ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY

DOCTORAL (PHD) THESIS

ÁGNES GYÖRFI

Automatic Detection and Segmentation of Brain Tumors in Multispectral MRI Records

Supervisors:

Prof. Dr. Levente Kovács
Prof. Dr. László Szilágyi

DOCTORAL SCHOOL OF APPLIED INFORMATICS AND APPLIED MATHEMATICS

Budapest, November 21, 2024

Contents

List of Figures

List of Tables

Abbreviations

ACC	Rate of correct decisions - Accuracy
ANN	Artificial Neural Network
BDT	Binary Decision Tree
BraTS	Brain Tumor Segmentation (Challenge)
CSF	Cerebrospinal Fluid
CNN	Convolutional Neural Network
CNS	Central Nervous System
CRF	Conditional Random Fields
CT	Computed Tomography
DCNN	Deep Convolutional Neural Network
DICOM	Digital Imaging and Communications in Medicine
DSC	Dice Similarity Score
ERT	Extremely Randomized Trees
FCM	Fuzzy C-means
FCNN	Fully Convolutional Neural Network
FLAIR	Fluid-Attenuated Inversion Recovery
GM	Grey Matter
GT	Ground Truth
HGG	High-Grade Glioma
INU	Intensity Non-Uniformity
KNN	K-Nearest Neighbours
LGG	Low-Grade Glioma
MICCAI	Medical Image Computation and Computer Assisted Intervention
MLP	Multi-Layer Perceptron
MCCNN	Multi-Cascaded Convolutional Neural Network
MRF	Markov Random Field
MRI	Magnetic Resonance Imaging
MRS	Magnetic Resonance Spectroscopy
OOB	Out-Of-Bag
PCA	Principal Component Analysis
PET	Positron Emission Tomography
PPV	Positive Predictive Value - Precision
RA	Rate of Accuracy
RF	Random Forest
ROI	Region of Interest

RU	Rate of Usage
SLIC	Simple Linear Iterative Clustering
SPECT	Single Photon Emission Computed Tomography
SVM	Support Vector Machine
TNR	True Negative Rate - Specificity
TPR	True Positive Rate - Sensitivity
WHO	World Health Organization
WM	White Matter

Preface

Cancers of the brain and the central nervous system result in over two hundred thousand deaths annually worldwide [1]. A significant part of these could be prevented with early diagnosis. The need for automated procedures that efficiently, quickly, and reliably pre-screen images and recommend cases suspected to be abnormal to physicians has become increasingly evident due to the sheer volume of image data produced by medical imaging equipment today.

The automated detection and segmentation of brain tumors in Magnetic Resonance Imaging (MRI) data has been a current research topic during the last three decades [2]. The Brain Tumor Segmentation Challenges (BraTS) organized annually since 2012 have intensified the research work in this field, providing standardized training, testing data, and uniform assessment frameworks [3, 4].

Understanding the prognosis of brain tumors is crucial for patient management. Factors such as tumor type, grade, and early detection play significant roles. The World Health Organization classifies brain tumors into different grades, each with its own prognosis and treatment approach.

This study aims to propose a solution for brain tumor segmentation that produces high-quality results with reduced computational requirements, utilizing classification via ensemble learning [2] assisted by several image processing tasks designed for MRI records likely to contain focal lesions. Further on, a U-Net based approach is proposed with the aim of producing segmentation of improved quality using state-of-the-art technologies of machine learning.

The first thesis group addresses important issues of MRI data preprocessing. It investigates and recommends the most suitable histogram equalization method for high-quality brain MRI segmentation with and without focal lesions. It proposed a feature generation and selection scheme to reduce the number of features required for segmentation without compromising quality. It explores the effect of spectral resolution on segmentation accuracy and recommends a reduction in storage space for MRI data representation. Finally it introduces an atlas-based [5, 6] data enhancement technique that improved segmentation accuracy for tumor volumes.

The second thesis group compares ensemble learning and deep learning methods [7, 8, 9] for brain tumor segmentation, highlighting the superior performance of deep learning methods. The study analyzes the accuracy and efficiency of various classifiers, showcasing the notable capabilities of deep learning in capturing complex patterns and spatial relationships within medical images. Unlike traditional handcrafted feature-based approaches, deep learning methods demonstrate a high degree of adaptability and scalability. The substantial improvement in segmentation accuracy highlight deep learning methods as a promising avenue for enhancing the efficiency and reliability of brain tumor analysis.

The third thesis group formulates a two-stage post-processing method and incorporates a secondary U-Net convolutional neural network [10] for refining and enhancing brain tumor segmentation accuracy. By leveraging classical machine learning-based and structural post-processing techniques, the study achieves significantly higher levels of accuracy and reliability in the segmentation results. This refined segmentation output serves as a valuable foundation for subsequent analyzes and clinical applications, ensuring accurate information is obtained for diagnosis and treatment planning. The integrated approach of classical machine learning and deep learning methods has proven its effectiveness in improving the quality of segmentation of brain tumors, ultimately benefiting both clinicians and researchers in the field of neuroimaging.

Introduction

Cancers of the brain and the central nervous system (CNS) cause the death of over two hundred thousand people every year. Between 1990 and 2016, the incidence of CNS cancer increased globally by 17.3% [1].

In 2023, an approximate 24,810 adults in the United States are projected to be diagnosed with primary cancerous brain and spinal cord tumors [11]. While primary brain tumors account for 85% to 90% of all primary central nervous system tumors, secondary brain tumors, or brain metastases, also exist. The latter occurs when a tumor originates elsewhere in the body and spreads to the brain. Additionally, the estimated number of children under 20 diagnosed with central nervous system tumors in the United States is 5,230 for the same year [12].

Today, medical imaging equipment is producing ever-increasing amounts of image data. Sooner or later, trained professionals working in the healthcare system will not be able to analyze every image in detail due to the sheer volume of data. Therefore, there is a need to develop automated procedures that can efficiently, quickly and reliably prescreen images and recommend cases that are suspected to be abnormal to the physician. The Brain Tumor Segmentation Challenges (BraTS), annually organized since 2012, have provided the research community with standardized training and testing data and a uniform assessment framework, leading to an intensification of the research field [3]. This has led to the spread of advanced techniques, particularly those based on convolutional neural networks (CNNs), which have been widely adopted in recent years for brain tumor segmentation.

Most tumors are diagnosed after their symptoms convince the patients to go to the doctor. In this case the tumor is detected in a certain advanced stage, when the chances of survival are reduced. The development of imaging devices and computers enable us to elaborate solutions that would allow for regular screening of a larger population and tracing most tumors in an earlier phase. Beside establishing the diagnosis, the automatic segmentation and quantitative analysis can assist therapy planning and evolution tracking of the tumor.

2.1 Brain tumor

Understanding the prognosis of brain tumors is crucial for patient management. Factors such as tumor type, grade, and early detection play a significant role. The World Health Organization (WHO) classifies brain tumors into different grades, each with its own prognosis and treatment approach. Table 2.1 provides an overview of the WHO Brain Tumor Grades, distinguishing between low-grade gliomas (LGG), which are slowly growing tumors classified as Grade I and II, and high-grade gliomas (HGG), which are rapidly growing, aggressive tumors classified as Grade III and IV [13].

Life expectancy after the diagnosis depends on several factors like: is it a primary tumor or a metastatic one, is it an aggressive form of tumor (also called high-grade glioma, HGG) or a less aggressive one (low-grade glioma, LGG), and of course, a key factor is how early the tumor is diagnosed [14]. Patients with HGG live fifteen months in average after diagnosis. With a LGG it is possible to live for several years, as this form of the tumor does not always require aggressive treatment immediately after the diagnosis. A detailed description of the number of deaths attributed to different types of tumors is provided by Ostrom *et al.* [15].

Table 2.1: World Health Organization (WHO) Brain Tumor Grades

LGG	Grade I	Pilocytic astrocytoma Craniopharyngioma Gangliocytoma Ganglioglioma
	Grade II	Diffuse astrocytoma Pineocytoma Pure oligodendroglioma
HGG	Grade III	Anaplastic astrocytoma Anaplastic ependymoma Anaplastic oligodendroglioma
	Grade IV	Glioblastoma multiforme Pineoblastoma Medulloblastoma Ependymoblastoma

2.2 Brain tumor imaging

Different imaging techniques are involved in brain tumor imaging, each of which has strengths and limitations. These methods include[16, 17, 18]:

- Computed Tomography (CT),
- Magnetic Resonance Imaging (MRI),

- Positron Emission Tomography (PET),
- Single Photon Emission Computed Tomography (SPECT),
- Magnetic Resonance Spectroscopy (MRS).

Magnetic resonance imaging (MRI) is the device that has become the most frequently utilized tool in the diagnosis of gliomas. MRI is preferred because it is much less invasive than other imaging modalities, like positron emission tomography, or computed tomography. With its high contrast and good resolution it can provide accurate data about the structure of the tumor. However, it also bears difficulties like the possible presence of intensity inhomogeneity [19], and the relative intensity values that vary from device to device and from patient to patient [20]. Multi-modal or multi-spectral MRI, through its T1-weighted (with or without contrast improvement) and T2-weighted (with or without Fluid Attenuated Inversion Recovery) data channels, can significantly contribute to the better visibility of intracranial structures [3].

Emphasis on crucial aspects such as segmentation, representation, and the incorporation of novel machine learning methodologies in decision-making processes characterizes the progress made in MRI processing for early brain tumor diagnosis and detection [21].

The quick development of computerized medical devices and the economic rise of several underdeveloped countries both contribute to the fast spreading of MRI equipment in hospitals worldwide. These MRI devices produce more and more image data. Training enough human experts to process these records would be very costly if possible at all. This is why there is a strong need for automatic algorithms that can reliably process the acquired image data and select those records which need to be inspected by the human experts, who have the final word in establishing the diagnosis. Although such algorithms are never perfect, they may contribute to cost reduction and allow for screening large masses of population, leading to early tumor detection.

2.3 Challenges

The brain tumor segmentation problem is a difficult task in medical image processing, having major obstacles like:

1. large variety of locations, shapes and appearances of the tumor;
2. displacement and distortion of normal tissues caused by the focal lesion;
3. the variety of imaging modalities and weighting schemes applied in MRI, which provide different types of biological information;
4. multi-channel data is not perfectly registered together;

5. numerical values produced by MRI do not directly reflect the observed tissues, they need to be interpreted in their context;
6. intensity inhomogeneity may be present in MRI measurements due to turbulence of the magnetic field.

Automatic tumor segmentation is not only utmost important task, but also a very challenging one, because of the high variety of anatomical structures and low contrast of current imaging techniques, which make the difference between normal regions and the tumor hardly recognizable for the human eye [2].

2.4 State of the art

The history of automatic brain tumor segmentation from MRI records can be divided in two eras: the pre-BraTS and the BraTS era, where BraTS refers to the Brain Tumor Segmentation challenges [3, 22] organized every year since 2012 by the Medical Image Computing and Computer-Assisted Intervention (MICCAI) society, which had an enormous impact with the introduction of a multi-spectral MRI data set that can be used as standard in the evaluation of segmentation methods. Earlier solutions were usually developed for single data channels and even for 2D data (or reduced number of distant slices), and mostly validated with private collections of MRI records that do not allow for objective comparison. A remarkable review on early segmentation methods, including manual, semi-automatic and fully automatic solutions, is provided by Gordillo *et al.* [2]. Methods developed in the BraTS era are usually fully automatic and employ either one or a combination of:

1. advanced general-purpose image segmentation techniques (mostly unsupervised);
2. classical machine learning algorithms (both supervised and unsupervised), or
3. deep learning convolutional neural networks (supervised learning methods).

All methods developed and published in the pre-BraTS era belong to the first two groups or their combination. They dominated the first years of the BraTS era as well, and they are still holding significant research interest. Unsupervised methods have the advantages that they do not require large amounts of training data and provide segmentation result in relatively short time. They organize the input data into several clusters, each consisting of highly similar data. However, they either have difficulty with correctly labeling the clusters, or they require manual interaction. Unsupervised methods involving active contours or region growing strongly depend on initialization as well. Sachdeva *et al.* [23] proposed a content-based active contour model that utilized both intensity and texture information to evolve the contour towards the tumor boundary. Njeh *et al.* [24]

proposed a quick unsupervised graph-cut algorithm that performed distribution matching and identified tumor boundaries from a single data channel. Li *et al.* [25] combined the fuzzy C-means (FCM) clustering algorithm with spatial region growing to segment the tumors, while Szilágyi *et al.* [26] proposed a cascade of FCM clustering steps placed into semi-supervised framework. Bal *et al.* [27] combines rough-fuzzy C-means (RFCM) with shape-based topological properties for automated brain tumor segmentation which outperforms hard C-means (HCM) and fuzzy C-means (FCM) methods.

Supervised learning methods deployed for tumor segmentation first construct a decision model using image based handcrafted features and use it for prediction in the testing phase. These methods mainly differ from each other in the employed classification technique, the features involved, and the extra processing steps that apply constraints to the intermediary segmentation outcome. Islam *et al.* [28] extracted so-called multi-fractional Brownian motion features to characterize the tumor texture, and compared its performance with Gabor wavelet features, using a modified AdaBoost classifier. Tustison *et al.* [29] built a supervised segmentation procedure by cascading two random forest (RF) classifiers and trained them using first order statistical, geometrical and asymmetry based features. They used a Markov random field (MRF) based segmentation to refine the probability maps provided by the random forests. Pinto *et al.* [30] deployed extremely randomized trees (ERT) and trained them with features extracted from local intensities and neighbourhood context, to perform a hierarchical segmentation of the brain tumor. Soltaninejad *et al.* [31] reformulated the simple linear iterative clustering (SLIC) [32] for the extraction of 3D superpixels, extracted statistical and texton feature from these, and fed them to a RF classifier to distinguish superpixels belonging to tumors and normal tissues. Imtiaz *et al.* [33] classified the pixels of MRI records with ERTs, using superpixel features from three orthogonal planes. Kalaiselvi *et al.* [34] proposed a brain tumor segmentation procedure that combined clustering techniques, region growing, and support vector machine (SVM) based classification. Supervised learning methods usually provide better segmentation accuracy than unsupervised ones and require longer processing time. Zhang *et al.* [35] employ support vector machine (SVM) classification integrated with feature selection in a kernel space. The selection criteria are based on kernel class separability, and this approach is extended to develop a framework for tracking brain tumor evolution over time. The follow-up framework involves learning the tumor and feature selection from the initial MRI examination, automatic segmentation of new data using SVM, and refinement of tumor contour through a region-growing technique.

Ensemble learning generally combines several classifiers and aggregates their predictions into a final one, thus allowing for more accurate decisions than the individual classifiers are capable of [36]. The ensembles deployed in this study consist of binary decision trees (BDT) [37] and are preferred because of their ability to learn any complex patterns that contain no contradiction.

In the recent years, convolutional neural networks (CNN) and deep learning are attempting to conquer a wide range of application fields in pattern recognition [38]. Features are not handcrafted anymore, as these architectures automatically build a hierarchy of increasingly complex features based on the learning data [39]. There are several successful applications in medical image processing, e.g. detection of kidney abnormalities [40], prostate cancer [41], lesions caused by diabetic retinopathy [42] and melanoma [43], and the segmentation of liver [44], cardiac structures [45], colon [46], renal artery [47], mandible [48], and bones [49], vertebrae [50], pancreas [51], lung [52], breast [53], anomalies of the heart [54], COVID-19 Detection in X-ray Images [55, 56] and hand vein network [57]. The brain tumor segmentation problem is not an exception. In this order, Pereira *et al.* [58] proposed a CNN architecture with small 3×3 sized kernels, and accomplished a thorough analysis of data augmentation techniques for glioma segmentation, attempting to compensate the imbalance of classes. Zhao *et al.* [59] applied conditional random fields (CRF) to process the segmentation output produced by a fully convolutional neural network (FCNN) and assure the spatial and appearance consistency. Wu *et al.* [60] proposed a so-called multi-feature refinement and aggregation scheme for convolutional neural networks that allows for a more effective combination of features and leads to more accurate segmentation. Kamnitsas *et al.* [61] proposed a deep CNN architecture with 3D convolutional kernels and a double pathway learning, which exploits a dense inference technique applied to image segments, thus reducing the computational burden. Ding *et al.* [62] successfully combined the deep residual networks with the dilated convolution, achieving fine segmentation results. Xue *et al.* [63] proposed a solution inspired by the concept of generative adversarial networks (GAN). They set up a CNN to produce pixel level labeling, complemented it with an adversarial critic network, and trained them to learn local and global features that were able to capture both short and long distance relationships between pixels. Chen *et al.* [64] proposed a dual force based learning strategy and employed the DeepMedic and U-Net architectures for glioma segmentation, and refined the output of deep networks using a multi-layer perceptron (MLP). In general, CNN architectures and deep learning can lead to slightly better accuracy than well designed classical machine learning techniques in the brain tumor detection problem, at the cost of much higher computational burden in all processing phases, especially at training.

Recently, Zhu *et al.* [65] proposed a CNN network that fuses the segmentation outcome based on deep semantic features with edge information extracted by a so-called edge spatial attention block. Cao *et al.* [66] proposed a so-called 3D shuffle attention module to be combined with a 3D convolutional neural network with multiple branches. Chang *et al.* [67] used a dual-path CNN architecture combined with a multi-scale attention fusion module and a 3D iterative dilated convolution merging module. Liu *et al.* [68] applied multiple changes to the 3D-U-Net architecture, replacing standard convolution by a hierarchical decoupled one, adding dilated convolution to the model, and introducing an attention mechanism to the output layer. Hu *et al.* [69] combines a Multi-Cascaded Convolutional Neural Network (MCCNN) with fully connected conditional random fields

(CRFs) to achieve robust segmentation results. The method demonstrates competitive performance across multiple databases and provides a promising direction for future research, particularly in the integration of information from more perspectives into 3D CNNs for enhanced segmentation.

Abut *et al.* [70] provides a comprehensive overview of the transition from Artificial Neural Networks (ANN) to Deep Convolutional Neural Networks (DCNN) in the field of medical image processing. It emphasizes the growing complexity and increasing dimensionality of medical data due to technological advancements and the proliferation of data sources. The article concludes by underlining the opportunities presented by DCNN and emphasizes the need for medical professionals to develop a deeper understanding of operational principles to effectively collaborate with data scientists.

Yu *et al.* [9] elaborates on various commonly used CNN models such as AlexNet, GoogleNet, ResNet, R-CNN, and FCNN, also presenting their applications in medical image analysis.

Ranjbarzadeh *et al.* [71] raises a critical question about the meaningfulness of utilizing extensive computational resources and deep learning technology. It suggests considering other learning techniques that may offer faster, higher interpretability, and comparable performance with fewer resources and parameterization. This insightful perspective opens avenues for future research to explore efficient and resource-friendly approaches in the field of brain tumor segmentation.

Fernando *et al.* [8] emphasizes the potential of combining statistical and deep learning approaches to develop automated systems in clinical oncology. It provides a critical overview of existing methodologies and opens avenues for future research directions, including the gathering of larger databases, improving classification accuracy, and developing hybrid systems.

CNN architectures and deep learning usually lead to slightly better accuracy in segmentation quality than well designed classical machine learning based techniques, because they use much more parameters to learn the details of the training examples. Of course, this better quality comes with a longer computation time.

2.5 Background of research

The research team at the Sapientia University, which I joined in 2017, elaborated an ensemble learning based solution for the brain tumor segmentation problem [72], that stands at the foundations of my doctoral research work. Preliminary works of the team were presented in [26, 73, 74, 75]. The flowchart of the procedure can be found in Figure 2.1.

Segmentation is achieved by classifying the pixels, which is performed by an ensemble (forest) of binary decision trees. Before classification, the MRI data is fed to preprocessing steps, aiming to make the raw MRI data suitable for classification. The necessary and

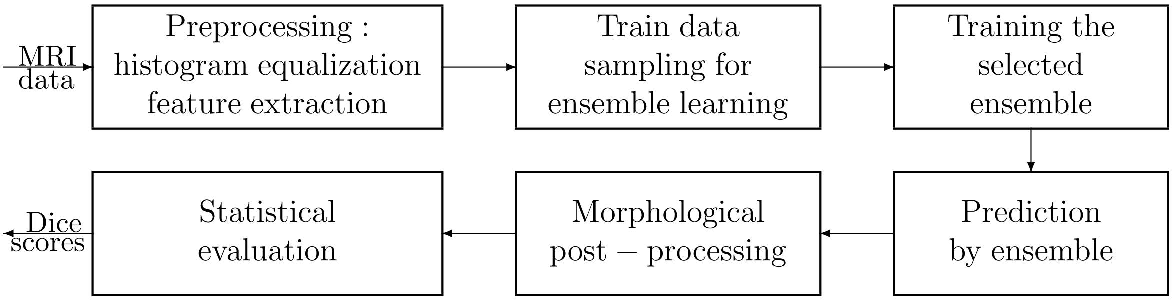

Figure 2.1: Block diagram of the initial brain tumor segmentation procedure.

possible steps of *preprocessing*, listed in order of execution are the following:

1. The intensity values of pixels in different MRI scans are not on the same scale, causing the same tissue types to appear differently in various scans. For a supervised learning intelligent algorithm to be able to generalize based on the rules learned from the training data, aligning the color scales of images becomes necessary. Hence, we incorporate histogram equalization during preprocessing, aligning the histograms of intensity values on each data channel separately. This step was implemented in the initial procedure through a context-dependent linear transformation.

2. Among the multi-spectral MRI data available at BraTS, not every pixel has measured intensity values on each data channel; in some cases, a zero value indicates missing data. We consider a pixel as part of the ROI (Region of Interest) if it has a measured value on at least one data channel. To compensate for missing data, we primarily use the conditional average intensity values of the 26 pixels that are direct spatial neighbours of the given pixel, or if this is not available, we use the average intensity value of the entire data channel for that particular scan.

3. Noise filtering is necessary when the available data require it. In this phase, we disregarded the inhomogeneous intensity commonly found in MRI scans, as the datasets published within the BraTS competition framework did not really contain this type of noise. Treating the inhomogeneous intensity can be considered a closed problem in MRI data processing, because it already has widely accepted and sufficiently accurate solutions [76].

4. Methods employing classical machine learning perform better in high-dimensional data spaces. However, it can also be said that not every possible detail of a voxel or the cubic millimeter of brain tissue it represents can be found in the four measured data channels provided by T1, T2, T1c, and FLAIR. Therefore, our initial procedure generated 100 additional features, thus using feature vectors with a total of 104 elements for training and testing. The generated features included various average, median, minimum, maximum, gradient, and Gabor wavelet features extracted from different planar and spatial neighbourhoods. Our initial procedure did not include feature selection.

Segmentation is performed by classifying the pixels using supervised machine learning techniques. The initial segmentation framework deployed an ensemble of binary decision trees (BDT) to accomplish the classification. A BDT of unlimited depth can describe any hierarchy of crisp (non-fuzzy) two-way decisions [37]. Further on, a BDT can perfectly learn the classification of any training data set if there is no contradiction in the data. The decision trees of the initial segmentation procedure learned to separate positive and negative pixels from the feature vectors of equal sized sets of randomly selected pixels. During learning, the minimum entropy criterion was used to select the choices and the depth of the decision trees was not limited. When estimating on test data, each vote in the forest was of equal value and the joint decision was made according to the majority of votes.

Post-processing represents a posterior relabeling scheme, which in the initial segmentation framework involved an unsupervised technique based on a simple morphological criterion. The input data of the post-processing step consisted in the labels provided by the BDT ensemble to all pixels in the test volume. For each pixel, the number of neighbours marked as positive within a predefined neighbourhood and the total number of neighbours were determined. The final label of a pixel became a tumor if and only if the proportion of positive neighbours exceeded the empirically determined threshold (1/3). The *evaluation* of the segmentation results involves statistical measures.

2.6 Proposed solutions

The vast majority of the papers published during my PhD studies optimized the operation of the brain tumor segmentation procedure used as a starting point, each paper focusing on a particular step in the processing. Each of these papers deals with the complete segmentation procedure, proposes some alternative solutions to the focused processing unit, and establishes the best performing version.

Paper [GyA4] identifies the histogram equalization method that best supports segmentation, comparing the simple linear transformation with the most commonly used method in the literature, the piecewise linear transformation [20] of László Nyúl, taking into account the possible parameters of both. Based on tests conducted on MRI scans with and without focal lesions, I determined that the piecewise linear transformation performs the best, and I have made suggestions for appropriate parameterization of the algorithm.

Paper [GyA1] introduced a feature selection procedure aimed at reducing the computational and memory requirements of the segmentation process without significantly diminishing the quality of the results. The proposed method performs the entire training and testing cycle iteratively using a gradually smaller set of features. During each iteration, I remove the least used features and those contributing to the worst decisions. The original feature set was successfully reduced to one-eighth of its original size, while the decrease in the average Dice score characterizing the final segmentation result remained below 0.3%.

Paper [GyA5] examined the impact of spectral resolution (color depth) of MRI data on the quality of segmentation. During histogram equalization, I had the opportunity to generate feature values with any number of bits in resolution. I investigated how the segmentation accuracy of MRI images with and without focal lesions changes in the 2-10 bit color depth range. Results show that classical machine learning methods can achieve their maximum segmentation accuracy using a 6-bit color depth, while any additional bits are largely redundant. Based on this observation, it is possible to reduce the memory space used for storing feature data without affecting segmentation accuracy. The [GyA2] paper is an early version of [GyA5] with fewer test cases, but similar results.

Paper [GyA3] proposed a multi-atlas based solution for enhancing the quality of MRI data. For a given learning task, I first aligned the complete MRI volumes using rigid registration, then considering only the pixels belonging to normal tissue types, I determined the local average and standard deviation of the intensity values of normal pixels, separately for each data channel. Subsequently, I modified the intensity values of every pixel in the learning and test dataset according to how much the local intensity differed from the local average value of normal intensities, how many times the difference is larger than the variance, and what is the sign of the difference. I used these modified intensity values for training and testing. Depending on the size of the atlas, I achieved an improvement of about 0.5-1% in the average Dice score indicators characterizing the quality of segmentation.

Papers [GyA10] and [GyA7] investigate the possible replacement of the BDT ensemble with other classical machine learning algorithms, in the role of the decision making unit used for segmentation. Both in case of testing with MRI records containing LGG and HGG tumors, I achieved the best segmentation accuracy with the random forest classifier.

Paper [GyA12], proposed an intelligent post-processing algorithm instead of the initially used morphological criterion, which re-evaluates the initially determined label of each pixel based on morphological features extracted from labels found in its predefined spatial neighbourhood. The proposed algorithm uses a random forest for decision-making. Compared to the simple morphological criterion, an improvement of up to 1% can be achieved in the final accuracy of the segmentation.

Paper [GyA11] introduced a structural post-processing operation. In the output of the initial segmentation, I first use spatial region growing to identify the contiguous regions classified as positive. I discard the very small positive regions and, for the others, I determine their three principal axes and corresponding sizes through principal component analysis. Using an empirically established criterion, the very flat positive regions (those with a small size in the third relevant dimension) are relabeled as negative, while the rest are definitively classified as positive.

Paper [GyA6] integrates all previously mentioned modifications simultaneously into the initial procedure, and a detailed evaluation process is carried out using various

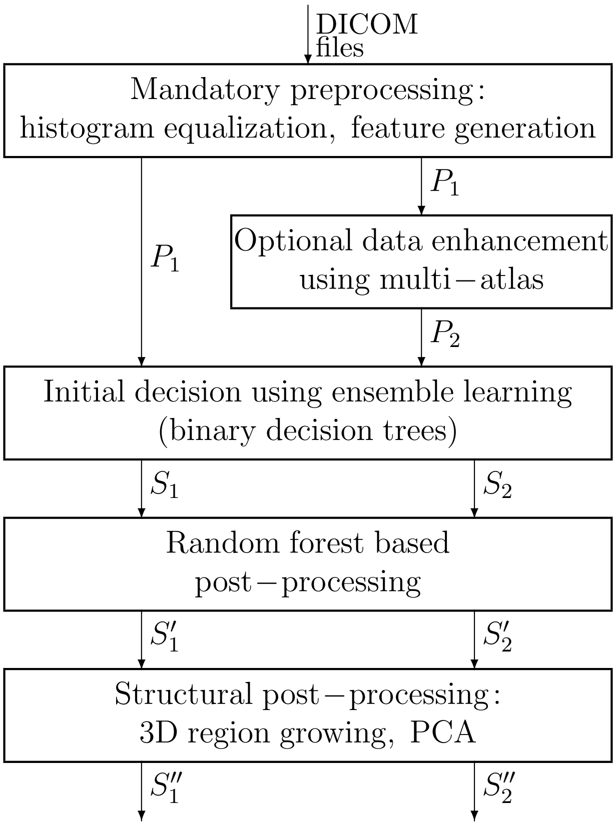

Figure 2.2: The input DICOM files are transformed into two variants of preprocessed data (P_1 and P_2). The intermediary segmentation results produced by the ensemble of binary decision trees (S_1 and S_2) are fine tuned in two post-processing steps to achieve high-quality segmentation (S''_1 and S''_2).

datasets. The flowchart of the modified procedure is shown in Figure 2.2. The modified procedure is competitive in terms of segmentation accuracy and efficiency with the best performing algorithms found in the literature.

During the later phase of my doctoral studies, I proposed a brain tumor segmentation procedure that incorporates U-Net [10] architectures. Paper [GyA8] introduced a U-Net cascade system, where two U-Net networks of identical type and size perform the classification and post-processing. The flowchart of the proposed procedure can be found in Figure 2.3. The first U-Net performs a primary segmentation on the MRI data, applying spatial convolution. It receives as input all four data channels of a multi-spectral MRI scan at once and outputs the estimated probability of each pixel belonging to the positive regions. From the output, I calculate four morphological features similarly to those described in paper [GyA12], and I provide the spatial distribution of these features as the input to the second U-Net network, which performs the post-processing and outputs the final estimated classification of each pixel. The segmentation accuracy achieved surpasses the results of my previous solution by an average of 1%. Paper [GyA9] involves an adaptive local histogram equalization algorithm (CLAHE) as a preprocessing step before the U-Net cascade, which further improves the segmentation accuracy.

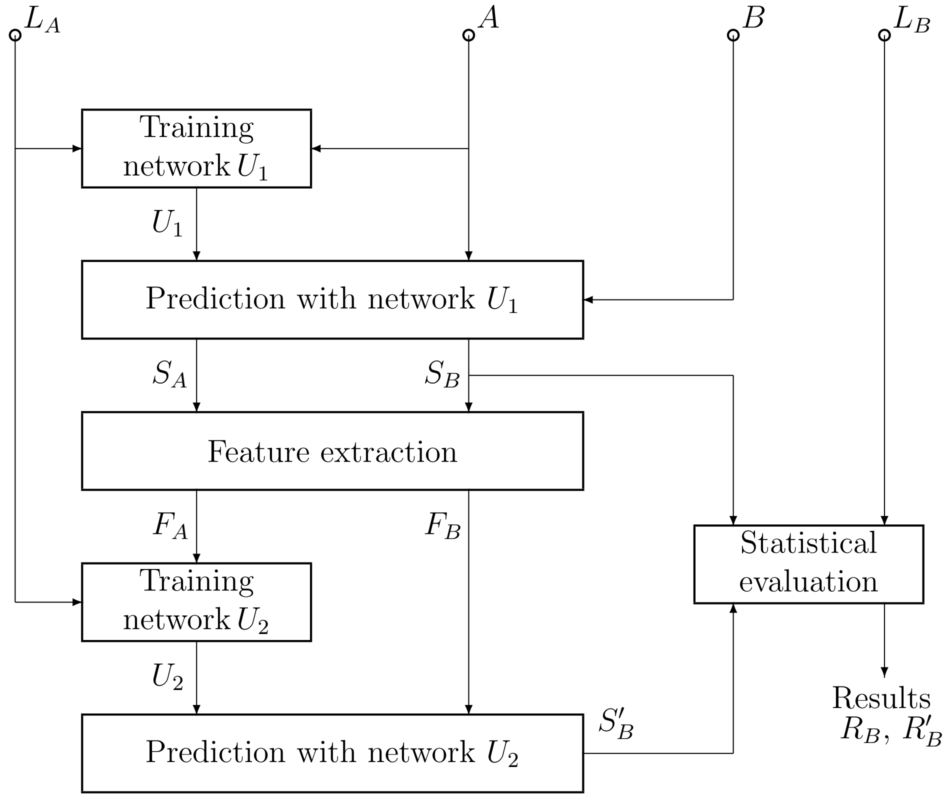

Figure 2.3: Block diagram of the proposed two-stage segmentation process. The first U-Net U_1 produces initial segmentation results, while the second U-Net U_2 performs a post-processing based on neighbourhood features extracted from the labels provided by U_1. A and B represent the set of pixels used for training and testing, respectively, while L_A are L_B stand for the corresponding ground truth labels.

Chapters 4-6 will concentrate on giving the necessary details about the proposed methods and achievements mentioned above.

Chapter 4 presents several attempts to optimize the preprocessing methods employed in the segmentation framework, having the goal to improve the overall segmentation accuracy produced by the framework. This involves revisiting histogram equalization, feature generation, spectral resolution, and the utilization of atlas-based enhancements to refine the segmentation process.

Chapter 5 summarizes my attempts to optimize the pixel classification process of the original segmentation framework, and some alternative segmentation methods proposed during my PhD research. This includes a comprehensive discussion of both classical methods and deep learning techniques.

Alternative post-processing techniques, mostly based on supervised machine learning methods are presented is Chapter 6.

Foundations

Throughout this section, I will present an overview of the datasets utilized in my study and establish the fundamental concepts and methodologies crucial for understanding and evaluating brain tumor segmentation algorithms.

3.1 Data

The datasets play a crucial role in training and evaluating the performance of the proposed segmentation methods. I utilized two distinct sets of MRI records, each with its own unique characteristics and challenges. The first dataset, referred to as 'BraTS data,' comprises volumes from the MICCAI Brain Tumor Segmentation Challenge, encompassing both low and high-grade glioma volumes from the years 2015 and 2019. The second dataset, known as 'i-Seg data,' encompasses infant brain MRI records sourced from the iSeg-2017 Challenge. This dataset poses specific challenges due to the overlapping intensity distributions of white and gray matter.

