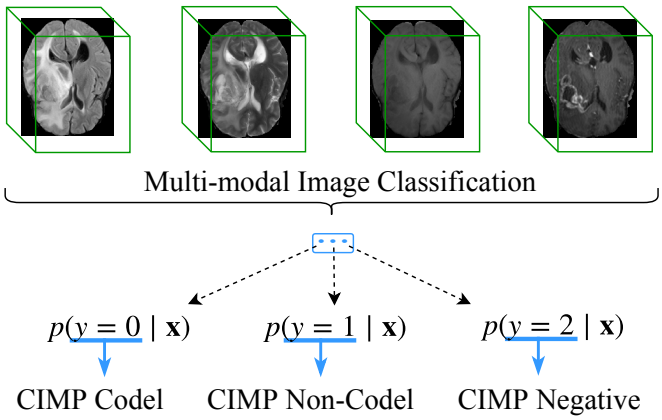


Figure 3.16: Illustration of the multi-modal image classification task studied in Chapter 6. A label is determined based on a multi-modal image set of 3-dimensional volumes.

3.5.1 Overview

In image classification, the task is traditionally divided in two parts: a first part extracts image features and a second part trains a classifier based on the extracted image features and learns to predict the image label. It should be noted that CNNs for image classification, in contrast to traditional computer vision approaches, combine feature extraction and classification, by learning features directly during the optimization of the classification task.

3.5.2 Texture Features

Classical texture descriptors include gray-level co-occurrence matrices (GLCM), Haralick features, Haar-like features and histogram of oriented gradients (HOG). GLCM are matrices calculated for a given offset $(\Delta x, \Delta y, \Delta z)$. First, the pixel values within the image need to be quantized into a finite set of gray values (*grey levels*), for example consisting of n gray levels. In that case, the matrix M will consist of n rows and n columns: $M \in \mathbb{N}^{n \times n}$. The element at (i, j) in the matrix will then indicate the number of pixels of gray level i which have a neighbor pixel at offset $(\Delta x, \Delta y, \Delta z)$ of gray level j. Haralick features consist of various metrics calculated on the GLCM [78]. Haar-like features are based on the difference in the sum of pixel values computed in certain rectangles (or cubes) [79]. Finally, HOG features are based on counting occurrences of gradient orientation in localized regions in the image [80]. Furthermore, filter banks, such as the Gabor filter bank [81], or wavelets, such as Riesz wavelets [82], have been used to extract texture features. For multi-modal medical image classification tasks, it is beneficial to consider short local image descriptors. Local binary patterns (LBP) [83] and BRIEF features [84] are short descriptors, written as binary strings calculated over image patches, and even propose an efficient 3D implementation in [85] and [63]. One last important pillar in texture feature analysis consists of deep learning frameworks. For example, restricted Bolzmann machines [86] or auto-encoders [11] have been successfully adopted for texture representation.

Table 3.4: Medical image classification for disease detection using various image features and different classifiers

	Studied disease	**Images**	**Features**	**Classifier**
[82]	Medulloblasoma	Histopathology	Riesz wavelets	softmax
[87]	Glioblastoma	MRI	polar S-transform	neural net
[88]	Breast cancer	Histopathology	Haralick	RF
[89]	Glaucoma	Retinal fundus	convolutional	neural net
[90]	Osteoarthritis	Radiography	intensity differences	RF
[91]	Antinuclear Autoantibodies	Indirect Immuno Fluorescence	LBP & SIFT & GLCM	elastic net
[92]	Prostate cancer	Ultrasound	deep belief network	SVM

3.5.3 Classifiers

For image classification, several classifiers have been suggested. In Chapter 6, experiments are conducted with a range of classifiers, including support vector machines (SVM), multi-layer perceptrons (MLP), k-nearest neighbors, logistic regression and random forests (RF). In this study, the RF performed best on the validation set. RFs have been proposed for image classification in a range of studies [62]. The authors in [88] propose to use *regenerative random forests* and the authors in [93] propose *scandent trees*, a random forest implementation which allows to cope with missing features.

3.6 Conclusion

Brain MR image preprocessing is an important part of image analysis. It usually consists of several steps, including image registration, biasfield correction, brain extraction and intensity normalization. In the case of brain MR images, it might be helpful to extract healthy brain tissue segmentations for WM, GM and CSF as well, for example by means of deformable image registrations. It is important to note that it can be necessary to preprocess new incoming images in the exact same way as training images were preprocessed, before they are fed to trained classification algorithms for further image analysis tasks. Such image analysis tasks can for example consist of image segmentation or image classification, for which a wealth of machine learning and computer vision methods can be adopted.

Part II

Research projects

4 A Nonparametric Growth Model for Brain Tumor Segmentation in Longitudinal MR Sequences

In this chapter, we investigate in retrospect whether a tumor has grown or shrunken in between two different scanning sessions. The framework is based on graph cuts calculated over a conditional random field (CRF) which includes growth constraints across all available time points. It assesses energy values of graph cuts obtained with different growth constraints in order to define a confidence in each growth constraint. The CRF is constructed based on input MR images, for the pairwise potentials, and on prior tumor probability maps, calculated by the segmentation framework from [8], for the unary potentials. This work has been published in [38], and is here only adapted to a limited extent.

4.1 Introduction

The assessment of disease progression after brain tumor treatment is very important in clinical practice for disease surveillance and treatment planning, but also in drug trials and clinical studies for evaluating drug or treatment efficacy.

Automatic tumor segmentation is well-suited for tumor volumetry. A longitudinal tumor segmentation is visualized in Figure 4.1. In contrast to expensive manual segmentations, automatic tumor segmentation obtains fast, reproducible and objective results. Over the past years, several automatic tumor segmentation methods have been developed (cf. [2] presenting an extended overview). Among these, longitudinal methods have been implemented to explicitly use time information. For example in [94], 4-dimensional (4D) spatio-temporal cliques are included in a CRF, enforcing regularization over time. However, this temporal regularization tends to smooth sudden growth

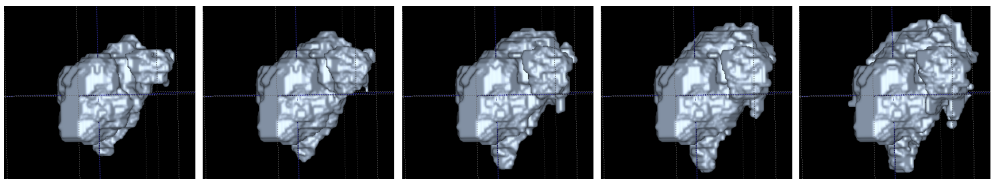


Figure 4.1: Illustration of tumor volumetry of a low-grade glioma along time. Longitudinal 3-dimensional tumor segmentation masks are shown chronologically from left to right, illustrating a steady tumor growth.

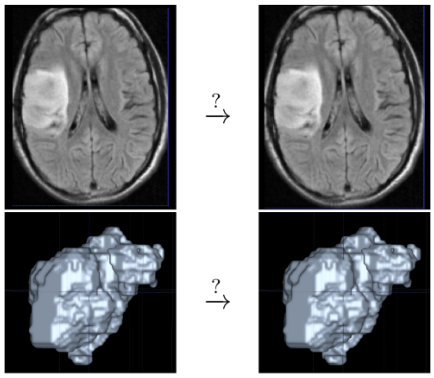


Figure 4.2: Identifying tumor growth or shrinkage in retrospect. Top: FLAIR images for two MR studies acquired for the same patient. Bottom: corresponding whole tumor segmentations.

events and the empirical temporal smoothness parameters are not easy to learn. The authors of [14] present a model based on a 4D CRF using infinite link functions that effectively constrain voxel classifications depending on predefined conditions, which allow to constrain tumor segmentations to grow or shrink for every time transition. This model can handle abrupt changes in tumor growth and only includes one parameter for spatial regularization.

In literature, tumor growth is often modeled by means of parametric models based on cell kinetics and reaction-diffusion processes, as reported in [95]. These models often aim to *predict* tumor growth in the future, rather than study it in retrospect, and do not calculate tumor segmentations in itself (prior tumor segmentations are included for initialization purposes). The authors of [56] were the first to use a parametric growth model to assist in brain tumor segmentation. However, parametric models are computationally expensive, make assumptions about tumor growth regularity and cannot easily handle post-operative tumor structures with resection cavities.

We believe tumor growth modeling and segmentation are inter-dependent, and aim to exploit this property by jointly optimizing both in the same framework. We adopt the longitudinal segmentation model developed in [14] and implement it as a nonparametric tumor growth segmentation model. We further develop the model to include a fast and robust estimation of the spatial regularization parameter and extend this model to detect tumor regrowth in longitudinal sequences, considering every two subsequent time points in a longitudinal sequence in retrospect (cf. Figure 4.2). We consider the clinical scenario where a tumor shrinks after therapy and automatically detect the time point at which tumor regrowth begins.

4.2 Methods

An overview of the model is depicted in Figure 4.3. The input consists of a set of 3D MR intensity images $\mathbf{I}$ and a set of 3D prior tumor probability maps $\mathbf{x}$. The set of 3D MR intensity images extends over M modalities (T1, T1c, T2 and FLAIR) and T time points: $\mathbf{I} = \{I_{stp} \mid s \in \{1,\ldots,M\}, t \in \{1,\ldots,T\}, p \in \mathcal{P}\}$, with s the modality index, t a time index and p a voxel index on the voxel grid $\mathcal{P}$ of the 3D volumes. Similarly, the

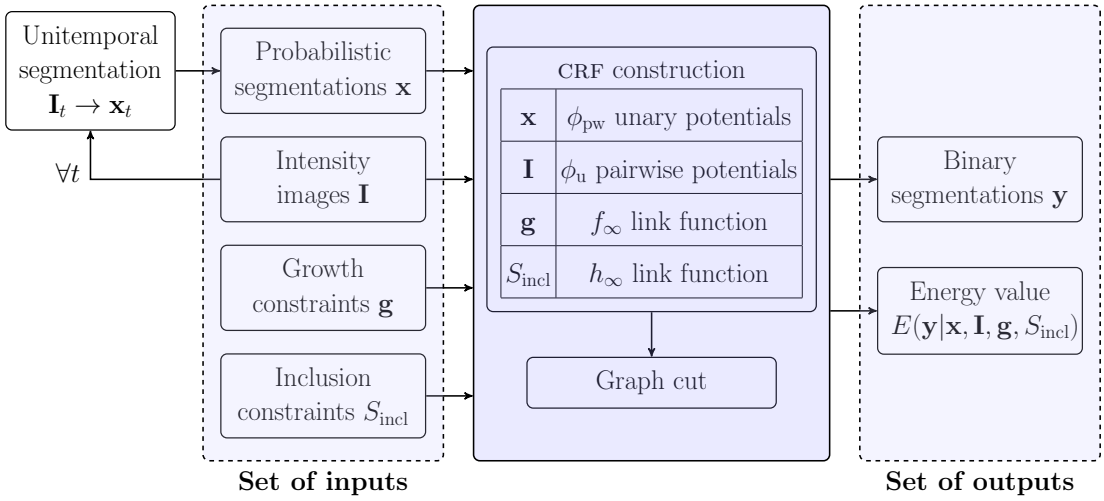


Figure 4.3: Overview of the proposed method. The inputs are listed and incorporated in certain terms of the energy function of the CRF. Inference is calculated by means of graph cut and outputs a binary segmentation. In this work we also use the energy value as an output of the CRF, in order to evaluate the likelihood of the growth constraints.

prior tumor probability maps are present for each modality and each time point and are indexed over the same voxel grid: $\mathbf{x} = \{x_{stp} \mid s \in \{1, \ldots, M\}, t \in \{1, \ldots, T\}, p \in \mathcal{P}\}$. These prior maps can for example be calculated by means of generative models lacking spatial, temporal and inter-modality coherence.

The growth model is specified through growth and inclusion constraints. The growth constraints specify whether the tumor is expected to grow or shrink for each time transition. They can either be specified in the model as to incorporate prior information (as in Experiment 1, Sect. 4.3.1) or they can be learned in a probabilistic framework (as in Experiment 2, Sect. 4.3.2). They are represented by a binary array indexed over all time transitions, $\mathbf{g} = [g_1, g_2, \ldots, g_{T-1}], \mathbf{g} \in \{0, 1\}^{T-1}$, where each element g_t imposes growth (1) or shrinkage (0) in between time points t and $t + 1$.

The inclusion constraints are represented by a set of pairs of modality indices, $(s', s'') \in S_{\text{incl}}$, such that all tumor voxels in the first modality, s', are a subset of the tumor voxels in the second modality, s''. The inclusion constraints allow us to incorporate our prior knowledge of tumor occurrence across different modalities: the tumor voxels in T1 and T1c are to be a subset of the tumor voxels in T2 and the tumor voxels in T2 are to be a subset of the tumor voxels in FLAIR.

In Section 4.2.1 we present the graph model and the implementation of the energy function and in Section 4.2.2 we explain how we use the energy values resulting from the graph cuts to define a confidence over different growth patterns.

4.2.1 A 4D CRF as a Nonparametric Growth Model (NPGM)

We first present the construction of the CRF graph and then proceed with the implementation of the energy function and the spatial regularization parameter.

4.2.1.1 Graph construction

A CRF is an undirected graphical model, represented by a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ with *vertices* $\mathcal{V}$ and *edges* $\mathcal{E}$ connecting related vertices. The edges are quantified by edge weights and represent the affinity between vertices.

In our implementation, the vertices include the voxel grid of the 3D volumes extended over M modalities and T time points, as well as the tumor/non-tumor labels. The edge weights are implemented in the energy function $E(\mathbf{y} \mid \ldots)$ in terms of the output segmentation $\mathbf{y} = \{y_{stp} \mid s \in \{1, \ldots, M\}, t \in \{1, \ldots, T\}, p \in \mathcal{P}\}$. The energy function quantifies how well the affinity between vertices (specified by the edge weights) is reflected by the output segmentation $\mathbf{y}$.

4.2.1.2 Implementation of the energy function

The energy function E is implemented as in [14]. It is modeled such that favorable states of $\mathbf{y}$ yield low energies. Minimization of the energy function E solves for the optimal output segmentation $\mathbf{y}$.

In general, the energy function consists of unary potentials ϕ_{u} and pairwise potentials ϕ_{pw}, weighted by a spatial regularization parameter λ. The unary potentials describe individual label preferences and the pairwise potentials describe voxel interactions encouraging spatial coherency. In this study, we extended the energy function by two functions, f_∞ and h_∞, to account for edges of infinite weight, which we introduce in order to exclude pairs of labels violating our growth or inclusion constraints. The energy function is then written as:

$$E(\mathbf{y} \mid \mathbf{x}, \mathbf{I}, \boldsymbol{\Theta}) = \sum_{t=1}^{T} \sum_{s=1}^{M} \left(\sum_{p \in \mathcal{P}} \phi_{\mathrm{u}}(x_{stp}, y_{stp}) + \lambda \sum_{(p,q) \in \mathcal{N}} \phi_{\mathrm{pw}}(I_{stp}, I_{stq}, y_{stp}, y_{stq}) \right)$$
$$+ \sum_{t=1}^{T-1} \sum_{s=1}^{M} \sum_{p \in \mathcal{P}} f_\infty(g_t, y_{stp}, y_{s(t+1)p}) + \sum_{t=1}^{T} \sum_{s'=1}^{M} \sum_{s''=1}^{M} \sum_{p \in \mathcal{P}} h_\infty(S_{\mathrm{incl}}, y_{s'tp}, y_{s''tp}) \quad , \tag{4.1}$$

where $\boldsymbol{\Theta} = \{\mathbf{g}, S_{\mathrm{incl}}\}$ the growth and inclusion constraints and $\mathcal{N}$ the set of voxel pairs within a spatial neighborhood. We briefly elaborate on the implementation of each term.

Unary potentials. The *unary potentials* are implemented based on the tumor probability maps $\mathbf{x}$:

$$\phi_{\mathrm{u}}(x_{stp}, y_{stp}) = y_{stp}\,(1 - x_{stp}) + (1 - y_{stp})\,(x_{stp}) \quad . \tag{4.2}$$

Pairwise potentials. The spatial *pairwise potentials* are implemented within the 3D volumes. They are quantified by a Gaussian, modeling the MR intensity difference between each voxel pair within a 3D neighborhood matrix $\mathcal{N}_{26}$:

$$\phi_{\mathrm{pw}}(I_{stp}, I_{stq}, y_{stp}, y_{stq}) = \begin{cases} \mathrm{d}(p,q)^{-1} \exp -\dfrac{(I_{stp} - I_{stq})^2}{2\sigma^2} & \text{if} \quad y_{stp} \neq y_{stq} \ , \\ 0 & \text{else} \ , \end{cases} \tag{4.3}$$

where $\mathrm{d}(p,q)$ is proportional to the voxel spacing and σ^2 is set to the variance of image intensities present in the 3D volume.

Growth constraints. The *growth constraints* are imposed on voxel pairs belonging to the same modality, having the same index within the 3D volumes, and being strictly consecutive in time. An infinite penalty is imposed if *a*) growth is imposed but the voxels switch from tumor ($y_{stp} = 1$) to non-tumor ($y_{s(t+1)p} = 0$), or *b*) shrinkage is imposed but the voxels switch from non-tumor ($y_{stp} = 0$) to tumor ($y_{s(t+1)p} = 1$):

$$f_{\infty}(g_t, y_{stp}, y_{s(t+1)p}) = \begin{cases} \infty & \text{if} \quad (g_t = 1) \ \wedge \ (y_{stp} > y_{s(t+1)p}) \ , \\ \infty & \text{if} \quad (g_t = 0) \ \wedge \ (y_{stp} < y_{s(t+1)p}) \ , \\ 0 & \text{else} \ . \end{cases} \tag{4.4}$$

Inclusion constraints. The *inclusion constraints* are imposed on voxel pairs of the same time point and having the same index within the 3D volumes. An infinite penalty is imposed if the voxels belong to two modalities in between which the inclusion constraint holds, $(s', s'') \in S_{\mathrm{incl}}$, and if the voxel in s' is tumor and the voxel in s'' is not:

$$h_{\infty}(y_{s'tp}, y_{s''tp}) = \begin{cases} \infty & \text{if} \quad ((s', s'') \in S_{\mathrm{incl}}) \ \wedge \ (y_{s'tp} > y_{s''tp}) \ , \\ 0 & \text{else} \ . \end{cases} \tag{4.5}$$

Once the edge weights have been assigned based on this energy function, the CRF is solved by graph cut, as described in [60].

4.2.1.3 Spatial regularization parameter λ

The regularization parameter, λ, is an important system parameter: an overly high value leads to under-segmentation and an overly low value leads to poor spatial regularization. Moreover, a good value for λ differs from one case to another. There are several methods to learn this parameter. A fairly easy, fast and robust method is adopted in [96], where the parameter is made spatially adaptable. That is, λ is set to lower values for voxels close to the edges of the images:

$$\lambda_{stp} = (1 - L_{stp})\lambda_{\max} \ , \tag{4.6}$$

where L_{stp} is the edge probability of a single voxel and $\lambda_{\max}$ is empirically set to 3. We calculate the edge probability map $\mathbf{L}$ based on the tumor probability maps $\mathbf{x}$, by applying an edge detector and subsequent Gaussian smoothing.

4.2.2 Switching from Tumor Shrinkage to Tumor Regrowth

Once the CRF is solved by graph cut, we obtain an energy value. In [97] these energy values are used to calculate the confidence in spatial voxel classifications. More precisely, the confidence in a single voxel classification in [97] is based on the energies acquired from graph cuts with and without a voxel classification constraint, which is imposed by an infinite link.

As our growth constraints are enforced by the same infinite link functions, we can transfer this spatial uncertainty measure to the temporal domain and quantify uncertainties – or confidences – in specific tumor growth constraints.

First, consider a growth constraint for a single time transition $t \to t+1$: $g_t = a$, with $a \in \{0, 1\}$. We define the min-marginal energy for this growth constraint, $\psi_{t,a}$, as the minimal energy within the family of energies obtained from graph cuts for all growth constraint patterns where g_t is kept equal to a:

$$\psi_{t,a} = \mathrm{C}^{-1} \min_{\mathbf{g},\mathbf{y}} E(\mathbf{y} \mid \mathbf{x}, \mathbf{I}, \mathbf{g}, S_{\mathrm{incl}}) \ , \quad \forall \mathbf{g} \in \{\{0,1\}^{T-1} \mid g_t = a\} \ , \tag{4.7}$$

with C as the number of voxels constrained with an infinite temporal link. Note that the calculation of $\psi_{t,a}$ requires 2^{T-2} graph cuts. The confidence in the growth constraint for this single time transition, $\sigma_{t,a}$, can then be calculated as a function of the min-marginal energies $\psi_{t,a}$, similar to [97]:

$$\sigma_{t,a} = \frac{\exp\left(-\psi_{t,a}\right)}{\exp\left(-\psi_{t,a}\right) + \exp\left(-\psi_{t,1-a}\right)} \ , \quad a \in \{0,1\} \ . \tag{4.8}$$

This calculation requires 2^{T-1} graph cuts. Note that this set of graph cuts covers all possible patterns of growth constraints. The energies of these graph cut solutions can be re-used to calculate $\sigma_{t',a'}$ for all other time points t' and all growth constraints $a' \in \{0, 1\}$.

The confidence in the entire pattern of growth constraints, $\sigma_{\mathbf{g}}$, is then calculated as the product of confidences over all time transitions: $\sigma_{\mathbf{g}} = \prod_{t=1}^{T-1} \sigma_{t,g_t}$.

4.3 Experiments

Data Specifications. We used ten patient-specific datasets acquired at the German Cancer Research Center (DKFZ), yielding a total of 248 images. Each patient-specific dataset contains four multi-modal sequences (T1, T1c, T2 and FLAIR) for three to nine time points, with time intervals of ± 90 days. Patients initially suffered from low grade gliomas, but some developed high grade gliomas in the course of the study.

All images within the same dataset are skull-stripped and affinely co-registered. For each image, manual ground truth segmentation is available in three orthogonal slices intersecting at the tumor center. The manual segmentations were acquired by a clinical expert who took images of several time points into account at once.

We calculated tumor probability maps with a generative model based on an Expectation-Maximization (EM) segmenter, as in [8]. Modalities belonging to the same time point

Table 4.1: FLAIR Dice scores for all ten datasets segmented by the EM segmenter and by the nonparametric growth model (NPGM) with different parameter settings concerning growth constraints and spatial regularisation parameter.

EM segmentation:									$[79\% \pm 8\%]$
63%	79%	89%	77%	67%	84%	80%	79%	84%	86%
NPGM - no growth constraints, adaptive λ:									$[81\% \pm 5\%]$
78%	80%	90%	82%	71%	84%	80%	80%	82%	87%
NPGM - constrained to grow, fixed λ:									$[82\% \pm 5\%]$
74%	78%	91%	81%	80%	85%	81%	82%	83%	86%
NPGM - constrained to grow, adaptive λ:									$[83\% \pm 4\%]$
81%	78%	93%	82%	81%	84%	83%	82%	83%	87%

are processed together, but time points are processed independently. The segmentation maps are concatenated over all time points, to obtain a valid input for the NPGM.

4.3.1 Experiment 1: Segmentation Accuracy

In this experiment we compare *a*) EM segmentations (i.e. acquired from generative model), *b*) NPGM segmentations (i.e. acquired from the nonparametric growth model) where no growth constraints are included, *c*) NPGM segmentations where the tumor is constrained to grow over all time transitions and *d*) NPGM segmentations where the spatial regularization parameter is voxel-adaptive as in (4.6).

Table 4.1 reports the FLAIR Dice scores for all ten datasets, for each of these segmentations. Dice scores of T2 and T1 are comparable and not all datasets are suitable for T1c segmentations. The Dice scores are highest for the segmentation where the tumor is constrained to grow along time and where the spatial regularization parameter is voxel-adaptive.

T2 and FLAIR segmentations are shown for three patients in Figure 4.4 and segmentations for all modalities are shown for one patient in Figure 4.5. These figures illustrate a clear improvement from EM to NPGM segmentations. Figure 4.6 shows tumor volumetry for three datasets along time. This figure illustrates that the use of growth constraints does not only attain higher Dice scores, but also results in a more realistic progress in tumor volume.

In terms of computation time, a NPGM segmentation of a dataset of eight time points and four modalities takes $\pm$ 10 s on a Intel® Xeon® Processor E3-1225 v3.