3.1.1 BraTS data

The Brain Tumor Segmentation Challenge is amongst the numerous challenges held in connection with the MICCAI (Medical Image Computing and Computer Assisted Intervention) conference [77]. Organized annually since 2012, this challenge has significantly intensified research in this field, resulting in several important solutions. These solutions typically leverage prior information and employ various image processing and pattern recognition methodologies.

Menze *et al.* [3] presents the results of the Multimodal Brain Tumor Image Segmentation Benchmark (BraTS) conducted in conjunction with the MICCAI 2012 and 2013 conferences. The study involved the application of twenty state-of-the-art tumor segmentation algorithms to a set of 65 multi-contrast MR scans of low- and high-grade glioma patients, manually annotated by up to four raters. The findings revealed considerable disagreement between human raters in segmenting various tumor sub-regions, highlighting the complexity of the task.

Bakas *et al.* [4] underscores the critical importance of accurate and reliable tumor classification for both clinical and computational studies. This precision can facilitate more accurate therapeutic planning and prognosis prediction. The article also highlights the

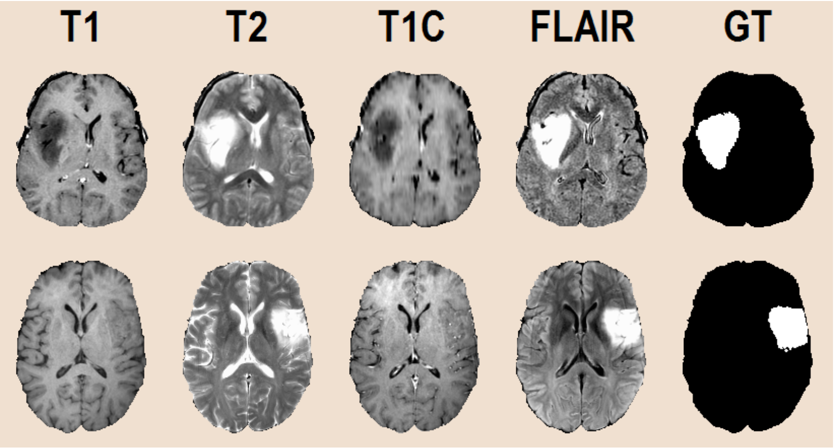

Figure 3.1: Input data after pre-processing (T1, T1C, T2, FLAIR) and the human expert-made ground truth (GT) for the whole tumor segmentation problem.

significance of publicly available data collections for the reproducibility and comparability of scientific research results.

The studies conducted during my PhD research use the MICCAI BraTS train data sets, both low and high grade glioma volumes of years 2015 and 2019. Some main attributes of these data sets are exhibited in Table 3.1. Only one out of these four data sets are used at a time, so in any scenario, the total number n_ρ of involved records varies between 54 and 259, as indicated in the first row of the table.

The records in each set have the same format. Records are multi-spectral, which means that every pixel in the volumes has four different observed intensity values (named T1, T2, T1C, and FLAIR after the weighting scheme used by the MRI device) recorded independently of each other and registered together afterwards with an automatic algorithm. Each volume contains 155 square shaped slices of 240 × 240 pixels. Pixels are isovolumetric as each of them represents brain tissues from a 1 mm^3 sized cubic region. Pixels were annotated by human experts using a semi-automatic algorithm, so they have a label that can be used as ground truth for supervised learning. Figure 3.1 shows two (preprocessed) arbitrary slices, with all four observed data channels, and the ground truth for the whole tumor, without distinguishing tumor parts. Since the adult human brain has a volume of approximately 1500 cm^3, records contain around 1.5 million pixels. Each record contains gliomas of total size between 7 and 360 cm^3. The skull is removed from all volumes so that the researchers can concentrate on classifying brain tissues only, but some of the records intentionally have missing or altered intensity values, in the amount of up to one third of the pixels, in one of the four data channels. An overview of these cases is also reported in Table 3.1.

Intensity non-uniformity (INU) is a low-frequency noise with possibly high magnitude [19, 78]. In the early years of BraTS challenges, INU was filtered from the train and test data provided by the organizers. However, in later data sets, this facility is not granted.

Tumors are present in each volume, but their position, size, and shape vary from

Table 3.1: The main properties of the MRI data sets involved in this study

Property	BraTS 2015 data		BraTS 2019 data	
	LGG	HGG	LGG	HGG
Number of records	54	220	76	259
Average size of whole tumor (cm^3)	101.1	110.6	111.5	95.1
Minimum size of whole tumor (cm^3)	18.9	8.5	17.1	7.3
Maximum size of whole tumor (cm^3)	256.3	318.3	361.8	256.9
Total number of negative pixels	71.4M	308.6M	101.6M	351.5M
Total number of positive pixels	5.46M	24.3M	8.48M	24.6M
Pixels with missing data	2.375%	9.855%	1.679%	0.426%
Pixels with more than 1 value missing	0.225%	4.838%	0.163%	0.092%
Records with missing data at $> 1\%$ of pixels	9	112	9	18
Records with missing data at $> 10\%$ of pixels	3	99	3	1

record to record. Further technical details of the BraTS data are presented in [3].

There are several non-brain pixels in the volume of any record, fact indicated by zero intensity on all four data channels. Furthermore, there are some brain pixels in most volumes, for which not all four nonzero intensity values exist, some of them are missing. Our first option to fill missing values is to replace them with the average computed from the $3 \times 3 \times 3$ cubic neighbourhood of the pixel, if there are any pixels in the neighbourhood with valid nonzero intensity. Otherwise, the missing value becomes $\gamma_{0.5}$ after preprocessing, which is the middle value of the target intensity interval, see definition in Section 4.2.1.

Within the BraTS 2019 dataset, a number of records exhibit instances of missing data in their pixels. In most cases, these gaps are isolated and can be remedied by substituting them with an averaged intensity value derived from neighbouring pixels. However, there exist certain volumes, particularly those from older sources such as the BraTS 2013 dataset, where significant portions lack valid data across one or more data channels. Notably, these volumes suffer from the absence of FLAIR data, a crucial component for precise segmentation [GyA1, 79]. While building the U-Net based segmentation procedure, to ensure robustness, we excluded 20 HGG and 10 LGG records from the training data of the BraTS 2019 dataset due to extensive missing data.

Ranjbarzadeh *et al.* [71](2023) specifically emphasizes in their overview article that the widespread utilization of the BraTS dataset is recognized, and the favorable metrics associated with it suggest its reliability for future studies.

3.1.2 i-Seg data

The infant brain MRI records involved in my research work originate from the iSeg-2017 Challenge [80]. Ten MRI records were published together with ground truth to serve as training data for machine learning based segmentation methods. Each record contains two data channels, namely T1 and T2, each with intensity values between 1 and 1000.

Zero intensity is reserved for external, non-brain pixels. The volume of each records may contain 256 slices of 144 × 192 pixels, but the number of slices that contain brain pixels is 107 at most. Each pixel represents a cubic millimeter of brain tissues. The number of brain pixels in a volume ranges between 700 and 900 thousand. Each pixel has a ground truth, indicating the tissue to which it belongs according to the human experts. In every volume there is maximum one pixel with a missing value. The intensity of such pixels was replaced with the average intensity of its immediate neighbours. The main difficulty of the infant brain segmentation problem is represented by the fact that the intensity of white matter (WM) and grey matter (GM) pixels is almost the same at the age of 6 months, the WM and GM histogram are severely overlapping each other. The i-Seg dataset is only involved in the histogram equalization study, to establish the most suitable algorithm that works on MRI data both with and without focal lesions.

3.2 Histogram equalization, uniformization

Magnetic resonance imaging (MRI) is a very popular technique in current medical diagnosis, due to its relatively high contrast and fine resolution, despite its drawback consisting in the fact that recorded numeric values do not directly reflect the observed tissues. The correct interpretation of observed images requires the adaptation of pixel intensities to their context, which is achieved via histogram equalization. The comparison of two intensity values from two different MRI records without having the histograms previously equalized, would be like comparing the value of two jewels by their weight and ignoring to check the precious metal they are made of. Figure 4.1 presents some brain MRI slices from different records before and after histogram equalization, demonstrating the necessity and the effect of this processing step.

Literature contains several attempts to standardize the intensity distributions of MRI records (e.g. [20, 81, 82, 83, 84]), but none of them were designed to specially handle cases with focal lesions, where the relative intensity of some relevant part of the pixels may seriously differ from the normal. Brain tumors may grow to up to 25% of the brain volume, causing strong distortions in pixel intensity distributions.

3.2.1 CLAHE

Whenever a machine learning based segmentation method classifies individual pixels, it is required to provide a uniform scale of pixel intensities in each data channel. On the other hand, convolutional neural networks do not require the strict uniform scale, all they need is to improve the visibility of different structures within their context. The Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm is a widely used preprocessing technique for enhancing contrast in medical imaging and works by dividing the input image into small, overlapping regions and then applying histogram

Table 3.2: Definition of statistical accuracy indicators

Accuracy indicator benchmark	Set based definition	Confusion matrix based definition
True positive rate Recall Sensitivity	$\mathrm{TPR}_i = \frac{\lvert\Gamma_i^{(\pi)} \cap \Lambda_i^{(\pi)}\rvert}{\lvert\Gamma_i^{(\pi)}\rvert}$	$\mathrm{TPR}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i}$
Positive predictive value Precision	$\mathrm{PPV}_i = \frac{\lvert\Gamma_i^{(\pi)} \cap \Lambda_i^{(\pi)}\rvert}{\lvert\Lambda_i^{(\pi)}\rvert}$	$\mathrm{PPV}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i}$
Dice similarity score F1 − score	$\mathrm{DSC}_i = \frac{2 \cdot \lvert\Gamma_i^{(\pi)} \cap \Lambda_i^{(\pi)}\rvert}{\lvert\Gamma_i^{(\pi)}\rvert + \lvert\Lambda_i^{(\pi)}\rvert}$	$\mathrm{DSC}_i = \frac{2 \cdot \mathrm{TP}_i}{2 \cdot \mathrm{TP}_i + \mathrm{FP}_i + \mathrm{FN}_i}$
Rate of correct decisions Accuracy	$\mathrm{ACC}_i = \frac{\lvert\Gamma_i^{(\pi)} \cap \Lambda_i^{(\pi)}\rvert + \lvert\Gamma_i^{(\nu)} \cap \Lambda_i^{(\nu)}\rvert}{\lvert\Gamma_i^{(\pi)}\rvert + \lvert\Gamma_i^{(\nu)}\rvert}$	$\mathrm{ACC}_i = \frac{\mathrm{TP}_i + \mathrm{TN}_i}{\mathrm{TP}_i + \mathrm{TN}_i + \mathrm{FP}_i + \mathrm{FN}_i}$
True negative rate Specificity	$\mathrm{TNR}_i = \frac{\lvert\Gamma_i^{(\nu)} \cap \Lambda_i^{(\nu)}\rvert}{\lvert\Gamma_i^{(\nu)}\rvert}$	$\mathrm{TNR}_i = \frac{\mathrm{TN}_i}{\mathrm{TN}_i + \mathrm{FP}_i}$

equalization separately to each region [85]. This adaptive approach ensures that the contrast enhancement is applied locally and does not lead to over-amplification of noise or sharp transitions, which are common issues with global histogram equalization methods.

In the context of the BraTS dataset, the CLAHE algorithm is often used as a preprocessing step to enhance the contrast of the multimodal brain images (e.g. [86],[87]). This helps to improve the visibility of tumor boundaries and other structures, making it easier for subsequent segmentation algorithms to accurately identify and delineate tumor regions.

One of the advantages of using CLAHE for preprocessing BraTS data is that it can be applied to each modality independently, allowing for contrast enhancement across all the different imaging modalities (T1, T2, FLAIR, and T1C). This multimodal enhancement further aids in improving the performance of subsequent segmentation algorithms, as it enhances the visibility of tumor regions across multiple image types.

3.3 Evaluation of segmentation quality

The performance of the brain tumor segmentation algorithms developed within the bounds of this thesis are evaluated using well-known statistical accuracy indicators, as utilized in [3]. These indicators can be defined various ways, but their meaning is the same in every case. Two sets of definitions are presented in this section: one is based on the definitions of the variables stored in the confusion matrix, while the other relies on set based definitions.

The whole set of pixels of MRI record with index i is Ω_i, which is separated by the ground truth into two disjoint sets: $\Gamma_i^{(\pi)}$ and $\Gamma_i^{(\nu)}$, the set of positive and negative pixels, respectively. The final segmentation result provides another separation of Ω_i into two disjoint sets denoted by $\Lambda_i^{(\pi)}$ and $\Lambda_i^{(\nu)}$, which represent the positive and negative pixels, respectively, according to the final decision (labeling) of the proposed procedure. Statistical accuracy indicators reflect in different ways how much the subsets $\Gamma_i^{(\pi)}$ and $\Lambda_i^{(\pi)}$, and their complementary subsets $\Gamma_i^{(\nu)}$ and $\Lambda_i^{(\nu)}$ overlap. The set based definitions of the main accuracy markers obtained for any record with index i ($i = 1 \dots n_\rho$) are presented in the second column of Table 3.2.

Any pixels of the testing MRI record with index i is Ω_i, can be either a case of true positive, a true negative, a false positive or a false negative. Let us denote by TP_i, TN_i, FP_i, and FN_i, the number of true positives, true negatives, false positives and false negatives, respectively, found within the volume number i. These values are obtained within the 2×2-sized confusion matrix. Based on these definitions, we can compute the statistical accuracy indicator values using the formulas given in the third column of Table 3.2.

To characterize the global accuracy, we may compute average values of the above defined indicators over a whole set of MRI records. We denote them by $\overline{\text{DSC}}$, $\overline{\text{TPR}}$, $\overline{\text{TNR}}$, $\overline{\text{PPV}}$, and $\overline{\text{ACC}}$. The average Dice similarity coefficient is given by the formula

$$\overline{\text{DSC}} = \frac{1}{n_\rho} \sum_{i=1}^{n_\rho} \text{DSC}_i \ . \tag{3.1}$$

The average values of other accuracy indicators are computed with analogous formulas.

The overall Dice similarity coefficient ($\widetilde{\text{DSC}}$), first introduced in [72], is a single Dice similarity coefficient extracted from the set of all pixels from Ω, given by the formula:

$$\widetilde{\text{DSC}} = \frac{2 \cdot \left| \bigcup_{i=1}^{n_\rho} \Gamma_i^{(\pi)} \cap \bigcup_{i=1}^{n_\rho} \Lambda_i^{(\pi)} \right|}{\left| \bigcup_{i=1}^{n_\rho} \Gamma_i^{(\pi)} \right| + \left| \bigcup_{i=1}^{n_\rho} \Lambda_i^{(\pi)} \right|} = \frac{2 \cdot \sum_{i=1}^{n_\rho} \text{TP}_i}{2 \cdot \sum_{i=1}^{n_\rho} \text{TP}_i + \sum_{i=1}^{n_\rho} \text{FP}_i + \sum_{i=1}^{n_\rho} \text{FN}_i} \ . \tag{3.2}$$

All accuracy indicators are defined in the $[0, 1]$ interval. Perfect segmentation sets all indicators to the maximum value of 1.

Preprocessing

The primary aim of this stage is to establish consistent histograms for all data channels within the MRI records and to generate supplementary features for each pixel. Moreover, preprocessing is standardizing, enhancing, and cleaning the data, laying a solid foundation for accurate and reliable brain tumor segmentation. These step in the segmentation process (Figure 2.2, 2.3) contribute to the overall effectiveness of the subsequent analysis and clinical interpretation of the images.

4.1 Background

A major drawback of MR imaging consists in the lack of a standard scale of image intensities. This is why we need to map the histogram of each data channel of BraTS volumes onto a uniform scale. Before my doctoral studies, I had previous experience with this issue and with the research group we addressed it as follows. Although literature contains various recommendation for this issue [20, 88], we opted to employ a simple linear transform $x \to \alpha x + \beta$ to all intensities, where parameters α and β were established separately for each volume and each data channel in such a way that the 25-percentile and 75-percentile values became 600 and 800, respectively. Further on, a minimum and a maximum intensity barrier was enforced at 200 and 1200, respectively. This approximately corresponds to a 10-bit resolution in each data channel.

Most voxels of the MRI records have valid nonzero intensity in all data channels. However, there are voxels with one or more missing values. We considered that the region of interest (ROI) in the BraTS volumes includes all voxels that have at least one nonzero value in any of the observed data channels. Missing values were replaced by the mean intensity value of existing neighbours within the 26-element immediate spatial neighbourhood, or the grand mean of the given data channel whenever no neighbours with correct intensity were found in the neighbourhood.

Although the four observed features of each voxel bear a lot more information than any one of them, there is an acute need to extend the feature vectors with further computed features. A total number of 100 computed features were added to the feature vector describing each voxel, according to the inventory given in Table 4.1. For each of the four observed intensities (T1, T2, T1C, FLAIR), six average, five median, one minimum, one maximum, four gradient values, and further eight Gabor features were extracted. All com-

Table 4.1: Inventory of computed features. All four data channels were involved equally. Spatial features are those twelve computed from $3 \times 3 \times 3$ neighborhood.

Neighborhood	$3 \times 3 \times 3$	3×3	5×5	7×7	9×9	11×11	Total
Average	4	4	4	4	4	4	24
Maximum	4						4
Minimum	4						4
Median		4	4	4	4	4	20
Gradient				16			16
Gabor wavelet						32	32
Total	12	8	8	24	8	40	100

puted feature values were linearly scaled into the $[200, 1200]$ interval. This way, together with the four observed features, each voxel is described by a 104-element feature vector. These feature vectors are used by the classification stage of the proposed segmentation procedure.

4.2 Methods

I have examined the issue of preprocessing from multiple perspectives. The investigations carried out are reported in four sections.

The first section delves into the process of histogram equalization, a crucial step in standardizing the intensity distributions of medical images for subsequent analysis.

The second section offers a comprehensive overview of the feature generation and selection methodology. It underscores the critical role of suitable features in automated decision tree-based procedures for distinguishing between normal tissues and complete tumors. The section outlines the steps involved in extracting and selecting features, with the aim of minimizing the number of utilized features while maintaining the accuracy of the system.

The third section extensively explores how spectral resolution (also known as color depth or intensity resolution) affects the accuracy of segmenting brain tissues and tumors in magnetic resonance imaging. The main objective was to define the optimal spectral resolution that would lead to the highest segmentation accuracy achievable by the given machine learning method, while requiring the lowest number of bits to store each pixel attribute.

The fourth section delves into the significance and utilization of spatial atlases in the segmentation of medical images. It underscores how atlases play a crucial role in enhancing the segmentation process by offering valuable prior information about the objects (organs) under consideration. The construction of spatial atlases based on the available data is discussed along with their application in refining feature values prior to the classification stage.

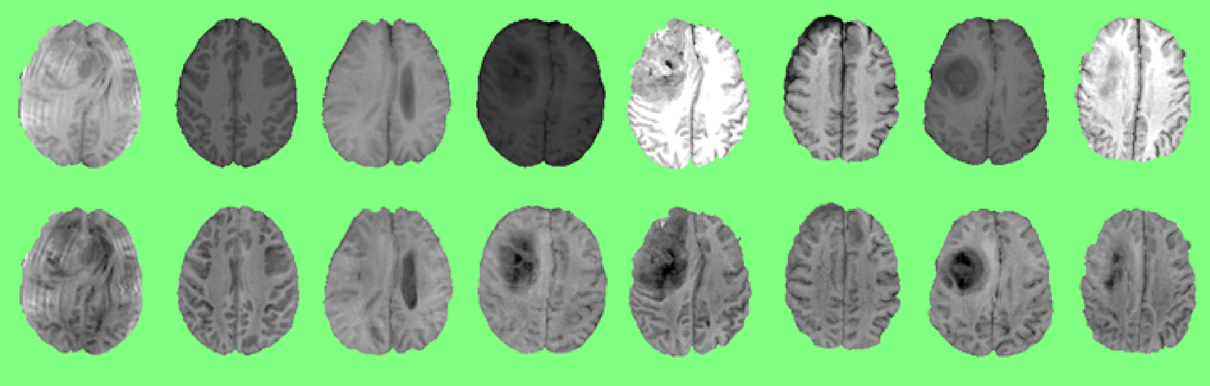

Figure 4.1: Eight T1-weighted slices from different records from the BraTS 2019 LGG data, before (top row) and after histogram equalization (bottom row).

4.2.1 Histogram equalization

MRI is widely utilized in medical diagnosis due to its high contrast and detailed resolution. However, it has a limitation in that the recorded numerical values do not directly represent the observed tissues. To address this issue, histogram equalization is employed to adapt pixel intensities to their context. This step is crucial for accurate image interpretation. Without prior histogram equalization, comparing intensity values from different MRI records would be akin to evaluating jewels solely based on their weight, without considering the precious metal they are made of. Figure 4.1 illustrates brain MRI slices before and after histogram equalization, underscoring the significance and impact of this preprocessing step.

The most popular histogram equalization technique used by current brain tumor segmentation solutions was proposed by Nyúl *et al.* [20]. This algorithm (referred to as Algorithm A2) works in batch mode: it adjusts the intensity distributions of all available records according to the averaged position of some milestones defined as certain percentiles of each input distribution. The more milestones are used, the stronger constraints are applied to the output distributions. However, applying very strong constraints to intensity distributions and expecting them to be similar no matter whether they contain focal lesions or not, may undermine the segmentation quality.

The most part of current brain tumor segmentation works (e.g. [4, 7, 89, 90, 91, 92, 93, 94, 95, 96]) only mention that they use the histogram equalization technique of Nyúl *et al.* [20], without giving details of the number of milestones or other parameters. There are few exceptions, where the number of landmark points is revealed: Soltaninejad *et al.* [31] proposed using 12 landmarks, while Pinto *et al.* [30] seem to be using the M12 setting of Algorithm A2, see details in Table 4.2. Alternately, Tustison *et al.* [29] noted that in their study, a simple linear transformation based histogram equalization method provided slightly better accuracy than Algorithm A2, but they did not make their linear transform method public. Other works that employed such linear techniques (e.g. [97, 72, GyA6]) did not compare their histogram equalization to any other method.

Table 4.2: Various landmark schemes for the Algorithm A2

Scheme	Landmark points
M01	$p_{\mathrm{Lo}}, p_{50}, p_{\mathrm{Hi}}$
M02	$p_{\mathrm{Lo}}, p_{25}, p_{75}, p_{\mathrm{Hi}}$
M03	$p_{\mathrm{Lo}}, p_{25}, p_{50}, p_{75}, p_{\mathrm{Hi}}$
M04	$p_{\mathrm{Lo}}, p_{10}, p_{50}, p_{90}, p_{\mathrm{Hi}}$
M05	$p_{\mathrm{Lo}}, p_{20}, p_{40}, p_{60}, p_{80}, p_{\mathrm{Hi}}$
M06	$p_{\mathrm{Lo}}, p_{10}, p_{25}, p_{75}, p_{90}, p_{\mathrm{Hi}}$
M07	$p_{\mathrm{Lo}}, p_{20}, p_{35}, p_{50}, p_{65}, p_{80}, p_{\mathrm{Hi}}$
M08	$p_{\mathrm{Lo}}, p_{10}, p_{25}, p_{50}, p_{75}, p_{90}, p_{\mathrm{Hi}}$
M09	$p_{\mathrm{Lo}}, p_{10}, p_{25}, p_{40}, p_{60}, p_{75}, p_{90}, p_{\mathrm{Hi}}$
M10	$p_{\mathrm{Lo}}, p_{10}, p_{25}, p_{40}, p_{50}, p_{60}, p_{75}, p_{90}, p_{\mathrm{Hi}}$
M11	$p_{\mathrm{Lo}}, p_{10}, p_{20}, p_{30}, p_{40}, p_{60}, p_{70}, p_{80}, p_{90}, p_{\mathrm{Hi}}$
M12	$p_{\mathrm{Lo}}, p_{10}, p_{20}, p_{30}, p_{40}, p_{50}, p_{60}, p_{70}, p_{80}, p_{90}, p_{\mathrm{Hi}}$

The aim is to explore which of the aforementioned histogram equalization techniques are most effective in aiding machine learning methods to achieve optimal segmentation quality and to determine the most convenient settings for this process.

Two approaches are compared in this study, both having several applications in current MRI data segmentation methods. The goal is to establish, which approach and what settings are needed to achieve best segmentation accuracy. Both approaches are formulated in such a way that the target set of normalized intensities is the continuous interval $[0, 1]$, which can later be resampled to any desired discrete spectral resolution. In the numerical evaluations use 8-bit coding, using values of 1 to 255 for valid brain pixel intensities and 0 for outer (non-brain) regions of the volume.

Method A1: linear transform with one parameter. The linear transform approach works on each volume separately, and treats data channels independently of each other. It transforms any intensity value y to $\min\{\max\{ay + b, 0\}, 1\}$, where coefficients a and b are identified in such a way, that the 25-percentile intensity value p_{25} is transformed to $\lambda_{25} \in (1/4, 1/2)$, while the 75-percentile p_{75} to $1 - \lambda_{25}$. The coefficients of the linear transform are:

$$a = \frac{1 - 2\lambda_{25}}{p_{75} - p_{25}} \qquad \text{and} \qquad b = \frac{\lambda_{25}p_{75} - (1 - \lambda_{25})p_{25}}{p_{75} - p_{25}} \ . \tag{4.1}$$

All input intensity values below $[p_{25}(1 - \lambda_{25}) - p_{75}\lambda_{25}]/(1 - 2\lambda_{25})$ are transformed to 0, while input intensities above $[p_{75}(1 - \lambda_{25}) - p_{25}\lambda_{25}]/(1 - 2\lambda_{25})$ are transformed to 1. So both tails of the input histogram are subject to cutting, but the thresholds are defined dynamically, they depend on the input parameter λ_{25} and the data through percentiles p_{25} and p_{75}. In previous works of our research team, and early works of my PhD studies [97, 72, GyA6], this approach was used with parameter setting $\lambda_{25} = 0.4$.

Method A2 by Nyúl *et al.* [20] The histogram equalization introduced by Nyúl *et al.* [20] treats data channels independently of each other, but uses the chosen data channel of all available volumes to establish the equalized histogram for each volume. It cuts a fixed amount of both tails of input histograms, defined by the percentiles p_{Lo} and $p_{\mathrm{Hi}} = p_{100-\mathrm{Lo}}$, where p_{Lo} is a parameter. Method A2 registers the histograms of different volumes together based on predefined milestones defined as percentiles. Table 4.2 presents 12 milestone schemes involved in this study. The steps of the algorithm are presented as follows:

1. For any record with index h ($h = 1 \dots H$), we establish a bounded linear mapping of original intensities $y^{(h)} \to \min\{\max\{a^{(h)}y^{(h)} + b^{(h)}, 0\}, 1\}$ in such a way that the percentile p_{Lo} is mapped to 0 and p_{Hi} is mapped to 1. The identified coefficients $a^{(h)}$ and $b^{(h)}$ are used to identify the mapped positions of each milestone $p_m^{(h)}$ ($m = 1 \dots M$, M stands for the number of milestones) from the chosen milestone scheme (see Table 4.2): $p_m^{(h)} \to \hat{y}_m^{(h)} = a^{(h)}p_m^{(h)} + b^{(h)}$.

2. The final transformation for any record h maps the milestone $p_m^{(h)}$ ($m = 1 \dots M$) to the averaged value $\overline{y}_m = \frac{1}{H}\sum_{i=1}^{H}\hat{y}_m^{(i)}$, and apply piecewise linear interpolation for any intensity value situated between consecutive milestones. So the final transformation of intensity y of record h is given by the formula:

$$y \to \begin{cases} 0 & \text{if } y < p_{\mathrm{Lo}}^{(h)} \\ 1 & \text{if } y > p_{\mathrm{Hi}}^{(h)} \\ \overline{y}_m + \frac{(\overline{y}_{m+1} - \overline{y}_m)(y - p_m^{(h)})}{p_{m+1}^{(h)} - p_m^{(h)}} & \text{otherwise} \end{cases} , \tag{4.2}$$

   where m is established in such a way that $p_m^{(h)} \le y < p_{m+1}^{(h)}$.

Findings and achievements are reported in Section 4.3.1.

4.2.2 Feature generation and selection

Accuracy is the most important quality marker in medical image segmentation. However, when the task is to handle large volumes of data, the relevance of processing speed rises. In machine learning solutions the optimization of the feature set can significantly reduce the computational load.

Feature generation is a key component of the procedure, because the four observed features, namely the T1, T2, T1C, and FLAIR intensities provided by the MRI device, do not contain all the possible discriminative information that can be employed to distinguish tumor pixels from normal ones. A major motivation for feature generation represents the fact that the automated registration algorithm used by the BraTS experts to align the four data channels never performs a perfect job, so for an arbitrary cubic millimeter of brain tissues, represented by the pixel situated at coordinates (x, y, z) in volume T1, the

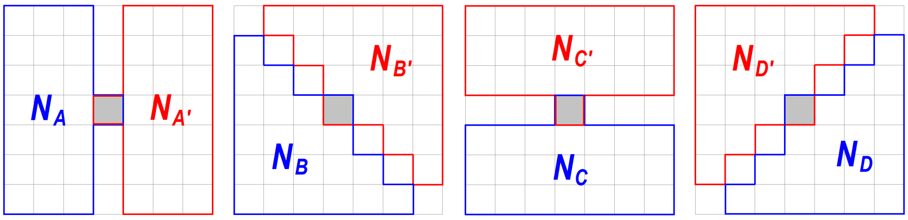

Figure 4.2: Neighborhoods used for the extraction of gradient features, with respect to the current pixel indicated by grey color.

corresponding information in the other volumes is not in the same place but somewhere in the close neighbourhood of (x, y, z). A further motivation consists in the usual property of image data that neighbour pixels correlate with each other.

All four MRI data channels equally contribute to the feature generation process. Observed features are treated independently of each other: 26 computed features are extracted from each of them. The inventory of extracted features is presented in Table 4.1. Minimum, maximum, and average values are extracted in such a way that only the brain pixels are taken into consideration from the neighbourhoods indicated in the table. The computation of the gradient features involves the masks presented in Figure 4.2. The current pixel is part of all masks to avoid division by zero in Eq. (4.3). The 16 gradient features of an arbitrary pixel p are computed with the formula

$$g_m^{(c)}(p) = \gamma_{0.5} + k_g \left(\frac{\sum\limits_{q \in N_m(p) \cap \overline{\Omega}} I_q^{(c)}}{|N_m(p) \cap \overline{\Omega}|} - \frac{\sum\limits_{q \in N_{m'}(p) \cap \overline{\Omega}} I_q^{(c)}}{|N_{m'}(p) \cap \overline{\Omega}|} \right) , \tag{4.3}$$

followed by $g_m^{(c)}(p) \leftarrow \min\{\max\{\alpha, g_m^{(c)}(p)\}, \beta\}$ to stay within the target interval $[\alpha, \beta]$, where $c \in \{\mathrm{T1}, \mathrm{T2}, \mathrm{T1C}, \mathrm{FLAIR}\}$ is the one of the four data channels, $m \in \{A, B, C, D\}$ is the index of the current gradient mask, and thus $I_q^{(c)}$ stands for the intensity of pixel q in data channel c. $\overline{\Omega}$ stands for the set of all brain pixels in the volume, $\gamma_{0.5}$ is the middle of the target interval of intensities, k_g is a globally constant scaling factor, and $|Q|$ represents the cardinality of any set denoted by Q.

The full set of generated features consists of 104 items. The procedure was first evaluated using the whole feature set, and some subsets of features. Table 4.3 (except last row) exhibits full details on the segmentation accuracy achieved, using the indicators defined in Section 3.3. The upper section of the table shows the accuracy achieved using combinations of the feature families presented in Table 4.1, while the lower section shows the outcome of segmentation when one ore more data channels are totally excluded from the process. These tests reveal that the GABOR and GRAD feature families possess lower discriminative power than AVG-MED and spatial ones (SPACE), regarding normal and tumor brain tissues. Further on, dropping any or both of the T1 and T1C data channels has little effect on the accuracy of the whole tumor segmentation problem. Despite these

Table 4.3: Performance with various subsets of the initial set of features

Features		Average DSC ($\overline{\mathrm{DS}}$)		Overall	Average		
Description	Count	Before PP	After PP	DSC ($\widetilde{\mathrm{DS}}$)	TPR	TNR	ACC
Observed ones only	4	50.54%	62.85%	69.16%	55.42%	99.63%	96.52%
GABOR	36	56.72%	67.73%	73.12%	60.50%	99.62%	96.87%
GABOR & GRAD	52	68.18%	76.65%	79.35%	71.78%	99.31%	97.37%
GRAD	20	69.62%	77.94%	80.48%	74.40%	99.21%	97.46%
GABOR & SPACE	48	70.10%	79.25%	81.65%	77.10%	99.11%	97.56%
AVG-MED	44	75.06%	79.40%	81.66%	81.13%	98.67%	97.44%
GABOR & AVG-MED	76	74.91%	79.58%	81.69%	80.44%	98.75%	97.46%
GABOR & GRAD & AVG-MED	92	74.96%	79.69%	81.78%	80.45%	98.77%	97.48%
GRAD & AVG-MED	60	75.25%	79.73%	81.95%	81.14%	98.72%	97.49%
SPACE	16	70.31%	79.64%	81.97%	77.38%	99.14%	97.61%
GABOR & GRAD & SPACE	64	71.05%	79.57%	81.97%	77.87%	99.08%	97.59%
GABOR & SPACE & AVG-MED	88	75.62%	80.37%	82.44%	81.41%	98.78%	97.56%
GRAD & SPACE	32	71.70%	80.09%	82.47%	78.77%	99.07%	97.64%
GRAD & SPACE & AVG-MED	72	76.12%	80.65%	82.78%	82.21%	98.76%	97.60%
SPACE & AVG-MED	56	76.12%	80.64%	82.80%	82.37%	98.75%	97.59%
Full set of features	104	75.70%	80.39%	82.48%	81.52%	98.78%	97.57%
T1 channel excluded	78	75.60%	80.30%	82.59%	81.92%	98.76%	97.57%
T1C channel excluded	78	75.55%	80.24%	82.44%	81.71%	98.75%	97.55%
T1 & T1C channels excluded	52	74.71%	79.40%	82.27%	81.84%	98.71%	97.52%
T2 channel excluded	78	67.60%	71.59%	74.99%	72.11%	98.47%	96.62%
FLAIR channel excluded	78	55.86%	64.55%	70.79%	62.39%	98.95%	96.38%
Final reduced feature set (Table 4.4)	13	75.19%	80.21%	82.52%	81.74%	98.76%	97.57%

remarks, the feature selection begins with the full set of 104 features.

The main goal of feature selection is to reduce the size of the employed feature set without significantly damaging the final average values of the quality indicators.

Features are selected using an iterative process, using a criteria based on the following two ratios:

- How frequently is a feature employed (actually used) while giving a label to a voxel? Or in other words, the decision of what part of voxels actually depends on the given feature? This ratio is referred to as the rate of usage (RU).
- Out of the cases when the given feature plays an actual role in the decision regarding the voxels, what is the rate of correct decisions? This ratio is referred to as the rate of accuracy (RA).

The proposed feature selection algorithm is summarized in the following:

1. Consider the initial set of features S the full one.
2. Perform a full training and testing using all MRI records, using S as the set of features.
3. Extract the average and overall Dice Scores ($\overline{\mathrm{DS}}$ and $\widetilde{\mathrm{DS}}$) from the current segmentation outcome.
4. Extract the rate of usage (RU_j) and the rate of accuracy (RA_j) for each feature f_j from S, $j = 1 \ldots |S|$.
5. Sort the features f_j, $j = 1 \ldots |S|$, by the composite criterion $\mathrm{RU}_j + \mathrm{RA}_j$.
6. If the latest extracted average and overall Dice Scores are acceptable, eliminate the worst ranked features (with lowest $\mathrm{RU}_j + \mathrm{RA}_j$) from S, and go back to Step 2. The number of eliminated features is $\alpha \times |S|$ rounded to the nearest integer, with $\alpha \in [0.1, 0.25]$.
7. The last set of features that provided solution with no damaged accuracy is the final set of features.

Results are reported in Section 4.3.2.

4.2.3 Spectral resolution

The number of medical imaging devices installed in hospitals and the amount of medical image data created day by day is continuously rising. The number of human experts available to process this data cannot follow this trend. This is why there is a strong need for fully automatic methods that can reliably process the image data. Further on, there is a need for efficient data storage as well.

Various techniques frequently deployed for medical image processing, like AdaBoost [28], random forest [29, 97, 98] or extremely random trees [30], use several dozens or even few hundreds of features extracted for each pixel separately, requiring an even larger data storage capacity. To store this large number of feature values, it would be useful to know, how many bits are necessary for each variable to maintain an acceptable accuracy of the segmentation. Using finer resolution than necessary leads to slower data processing and larger storage requirements. On the other hand, if the spectral resolution is too coarse, it may strongly oppose the accurate segmentation, significantly reducing the quality of the obtained partition.

The following study investigates how the spectral resolution (also referred to as color depth or intensity resolution) influences the segmentation accuracy of brain tissues and tumors from multi-spectral magnetic resonance imaging data. The main goal is to determine the optimal spectral resolution that maximizes segmentation accuracy using the given machine learning method, while minimizing the number of bits needed to store each attribute of a pixel. Additionally, we aim to examine the impact of spectral resolution on the segmentation accuracy of MRI data.

After having intensities normalized and transformed to the $[0, 1]$ interval, as indicated in the histogram equalization process, it is an easy step to generate data with the desired spectral resolution. Let us suppose, that we wish to use β bits to represent each intensity value. This means that we can use integer intensity values between 1 and $2^\beta - 1$. The value 0 is reserved for non-brain (external) pixels. So in case of β-bit spectral resolution, we can have $2^\beta - 1$ different intensity values. So each intensity value is transformed as $y \in [0, 1]$ becomes $1 + \langle y \times (2^\beta - 2) \rangle$, where $\langle \cdot \rangle$ is the sign of rounding to the nearest integer. Figure 4.3 presents some brain MRI slices represented at various spectral resolution.

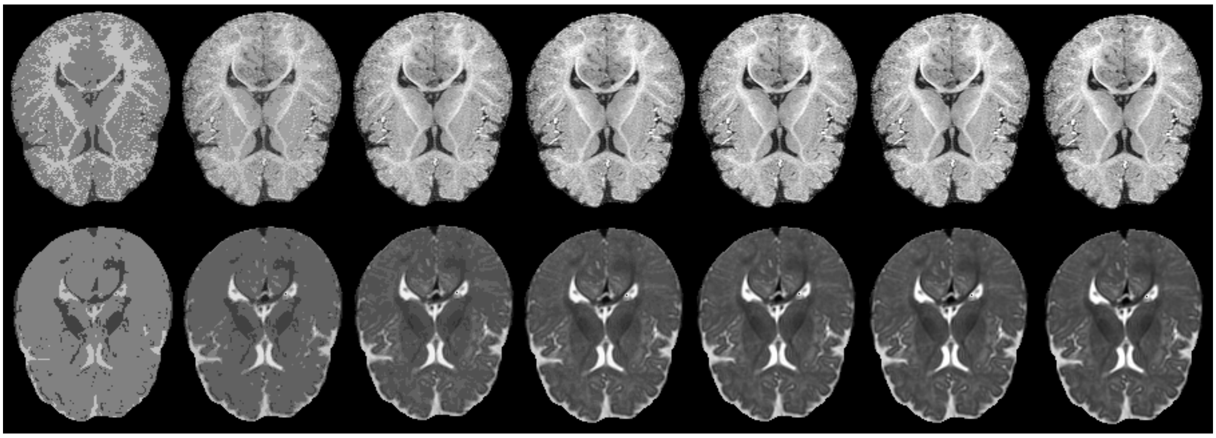

Figure 4.3: T1-weighted (top row) and T2-weighted (bottom row) MRI images transformed to spectral resolution of 2, 3, 4, ..., and 8 bits.

Two datasets are involved in this study, so that we can establish the segmentation quality in case of MRI data with and without focal lesions. Results are reported in Section 4.3.3.

4.2.4 Atlas based data enhancement

Atlases and multi-atlases employed in medical image segmentation problems attempt to enhance the quality of the outcome using prior information regarding the object (organ) being segmented. Without atlases and shape models, segmentation methods can only use global and local properties of pixels (or voxels), like intensity distributions and textures. The use of atlases enables us to add further information to the segmentation process, for example, what is usually present in the same place in other similar image records, or what intensities are usually present in the same place in other normal records.

Atlases have recently been involved in several medical imaging problems, including the segmentation of brain tissues and lesions [5, 99, 100], prostate [101], lung [102], cardiac structures (e.g. myocardium) [103, 104], pancreas [105, 106], bones [107], cartilage [108], and multiple abdominal organ [109]. Atlases are used in segmentation problems based on image data originating from virtually all imaging modalities, including magnetic resonance images (MRI) [5, 99, 101], computed tomography (CT) [102, 106, 107], CT angiography [103], positron emission tomography (PET) [104], X-ray [110] and mammography [111]. A systematic review of earlier image segmentation solutions based on atlases and multi-atlases is given by Cabezas *et al.* [6]. Several earlier atlas-based solutions are summarized in the review paper of Gordillo *et al.* [2]. A more recent summary of such methods can be found in the work of Sun *et al.* [112].

In image segmentation, atlases are usually used as approximate maps of the objects that should be present in the image in normal case. Atlases are usually established based on prior measurements. The image data before the current segmentation is registered to the atlas, and the current segmentation outcome is fused with the atlas to obtain a final, atlas assisted segmentation result.

In image segmentation problems, atlases and multi-atlases are generally employed to provide shape or texture priors and to guide the segmentation toward a regularized solution via label fusion [113, 114, 115]. My solution uses multi-atlases in a different manner: before proceeding to ensemble learning and testing, atlases are trained to characterize the local appearance of normal tissues, and applied to transform all feature values to emphasize their deviation from normal.

The main contributions consist in:

1. the way the multi-atlases are involved in the preprocessing, to prepare the feature data for segmentation via ensemble learning;
2. the ensemble model built from binary decision trees with unconstrained depth;
3. the two-stage post-processing scheme that discards a large number of false positives.

A spatial atlas for each feature is constructed using the train records, which contains the local average and standard deviation of the intensity values taken from normal pixels only. Of course, the train records need to be registered to each other so that we can build consistent atlases. These atlases are then used to refine the feature values in both train and test records, before proceeding to the ensemble learning based classification.

First, it is important to define the spatial resolution of the atlases by introducing a single parameter S, which represents the size of the atlas. Each atlas is a discrete spatial array defined on $\mathcal{S}^3$, where $\mathcal{S} = \{-S, -S+1, \dots 0 \dots, S-1, S\}$. Within the atlas array, the neighbourhood of atlas point $\hat{\pi}$ having the coordinates $(\hat{x}, \hat{y}, \hat{z})$, is defined as:

$$\mathcal{C}_\delta(\hat{\pi}) = \{(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) \in \mathcal{S}^3, |\hat{\alpha} - \hat{x}| \le \delta \quad \wedge \quad |\hat{\beta} - \hat{y}| \le \delta \quad \wedge \quad |\hat{\gamma} - \hat{z}| \le \delta\} \ , \tag{4.4}$$

where δ is a small positive integer, typically one ore two, which determines the size of the neighbourhood.

The mathematical description of the atlas building process requires the introduction of several notations. Let $\mathcal{M}$ stand for the set of all MRI records of the chosen data set, while $\mathcal{M}^{(T)}$ and $\mathcal{M}^{(E)}$ represent the current set of train and evaluation (or test) records, respectively. Obviously, $\mathcal{M} = \mathcal{M}^{(T)} \cup \mathcal{M}^{(E)}$, and $\mathcal{M}^{(T)} \cap \mathcal{M}^{(E)} = \Phi$. Let M_i be the record with index i, belonging to either $\mathcal{M}^{(T)}$ or $\mathcal{M}^{(E)}$ in a certain scenario. The set of pixels belonging to record M_i are denoted by Ω_i, for any $i = 1 \dots n_\rho$. The set of all pixels from all MRI records is $\Omega = \bigcup_{i=1}^{n_\rho} \Omega_i$. Any pixel $\pi \in \Omega$ has a feature vector with n_φ elements. For any pixel π, the value of feature with index $\varphi \in \{1 \dots n_\varphi\}$ is denoted by $I_\pi^{(\varphi)}$. Further on, let $\Gamma^{(\nu)}$ and $\Gamma^{(\tau)}$ be the set of negative and positive pixels, respectively, as indicated by the ground truth. Obviously, $\Omega = \Gamma^{(\nu)} \cup \Gamma^{(\tau)}$, and $\Gamma^{(\nu)} \cap \Gamma^{(\tau)} = \Phi$.

A rigid registration is defined in the following, so that we can map all volumes onto the atlas. For each record M_i $(i = 1 \dots n_\rho)$ a function $f_i : \Omega_i \to \mathcal{S}^3$ is needed that maps the pixels onto the atlas, to find the corresponding atlas position for all brain pixels. These functions map the gravity center of each brain to the atlas origin, and the standard deviations of x, y and z coordinates are all transformed to S/ξ, where $\xi = 2.5$ is a predefined constant. From the pixel coordinates we compute averages and standard deviations as follows:

$$\begin{cases} \mu_x^{(i)} = \frac{1}{|\Omega_i|} \sum\limits_{\pi \in \Omega_i} x_\pi \\ \mu_y^{(i)} = \frac{1}{|\Omega_i|} \sum\limits_{\pi \in \Omega_i} y_\pi \\ \mu_z^{(i)} = \frac{1}{|\Omega_i|} \sum\limits_{\pi \in \Omega_i} z_\pi \end{cases} , \tag{4.5}$$

and then

$$\begin{cases} \sigma_x^{(i)} = \sqrt{\frac{1}{|\Omega_i|-1} \sum\limits_{\pi \in \Omega_i} \left(x_\pi - \mu_x^{(i)}\right)^2} \\ \sigma_y^{(i)} = \sqrt{\frac{1}{|\Omega_i|-1} \sum\limits_{\pi \in \Omega_i} \left(y_\pi - \mu_y^{(i)}\right)^2} \\ \sigma_z^{(i)} = \sqrt{\frac{1}{|\Omega_i|-1} \sum\limits_{\pi \in \Omega_i} \left(z_\pi - \mu_z^{(i)}\right)^2} \end{cases} , \tag{4.6}$$

where x_π, y_π, and z_π are the coordinates of pixel π in the brain volume. The formula of

```
Data: Set of MRI records involved in the study $\mathcal{M}$, its set of pixels $\Omega$, each pixel
      with 4 observed and $n_\varphi - 4$ computed features.
Data: Parameters $S$ and $\delta$
Result: Atlas functions $A_\varphi$, $\varphi = 1 \dots n_\varphi$
Result: Updated intensities in all records of $\mathcal{M}$
Define set of train and test records, $\mathcal{M}^{(T)}$ and $\mathcal{M}^{(E)}$, respectively.
for each $M_i \in \mathcal{M}$ do
    Compute the mapping function $f_i(\pi)$ for every pixel $\pi \in \Omega_i$ using Eqs.
      (4.5)-(4.7).
end
for $\varphi = 1 \dots n_\varphi$ do
    Compute the atlas function $A_\varphi(\hat{\pi})$ for every discrete point $\hat{\pi} \in \mathcal{S}^3$ having
    $\hat{\nu}_{\hat{\pi}}^{(\varphi)} > 1$, using Eqs. (4.8)-(4.10).
    for each $M_i \in \mathcal{M}$ do
        for each $\pi \in \Omega_i$ do
            Update the value of feature $\varphi$ of pixel $\pi$ using the formula given in Eq.
              (4.11).
        end
    end
end
```

Algorithm 1: Build the atlas function A_φ for each feature $\varphi = 1 \dots n_\varphi$, and apply it to all feature data

the mapping f_i is:

$$f_i(\pi) \quad = \quad \left(\left\langle \frac{S\left(x_\pi - \mu_x^{(i)}\right)}{\xi \sigma_x^{(i)}} \right\rangle , \left\langle \frac{S\left(y_\pi - \mu_y^{(i)}\right)}{\xi \sigma_y^{(i)}} \right\rangle , \left\langle \frac{S\left(z_\pi - \mu_z^{(i)}\right)}{\xi \sigma_z^{(i)}} \right\rangle \right) \quad , \tag{4.7}$$

where $\langle\cdot\rangle$ stands for the operation of rounding a floating point variable to the closest integer.

For any feature with index $\varphi \in \{1 \dots n_\varphi\}$, the atlas function has the form $A_\varphi : \mathcal{S}^3 \rightarrow \mathbb{R}^3$, which for any atlas point $\hat{\pi} \in \mathcal{S}^3$ is defined as

$$A_\varphi(\hat{\pi}) = \left(\hat{\mu}_{\hat{\pi}}^{(\varphi)}, \hat{\sigma}_{\hat{\pi}}^{(\varphi)}, \hat{\nu}_{\hat{\pi}}^{(\varphi)} \right) \quad , \tag{4.8}$$

where the components $\hat{\mu}_{\hat{\pi}}^{(\varphi)}$, $\hat{\sigma}_{\hat{\pi}}^{(\varphi)}$, and $\hat{\nu}_{\hat{\pi}}^{(\varphi)}$ are established with the following formulas:

$$\begin{cases} \hat{\nu}_{\hat{\pi}}^{(\varphi)} = \sum\limits_{M_i \in \mathcal{M}^{(T)}} \left| \Psi(i, \hat{\pi}) \right| \\ \hat{\mu}_{\hat{\pi}}^{(\varphi)} = \left(\hat{\nu}_{\hat{\pi}}^{(\varphi)} \right)^{-1} \sum\limits_{M_i \in \mathcal{M}^{(T)}} \left(\sum\limits_{\pi \in \Psi_i(\hat{\pi})} I_\pi^{(\varphi)} \right) \\ \hat{\sigma}_{\hat{\pi}}^{(\varphi)} = \sqrt{ \left(\hat{\nu}_{\hat{\pi}}^{(\varphi)} - 1 \right)^{-1} \sum\limits_{M_i \in \mathcal{M}^{(T)}} \left(\sum\limits_{\pi \in \Psi_i(\hat{\pi})} \left(I_\pi^{(\varphi)} - \hat{\mu}_{\hat{\pi}}^{(\varphi)} \right)^2 \right) } \end{cases} \tag{4.9}$$

where

$$\Psi_i(\hat{\pi}) = \left\{ \pi \in \Omega_i \cap \Gamma^{(\nu)}, f_i(\pi) \in \mathcal{C}_\delta(\hat{\pi}) \right\} \quad . \tag{4.10}$$

The feature values of each pixel $\pi \in \Omega_i$ $(i = 1 \dots n_\rho)$, no matter whether π belongs to a record of the train or test data, are updated with the following formula:

$$\widetilde{I}_\pi^{(\varphi)} \leftarrow \min \left\{ \max \left\{ \alpha, \left\langle \overline{\mu} + \overline{\sigma} \frac{I_\pi^{(\varphi)} - \hat{\mu}_{f_i(\pi)}^{(\varphi)}}{\hat{\sigma}_{f_i(\pi)}^{(\varphi)}} \right\rangle \right\}, \beta \right\} \quad , \tag{4.11}$$

where parameters $\overline{\mu}$ and $\overline{\sigma}$ represent the target average and standard deviation, respectively, and their recommended values are:

$$\begin{cases} \overline{\mu} = (\alpha + \beta)/2 \\ \overline{\sigma} = (\beta - \alpha)/10 \end{cases} , \tag{4.12}$$

where α and β represents the lower and higher limit of the target interval of feature values.

Algorithm 1 summarizes the construction and usage of the atlas. The updated intensity values compose the preprocessed data denoted by P_2 in Figure 2.2. Both P_1 and P_2 follow the very same classification and post-processing steps.

4.3 Results and discussion

The methods presented above underwent a thorough evaluation process that involved the two MRI data sets presented in Section 3.1 evaluated with the statistical indicators presented in Section 3.3.

4.3.1 Histogram equalization

Two brain MRI segmentation problems are considered, one without and the other with focal lesions. Both datasets consist of MRI records, so it is not surprising that they are fed to very similar procedures. However, due to the nature of the imaged organs, there are some differences as well. The processing steps are briefly presented in the following:

Three classification algorithms were included in this study, approaches that functionally strongly differ from each other: random forest (RF), K-nearest neighbours (KNN), and ensembles of support vector machines (SVM). In case of both datasets and all classification approaches, the available data was divided to training and testing data in proportion of 90% vs. 10%. This is explained by the fact that only 10 infant brain records were available that determined us to deploy the “leave-one-out” scheme. We kept the same ratio for the BraTS data as well: the 50 records were randomly divided into ten groups of five, and each group took its turn to serve as testing data while using the other nine groups for training. The classification algorithms were deployed with various settings listed below:

1. RF$_1$: RF using 2000 randomly chosen pixels from each training volume of infant brains, or 1000 pixels from training volumes of BraTS, allowing maximum tree depth

of 18 in both cases.

2. RF_2: RF using 20,000 pixels from each training volume of infant brains, or 10,000 pixels from training volumes of BraTS, maximum tree depth of 22.
3. RF_3: RF using 100,000 pixels from each training volume of infant brains, or 50,000 pixels from training volumes of BraTS, maximum tree depth of 26.
4. KNN_1: KNN using 500 randomly chosen pixels from each training volume of infant brains, or 200 pixels from training volumes of BraTS.
5. KNN_2: KNN using 2000 pixels from each training volume of infant brains, or 1000 pixels from training volumes of BraTS.
6. KNN_3: KNN using 10,000 pixels from each training volume of infant brains, or 5000 pixels from training volumes of BraTS.
7. SVM_1: Ensemble of 15 SVM units, each trained with 30 randomly chosen pixels from each training volume of infant brains or BraTS.

All RF approaches were set to maximum tree count of 45, while all KNN approaches made decisions based on the votes of $k = 11$ neighbours.

Postprocessing is only applied to BraTS records, the details of which are described in Section 6.1. The three classifier algorithms with the 7 settings are listed in Section 4.2.1.

Figure 4.4 exhibits the global accuracy obtained by the random forest classifier, when using Algorithm A2 for histogram equalization. The plot on the left side shows that at low amount of training data (setting RF_1), the milestone schemes M01, M03, and M07 performed the best, while parameter setting $p_{\mathrm{Lo}} = 0.5\%$ is the optimal choice in most cases. Using a larger amount of training data (setting RF_2) leads to higher global accuracy values, while the best performing settings remains the same. The plot on the right side shows the segmentation accuracy achieved by the setting RF_3 that uses a very large training data set. Here the red columns indicate the accuracy obtained by the use of Algorithm A1 for histogram equalization and the blue ones for Algorithm A2 with $p_{\mathrm{Lo}} = 0.5\%$. Apparently M01 is the best performing milestone scheme, and it gives 0.5% higher global accuracy than Algorithm A1 at any value of λ_{25}. This difference can be seen at lower training data sets as well, if we compare the global accuracy values with the ones presented in Figure 4.7.

Analogously to Figure 4.4, Figure 4.5 presents the segmentation accuracy achieved by the KNN classifier, using three different training data sizes. Here the milestone schemes M01, M02, M03, M05, and M07 seem to perform better than others, and the global accuracy values seem to be 0.2-0.5% higher than in case of linear transform based histogram equalization. Similarly to the RF classifier, KNN seems to perform best when using Algorithm A2 for histogram equalization, M01 or M03 milestone scheme, and $p_{\mathrm{Lo}} = 0.5\%$.

Figure 4.4: Global accuracy achieved by the random forest classifiers RF_1 (top), RF_2 (middle), and RF_3 (bottom) in the segmentation of infant brain tissues, after histogram equalization via Algorithm A2 using various milestone settings and p_{Lo} parameter values, or Algorithm A1 with various values of λ_{25}. RF_3 combined with Algorithm A2 used $p_{\mathrm{Lo}} = 0.5\%$.

Figure 4.5: Global accuracy achieved by the KNN classifiers KNN_1 (top), KNN_2 (middle), and KNN_3 (bottom) in the segmentation of infant brain tissues, after histogram equalization via Algorithm A2 using various milestone settings and p_{Lo} parameter values, or Algorithm A1 with various values of λ_{25}. KNN_3 combined with Algorithm A2 used $p_{\mathrm{Lo}} = 0.5\%$.

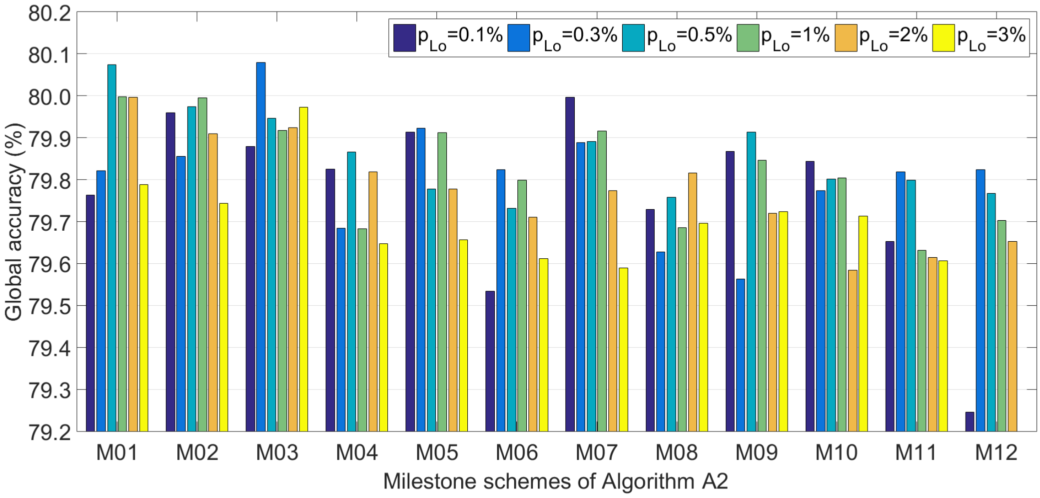

Figure 4.6: Global accuracy achieved by the SVM_1 classifier in the segmentation of infant brain tissues, after histogram equalization via Algorithm A2 using various milestone settings and p_{Lo} parameter values.

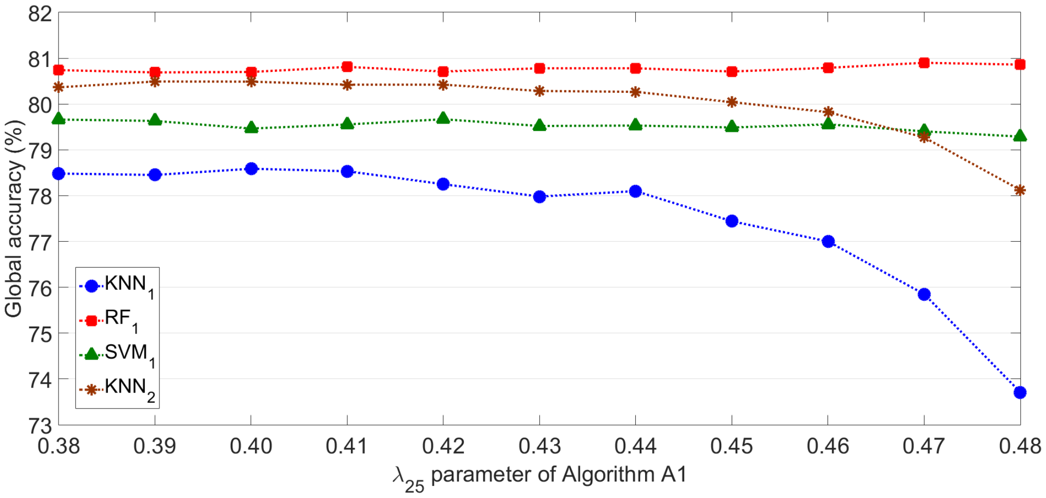

Figure 4.7: Global accuracy achieved by the SVM_1, KNN_1, KNN_2 and RF_1 classifiers in the segmentation of infant brain tissues, after histogram equalization via Algorithm A1 using various values of parameter λ_{25}.

Figure 4.6 presents the segmentation accuracy results obtained by the ensemble of SVM classifiers using setting SVM_1. The best performing milestone schemes of Algorithm A2 are the same as in case of KNN or RF classifiers. As the best result obtained by SVM in combination with linear transform based histogram equalization is 79.6% (see Figure 4.7), Algorithm A2 with its best setting has a lead of 0.5% in accuracy. Figure 4.7 presents the segmentation accuracy achieved by various classifiers combined with Algorithm A1, plotted against the value of parameter λ_{25}. Whenever $\lambda_{25} < 0.42$, the value of λ_{25} does not influence the accuracy. As λ_{25} approaches its theoretical maximum of 0.5, the accuracy achieved by KNN classifier is damaged. This is because at such high values of λ_{25}, hardly anything is cut in the tails of the histograms, and KNN is sensitive to extreme feature values.

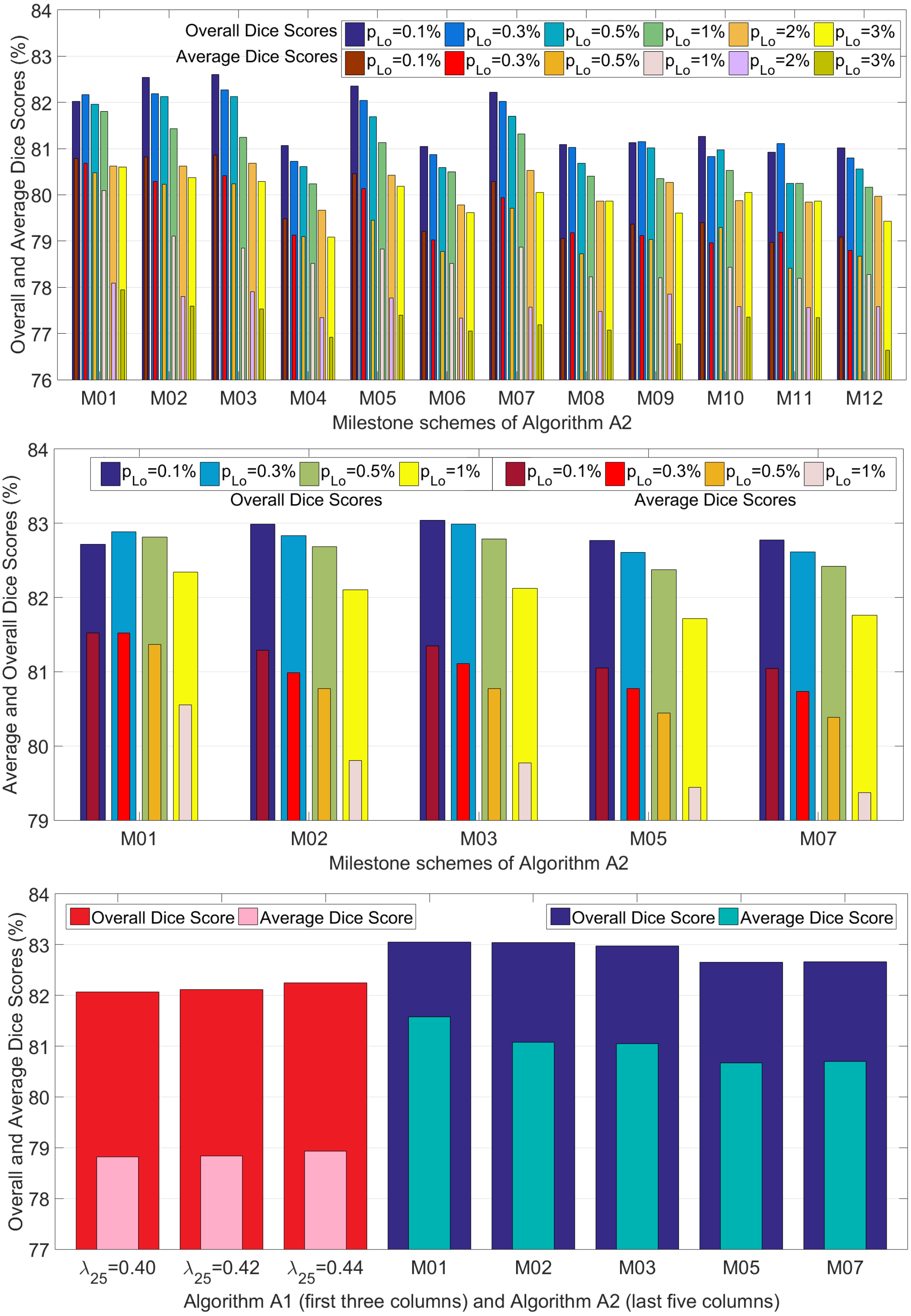

Figure 4.8: Average and overall Dice Scores achieved by the random forest classifiers RF_1 (top), RF_2 (middle), and RF_3 (bottom) in the segmentation of whole tumors, after histogram equalization via Algorithm A2 using various milestone settings and p_{Lo} parameter values, or Algorithm A1 with various values of λ_{25}. RF_3 combined with Algorithm A2 used $p_{\mathrm{Lo}} = 0.5\%$.

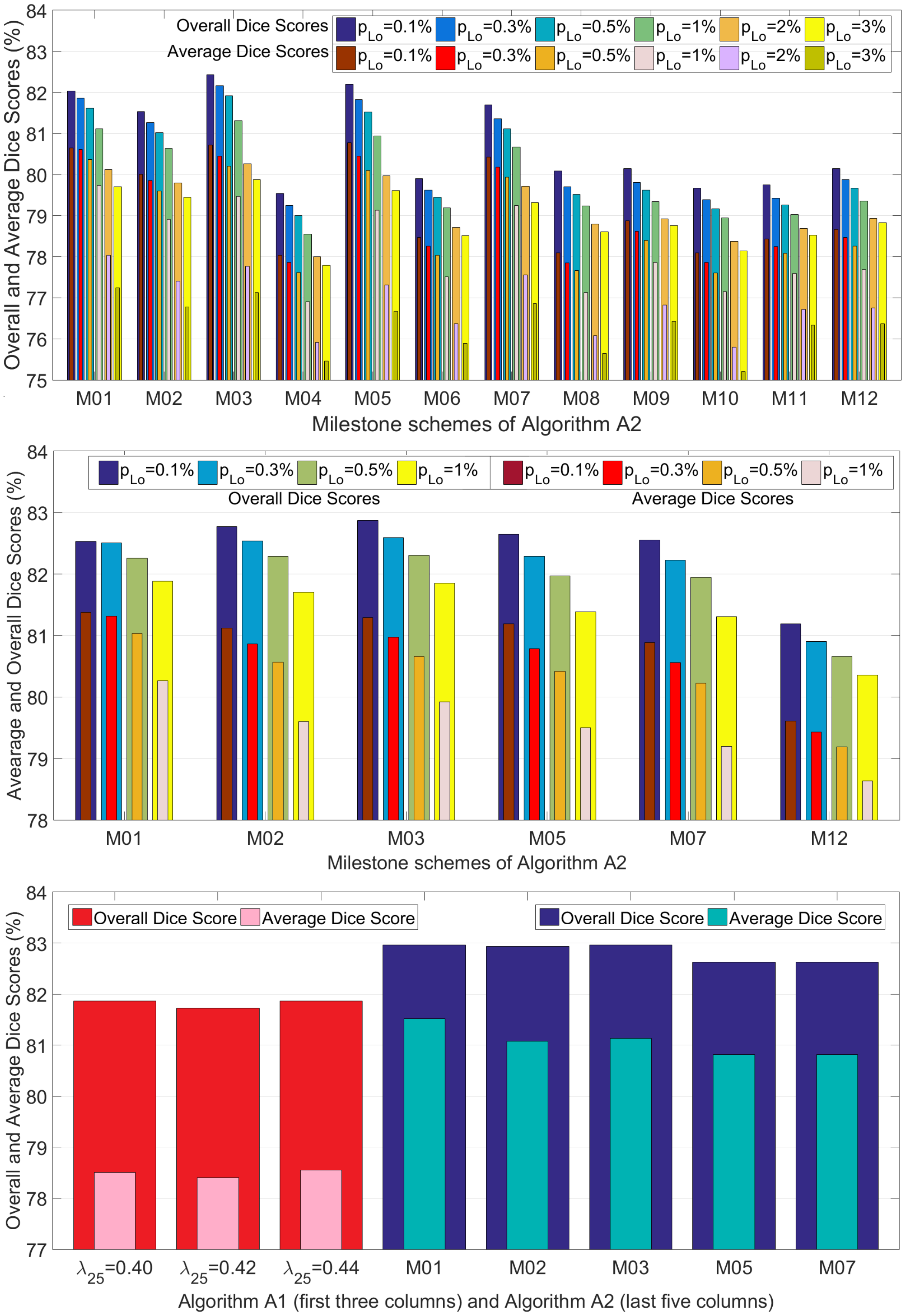

Figure 4.9: Average and overall Dice Scores achieved by the random forest classifiers RF_1 (top), RF_2 (middle), and RF_3 (bottom) in the segmentation of whole tumors, after histogram equalization via Algorithm A2 using various milestone settings and p_{Lo} parameter values, or Algorithm A1 with various values of λ_{25}. KNN_3 combined with Algorithm A2 used $p_{\text{Lo}} = 0.5\%$.