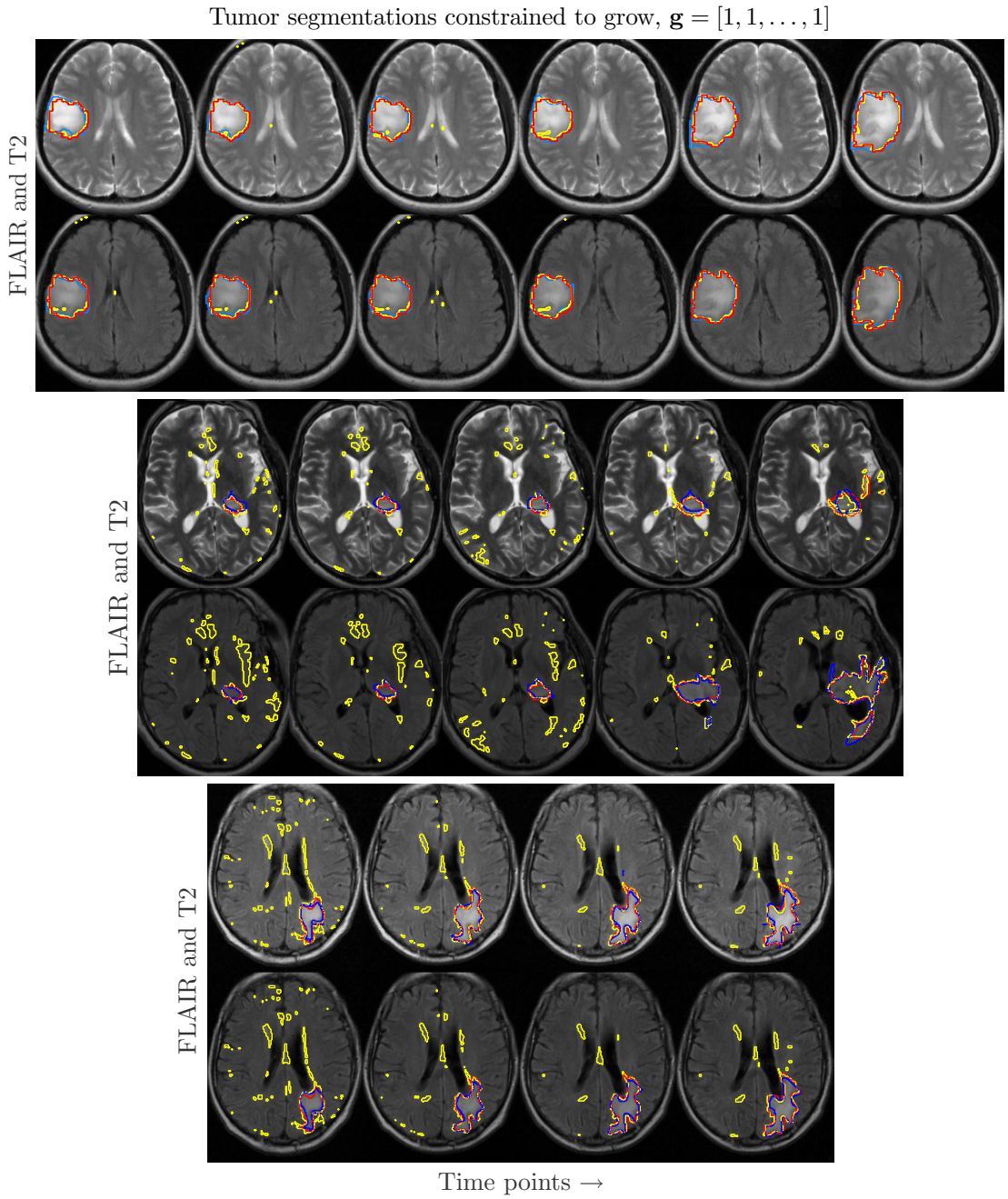


Figure 4.4: FLAIR and T2 images for three patients with 6, 5 and 4 time points, annotated with EM segmentations (*yellow*), NPGM segmentations with a strict growth constraint along time (*red*) and ground truth (*blue*).

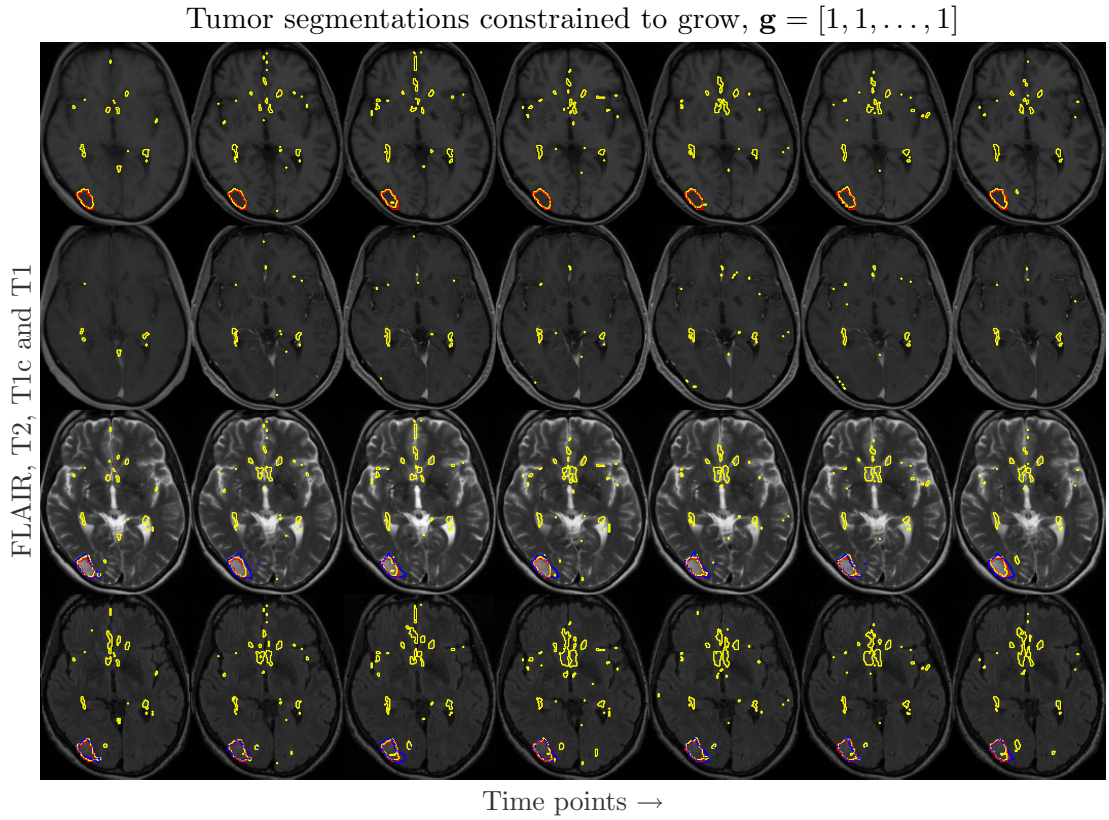


Figure 4.5: FLAIR, T2, T1c and T1 images for one patient with 7 time points, annotated with EM segmentations (*yellow*), NPGM segmentations with a strict growth constraint along time (*red*) and ground truth (*blue*). Ground truth for T1 and T1c is not shown. These segmentations illustrate the spatial regularization of the CRF.

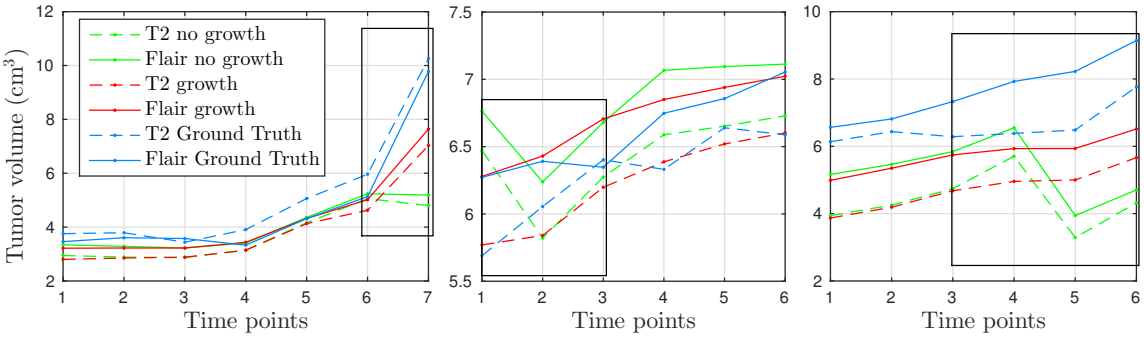


Figure 4.6: Tumor volumetry of T2 (*dashed lines*) and FLAIR (*solid lines*) showing a clear advantage in the application of growth constraints (*red*) rather than leaving them out (*green*) when comparing with ground truth (*blue*).

4.3.2 Experiment 2: Detection of Tumor Regrowth

We adopt the probabilistic formulation for different patterns of growth constraints (Sect. 4.2.2) to detect at which point tumor regrowth begins. We shorten the datasets to include three time points. Based on the ground truth volumes, we rearranged the order of the three time points in order to get 84 sequences with:

1. tumor shrinking for both time increments, that is: $\mathbf{g}_0 = [0, 0]$,
2. tumor shrinking for the first time increment and growing for the second time increment, that is: $\mathbf{g}_1 = [0, 1]$.

Using the probabilistic growth framework explained in Sect. 4.2.2, we now calculate the probabilities of $\mathbf{g}_0$ or $\mathbf{g}_1$ for each sequence. This experiment is of clinical relevance: tumors tend to shrink temporarily after therapy and tumor regrowth needs to be detected as soon as possible. For each sequence, the algorithm will estimate confidence measures in $\mathbf{g}_0$ and $\mathbf{g}_1$. We obtain probabilities for both tumor growth patterns by normalizing these confidence measures:

$$[p_{\mathbf{g}_0}, p_{\mathbf{g}_1}] = [\sigma_{\mathbf{g}_0}, \sigma_{\mathbf{g}_1}]/(\sigma_{\mathbf{g}_0} + \sigma_{\mathbf{g}_1}) \ . \tag{4.9}$$

Figure 4.7 illustrates the amount of correctly classified tumor growth patterns. Of 168 datasets, 128 datasets were correctly classified, 35 datasets were falsely estimated to grow after the second time point (false positives) and only 5 datasets were falsely estimated to keep shrinking after the second time point (false negatives). To the right in Figure 4.7, one can see that the accuracy of tumor regrowth detection is highly related to the relative increase in tumor volume between the last time points. As expected, the difference in the tumor growth pattern probabilities ($|p_{\mathbf{g}_1} - p_{\mathbf{g}_0}|$) tends to be lower for misclassified tumor growth patterns. Note that our classification detects

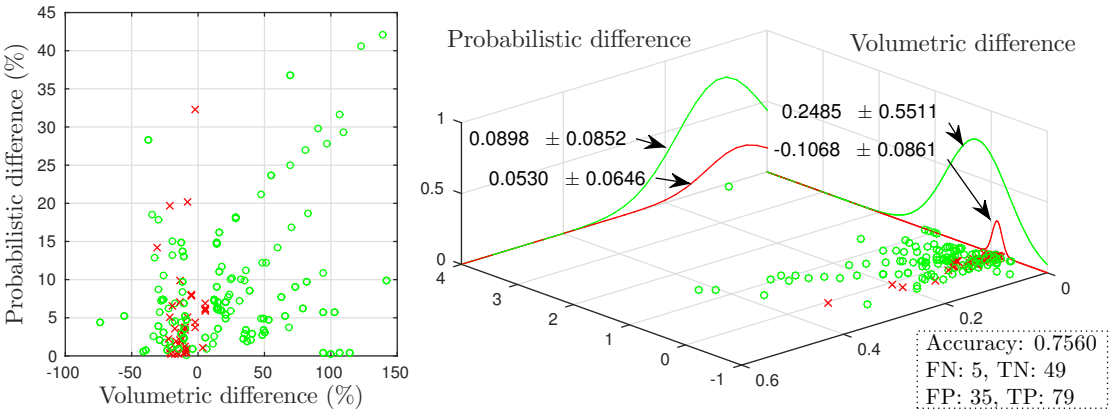


Figure 4.7: Distribution of correctly (○) and incorrectly (×) classified tumor growth patterns as a function of the difference in the growth pattern probabilities ($|p_{\mathbf{g}_1} - p_{\mathbf{g}_0}|$) and as a function of the relative increase in tumor volumes between the last time points.

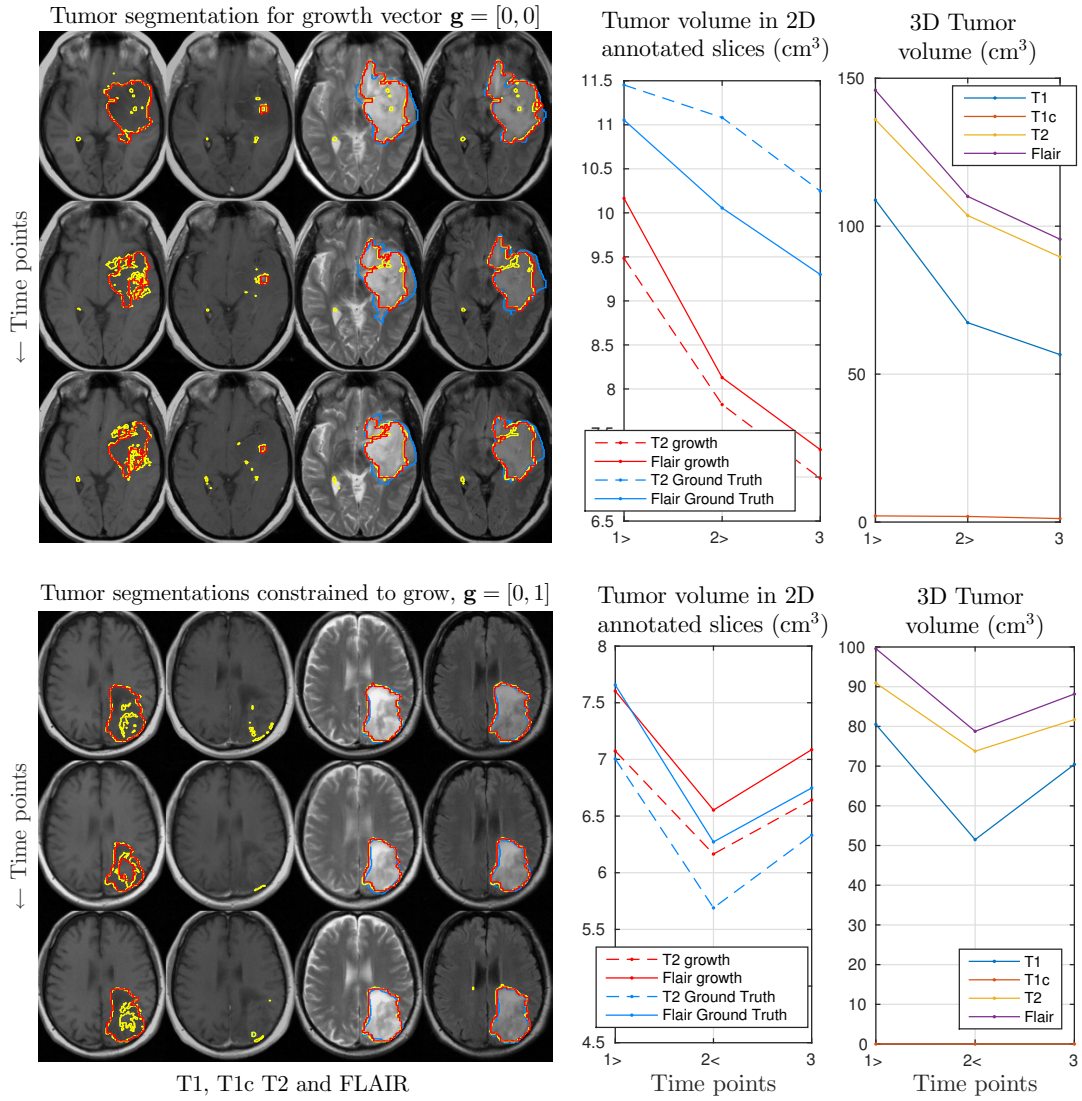


Figure 4.8: *Upper row*: dataset depicting tumor shrinkage over both time transitions, *lower row*: dataset depicting tumor regrowth occurring at the second time point. The brain image slices are annotated with EM segmentations (*yellow*), NPGM segmentations with $\mathbf{g} = [0, 0]$ and $\mathbf{g} = [0, 1]$ (*red*) and ground truth (*blue*). Ground truth for T1 is not shown and tumor is not present in T1c. Volumes are given within the 2D ground truth annotated slices (*in the middle*) and for the entire 3D volumes (*to the right*).

either shrinkage or growth. In other words, it does not account for cases of 'stable disease', where the tumor is neither shrinking nor growing. This injects noise in our classification model, which gives rise to misclassifications.

Figure 4.8 illustrates segmentations of two rearranged datasets – one dataset with a tumor shrinking over two time increments and one with a tumor shrinking for the

first time increment, but growing for the second time increment – together with tumor volumetry of T1, T1C, T2 and FLAIR.

4.4 Conclusion

In this study, we present a nonparametric model to segment brain tumors and to estimate the occurrence of tumor growth and/or shrinkage along time. We show the advantage of including longitudinal information in order to acquire more accurate tumor segmentations and volumetry. Furthermore, we adopt a fast and practical solution for the estimation of the spatial regularization parameter in the CRF energy function. Our model was extended to include probabilistic formulations for tumor regrowth after therapy, and it was shown to succeed in accurately estimating the occurrence of tumor regrowth.

5 Uncertainty Quantification in Brain Tumor Segmentation Using CRFs and Random Perturbation Models

In this chapter, we further investigate how to access brain tumor segmentation uncertainties using the segmentation framework presented in Chapter 4. In particular, we assess uncertainties in the context of a conditional random field (CRF), making use of MAP perturbation models presented in [15] and [16]. This work has been published in [39], and is here only adapted to a limited extent.

5.1 Introduction

Medical image segmentation is a well-studied field involving the delineation of cells, tissues, organs and pathological structures. The major difficulties in this field are the inherent ill-posed nature of the segmentation task as well as the high variability of the data observed across subjects (due to tissue heterogeneities), across data acquisitions (due to imaging artifacts) and across medical sites (due to different scanning protocols). Therefore, segmentation algorithms often have to make decisions in the presence of uncertainty. However, validation usually only relies on the hard segmentations generated by the algorithm, and little prior work has been done to estimate and report the uncertainty of the results. In this paper, we want to emphasize that uncertainty in medical image segmentation is a valuable concept and can play a critical role in the evaluation and validation of an algorithm, as well as in clinical practice, where uncertainty can provide useful information in several applications such as disease diagnosis, therapy guidance and radiotherapy planning.

The conventional method to assess segmentation uncertainty is by using probabilistic segmentation models. In the past years, graphical models have become very popular, as they allow to incorporate spatial, temporal and inter-image coherence. However, they make it difficult to go beyond hard segmentation. Traditionally, Markov chain Monte Carlo (MCMC) methods are used for sampling purposes. These methods are computationally expensive, suffer from burn-in periods and lack scalability. Therefore, sampling directly over the entire voxel grid is to be avoided, and instead, one would rather sample from parametric representations as in [98] or use blockwise updates as in [99]. Alternatively, [97] proposed to use min-marginal energies to quantify voxel-specific uncertainties, but it becomes expensive for volumetric voxel grids. In [15] and [16], the principle of random MAP perturbations is introduced. These perturbations

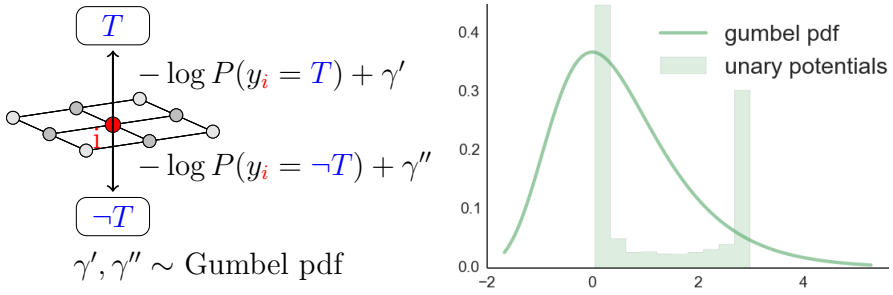


Figure 5.1: *Left:* Gumbel samples perturb the unary potentials in between graph vertices i and label terminals T (tumor) and $\neg T$ (non-tumor). *Right:* range of Gumbel samples compared to the non-perturbed unary potentials.

allow sampling from the Gibbs distribution, which is inherently defined in CRF models or equivalent energy minimization approaches.

In this work, we extend the framework presented in Chapter 4 for brain tumor segmentation (composed of a probabilistic local classification scheme and a CRF model), with random MAP perturbations in order to sample brain tumor segmentations. With this approach, we are able to go beyond the typical hard segmentation, ie. the MAP solution, and quantify the uncertainty of the segmentation without additional model assumptions. To our knowledge, we are the first to employ this theoretical framework to obtain uncertainties in medical image segmentation.

5.2 Methods

In the following, we briefly review the background on CRFs and the recent work on perturbation models.

The presented work is an extension to the framework presented in Chapter 4. Starting from a CRF that simultaneously segments brain tumors visible in several standard MR modalities such as T1, T1c and FLAIR, we make use of the perturbation model to quantify the uncertainties in the segmentations.

5.2.1 CRFs and perturbation models

The CRF is defined over a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $i \in \mathcal{V}$ for each voxel and edges $(i, j) \in \mathcal{E}$ between related vertices. In the particular model that we employ, edges will be introduced for all neighburing voxels and between the corresponding voxels in all modalities. A variable vector $\mathbf{y}$ describes the assigned labels for all voxels. The energy function $E(\mathbf{y})$ of the CRF is then defined over the graph $\mathcal{G}$ as:

$$E(\mathbf{y}) = \sum_{i \in \mathcal{V}} \phi_i \left(y_i\right) + \sum_{(i,j) \in \mathcal{E}} \phi_{ij} \left(y_i, y_j\right) \quad , \tag{5.1}$$

where $\phi_i \left(y_i\right)$ and $\phi_{ij} \left(y_i, y_j\right)$ are the unary and pairwise potentials, respectively. It is modeled such that favorable states of $\mathbf{y}$ yield low energies. For many practically

relevant cases, the energy function can be minimized (approximately) with efficient optimizers such as graph-cut algorithms [59]. For this study, we use the unary and pairwise potential functions from (4.2) and (4.3) and adopt the inference algorithm of [61], which is based on a mean field approximation, iteratively optimized through a series of message passing steps. Note that the perturbation approach is not restricted to this particular choice of potentials.

It is important to notice that the CRF implicitly defines a probability distribution over the segmentations $\mathbf{y}$ in terms of a Gibbs distribution:

$$P(\mathbf{y}) = \frac{1}{Z} \exp\big(-E(\mathbf{y})\big) \ , \tag{5.2}$$

where $Z = \sum_{\mathbf{y}\in\mathcal{Y}} \exp(-E(\mathbf{y}))$ is the partition function, which acts as a normalization constant ensuring that the distribution sums to 1. This gives us directly a way of quantifying the uncertainty in our segmentation by computing the marginal distributions. Unfortunately, calculating these marginals is intractable in loopy graphs [57] and sampling with MCMC approaches is computationally expensive. To this end, recent work of Papandreou and Yuille [16] has shown that the Gibbs distribution can be approximated by perturbing its energy function with Gumbel noise (cf. Section 5.2.1.1 and Section 5.2.1.2), and solving for its MAP state repeatedly (as explained in Section 5.2.1.3). This allows us to leverage powerful solvers for sampling segmentations from the CRF to approximate the marginal distribution for all voxels.

Next, we elaborate on two particular ways to perturb the energy function that we investigate in this work.

5.2.1.1 Voxel-specific Gumbel perturbations

Originally, the authors of [16] have proven that samples drawn from the perturbation model coincide in the limit case with the Gibbs distribution if the noise γ is drawn IDD from a Gumbel distribution with zero mean for the full state table of $\mathbf{y}$. While it is not feasible to do this in practice – the full state table has exponentially many entries – several studies [15, 16] have empirically shown that applying low-order perturbations of Gumbel noise yields sufficient results. Our first approach is therefore to perturb each of the unary potentials of the CRF with a sample drawn from the zero-mean Gumbel distribution, as illustrated in Figure 5.1. For this case, we can write down the whole perturbation as:

$$\gamma(\mathbf{y}) = \sum_{i\in\mathcal{V}} \sum_{k\in\{T,\neg T\}} \gamma_i^k \mathbb{1}(y_i = k) \ , \tag{5.3}$$

where $\mathbb{1}(.)$ is the indicator function and γ_i^k are samples of the Gumbel distribution with zero mean.

5.2.1.2 Context-sensitive Gumbel perturbations

Inspired by the Swendsen-Wang cluster-specific updates [99], we explore perturbations on a supervoxel scale in order to detect context-sensitive uncertainties. We parcel the

voxel grid into supervoxels (of $\approx$ 1ml) using MonoSLIC [100] and draw perturbations γ from the Gumbel distribution for each supervoxel, resulting in identical perturbations for the unary potentials of all voxels within the same supervoxel. We note that doing so violates the assumptions of the original proof in [16] and the resulting segmentation samples do not follow the original Gibbs distribution of the CRF anymore. Instead, we can interpret it as an additional, context-sensitive correlation prior on the supervoxel grid that is added to the original CRF.