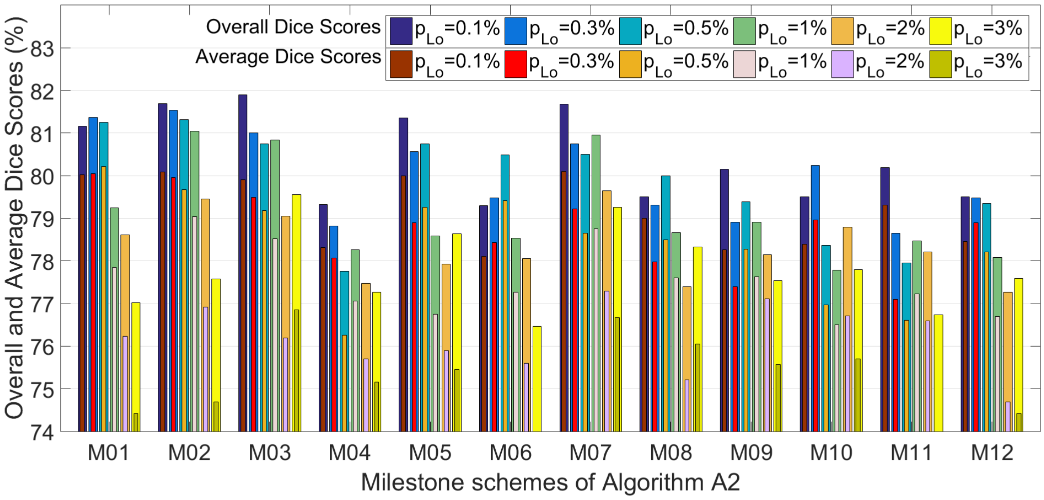

Figure 4.10: Global accuracy achieved by the SVM_1 classifier in the segmentation of whole tumors, after histogram equalization via Algorithm A2 using various milestone settings and p_{Lo} parameter values.

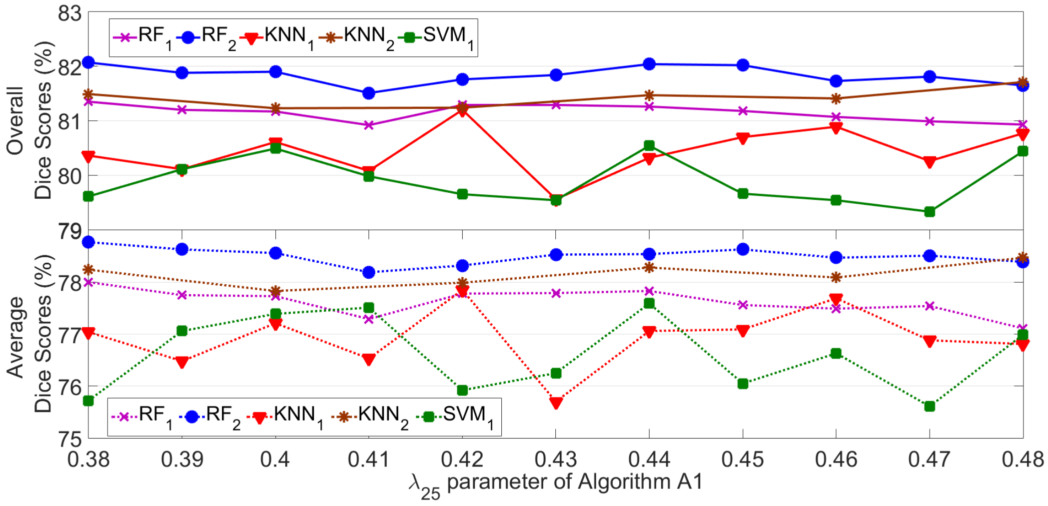

Figure 4.11: Average and overall Dice Scores achieved by the SVM_1, KNN_1, KNN_2, RF_1 and RF_2 classifiers in the segmentation of whole tumors, after histogram equalization via Algorithm A1 using various values of parameter λ_{25}.

For the whole brain tumor segmentation problem, the obtained Dice scores are represented in Figures 4.8-4.11 in an analogous way to Figures 4.4-4.7. The difference here is that instead of a single global accuracy value we have an average and an overall Dice score. The latter is always greater by 2-3%, due to the fact that larger tumors are likely to be detected with better accuracy than smaller ones.

Figure 4.8 presents the Dice scores obtained by the random forest classifier when using Algorithm A2 for histogram equalization. The plot on the left side shows that at low amount of training data (setting RF_1), the milestone schemes M01, M02, M03, M05 and M07 performed better than the others, while parameter p_{Lo} should be kept below 0.5% to achieve fine accuracy. Using a larger amount of training data (with RF_2 and RF_3) leads to higher global accuracy values, while the best performing settings prove to be milestone schemes M01, M02 and M03 with p_{Lo} around 0.3%. The plot on the right side also includes three columns that represent best Dice scores obtained with histogram equalization via Algorithm A1, having values 0.8% below the highest ones provided by Algorithm A2.

It is not surprising at all that KNN classifier provides similar results (Figure 4.9). Dice scores are somewhat lower than in case of using RF classifier, mainly due to using smaller sets of training data because KNN becomes prohibitively slow when the training data exceeds certain size. KNN gives best Dice scores when combined with Algorithm A2 using milestone scheme M03 and $p_{\mathrm{Lo}} = 0.1\%$.

Figure 4.10 exhibits the Dice scores obtained by the ensemble of SVM classifiers using setting SVM_1. The best performing milestone schemes of Algorithm A2 are the same ones as in case of KNN or RF classifiers, but not always in the same order. The comparison with the Dice scores achieved using Algorithm A1 for histogram equalization (Figure 4.11) reveals that Algorithm A2 at its best settings can provide Dice scores that are greater by 1%. Figure 4.11 presents the Dice scores achieved by various classifiers combined with Algorithm A1, plotted against the value of parameter λ_{25}. None of the classifiers is really influenced by the value of parameter λ_{25}. The absolute value of the Dice scores are below the ones provided by Algorithm A2.

The recommendations formulated based on the experiments presented above are listed below:

- Algorithm A2, proposed in general form by Nyúl *et al.* [20], can lead to better accuracy than a well-designed linear transform in machine learning based segmentation of MRI data, if it is properly adjusted.
- The best performing milestone schemes contain no more than five milestones, including the ones situated at the two ends of the intensity range, p_{Lo} and p_{Hi}.
- The best performing milestone schemes do not use milestones at p_{10} and p_{90} percentiles. The milestones situated closest to p_{Lo} and p_{Hi} should be at least as far as p_{20} and p_{80}, respectively.
- The experiments showed that p_{Lo} should be set at the 0.5% percentile or below that, and accordingly, p_{Hi} at the 99.5% percentile or above that.
- The above recommendations are valid for histogram equalization of MRI data, no matter whether it contains focal lesions or not.

The results of the investigation suggests that studies like Tustison *et al.* [29] may have achieved up to 1% higher Dice scores by using the histogram equalization of Nyúl *et al.* [20] with the above recommended settings, instead of deploying the simple linear transform. Similarly, studies like Soltaninejad *et al.* [31] and Pinto *et al.* [30] may have obtained half to one percent higher Dice scores if they used the Algorithm A2 with considerably less

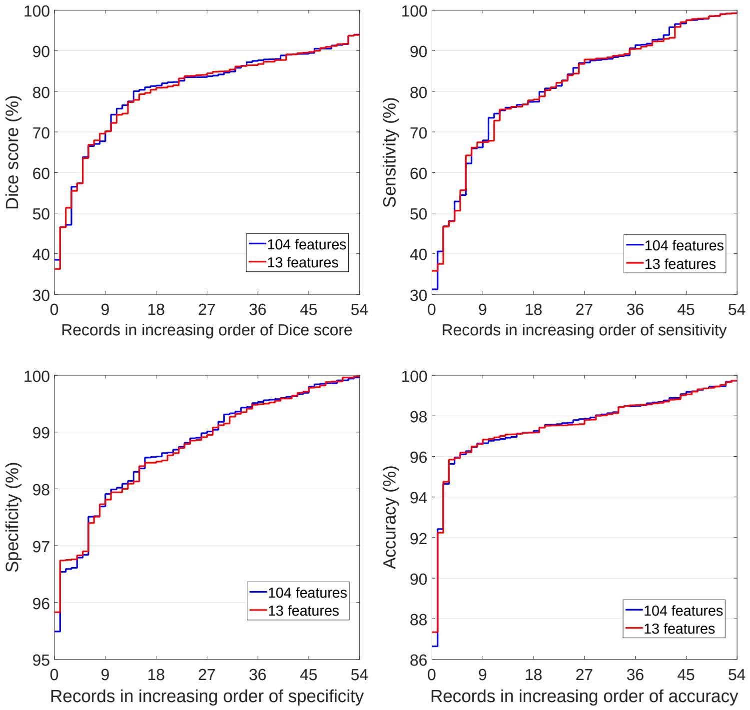

Figure 4.12: Segmentation quality indicators obtained for each of the 54 LGG tumor volumes using the reduced set of 13 features, and the original set of 104 features.

milestones. Further on, the uncountable amount of brain MRI segmentation papers within the BraTS mainstream, which only mention having employed the histogram equalization of Nyúl *et al.* [20], may improve their segmentation accuracy by adjusting their work according to the recommendations.

4.3.2 Feature generation and selection

The proposed algorithm was evaluated using the $n_V = 54$ LGG tumor records of the BraTS 2015 data set. The cardinality of decision tree ensembles was set to $n_T = 125$. Each decision tree was trained using the feature vectors of $N = 10000$ voxels, out of which $p_N = 92\%$ were negatives. The morphological post-processing set those voxels positive, which had at least 1/3 of its neighbours labeled as positive by the ensemble.

Table 4.4 exhibits the items of the final reduced feature set, obtained with the proposed method, using parameter $\alpha = 0.15$, while the last row of Table 4.3 shows the quality indicators of this reduced feature set. This result is not surprising, it is in concordance to most expectations after having seen Table 4.3. Two of the data channels, namely T1 and T1C are scarcely represented in the reduced feature vector, their presence is not required if we only want to separate the whole tumor from normal tissues. The five most important features represent data channels FLAIR and T2. Further on, it is important to remark, that median intensities are not at all present in the reduced feature set. The extraction of medians represents a high computational load, so the elimination of such features strongly accelerated the whole segmentation process. It is also visible, that the

Table 4.4: The final reduced feature set

Feature name			Rate of usage (RU)	Rate of accuracy (RA)
Data channel	Operation	Neighborhood		
FLAIR	Average	11×11	95.81%	95.88%
T2	Average	11×11	77.15%	96.16%
T2	Minimum	$3 \times 3 \times 3$	56.12%	96.26%
FLAIR	Average	$3 \times 3 \times 3$	30.89%	96.76%
T2	Average	$3 \times 3 \times 3$	24.64%	96.58%
T1C	Average	11×11	23.84%	95.09%
FLAIR	Average	3×3	16.59%	95.85%
T1C	Average	$3 \times 3 \times 3$	15.59%	94.37%
FLAIR	Maximum	$3 \times 3 \times 3$	12.85%	93.33%
FLAIR	Minimum	$3 \times 3 \times 3$	12.42%	91.88%
T2	Maximum	$3 \times 3 \times 3$	9.99%	91.56%
T1C	Maximum	$3 \times 3 \times 3$	8.47%	90.27%
T1	Minimum	$3 \times 3 \times 3$	7.64%	89.28%

observed four features were eliminated. This does not mean that they are not important, since all computed features were extracted from the observed ones.

Figure 4.12 presents the quality indicator values obtained for individual LGG tumor records, namely DSC_i $(i = 1 \dots n_V)$, plotted in increasing order of the quality indicator. The blue and the red curve shows the outcome of segmentation in case of the original set of 104 features, and the reduced set of 13 features, respectively. There is no visible difference between the two feature sets, in terms of these quality indicators.

The reduced feature set can be extracted in 7-8% of the time necessary for the full feature set. Using the reduced set of features instead of the full one, the duration of the whole data processing starting from feature extraction and finishing with the post-processing, reduces from 48 to 15 seconds, which means an acceleration of the whole segmentation process over 3 times. These benchmarks were measured on a notebook computer with quad-core i7 processor running at 3.4GHz, using a single core of the microprocessor.

The results of this study does not imply that the final feature set is ideal for all brain tumor segmentation problems. For example, a different study will be necessary for the multiple class segmentation of brain tumors (edema, tumor core, necrotic and active parts of the tumor), for which there is available ground truth in the BraTS data.

4.3.3 Spectral resolution

Two MRI data processing frameworks are involved in this study. The first one the iSeg-2017 dataset uses various machine learning models to classify pixels of 6-month old baby brains into three classes that represent the main tissue types: white matter (WM), grey matter (GM), and cerebro-spinal fluid (CSF) [116]. The second frameworks 50 LGG from

Table 4.5: Observed and generated features for both datasets

Feature		iSeg-2017		BraTS 2019 data			
Neighborhood	Operation	T1	T2	T1	T2	T1C	FLAIR
Observed (1×1)	–	1	1	1	1	1	1
Planar 3×3	AVG	1	1	1	1	1	1
Planar 5×5	AVG	1	1	1	1	1	1
Planar 7×7	AVG	1	1	1	1	1	1
Planar 9×9	AVG	1	1	1	1	1	1
Planar 11×11	AVG	1	1	1	1	1	1
Spatial $3 \times 3 \times 3$	AVG	1	1	1	1	1	1
Spatial $3 \times 3 \times 3$	MIN	1	1	1	1	1	1
Spatial $3 \times 3 \times 3$	MAX	1	1	1	1	1	1
Features per data channel		9	9	9	9	9	9
Relative coordinates (x, y, z)		3		–			
Total features		21		36			

BraTS 2019 performs two-class classification of MRI records that contain focal lesions (gliomas) [117]. Here the pixels are assigned to classes called negative and positive, where the latter includes the whole tumor (union of necrotic tumor, tumor core and edema). The original MRI data is stored with 16 bit spectral resolution, meaning that each pixel in each data channel is represented by an integer value stored in two bytes.

The two or four observed features in case of the iSeg-2017 and BraTS-2019 records, respectively, are not enough for accurate classification of pixels. The neighbourhood of each pixel contains useful information regarding the correct label of the pixel. This is why, a series of further features are extracted from various neighbourhoods of the pixels, as presented in Table 4.5. In each data channel, average intensity values are extracted from planar neighbourhoods of sizes ranging from 3×3 to 11×11. Further on, average, minimum, and maximum intensity is extracted from $3 \times 3 \times 3$ sized spatial neighbourhood. In case of the iSeg-2017 data, we also included relative x, y and z spatial coordinates of the pixel into the feature vector. These three features represent atlas information, they can provide additional information to the classification process [118, GyA3], as brain tissues correlate to a certain extent to a general map of the brain. In case of MRI volumes with focal lesions it is not useful to use atlas information, since the position, shape, and size of the tumor differ a lot from patient to patient. So the feature vector has 21 elements in case of the infant brain MRI volumes, and 36 features in case of the BraTS MRI volumes.

Each data channel of all records was fed to histogram equalization and was first transformed to the $[0, 1]$ interval. Further on, the records were resampled at various spectral resolutions from 2 bits (3 intensity levels) up to 10 bits (1023 intensity levels).

Figure 4.13 exhibits the accuracy benchmarks obtained for the iSeg-2017 dataset. In panels (a), (b), (c), and (d) of Figure 4.13, the average value of the represented statistical indicator obtained for the three tissue types were averaged to have a single indicator that

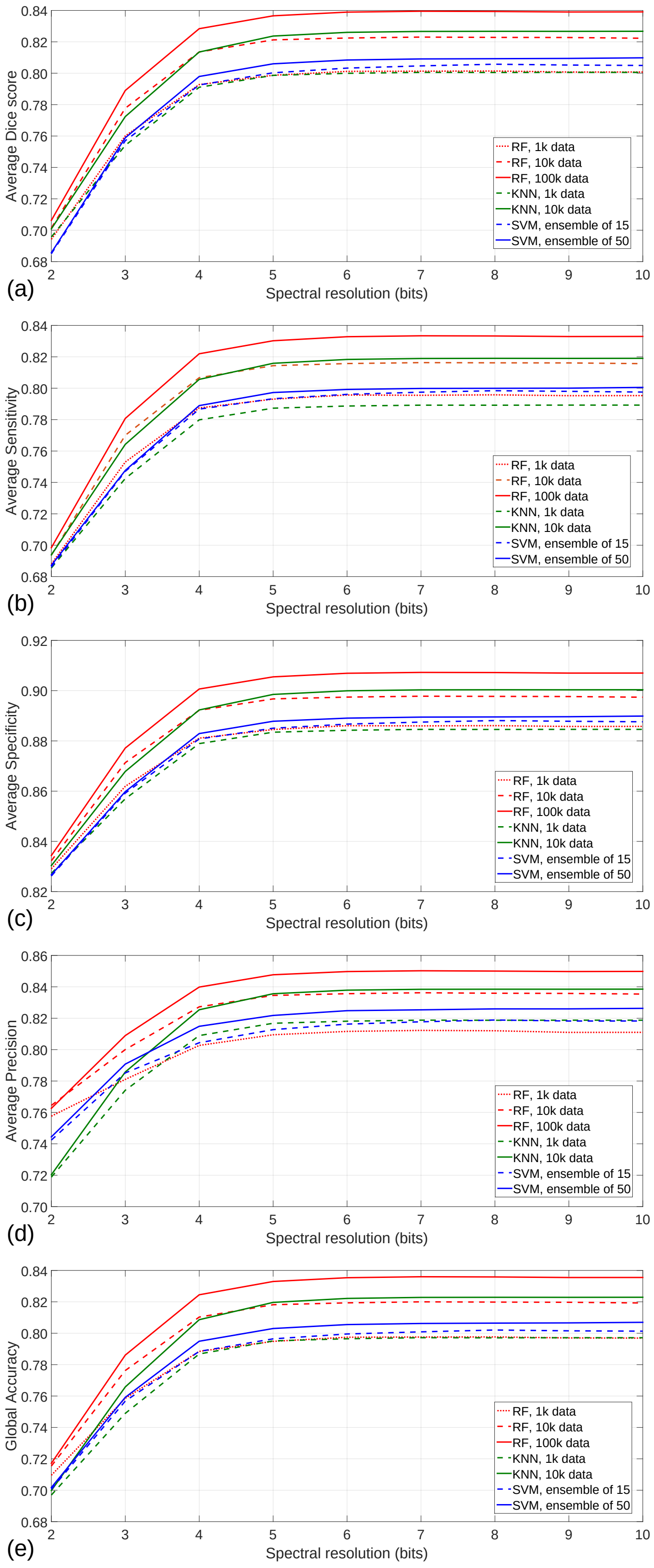

Figure 4.13: Statistical accuracy indicators obtained by various classification algorithms for the iSeg-2017 data, plotted against spectral resolution: (a) the average of the three $\overline{\mathrm{DSC}}_\tau$ values obtained for the three tissue types ($\tau \in \{\mathrm{CSF}, \mathrm{GM}, \mathrm{WM}\}$); (b) the average of three $\overline{\mathrm{TPR}}_\tau$ values; (c) the average of three $\overline{\mathrm{TNR}}_\tau$ values; (d) the average of three $\overline{\mathrm{PPV}}_\tau$ values; (e) the global accuracy ($\overline{\mathrm{ACC}}$).

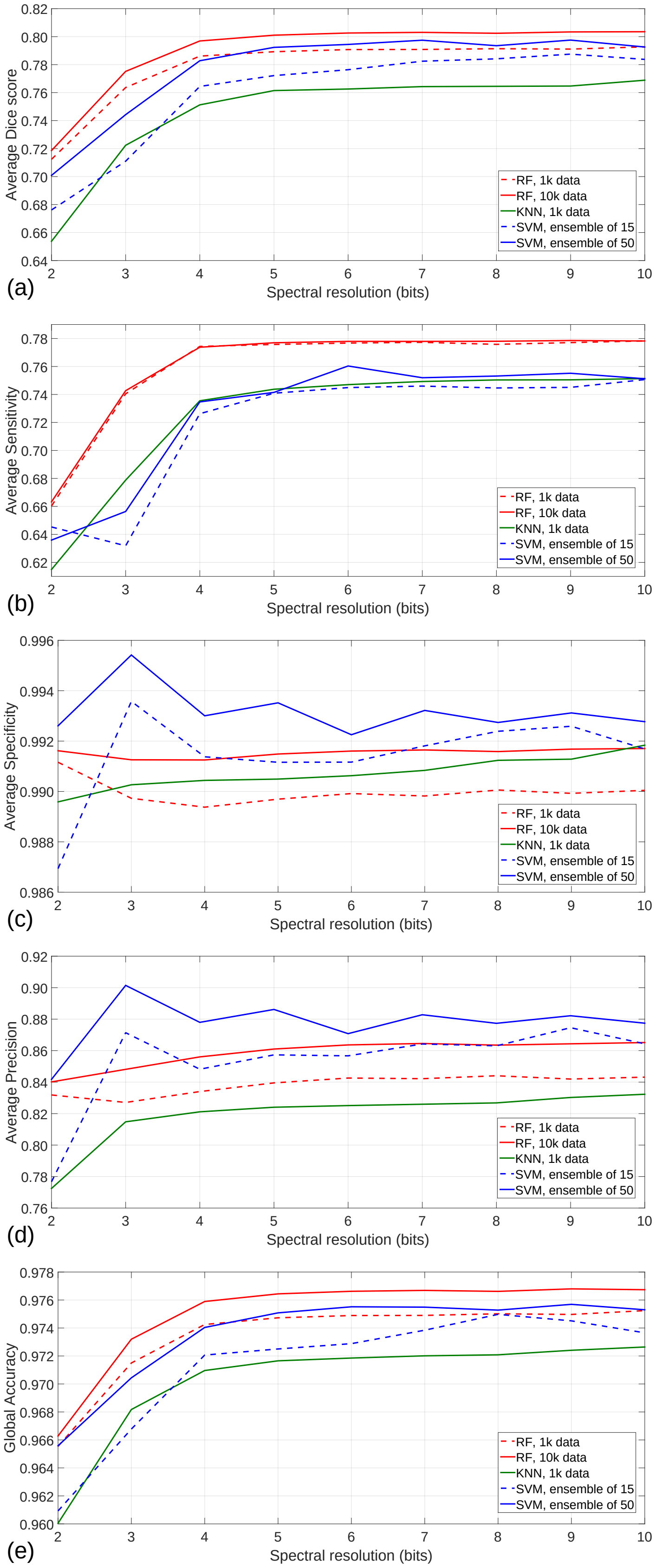

Figure 4.14: Statistical accuracy indicators obtained by various classification algorithms for the BraTS-2019 data, plotted against spectral resolution: (a) the average $\overline{\mathrm{DSC}}_{\mathrm{POS}}$ Dice score of the positive class; (b) the average sensitivity values $\overline{\mathrm{TPR}}_{\mathrm{POS}}$; (c) the average specificity values $\overline{\mathrm{TNR}}_{\mathrm{POS}}$; (d) the average precision values $\overline{\mathrm{PPV}}_{\mathrm{POS}}$; (e) the global accuracy ($\overline{\mathrm{ACC}}$).

characterizes the whole set of pixels in a record. Global accuracy shown in Figure 4.13(e) by definition refers to the rate of correct decisions obtained for all pixels, so here we did not have to average three value obtained for various tissues. Apparently all plots in the graphs of Figure 4.13 look like some saturation curves, which indicate that accuracy is very low at 2 bits spectral resolution, and it visibly rises up to 6 bits. Above this limit, at finer resolutions the accuracy of segmentation does not improve, it seems to stagnate. Since all indicators in case of all investigated classification methods comply to this criterion, we may assert that the iSeg-2017 data does not need more than 6 bits spectral resolution to grant the best achievable segmentation accuracy.

Figure 4.14 exhibits the accuracy benchmarks obtained for the BraTS-2019 dataset. Here the statistical indicators are presented for the positive class, because that is the real aim of segmentation. Again here, the most important benchmarks, those which account for both kinds of misclassifications (false positives and false negatives), namely the average Dice score and the global accuracy, are represented by some saturation curves with their saturation point at 6 bits resolution. Due to the imbalance of positive and negative pixels in the volumes – only 8% of all pixels are positive – the numerical values of statistical indicators significantly differ between the two datasets. We could not call the tumor segmentation accurate without having specificity around 99%. The global accuracy also has much higher values at the tumor segmentation, due to the correctly identified high majority of the negative pixels. The average specificity and precision plots do not have the specific saturation curve shape, in case of SVM based classification these benchmarks have high values at a coarse spectral density of 3 bits.

Among all investigated classification methods, random forest seems to provide the best accuracy, regardless to the dataset and spectral resolution. Larger amount of training data usually leads to better decisions, and larger ensembles of SVM units also give better accuracy than smaller ones.

Benchmarks obtained from the experiments using both datasets suggest that a 6-bits spectral resolution is necessary requirement for fine quality in the segmentation problem of MRI data. This also means that an 8-bit representation of each attribute of the pixels in the MRI data is more than enough, there is no need to maintain during the whole data processing the 16-bit resolution used by the MRI devices at image creation. Thus it is possible to save half of the storage space required in those environments where hundreds of feature values are extracted for each pixel.

4.3.4 Atlas based data enhancement

The whole set of 54 LGG tumor volumes was divided into two disjoint set of records, each of which played the role of train and test data set in turns. In each turn, atlases of various sizes were built using values of parameter S ranging from 60 to 120 in steps of 10, while parameter δ was set to one. Ensembles were trained using feature vector sets ranging from

Table 4.6: Average values of accuracy indicators in case of various sizes of the train data set, obtained with no atlas or with atlas of different sizes.

Accuracy	Data	Atlas size							
indicator	size	no atlas	$S = 60$	$S = 70$	$S = 80$	$S = 90$	$S = 100$	$S = 110$	$S = 120$
Average Dice score	10k	81.458	82.159	82.278	82.200	**82.333**	82.326	82.288	82.160
	100k	82.519	82.926	82.975	83.020	83.048	**83.062**	83.033	82.964
	500k	82.473	83.187	83.241	83.253	83.336	**83.352**	83.290	83.223
	1000k	82.480	83.203	83.270	83.250	83.348	**83.354**	83.313	83.221
Overall Dice score	10k	83.526	83.993	84.101	84.045	84.169	**84.173**	84.115	84.004
	100k	84.320	84.584	84.605	84.642	84.673	**84.688**	84.649	84.563
	500k	84.322	84.721	84.774	84.764	84.857	**84.873**	84.818	84.731
	1000k	84.346	84.784	84.808	84.774	84.864	**84.881**	84.830	84.740
Average Sensitivity	10k	83.480	83.432	**83.598**	83.365	83.431	83.466	83.507	83.263
	100k	**85.332**	84.967	85.005	84.884	84.987	85.034	84.978	84.813
	500k	**86.266**	85.760	85.846	85.695	85.907	85.957	85.831	85.688
	1000k	**86.860**	86.265	86.351	86.224	86.400	86.454	86.380	86.179
Average Specificity	10k	98.775	98.832	98.832	98.844	**98.857**	98.854	98.840	98.848
	100k	98.733	98.774	98.773	**98.792**	98.786	98.782	98.782	98.787
	500k	98.638	98.715	98.715	**98.731**	98.723	98.719	98.724	98.725
	1000k	98.575	98.669	98.670	**98.677**	98.675	98.669	98.670	**98.677**

ten thousand to one million items, out of which 93% were negatives and 7% positives. The training and testing cycle was performed for all cases and accuracy indicators recorded for each volume separately. Statistical evaluation was performed based on these recorded accuracy indicator values.

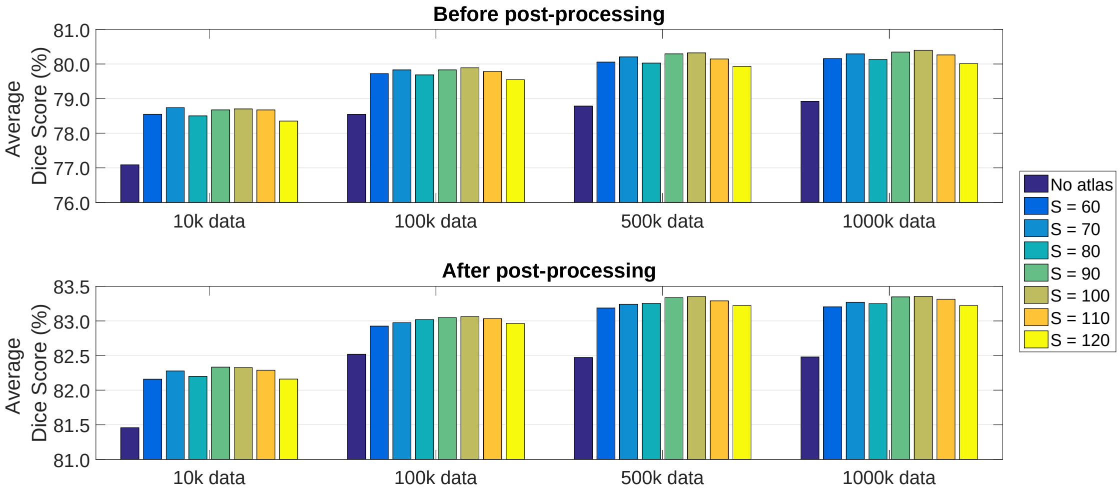

Figure 4.15: Average Dice scores obtained for the 54 LGG tumor volumes, before and after post-processing, without atlas and with atlas of size ranging in the set $S \in \{60, 70, \ldots, 120\}$. The introduction of the atlas improves all DSC values by approximately 1.5%. The highest average Dice scores are achieved using the atlas of size $S = 100$.

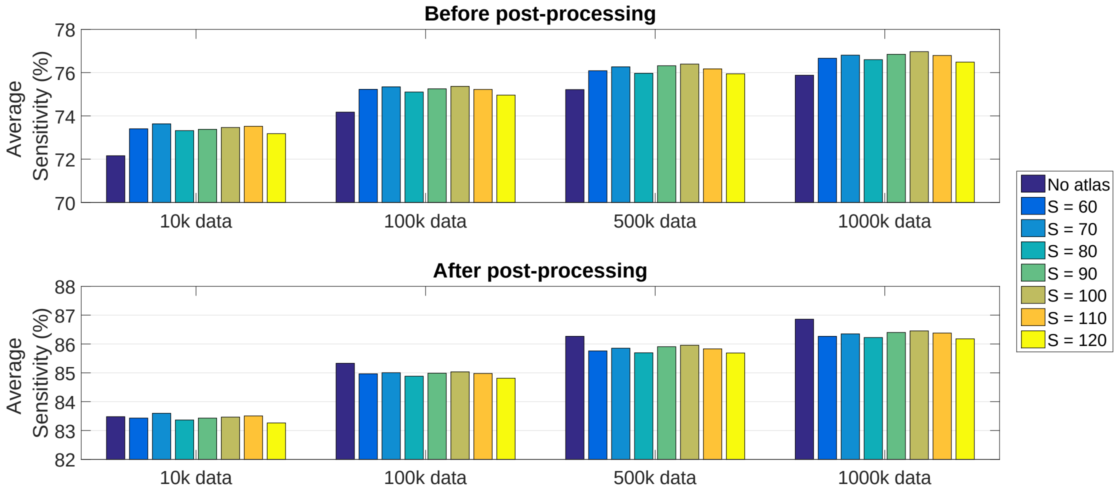

Figure 4.16: Average Sensitivity values obtained for the 54 LGG tumor volumes, before and after post-processing, without atlas and with atlas of size ranging in the set $S \in \{60, 70, \ldots, 120\}$. Post-processing improves all average Sensitivity values by approximately 10%, however, the atlases seem to slightly damage this improvement.

Table 4.6 reports average and overall values of the most important accuracy indicators obtained over all the LGG tumor volumes, in all the scenarios depicted in the previous paragraph. The highest indicator values are highlighted in each row of the table. Dice scores are at their highest when using the $S = 100$ atlas. Sensitivity and specificity values by themselves only show that using the atlas yields the prediction of more positives, but

Dice scores become definitely higher by $0.5 - 1.0\%$. Figures 4.15, 4.16, and 4.17 exhibit the same average accuracy indicator values in graphical representation, before and after post-processing. These graphs reveal further phenomena:

1. larger train data sets provide higher DSC and TPR values, and lower TNR values;
2. TPR values become higher at the ensemble output when we use the atlas, but not after post-processing, suggesting the need for an intelligent, machine learning based post-processing instead of the currently used morphological one;
3. the beneficial effect of the atlas seems to be at its highest at $90 \leq S \leq 100$.

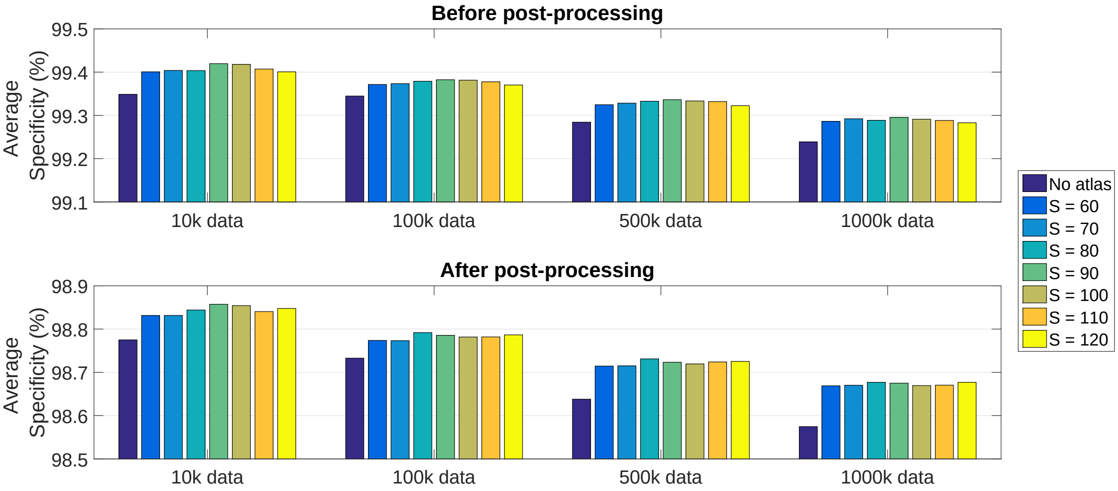

Figure 4.17: Average Specificity values obtained for the 54 LGG tumor volumes, before and after post-processing, without atlas and with atlas of size ranging in the set $S \in \{60, 70, \ldots, 120\}$. Atlases seem to contribute to the improvement of average Specificity values by approximately 0.1%.