5.2.1.3 Sampling with perturbation models

Having defined both CRF and the type of perturbation, we can draw samples $\mathbf{y}$ of segmentations with the following procedure:

1. Draw random perturbations $\gamma(\mathbf{y})$ as described in Sect. 5.2.1.1 or 5.2.1.2.
2. Draw a new sample $\hat{\mathbf{y}}$ by computing the MAP state of the perturbed CRF:

$$\hat{\mathbf{y}} = \arg\min_{\mathbf{y}\in\mathcal{Y}} E(\mathbf{y}) + \gamma(\mathbf{y}) \ . \tag{5.4}$$

The two steps are repeated until the desired amount of samples has been created to approximate the marginal distributions (also called *soft segmentation* in our setting). Other than MCMC methods, we do not have to throw away initial samples as there is no burn-in period.

5.3 Experiments and Results

We conducted experiments on 15 patient-specific datasets, each dataset consisting of 3 time points, and each time point containing 3 standard MR modalities (T1, T1c and FLAIR) as illustrated in Figure 5.2. In all T1c and FLAIR volumes, indicating active tumor core and whole tumor similar to the BRATS standards [2], 3-dimensional ground truth annotations have been acquired by a clinical expert using a interactive segmentation tool (SmartBrush, by Brainlab).

Note that this dataset poses a complicated segmentation task. The data contains pre- and post-operative scans depicting resection cavities, internal bleedings and scar tissue.

The datasets are preprocessed by intra-subject registrations, skull extraction and isotropic resampling. We calculated initial brain tumor regions using a generative model as in [8]. Inference is calculated on all modalities at once, as in Chapter 4, but time points are processed individually. Using the perturbation models, we sampled 100 segmentations for each time point for each patient.

5.3.1 Quantitative evaluation

Based on the marginal distributions computed by the voxel-specific perturbation model, we calculated precision-recall curves for each segmentation task. Figure 5.3 illustrates

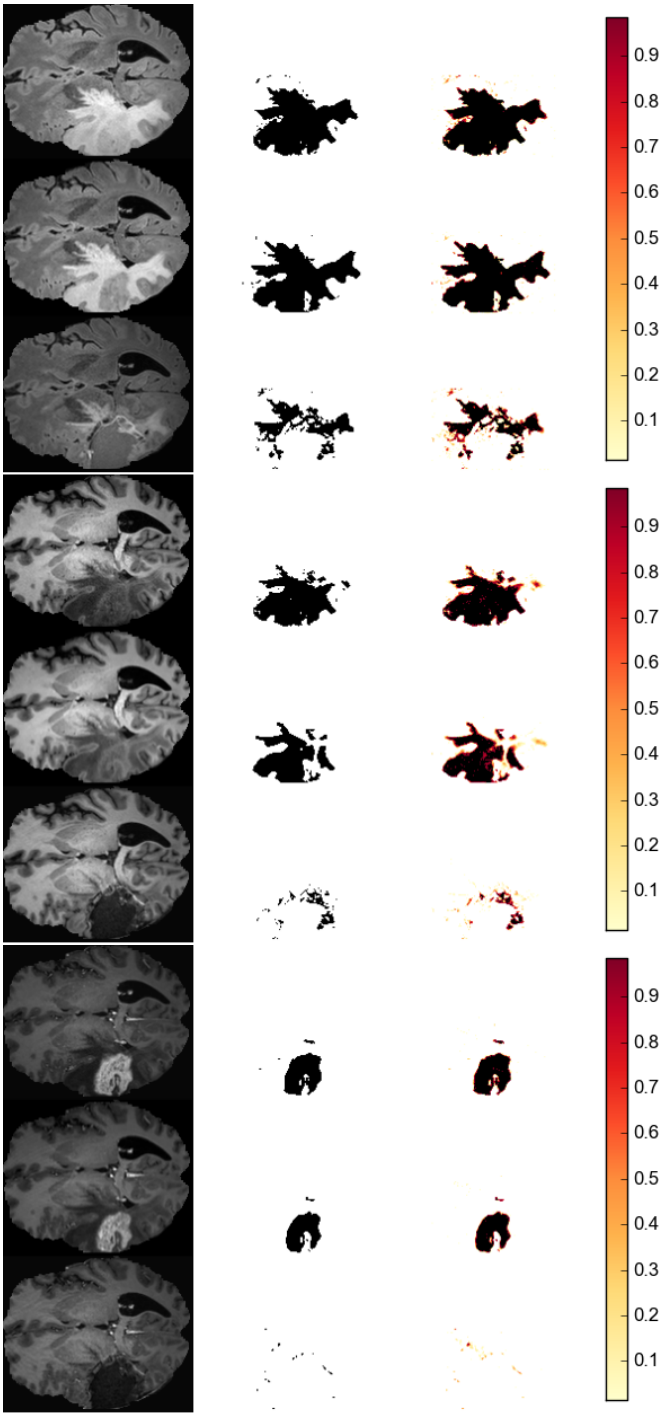


Figure 5.2: Patient-specific dataset containing FLAIR (top), T1 (middle) and T1c (bottom) modalities, each available for 3 time points (*top to bottom*). For each modality, input data (*left*), hard segmentations, i.e. MAP solutions, (*middle*) and soft segmentations, i.e. marginals, resulting from voxel-specific perturbations (*right*) are shown.

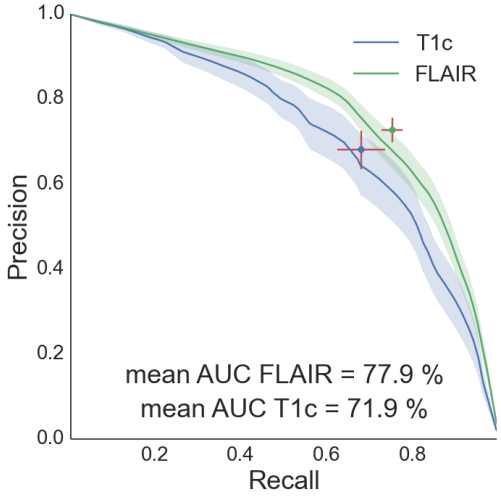


Figure 5.3: 68% confidence intervals of the mean precision-recall curve (*blue* and *green*) and the mean MAP solution (*red*).

the confidence bands of the mean precision-recall curve for all T1c and FLAIR segmentation tasks (ground truth was not available for T1), together with the confidence intervals of precision and recall scores for the mean MAP solution of all segmentation tasks. The mean area under curve (AUC) is equal to 71.9% for T1c and 77.9% for FLAIR. For the soft segmentations acquired by the supervoxel-specific perturbation model, the mean AUC is equal to 71.6% for T1c and 70.1% for FLAIR, indicating an overall decrease in segmentation accuracy, especially in FLAIR.

5.3.2 Qualitative evaluation

Soft segmentation maps resulting from the voxel-specific perturbation model are illustrated in Figure 5.2 for one patient-specific dataset containing several time points. Figure 5.4 further illustrates the benefit of soft segmentation maps for a single time point for three other patients.

Furthermore, we extended the voxel-specific segmentation uncertainties towards volumetric uncertainties. Figure 5.5 illustrates T1c and FLAIR segmentations (from two different patients) over 3 time points, together with the tumor volumetry uncertainties. The T1c segmentations show a low variation in volumes across samples for the first time point (where a very clear active tumor rim is visible), while the second time point shows quite a high variation (where the tumor was resected and false positives are present in scar tissue and vessels). The third time point depicts a very clear decrease in active tumor core volume. The FLAIR segmentations show quite high volumetric uncertainties along time, which might be related to the smooth tumor boundaries.

Soft segmentations resulting from supervoxel-specific perturbations are illustrated in Figure 5.6. They generally depict larger areas of uncertainty (lower and upper row in Figure 5.6) and, in some cases, reflect the underlying structure of the input data better (middle row in Figure 5.6).

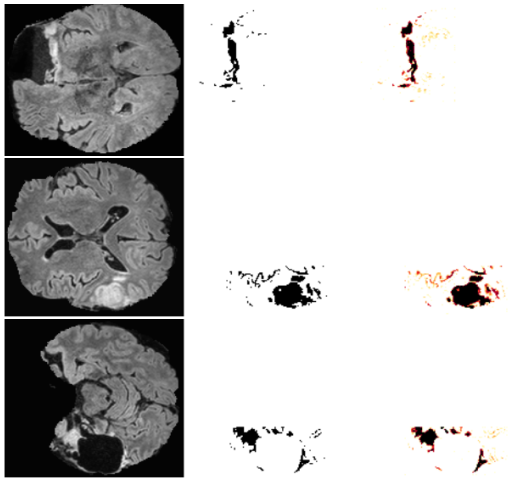


Figure 5.4: Hard and soft edema tumor segmentations resulting from voxel–specific perturbations. *Left to right*: FLAIR images, hard segmentations (MAP solutions) and soft segmentations.

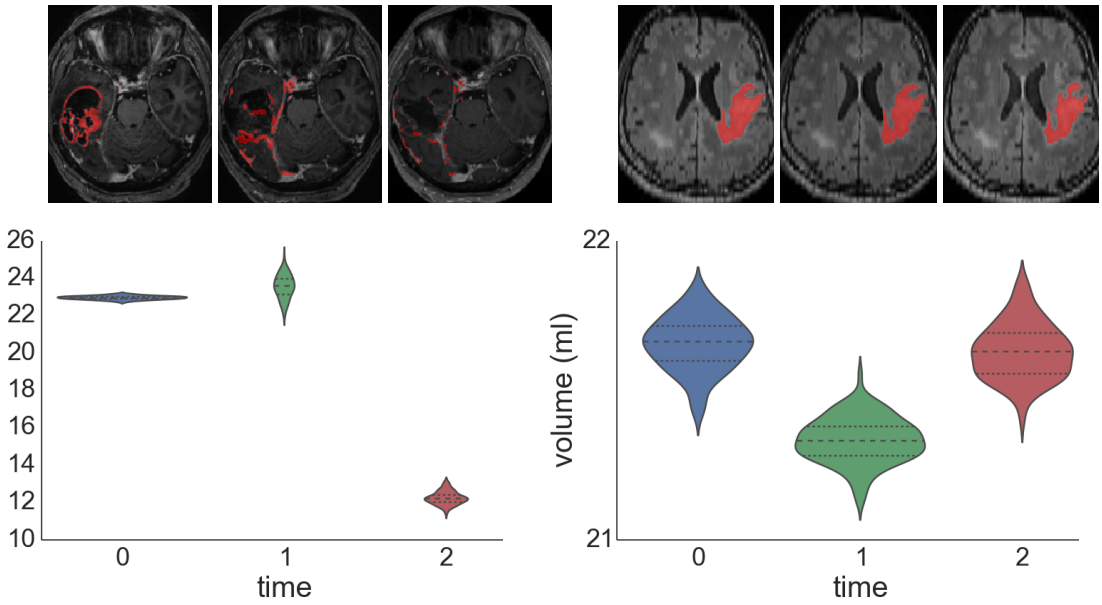


Figure 5.5: Uncertainty in brain tumor volumetry. The top row visualizes the MAP solutions and the plots show the distribution of volumes calculated for the samples obtained by the perturbation model. *Left:* T1C tumor segmentations depicting similar volumes for time point 1 and 2, but different levels of uncertainty (the MAP solution for time point 1, with a low uncertainty level, correctly identifies active tumor, whereas the segmentation for time point 2 is much less accurate). *Right:* FLAIR tumor segmentations with high volumetric uncertainties along time, although the segmentations seem to correctly delineate edema. This can be related to the smooth tumor boundaries.

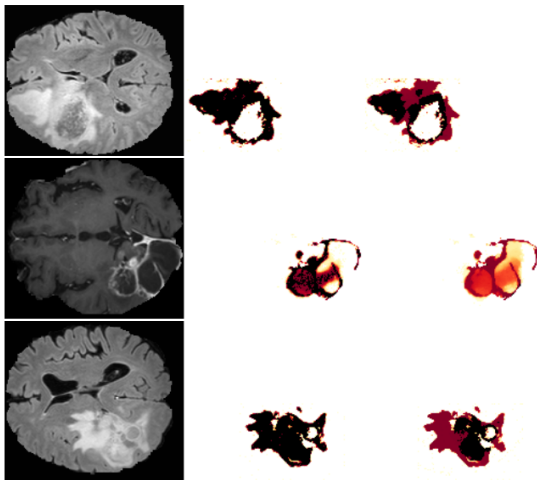


Figure 5.6: Soft segmentations resulting from voxel- and supervoxel-specific perturbations. *Left to right*: FLAIR or T1c image, soft tumor segmentation resulting from voxel-specific perturbations (middle) and from supervoxel-specific perturbations (rightmost).

5.4 Discussion and conclusion

This work has demonstrated the benefits of probabilistic segmentations in a CRF framework for brain tumor segmentation. Recent work on perturbation models has shown to be well-suited for obtaining samples of segmentations from the CRF model. It introduces minimal overhead and can be applied to virtually all segmentation approaches that rely on graphical models and energy minimization schemes.

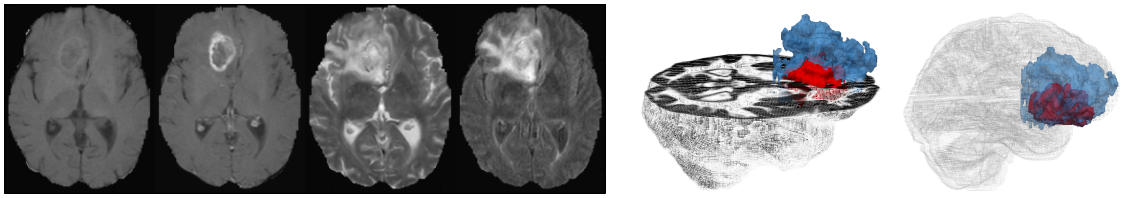
6 Image Classification using Low-Dimensional Texture Features for Genomic Brain Tumor Recognition

Chapter 4 and Chapter 5 are both built on brain tumor segmentation frameworks. Chapter 4 further extends towards longitudinal frameworks enabling tumor growth assessment and Chapter 5 provides a framework to assess tumor segmentation uncertainties. This chapter explores another type of computer vision task, shifting from image segmentation towards image classification. As seen in the previous chapters, in image segmentation, labels need to be assigned to each of the voxels in the image grid. In image classification, labels need to be assigned to images – not to voxels. Because of the high-dimensional feature space (i.e. multiple three-dimensional images), this task can be particularly challenging. In this chapter, a framework is presented to identify tumor types, based on multi-modal MR sequences. This work has been published in [11], and is here only adapted to a limited extent. The framework that is here presented has also been used to test image classification based on diffusion MR images, published in [12].

6.1 Introduction

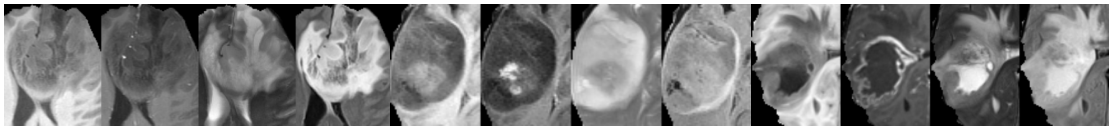
Currently the most aggressive brain tumors are categorized in World Health Organization (WHO) Grade III (anaplastic gliomas) and Grade IV (glioblastomas). [101] proposes to group brain tumors in three genomic classes based on DNA methylation patterns: CIMP Codel, CIMP Non-Codel and CIMP Neg. These genomic classes are based on the presence of IDH mutations and 1p/19q codeletion in the genomic profile of the tumor. Several studies have shown that the presence of these two genetic parameters have clear links with underlying biology and a significant association with clinical outcome [17], [18]. In this study, we aim to detect these genomic groups using image-based biomarkers such as tumor texture and shape features within standard MR modalities: T1, T1c, T2 and FLAIR. We segment active tumor and edema, and make use of the segmentation masks to generate tumor shape and texture features. Image modalities and segmentation masks are illustrated for one data sample in Figure 6.1. Examples of the tumor texture in MR images for tumors of the three genomic classes are depicted in Figure 6.2.

Predicting labels from image evidence is a hot topic in medical image analysis, studied with the purpose of disease detection [88, 82, 89], genetic labeling [87] or outcome prediction [102]. Medical image classification methods differ in feature generation and



T1, T1c, T2 and FLAIR

Figure 6.1: Illustration of input data for one data sample i. The four image modalities, $\{I_i^m\}_{m=1}^4$, are shown to the left and the two segmented ROIs for active tumor (•) and edema (•), $\{S_i^k\}_{k=1}^2$, are shown to the right.



CIMP Codel CIMP Non-Codel CIMP Negative

Figure 6.2: Illustration of tumor texture for three different genetic tumor classes. From left to right: Close-up of tumor texture in T1, T1c, T2 and FLAIR resp. for samples with genetic labels CIMP Codel, Non-Codel and Negative.

classification algorithms. Features reflecting image evidence are often implemented as texture and shape representations. *Classical texture descriptors* based on filter banks, Haar wavelets, Haralick features of gray-level co-occurrence matrices (GLCM) and histogram of oriented gradients (HOG) are still popular [88], but are known to suffer from normalization problems, lack of (local) rotation invariance, poor local organization of image directions and non-trivial 3D extensions. Wavelet and space-frequency representations, such as the ones generated by Riesz wavelets [82] and polar S-transforms [87] have been forwarded as an attempt to deal with these caveats. However, when working with multi-modal 3D medical images for which one needs to calculate, store and match a large number of descriptors, adopting *short descriptors* has prominent advantages in terms of memory-efficiency and computation time. Short descriptors can be generated from full descriptors (such as wavelet representations or traditional descriptors such as SIFT [103] and SURF [104]), by means of dimensionality reduction, quantization into integers or binarization. A computationally efficient alternative is to directly extract short descriptors as binary strings calculated over image patches, as proposed for local binary patterns (LBP) [83] and BRIEF features [84]. Next to classical texture descriptors, we want to consider another strong category of texture descriptors generated by *deep learning*. Neural nets have been shown to learn meaningful texture representations, either in supervised contexts (using convolutional neural networks (CNN)) or unsupervised contexts (using auto-encoders (AE), deep belief networks (DBN) or restricted Boltzmann machines (RBM)) [89].

For medical image classification, traditional classifiers typically obtain good accuracies, as shown in [82] using softmax classifiers, [87] using l1-regularized neural nets

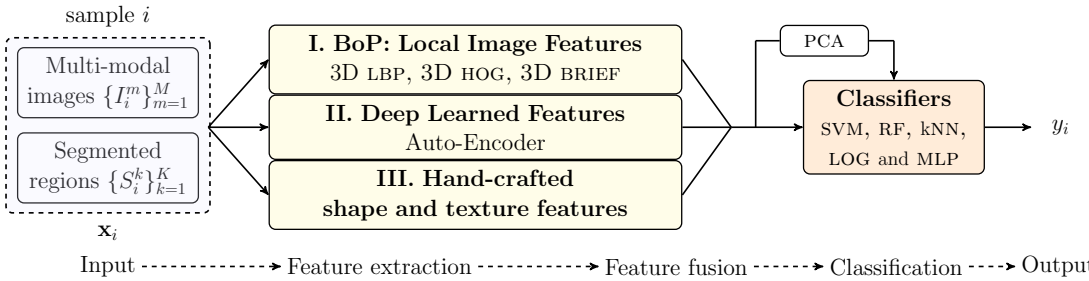


Figure 6.3: Overview of the proposed framework. Each sample consist of several image modalities and segmented regions. Features are generated by a Bag-of-Patterns (BoP) model performed on local image features, a deep learned auto-encoder and traditional shape and texture features. Feature dimension reduction through principal component analysis (PCA) is investigated before the features are being fed to the classifiers.

and [88] using random forests (RF). The high-dimensional nature of the feature space p and the sparse availability of data samples N, denoted as $p \gg N$, makes the classification task particularly challenging, leading to overfitting. Therefore, most studies include feature dimensionality reduction or specialized task-dependent feature selection models [88, 102, 105] in order to facilitate the classification task.

In this paper, we build a multi-modal image classifier making use of traditional hand-crafted features, modern local image texture features and features learned through a deep neural network. More in particular, we adopt a mathematically sound extension of LBPs towards 3D based on spherical harmonics [85], as well as a 3D version of the BRIEF [63] and HOG [106] features. In parallel, inspired by the success of deep learning in texture analysis [89], we also include features learned by unsupervised auto-encoders (AEs). We validate the framework on a brain tumor dataset with different genetic classes, using several state-of-the-art classifiers in conjunction with feature dimensionality reduction by means of principal component analysis (PCA).

6.2 Methods

An overview of the framework is depicted in Figure 6.3. An input sample consists of T1, T1 contrast-enhanced (T1c), T2 and FLAIR MR images as well as segmentation masks for active tumor and edema (cf. Figure 6.1). In this section, we discuss the feature extraction based on local image features (I.) and deep learned features (II.). Hand-crafted shape and texture features (III.) are directly calculated on segmented regions of interest (ROIs) using criteria such as volume, surface, flatness, roundness and elongation and on the image intensities within the ROIs using the first four statistical moments (mean, variance, skewness and kurtosis) for each of the image modalities.

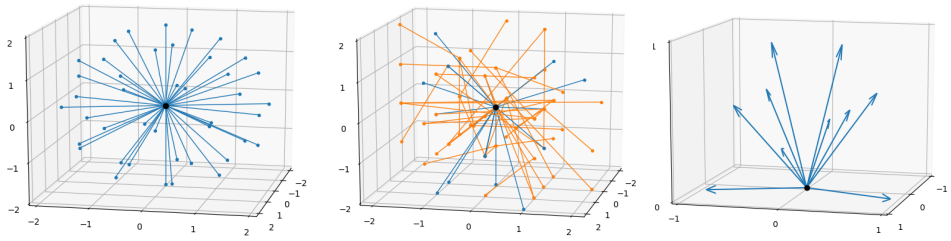


Figure 6.4: Sampled offset locations in the voxel grid for the LBP (left) and BRIEF (middle) features ($r = 2\,\text{mm}$) and HOG gradient orientation quantization (right). Binary intensity comparisons are made between offset locations and the center (blue lines) and among pairs of offset locations (orange lines).

Notation. Given a training set of N samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, with $\mathbf{x}_i = \{I_i^m|_{m=1}^{M}, S_i^k|_{k=1}^{K}\}$ the image data consisting of M image modalities and K segmentation ROIs, and $y_i \in \{1, \ldots, C\}$ the labels associated with each sample.

6.2.1 Bag-of-Patterns: from Local Image Features to Image Region Descriptors

We investigate three types of *local image features*: LBP, BRIEF and HOG. To compute the LBP features, 42 offsets are taken on an icosahedron of radius r centered around the center voxel (Figure 6.4, left), each offset denoted by its spherical coordinates (θ, ϕ). Next, binary intensity comparisons $f(\theta, \phi) \in \{0, 1\}$ are made between the discrete offsets (θ, ϕ) and the center voxel, and a continuous approximation $\tilde{f}(\theta, \phi)$ is made by a linear combination of spherical harmonics Y_l^m of degree l and order m: $\tilde{f}(\theta, \phi) = \sum_{l=1}^{n} \sum_{m=-l}^{l} c_l^m Y_l^m(\theta, \phi)$. The 3D rotationally invariant LBP features are then set to $(\|f_0\|, \ldots, \|f_n\|, k)$ with $\|f_l\| = (\sum_\theta \sum_\phi (\sum_{m=-l}^{l} c_l^m Y_l^m(\theta, \phi))^2)^{1/2}$ and k the kurtosis over the intensities. We use spherical harmonics up to degree $n = 3$. To compute the BRIEF features, images where smoothed with a Gaussian kernel and 64 binary intensity comparisons are made around the center voxel, including comparisons between pairs of offset locations (Figure 6.4, middle). A binary string is constructed from the results of the binary comparisons and is stored as a 64-bit float. Next to the binary descriptors, we explore histogram of oriented gradients (HOG) features as presented in [106]. We quantized gradient orientations on an icosahedron, categorizing opposite directions in one bin, which leaves us with 10 gradient orientations as visualized in the rightmost plot in Figure 6.4. Gradients are first calculated along the three dimensions using point derivative discrete masks and are then projected on the 10 aforementioned gradient orientations to get the local image descriptor.

Local image features are calculated over image patches centered at voxel locations. We use a Bag-of-Patterns (BoP) model to generate *image region descriptors*, describing texture over image regions of arbitrary size. Analogous to the Bag-of-Words (BoW) model in computer vision [107], the BoP model consists of two steps: first, a *codebook of patterns* is learned and secondly, a *bag of visual patterns* is calculated over an

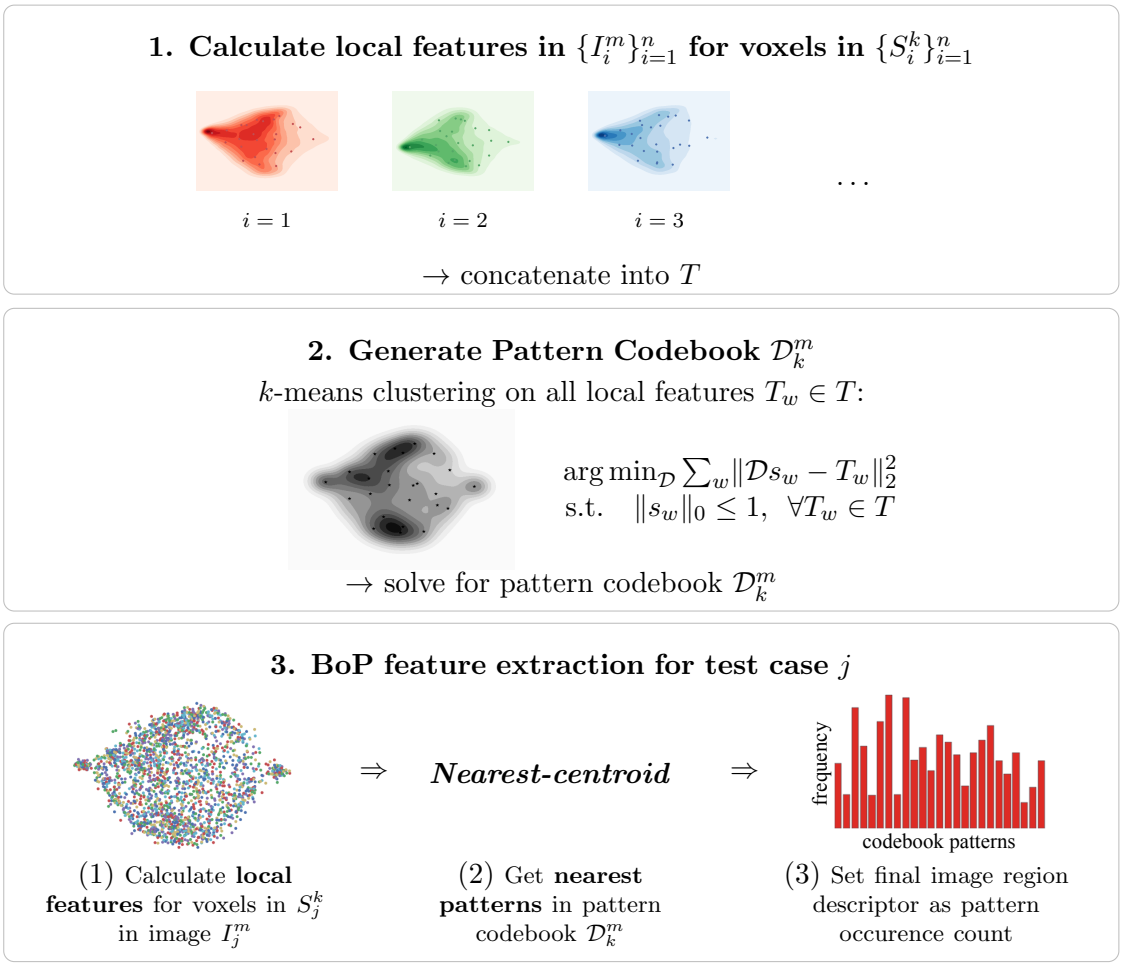


Figure 6.5: Generation of image region descriptors based on the BoP model. Local image features are calculated on n images of modality m for all voxels j in a ROI $S_i^k(j) > 0$. A codebook pattern is then generated by means of k-means clustering. The image region descriptor for an (unseen) image i is set as the codebook pattern occurrence count predicted over its local image features.

image region, as visualized in Figure 6.5. The patterns in the codebook are learned, for example by means of dictionary learning, on a (large) set of local image features acquired over several images. Each of the codebook patterns represent a subspace of the local image feature space. The bag of visual patterns can be obtained by collecting the closest pattern for each of the local image features within an image region. The image region descriptor is then set as the occurrence count of each of the patterns in the bag of visual patterns and is fed to the classifier.

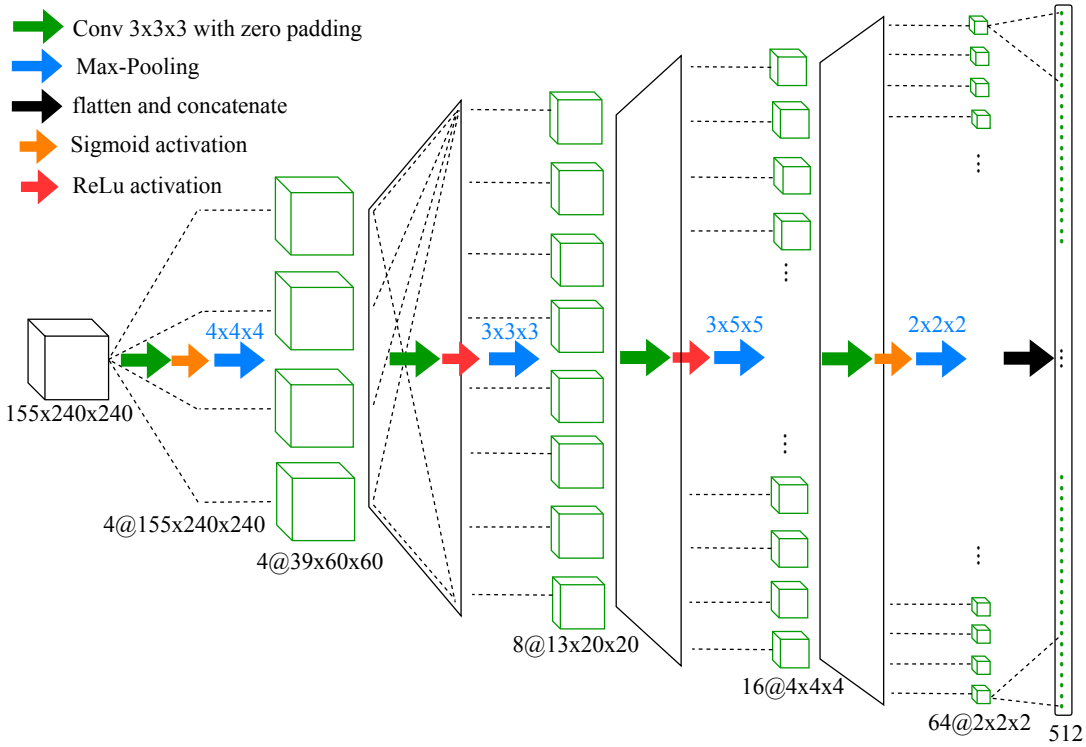


Figure 6.6: Network architecture of the auto-encoder: the encoder part and its 4 hidden layers are shown, generating a feature vector of 512 features for an input volume of $155 \times 240 \times 240$. Each hidden layer consists of a 3D convolutional layer with $3 \times 3 \times 3$ kernels applied with zero padding, an activation layer with either the Sigmoid or the Rectified Linear (ReLu) activation function and a pooling layer with indicated size and stride. Feature maps are visualized as green cubes (their dimensions are indicated below them) and the final features as green dots.

6.2.2 Unsupervised Feature Learning Based on Deep Auto-Encoders

An auto-encoder (AE) is an unsupervised deep neural network used for non-linear dimensionality reduction, image reconstruction or image denoising. By means of stacked convolutional, activation and pooling layers, the input image in space $\mathcal{X}$ is reduced to a feature vector in space $\mathcal{F}$. This part of the AE is denoted as the encoder: $\phi : \mathcal{X} \to \mathcal{F}$, whereas the decoder maps the feature vector back to the input image space: $\psi : \mathcal{F} \to \mathcal{X}$. The filters in the convolutional layers of the AE architecture are learned trough back-propagation of the reconstruction error $\mathcal{L}(I_i^m, \tilde{I_i^m})$, defined between the original image I_i^m and the reconstructed image $\tilde{I}_i^m = (\psi \circ \phi)(I_i^m)$.

Network Architecture. Let $\texttt{C}_\texttt{k}$ denote a $(3 \times 3 \times 3)$ convolutional layer with $\texttt{k}$ filter banks, $\texttt{S}$ a sigmoid activation layer, $\texttt{R}$ a ReLu activation layer and $\texttt{P}_{\texttt{s}_1\texttt{s}_2\texttt{s}_3}$ a pooling layer with $(\texttt{s}_1 \times \texttt{s}_2 \times \texttt{s}_3)$ filters. We implemented an AE of four layers, which we can then write as: $\texttt{C}_4\texttt{SP}_{444}$ - $\texttt{C}_8\texttt{RP}_{333}$ - $\texttt{C}_{16}\texttt{RP}_{355}$ - $\texttt{C}_{64}\texttt{SP}_{333}$. Taking a 3D volume with

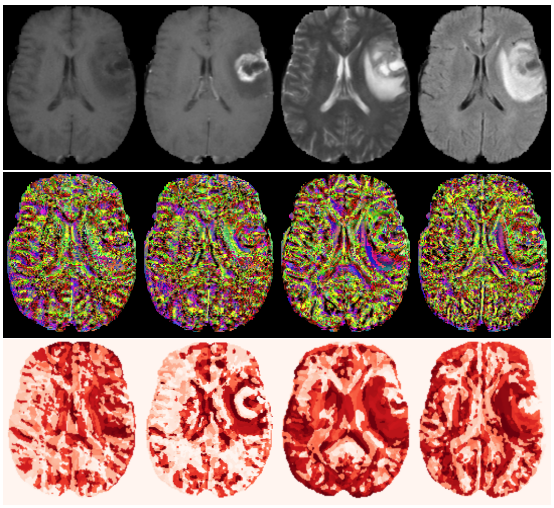


Figure 6.7: Local image features ($r = 5\,\text{mm}$) calculated in the full brain mask for modalities T1, T1c, T2 and FLAIR (top). LBP features are visualised as RGB channels over the first three LBP features corresponding to the order of the spherical harmonics: $[\|f_0\|, \|f_1\|, \|f_2\|]$ (middle row). BRIEF features are visualised as their 64-bit float representation.

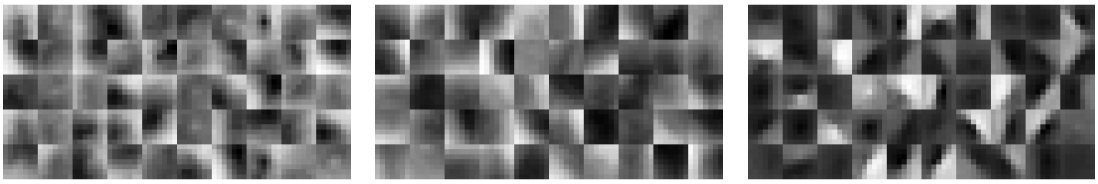
fixed dimensions as an input, the AE generates a feature vector of length 512. The network architecture and the dimensions of the kernels and feature maps for each layer are illustrated in Figure 6.6.

Implementation Details. The AE is trained with tied weights and input images are corrupted with noise in order to force the AE to generate features that are robustly representing the input distribution, similar to denoising AE [108]. Moreover, during training, the reconstruction error is only calculated within the segmentation masks, as we are only interested in reconstructing features representing texture within the segmentation masks.

6.3 Experiments

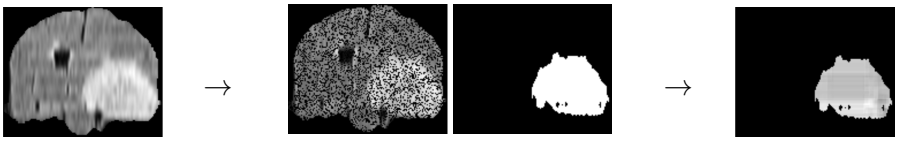
We conduct experiments on a brain tumor dataset from "The Cancer Imaging Archive" (TCIA). For each patient genetic labels, CIMP Codel, CIMP Non-Codel or CIMP Negative, are available. These genetic classes are originally acquired by biopsies. Medical experts are not able to differentiate between these classes solely based on image evidence, making the classification task particularly challenging. We constructed a training set of 92 patients and a test set of 24 patients by means of stratified sampling. The genetic classes are rather unbalanced with 19, 25 and 48 respective label occurrences in the training dataset and 5, 7 and 12 in the test set. T1, T1c, T2 and FLAIR are skull stripped, co-registered and resampled to the same resolution and image dimensions. Brain tumor segmentations for whole tumor and active tumor are calculated using a brain tumor Expectation-Maximization algorithm [54] for which the code is online available[1].

[1] `https://bitbucket.org/s0216660/brain_tumor_segmentation_em`



LBP BRIEF HOG

Figure 6.8: Pattern codebooks learned over 50 FLAIR images within the whole tumor regions for the three local image features. Patterns are visualized by averaging patches (⌀7 mm) collected in the normalized FLAIR images for which the local image features of the center voxel are nearest to the pattern centroids (2D cross-sections are shown).



FLAIR AE input AE reconstruction

Figure 6.9: Auto-encoder reconstruction results learned on FLAIR and whole tumor masks. Inputs are corrupted by a noise level of 30%.

LBP, BRIEF and HOG feature generation. LBP and BRIEF features are generated for radii of $2, 3, 4$ and 5 mm. Figure 6.7 visualizes LBP and BRIEF features calculated over the entire brain region. Using 50 patients randomly sampled from the training set, pattern codebooks are learned for each modality for LBP, BRIEF and HOG. Only whole tumor masks were used, because active tumor regions do not contain much texture information. Image patches representing the codebook patterns learned for LBP, BRIEF and HOG in FLAIR images are visualized in Figure 6.8. Each codebook consists of 50 patterns, resulting in image region descriptors of length $4 \times 50 = 200$ of for each local image feature type. Average runtimes[2] for LBP, BRIEF and HOG feature calculation for one sample $\mathbf{x}_i$ (including all four modalities masked over the whole tumor region) are about 81, 5 and 9 s.

AE feature generation. We train AEs on FLAIR and on T1c images separately, each time using the whole tumor masks when calculating the reconstruction error. We denote the concatenation of the features generated by these two AEs as $\text{AE}_{\text{texture}}$. A third AE, generating features denoted as AE_{shape}, is trained on ternary tumor mask images (background, edema and active tumor encoded as 0, 1 and 2 resp.) to learn tumor shape features. Training was done on 50 patients randomly sampled from the training set. Each AE was randomly initialized and ran during 200 iterations. The reconstruction error is set to the mean squared error and 30% of the voxels in each input volume were corrupted (setting voxel intensities to zero). An example of the

[2] Intel® Xeon® Processor E3-1225 v3

Table 6.1: Three-class classification accuracies (training/test score) acquired with different sets of features (columns) using different classifiers (rows). For comparison, a majority-vote classifier would obtain a test score of **0.50**. Classifiers include nearest neighbor (NN), logistic regression (LOG), multi-layer perceptron (MLP), random forests (RF) and support vector machines (SVM), optionally in combination with principal component analysis (PCA).

train/test	LBP	BRIEF	HOG	shape	$AE_{texture}$	AE_{shape}	All
k-NN	0.66/0.50	0.66/0.50	0.66/0.50	0.70/0.42	0.74/**0.75**	0.74/**0.75**	0.66/0.46
- w/ PCA	0.72/0.50	0.72/0.50	0.72/0.50	0.72/0.38	0.73/**0.75**	0.73/**0.71**	0.66/0.42
LOG	1.00/0.67	1.00/0.67	1.00/0.67	0.65/0.50	0.64/0.63	0.55/0.54	1.00/0.71
- w/ PCA	0.99/**0.75**	0.99/**0.75**	1.00/**0.75**	0.70/0.38	0.64/0.63	0.55/0.54	0.89/0.63
MLP	1.00/0.67	0.92/0.54	1.00/0.63	0.74/0.33	0.69/0.63	0.52/0.50	0.63/0.42
- w/ PCA	1.00/0.63	1.00/0.67	1.00/0.63	0.46/0.42	0.64/0.54	0.94/0.67	0.73/0.42
RF	1.00/0.58	1.00/0.67	1.00/0.63	1.00/0.42	0.97/0.63	1.00/0.67	1.00/**0.83**
- w/ PCA	1.00/0.54	0.99/0.50	1.00/0.54	1.00/0.50	1.00/**0.79**	1.00/0.58	1.00/0.54
SVM	1.00/0.67	1.00/0.67	1.00/0.67	0.54/0.33	0.52/0.50	0.52/0.50	1.00/0.67

Table 6.2: Confusion matrix acquired on the test set of 24 samples, when using the best estimator setting (using all features and RF classification). The rows contain indicate the predicted labels $\hat{y}$ whereas the columns indicate the true labels y.

$\hat{y} \downarrow y \rightarrow$	CIMP Codel	CIMP Non-Codel	CIMP Neg
CIMP Codel	12	1	1
CIMP Non-Codel	0	5	1
CIMP Neg	0	1	3

input and reconstruction of an AE trained on FLAIR images is shown in Figure 6.9. Average runtime[3] for the AE feature calculation (including $AE_{texture}$ and AE_{shape}) for one sample $\mathbf{x}_i$ is $< 1\,\text{s}$.

Classification results. Table 6.1 shows the results of all classifiers using different feature sets. For each classifier and feature set, classifier-specific parameters were learned by cross-validation on the training set using grid search. Best accuracies on the test set are obtained by using all features together with random forests (0.83). LBP, BRIEF or HOG features do remarkably well in combination with PCA and logistic regression (0.75), while the AE features perform well in combination with k-NN $(0.71, 0.75)$ or with random forests in combination with PCA (0.75).

Best estimator. According to Table 6.1, the best estimator is the RF classifier trained on all extracted features. Table 6.2 shows the confusion matrix acquired on the test

[3]GeForce GTX Titan X (VRAM 12 GB)

Table 6.3: Feature importances of the RF classifier trained on all features (best estimator).

LBP	23.69%
BRIEF	23.00%
HOG	23.17%
shape	0.72%
$AE_{texture}$	19.73%
AE_{shape}	9.68%

set using the best estimator. All CIMP Codel samples have been correctly identified. Two CIMP Non-Codel samples have been falsely classified (one as CIMP Codel and the other as CIMP Neg) and two CIMP Neg samples have been falsely classified (one as CIMP Codel and one as CIMP Non-Codel). Since the best estimator is a RF classifier, feature importances can be evaluated by averaging the impurity decrease that each feature establishes in the trained trees. The feature importances attributed to all LBP, BRIEF, HOG, shape and AE features are listed in Table 6.3. The shape features seem to have the least importance, whereas the LBP, BRIEF, HOG and AE features seem to have similar importances.

6.4 Conclusion

In this study, we present a multi-modal medical image classifier making use of modern 3D implementations of LBP, BRIEF and HOG features, as well as deep learned auto-encoder features. The framework is validated against a highly difficult classification task: the classification of brain tumors in genetic classes. Although medical experts are not able to differentiate between the genetic classes using only image evidence, the presented framework obtained an accuracy of 0.83 when using all features and RF classification.

7 Multi-modal Brain Tumor Segmentation using High-Dimensional Clustering with Copulas and Supervoxel Classifiers

This chapter presents a novel and flexible method to segment brain tumors in multi-modal magnetic resonance images. Segmentation is performed in a multi-scale fashion, moving between a coarse supervoxel grid and a fine voxel grid. On the voxel grid, a probabilistic model based on Bayesian clustering is implemented, optimized by means of the Expectation-Maximization algorithm (similar to the probabilistic framework adopted in Chapter 4 and 5). In this chapter, the likelihood function in the probabilistic model is modeled with the use of Gaussian copulas. By using Gaussian copulas, we tailor the probabilistic model to effectively capture the dependencies between the intensities in all input modalities, while being able to cope with non-Gaussian univariate intensity distributions in each input modality. On the supervoxel grid, we train random forests (RF): first to generate robust label prior maps before the probabilistic model and secondly, for regularization purposes, after the probabilistic model to ensuring spatial coherence and consistent label co-occurrence patterns. Recall that, in Chapter 4 and 5 a CRF was used for this purpose. Finally, two practical semi-automated extensions are presented, allowing a user to interact with the segmentation process. The method is shown to generate accurate image segmentations, even when not all modalities are at hand or in complex cases containing resection cavities, and the semi-automated extensions are shown to be extremely valuable tools for clinical practice. The work presented in this chapter has been submitted for publication.

7.1 Introduction

Medical image segmentation is emerging towards an indispensable tool in clinical practice. Brain tumor segmentation is of high value both in clinical research and clinical practice, aiding in diagnosis, treatment planning and follow-up assessments ([2]). Gliomas are the most common type of primary malignant brain tumors, often classified into low-grade gliomas (oligodendrogliomas, oligoastrocytomas or astrocytomas) and more aggressive high-grade gliomas, such as glioblastomas ([1]). Depending on their genotype, gliomas are usually treated with immediate surgical treatment followed with radio- and/or chemotherapy ([9]). Post-operative magnetic resonance (MR) scans with resection cavities and possibly internal bleedings, are typically very challenging segmentation cases which have received little attention, although they are very common in clinical practice. Volumetric measures quantifying disease progression (i.e. tumor

growth or tumor recurrence after resection) or treatment success (i.e. tumor shrinking in response to therapy) are fundamental indicators for further treatment planning, both for treated and untreated low- and high-grade glioma cases ([109]). In general, robust, reliable and flexible segmentation models are currently still in high demand for several clinical applications.

Strongly supervised models have become very successful in medical image segmentation challenges, which typically offer training datasets growing larger and more accurate every year ([2]). The current top state-of-the-art brain tumor segmentation methods almost exclusively consist of deep neural network implementations: [71, 72, 75, 76, 77]. Convolutional neural networks (CNN) learn feature maps, by having several cascading convolutional layers on an image-like input, where the receptive fields of the features grow larger with every deeper layer. Having features learned automatically on different scales of the image in an end-to-end matter, seems to play a key role in the success of deep learning architectures ([74]). The first CNN-based brain tumor segmentation methods were acting on a patch-level: [71] classifies voxels based on 2D multi-scale patches and [72] classifies 3D patches followed by a fully-connected conditional random field (CRF). Later on, fully-convolutional CNNs acting on entire MR images were introduced: [75] proposed a multi-task setup to create a boundary-aware network, [76] proposed a fully-connected 3D residual network and [77] proposed a scalable network allowing to efficiently refine the network to take an arbitrary number of MR input modalities. Although deep learning implementations are amongst the top-ranked methods in most medical segmentation challenges, they remain dependent on the size and the quality of the training dataset, on the set of input modalities the models are trained on and on the computational resources that are available. In this paper, we investigate an alternative method based on generative probabilistic models, aiming to obtain good segmentation accuracies while being able to accommodate for missing modalities and offering strong robustness in terms of training data size. Moreover, we aim to build a model that allows for semi-automated extensions, such that a user can guide the segmentation process if needed.

7.1.1 Generative Probabilistic Models

In comparison to strongly supervised models, generative probabilistic models offer a few interesting flexibilities. They have good generalization performance, generalizing well to test images acquired with different imaging protocols ([54]). As they learn local intensity variation during the segmentation process, they do not require any image intensity calibration, which is known to be very cumbersome for MR images, especially in the presence of lesions ([50]). Furthermore, they are often flexible in terms of input modalities and do not have to be retrained from scratch when modalities are missing or altered. This is an important property for clinical practice, as acquiring complete multi-modal imaging data might be too costly or time consuming. The importance of the missing modality problem has been addressed in several studies that developed strongly supervised methods that are able to cope with missing modalities: [77] proposes to use *scalable* multi-modal CNNs, [110] developed a *Hetero-modal* Im-

age Segmentation (HeMIS) framework and [93] implemented *scandant* trees. Lastly, generative probabilistic models tend to be cheaper in terms of computation time and resources, even though deep learning test phase computation times are hard to beat, if the appropriate resources are available.

Probabilistic models optimized by the Expectation-Maximization (EM) algorithm have been proposed for several medical applications, with a strong prevalence in brain analysis: cf. [53, 111] for brain tissue segmentation and brain lesion segmentation, [8, 54] for multi-modal brain tumor segmentation and [112] for multi-modal glioma segmentation, registration and tumor growth modeling (extended with gradient boosting machines in [113]).

One of the major difficulties with EM optimization is its tendency to get stuck in local minima, which often finds its cause in poor initializations or inadequate priors ([114]). Flat priors, in which all voxels have an equal prior class probability, will almost always converge to local minima for medical segmentation applications. Instead, medical image atlases, such as the SRI24 Atlas developed by [3], can offer stable tissue-specific location priors. In atlas-based methods, priors are generated by registering the population atlases to subject space, and are commonly also used to initialize the tissue-specific segmentations in the EM algorithm. They can only be generated for a number of healthy tissue classes, such as white matter (WM), grey matter (GM) and cerebro-spinal fluid (CSF), for which population atlases are available. Priors for tissue classes characterized by irregular locations, extensions or intensity patterns, such as brain tumors and other pathological tissues, cannot be directly generated by registration of a population atlas. In [112, 113], a diffusion-reaction tumor growth model is incorporated in the EM iterations, to adopt a synthetically generated tumor probability map as well as a patient-specific atlas. Other generative methods ([54, 115]) combine atlas-based initialization for healthy tissues with modality-specific outlier-based initializations for pathological tissues.