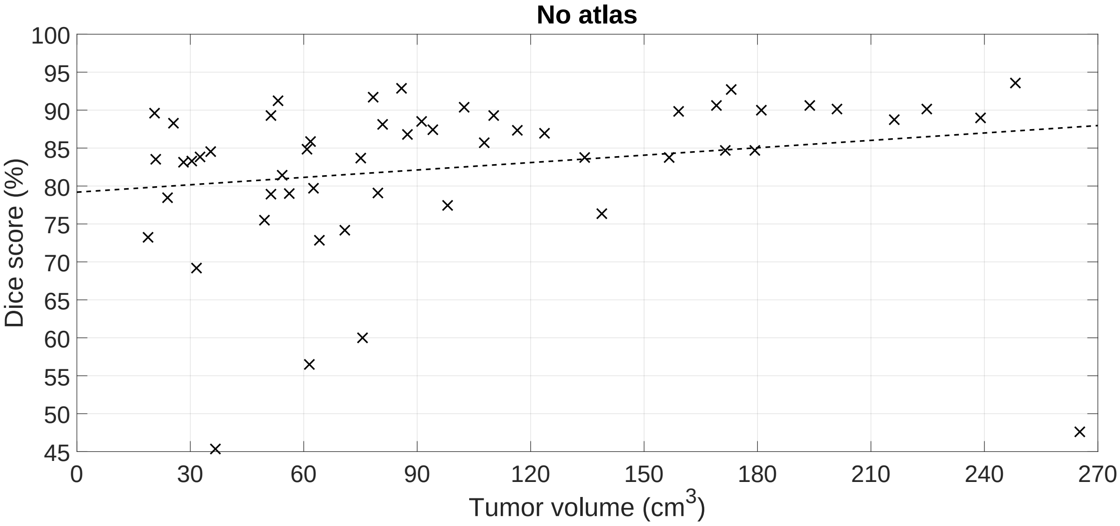

Figure 4.18: Dice scores obtained for individual LGG tumor volumes with no atlas. The trees of the ensemble were trained with feature vectors of 500k voxels.

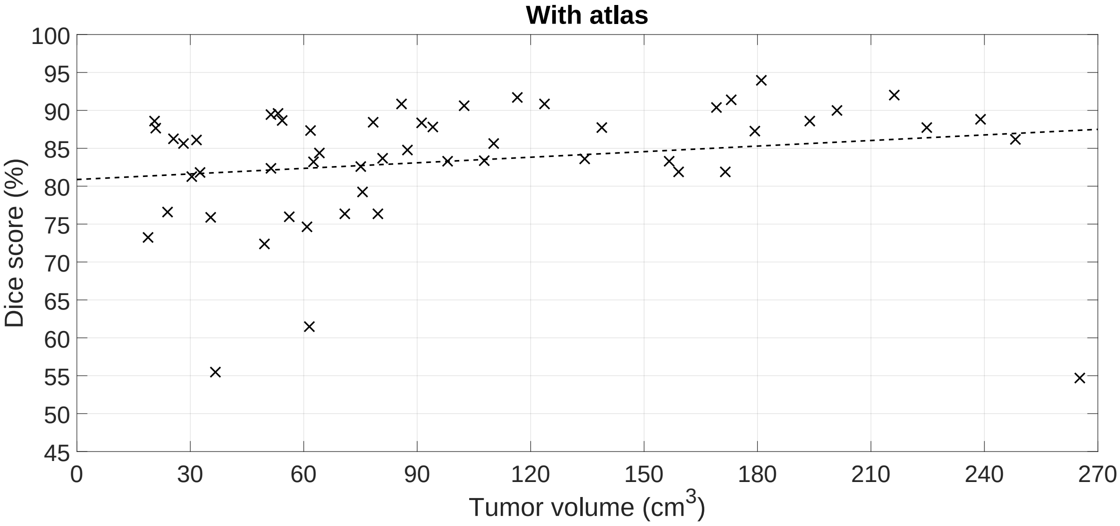

Figure 4.19: Dice scores obtained for individual LGG tumor volumes with the atlas of size $S = 100$. The trees of the ensemble were trained with feature vectors of 500k voxels.

Linear regression was employed to establish the trend of Dice scores obtained for individual LGG tumor volumes, without using atlas and with using the best performing atlas. Figures 4.18, 4.19 and 4.20 exhibit the main findings. Figure 4.18 plots the obtained individual Dice scores against the true size of the tumor, and indicates the detected linear trend with the dashed line, for the no atlas case. Figure 4.18 plots the same thing in case of atlas usage. The difference is best visible for smaller tumors: using the atlas increases the expected Dice score by 1.5%, it is well over 80% while it was below that when no atlas was employed. Figure 4.20 shows the effect of the atlas according to the linear trend for two selected sizes of the tumor, in case of various sizes of the ensemble training data set. Trends predict a $1.0 - 1.5\%$ and $0.7 - 1.0\%$ increase of the Dice score for a $10\,cm^3$ and a $100\,cm^3$-volume tumor, respectively.

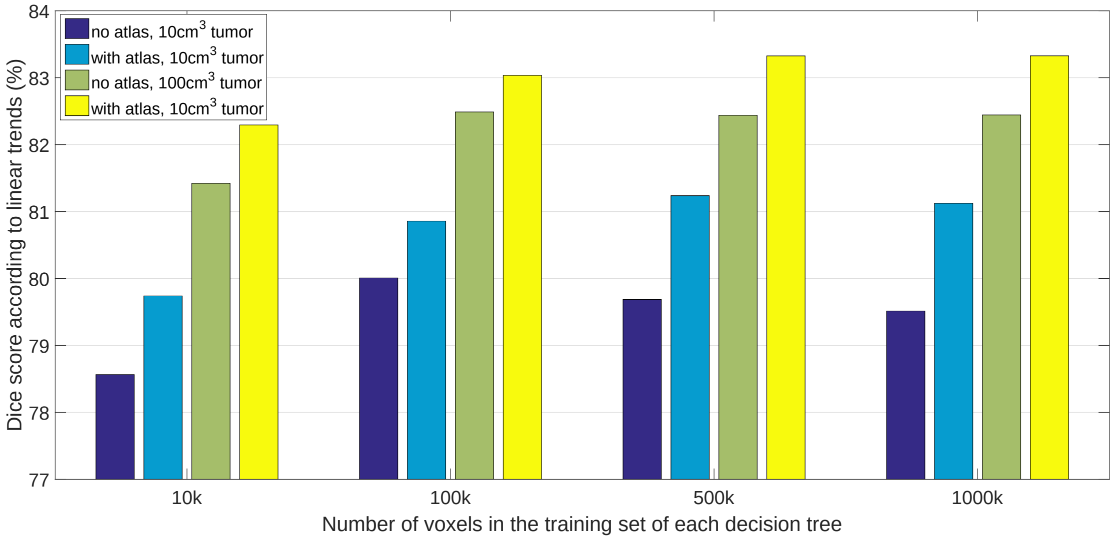

Figure 4.20: Expected Dice scores for tumors of $10\ cm^3$ and $100\ cm^3$ computed with linear regression, at various sizes of the train data set. The atlas used here is the one with size $S = 100$. The atlas improves Dice scores by 2% in case of small tumors and by 1% case of large tumors.

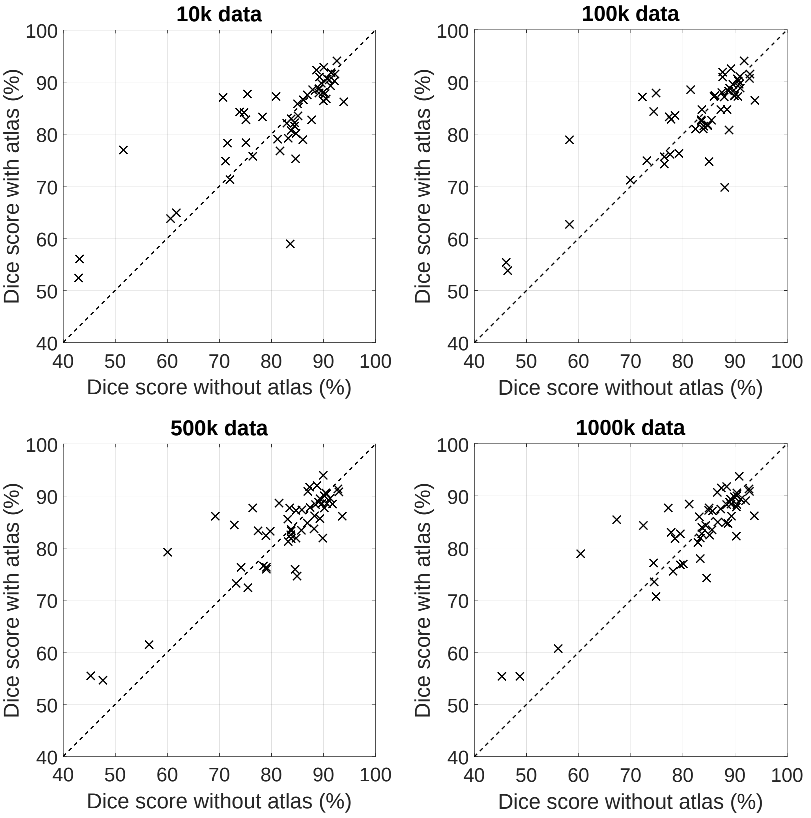

Figure 4.21: Dice scores obtained for individual LGG tumor volumes, values obtained with atlas ($S = 100$) plotted against values obtained without atlas. The four graphs represent the cases of ensembles whose trees were trained with 10k, 100k, 500k, and 1000k feature vectors, respectively. Average Dice score is higher when the atlas is used, but individual values can be lower.

Figure 4.21 shows the effect of the atlas upon Dice scores obtained for individual LGG tumor volumes, plotting DSC values achieved with atlas ($S = 100$) against DSC values achieved without atlas. The four graphs exhibit the cases of different train data sizes, from 10k to 1000k feature vectors for each decision tree of the ensemble. These graphs also indicate that the beneficial effect of the atlas is better visible in case of lower Dice score achieved without atlas. The use of the atlas improves the average quality of segmentation, but there are several MRI records which get higher Dice scores without atlas.

4.4 Conclusion

This section presented several proposals that aimed at improving the quality of the MRI data before being utilized in machine learning based segmentation. The proposed methods have provided relevant contribution in feature generation and selection, histogram equalization, and atlas-based data enhancement. Moreover, an investigation was carried out to establish the minimum spectral resolution required by machine learning based classifiers to achieve their best performance. These contributions stand at the basis of Thesis group 1.

Classification

Brain tumor segmentation is a crucial task in the initial segmentation processes (Figure 2.2, 2.3) helping to accurately separate tumors from surrounding tissues. Classical machine learning and deep learning methods each bring unique techniques and tools to address this task.

5.1 Background

Prior to my doctoral studies, I actively participated in a research project that employed Binary Decision Trees (BDT) for the segmentation of brain structures.

Binary decision trees (BDT) of unlimited depth can describe any hierarchy of crisp (non-fuzzy) two-way decisions [37]. Given an input data set of vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, where $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,m}]^T$, a BDT can be employed to learn the classification that corresponds to any set of labels $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. The classification learned by the BDT can be perfect if there are no identical training vectors with different labels, that is, $\mathbf{x}_i = \mathbf{x}_j$ implies $\lambda_i = \lambda_j$, $\forall i, j \in \{1, 2, \ldots, n\}$. The BDT is built during the training process. Initially the tree consists of a single node, the root, which has to make a decision regarding all n train data vectors. If not all n vectors have the same label, which is likely to be so, then the set of data is not homogeneous, and there is a need for a separation. The decision will compare a single chosen feature, the one with index k ($1 \leq k \leq m$), of the input vectors with a certain threshold α, and the comparison will separate the vectors into two subgroups: those with $x_{i,k} < \alpha$ ($i = 1 \ldots n$), and those with $x_{i,k} \geq \alpha$ ($i = 1 \ldots n$). The root will then have two child nodes, each corresponding to one of the possible outcomes of the above decision. The left child will further classify those n_1 input vectors, which satisfied the former condition, while the right child those n_2 ones that satisfied the latter condition. Obviously, we have $n_1 + n_2 = n$ with $n_1 > 0$ and $n_2 > 0$. For both child nodes, the procedure is the same as it was for the root. When at a certain point of the learning algorithm, all vectors being classified by a node have the same label λ_p, then the node is declared a leaf node, which is attributed to the class with index p. Another case when a node is declared leaf node is when all vectors to be separated by the node are identical, so there is no possible condition to separate the vectors. In this case, the label of the node is decided by the majority of labels, or if there is no majority, a label should be chosen from the present ones. In our application, this kind of rare leaves are labeled as tumor.

The separation of a finite set of data vectors always terminates in a finite number of steps. The maximum depth of the tree highly depends on the way of establishing the separation condition in each node. Our application uses an entropy based criterion to choose the separation condition. Whenever a node has to establish its separation criterion for a subset of vectors $\overline{\mathbf{X}} \subseteq \mathbf{X}$ containing $\overline{n}$ items with $1 < \overline{n} \leq n$, the following algorithm is performed:

1. Find all those features which have at least two different values in $\overline{\mathbf{X}}$.

2. Find all different values for each feature and sort them in increasing order.

3. Set a threshold candidate at the middle of the distance between each consecutive pair of values for each feature.

4. Choose that feature and that threshold, for which the entropy-based criterion

$$E = \overline{n}_1 \log \frac{\overline{n}_1}{\overline{n}} + \overline{n}_2 \log \frac{\overline{n}_2}{\overline{n}} \tag{5.1}$$

   gives the minimum value, where $\overline{n}_1$ ($\overline{n}_2$) will be the cardinality of the subset of vectors $\overline{\mathbf{X}}_1$ ($\overline{\mathbf{X}}_2$), for which the value of the tested feature is less than (greater or equal than) the tested threshold value.

After having the BDT trained, it can be applied for the classification of test data vectors. Any test vector is first fed to the root node, which according to the stored condition and the feature values of the vector, decides towards which child node to forward the vector. This strategy is followed then by the chosen child node, and the vector will be forwarded to a further child. The classification of a vector terminates at the moment when it is forwarded to a leaf node of the tree. The test vector will be attributed to the class indicated by the labeling of the reached leaf node.

Binary decision trees were trained to separate tumor voxels from negative ones. Due to practical reasons, negative voxels were randomly subsampled to 12% for the BDT training process. Forests were trained according to the following parameters:

1. The number of trees in the forest denoted by n_T. This parameter was usually set to 255. Experiments proved this number of trees more than necessary for good accuracy.

2. The number of data vectors used to train each tree of the forest, denoted by n_P. Typical values of this parameter ranged from 10 thousand to 500 thousand.

3. The rate (percentage) of negative labeled data within the training set, denoted by p_n.

4. The threshold of positive votes θ_p (expressed in percentage) necessary to assign a voxel to the class of positives. Making a decision according to majority voting would mean using a $\theta_p = 50\%$ threshold, but slightly shifted values of θ may lead to better accuracy.

Ideal parameter settings were identified using the so-called out-of-bag (OOB) data, as recommended by Breiman in [119]. Testing on OOB data allowed us to preselect those forests that were likely to produce high accuracy, and discard those that were prone to severe misclassifications. The best performing trees achieved $93 - 95\%$ correct decisions, while the most accurately classifying forests scored $96 - 98\%$ in labeling the OOB data.

5.2 Methods

This stage of my work investigates two distinct approaches for brain tumor segmentation: classical machine learning and deep learning methods. Each approach brings its own unique set of techniques and tools to address the intricate task of accurately delineating brain tumors from medical imaging data.

Within the realm of classical machine learning methods, I explored ensemble learning techniques that leverage sets of weak classifiers to enhance accuracy through majority voting. These algorithms encompass the Random Forest (RF) classifier, an ensemble of real AdaBoost classifiers, an ensemble of artificial neural networks (ANN), and an ensemble of binary decision trees (BDT).

On the other hand, deep learning methods will take a neural network approach, with a specific focus on convolutional neural networks (CNNs). In this category, the spotlight will be on the widely adopted U-Net architecture, originally designed for biomedical image segmentation, particularly in the context of brain tumor segmentation. The adaptation of the U-Net architecture will incorporate 3D convolutional layers, enabling effective handling of volumetric MRI data.

5.2.1 Classical methods

Ensemble learning methods use sets of weak classifiers to produce improved accuracy via majority voting. The algorithms involved in this study are:

- Random forest (RF) classifier, using the implementation given in OpenCV ver. 3.4.0. RF represents an ensemble of decision trees. The main parameter, beside the number of trees, is the maximum depth of each tree. Experiments showed that train data sets of 10000 items were best learned using the maximum depth set to 7.

- Ensemble of real Adaboost classifiers, using the implementation given in OpenCV ver. 3.4.0.

- Ensemble of artificial neural networks (ANN), using the implementation given in OpenCV ver. 3.4.0. Each ANN had the same architecture: input layer of size that corresponds to the number of features, two hidden layers of 7 and 5 neurons, respectively, and one neuron in the output layer. ANNs were trained using the backpropagation rule implemented in OpenCV.

- Ensemble of binary decision trees (BDT), using an own implementation presented in Section 5.1. BDTs can learn to perfectly separate negative from positive train samples unless there are two coincident vectors with different ground truth. Training sets of randomly chosen 10000 samples are usually learned by BDTs of maximum depth 19.3 ± 3.6 (AVG±SD). Decisions are made at average depth of 7.29 ± 2.67.

5.2.2 Deep learning methods

The image segmentation task requires a decision regarding each pixel separately. CNN networks are usually deployed for the classification of images, for example when a diagnosis is established based on a whole record. U-Net is a widely applied deep learning architecture used for brain tumor segmentation on the BraTS datasets. It was originally proposed for biomedical image segmentation tasks, particularly in the field of medical imaging. U-Net is known for its effective performance in segmenting gliomas and other brain tumors, making it a popular choice among researchers and medical professionals. The architecture of U-Net consists of an encoder-decoder structure, where the encoder learns high-level features from the input data, and the decoder reconstructs the segmented image based on the encoded features. The skip connections of the network allow the efficient transfer of low-level features, contributing to the preservation of fine details during segmentation.

I adapted the U-Net convolutional neural network architecture introduced by Ronneberger *et al.* [10] to the requirements. The architecture of the neural network, presented in Figure 5.1, is symmetric. Unlike the original U-Net that worked with 2D convolution, my model uses 3D convolutional layers. 3D convolution works in a spatial neighbourhood of each pixel instead of a planar one, each filter dealing with several slices from the MRI volume. This approach is likely to be more effective in lesion detection and produce better segmentation results.

The contracting path consists of three encoder blocks. Each encoder block doubles the number of the filters and halves the volume in each direction. These filters are responsible for the feature extraction from the 3D volume. When the filter coefficients are well trained, they can extract the important features from the volumetric data that will support the discrimination between normal and lesion pixels. Each encoder block contains a normalization, ReLU activation function and a dropout operation. The dropout helps prevent overfitting. A max pooling operation is the last one in each encoder block, which reduces the size of the volume before passing it to the next encoder level. At the end of

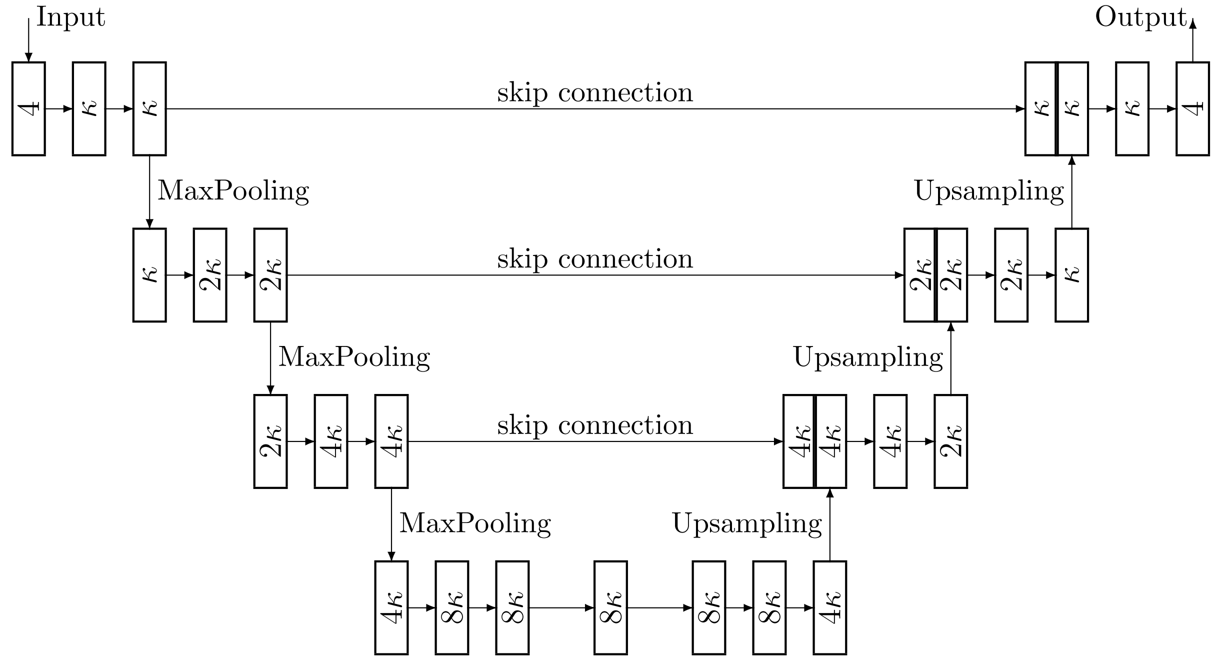

Figure 5.1: The structure of the U-Net networks used in both stages of the proposed segmentation procedure; κ is a parameter, which was tested as multiples of 16 ranging up to 112.

the contracting path there is a bottom layer, which performs the detection of tumor pixels within the reduced sized volume. The output of the bottom layer is propagated toward the decoder blocks, which build up the original sized segmented volume, containing the label information at the level of pixels. Each decoder block receives input from two sources. The output from the previous decoder or the bottom layer serves as the first source, while the second is the skip connection from the encoder situated at the same level. This mechanism helps in restoring the boundary of the detected tumor in its correct site. An important difference between the encoder and decoder blocks is that the latter performs a trilinear upsampling of the volume instead of reducing it via the max pooling operation. The final output is fed to a SoftMax operator, which is a normalized exponential function that provides predicted probabilities for classification.

The proposed architecture can work with different filter (or channel) counts depending on parameter κ, as shown in Figure 5.1. The number of filters in the first encoder block (denoted by κ) determines the filter count in all other blocks following the rule that encoder blocks double the filters count of the previous block, while the decoder blocks halve it. The number of filters should be chosen knowing that too many filters require too much storage and rise the computational burden. On the other hand, a too low number of filters cannot produce fine accuracy of the segmentation.

5.3 Results

In the current section I present the results obtained with the methods described in Section 5.2. All the above mentioned algorithms underwent a thorough evaluation process using the statistical indicator described in Section 3.3. In this stage, no post-processing had been applied yet. All tests involving classical machine learning were executed on a notebook computer with quad-core i7 processor running at 3.4GHz, using a single core of the microprocessor. Ensembles of RF and BDT performed quickly, in virtually same time. ANN and Adaboost required 2 to 5 times longer execution time, especially in larger ensembles. In the case of the deep learning methods all the tests were executed on a desktop computer with octa-core i9 processor running at 3.6GHz, and an NVIDIA GeForce RTX 3090 graphics card with 24GB of memory.

5.3.1 Classical methods

The dataset utilized for this study comprises the 54 low-grade and 220 high-grade tumor volumes of the BraTS 2015 database. The size of the ensemble varied in four steps, using values of 5, 25, 125, and 255. The quality indicators exhibited in Table 3.2 were extracted for each scenario and each individual MRI record. The average and median value for each indicator was established for overall quality characterization. The algorithms were compared in group and one against one, through tests performed with individual data volumes and extracted overall quality benchmarks. Results are exhibited in the following.

Tests on LGG tumor records

Table 5.1 presents the global average and median values for all four quality indicators, obtained using the four evaluated classification algorithms and the above mentioned four ensemble sizes. The median values are always greater than the average, since there are a few data records of reduced quality that are usually segmented much worse than all others. The best values, which are highlighted in each column, suggest that the random forest performed slightly better than any other tested algorithm. Segmentation quality rises together with the size of the ensemble up to 125 units, and seems to saturate above this value. Best achieved average Dices scores are slightly above 81%, while median values approach 85%. The accuracy of all algorithms is around 97.5%, meaning that about one pixel out of 40 is misclassified by these algorithms.

Figure 5.2 presents the Dice score and Sensitivity, respectively the Specificity and Accuracy indicator values obtained by the most accurately performing ensemble of 125 random forests, evaluated on the 54 individual LGG volumes. Apparently there are approximately 10% of the records that lead to mediocre result. In these cases the classification algorithm did not succeed to capture the main specific features of the data, probably due to the reduced quality of the recorded images.

Table 5.3 presents for each algorithm the number of records that led to Dice scores over predefined threshold values between 50% and 90%. The best values highlighted in each row of the table indicate again that random forest achieved the best segmentation quality, followed by Adaboost and ANN.

Table 5.1: Statistical accuracy indicator values of the 54 LGG tumor volumes obtained for various classifiers and ensemble sizes. The best achieved performance is highlighted in each column.

Classifiers in ensemble	Ensemble size	Dice score		Sensitivity		Specificity		Accuracy	
		AVG	MED	AVG	MED	AVG	MED	AVG	MED
ANN	5	79.60	83.57	**83.36**	86.77	98.39	98.78	97.37	97.61
	25	80.53	84.39	82.39	85.75	98.60	98.92	97.51	97.79
	125	80.52	84.07	81.47	85.20	98.70	98.92	97.54	97.84
	255	80.55	83.99	81.54	84.93	98.69	98.94	97.53	97.78
AdaBoost	5	80.20	84.28	80.63	85.46	98.85	99.03	97.63	97.74
	25	80.55	84.13	80.96	86.68	**98.86**	**99.12**	97.65	97.76
	125	80.76	84.11	81.38	86.96	98.84	99.11	97.66	97.87
	255	80.78	84.13	81.44	87.00	98.83	99.10	97.66	97.85
Random forest	5	80.02	82.94	80.39	83.38	98.79	99.03	97.57	97.90
	25	81.10	**84.44**	81.90	86.63	98.78	98.98	97.67	**98.01**
	125	**81.33**	84.36	82.84	87.31	98.73	98.93	**97.69**	97.71
	255	81.29	84.32	82.72	**87.35**	98.74	98.93	97.68	97.72
Binary decision trees	5	78.83	83.59	81.50	85.80	98.51	98.86	97.37	97.53
	25	80.13	83.83	81.17	84.72	98.75	98.95	97.57	97.59
	125	80.21	84.25	81.41	87.30	98.75	98.93	97.56	97.70
	255	80.22	84.21	81.44	87.20	98.75	98.92	97.57	97.69

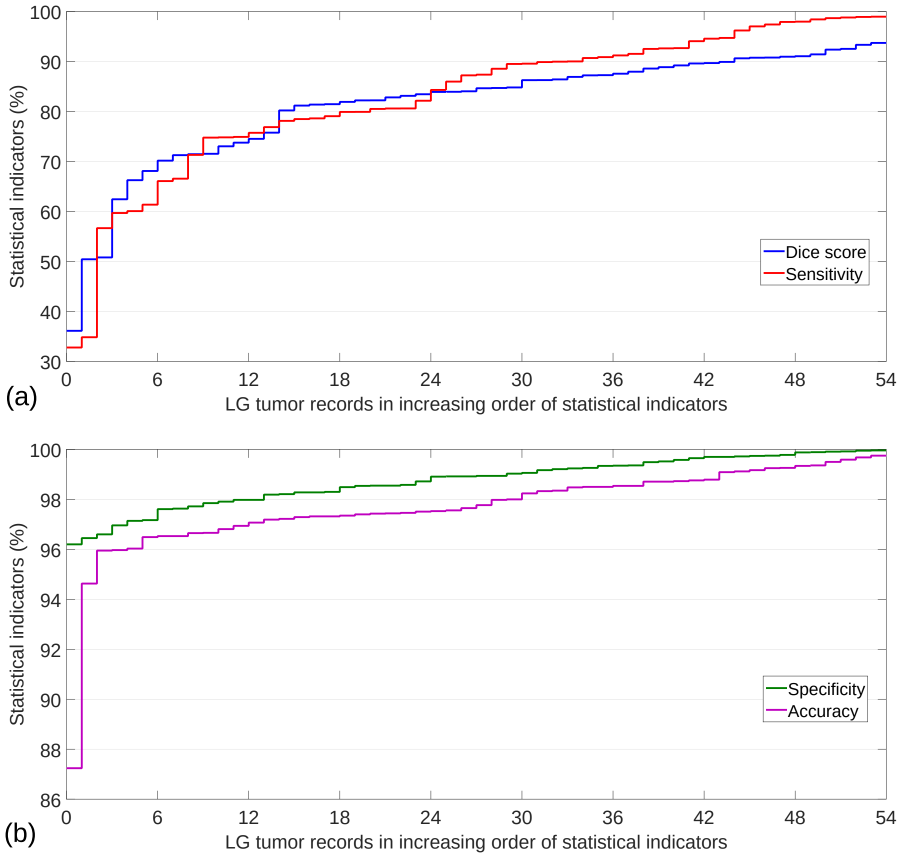

Figure 5.2: Accuracy indicator values obtained for each of the 54 LGG tumor volumes using the random forest algorithm in ensemble of 125, plotted in increasing order of the quality indicators.

Table 5.2: Dice score tournament using the 54 LGG volumes: algorithms against each other, each using ensembles of size 125. Here ANN proved to be the weakest.

Algorithm	ANN	Adaboost	RF	BDT	Won:Lost
ANN	N/A	23:31	24:30	26:28	0:3 (73:89)
Adaboost	31:23	N/A	22:32	30:24	2:1 (83:79)
RF	30:24	32:22	N/A	25:29	2:1 (87:75)
BDT	28:26	24:30	29:25	N/A	2:1 (81:81)

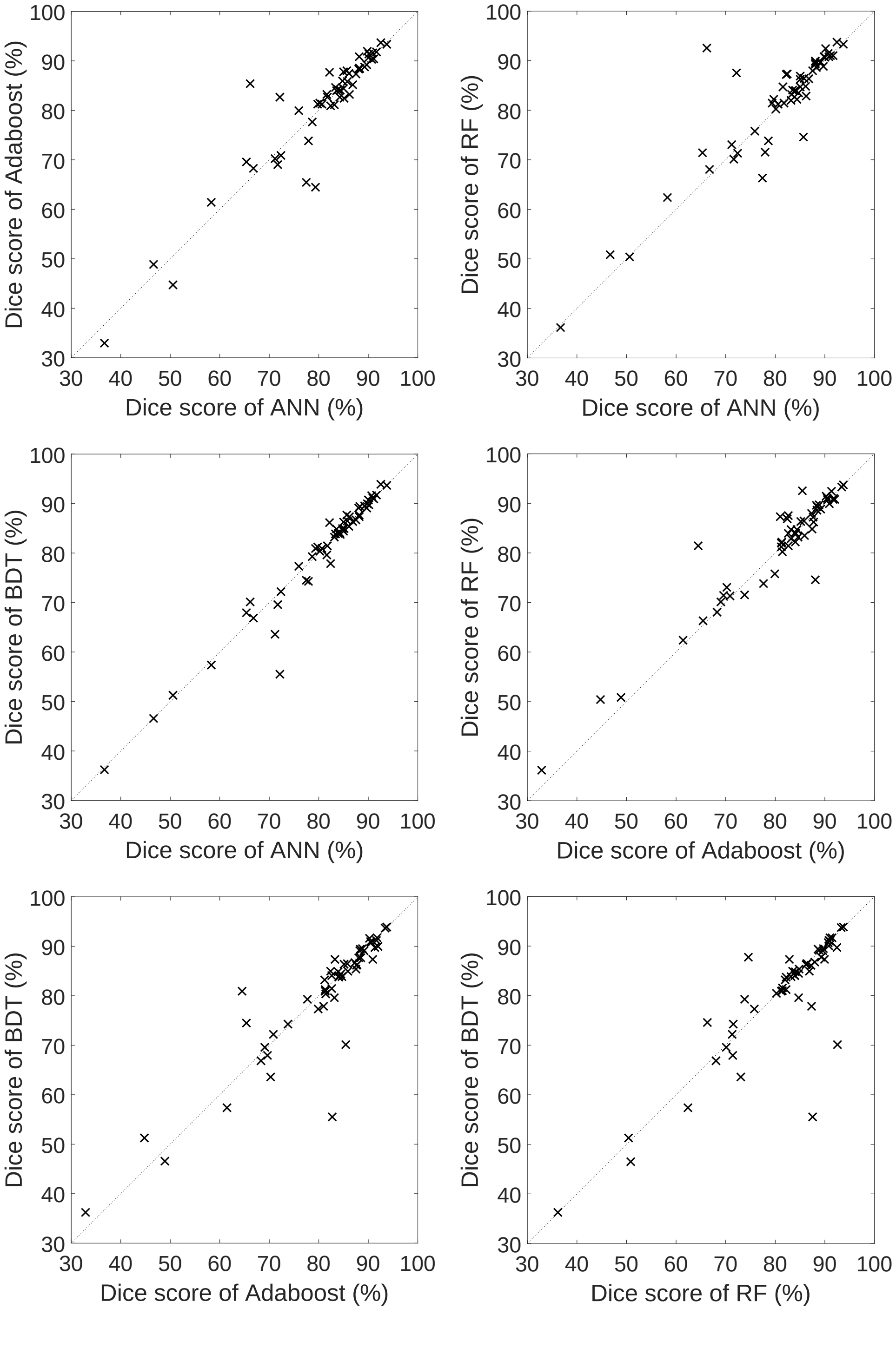

Figure 5.3: Dice scores obtained for individual volumes (LGG) by the four algorithms using ensembles of size 125, plotted one algorithms vs. another, in all possible six combinations.