Furthermore, generative probabilistic models developed for segmentation purposes usually require a regularization step, for example to enforce spatial consistency in the segmentation results. In [113], for example, gradient boosting machines are used to refine the tumor labels generated by a generative model. In [54] a full range of discriminative extensions to generative probabilistic models is studied, aiming to generate more spatially consistent segmentations. Regularization is also common for strongly supervised models, for example by applying a CRF on the segmentation results ([72, 116]).

An important modeling choice in generative probabilistic models is the selection of appropriate multivariate distributions to model the joint data likelihood over all input modalities. In [111], the joint probability density of the input modalities is modeled as a product of marginal univariate Gaussian densities, one for each modality and each label, assuming that the input modalities are statistically independent. In [8, 54], efforts have been made to break this pattern, modeling the dependencies between input modalities by heuristically hard-coding certain tumor co-occurrence patterns along the input modalities. Active tumor, for example, is hard-coded to be hyper-intense in T1 with contrast (T1c) and hypo-intense in T1. Another common approach to model the dependency between modalities is to use multivariate Gaussians in the data likelihood,

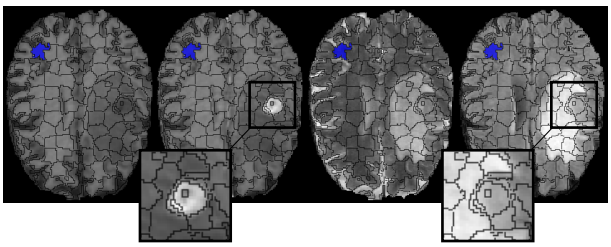


Figure 7.1: Illustration of supervoxels calculated on a multi-modal sequence (T1, T1c, FLAIR and T2) using SLIC, with $n = 20000$, the number of segments, and compactness $c = 0.1$.

as originally proposed in [53]. However, unless several multivariate Gaussians are used to model one tissue class (as in mixture models), one is not able to accurately model asymmetric outlier distributions, such as tumor intensity distributions. In mixture models, on the other hand, the number of mixture components needs to be estimated or hard-coded, and *several* covariance matrices need to be calculated *for each* label, accordingly. The parameters grow linearly in size with the number of components, and the problem oftentimes becomes ill-defined and slow to solve. In general, modeling the joint data likelihood is usually done by a combination of (multivariate) Gaussian densities, which remains insufficient when modeling non-Gaussian distributions and is inefficient for high-dimensional input spaces.

To summarize, probabilistic models have important advantages, but come with a few difficulties: their performance depends on robust and adequate priors, they need a good regularization and they need to allow more flexibility in the choice of distribution used to model the joint probability density over the input variables, such that the dependencies in a high-dimensional input space are modeled in an efficient and transparent way.

7.1.2 Proposed method

In this paper, we propose to use a multi-scale approach, in which we iterate between a coarse supervoxel grid and a fine voxel grid. On the supervoxel grid, we train two classifiers: once to generate reliable label prior maps for the generative probabilistic model and once to regularize the segmentation maps of the probabilistic model. Supervoxels are atomic regions, consisting of voxels that have been aggregated based on similar appearance and spatial proximity (Figure 7.1). In contrast to the perceptive fields of the neurons in CNN architectures, supervoxels have perceptually meaningful shapes. We use simple linear iterative clustering (SLIC) to calculate the supervoxels directly on the multi-modal image sequence, as presented in [117], and propose to train random forests (RF), inspired by [62] and [118], to classify supervoxels. In the first classifier, aiming to generate label prior maps, supervoxels are classified based on their multi-modal regional image features. In the second classifier, designed to regularize the segmentation maps of the probabilistic model, supervoxels are classified based on their

regional image and regional label context features. On the fine voxel grid, we adopt a probabilistic model optimized by the EM algorithm, which we robustly initialize with the label prior maps generated by the RF supervoxel classification. By leveraging recent advances in statistical analysis, we introduce Gaussian copulas ([119]) to handle the dependencies between the numerous input channels in the probabilistic model. Copulas allow to model a joint density by a product of marginals and a dependence structure, captured by the copula density. They do not impose any restrictions on the marginals to be used, offering an unseen flexibility for the choice of distribution used to model the univariate densities. Their favorable properties, including a low-dimensional parameter space, have been reported for a few high-dimensional classification problems, for example in discriminant analysis ([120]) and Bayesian networks ([121]). Gaussian copulas in particular have been shown to be an efficient and robust alternative to classical multivariate Gaussians for estimating high-dimensional undirected graphical models, notably when the distributions of the data are non-Gaussian or noisy ([122]).

We present the multi-scale framework (Section 7.2) and address its keystones, starting with the probabilistic model (Section 7.2.1), in particular the incorporation of the Gaussian copulas in the joint density (Section 7.2.1.1) and its EM optimization (Section 7.2.1.2). We then present the supervoxel classifiers (Section 7.2.2) and offer two interactive extensions to the proposed method in order to allow for user-interaction on difficult segmentation cases (Section 7.2.3). In Section 7.3, we conduct experiments on brain tumor segmentation. We assess performance, dependence on training data and missing input modalities (Section 7.3.1), compare with the state-of-the-art (Section 7.3.2) and demonstrate clinical relevance by testing our method on a highly challenging clinical dataset, with missing modalities and including resection cavities (Section 7.3.3).

7.2 Methods

We implement a brain tumor segmentation framework that takes m input modality images and generates a segmentation image with $k = 7$ segmentation classes, including three healthy tissue classes (WM, GM, CSF) and four tumor classes ('enhanced tumor', 'non-active tumor', 'necrotic tumor' and 'edema'), conform with the BRATS dataset annotations [2]. The proposed framework consists of a probabilistic model and two supervoxel classifiers shown in Figure 7.2: (i) the *initializing supervoxel classifier* generates coarse segmentations on a supervoxel grid, adopted as a prior in (ii) the *generative probabilistic model*, which generates segmentations on a fine voxel grid and feeds them to (iii) the *label context supervoxel classifier*, which uses the voxel-level predictions of the probabilistic model to generate regularized segmentations, aiming to adopt spatial coherence and capture label co-occurrence patterns. We first elaborate on the probabilistic model, then explain the role and implementation of the supervoxel classifiers and finally offer two semi-automated extensions of the proposed method.

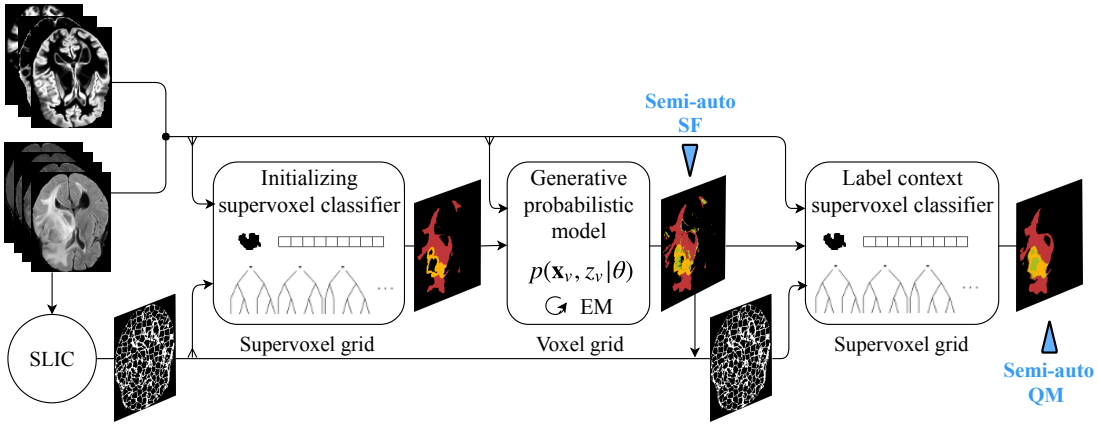


Figure 7.2: Overview of the proposed framework. The initializing supervoxel classifier generates a prior for the probabilistic model, which generates voxel-level predictions that are fed to the label context supervoxel classifier for spatial coherence. Image modalities and registered atlas tissues are fed to all segmenters. Supervoxels are generated based on SLIC and are fed to the two supervoxel classifiers. Tumor segmentations maps are visualized in color and the interactive extensions semi-auto SF and QM (Section 7.2.3) are indicated in blue.

7.2.1 Voxel grid: a Generative Probabilistic Model with EM Optimization

On the voxel grid, we adopt a generative probabilistic model tailored specifically to cope with non-Gaussian outlier distributions by incorporating Gaussian copulas. We first present the probabilistic model and its peculiarities for medical image segmentation, then elaborate on Gaussian copulas and how they are incorporated in the probabilistic model, and lastly present its EM optimization.

7.2.1.1 A Generative Probabilistic Model with Gaussian Copulas

Given a set of observed variables $\mathbf{x}$ and one hidden variable z, both indexed over a voxel grid $\mathcal{V}$; the observed variables $\mathbf{x}$ are the input data supplied as a set of m input modalities or feature maps: $\mathbf{x} = \{x_1 \dots x_m\}$, and the hidden variable z is the label map to be estimated with k possible labels in every voxel v: $z_v \in \{1 \dots k\}$, $\forall v \in \mathcal{V}$. For each label l, the generative model estimates the joint probability density function over the observed and the hidden variables in a voxel v:

$$p(\mathbf{x}_v, z_v{=}l|\boldsymbol{\theta}_l, \alpha_{v,l}) = p(\mathbf{x}_v|z_v{=}l, \boldsymbol{\theta}_l)p(z_v{=}l|\alpha_{v,l}) \ , \tag{7.1}$$

with $\boldsymbol{\theta}_l$ the model parameters for label l and $\alpha_{v,l}$ the prior for label l in voxel v. The conditional dependencies as factorized in this joint probability density function are expressed in a graphical model, depicted in Figure 7.3, which also illustrates how the variables are indexed. To get probabilistic label segmentations, we estimate the

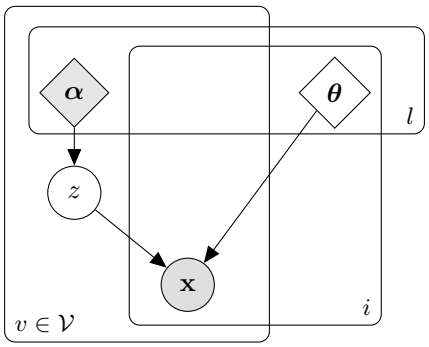


Figure 7.3: Bayesian network of the proposed generative model. The prior maps $\boldsymbol{\alpha}$ and the input maps $\mathbf{x}$ are observed, whereas the label map z and the parameters $\boldsymbol{\theta}$ are latent variables updated through iterative Bayesian inference (cf. Section 7.2.1.2). Plates denote the range of indices: the variables are indexed over a voxel space $\mathcal{V}$, the input modalities i and the segmentation labels l. Stochastic variables are encircled, whereas deterministic variables have a square shape.

posterior probabilities on the hidden variables:

$$p(z_v{=}l|\mathbf{x}_v, \boldsymbol{\theta}_l) = \frac{p(\mathbf{x}_v, z_v{=}l|\boldsymbol{\theta}_l)}{\sum_{l=1}^{k} p(\mathbf{x}_v, z_v{=}l|\boldsymbol{\theta}_l)} \quad . \tag{7.2}$$

All model decisions are made modeling the two factors on the right-hand side in (7.1). In existing implementations of generative models ([53, 111, 8, 54]), the observed variables are typically either modeled by means of multivariate Gaussians, $p(\mathbf{x}_v|z_v{=}l, \boldsymbol{\theta}_l) = \mathcal{N}(\mu_l, \boldsymbol{\Sigma}_l)$, or they are assumed to be independent, in which case they are modeled as a product of their marginals, $p(\mathbf{x}_v|z_v{=}l, \boldsymbol{\theta}_l) = \prod_{i=1}^{m} p(x_{v,i}|z_v{=}l, \boldsymbol{\theta}_l)$. In these works, the label priors, $p(z_v{=}l|\alpha_{v,l})$, are set to population atlases registered to subject space, or to a spatially smoothed version of the current label segmentation estimates, in order to attempt spatial coherence.

In our model, we make use of Gaussian copulas in $p(\mathbf{x}_v|z_v{=}l, \boldsymbol{\theta}_l)$, allowing to flexibly model dependencies between the input modalities. We first discuss the concept of copulas, then explain how they are integrated in $p(\mathbf{x}_v|z_v{=}l, \boldsymbol{\theta}_l)$ and finally explain how the prior $p(z_v{=}l|\alpha_{v,l})$ is modeled.

Copulas and Sklar's theorem. A copula C can be defined as a joint cumulative density function over d random variables with uniform marginals: $C : [0,1]^d \to [0,1]$. Their application is justified by Sklar's theorem, which states that for any joint cumulative density function $F(\mathbf{x})$, there exists a copula C that expresses F in terms of its univariate cumulative marginals, i.e. s.t. $F(\mathbf{x}) = C\big(F_1(x_1), \ldots, F_d(x_d)\big)$. It also holds that, if a joint probability density function $p(\mathbf{x})$ exists for $F(\mathbf{x})$, there exists a copula density c, such that the joint density $p(\mathbf{x})$ is equal to the product of its univariate marginals and the density of the copula c, expressed in terms of its univariate cumulative marginals $F_1(x_1), \ldots, F_d(x_d)$:

$$p(\mathbf{x}) = p_1(x_1) \cdots p_d(x_d) \cdot c\big(F_1(x_1), \ldots, F_d(x_d)\big) \quad . \tag{7.3}$$

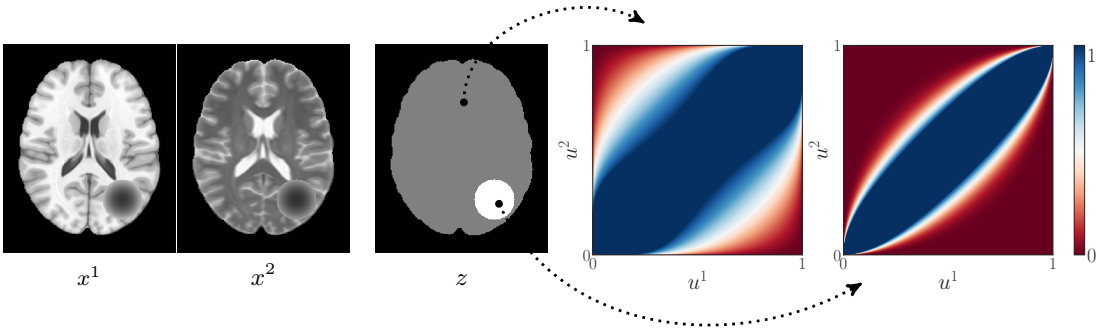


Figure 7.4: Illustration of Gaussian copula densities: two synthetic input images, a label map with healthy brain in grey and a simulated artifact in white, and the bivariate Gaussian copula densities across the two input channels for healthy brain (moderate correlation, $\rho = 0.70$) and for the artifact (strong correlation, $\rho = 0.97$), expressed in function of the cumulative densities in the input images, u^1 and u^2 (as defined in (7.5)).

Modeling the conditional joint density using Gaussian copulas. We now model the first factor in (7.1), i.e. the joint density of the observed variables $\mathbf{x}$ conditioned on the hidden variable z: $p(\mathbf{x}_v|z_v{=}l, \boldsymbol{\theta}_l)$. Based on Sklar's theorem in (7.3), we can write it as follows:

$$p(\mathbf{x}_v|z_v{=}l, \boldsymbol{\theta}_l) = \left[\prod_{i=1}^{m} p(x_{v,i}|z_v{=}l, \boldsymbol{\theta}'_{l,i})\right] \; c_{\boldsymbol{\theta}''_l}(u_{v,1}, \ldots, u_{v,m}) \;\; , \tag{7.4}$$

with $c_{\boldsymbol{\theta}''_l}(\cdot)$ the density of the copula and $u_{v,i}$ the univariate cumulative density function of the marginal probability density function $p(x_{v,i}|z_v{=}l, \boldsymbol{\theta}'_{l,i})$:

$$u_{v,i} = F(x_{v,i}|z_v{=}l, \boldsymbol{\theta}'_{l,i}) = \int_{-\infty}^{x_{v,i}} p(t_{v,i}|z_v{=}l, \boldsymbol{\theta}'_{l,i}) dt_{v,i} \;\; . \tag{7.5}$$

To model the marginal densities in (7.4), i.e. $p(x_{v,i}|z_v{=}l, \boldsymbol{\theta}'_{l,i})$ for all modalities i and a label l, any univariate distribution can be used, as long as we have access to the full sufficient statistics for their distribution parameters. Asymmetric distributions such as the Gumbel, Fréchet or Gamma distribution might be favorable to fit skewed data, and logistic distributions are a good choice to model distributions with excess kurtosis. Note that, when all marginals are modeled with normal distributions, (7.4) becomes mathematically equivalent to a multivariate Gaussian. For brain tumor segmentation, we propose to use Gumbel distributions to model the marginal densities. Tumor intensity distributions are typically asymmetric outlier distributions, being either hypo- or hyper-intense in the input modalities. Figure 7.5 shows empirical tumor intensity distributions in each modality together with fitted Gumbel distributions, demonstrating that Gumbel distributions are indeed a good fit. The Gumbel probability density function is parametrized by a location and a scale parameter, μ_{gum} and β_{gum}:

$$p_{\text{gum}}(x) = \frac{1}{\beta_{\text{gum}}} \exp -\Big(\delta \;\; \frac{x - \mu_{\text{gum}}}{\beta_{\text{gum}}} + \exp -\big(\delta \;\; \frac{x - \mu_{\text{gum}}}{\beta_{\text{gum}}}\big)\Big) \;\; , \tag{7.6}$$

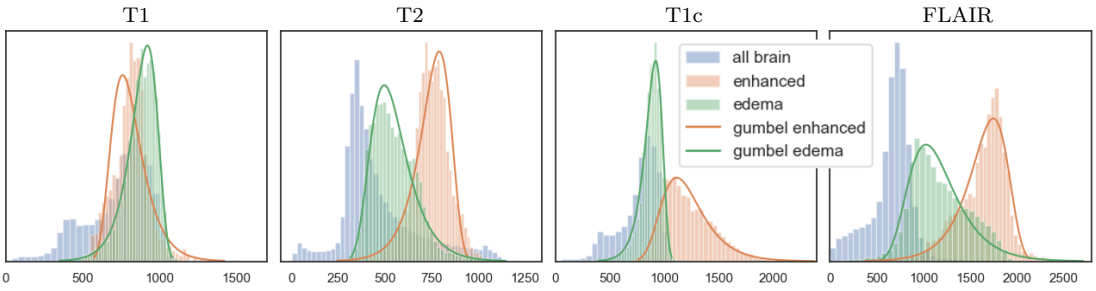


Figure 7.5: Empirical intensity distributions, for a BRATS 2016 training case, together with fitted left- and right-skewed Gumbel distributions. Intensity histograms are given for all voxels in the brain mask (•) as well as for the enhanced tumor voxels (•) and the edema tumor voxels (•). In addition, Gumbel distributions are shown, fitted to the enhanced and edema distributions by means of moment matching. All histograms are normalized for visualization purposes.

where δ=1 in the right-skewed and δ=-1 in the left-skewed Gumbel distribution. We calculate the sample skewness, $\hat{\kappa}$, in order to determine whether to adopt a right-skewed ($\hat{\kappa} \geq 0$) or left-skewed ($\hat{\kappa} < 0$) Gumbel distribution. The remaining parameters in (7.6) can both be expressed in terms of the sample mean and variance, $\hat{\mu}$ and $\hat{\sigma}^2$, based on moment matching:

$$\mu_{\text{gum}} = \begin{cases} \hat{\mu} - \gamma\beta_{\text{gum}} & \text{if } \hat{\kappa} \geq 0 \text{ (right-skewed)} \\ \hat{\mu} + \gamma\beta_{\text{gum}} & \text{if } \hat{\kappa} < 0 \text{ (left-skewed)} \end{cases} \quad \text{and} \quad \beta_{\text{gum}} = \frac{\sqrt{6}\hat{\sigma}}{\pi} \ , \tag{7.7}$$

where γ is the Euler-Mascheroni constant.

The Gumbel distribution parameters can thus be calculated based on the sample mean, variance and skewness, $\hat{\mu}_{l,i}$, $\hat{\sigma}_{l,i}$ and $\hat{\kappa}_{l,i}$, observed in the data for label distribution l in modality i. The sample mean, variance and skewness are calculated as weighted population statistics of the data observed in modality i, based on the posterior probabilities for label l, $p(z_v{=}l|\mathbf{x}_v, \boldsymbol{\theta}_l)$, as follows:

$$\hat{\mu}_{l,i} = \frac{\sum_{v\in\mathcal{V}} x_{v,i} p(z_v{=}l|\mathbf{x}_v, \boldsymbol{\theta}_l)}{\sum_{v\in\mathcal{V}} p(z_v{=}l|\mathbf{x}_v, \boldsymbol{\theta}_l)} \tag{7.8}$$

$$\hat{\sigma}^2_{l,i} = \frac{\sum_{v\in\mathcal{V}} (x_{v,i} - \hat{\mu}_{l,i})^2 p(z_v{=}l|\mathbf{x}_v, \boldsymbol{\theta}_l)}{\sum_{v\in\mathcal{V}} p(z_v{=}l|\mathbf{x}_v, \boldsymbol{\theta}_l)} \tag{7.9}$$

$$\hat{\kappa}_{l,i} = \frac{\sum_{v\in\mathcal{V}} \left(\frac{x_{v,i} - \hat{\mu}_{l,i}}{\hat{\sigma}_{l,i}} \right)^3 p(z_v{=}l|\mathbf{x}_v, \boldsymbol{\theta}_l)}{\sum_{v\in\mathcal{V}} p(z_v{=}l|\mathbf{x}_v, \boldsymbol{\theta}_l)} \ . \tag{7.10}$$

The total set of parameters of the univariate distribution for label l modeled in modality i can then be written as: $\boldsymbol{\theta}'_{l,i} = \{\hat{\mu}_{l,i}, \hat{\sigma}_{l,i}, \hat{\kappa}_{l,i}\}$. They are part of the generative model parameters, $\forall$ modalities $i : \boldsymbol{\theta}'_{l,i} \in \boldsymbol{\theta}_l$.

To model the copula $c_{\boldsymbol{\theta}''_l}(\cdot)$ in (7.4), we use Gaussian copulas, motivated by their prominent use in literature ([119]). In Figure 7.4, bivariate Gaussian copulas generated

for two synthetic images are illustrated for two image regions. In contrast to other types of copulas, we have access to an explicit formula for the Gaussian copula density $c_{\boldsymbol{\theta}_l''}$:

$$c_{\boldsymbol{\theta}_l''}\begin{pmatrix} u_{v,1} \\ \vdots \\ u_{v,m} \end{pmatrix} = \frac{1}{\sqrt{|\hat{R}_l|}} \exp -\frac{1}{2} \begin{pmatrix} \Phi^{-1}(u_{v,1}) \\ \vdots \\ \Phi^{-1}(u_{v,m}) \end{pmatrix}^T \left(\hat{R}_l^{-1} - \mathbf{I}\right) \begin{pmatrix} \Phi^{-1}(u_{v,1}) \\ \vdots \\ \Phi^{-1}(u_{v,m}) \end{pmatrix} \ , \tag{7.11}$$

with $\hat{R}_l$ the correlation matrix calculated for label l, Φ^{-1} the inverse cumulative density function of a standard normal variable and $\mathbf{I}$ the identity matrix. The correlation matrix $\hat{R}_l$ is a sample correlation matrix calculated over the transformed vectors $\left(\Phi^{-1}(u_{v,1}), \ldots, \Phi^{-1}(u_{v,m})\right)$. It is the only parameter of the Gaussian copula: $\boldsymbol{\theta}_l'' = \{R_l\}$, and is also part of the generative model parameters: $\boldsymbol{\theta}_l'' \in \boldsymbol{\theta}_l$. It can be derived calculating,

$$\hat{S}_l = \sum_{v \in \mathcal{V}} \frac{p(z_v{=}l|\mathbf{x}_v, \boldsymbol{\theta}_l)}{\sum_{v \in \mathcal{V}} p(z_v{=}l|\mathbf{x}_v, \boldsymbol{\theta}_l)} \begin{pmatrix} \Phi^{-1}(u_{v,1}) \\ \vdots \\ \Phi^{-1}(u_{v,m}) \end{pmatrix} \begin{pmatrix} \Phi^{-1}(u_{v,1}) \\ \vdots \\ \Phi^{-1}(u_{v,m}) \end{pmatrix}^T \ , \tag{7.12}$$

such that $\hat{R}_l = D^{-\frac{1}{2}} \hat{S}_l D^{-\frac{1}{2}}$, with $\hat{S}_l$ the sample covariance matrix and D a diagonal matrix with its diagonal entries. To summarize, all parameters $\boldsymbol{\theta}_l$ in (7.4) can now be written as:

$$\boldsymbol{\theta}_l = \boldsymbol{\theta}_l' \cup \boldsymbol{\theta}_l'' = \{\hat{\mu}_{l,i}, \hat{\sigma}_{l,i}^2, \hat{\kappa}_{l,i}\}_{\forall i} \cup \{\hat{R}_l\} \ . \tag{7.13}$$

A stable label prior. The second factor in (7.1) is the label prior $p(z_v{=}l|\alpha_{v,l})$. The prior allows to incorporate information present in the training set, assists in stable segmentation outcomes and can prevent the model from converging towards implausible situations. We identify the prior for label l with the probabilistic segmentation map calculated by the initializing supervoxel classifier (cf. Section 7.2.2.2), denoted as $\alpha_{v,l}$ in a voxel v for label l:

$$p(z_v{=}l|\alpha_{v,l}) \leftarrow \alpha_{v,l} \ . \tag{7.14}$$

7.2.1.2 Optimization by the EM Algorithm

The EM algorithm iteratively maximizes the data log-likelihood of the observed variables by updating a set of parameters $\boldsymbol{\theta}$. The data log-likelihood is written as a function of the probability density function of the observed variables $\mathbf{x}$ conditioned on the parameters $\boldsymbol{\theta}$:

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = \log p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{v \in \mathcal{V}} \log p(\mathbf{x}_v|\boldsymbol{\theta}) \ . \tag{7.15}$$

The hidden variable z is introduced using the law of total probability, writing the data likelihood in a voxel v as a sum over all labels $l \in 1 \ldots k$:

$$p(\mathbf{x}_v|\boldsymbol{\theta}) = \sum_{l=1}^{k} p(\mathbf{x}_v, z_v{=}l|\boldsymbol{\theta}_l) \ , \tag{7.16}$$

with $\boldsymbol{\theta} = \{\boldsymbol{\theta}_l | l \in 1 \dots k\}$, the complete set of parameters to be estimated.