Table 5.3: Comparison of classifier algorithms using the Dice scores obtained for the 54 individual LGG tumor volumes. The best achieved performance is highlighted in each row.

Classifier	ANN				Adaboost				Random forest				BDT			
Ensemble size	5	25	125	255	5	25	125	255	5	25	125	255	5	25	125	255
DSC > 50%	52	52	52	52	52	51	51	51	51	51	**53**	52	51	52	52	52
DSC > 60%	50	50	50	50	50	50	**51**	**51**	50	**51**	**51**	**51**	48	48	49	49
DSC > 70%	45	45	47	47	45	44	45	45	44	47	**48**	**48**	42	46	45	46
DSC > 75%	41	**43**	**43**	**43**	41	42	42	42	**43**	41	41	41	40	42	41	41
DSC > 80%	38	37	37	37	39	**40**	**40**	**40**	39	**40**	**40**	**40**	34	36	37	37
DSC > 85%	19	24	24	22	24	22	24	24	20	**25**	24	**25**	22	22	23	24
DSC > 90%	5	8	8	7	7	10	10	10	**12**	10	10	10	8	8	7	7

Table 5.4: Comparison of the tested ensemble learning techniques using the DSC obtained for the 220 individual HGG tumor records. Bests scores were identified and highlighted in each row of the table.

Classifier	ANN				AdaBoost				Random forest				BDT			
Ensemble size	5	25	125	255	5	25	125	255	5	25	125	255	5	25	125	255
DSC > 50%	210	210	210	210	208	208	208	209	**213**	212	212	212	208	208	209	209
DSC > 60%	195	201	**202**	201	198	200	200	201	197	200	201	201	194	198	200	201
DSC > 70%	173	175	174	174	173	175	175	175	177	180	**181**	**181**	173	174	175	175
DSC > 75%	159	164	164	164	161	163	164	164	164	**168**	166	166	160	163	164	163
DSC > 80%	135	143	145	141	141	143	143	143	142	144	**146**	**146**	138	140	143	143
DSC > 85%	109	114	114	114	112	116	116	116	117	119	**120**	**120**	114	111	117	116
DSC > 88%	67	74	74	74	76	76	76	78	75	**82**	81	**82**	70	76	75	76
DSC > 90%	38	44	44	42	46	51	51	**53**	47	49	51	50	38	47	52	51
DSC > 92%	18	23	24	23	25	26	26	25	23	24	**27**	26	21	26	25	25

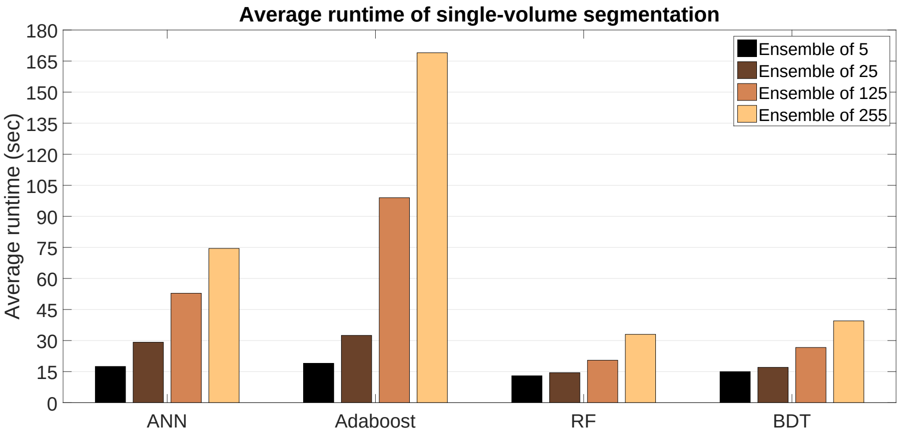

Figure 5.4: Efficiency benchmarks of the four classification algorithms: the average value of the total processing time in a single record testing problem (LGG).

Figure 5.3 exhibits in a different format the Dice scores obtained by each algorithm using ensembles size of 125 units, tested on each individual LGG tumor volume. Each graph presents the Dice scores obtained by two algorithms, plotted one against the other. Each cross (×) in these graph shows the Dice score of the two algorithms achieved on the very same data. Most crosses on each graph are close to the diagonal, indicating that the Dice score achieved by both algorithms were pretty much the same, but there are also crosses far from the diagonal, showing cases when one of the algorithms produced significantly better segmentation quality.

Table 5.2 exhibits the same data as Figure 5.3, but in a tournament format. Surprisingly, this table suggests that ANN is the worst performing algorithm in term of accuracy, because all other three classification methods achieved better Dice score in case of the majority of the data records. Again in this case, random forest proved the most accurate one, despite losing the direct comparison to BDT.

Figure 5.4 exhibits the efficiency benchmarks of the four algorithms. When we get a new data record of average size, we need to extract the 13 features for each pixel from its neighbourhood, feed the obtained data to ensemble of classification algorithms of the chosen size, and finally apply the post-processing that leads to final segmentation. The total necessary processing time, needed by each algorithm is shown in the figure.

Tests on HGG tumor records

Overall average and median values of the four main quality indicators are exhibited in Table 5.5, for all evaluated ensemble learning algorithms and various ensemble sizes. Median values were found greater than the average, for all indicators and scenarios, because there are a few records of reduced or damaged quality that are likely to be segmented considerably worse than all others. Highest values highlighted in each column of the table indicate that the random forest achieved slightly better results than any other evaluated technique.

Table 5.5: Various statistical accuracy indicator values achieved by tested techniques and ensemble sizes, expressed in percentage (%) of the 220 HGG tumor volumes. Best performance is highlighted in all columns. AVG stands for average, MED stands for median.

Classifiers in ensemble	Ensemble size	Dice score		Sensitivity		Specificity		Accuracy	
		AVG	MED	AVG	MED	AVG	MED	AVG	MED
ANN	5	79.11	84.82	83.00	**90.62**	98.40	98.88	97.50	97.98
	25	80.02	85.73	83.15	90.48	98.40	98.86	97.52	97.97
	125	80.09	85.62	**83.33**	90.58	98.40	98.85	97.52	97.96
	255	80.05	85.49	83.31	90.56	98.39	98.56	97.52	97.97
AdaBoost	5	79.69	85.37	82.24	90.00	98.48	98.92	97.53	97.97
	25	80.00	85.59	82.02	89.58	98.55	99.00	97.58	98.01
	125	79.98	85.54	81.89	89.34	**98.56**	**99.02**	97.59	98.03
	255	80.04	85.77	81.96	89.49	98.55	99.00	97.59	98.03
Random forest	5	80.29	85.59	82.95	89.83	98.45	98.93	97.56	97.99
	25	80.59	85.93	82.73	89.74	98.53	98.98	97.62	98.06
	125	80.71	86.22	82.77	89.94	98.55	99.00	**97.64**	98.08
	255	**80.74**	**86.27**	82.77	89.88	98.55	99.01	**97.64**	**98.09**
Binary decision trees	5	79.22	85.52	82.27	89.93	98.39	98.91	97.46	97.86
	25	79.80	85.36	82.12	90.05	98.50	98.94	97.54	98.04
	125	80.05	85.68	82.03	89.37	**98.56**	99.00	97.59	98.02
	255	80.03	85.73	81.95	89.46	**98.56**	98.99	97.59	98.03

The accuracy of segmentation rises together with the ensemble size up to 125 units, above which it seems to stabilize or fall slightly. Highest achieved average Dices scores approached 81%, while median values surpass 86%. The accuracy of all evaluated ensemble learning techniques is around 98%, meaning that approximately one pixel out of 50 is misclassified.

Figure 5.5 exhibits the Dice score and Sensitivity in the left panel, respectively the Specificity and Accuracy in the right panel, indicator values obtained by the random forest using ensemble of 125, which was identified as the most accurately performing algorithm. Approximately 10% of the records lead to mediocre result. In these cases the classification methods failed to capture the main specific characteristics of the data, probably because the recorded images were of low quality.

Table 5.4 shows for each test scenario (algorithm and ensemble size) the number of successfully segmented records, where the Dice score exceeded predefined threshold values ranging from 50% to 92%. Highest values highlighted for each threshold value indicate again that random forest achieved the best segmentation quality.

Figure 5.6 presents the outcome of one-vs-one comparison of the tested algorithms, each using ensembles of 125 units. Dice scores shown here were obtained on each individual HGG tumor records. Each cross (×) in the graph shows the Dice score achieved by the two ensemble learning techniques on the very same data. Most crosses are situated in the proximity of the diagonal, indicating that both algorithms obtained pretty much the same

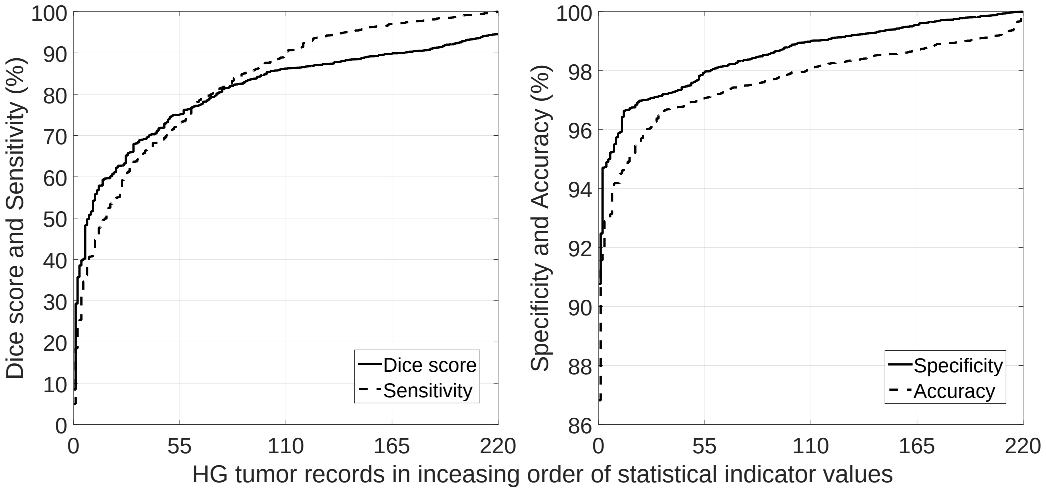

Figure 5.5: Main quality indicator values obtained for individual HGG tumor volumes, using the random forest method in ensemble of 125, sorted in increasing order.

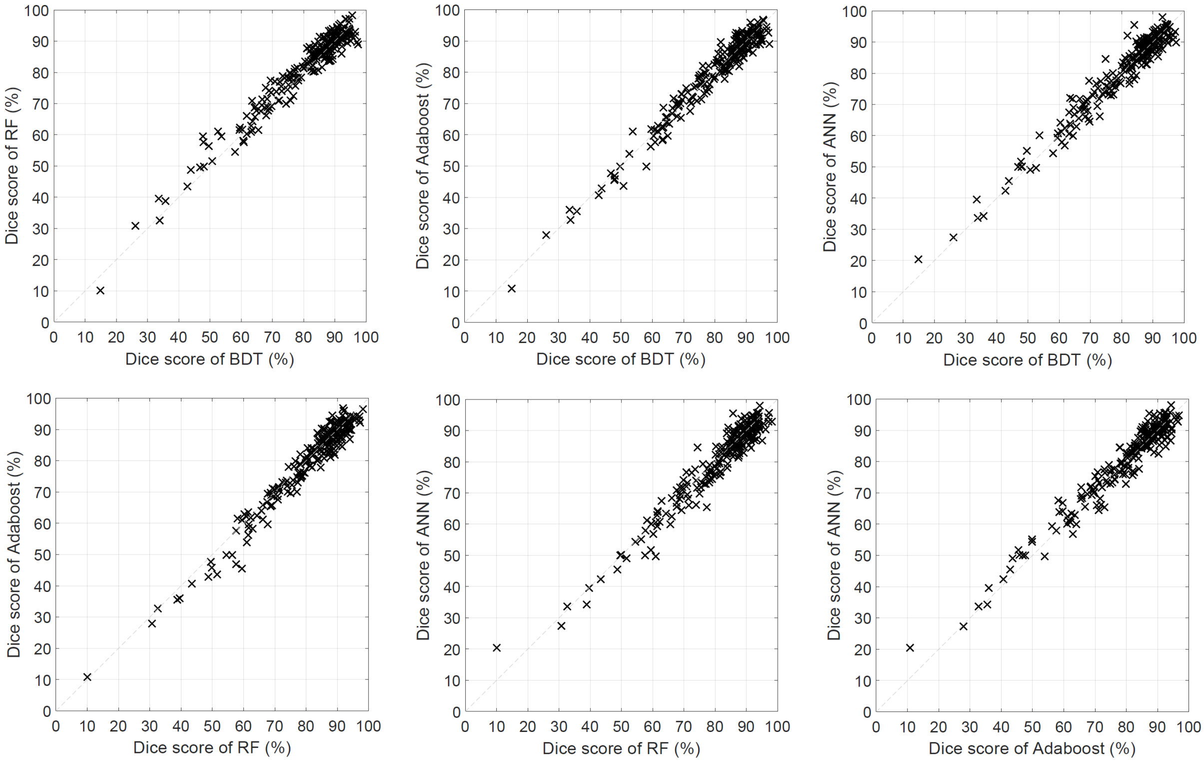

Figure 5.6: Dice scores obtained for individual volumes by the four algorithms using ensembles of size 125, plotted one algorithms vs. another, in all possible six combinations.

Table 5.6: Dice score tournament using the 54 LGG volumes: algorithms against each other, each using ensembles of size 125. Here ANN proved to be the weakest.

Algorithm	ANN	AdaBoost	RF	BDT	Won:Lost
ANN	N/A	89:131	57:163	86:134	0:3 (232:428)
AdaBoost	131:89	N/A	71:149	102:118	1:2 (304:356)
RF	163:57	149:71	N/A	147:73	3:0 (459:201)
BDT	134:86	118:102	73:147	N/A	2:1 (325:335)

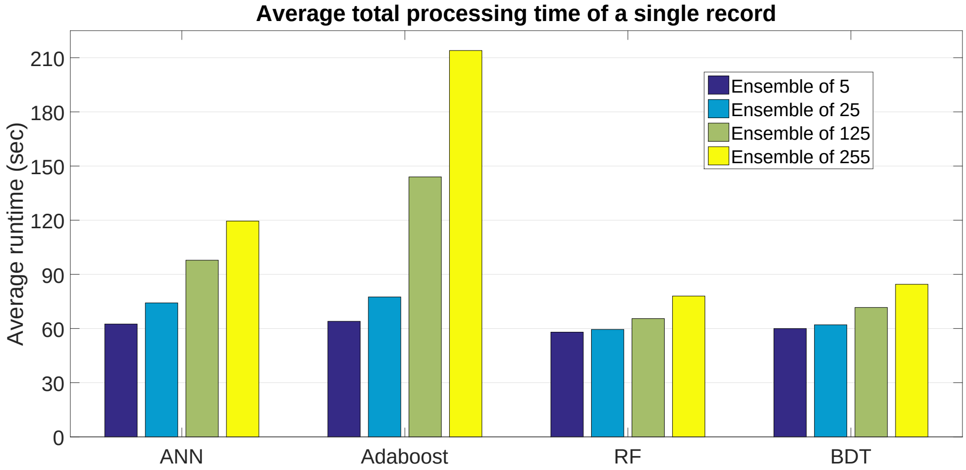

Figure 5.7: Runtime benchmarks of the four classification algorithms: the average value of the total processing time in a single record testing problem.

accuracy. There are also crosses apart from the diagonal, representing scenarios where one of the methods led to significantly better segmentation quality.

Table 5.6 exhibits the same results as Figure 5.6, but here the one-against-one outcome of tests is organized in a tournament format. The tournament was won by the random forest algorithm, followed by BDT, AdaBoost, and ANN. Figure 5.7 compares the efficiency of the four evaluated algorithms. Total runtimes exhibited here include the duration of histogram equalization and feature generation, segmentation and post-processing of an average sized never seen MR data volume.

5.3.2 Deep learning methods

Several records in the BraTS 2019 data set contain pixels with missing data. Most of them have missing data in isolated pixels only, which can be easily replaced with an averaged intensity value from the neighbourhood of the pixel. There are some volumes, especially old ones originating from the BraTS 2013 data set, which contain large areas without valid data in one or even more data channels. Especially the FLAIR data is missing from these volumes, which are the most important for accurate segmentation [GyA1, 79]. We excluded 20 HGG and 10 LGG records from the BraTS 2019 training data sets, which had lots of missing data. Consequently, in this study, the data involved in U-Net training and evaluation contained 239 HGG and 66 LGG records.

Whenever the segmentation is performed via classification of individual pixels based on a handcrafted feature set, it is necessary to align the histograms of all records so that they contain comparable intensity values [GyA4]. However, since this study uses convolutional networks that capture the context of pixels beside its own intensity values, we decided to employ an adaptive spatial histogram equalization method that optimizes the visibility of structures based on local histograms presented in Section 3.2.1. The segmentation outcome was analyzed based on the statistical metrics exhibited in Table 3.2.

Table 5.7: Accuracy benchmarks obtained with various values of the layer depth parameter κ, without post-processing, were evaluated using HGG and LGG data.

		Layer depth parameter κ						
	DSC	16	32	48	64	80	96	112
HGG results	average	0.8568	0.8702	0.8725	0.8629	0.8707	**0.8768**	0.8688
	stdev	0.0868	0.0757	0.0769	0.0944	0.0780	0.0774	0.0798
	1st quartile	0.8255	**0.8509**	0.8464	0.8386	0.8379	0.8503	0.8419
	median	0.8809	0.8942	0.8956	0.8921	0.8942	**0.9024**	0.8910
	3rd quartile	0.9164	0.9215	0.9260	0.9240	0.9263	**0.9302**	0.9259
	ranked 1st	15	29	33	19	23	58	**62**
	ranked 2nd	20	39	26	31	52	52	19
	ranked 3rd	15	23	41	33	59	33	35
	Total	50	91	100	83	134	**143**	116
	$\text{DSC}_i > 0.93$	31	40	52	47	52	**61**	51
	$\text{DSC}_i > 0.9$	86	109	110	105	112	**122**	107
	$\text{DSC}_i > 0.85$	160	**181**	177	170	175	180	169
	$\text{DSC}_i > 0.8$	193	201	203	196	**204**	202	200
LGG results	average	0.7442	0.7795	0.7930	**0.7991**	0.7931	0.7816	0.7864
	stdev	0.2215	0.1348	0.1753	0.1605	0.1731	0.1862	0.1888
	1st quartile	0.6673	0.7373	**0.7713**	0.7706	0.7493	0.7543	0.7159
	median	0.8261	0.8096	0.8601	0.8587	0.8504	0.8518	**0.8745**
	3rd quartile	0.8873	0.8747	0.8996	**0.9047**	0.8981	0.9005	0.9006
	ranked 1st	0	5	10	11	9	5	**26**
	ranked 2nd	0	8	26	13	10	4	5
	ranked 3rd	1	1	6	11	17	22	8
	Total	1	14	**42**	35	36	32	39
	$\text{DSC}_i > 0.93$	5	2	5	5	**8**	4	5
	$\text{DSC}_i > 0.9$	13	8	16	**18**	16	17	**18**
	$\text{DSC}_i > 0.85$	30	24	35	35	34	34	**38**
	$\text{DSC}_i > 0.8$	39	38	43	44	**45**	44	43

Segmentation quality is strongly influenced by a structural parameter of the U-Net architecture, namely the number of feature channels or filters situated in the first encoder block, which also determines the number of channels in further encoder and all decoder blocks. Best results in case of HGG, as shown in the upper half of Table 5.7, were obtained with $\kappa = 96$ feature channels in the first encoding block, which is an optimal value established empirically. The best performances are confirmed at the level of mean and main quartile benchmark values, but it is also visible if we rank the performance of networks in case of individual MRI records, or if we establish, how many of the 239 HGG volumes were segmented at DSC_i values exceeding predefined thresholds (e.g. 0.8 or 0.9). For LGG, the most favorable outcomes were achieved with $\kappa = 64$, as indicated in the lower half of Table 5.7

Figure 5.8 displays the accuracy metric values obtained for individual records of HGG. The left panel of Figure 6.5 presents TPR, PPV, and DSC values for individual HGG tu-

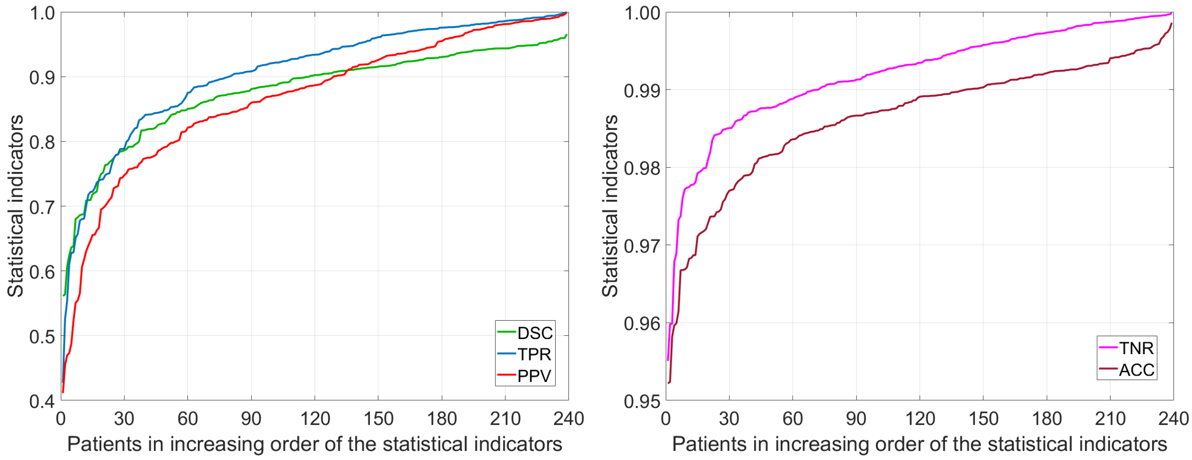

Figure 5.8: Accuracy indicator values obtained for individual records, for HGG volumes before post-processing, plotted in increasing order: sensitivity, precision and Dice similarity score values are shown in the left panel, while specificity and accuracy values in the right panel.

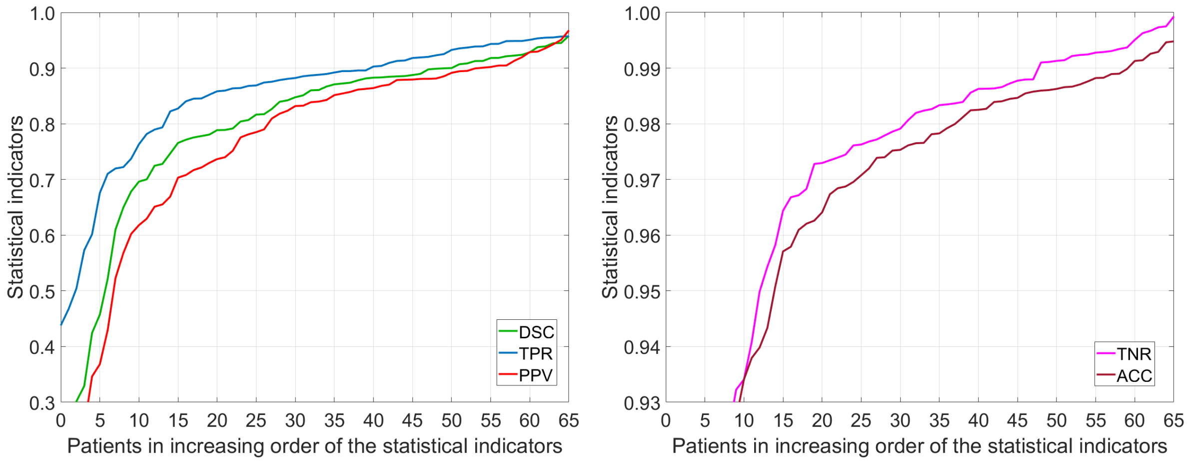

Figure 5.9: Accuracy indicator values obtained for individual records, for LGG volumes before post-processing, plotted in increasing order: sensitivity, precision and Dice similarity score values are shown in the left panel, while specificity and accuracy values in the right panel

mor volumes, organized in ascending order. Likewise, the right panel depicts the TNR and ACC values achieved for individual records. These curves reveal a distribution of accuracy indicators where the median is substantially higher than the mean, indicating the presence of a few poorly segmented records alongside the majority that were accurately segmented. Similarly, Figure 5.9 showcases the statistical indicator results for LGG records.

5.4 Conclusion

The chapter provides a comprehensive analysis of classifiers, including ensemble learning and deep learning methods and discusses the effectiveness and efficiency of these methods in the context of brain tumor segmentation using multispectral magnetic resonance image data. Additionally, it provides insights into the comparative performance of different algorithms and their suitability for specific tasks and datasets. Furthermore, the study explores the architecture of the U-Net convolutional neural network, specifically tailored to handle brain tumor segmentation on BraTS datasets. The adaptation of the 3D convolutional layers and the symmetric encoder-decoder structure is highlighted, emphasizing its potential for accurate and detailed segmentation. This detailed architecture introduces a practical approach to preserving fine details and discerning subtle features, crucial for the accurate delineation of tumor boundaries, even in cases of irregular shapes or heterogeneous appearances. This part of the thesis encapsulates the significance of these findings, positioning deep learning methods as a promising avenue for enhancing the efficiency and reliability of brain tumor analysis. The scientific formulation of the conclusion provides a cohesive summary of the Thesis II, showcasing the analytical contributions and their implications in the field of MRI data processing and segmentation.

Post-processing

Refinement techniques are vital for improving the outcomes of brain tumor segmentation. Following the initial segmentation processes (Figure 2.2, 2.3), it is typical to come across imperfections like isolated regions, noise, or irregularities in the tumor boundaries. These imperfections can introduce inaccuracies in subsequent analyzes or clinical applications. Therefore, employing specific steps after the initial segmentation is crucial for rectifying these issues and elevating the overall quality of the segmentation results.

One of the key objectives of post-processing is to achieve spatial coherence within the segmented regions. This involves smoothing out irregularities and ensuring that the segmented regions exhibit realistic anatomical shapes. Various techniques, such as morphological operations and connected component analysis, can be applied to achieve this spatial coherence.

Additionally, post-processing helps in removing small, isolated regions that may have resulted from noise or artifacts in the original imaging data. By setting a minimum threshold for the size of segmented regions, these small artifacts can be effectively eliminated, leading to a more accurate representation of the tumor.

Another critical aspect of post-processing is refining the boundaries of the segmented tumor. This is important for applications such as radiation therapy planning, where precise tumor boundaries are crucial. Techniques like active contours or level-set methods can be employed to smoothen and refine the segmented boundaries.

Overall, post-processing is essential in ensuring that the segmented regions accurately represent the underlying anatomy and pathology. It enhances the reliability of subsequent analyzes and clinical decision-making processes, making it a critical component in the pipeline of brain tumor segmentation.

6.1 Background

Before embarking on my doctoral studies, I was actively involved in a research project where we employed a posterior relabeling scheme. The input data of the post-processing step consisted in the labels provided by the random forest to all voxels in the test volume. For each voxel, the number of tumor labeled neighbours (ν^+) and the number of all neighbours (ν) were extracted, using a predefined neighbourhood. The final label of a voxel was set to tumor if and only if $\nu^+/\nu > \theta$. The ideal neighbourhood to be employed in post-processing was identified as the cubic $11 \times 11 \times 11$ sized one for LGG tumor volumes, and $9 \times 9 \times 9$ for HGG tumor volumes. The most suitable threshold value established empirically was $\theta = 1/3$ [N3].

6.2 Methods

I introduced three approaches to improve the initial segmentation results.

In the first approach, classical machine learning-based post-processing is employed to reevaluate the initial labels assigned to each pixel. This is achieved through a random forest classifier that leverages morphological features extracted from the intermediary labels produced by the ensemble step. The features include the rate of positives within concentric cubic neighbourhoods around each pixel, providing valuable contextual information for classification. By training the random forest on a substantial volume of data, the algorithm can effectively distinguish between positive and negative pixels, enhancing the segmentation results.

In the second approach, structural post-processing focuses on positively labeled pixels from the initial segmentation. This step employs a region-growing algorithm to identify contiguous spatial regions formed by positive pixels, effectively consolidating clusters of interest. Small lesions with limited cardinality are excluded from further consideration, ensuring that only substantial regions are subjected to shape-based validation. The application of Principal Component Analysis (PCA) allows for the assessment of lesion dimensions and shapes, with particular emphasis on eigenvalues derived from the analysis. This information is crucial in discriminating between likely glioma shapes and outliers. Ultimately, the structural post-processing phase refines the positive labels, providing a more accurate depiction of glioma locations within the brain.

In the third approach, U-Net based post-processing is employed to further refine the segmentation results obtained from the initial process. This technique involves the use of a second U-Net, which leverages information from the spatial neighbourhood of pixels to reevaluate their segmentation status. The approach employs two identical U-Nets, referred to as the initial segmentation network and the post-processing network. The initial segmentation network is trained on a designated dataset, generating preliminary segmentation results for both the training and testing datasets. Subsequently, neighbourhood features are extracted from these segmented datasets. The post-processing network is trained on the extracted features from the training dataset and then applied to the feature data of the testing dataset, resulting in the final segmentation output. The U-Nets share the same architecture. The input features for both networks include information derived from the MRI records, and the methodology aims to refine segmentation outcomes based on spatial context.

6.2.1 Random forest based and structural post-processing

Random forest based post-processing

The first post-processing step reevaluates the initial label received by each pixel of the test data records. The decision is made by a random forest classifier that relies on morphological features. Details of this post-processing step are listed in the following:

1. The RF is trained to separate positive and negative pixels using six features extracted from the intermediary labels produced by the ensemble. Let us consider $K = 5$ concentric cubic neighbourhoods of the current pixel, denoted by $\mathcal{N}_k$ ($k = 1 \dots K$), each having the size $(2k+1) \times (2k+1) \times (2k+1)$. Inside the neighbourhood $\mathcal{N}_k$ of the current pixel, the count of brain pixels ν_k, and the count of positive intermediary labeled brain pixels ν_k^+ is extracted. The ratio $\rho_k = \nu_k^+/\nu_k$ is called the rate of positives within neighbourhood $\mathcal{N}_k$. The feature vector has the form $(\rho_1, \rho_2, \dots \rho_K, \eta)$, where η is the normalized value of the number of complete neighbourhoods of the current pixel, determined as:
$$\eta = \frac{1}{K} \sum_{k=1}^{K} \delta\left(\nu_k, (2k+1)^3\right) \quad , \tag{6.1}$$
where
$$\delta(u, v) = \begin{cases} 1 & \text{if } u = v \\ 0 & \text{otherwise} \end{cases} \quad . \tag{6.2}$$
2. Each forest is trained using the feature vectors of 10^7 randomly selected voxels, whose feature values fulfil $\sum_{k=1}^{K} \rho_k > 0$.
3. Pixels whose features satisfy $\sum_{k=1}^{K} \rho_k = 0$ are declared negatives by default, they are not used for training the RF, and are not tested with the RF either.
4. The number of trees in the RF is set to 100, while the maximum allowed depth of the trees is eight.

The result of this post-processing step can be seen in Figure 2.2, represented by segmentation results S_1' and S_2'.

Structural post-processing

The structural post-processing handles only pixels that are labeled positive in segmentation results S_1' and S_2', consequently it has the option to approve or discard the current positive labels. As a first operation, it searches for contiguous spatial regions formed by positive pixels within the volume using a region growing algorithm. Contiguous regions of positive labeled pixels with a cardinality below 100 are discarded, because such small lesions cannot be reliably declared gliomas. Larger lesions are subject to shape based validation. For this purpose, the coordinates of all positive pixels belonging to the current contiguous region undergo a Principal Component Analysis (PCA), which establishes the three main axis determined by the three eigenvectors, and the corresponding radii represented by the square root of the three eigenvalues provided by PCA. We denote by $\lambda_1 > \lambda_2 > \lambda_3$ the three eigenvalues in decreasing order. Lesions having the third radius below a predefined threshold ($\sqrt{\lambda_3} < 2$) are discarded, as they are considered unlikely shapes for a glioma. All those detected lesions that are not discarded by the criteria presented above are finally declared gliomas, and all their pixels receive final positive labels. This is the final solution denoted by S_1'' and S_2'' in Figure 2.2.

6.2.2 U-Net based post-processing

I used a second U-Net for post-processing which reevaluates the segmentation based on information collected from the spatial neighbourhood of the pixels.

Without loss of generality, let us denote by A the subset of records used as training data and by B the testing subset. The corresponding set of labels (annotations) are denoted by L_A and L_B, respectively. Labels L_B can only be used for the statistical evaluation of the segmentation results obtained on the testing data.

The block diagram of the study is presented in Figure 2.3. The whole study involves a cascade of two identical U-Net architectures, denoted by U_1 and U_2. U_1 is trained to perform an initial segmentation, while U_2 produces a post-processing, which reevaluates the status of each pixel based on the labels given by U_1 to the neighbours of the pixel.

As a first step, U-Net U_1 is trained using the data set A and corresponding expected values L_A. Then the trained network U_1 gives a prediction for all pixels in both data sets A and B, thus providing an initial segmentation on both sets, which we denote by S_A and S_B. Four neighbourhood features are extracted for each pixel of both segmented data sets, and we denote the obtained feature data by F_A and F_B. F_A and F_B contain volumes of the same size as the initial data, but instead of pixel intensities they store the newly extracted features representing the ratio of neighbour pixels declared positive in the first stage. Neighborhood feature data set F_A is used to train the second U-Net U_2, which is finally used to give prediction on feature data set F_B, and this way we obtain the final segmentation result for the testing data set that is denoted by S_B'.