If we would have a solution for $\boldsymbol{\theta}$, we could estimate the hidden variable z via (7.2), using the aforementioned equations (Section 7.2.1.2). However, due to the presence of the hidden variable, setting the gradient of the data log-likelihood in (7.15) to zero will not leave us with a closed-form solution for the parameters $\boldsymbol{\theta}$. The EM algorithm solves this problem by iterative decoupling, splitting each iteration in E(xpectation) and M(aximization) steps:

$$\text{E-step:} \quad Q_v(z_v) \leftarrow p(z_v|\mathbf{x}_v, \boldsymbol{\theta}^{(i)}) \tag{7.17}$$

$$\text{M-step:} \quad \boldsymbol{\theta}^{(i+1)} = \arg\max_{\boldsymbol{\theta}} \sum_{v \in \mathcal{V}} E_{Q_v} \left[\log p(\mathbf{x}_v, z_v|\boldsymbol{\theta})\right] \tag{7.18}$$

More in particular, the E-step approximates the posterior probabilities of the hidden variable, while the M-step updates the parameters by means of a relatively cheap point estimate. The calculations associated with the E- and M-steps present in each EM-iteration, are summarized in Algorithm 1.

EM initialization. A well-known problem of the EM algorithm is its dependence on initialization. Although the marginal likelihood of the data, $p(\mathbf{x}|\boldsymbol{\theta})$, is guaranteed to be non-decreasing along the iterations, the EM does not guarantee to converge to the global minimum. One technique to get closer to the global minimum of the often high-dimensional, heavy-tailed and multi-modal likelihood function, is to refine the initialization ([114]). Depending on the application, different initializations have been proposed. The posterior probabilities in the E-step, $p(z_v{=}l|\mathbf{x}_v, \boldsymbol{\theta}_l)$, are commonly initialized by the prior generated by population atlases or they are taken as flat priors. Alternatively, the parameters $\boldsymbol{\theta}_l$ can be initialized, for example as their expected values calculated based on the training set. We initialize the EM algorithm by setting the initial posterior label probabilities in (7.17), to the prior on the hidden variable, which we already set to the probabilistic segmentation maps α calculated by the initializing supervoxel classifier in (7.14):

$$Q_v^{(0)}(z_v{=}l) \leftarrow \alpha_{v,l} \ . \tag{7.19}$$

7.2.2 Supervoxel grid: Supervoxel Classifiers

In order to cope with the two main limitations of the generative probabilistic model, we propose to use supervoxel classifiers: the initializing supervoxel classifier generates a prior for the probabilistic model, which also serves as the EM initializations, and the label context supervoxel classifier aims to spatially regularize the segmentations generated by the probabilistic model.

7.2.2.1 Supervoxel Generation, Feature Extraction and Classification

To operate on a supervoxel grid, supervoxels are calculated in a multivariate fashion, over all input modalities at once, using SLIC as proposed by [117]. An example of

Algorithm 1 Expectation-Maximization (EM) iterations

function EM-ITERATION($\theta^{(t)}, \alpha$) ▷ t^{th} iteration, prior maps α
  $\{p(z_v{=}l|\mathbf{x}_v, \theta_l^{(t)})\}_{\forall l} \leftarrow$ EXPECTATION($\theta^{(t)}, \alpha$) ▷ E-step
  $\theta^{(t+1)} \leftarrow$ MAXIMIZATION$\big(\{p(z_v{=}l|\mathbf{x}_v, \theta_l^{(t)})\}_{\forall l}\big)$ ▷ M-step
  return $\theta^{(t+1)}$
end function

function EXPECTATION(θ, α) ▷ $\theta = \{\theta_l\}_{\forall l}$
  for $l \leftarrow 1, k$ **do**
    calculate $p(\mathbf{x}_v|z_v{=}l, \theta_l)$ ▷ using (7.4)
    calculate $p(z_v{=}l|\alpha_{v,l})$ ▷ using (7.14)
    $p(\mathbf{x}_v, z_v{=}l|\theta_l) \leftarrow p(\mathbf{x}_v|z_v{=}l, \theta_l)p(z_v{=}l|\alpha_{v,l})$ ▷ (7.1)
  end for
  for $l \leftarrow 1, k$ **do**
    $p(z_v{=}l|\mathbf{x}_v, \theta_l) \leftarrow \frac{p(\mathbf{x}_v, z_v{=}l|\theta_l)}{\sum_{l=1}^{k} p(\mathbf{x}_v, z_v{=}l|\theta_l)}$ ▷ (7.2)
  end for
  return $\{p(z_v{=}l|\mathbf{x}_v, \theta_l)\}_{\forall l}$
end function

function MAXIMIZATION$\big(\{p(z_v{=}l|\mathbf{x}_v, \theta_l)\}_{\forall l}\big)$
  for $l \leftarrow 1, k$ **do**
    for $i \leftarrow 1, m$ **do**
      calculate $\{\hat{\mu}_{l,i}, \hat{\sigma}^2_{l,i}, \hat{\kappa}_{l,i}\}$ ▷ using (7.8), (7.9), (7.10)
    end for
    calculate $\hat{R}_l$ ▷ using (7.12)
    $\theta_l \leftarrow \{\hat{\mu}_{l,i}, \hat{\sigma}^2_{l,i}, \hat{\kappa}_{l,i}\}_{\forall i} \cup \{\hat{R}_l\}$ ▷ (7.13)
  end for
  $\theta \leftarrow \{\theta_l\}_{\forall l}$
  return θ
end function

supervoxels generated based on four brain MR images is visualized in Figure 7.1. SLIC is based on a k-means clustering approach that defines distances D between supervoxel cluster centers and surrounding voxels:

$$D = \sqrt{\left(\frac{d_m}{m}\right)^2 + \left(\frac{d_s}{S}\right)^2} \ , \tag{7.20}$$

with d_m, the m-dimensional Euclidean distance between the grey level intensities in the m modalities and d_s the 3D Euclidean distance in space (m and S normalization factors as defined in [117]). In order to ensure each modality is equally weighted in d_m, all images are 1) standard normalized to zero mean and unit standard deviation, 2) cut off at mean $\pm$ four standard deviations and 3) rescaled to be in $[0, 1]$, prior to supervoxel generation.

After supervoxel generation, a feature vector $\mathbf{f}_S$ is assembled for each supervoxel S by extracting supervoxel features, based on regional image features and segmentation labels. A RF classifier is then trained to predict the dominant tumor label in the supervoxel, $z_S = \arg\max_l \ |\{v|z_v{=}l, \forall v \in S\}|$, where we use the ground truth labels for z_v. The trained RF classifier then generates probabilistic supervoxel predictions for each label l: $p(z_S = l|\mathbf{f}_S)$.

7.2.2.2 Initializing Supervoxel Classifier

For the initializing supervoxel classifier, the supervoxel features $\mathbf{f}_S^{init}$ are calculated based on the regional image features in $\mathbf{x}$. For each supervoxel, we extract: *a*) its intensity mean, median and standard deviation in each standard-normalized modality, *b*) the roundness of the supervoxel: $R_S = A_{\text{sphere}}/A_S$, with A_S the surface area of the supervoxel and A_{sphere} the surface area of a sphere with the same volume as the supervoxel, and *c*) tissue-specific information such as the percentage of GM, WM and CSF present in the supervoxel.

The probabilistic supervoxel predictions are then mapped on a voxel-level and are used as a prior for the probabilistic model described in Section 7.2.1:

$$\alpha_{v,l} \leftarrow p(z_v{=}l|\mathbf{f}_S^{\text{init}}, v{\in}S) = p(z_S{=}l|\mathbf{f}_S) \ . \tag{7.21}$$

7.2.2.3 Label Context Supervoxel Classifier

The label context supervoxel classifier is trained to improve spatial label coherence. The supervoxel features $\mathbf{f}_S^{label}$ are now calculated based on the regional image features in $\mathbf{x}$ and the probabilistic segmentations of the probabilistic model $p(z_v|\mathbf{x}_v, \boldsymbol{\theta})$ described in Section 7.2.1. Prior to classification, the previously calculated supervoxels are split until each supervoxel contains only one label in the label segmentations of the probabilistic model. This tends to make the supervoxel grid more precise around the segmentation borders. For each supervoxel, we extract: *a*) its intensity mean in each standard-normalized modality, *b*) the predicted labels (in %) in the supervoxel, *c*) the predicted labels (in %) directly neighboring the supervoxel, *d*) the normalized

difference between the mean intensity in the supervoxel and the mean intensity in the predicted labels, and e) the distance of the supervoxel to the center-of-mass of each predicted label.

The probabilistic supervoxel predictions are then mapped on a voxel-level to acquire the spatially regularized, final segmentation:

$$p(z_v|\mathbf{f}_S^{label}, v \in S) = p(z_S|\mathbf{f}_S^{label}) \ . \tag{7.22}$$

7.2.3 Semi-Automated Extensions

Because of the challenging nature of brain tumor datasets and real-world clinical cases, we extended our proposed method with two interactive, semi-automated alternatives, indicated in blue in the pipeline in Figure 7.2. They allow a user to guide the segmentation process for cases where the fully automated results aren't sufficient. The interactive framework is visualized in Figure 7.6.

In the first interactive extension, denoted as semi-automated super-fast '**Semi-auto SF**', we allow the user to interact with the pipeline after the probabilistic model and before the label context supervoxel classifier (cf. Figure 7.2). The user is asked to click on supervoxels, close to the center-of-mass of the tumor. As described in Section 7.2.2.3, the label context supervoxel classifier calculates the distance of each supervoxel to the center-of-mass of each predicted label as part of the supervoxel features $\mathbf{f}_S^{label}$. In the fully-automated model, the center-of-mass of the tumor is calculated on the label segmentations of the probabilistic model. In this interactive extension, the center-of-mass is preset to the center-of-mass of the supervoxels selected by the user. This will generate more reliable features in cases where the segmentations of the probabilistic model are unreliable. In a second extension, denoted as semi-automated quick manual '**Semi-auto QM**', we allow the user to interact with the final segmentations. The user is asked to change supervoxel labels in case of misdetection. An interactive framework is built, visualizing the final segmentations and the supervoxels, and the user is able to click on supervoxels and change their respective labels, for example by setting false positive supervoxels to background. The process of changing supervoxel labels within the interactive framework is visualized in Figure 7.6.

7.3 Experiments

We conduct several experiments on brain tumor segmentation datasets. We first validate our model choices (Section 7.3.1) and then test our model by reporting its performance on a test set and comparing it to the state-of-the-art methods (Section 7.3.2). We finally conduct experiments on a clinical dataset, containing mostly pre- and post-operative scans with resection cavities and missing modalities, to test our method on a highly challenging real-world clinical dataset and test its generalization to incomplete datasets (Section 7.3.3).

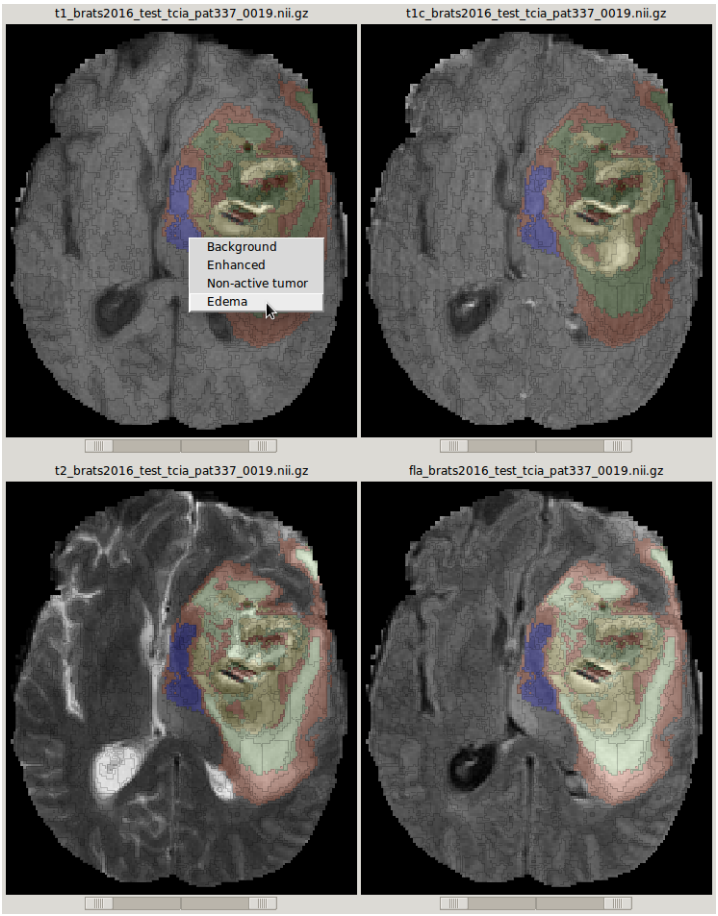


Figure 7.6: Illustration of interactive framework used for the semi-automated extensions for a complex case within the BRATS 2016 dataset. Automatically calculated tumor segmentations are shown in yellow, green and red for enhanced, non-active and edema. The user can select multiple supervoxels (shown in blue) and get the center-of-mass of the selected supervoxels – within the context of the **Semi-auto SF** extension – or change their label (via keyboard shortcuts or via the shown drop down list) – within the context of the **Semi-auto QM** extension. The complete interactive framework is publicly available (`https://github.com/s0216660/gui_supervoxel_tumor_segmentation`).

Data. For validation and testing, we use the **BRATS 2016 dataset**, consisting of a training set with 244 cases and a test set with 191 cases. We adopt the BRATS training set for the validation of our model (Section 7.3.1) and test our model on the BRATS test set (Section 7.3.2) for comparison with the state-of-the-art. The BRATS dataset contains four standard MR modalities: FLAIR, T2, T1 and T1 contrast-enhanced (T1c). Ground truth segmentations are available for the training set and contain labels for 'background', 'edema', 'enhanced', 'non-active' and 'necrotic'. For the test set, no ground truth is available: all evaluations are done by the BRATS organizers.

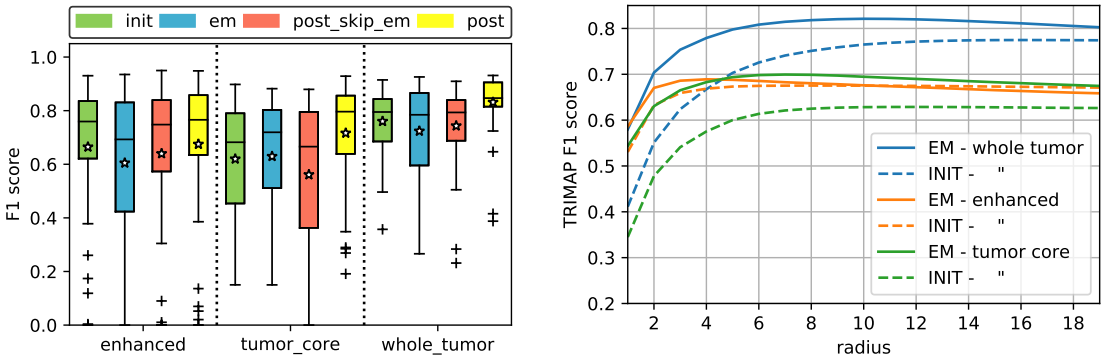
To evaluate our algorithm in a complex real-world scenario (Section 7.3.3), we adopt a challenging **clinical dataset from [4]**, consisting of 97 cases. In contrast to the BRATS dataset, this clinical dataset contains only three standard MR modalities for each case: FLAIR, T1 and T1c. T2 modalities, on the other hand, are *not* available in this dataset.

Experimental set-up. WM, GM and CSF probability maps are generated for each case individually, by non-rigid registration of the WM, GM and CSF atlas maps from the SRI24 Atlas ([3]) to subject space, using the fast free-form deformation provided by NiftiReg ([123]). The initializing supervoxel classifier is trained on the 200 training subjects we separated from the BRATS 2016 training set, predicting 'background', 'enhanced tumor', 'non-active tumor', 'necrotic tumor' and 'edema' for each supervoxel. To model the univariate distributions of the tumor classes in the generative probabilistic model in (7.4), we use Gumbel distributions, as described in Section 7.2.1.1 and motivated in Figure 7.5. Finally, the label context supervoxel classifier is trained on the same 200 training subjects. For both RF supervoxel classifiers, we use 50 trees, inversely weight the class frequencies in the leaves with the class frequency in the training data (the dataset is highly unbalanced), bootstrap the training data used to grow each tree, while preserving training sample size, and stop splitting nodes in the trees if less then 10 supervoxel training samples remain, to avoid overfitting.

Evaluation criteria. Similar to the BRATS evaluations, we evaluate segmentation accuracies for enhanced tumor, tumor core (including enhanced, necrotic and non-active tumor) and whole tumor (including tumor core and edema). We calculate F1 scores, also known as Dice scores, to quantify the percentage of voxel-wise overlap between the segmented tumor regions and the ground truth. Additionally, we also calculate TRIMAP scores, to assess the accuracy of the segmentation borders. TRIMAP scores are calculated as the F1 score between the segmented region and the ground truth region, while only considering the voxels within a certain margin around the ground truth border. Lastly, we also report Hausdorff distances on the clinical dataset. The Hausdorff distance quantifies the surface distance between two segmentations. Similar to the BRATS challenge evaluations ([2]), we report 'Hausdorff 95', i.e. the 95 percentile instead of the maximum distance between the segmented and ground truth tumor surfaces.

Computation time and resources. All experiments are conducted on a workstation with a modern CPU[1] and 31,4 GB RAM. Average runtime for a single multi-modal segmentation case is 72.5 s for the initializing supervoxel classifier, 23.8 s for the generative probabilistic model and 81.3 s for the label-context supervoxel classifier (runtime is calculated and averaged over the 44 BRATS validation cases used in Section 7.3.1).

[1] Intel® Xeon® CPU E3-1225 v3 @ 3.20 GHz × 4



(a) (b)

Figure 7.7: Results on the BRATS 2016 validation set. (a) F1 scores after the initializing supervoxel classifier (■), the probabilistic model (■) and the label context supervoxel classifier (■: when leaving out the probabilistic model, ■: as proposed). (b) TRIMAP F1 scores for whole tumor (blue), enhanced (orange) and tumor core (green), after initialization (dotted line) and after the probabilistic model (solid line). Each step is crucial and contributes to the final segmentation accuracy. The probabilistic model plays a key-role and improves the accuracy of the segmentation borders.

7.3.1 Validation of Model Choices and Robustness

To motivate our model choices and assess the robustness of our model, we report performance after the supervoxel classifiers and after the probabilistic model, conduct experiments with different initializations on the probabilistic model, with missing modalities and with limited training data sizes. We adopt the BRATS 2016 training set and randomly subsample 44 validation cases, referred to as the BRATS 2016 validation set. The remaining 200 training cases are used to train our models.

Performance obtained by the supervoxel classifiers and the probabilistic model. Figure 7.7a reports the F1 scores of the label maps obtained after the initializing supervoxel classifier (init), the probabilistic model (em) and after the label context supervoxel classifier (post). These evaluations correspond to the segmentation maps shown in Figure 7.2. Observe that the average F1 score does not considerably increase after the probabilistic model. However, based on Figure 7.7, the probabilistic model seems to be crucial for the performance of the proposed method: 1) rather than increasing the overall F1 score, it increases the TRIMAP scores (Figure 7.7b), indicating that it increases the border segmentation accuracy and 2) it ensures a significant increase in F1 score after the label context supervoxel classifier (mean F1 scores of .83, .72, .68 for whole tumor, tumor core and enhanced) compared to when it is just left out (mean F1 scores of .74, .56, .64 resp.), cf. two rightmost boxplots in Figure 7.7a for each tumor label. Qualitative segmentation results are shown in Figure 7.8, including the two least

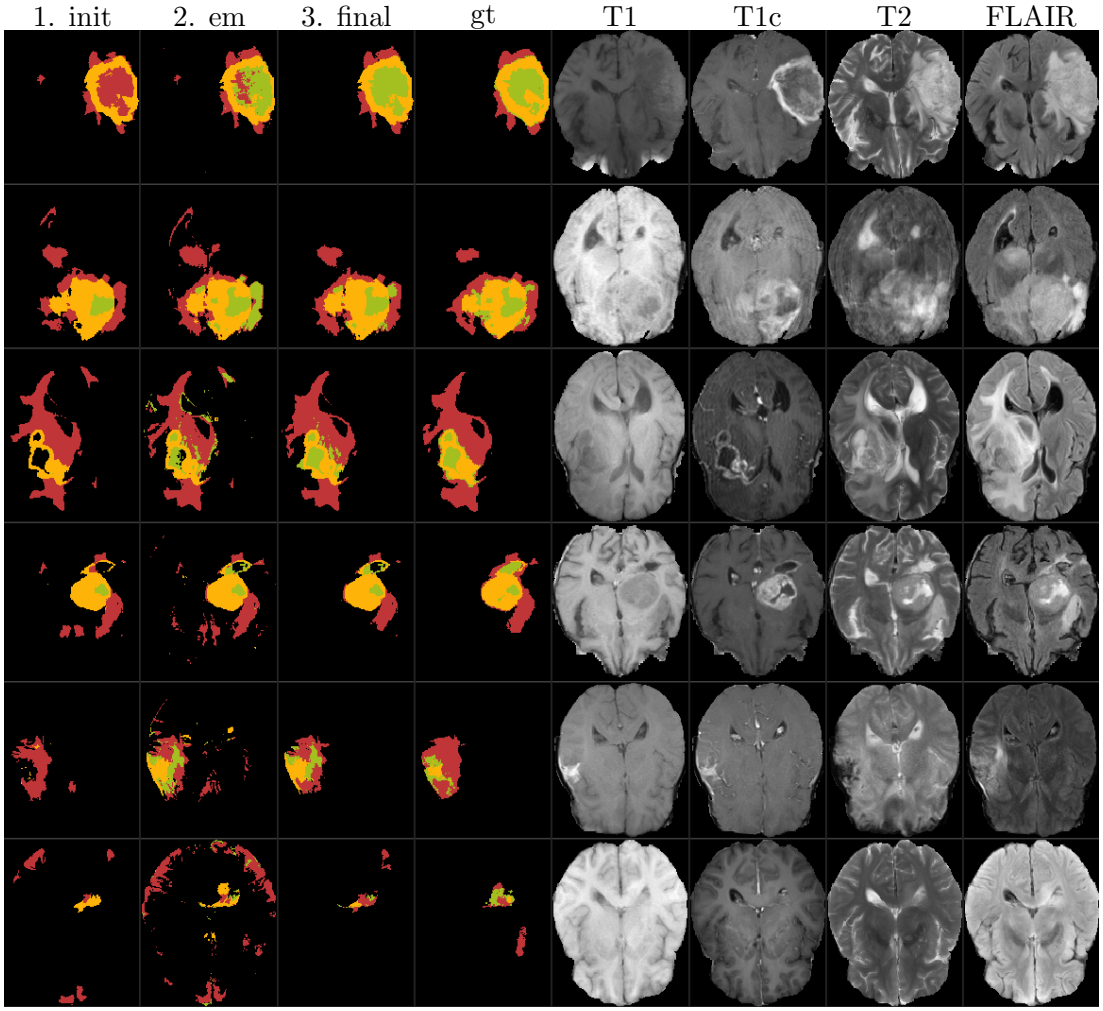


Figure 7.8: Segmentation results on the BRATS 2016 validation set. The four upper rows correspond to F1 scores within the interquartile ranges, while the two lower rows correspond to the lower outliers in the boxplots in Figure 7.7a. Left to right: segmentation results of the proposed method, acquired after the initializing supervoxel classifier (init), the probabilistic model (em) and the label context supervoxel classifier (final), ground truth (gt) and four MR images (T1, T1c, T2 and FLAIR). Segmentations are shown for enhanced (●), non-active tumor (●) and edema (●).

accurate segmentation cases, and reveal accurate tumor segmentations and an obvious improvement after each step.