U-Nets U_1 and U_2 are identical architectures. The input presented to both U-Nets consists of four $155 \times 240 \times 240$ sized volumes, which contain four features of all pixels from an MRI record. In the first stage, the four features are the T1, T2, T1c and FLAIR intensity values, not the observed ones, but the ones after having undergone the preprocessing described in Section 5.2.2. Alternately, the second stage relies on four features which represent the ratio of positively labeled neighbours according to the prediction outcome of the first U-Net, extracted from cubic neighbourhoods of sizes $(2q+1) \times (2q+1) \times (2q+1)$, with $q = 1 \ldots 4$. Both U-Nets are trained based on a tailored Dice score loss function.

6.3 Results

The accuracy of the different post-processing methods was evaluated using the statistical indexes presented in Section 3.3.

6.3.1 Random forest based and structural post-processing

The above presented Random forest based and structural post-processing procedures 6.2.1, underwent a thorough evaluation involving all low-grade and high-grade tumor volumes of the BraTS 2015 data set. The train data size varied in four steps, using values of 100k, 200k, 500k, and 1000k. Procedure variants using no atlas and involving atlases of various spatial resolutions ($S \in \{60, 80, 100, 120\}$) were evaluated.

The morphological (Section 6.1) and the RF-based post-processing (Section 6.2.1) were evaluated in parallel so that we can formulate comparative assertions. The quality indicators exhibited in Section 3.3 were extracted for each scenario and each individual MRI record. The average and overall value for each indicator was established for global quality characterization. Table 6.1 presents the main overall accuracy indicator values obtained in case of various train data sizes and atlas resolutions. In each scenario and for each indicator, three values are given: the one obtained without post-processing, the one given by the morphological post-processing, and the one produced by the proposed RF-based post-processing. Dice Score, which is the most important accuracy indicator, shows the superiority of the random forest in all cases. The differences rise together with the train data size, from 0.1% at 10k feature vectors per BDT to 0.5% at 1000k data. The best performing atlas is the one represented by $S = 100$.

Figure 6.1 exhibits the overall Dice Scores obtained in case of various train data sizes and atlas resolutions, and gives a visual comparison for the values obtained with RF-based and morphological post-processing. Although the difference of only 0.5% may seem small at first sight, that improvement represents the elimination of 9-10% of mistaken labels produced by the morphological post-processing.

Table 6.2 presents the average ($\overline{\text{DSC}}$) values of the Dice similarity coefficients obtained for various data sets and training data sizes, at two different phases of the segmentation process. The ensemble output refers to the outcome achieved prior to any post-processing steps. The ultimate outcome comprises the outputs following both Classical machine learning-based post-processing and subsequent structural post-processing. In this case, I examined the performance of the aforementioned methods not only on the BraTS 2015 dataset but also on the 2019 dataset to assess their effectiveness. A larger training data size always led to better accuracy: the value rose by 0.5–0.8% if the training data size changed from $100k$ to $1000k$. If we consider segmentation accuracy the only important

Table 6.1: Main overall accuracy indicators without post-processing, with morphological post-processing, and with random forest based post-processing (BraTS 2015)

Train data	Atlas	Dice Score			Sensitivity			Specificity		
size	size	BEFORE	MORPH	RF-BASED	BEFORE	MORPH	RF-BASED	BEFORE	MORPH	RF-BASED
10k	No atlas	0.8024	0.8353	0.8364	0.7277	0.8311	0.8096	0.9934	0.9879	0.9903
	$S = 60$	0.8116	0.8388	0.8399	0.7382	0.8157	0.7943	0.9938	0.9901	0.9926
	$S = 80$	0.8137	0.8403	0.8414	0.7383	0.8158	0.7942	**0.9942**	0.9904	0.9928
	$S = 100$	0.8141	0.8415	0.8426	0.7387	0.8166	0.7951	**0.9942**	**0.9905**	**0.9929**
	$S = 120$	0.8112	0.8380	0.8391	0.7369	0.8143	0.7928	0.9939	0.9901	0.9926
100k	No atlas	0.8147	0.8432	0.8464	0.7466	0.8481	0.8278	0.9934	0.9875	0.9902
	$S = 60$	0.8207	0.8454	0.8484	0.7542	0.8308	0.8104	0.9936	0.9897	0.9924
	$S = 80$	0.8236	0.8468	0.8502	0.7572	0.8329	0.8126	0.9938	0.9897	0.9924
	$S = 100$	0.8235	0.8469	0.8503	0.7573	0.8330	0.8128	0.9937	0.9897	0.9924
	$S = 120$	0.8206	0.8446	0.8478	0.7541	0.8304	0.8100	0.9936	0.9896	0.9923
500k	No atlas	0.8175	0.8432	0.8475	0.7565	0.8573	0.8375	0.9928	0.9865	0.9893
	$S = 60$	0.8242	0.8476	0.8517	0.7646	0.8408	0.8210	0.9931	0.9891	0.9918
	$S = 80$	0.8270	0.8489	0.8534	0.7672	0.8422	0.8225	0.9933	0.9891	0.9920
	$S = 100$	0.8269	0.8495	0.8538	0.7667	0.8423	0.8226	0.9933	0.9892	0.9920
	$S = 120$	0.8245	0.8472	0.8514	0.7652	0.8408	0.8210	0.9931	0.9890	0.9918
1000k	No atlas	0.8194	0.8435	0.8484	0.7637	**0.8637**	**0.8443**	0.9923	0.9859	0.9888
	$S = 60$	0.8254	0.8483	0.8530	0.7705	0.8462	0.8267	0.9926	0.9886	0.9915
	$S = 80$	0.8278	0.8495	0.8547	0.7723	0.8471	0.8278	0.9929	0.9887	0.9916
	$S = 100$	**0.8279**	**0.8500**	**0.8549**	**0.7727**	0.8477	0.8284	0.9928	0.9888	0.9916
	$S = 120$	0.8272	0.8497	0.8545	0.7715	0.8468	0.8273	0.9928	0.9888	0.9917

BEFORE - output of first ensemble, MORPH = output morphological post-processing
RF-BASED - output of RF-based post-processing

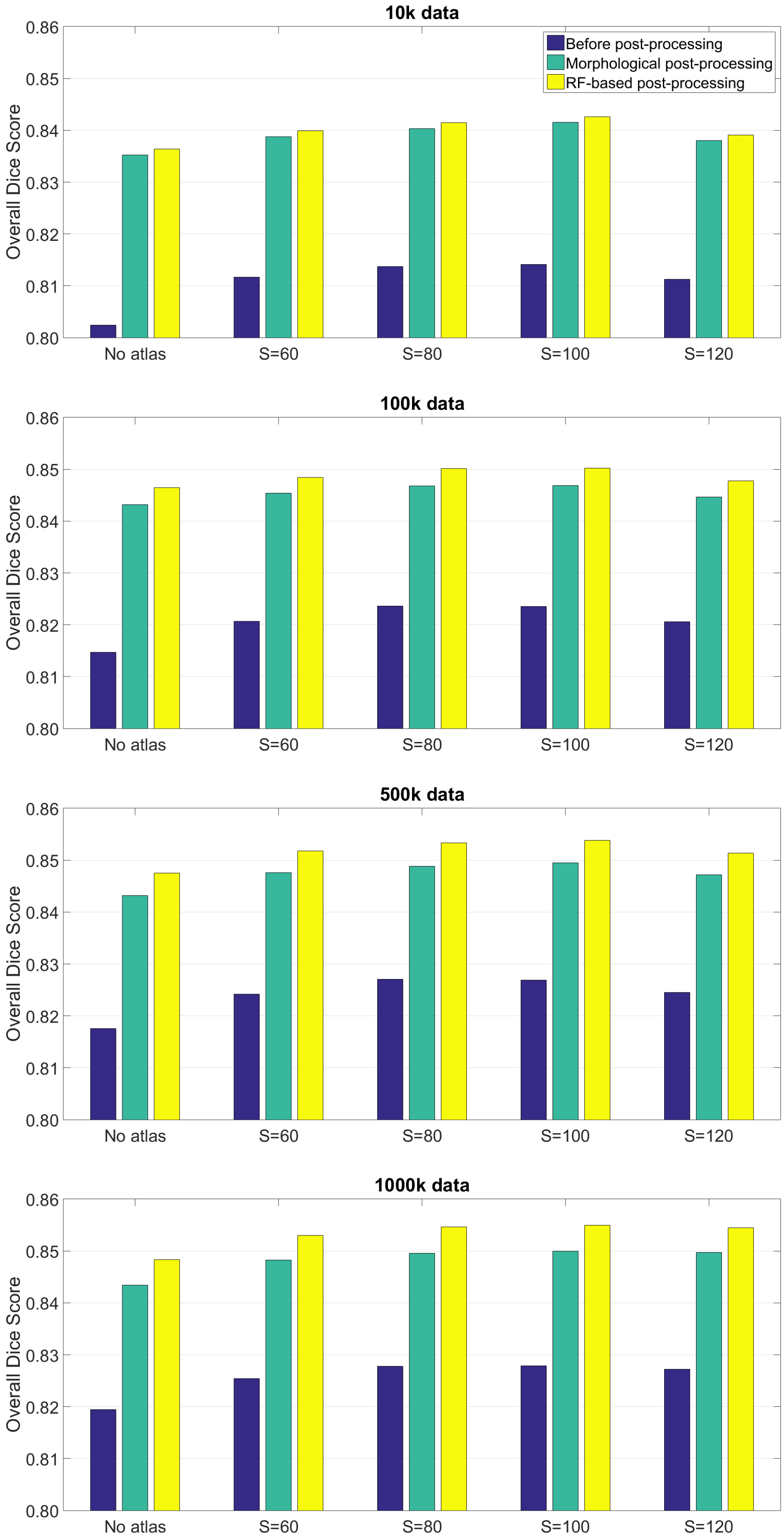

Figure 6.1: Overall Dice Score values without post-processing, with morphological post-processing, and with random forest based post-processing, for various train data sizes.

Table 6.2: Average Dice similarity coefficients obtained for various data sets and train data sizes, with and without atlas based data enhancement at preprocessing. The final results are after random forest based and structural post-processing

Segmentation result	Train data size	BraTS 2015 LGG	BraTS 2015 HGG	BraTS 2019 LGG	BraTS 2019 HGG
ensemble output without atlas (S_1)	100k	0.7904	0.7827	0.7693	0.8232
	200k	0.7928	0.7858	0.7748	0.8264
	500k	0.7961	0.7903	0.7795	0.8300
	1000k	0.7980	0.7935	0.7828	0.8324
final result without atlas (S_1'')	100k	0.8470	0.8217	0.8266	0.8421
	200k	0.8490	0.8241	0.8309	0.8436
	500k	0.8515	0.8276	0.8347	0.8452
	1000k	0.8536	0.8300	0.8378	0.8464
ensemble output with atlas (S_2)	100k	0.8040	0.7943	0.8015	0.8259
	200k	0.8063	0.7963	0.8047	0.8281
	500k	0.8089	0.7990	0.8074	0.8307
	1000k	0.8107	0.8013	0.8091	0.8326
final result with atlas (S_2'')	100k	0.8503	0.8299	0.8412	0.8464
	200k	0.8519	0.8317	0.8442	0.8485
	500k	0.8547	0.8337	0.8470	0.8504
	1000k	**0.8564**	**0.8355**	**0.8479**	**0.8516**

quality marker, then it is worth using the largest training data size.

Table 6.3: Average sensitivity values ($\overline{\text{TPR}}$) – final result with atlas (S_2'')

Train data size	BraTS 2015 data		BraTS 2019 data	
	LGG	HGG	LGG	HGG
100k	0.8155	0.7990	0.8131	0.8366
200k	0.8191	0.8016	0.8179	0.8401
500k	0.8248	0.8056	0.8234	0.8444
1000k	**0.8299**	**0.8094**	**0.8279**	**0.8473**

Table 6.4: Average positive predictive values ($\overline{\text{PPV}}$) – final result with atlas (S_2'')

Train data size	BraTS 2015 data		BraTS 2019 data	
	LGG	HGG	LGG	HGG
100k	**0.9071**	0.8935	**0.9055**	**0.8888**
200k	0.9064	**0.8936**	0.9046	0.8878
500k	0.9043	0.8923	0.9032	0.8857
1000k	0.9019	0.8909	0.8986	0.8842

Tables 6.3, 6.4, 6.5, and 6.6 exhibit the average sensitivity, positive predictive value, specificity, and correct decision rate (accuracy) values, respectively, obtained for different data sets and various train data sizes. Not all these indicators increase together with

Table 6.5: Average specificity values ($\overline{\text{TNR}}$) – final result with atlas (S_2'')

Train data size	BraTS 2015 data LGG	BraTS 2015 data HGG	BraTS 2019 data LGG	BraTS 2019 data HGG
$100k$	**0.9936**	**0.9913**	**0.9933**	**0.9923**
$200k$	0.9935	**0.9913**	0.9932	0.9922
$500k$	0.9933	0.9912	0.9931	0.9920
$1000k$	0.9931	0.9910	0.9927	0.9919

Table 6.6: Average accuracy values ($\overline{\text{ACC}}$) – final result with atlas (S_2'')

Train data size	BraTS 2015 data LGG	BraTS 2015 data HGG	BraTS 2019 data LGG	BraTS 2019 data HGG
$100k$	0.9812	0.9792	0.9796	0.9837
$200k$	0.9814	0.9794	0.9799	0.9838
$500k$	0.9816	0.9796	**0.9803**	**0.9840**
$1000k$	**0.9817**	**0.9797**	**0.9803**	**0.9840**

the train data size, but the rate of correct decisions does, showing that it is worth using larger amount of train data to achieve better segmentation quality. Sensitivity values are in the range $0.8 - 0.85$, while positive predictive values around 0.9, which indicate a fine recognition of true tumor pixels. Specificity rates are well above 0.99, which is highly important because it grants a reduced number of false positives. The average rate of correct decisions, with its values mostly above 0.98 indicates that one out of fifty or sixty pixels is misclassified by the proposed segmentation procedure. All values from Tables 6.3, 6.4, 6.5, and 6.6 reflect the evaluation of final segmentation outcomes denoted by S_2'', obtained from atlas enhanced preprocessed data.

Figure 6.2 exhibits the Dice similarity coefficients obtained for individual MRI records (DSC_i, for $i = 1 \dots n_\rho$) plotted against the true size of the tumor according to the ground truth. Each cross represents one of the MRI records, while the dashed lines indicate the linear trend of the DSC_i values identified with linear regression. As it was expected, the linear trends indicate better segmentation accuracy for larger tumors. It is also visible that for most tumors we achieve a Dice score above 0.8, and there are some records below that limit where the Dice scores can be very low. This is mostly because not all records in the BraTS data have the same image quality, some of them even contain artificially created obstacles. This distribution of the DSC_i values is also the reason why the overall Dice similarity value is higher than the average ($\widetilde{\text{DSC}} > \overline{\text{DSC}}$) in case of all four data sets. In fact, $\widetilde{\text{DSC}}$ is close to the median value of individual Dice scores DSC_i, $i = 1 \dots n_\rho$.

Figure 6.3 presents the Dice similarity scores we may expect for a $10\,cm^3$ and a $100\,cm^3$ sized tumor according to the linear trends identified from the data presented in Figure 6.2, in case of the four data sets separately. Small tumors of $10\,cm^3$ are not even present in all data sets, but the proposed method apparently learned how to segment them ac-

Table 6.7: Global accuracy indicator values were assessed using HGG data. Benchmarks improved by the post processing are highlighted in bold.

	Value	DSC	TPR	PPV	TNR	ACC
Before PP	average	0.8768	0.8662	0.9057	0.9920	0.9864
	stdev	0.0776	0.1154	0.0961	0.0073	0.0094
	1st quartile	0.8503	0.8214	0.8751	0.9888	0.9836
	median	0.9024	0.8868	0.9338	0.9935	0.9891
	3rd quartile	0.9302	0.9548	0.9757	0.9973	0.9921
After PP	average	**0.8879**	**0.8665**	**0.9251**	0.9913	**0.9870**
	stdev	0.0767	0.1108	0.0819	0.0085	0.0094
	1st quartile	**0.8612**	**0.8303**	**0.9070**	**0.9891**	0.9834
	median	**0.9116**	**0.8957**	**0.9531**	**0.9939**	**0.9901**
	3rd quartile	**0.9380**	0.9398	0.9749	0.9964	**0.9932**

Table 6.8: Global accuracy indicator values were assessed using LGG data. Benchmarks improved by the post processing are highlighted in bold.

	Value	DSC	TPR	PPV	TNR	ACC
Before PP	average	0.7991	0.9583	0.7159	0.9677	0.9669
	stdev	0.1605	0.0494	0.2017	0.0283	0.0255
	1st quartile	0.7706	0.9399	0.6562	0.9561	0.9549
	median	0.8587	0.9735	0.7757	0.9772	0.9746
	3rd quartile	0.9047	0.9937	0.8423	0.9868	0.9845
After PP	average	**0.8575**	0.9004	**0.8442**	**0.9864**	**0.9787**
	stdev	0.1062	0.0800	0.1544	0.0144	0.0144
	1st quartile	**0.8446**	0.8623	**0.8241**	**0.9817**	**0.9686**
	median	**0.8853**	0.9122	**0.8889**	**0.9924**	**0.9829**
	3rd quartile	**0.9130**	0.9620	**0.9462**	**0.9957**	**0.9896**

curately, with an expected Dice similarity value around 0.8, which reportedly represents fine segmentation [27]. The Dice score obtained for an average sized tumor from the data sets is expected in the proximity of 0.85.

Figure 6.4 exhibits some selected segmentation results. Four MRI volumetric records are selected from each of the four data sets, out of which that slice is chosen, which contains the highest number of positive pixels, ground truth positives and positive labeled pixels combined. Thus, each row in this figure exhibits one slice from a volume in five images, which represent the four observed data channels T1, T2, T1C, and FLAIR, and the segmentation outcome, respectively. In the segmentation outcome, true positives are drawn in green, false negatives in red, false positives in blue, and true negatives in grey. These segmentation results suggest that there is still place for improvement in establishing the exact boundary of the tumor, and also in suppressing the patches of false positives.

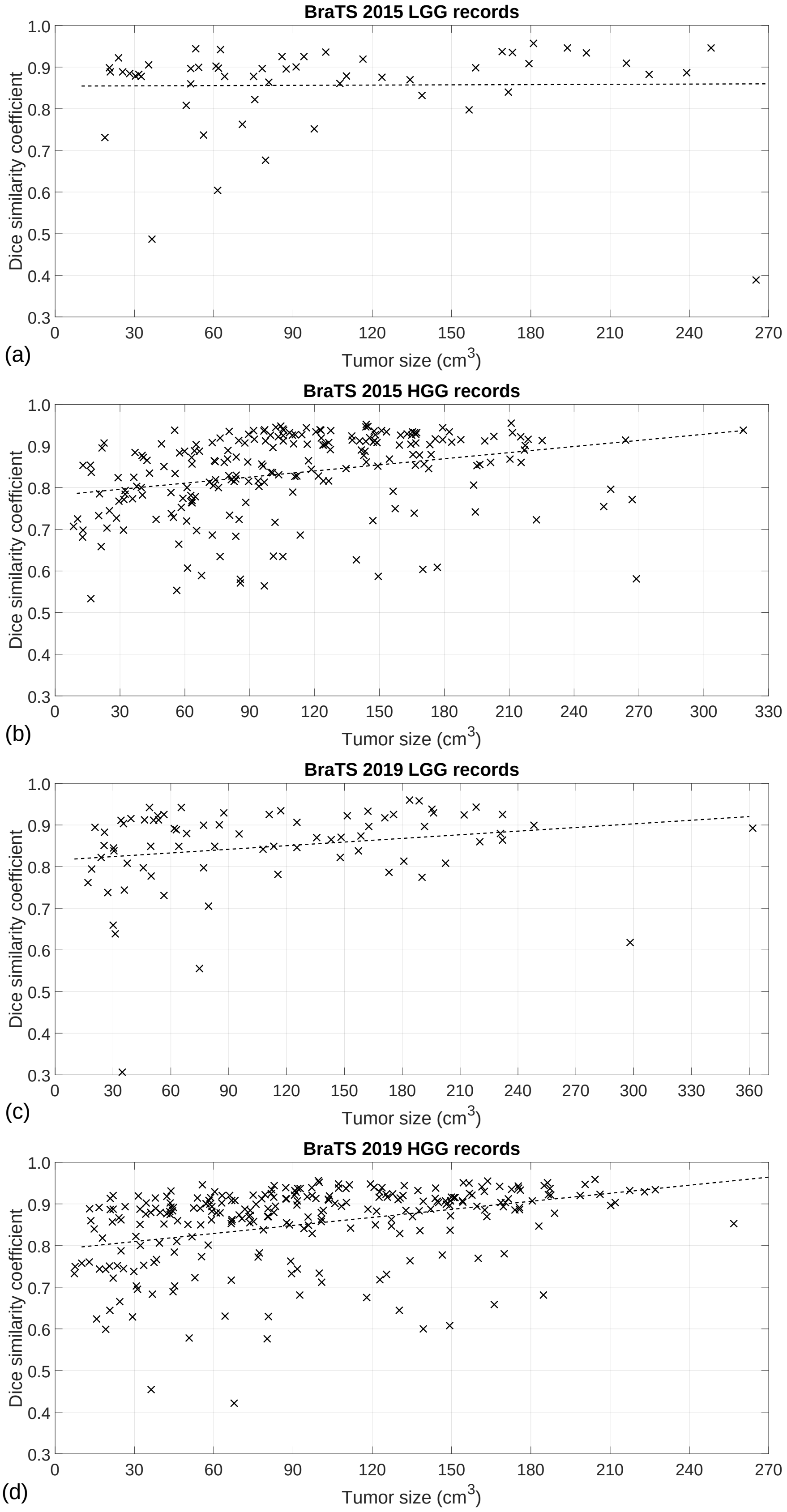

Figure 6.2: Individual Dice scores plotted against the true size of the tumor: (a) 54 records of BraTS 2015 LGG data; (b) 220 records of BraTS 2015 HGG data; (c) 76 records of BraTS 2019 LGG data; (d) 259 records of BraTS 2019 HGG data.

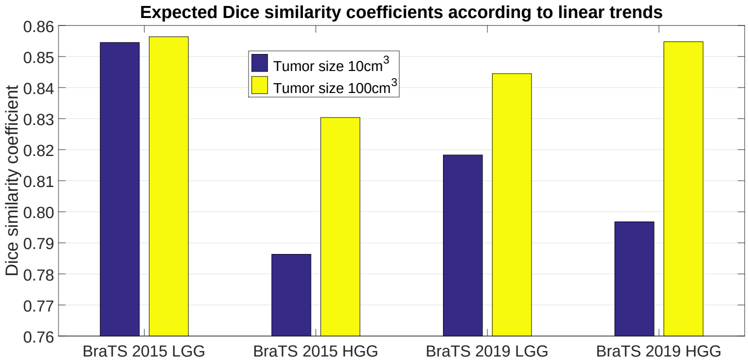

Figure 6.3: Expected Dice similarity coefficients for a $10\,cm^3$ and a $100\,cm^3$ sized tumor, in case of the four BraTS data sets, according to the identified linear trends.

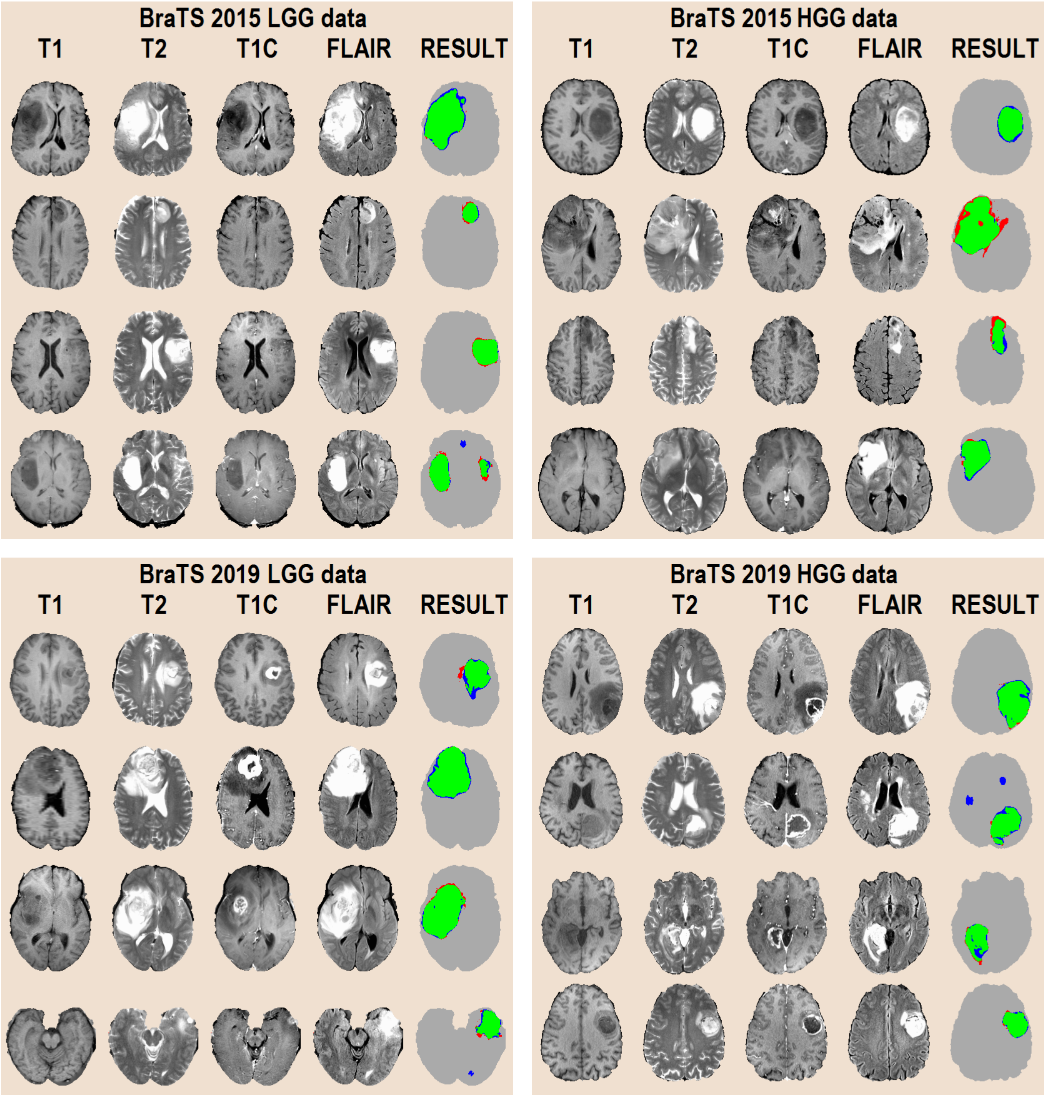

Figure 6.4: One slice from 16 different MRI records, four from each data set, showing the four observed data channels and the segmentation result. The first four columns present the T1, T2, T1C and FLAIR channel data of the chosen slices. The last column shows the segmented slice, representing true positives in green, false negatives in red, false positives in blue, and true negatives in grey, where i is the index of the current MRI record.

Table 6.9: Expected DSC values obtained for tumors of various sizes, according to the identified linear trends in case of HGG data

Tumor size	10 cm^3	20 cm^3	50 cm^3	100 cm^3
Before PP	0.8147	0.8212	0.8405	0.8728
After PP	0.8622	0.8643	0.8706	0.8811

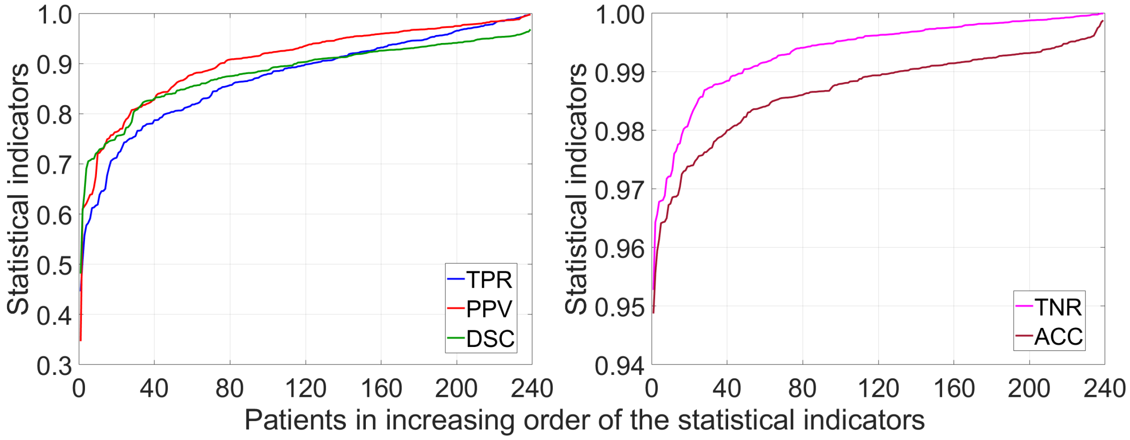

Figure 6.5: Accuracy indicator values obtained for individual records, for HGG volumes with post-processing plotted in increasing order: sensitivity, precision and Dice similarity score values are shown in the left panel, while specificity and accuracy values in the right panel.

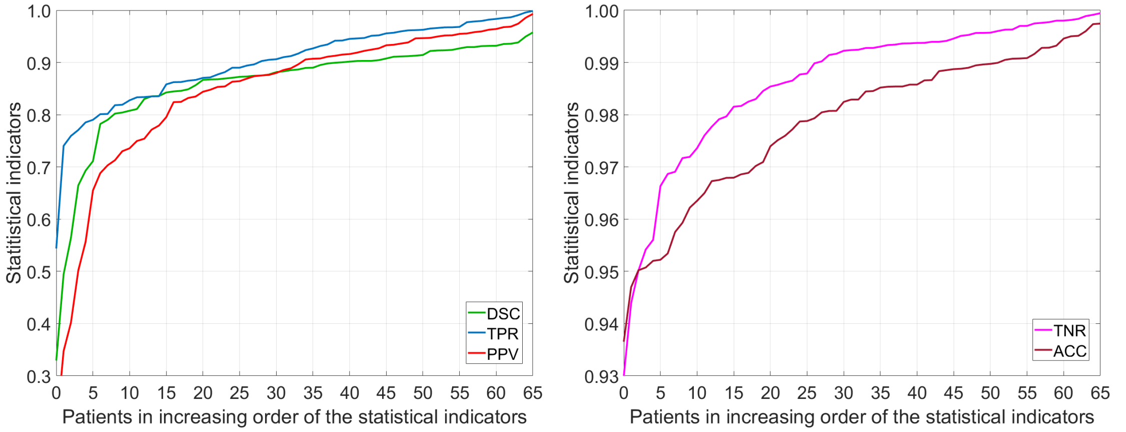

Figure 6.6: Accuracy indicator values obtained for individual records, for LGG volumes with post-processing plotted in increasing order: sensitivity, precision and Dice similarity score values are shown in the left panel, while specificity and accuracy values in the right panel.

6.3.2 U-Net based post-processing

The proposed procedure underwent a thorough evaluation process using 239 HGG volumes and 66 LGG volumes from the BraTS 2019 train data set. All records were preprocessed as described in Section 3.2.1. Both data sets were randomly separated into two equal groups, which took turns in serving as training data and testing data. This way all records were segmented by networks trained on the complementary group of records.

Table 6.7 and 6.8 provide a comprehensive overview of the achieved benchmarks for

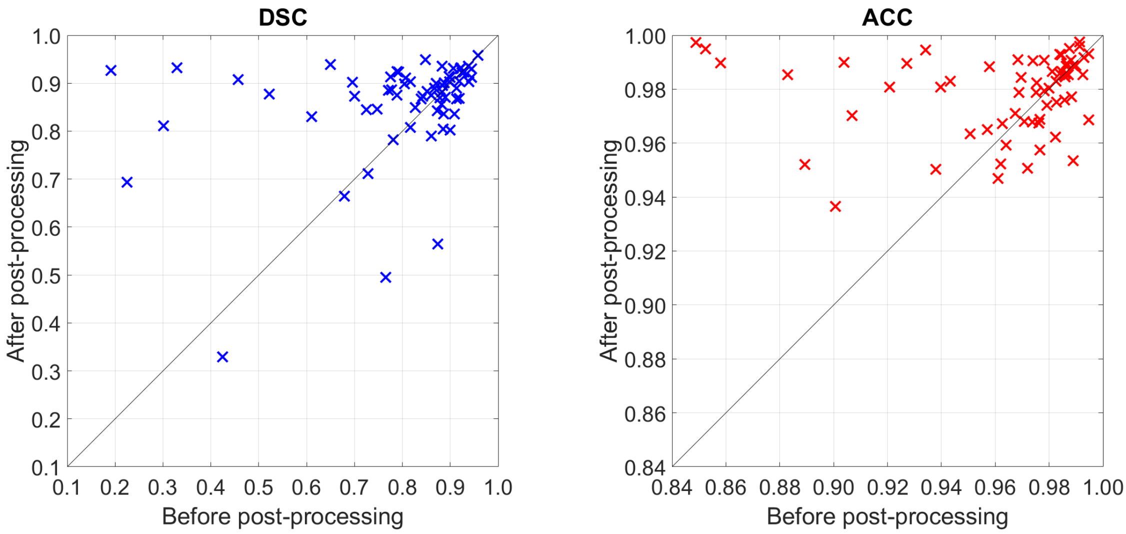

Figure 6.7: Dice scores and accuracy obtained for individual LGG tumor volumes, values obtained before post-processing plotted against values obtained after post-processing

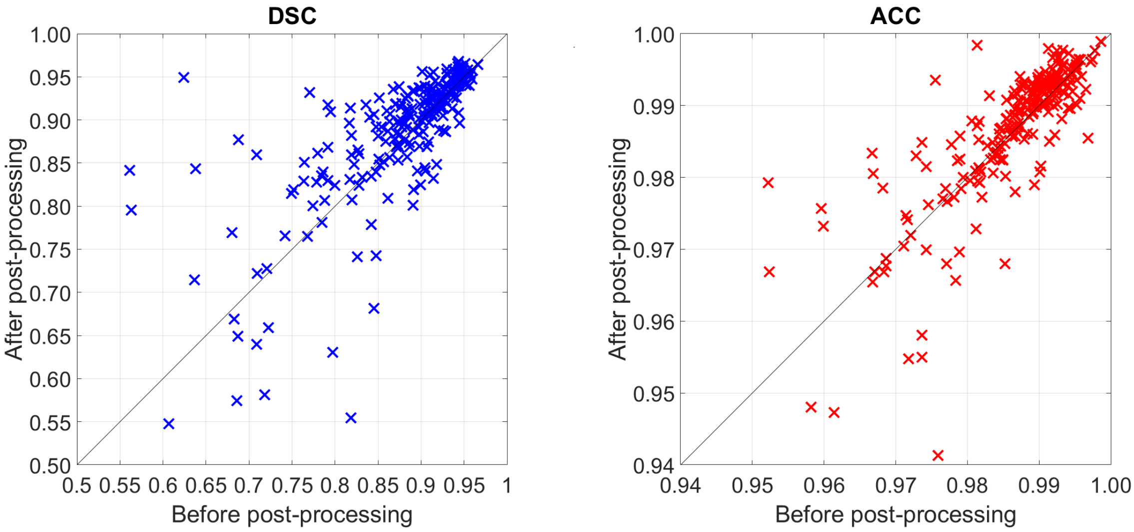

Figure 6.8: Dice scores and accuracy obtained for individual HGG tumor volumes, values obtained before post-processing plotted against values obtained after post-processing

overall accuracy indicators on LGG and HGG cases, before and after post-processing as well. These tables present a detailed breakdown of the performance metrics, offering valuable insights into the accuracy levels attained in the respective categories. The post-processing visibly improves the main indicator values, namely the DSC and ACC, which directly reflect both kinds of mistakes.