Prior tumor label maps generated by the initializing supervoxel. The initializing supervoxel classifier generates probabilistic label maps, $\boldsymbol{\alpha}$, adopted as a label prior in the generative probabilistic model. Figure 7.10 visualizes the probabilistic prior maps

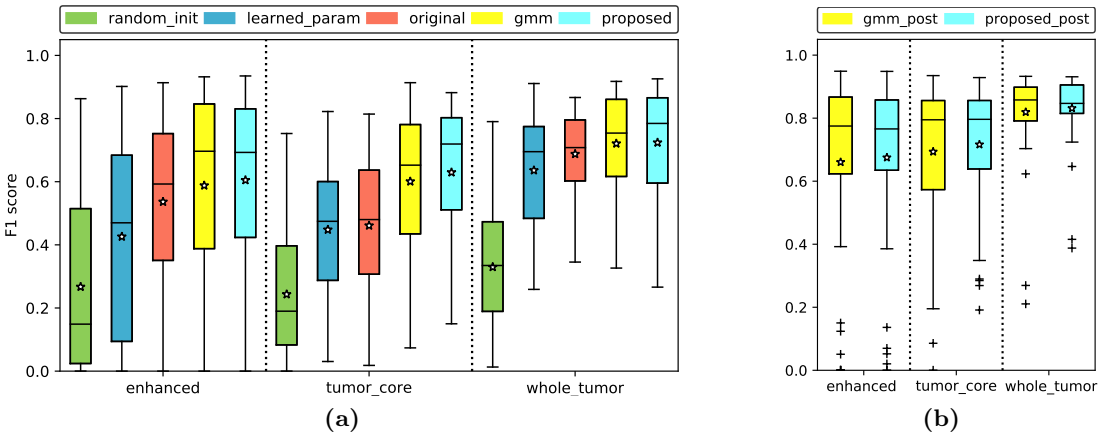


Figure 7.9: Initialization and use of copulas in the probabilistic model, tested on the BRATS 2016 validation set. (a) F1 scores after the probabilistic model for flat initialization (■), using expected parameters for $\boldsymbol{\theta}$, as calculated on the training set (■), using multivariate Gaussians as in [8] (■), using a mixture of multivariate Gaussians (■) and initialized with the supervoxel classifier using univariate Gumbel distributions and Gaussian copulas (■). (b) Final F1 scores when using a mixture of Gaussians in the probabilistic model (■) and as proposed (■).

$\boldsymbol{\alpha}$ for enhanced, non-active tumor and edema, generated for a subset of the BRATS 2016 validation set.

EM initialization and use of Gaussian copulas. Figure 7.9a reports the F1 scores acquired after the EM with different initializations and compares the performance when using different likelihood distributions. We perform experiments with a) random initialization, where we initialize the parameters $\boldsymbol{\theta}$ in (7.13) randomly, with b) learned parameters, based on the training data statistics (i.e. images are normalized to zero mean and unit standard deviation, $\boldsymbol{\theta}$ is calculated for every training case using the ground truth and their averages are taken as the initial parameters for unseen cases) and with c) the initializing supervoxel classifier (proposed), where we initialize the probabilistic segmentation maps rather than the parameters $\boldsymbol{\theta}$ (cf. Figure 7.10 for visualization of the label maps generated by the initializing supervoxel classifier). To evaluate different likelihood models, we consistently use the same initialization as proposed and perform experiments using a) multivariate Gaussians with diagonal covariance matrices (one Gaussian per tumor class) as in [8] (Figure 7.9a – original), b) mixtures of three multivariate Gaussians for each tumor class (Figure 7.9a – gmm) and c) Gaussian copulas with univariate Gumbel distributions (Figure 7.9a – proposed).

Initialization with the initializing supervoxel classifier outperforms the other initialization methods considerably. By using Gaussian copulas with Gumbel distributions we also considerably increase the performance (mean F1 scores of .72, .63, .60 for whole tumor, tumor core and enhanced) compared to the original method which uses mul-

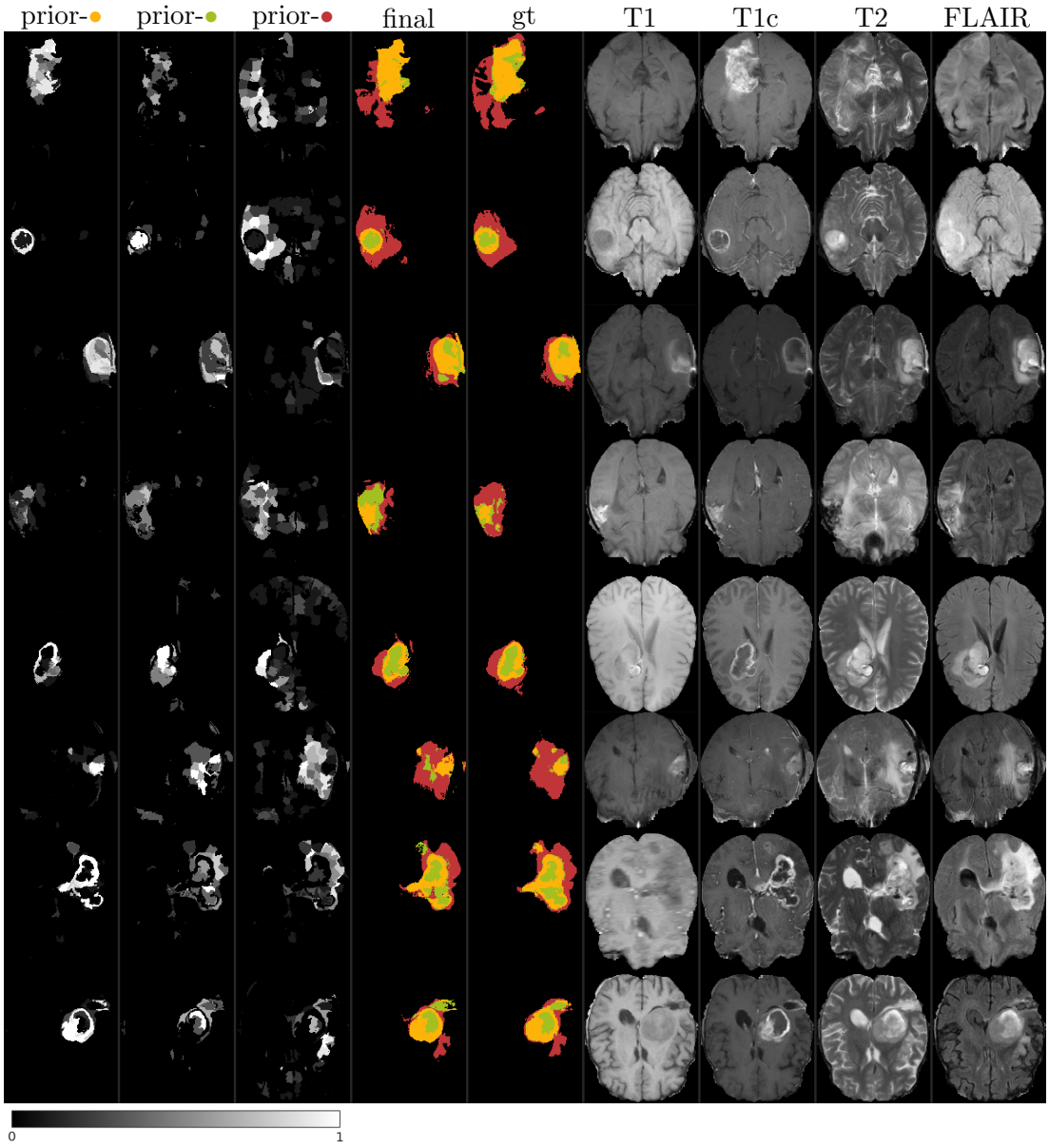


Figure 7.10: Label prior maps generated by the initializing supervoxel classifier and adopted in the generative probabilistic model. From left to right: prior label maps $\boldsymbol{\alpha}$ for enhanced (●), non-active tumor (●) and edema (●), final segmentation (after label-context classifier), ground truth and input MR modalities (T1, T1c, T2 and FLAIR).

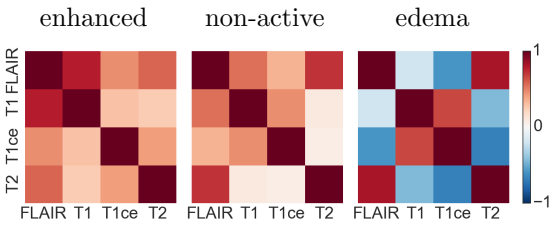


Figure 7.11: Gaussian copula correlation matrices $\hat{R}_l$ for enhanced, non-active tumor and edema. Strong correlations include T1 and FLAIR for enhanced tumor and T1c and T2 for whole tumor. T1c and T2 are weakly correlated for tumor core.

tivariate Gaussians (mean F1 scores of .69, .46, .54 resp.). When using the mixture of three multivariate Gaussians per tumor class, the F1 scores are comparable to our proposed method, at the expense that more parameters need to be calculated (longer computation time). Figure 7.11 visualizes the correlation matrices $\hat{R}_l$ of the Gaussian copulas for different labels l, as calculated by the probabilistic model for one of the validation cases.

Missing modalities. The probabilistic model models the joint density of the input modalities on the fly, using only the prior label maps to initialize the EM algorithm, so it is able to generate results when modalities are missing or even when extra input modalities are provided. Because of this flexibility, we performed experiments in the presence of missing modalities, each time leaving out one of the four modalities. For the RF supervoxel classifiers, where we need all modalities to calculate the supervoxel features, we replaced the missing modality with an image with white noise (zero mean, unit standard deviation). The resulting F1 scores are reported in Figure 7.12, and show that performance is only little affected by leaving out T1 or T2, but is considerably reduced for enhanced tumor when T1c is missing, showing the importance of the T1c modality with respect to delineating enhanced tumor. This demonstrates the flexibility of our proposed automated method, as well as its robustness towards missing modalities.

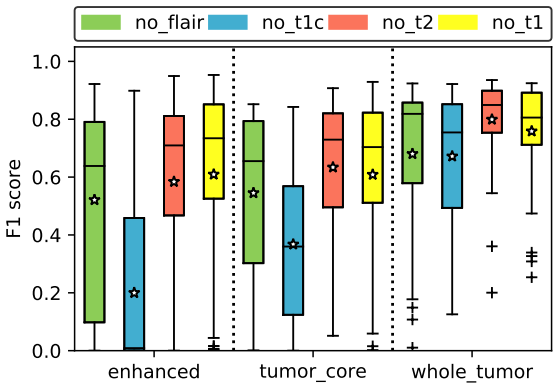


Figure 7.12: Effect of missing modalities on the BRATS 2016 validation set. Reported are F1 scores when either FLAIR (■), T1c (■), T2 (■) or T1 (■) are missing. The obtained accuracies demonstrate a good model robustness against missing modalities, especially when T1 or T2 are missing. As expected, T1c is crucial to obtain good accuracies for enhanced tumor.

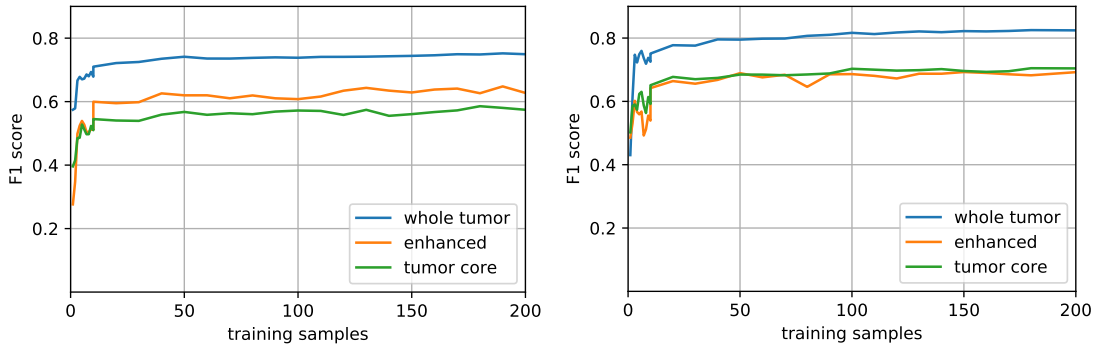


Figure 7.13: Performance in function of training sample size. Reported is the mean F1 score acquired on the BRATS 2016 validation set, after each of the supervoxel classifiers, trained on different numbers of training samples. Left: mean F1 scores on the prior maps, generated by the initializing supervoxel classifier. Right: mean F1 scores on the final segmentations, generated by the label context supervoxel classifier.

Supervoxel classifiers: dependence on training data. In Figure 7.13, we report the performance of the proposed method in function of the training sample size, both for the initializing supervoxel classifier and the label context supervoxel classifier, when training them on a limited number of training samples, randomly subsampled from the training set. For both classifiers, the mean F1 score on the validation set seems to quickly reach good performance for all tumor structures. Note that the label context classifier generates the final segmentations. Mean F1 scores of .75, .65 and .64 for whole tumor, tumor core and enhanced tumor can already be reached for the final segmentations with only 10 training samples available. This experiment reveals that the proposed method is not strongly dependent on the training dataset and does not require an extensive training data set.

7.3.2 Testing Performance and Comparison to State-of-the-Art

For testing, we use our model trained on the BRATS 2016 training set and calculate segmentations for the BRATS 2016 test set. We compare with two state-of-the-art methods, one using CNNs ([72]) and one using RFs, on a voxel level, extended with a CRF ([116]). F1 scores are calculated and provided by the BRATS organizers for enhanced tumor, tumor core and whole tumor, for all 191 segmentation volumes. The mean F1 scores obtained by the state-of-the-art methods are reported in Table 7.1 and indicate that this set is more challenging in comparison to the BRATS 2013 and BRATS 2015 test sets (on which participants reach mean F1 scores in the range of 0.8-0.9 for whole tumor structures [2]). Visual inspection of the data, reveals that this is most probably due to the diversity in imaging protocols, tumor shape and appearance (including small and big tumors, resection cavities, necrosis, internal bleedings and cysts). Table 7.2 reports inter-rater variability in the BRATS 2016 test set (provided by the BRATS organizers) and confirms the challenging nature of the test set

Table 7.1: Mean F1 score (in %) acquired over the 191 cases of the BRATS 2016 test set.

	whole tumor	tumor core	enhanced
[72]	73.9	63.3	67.2
[116]	71.8	54.6	62.3
Proposed			
– fully automated	69.7	57.0	60.9
– semi-auto SF	73.6	60.0	64.8
– semi-auto QM	77.0	65.9	69.4

Table 7.2: Inter-rater variability in terms of F1 score (mean ± std in %) acquired between two groups of experts, each generating one segmentation in consensus.

	whole tumor	tumor core	enhanced
BRATS 2016 test set	77.3 ± 18.1	59.9 ± 30.6	63.7 ± 25.4
BRATS 2013 [2]	85 ± 8	75 ± 24	74 ± 13

in comparison to former BRATS datasets (e.g. mean inter-rater F1 score of 63.7% in comparison to 74% for enhanced). Because of the complexity of this test set, we conduct experiments using the two interactive, semi-automated extensions of our model, as presented in Section 7.2.3 and visualized in Figure 7.6. We report the mean F1 scores acquired over the BRATS 2016 test set by the state-of-the-art methods, by our proposed fully-automated method an by the semi-automated extensions of our method in Table 7.1. The CNN-based method is outperforming the proposed fully automated method, whereas the RF-based method compares well to our proposed method. Our semi-auto QM interactive extension outperforms the fully-automated state-of-the-art methods and seems to offer a quick and transparent way to establish accurate and reliable tumor segmentations for difficult cases where the fully-automated algorithm needs interaction (valuable for example for cases containing very small tumors or irregularities such as internal bleedings and cysts).

7.3.3 Challenging Clinical Pre- and Post-Operative Datasets

To test our model on a real-world clinical dataset, we use our models trained on the BRATS 2016 training set and test them on the clinical dataset from [4]. This set contains 97 cases from 20 patients, each with one or two pre-operative cases and one or several post-operative cases. This is a particularly challenging test because (i) this dataset contains many post-operative cases containing resection cavities and internal bleedings, (ii) it doesn't contain any T2 modalities, making it a perfect candidate to demonstrate the importance of the missing modality problem as well as to test the ability of our model to cope with missing modalities and (iii) it contains images that

Table 7.3: F1 scores (in %) and Hausdorff 95 distances (in mm) calculated for whole tumor on the clinical pre- and post-operative datasets.

	F1 score	Haussdorf 95
median	61.3	17.3
mean ± std	59.1 ± 20.1	20.6 ± 15.3
[min, max]	$[28.3, 90.5]$	$[1.4, 58.0]$

are acquired with different scanners[2] than those contained in the BRATS set, which we used to train our model on.

Preprocessing. In contrast to the BRATS dataset, the clinical dataset is not preprocessed. Therefore, we co-register the MR modalities using the affine registrations based on block-matching provided by NiftiReg ([123]) and skull strip the MR images based on the brain masks calculated on the T1 images using the skull-stripping module from [7]. Ground truth is taken as the semi-automated segmentations performed by two experienced radiologists as described in [4], and is only available for 'enhanced' and 'edema'.

Results. We calculate the segmentations identifying enhanced, tumor core and whole tumor with the proposed method in a fully-automated fashion. Since cases containing resection cavities tend to show fine, tubular- and sheetlike structures, F1 scores can be misleading, as they quantify voxelwise overlap. Therefore, we also report Hausdorff distances, quantifying surface distance. The F1 scores and Hausdorff 95 scores are reported in Table 7.3 for whole tumor, showing good Hausdorff scores, especially considering that the dataset contains a lot of resection cavities and given T2 is missing for this dataset. Segmentations for all tumor types are visualized for pre- and post-operative scans of three patients in Figure 7.14, demonstrating reliable segmentations with great potential value for clinical practice.

7.4 Conclusion

In this study, we present a multi-scale segmentation approach, making use of a probabilistic model optimized by the EM algorithm on a voxel grid, and using supervoxel classifiers for initialization and regularization on a supervoxel grid. We tailored the probabilistic model specifically to deal with outlier-like tumor intensity distributions by combining univariate asymmetric Gumbel distributions and multivariate Gaussian copulas to model the dependencies between the input modalities. We show that our method produces brain tumor segmentations comparable to the state-of-the art, offer

[2] 3 Tesla (T) Achieva scanner (Philips Medical Systems) and 3T Verio scanner (Siemens Healthcare) using 3D-spoiled gradient echo sequences either with an 8- or 16-channel phased array head coil. TR, TE and image resolution are reported in [4].

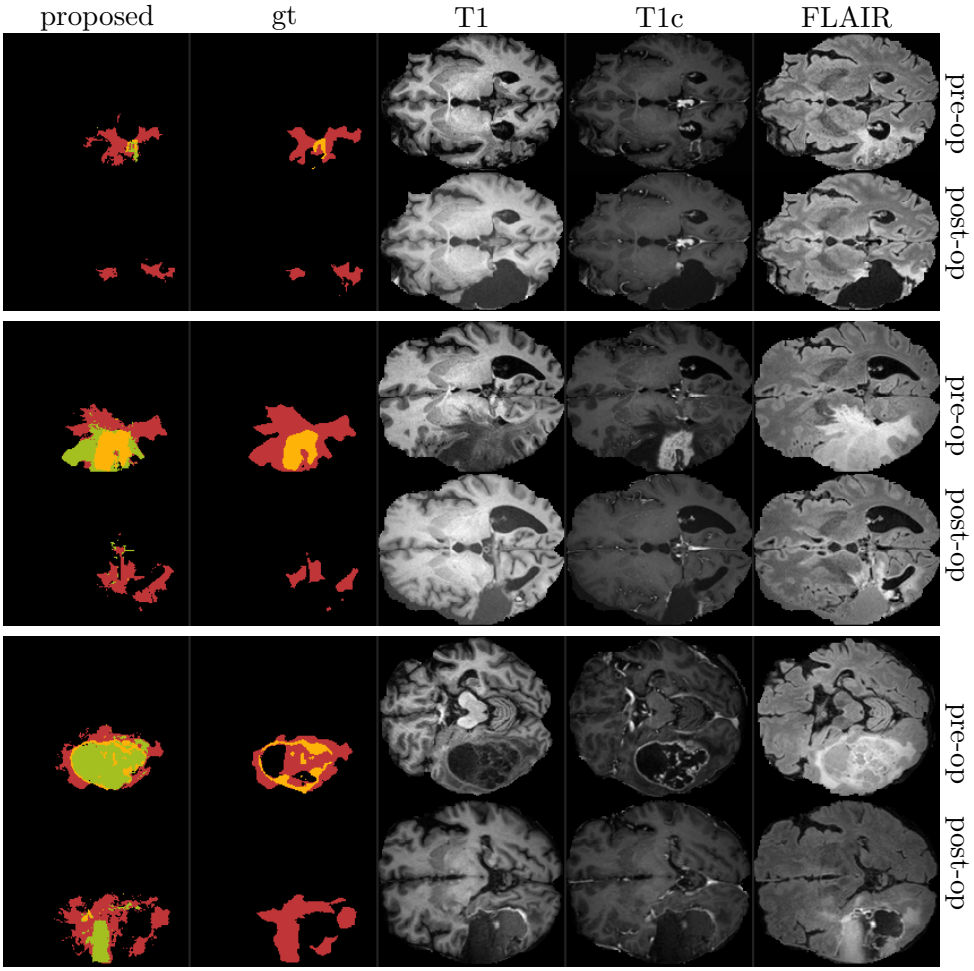


Figure 7.14: Segmentation results on a challenging clinical test set with missing T2 modalities. For three patients, pre- and post-operative scans with resection cavities are shown. Left to right: segmentation results of the proposed method, ground truth (only enhanced and edema are annotated), and three MR images (T1, T1c and FLAIR). Segmentations are shown for enhanced (●), non-active tumor (●) – not annotated in the ground truth – and edema (●).

a flexible semi-automated extension and demonstrate a number of valuable flexibilities such as being able to deal with small amounts of training data, missing input modalities and resection cavities, while only requiring ordinary computation time and resources.

8 Conclusion & Outlook

This thesis covers several medical image analysis topics related to brain tumor segmentation, growth analysis and classification.

Brain tumor segmentation is a complex topic which has received a lot of interest from the technical community and is still an ongoing field of research. State-of-the-art brain tumor segmentation methods are generally performing pretty well, but lack a few flexibilities which are very relevant for clinical practice. The brain tumor segmentation method presented in this thesis has a strong focus on clinical usability. Firstly, it only takes a few minutes to run on a modern desktop computers. Secondly, it can still generate results if a few scans are missing. And lastly, it provides semi-automated extensions during which a user (for example an experienced radiologist) can adapt the automatically generated segmentations in a quick, easy and transparent way. For this purpose, a semi-interactive framework is build in which segmentations can be automatically calculated, visualized and manually adapted. The automatically generated segmentations perform well and almost obtain accuracies comparable to the stat-of-the-art brain tumor segmentation methods (tested on the BRATS 2016 test set). Moreover, they are able to generate reliable segmentations on datasets acquired on different scanners and even on datasets where some MR scans are missing. The part that allows for manual adaption is based on dividing the scans in small subregions, called supervoxels, which the user can select and assign a certain label. These supervoxels are already calculated in the fully-automated part of the segmentation pipeline, and therefore require no additional computation costs. This manual interaction is proven to be especially valuable and appreciated by experienced radiologists, as it allows to do a quick 'check' on the automated calculated segmentations and, if necessary, click on a few supervoxels to correct or improve the calculated tumor segmentations. Further research is to be conducted, studying how to optimize the size and shape of the supervoxels - this is maybe a parameter that should be interactively adapted 'on-the-fly' - and how to speed up or parallelize the computation of the supervoxels.

Several side projects have been investigated related to brain tumor segmentation. One project investigates longitudinal brain tumor segmentation: instead of studying multi-modal brain tumor segmentation for single time points, it is investigated whether more reliable segmentations can be generated if MR scans from other time points can be taken into account. For this purpose, a conditional random field (CRF) is implemented, spanning over multi-temporal multi-modal image sequences at once. A growth model is incorporated in the CRF by including directed links with infinite weight encoding tumor growth or shrinkage in between every two subsequent time points. The framework is shown to generate accurate longitudinal brain tumor segmentations and is able, to some extent, to identify at which time point tumor regrowth occurs after therapy. The

adopted growth model can only identify tumor regrowth in between two subsequent time points: it doesn't predict for any future time point. Furthermore, it doesn't identify in what direction and at which location the tumor grows or shrinks. For this purpose, further research could be conducted, investigating how more elaborated tumor growth models, for example a diffusion-reaction tumor growth models as presented in [112], could be implemented in the context of a CRF.

In another project, a CRF is also adopted as a spatial regularizer, directly applied on the segmentation generated by a generative probabilistic model, but this time on uni-temporal multi-modal datasets. In this project, the uncertainty of the maximum a posteriori (MAP) solution obtained by the CRF is assessed by means of random MAP perturbations, directly applied on the unary potentials in the energy function. The project is shown to be valuable in the context of tumor volume measurements, and allows to separate stable and robust segmentations from uncertain segmentations, usually generated for difficult cases containing blurry tumor borders or resection cavities with internal bleedings. Uncertainty depends on the algorithms adopted for segmentation. This project demonstrated that segmentation uncertainty assessment is an important aspect to communicate to the end-user and seems to be particularly well-received by neuroradiologists. Therefore, modern segmentation frameworks should incorporate uncertainty assessment. Further research could be conducted investigating how to asses segmentation uncertainty in other state-of-the-art segmentation frameworks.

In the last project, an image classification task is studied in the field of radiogenomics. In this project, a tumor type is to be automatically recognized based on a set of multi-model MR images. Several image features are investigated – including short binary local image descriptors such as linear binary patterns (LBPs) and BRIEFs and low-dimensional image representations generated by deep convolutional denoising auto-encoders – in combination with principal component analysis (PCA) and a range of classical machine learning algorithms. Random forests (RFs) are shown to generate the best estimator, trained on all features without the application of PCA. Random forests are an ensemble machine learning method – which could explain why they generated the best estimator on the adopted – highly unbalanced – dataset. Further research should be conducted, investigated whether 3-dimensional convolutional neural networks (CNNs) could obtain better classification accuracies, as they are able to optimize feature extraction and classification at once. In the end, it should be noted that the classification task remains very challenging: experienced neuroradiologists are not able to identify these genetic tumor classes solely based on MR images. Our best estimator, in comparison, obtained accuracies which largely exceeds the accuracy obtained by a majority voter.

Bibliography

[1] Q. T. Ostrom, H. Gittleman, J. Fulop, M. Liu, R. Blanda, C. Kromer, Y. Wolinsky, C. Kruchko, and J. S. Barnholtz-Sloan. CBTRUS Statistical Report: Primary brain and central nervous system tumors diagnosed in the United States in 2008-2012. *Neuro-Oncology*, 17(Suppl 4):iv1–iv62, Oct 2015.

[2] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M. Weber, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, Oct 2015.

[3] T. Rohlfing, N. M. Zahr, E. V. Sullivan, and A. Pfefferbaum. The SRI24 multichannel atlas of normal adult human brain structure. *Human Brain Mapping*, 31(5):798–819, 2010.

[4] T. Huber, G. Alber, S. Bette, J. Kaesmacher, T. Boeckh-Behrens, J. Gempt, F. Ringel, H. M. Specht, B. Meyer, C. Zimmer, B. Wiestler, and J. S. Kirschke. Progressive disease in glioblastoma: Benefits and limitations of semi-automated volumetry. *PLOS ONE*, 12(2):1–15, Feb 2017.

[5] L. Scarpace, A. E. Flanders, R. Jain, T. Mikkelsen, and D. W. Andrews. Data from REMBRANDT. *The Cancer Imaging Archive*, 2015.

[6] S. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, Nov 2002.

[7] S. Bauer, T. Fejes, and M. Reyes. A skull-stripping filter for ITK. Jan 2013. URL: `http://hdl.handle.net/10380/3353`.

[8] B. H. Menze, K. van Leemput, D. Lashkari, M.-A. Weber, N. Ayache, and P. Golland. A generative model for brain tumor segmentation in multi-modal images. In T. Jiang, N. Navab, J. P. W. Pluim, and M. A. Viergever, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*, volume 6362, pages 151–159, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[9] D. N. Louis, H. Ohgaki, O. D. Wiestler, W. K. Cavenee, P. C. Burger, A. Jouvet, B. W. Scheithauer, and P. Kleihues. The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathologica*, 114(2):97–109, Aug 2007.

[10] P. Suetens. *Fundamentals of Medical Imaging*. Cambridge University Press, 2 edition, 2009.

[11] E. Alberts, G. Tetteh, S. Trebeschi, M. Bieth, A. Valentinitsch, B. Wiestler, C. Zimmer, and B. H. Menze. Multi-modal image classification using low-dimensional texture features for genomic brain tumor recognition. In M. J. Cardoso, T. Arbel, E. Ferrante, X. Pennec, A. V. Dalca, S. Parisot, S. Joshi, N. K. Batmanghelich, A. Sotiras, M. Nielsen, M. R. Sabuncu, T. Fletcher, L. Shen, S. Durrleman, and S. Sommer, editors, *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*, pages 201–209, Cham, 2017. Springer International Publishing.

[12] P. Eichinger, E. Alberts, C. Delbridge, S. Trebeschi, A. Valentinitsch, S. Bette, T. Huber, J. Gempt, B. Meyer, J. Schlegel, C. Zimmer, J. S. Kirschke, B. H. Menze, and B. Wiestler. Diffusion tensor image features predict IDH genotype in newly diagnosed WHO grade II/III gliomas. *Scientific Reports*, 7(13396), Oct 2017.

[13] P. A. McKinney. Brain tumours: incidence, survival, and aetiology. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(suppl 2):ii12–ii17, 2004.

[14] Y. Tarabalka, G. Charpiat, L. Brucker, and B. H. Menze. Spatio-temporal video segmentation with shape growth or shrinkage constraint. *IEEE Transactions on Image Processing*, 23(9):3829–3840, 2014.

[15] T. Hazan, S. Maji, and T. Jaakkola. On sampling from the Gibbs distribution with random maximum a-posteriori perturbations. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, pages 1268–1276, USA, 2013. Curran Associates Inc.

[16] G. Papandreou and A. Yuille. Perturb-and-MAP random fields: Using discrete optimization to learn and sample from energy models. In *IEEE International Conference on Computer Vision (ICCV)*, pages 193–200. Nov 2011.

[17] D. E. Reuss, F. Sahm, D. Schrimpf, B. Wiestler, D. Capper, C. Koelsche, L. Schweizer, A. Korshunov, D. T. W. Jones, V. Hovestadt, M. Mittelbronn, J. Schittenhelm, C. Herold-Mende, A. Unterberg, M. Platten, M. Weller, W. Wick, S. M. Pfister, and A. von Deimling. ATRX and IDH1-R132H immunohistochemistry with subsequent copy number analysis and IDH sequencing as a basis for an "integrated" diagnostic approach for adult astrocytoma, oligodendroglioma and glioblastoma. *Acta Neuropathologica*, 129(1):133–146, Jan 2015.

[18] A. Olar, K. M. Wani, K. D. Alfaro-Munoz, L. E. Heathcock, H. F. van Thuijl, M. R. Gilbert, T. S. Armstrong, E. P. Sulman, D. P. Cahill, E. Vera-Bolanos, Y. Yuan, J. C. Reijneveld, B. Ylstra, P. Wesseling, and K. D. Aldape. IDH mutation status and role of WHO grade and mitotic index in overall survival in grade II–III diffuse gliomas. *Acta Neuropathologica*, 129(4):585–596, Apr 2015.

[19] A. G. Fritz. *International classification of diseases for oncology: ICD-O*. World Health Organization, 2000.

[20] M. L. Goodenberger and R. B. Jenkins. Genetics of adult glioma. *Cancer Genetics*, 205(12):613–621, Dec 2012.

[21] Z.-X. Lin. Glioma-related edema: new insight into molecular mechanisms and their clinical implications. *Chinese Journal of Cancer*, 32(1):49–52, 2013.

[22] A. Lieu, S. Hwang, S. Howng, and C. Chai. Brain tumors with hemorrhage. *Journal Of The Formosan Medical Association*, 98(5):356–363, May 1999.

[23] S. Kumar, A. Handa, R. Sinha, and R. Tiwari. Bilateral cystic glioblastoma multiforme. *Journal of Neurosciences in Rural Practice*, 4(4):476–477, 2013.

[24] D. N. Louis, A. Perry, G. Reifenberger, A. von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, and D. W. Ellison. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathologica*, 131(6):803–820, Jun 2016.

[25] B. Stewart and C. Wild. *World Cancer Report 2014*. World Health Organization, 2014.

[26] V. Dunet, C. Rossier, A. Buck, R. Stupp, and J. O Prior. Performance of F-18-Fluoro-Ethyl-Tyrosine (F-18-FET) PET for the differential diagnosis of primary brain tumor: A systematic review and metaanalysis. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, 53:207–14, Feb 2012.

[27] J. B. Little. Radiation carcinogenesis. *Carcinogenesis*, 21(3):397–404, 2000.

[28] P. Korfiatis and B. Erickson. The basics of diffusion and perfusion imaging in brain tumors. *Applied Radiology*, Jul 2014.

[29] F. C. Ghesu, B. Georgescu, Y. Zheng, S. Grbic, A. Maier, J. Hornegger, and D. Comaniciu. Multi-scale deep reinforcement learning for real-time 3D-landmark detection in CT scans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018.

[30] T. Huang. Computer Vision: Evolution And Promise. *CERN european organization for nuclear research report – CERN*, 8:21–26, 1996.

[31] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer, 2010.

[32] P. McCullagh. What is a statistical model? *The Annals of Statistics*, 30(5):1225–1310, Oct 2002.

[33] A. M. Walker. On the asymptotic behavior of posterior distributions. *Journal of the Royal Statistical Society, Series B*, 31:80–88, 1969.

[34] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[35] J. Uffink. Boltzmann's work in statistical physics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017 edition, 2017.

[36] H. Ritzema, I. I. for Land Reclamation, and Improvement. *Drainage Principles and Applications*. ILRI Publication. International Institute for Land Reclamation and Improvement, 1994.

[37] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior. The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, Dec 2013.

[38] E. Alberts, G. Charpiat, Y. Tarabalka, T. Huber, M.-A. Weber, J. Bauer, C. Zimmer, and B. H. Menze. A nonparametric growth model for brain tumor segmentation in longitudinal MR sequences. In A. Crimi, B. Menze, O. Maier, M. Reyes, and H. Handels, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 69–79, Cham, 2016. Springer International Publishing.

[39] E. Alberts, M. Rempfler, G. Alber, T. Huber, J. Kirschke, C. Zimmer, and B. H. Menze. Uncertainty quantification in brain tumor segmentation using CRFs and random perturbation models. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 428–431, Apr 2016.

[40] W. Bidgood, S. Horii, F. Prior, and D. Van Syckle. Understanding and using DICOM, the data interchange standard for biomedical imaging. *Journal of the American Medical Informatics Association*, 4(3):199–212, May-Jun 1997.

[41] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006.

[42] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J. V. Miller, S. Pieper, and R. Kikinis. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magnetic Resonance Imaging*, 30(9):1323 – 1341, 2012. Quantitative Imaging in Cancer.

[43] S. Ourselin, A. Roche, G. Subsol, X. Pennec, and N. Ayache. Reconstructing a 3D structure from serial histological sections. *Image and Vision Computing*, 19(1):25 – 31, 2001.

[44] S. Ourselin, R. Stefanescu, and X. Pennec. Robust registration of multi-modal images: Towards real-time clinical applications. In T. Dohi and R. Kikinis, editors, *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2002*, pages 140–147, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[45] B. Avants, N. Tustison, and G. Song. Advanced Normalization Tools: V1.0. July 2009. URL: `http://hdl.handle.net/10380/3113`.

[46] J. Juntu, J. Sijbers, D. Van Dyck, and J. Gielen. Bias field correction for MRI images. In M. Kurzyński, E. Puchała, M. Woźniak, and A. żołnierek, editors, *Computer Recognition Systems*, pages 543–551, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[47] U. Vovk, F. Pernus, and B. Likar. A review of methods for correction of intensity inhomogeneity in MRI. *IEEE Transactions on Medical Imaging*, 26(3):405–421, Mar 2007.

[48] N. Tustison, B. Avants, P. Cook, Y. Zheng, A. Egan, P. Yushkevich, and J. Gee. N4itk: Improved n3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, 2010.

[49] J. Iglesias, C. Liu, P. Thompson, and Z. Tu. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*, 30(9):1617–1634, 2011.

[50] M. Shah, Y. Xiao, N. Subbanna, S. Francis, D. L. Arnold, D. L. Collins, and T. Arbel. Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. *Medical Image Analysis*, 15(2):267–282, 2011.

[51] X. Sun, L. Shi, Y. Luo, W. Yang, H. Li, P. Liang, K. Li, V. C. T. Mok, W. C. W. Chu, and D. Wang. Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions. *BioMedical Engineering OnLine*, 14(1):73, Jul 2015.

[52] J. Kleesiek, A. Biller, G. Urban, U. Köthe, M. Bendszus, and F. A. Hamprecht. ilastik for multi-modal brain tumor segmentation. In *Proceedings MICCAI BraTS (Brain Tumor Segmentation) Challenge*, pages 12–17, 2014.

[53] K. V. Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated model-based tissue classification of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):897–908, Oct 1999.

[54] B. H. Menze, K. V. Leemput, D. Lashkari, T. Riklin-Raviv, E. Geremia, E. Alberts, P. Gruber, S. Wegener, M. A. Weber, G. Székely, N. Ayache, and P. Golland. A generative probabilistic model and discriminative extensions for brain lesion segmentation – with application to tumor and stroke. *IEEE Transactions on Medical Imaging*, 35(4):933–946, Apr 2016.

[55] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, Aug 1999.

[56] A. Gooya, K. M. Pohl, M. Bilello, G. Biros, and C. Davatzikos. Joint segmentation and deformable registration of brain scans guided by a tumor growth model. In G. Fichtinger, A. Martel, and T. Peters, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011*, pages 532–540, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[57] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.

[58] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[59] J. H. Kappes, B. Andres, F. A. Hamprecht, C. Schnörr, S. Nowozin, D. Batra, S. Kim, B. X. Kausler, T. Kröger, J. Lellmann, N. Komodakis, B. Savchynskyy, and C. Rother. A comparative study of modern inference techniques for structured discrete energy minimization problems. *International Journal of Computer Vision*, 115(2):155–184, 2015.

[60] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

[61] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 109–117. Curran Associates, Inc., 2011.

[62] A. Criminisi and J. Shotton. *Decision Forests for Computer Vision and Medical Image Analysis*. Springer Publishing Company, Incorporated, 2013.

[63] M. P. Heinrich and M. Blendowski. Multi-organ segmentation using vantage point forests and binary context features. In S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 598–606, Cham, 2016. Springer International Publishing.

[64] R. Meier, S. Bauer, J. Slotboom, R. Wiest, and M. Reyes. Appearance-and context-sensitive features for brain tumor segmentation. In *Proceedings MICCAI BraTS (Brain Tumor Segmentation) Challenge*, pages 20–26, 2014.

[65] K. Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1(2):119 – 130, 1988.

[66] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, December 1989.

[67] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. `arXiv:1412.6980`.

[68] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. `arXiv:1502.03167`.

[69] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[70] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2843–2851. Curran Associates, Inc., 2012.

[71] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.

[72] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017.

[73] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014. `arXiv:1411.4038`.

[74] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[75] H. Shen, R. Wang, J. Zhang, and S. J. McKenna. Boundary-aware fully convolutional network for brain tumor segmentation. In M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017*, pages 433–441, Cham, 2017. Springer International Publishing.

[76] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage*, 170:446–455, 2018.

[77] L. Fidon, W. Li, L. C. Garcia-Peraza-Herrera, J. Ekanayake, N. Kitchen, S. Ourselin, and T. Vercauteren. Scalable multimodal convolutional networks

for brain tumour segmentation. In M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017*, pages 285–293, Cham, 2017. Springer International Publishing.

[78] R. M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, May 1979.

[79] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision & Pattern Recognition (CVPR '01)*, volume 1, pages I–I, Dec 2001.

[80] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In C. Schmid, S. Soatto, and C. Tomasi, editors, *International Conference on Computer Vision & Pattern Recognition (CVPR '05)*, volume 1, pages 886–893, San Diego, United States, June 2005. IEEE Computer Society.

[81] M. Idrissa and M. Acheroy. Texture classification using Gabor filters. *Pattern Recognition Letters*, 23(9):1095 – 1102, 2002.

[82] S. Otálora, A. Cruz-Roa, J. Arevalo, M. Atzori, A. Madabhushi, A. R. Judkins, F. González, H. Müller, and A. Depeursinge. Combining unsupervised feature learning and Riesz wavelets for histopathology image representation: Application to identifying anaplastic medulloblastoma. In N. Navab, J. Hornegger, W. M. Wells, and A. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 581–588, Cham, 2015. Springer International Publishing.

[83] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation-invariant texture classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[84] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua. BRIEF: Computing a local binary descriptor very fast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1281–1298, Jul 2012.

[85] J. Banerjee, A. Moelker, W. J. Niessen, and T. van Walsum. 3D LBP-based rotationally invariant region description. In J.-I. Park and J. Kim, editors, *Computer Vision - ACCV 2012 Workshops*, pages 26–37, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[86] J. Kivinen and C. Williams. Multiple texture Boltzmann machines. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research (PMLR)*, pages 638–646, La Palma, Canary Islands, Apr 2012.

[87] I. Levner, S. Drabycz, G. Roldan, P. De Robles, J. G. Cairncross, and R. Mitchell. Predicting MGMT methylation status of glioblastomas from MRI texture. In G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, and C. Taylor, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*, pages 522–530, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[88] A. Paul, A. Dey, D. P. Mukherjee, J. Sivaswamy, and V. Tourani. Regenerative random forest with automatic feature selection to detect mitosis in histopathological breast cancer images. In N. Navab, J. Hornegger, W. M. Wells, and A. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 94–102, Cham, 2015. Springer International Publishing.

[89] X. Chen, Y. Xu, S. Yan, D. W. K. Wong, T. Y. Wong, and J. Liu. Automatic feature learning for glaucoma detection based on deep learning. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 669–677, Cham, 2015. Springer International Publishing.

[90] M. Kotti, L. D. Duffell, A. A. Faisal, and A. H. McGregor. Detecting knee osteoarthritis and its discriminating parameters using random forests. *Medical Engineering & Physics*, 43:19 – 29, 2017.

[91] A. Taalimi, S. Ensafi, H. Qi, S. Lu, A. A. Kassim, and C. L. Tan. Multimodal dictionary learning and joint sparse representation for HEp-2 cell classification. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 308–315, Cham, 2015. Springer International Publishing.

[92] S. Azizi, F. Imani, B. Zhuang, A. Tahmasebi, J. T. Kwak, S. Xu, N. Uniyal, B. Turkbey, P. Choyke, P. Pinto, B. Wood, M. Moradi, P. Mousavi, and P. Abolmaesumi. Ultrasound-based detection of prostate cancer using automatic feature selection with deep belief networks. In N. Navab, J. Hornegger, W. M. Wells, and A. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 70–77, Cham, 2015. Springer International Publishing.

[93] S. Hor and M. Moradi. Scandent Tree: A random forest learning method for incomplete multimodal datasets. In N. Navab, J. Hornegger, W. M. Wells, and A. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 694–701, Cham, 2015. Springer International Publishing.

[94] S. Bauer, J. Tessier, O. Krieter, L.-P. Nolte, and M. Reyes. Integrated spatio-temporal segmentation of longitudinal brain tumor imaging studies. In B. Menze, G. Langs, A. Montillo, M. Kelm, H. Müller, and Z. Tu, editors, *Medical Computer Vision. Large Data in Medical Imaging*, pages 74–83, Cham, 2014. Springer International Publishing.

[95] E. Angelini, O. Clatz, E. Mandonnet, E. Konukoglu, L. Capelle, and H. Duffau. Glioma dynamics and computational models: A review of segmentation, registration, and in silico growth algorithms and their clinical applications. *Current Medical Imaging Reviews*, 3(4):262–176, 2007.

[96] S. Candemir and Y. S. Akgül. Adaptive regularization parameter for graph cut segmentation. In A. Campilho and M. Kamel, editors, *Image Analysis and Recognition*, pages 117–126, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[97] P. Kohli and P. Torr. Measuring uncertainty in graph cut solutions. *Computer Vision and Image Understanding*, 112(1):30–38, Oct 2008.

[98] A. C. Fan, J. W. Fisher, W. M. Wells, J. J. Levitt, and A. S. Willsky. MCMC curve sampling for image segmentation. In N. Ayache, S. Ourselin, and A. Maeder, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007*, volume 4792 of *LNCS*, pages 477–485. Springer Berlin Heidelberg, 2007.

[99] A. Barbu and S.-C. Zhu. Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1239–1253, Aug 2005.

[100] M. Holzer and R. Donner. Over-segmentation of 3D medical image volumes based on monogenic cues. In Z. Kúkelová and J. Heller, editors, *Proceedings of the Computer Vision Winter Workshop (CVWW) 2014*, pages 35–42, 2014.

[101] B. Wiestler, D. Capper, M. Sill, D. T. W. Jones, V. Hovestadt, D. Sturm, C. Koelsche, A. Bertoni, L. Schweizer, A. Korshunov, et al. Integrated DNA methylation and copy-number profiling identify three clinically and biologically relevant groups of anaplastic glioma. *Acta Neuropathologica*, 128(4):561–571, Oct 2014.

[102] C. Lian, S. Ruan, T. Denœux, H. Li, and P. Vera. Dempster-Shafer theory based feature selection with sparse constraint for outcome prediction in cancer therapy. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 695–702, Cham, 2015. Springer International Publishing.

[103] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov 2004.

[104] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In A. Leonardis, H. Bischof, and A. Pinz, editors, *Computer Vision – ECCV 2006*, pages 404–417, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[105] X. Chen, Y. Xu, S. Yan, T.-S. Chua, D. W. K. Wong, T. Y. Wong, and J. Liu. Discriminative feature selection for multiple ocular diseases classification by sparse

induced graph regularized group Lasso. In N. Navab, J. Hornegger, W. M. Wells, and A. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 11–19, Cham, 2015. Springer International Publishing.

[106] A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In M. Everingham, C. Needham, and R. Fraile, editors, *BMVC 2008 - 19th British Machine Vision Conference*, pages 275:1–10, Leeds, United Kingdom, September 2008. British Machine Vision Association.

[107] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[108] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, December 2010.

[109] P. Therasse, S. G. Arbuck, E. A. Eisenhauer, J. Wanders, R. S. Kaplan, L. Rubinstein, J. Verweij, M. Van Glabbeke, A. T. van Oosterom, M. C. Christian, and S. G. Gwyther. New guidelines to evaluate the response to treatment in solid tumors. *JNCI: Journal of the National Cancer Institute*, 92(3):205–216, 2000.

[110] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio. HeMIS: Hetero-Modal Image Segmentation. In S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 469–477, Cham, 2016. Springer International Publishing.

[111] K. V. Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE Transactions on Medical Imaging*, 20(8):677–688, Aug 2001.

[112] A. Gooya, K. M. Pohl, M. Bilello, L. Cirillo, G. Biros, E. R. Melhem, and C. Davatzikos. GLISTR: Glioma Image Segmentation and Registration. *IEEE Transactions on Medical Imaging*, 31(10):1941–1954, Oct 2012.

[113] S. Bakas, K. Zeng, A. Sotiras, S. Rathore, H. Akbari, B. Gaonkar, M. Rozycki, S. Pati, and C. Davatzikos. GLISTRboost: Combining multimodal mri segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation. In A. Crimi, B. Menze, O. Maier, M. Reyes, and H. Handels, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 144–155, Cham, 2016. Springer International Publishing.

[114] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques – Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[115] M. Prastawa, E. Bullitt, S. Ho, and G. Gerig. A brain tumor segmentation framework based on outlier detection. *Medical Image Analysis*, 8(3):275–83, Oct 2004.

[116] R. Meier, U. Knecht, R. Wiest, and M. Reyes. CRF-based brain tumor segmentation: Alleviating the shrinking bias. In A. Crimi, B. Menze, O. Maier, M. Reyes, S. Winzeck, and H. Handels, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 100–107, Cham, 2016. Springer International Publishing.

[117] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, Nov 2012.

[118] S. Pereira, R. Meier, R. McKinley, R. Wiest, V. Alves, C. A. Silva, and M. Reyes. Enhancing interpretability of automatically extracted machine learning features: application to a RBM-random forest system on brain lesion segmentation. *Medical Image Analysis*, 44:228–244, 2018.

[119] C. A. Klaassen and J. A. Wellner. Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, 3(1):55–77, Feb 1997.

[120] F. Han, T. Zhao, and H. Liu. CODA: High dimensional copula discriminant analysis. *Journal of Machine Learning Research*, 14(1):629–671, February 2013.

[121] G. Elidan. Copula Bayesian networks. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems (NIPS) – Volume 1*, NIPS'10, pages 559–567, USA, 2010. Curran Associates Inc.

[122] H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, Aug 2012.

[123] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin. Fast free-form deformation using graphics processing units. *Computer Methods and Programs in Biomedicine*, 98(3):278–284, 2010.