Figure 6.5 exhibits the accuracy marker values achieved in case of individual records in case of HGG. The left panel of Figure 6.5 plots TPR, PPV, and DSC values obtained for individual HGG tumor volumes, arranged in increasing order. Similarly, the right panel shows the TNR and ACC values achieved for individual records. All these curves indicate a distribution of accuracy indicator values where the median is well above the average, meaning that there are a few records with poor segmentation beside the majority of records that were accurately segmented. Similarly, Figure 6.6 plots the results of the

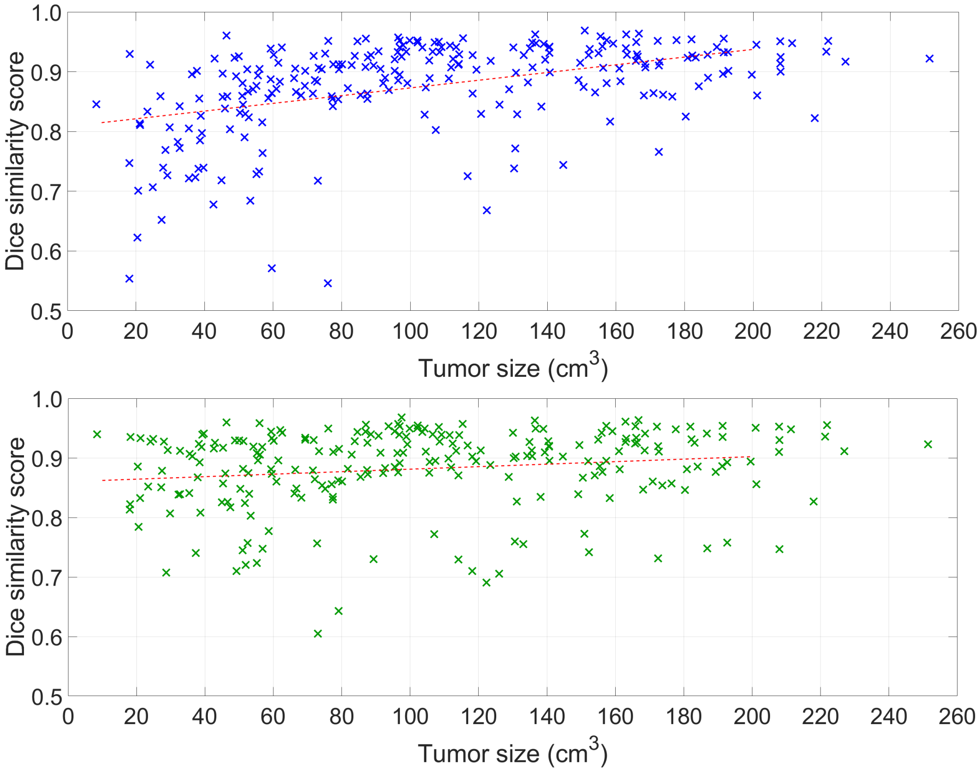

Figure 6.9: Dice similarity scores obtained for individual HGG tumor volumes before post-processing (a), and after post processing (b), plotted against the actual tumor size (according to the ground truth). The dashed red lines indicate the linear trend of Dice scores, extracted by the use of linear regression. The identified trends suggest that the segmentation outcome of a typical small sized HGG tumor is characterized by a DSC value of 0.81 before and 0.86 after post-processing.

statistical indicators for LGG.

Figure 6.7 and 6.8 exhibits the Dice scores and accuracy obtained before and after post-processing tested on each individual LGG and HGG tumor volume. Most crosses (×) on each graph are close to the diagonal, indicating that the Dice score achieved by both tests were pretty much the same, but there are also crosses far from the diagonal, showing cases when post-processing either improved or deteriorated segmentation outcomes. In the case of LGG, it is very noticeable that for smaller tumors, the post-processing improved the final result significantly.

Figure 6.9 and 6.10 exhibits the DSC values obtained for each HGG and LGG tumor record, plotted against the true tumor size. Figure 6.9(a) shows DSC values before post-processing, while Figure 6.9(b) reflects final DSC values achieved after post-processing. Linear regression was employed to identify the linear trend of Dice scores in both graphs. These trends are represented by the dashed red lines, which clearly indicate that larger tumors are likely to be segmented with higher accuracy. The linear trends of the two graphs representing LGG records and two graphs representing HGG also prove that more relevant improvement is obtained for tumors of reduced size.

Table 6.9 exhibits theoretical expected Dice scores that the proposed procedure can

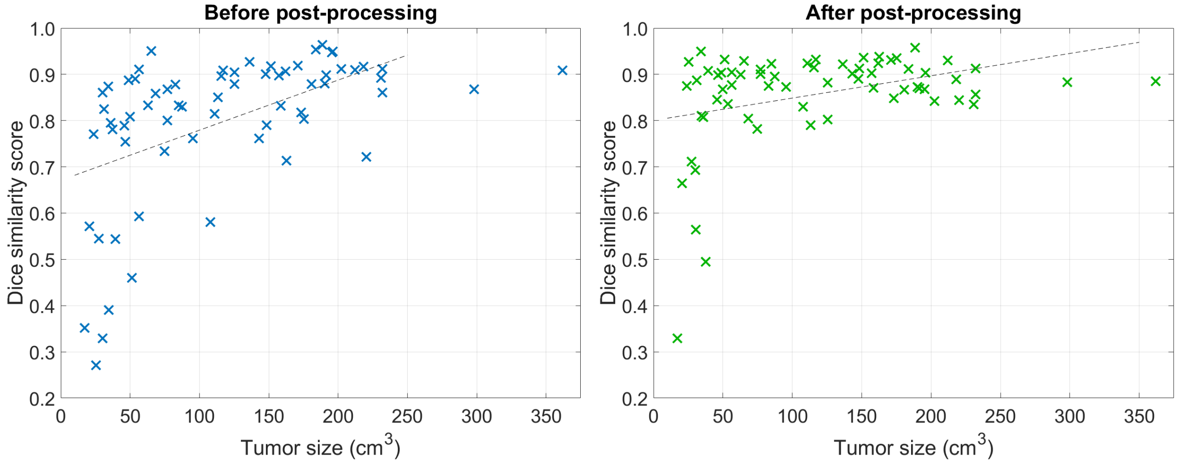

Figure 6.10: Dice scores similarity scores obtained for individual LGG tumor volumes before postprocessing and after post processing plotted against the actual tumor size (according to the ground truth).

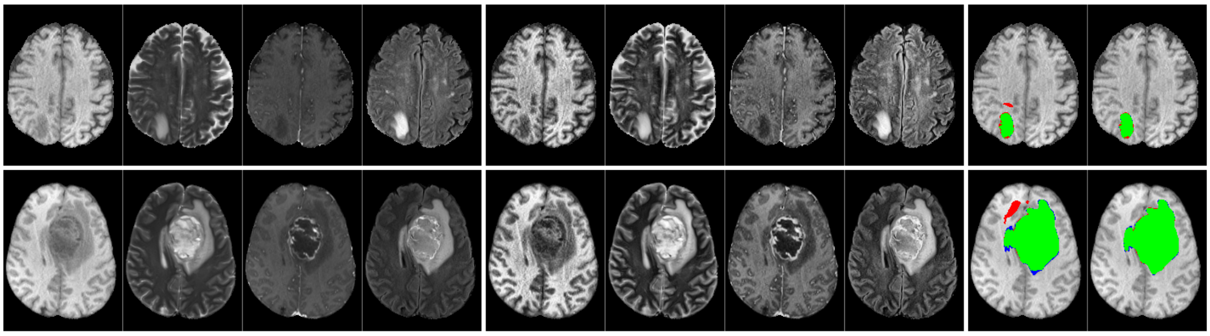

Figure 6.11: Segmentation outcome demonstrated on two selected slices: the first four images in each row present the input data (channels T1, T2, T1c and FLAIR), the next four images show the preprocessed data after CLAHE, while the last two images exhibit the segmented tumor after the first and the second U-Net stage. Color codes: green - TP, blue - FP, red - FN.

achieve in case of typical tumors of size ranging from 10 cm^3 to 100 cm^3, according to the identified linear trends presented in Figure 6.9. The effect of post-processing is stronger for small sized tumors, namely, post-processing helps more where it is necessary.

Figure 6.11 presents the segmentation outcome of two selected slices from two different HGG tumor records. The first four columns in each row represent the observed data channels T1, T2, T1c, and FLAIR of the given slice, the next four columns show the CLAHE-based (Section 3.2.1) enhanced version of the image data channels, while the last two exhibit the segmented tumor after the first stage and second stage of the U-Net cascade, respectively. Green, red, and blue pixels represent TP, FN, and FP cases, respectively.

Table 6.10 shows a collection of solutions given to the brain tumor segmentation problem that belong to the state-of-the-art, grouped into three clusters according to the BraTS data set involved in the respective studies. The average DSC is presented for each study, using the number of decimals given by their authors. Giving only two decimals

Table 6.10: Comparison with state-of-the-art methods using various versions of the BraTS data sets

Paper	Classifier	Data	Mean DSC
Tustison *et al.* [29] (2015)	RF, MRF	BraTS 2013	0.87
Pereira *et al.* [58] (2016)	CNN		0.88
Lefkovits *et al.* [97] (2017)	RF		0.868
Havaei *et al.* [120] (2017)	deep CNN		0.88
Pinto *et al.* [30] (2018)	ERT		0.85
Pereira *et al.* [121] (2019)	FCNN		0.86
Pereira *et al.* [58] (2016)	CNN	BraTS 2015	0.78
Kamnitsas *et al.* [61] (2017)	deep CNN		0.849
Zhao *et al.* [59] (2018)	FCNN, CRF		0.84
Chen *et al.* [64] (2019)	CNN		0.85
Ding *et al.* [62] (2019)	deep ResNet		0.86
Wu *et al.* [60] (2020)	CNN		0.83
Györfi *et al.* [GyA6] (2021)	BDT ensemble		0.8355
Bhalerao *et al.* [122] (2020)	3D Residual U-Net	BraTS 2019/20	0.85269
Wang *et al.* [123] (2020)	3D U-Net		**0.894**
Guo *et al.* [124] (2020)	CNN + fusion		0.872
Györfi *et al.* [GyA6] (2021)	BDT ensemble		0.8516
Lefkovits *et al.* [125] (2022)	CNN ensemble		0.8780
Proposed method	U-Net cascade		0.8879

may not be appropriate because it may introduce invisible differences of half percent.

Brain tumor segmentation algorithms perform differently on the three data sets. For example, the solution provided by Pereira *et al.* [58] achieved DSC = 0.88 on BraTS 2013 data set and DSC = 0.78 on BraTS 2015. Apparently, it is not fair to directly compare the average Dice scores from different clusters in Table 6.10. Accuracy benchmarks suggest that the proposed method can compete with the finest existing solutions.

6.4 Conclusion

This chapter encapsulates the incorporation of both classical machine learning-based and structural post-processing techniques significantly elevates the accuracy and reliability of segmentation results. The merge of Random forest based and structural post-processing adeptly addresses various imperfections that may arise in initial segmentation, including noise, irregularities, and small isolated regions. The integrated approach showcases its effectiveness in enhancing the quality of brain tumor segmentation, benefitting both clinical practitioners and researchers in the field of neuroimaging.

Leveraging information from neighbouring pixels, U-Net contributes to a more comprehensive and contextually informed evaluation of each pixel status, resulting in a refined and clinically relevant depiction of glioma locations within the brain. Additionally, the integration of U-Net-based post-processing in the segmentation pipeline proves to be a powerful refinement step, enhancing the accuracy and reliability of the final segmentation results, consequently shaping the landscape of brain tumor segmentation. Thesis group III encapsulates the impactful contributions of Random forest-based and structural post-processing techniques and deep learning post-processing techniques.

New scientific results

Thesis group I. – Results concerning the preprocessing of MRI data

This study aims to investigate which of the commonly used histogram equalization methods provides the most suitable preprocessed brain MRI data for accurate segmentation via machine learning techniques, and which parameter settings are likely to assure high quality results. Two publicly available brain MRI data sets were involved in the investigation, one without and the other with focal lesions. Based on the results obtained via experimental evaluation, I recommend using the histogram equalization method of Nyúl *et al.* [20], with milestone schemes containing 3-5 landmarks, which ensure an effective adjustment of intensity distributions for optimal segmentation quality.

In the case of feature selection solution the achieved reduced feature set can accelerate the whole processing over three times. Further works will aim at providing more general solutions and validating them in various multiple class segmentation problems.

Regarding the effect of spectral resolution upon the segmentation quality of MRI brain image data achieved by machine learning based methods, two datasets with two and four features, respectively, underwent a morphological feature generation procedure creating final feature vectors of 21 and 36 features, respectively. Three classification methods with different foundation were involved in the segmentation process. All three methods revealed that a coarse spectral resolution below 6 bits color depth damage the segmentation accuracy, while at finer spectral resolution the segmentation quality benchmarks saturate. This result supports the idea that the 16-bit unsigned integer data representation, which is widely used in medical imaging, is a waste of storage space when MRI feature values are stored. Using a single byte to represent each attribute can reduce the necessary storage space by up to 50%.

In terms of the evaluation of atlas-based data enhancement on the segmentation quality of tumor volumes, I explored various combinations of training and testing sets, as well as different sizes of atlases. The results consistently demonstrated that the introduction of an atlas led to improvements in segmentation accuracy. Additionally, the linear regression analysis revealed that the use of an atlas resulted in notable enhancements for smaller tumors, with an expected Dice score increases. This suggests that the atlas-based approach is particularly beneficial for segmenting smaller tumors.

Thesis group 1. - Results concerning the preprocessing of MRI data

Thesis 1.1
I have accomplished an investigation regarding the most suitable histogram equalization method that would support high-quality segmentation of brain MRI records, both with and without focal lesions. I have formulated new recommendations for the use of the best performing algorithm, originally published by Nyúl *et al.* [20], which yields 0.5-3.0% improvement of average Dice similarity coefficients in comparison with previously published scenarios.

Relevant own publication pertaining to this thesis: [GyA4].

Thesis 1.2
I have proposed a feature generation and selection scheme to support the segmentation of brain tumors from multi-spectral MRI records, to establish an optimal sized handcrafted feature vector that enables classical machine learning based methods to achieve high-quality segmentation. The proposed method can significantly reduce the number of necessary features, and consequently the computational load of training and testing, without any loss in terms of the segmentation quality.

Relevant own publication pertaining to this thesis: [GyA1].

Thesis 1.3
As a spin-off of the histogram equalization study, I have investigated the effect of spectral resolution used when storing the preprocessed data, upon the achievable segmentation quality. The evaluation using MRI data with and without focal lesions revealed that 6-bit spectral resolution is sufficient for classical machine learning based classifiers to obtain the best segmentation quality they are capable of. This way we may significantly reduce the necessary storage space, taking into consideration that the MRI devices produce data at 16-bit resolution.

Relevant own publications pertaining to this thesis: [GyA2], [GyA5].

Thesis 1.4

I have proposed an atlas-based data enhancement technique relying on the statistical analysis of pixels from the training dataset that belong to healthy tissues. Locally extracted means and variances are used to update the pixel intensities in all volumes of the training and testing datasets, before proceeding to the classification. The proposed method caused up to 1% improvement of average Dice scores in case of both HGG and LGG records.

Relevant own publications pertaining to this thesis: [GyA3], [GyA6].

Thesis group II. – Results concerning the classification of pixels

This thesis attempted to compare the accuracy and efficiency of various ensemble learning algorithms involved in a brain tumor segmentation based on multispectral magnetic resonance image data. The performed investigation indicates that publicly available implementations of ensemble learning methods are all capable to detect and segment the tumor with an acceptable accuracy. The small differences in terms of accuracy, and larger ones in terms of efficiency together revealed that random forest is the best decision making algorithm from the investigated ones.

In contrast, evaluating the performance of deep learning methods in brain tumor segmentation, a significant leap in accuracy and efficiency was observed. The model demonstrated a superior ability to capture complex patterns and spatial relationships within the medical images. Unlike traditional handcrafted feature-based approaches, the deep learning method showcased a high degree of adaptability and scalability.

This substantial improvement in segmentation accuracy positions deep learning methods as a promising avenue for enhancing the efficiency and reliability of brain tumor analysis.

Thesis 2. - Results concerning the classification of pixels

Thesis 2

I conducted a thorough analysis of commonly employed classifiers, including both classical machine learning and deep learning approaches, to determine their performance in relation to the brain tumor segmentation task and the dataset at hand. The findings revealed that deep learning methods operate more effectively, showcasing superior performance in the context of the designated task and dataset.

Relevant own publications pertaining to this thesis: [GyA7], [GyA8], [GyA9], [GyA10].

Thesis group III. – Results concerning the post-processing of segmentation outputs

By incorporating both classical machine learning-based and structural post-processing techniques, the segmentation results achieve a significantly higher level of accuracy and reliability. The combination of these approaches addresses various imperfections that may arise in the initial segmentation, including noise, irregularities, and small isolated regions. This refined segmentation output serves as a valuable foundation for subsequent analyzes and clinical applications, ensuring that accurate information is obtained for diagnosis and treatment planning. The integrated approach of leveraging both classical machine learning and structural post-processing demonstrates its effectiveness in enhancing the quality of brain tumor segmentation, ultimately benefiting both clinical practitioners and researchers in the field of neuroimaging.

The integration of U-Net based post-processing in the segmentation pipeline serves as a powerful refinement step, enhancing the accuracy and reliability of the final segmentation results. By leveraging information from neighbouring pixels, U-Net contributes to a more comprehensive and contextually-informed evaluation of each pixel status, resulting in a refined and clinically relevant depiction of glioma locations within the brain. This approach demonstrates its efficacy in complementing the initial segmentation process, ultimately benefiting both clinical practice and neuroimaging research.

Thesis group 3. Post-processing - Results concerning the post-processing of segmentation outputs

Thesis 3.1

I have proposed a two-stage post-processing for the classical machine learning based segmentation methods. The first stage consists of a random forest that predicts the label of pixels based on morphological features extracted from the initial decision. The second stage, relying on spatial region growing and principal component analysis, establishes the size and shape of contiguous ranges of lesions, and provides final decision based on the identified structures. The proposed method improved the final segmentation outcome of HGG and LGG records by 0.5-1.0% and 0.5-0.8%, respectively, compared to the baseline morphological method.

Relevant own publications pertaining to this thesis: [GyA11], [GyA12].

Thesis 3.2
I have developed a novel approach involving the incorporation of a secondary U-Net convolutional neural network for post-processing. This method refines the results obtained from the initial segmentation process and enhances the overall segmentation accuracy.

Relevant own publications pertaining to this thesis: [GyA8], [GyA9].

Practical Applicability of the Results

The results offer numerous potential applications in medical practice and scientific research. The developed segmentation methods can contribute to faster and more accurate diagnosis of brain tumors, reducing the workload of physicians and improving the quality of patient care. Automatic segmentation allows for tracking changes in the size and shape of the tumor during treatment, providing important information about the effectiveness of the therapy. Moreover, the new methods and algorithms developed during the research can be applied to the processing of other medical imaging data as well, contributing to the further advancement of machine learning and deep learning methods, thereby supporting research and development activities. Additionally, the developed procedures and achieved results can be utilized in education within the fields of medical imaging and machine learning, helping to expand the practical knowledge of students and to acquire modern technological skills.

Bibliography

[1] A. P. Patel, J. L. Fisher, E. Nichols, F. Abd-Allah, J. Abdella, A. Abdelalim, and et al. Global, regional, and national burden of brain and other CNS cancer, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.*, 18:376–393, 2019. `doi:10.1016/S1474-4422(18)30468-X`.

[2] N. Gordillo, E. Montseny, and P. Sobrevilla. State of the art survey on MRI brain tumor segmentation. *Magn. Reson. Imaging*, 31:1426–1438, 2013. `doi:10.1016/j.mri.2013.05.002`.

[3] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, and et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imag.*, 34:1993–2024, 2015. `doi:10.1109/TMI.2014.2377694`.

[4] S. Bakas, H. Akbari, A. Sotiras, and et al. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4:170117, 2017. `doi:10.1038/sdata.2017.117`.

[5] Y. Y. Yang, W. J. Jia, and Y. N. Yang. Multi-atlas segmentation and correction model with level set formulation for 3D brain MR images. *Pattern Recognition*, 90:450–463, 2019. `doi:10.1016/j.patcog.2019.01.031`.

[6] M. Cabezas, A. Oliver, X. Lladó, J. Freixenet, and M. Bach Cuadra. A review of atlas-based segmentation for magnetic resonance brain images. *Computer Methods and Programs in Biomedicine*, 104(3):e158–e177, 2011. `doi:10.1016/j.cmpb.2011.07.015`.

[7] S. Pereira, A. Pinto, V. Alves, and C. A. Silva. Deep Convolutional Neural Networks for the Segmentation of Gliomas in Multi-sequence MRI. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 131–143, 2016. `doi:10.1007/978-3-319-30858-6_12`.

[8] K. R. M. Fernando and Ch. P. Tsokos. Deep and statistical learning in biomedical imaging: State of the art in 3D MRI brain tumor segmentation. *Information Fusion*, 92:450–465, 2023. `doi:10.1016/j.inffus.2022.12.013`.

[9] H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, and M. Jamal Deen. Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives. *Neurocomputing*, 444:92–110, 2021. `doi:10.1016/j.neucom.2020.04.157`.

[10] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer International Publishing, 2015. `doi:10.1007/978-3-319-24574-4_28`.

[11] Cancer Stat Facts: Brain and Other Nervous System Cancer. `https://seer.cancer.gov/statfacts/html/brain.html`. Accessed: 2023-08-18.

[12] Cancer.Net - Brain Tumor: Statistics. `https://www.cancer.net/cancer-types/brain-tumor/statistics`. Accessed: 2023-08-18.

[13] D. N. Louis, A. Perry, P. Wesseling, D. J. Brat, I. A. Cree, D. Figarella-Branger, C. Hawkins, H. K. Ng, S. M. Pfister, G. Reifenberger, R. Soffietti, A. von Deimling, and D. W. Ellison. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology*, 23(8):1231–1251, 06 2021. `doi:10.1093/neuonc/noab106`.

[14] G. Mohan and M. M. Subashini. MRI based medical image analysis: Survey on brain tumor grade classification. *Biomed. Sign. Proc. Contr.*, 39:139–161, 2018. `doi:10.1016/j.bspc.2017.07.007`.

[15] Q. T. Ostrom, G. Cioffi, K. Waite, C. Kruchko, and J. S. Barnholtz-Sloan. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2014–2018. *Neuro-Oncology*, 23(3):iii1–iii105, 10 2021. `doi:10.1093/neuonc/noab200`.

[16] M. C. Mabray, R. F. Jr Barajas, and S. Cha. Modern brain tumor imaging. *Brain Tumor Res Treat*, 3, 2015. URL: `https://doi.org/10.14791/btrt.2015.3.1.8`.

[17] J. R. Fink, M. Muzi, M Peck, and K. A. Krohn. Multimodality Brain Tumor Imaging: MR Imaging, PET, and PET/MR Imaging. *Journal of Nuclear Medicine*, 56(10):1554–1561, 2015. `doi:10.2967/jnumed.113.131516`.

[18] J. Zhang, K. S. Traylor, and J. M. Mountz. PET and SPECT Imaging of Brain Tumors. *Seminars in Ultrasound, CT and MRI*, 41(6):530–540, 2020. `doi:10.1053/j.sult.2020.08.007`.

[19] U. Vovk, F. Pernuš, and B. Likar. A review of methods for correction of intensity inhomogeneity in MRI. *IEEE Trans. Med. Imag.*, 26:405–421, 2007. `doi:10.1109/TMI.2006.891486`.

[20] L. G. Nyúl, J. K. Udupa, and X Zhang. New variants of a method of MRI scale standardization. *IEEE Trans. Med. Imag.*, 19:143–150, 2000. `doi:10.1109/42.836373`.

[21] A. Sravanthi Peddinti, S. Maloji, and K. Manepalli. Evolution in diagnosis and detection of brain tumor – review. *Journal of Physics: Conference Series*, 2115(1):012039, nov 2021. `doi:10.1088/1742-6596/2115/1/012039`.

[22] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, and et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv: 1181.02629v2, 19 Mar 2019, 2019. URL: `https://arxiv.org/abs/1811.02629`.

[23] J. Sahdeva, V. Kumar, I. Gupta, N. Khandelwal, and C. K. Ahuja. A novel content-based active countour model for brain tumor segmentation. *Magn. Reson. Imaging*, 30:694–715, 2012. `doi:10.1016/j.mri.2012.01.006`.

[24] I. Njeh, L. Sallemi, I. Ben Ayed, K. Chtourou, S. Lehericy, D. Galanaud, and A. Ben Hamida. 3D multimodal MRI brain glioma tumor and edema segmentation: a graph cut distribution matching approach. *Comput. Med. Imag. Graph.*, 40:108–119, 2015. `doi:10.1016/j.compmedimag.2014.10.009`.

[25] Q. N. Li, Z. F. Gao, Q. Y. Wang, J. Xia, H. Y. Zhang, H. L. Zhang, H. F. Liu, and S. Li. Glioma segmentation with a unified algorithm in multimodal MRI images. *IEEE Access*, 6:9543–9553, 2018. `doi:10.1109/ACCESS.2018.2807698`.

[26] L. Szilágyi, L. Lefkovits, and B. Benyó. Automatic brain tumor segmentation in multispectral MRI volumes using a fuzzy *c*-means cascade algorithm. In *12th International Conference on Fuzzy Systems and Knowledge Discovery*, pages 285–291, 2015. `doi:10.1109/FSKD.2015.7381955`.

[27] A. Bal, M. Banerjee, A. Chakrabarti, and P. Sharma. MRI brain tumor segmentation and analysis using rough-fuzzy *c*-means and shape based properties. *J. King Saud Univ. - Comput. Inf. Sci.*, 2020. `doi:10.1016/j.jksuci.2018.11.001`.

[28] A. Islam, S. M. S. Reza, and K. M. Iftekharuddin. Multifractal texture estimation for detection and segmentation of brain tumors. *IEEE Trans. Biomed. Eng.*, 60:3204–3215, 2013. `doi:10.1109/TBME.2013.2271383`.

[29] N. J. Tustison, K. L. Shrinidhi, M. Wintermark, C. R. Durst, B. M. Kandel, J. C. Gee, M. C. Grossman, and B. B. Avants. Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (simplified) with ANTsR. *Neuroinform.*, 13:209–225, 2015. `doi:10.1007/s12021-014-9245-2`.

[30] A. Pinto, S. Pereira, D. Rasteiro, and C. A. Silva. Hierarchical brain tumour segmentation using extremely randomized trees. *Patt. Recogn.*, 82:105–117, 2018. `doi:10.1016/j.patcog.2018.05.006`.

[31] M. Soltaninejad, G. Yang, T. Lambrou, N. Allinson, T. L. Jones, T. R. Barrick, F. A. Howe, and X. J. Ye. Supervised learning based multimodal MRI brain tumour segmentation using texture features from supervoxels. *Comput. Meth. Prog. Biomed.*, 157:69–84, 2018. `doi:10.1016/j.cmpb.2018.01.003`.

[32] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Patt. Anal. Mach. Intell.*, 34:2274–2282, 2012. `doi:10.1109/TPAMI.2012.120`.

[33] T. Imtiaz, S. Rifat, S. A. Fattah, and K. A. Wahid. Automated brain tumor segmentation based on multi-planar superpixel level features extracted from 3D MR images. *IEEE Access*, 8:25335–25349, 2020. `doi:10.1109/ACCESS.2019.2961630`.

[34] T. Kalaiselvi, P. Kumarashankar, and P. Sriramakrishnan. Three-phase automatic brain tumor diagnosis system using patches based updated run length region growing technique. *J. Digit. Imag.*, 33:465–479, 2020. `doi:10.1007/s10278-019-00276-2`.

[35] N. Zhang, S. Ruan, S. Lebonvallet, Q. Liao, and Y. Zhu. Kernel feature selection to fuse multi-spectral MRI images for brain tumor segmentation. *Computer Vision and Image Understanding*, 115(2):256–269, 2011. `doi:10.1016/j.cviu.2010.09.007`.

[36] J. Waring, C. Lindvall, and R. Umeton. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artif. Intell. Med.*, 104:101822, 2020. `doi:10.1016/j.artmed.2020.101822`.

[37] S. B. Akers. Binary decision diagrams. *IEEE Trans. Computers*, C-27(6):509–516, 1978. `doi:10.1109/TC.1978.1675141`.

[38] S. Sengupta, S. Basak, P. Saikia, S. Paul, V. Tsalavoutis, F. Atiah, V. Ravi, and A. Peters. A review of deep learning with special emphasis on architectures, applications and recent trends. *Knowl.-Based Syst.*, 194:105596, 2020. `doi:10.1016/j.knosys.2020.105596`.

[39] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. `doi:10.1038/nature14539`.

[40] Y. Wu and Z. Yi. Automated detection of kidney abnormalities using multi-feature fusion convolutional neural networks. *Knowl.-Based Syst.*, 200:105873, 2020. `doi:10.1016/j.knosys.2020.105873`.

[41] R. R. Wildeboer, R. G. J. van Sloun, H. Wijkstra, and M. Mischi. Artificial intelligence in multiparametric prostate cancer imaging with focus on deep-learning methods. *Comput. Meth. Prog. Biomed.*, 189:105316, 2020. `doi:10.1016/j.cmpb.2020.105316`.

[42] J. Xue, S. Yan, J. H. Qu, F. Qi, C. G. Qiu, H. Y. Zhang, M. R. Chen, T. T. Liu, D. W. Li, and X. Y. Liu. Deep membrane systems for multitask segmentation in diabetic retinopathy. *Knowl.-Based Syst.*, 183:104887, 2019. `doi:10.1016/j.knosys.2019.104887`.

[43] T. Y. Tan, L. Zhang, and C. P. Lim. Adaptive melanoma diagnosis using evolving clustering, ensemble and deep neural networks. *Knowl.-Based Syst.*, 187:104807, 2020. `doi:10.1016/j.knosys.2019.06.015`.

[44] M. Y. Chung, J. Y. Li, M. K. Lee, J. J. Lee, and Y. G. Shin. Deeply self-supervised contour embedded neural network applied to liver segmentation. *Comput. Meth. Prog. Biomed.*, 192:105447, 2020. `doi:10.1016/j.cmpb.2020.105447`.

[45] X. Q Du, Y. H. Song, Y. G. Liu, Y. P. Zhang, H. Liu, B. Chen, and S. Li. An integrated deep learning framework for joint segmentation of blood pool and myocardium. *Med. Image Anal.*, 62:101685, 2020. `doi:10.1016/j.media.2020.101685`.

[46] B. Orellana, E. Monclús, P. Brunet, I. Navazo, A. Bendezú, and F. Azpiroz. A scalable approach to T2-MRI colon segmentation. *Med. Image Anal.*, 63:101697, 2020. `doi:10.1016/j.media.2020.101697`.

[47] C. L Wang, M. Oda, Y. Hayashi, Y. Yoshino, T. Yamamoto, A. F. Frangi, and K. Mori. Tensor-cut: A tensor-based graph-cut blood vessel segmentation method and its application to renal artery segmentation. *Med. Image Anal.*, 60:101623, 2020. `doi:10.1016/j.media.2019.101623`.

[48] M. Yan, J. X. Guo, W. D. Tian, and Z. Yi. Symmetric convolutional neural network for mandible segmentation. *Knowl.-Based Syst.*, 159:63–71, 2018. `doi:10.1016/j.knosys.2018.06.003`.

[49] S. Noguchi, M. Nishio, M. Yakami, K. Nakagomi, and K. Togashi. Bone segmentation on whole-body CT using convolutional neural network with novel data augmentation techniques. *Comput. Biol. Med.*, 121:103767, 2020. `doi:10.1016/j.compbiomed.2020.103767`.

[50] F. Rehman, S. I. A. Shah, M. N. Riaz, Gilani O. S., and Faiza R. A region-based deep level set formulation for vertebral bone segmentation of osteoporotic fractures. *Journal of Digital Imaging*, 33:191–203, 2020. `doi:10.1007/s10278-019-00216-0`.

[51] H. Y. Zheng, Y. F. Chen, X. D. Yue, C. Ma, X. H. Liu, P. P. Yang, and J. P. Lu. Deep pancreas segmentation with uncertain regions of shadowed sets. *Magnetic Resonance Imaging*, 68:45–52, 2020. `doi:10.1016/j.mri.2020.01.008`.

[52] S. P. Pawar and S. N. Talbar. LungSeg-Net: Lung field segmentation using generative adversarial network. *Biomedical Signal Processing and Control*, 64:102296, 2021. `doi:10.1016/j.bspc.2020.102296`.

[53] M. Kallenberg, K. Petersen, M. Nielsen, and et al. Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring. *IEEE Trans Med Imaging*, 35(5):1322–1331, 2016. `doi:doi:10.1109/TMI.2016.2532122`.

[54] S. Kiranyaz, T. Ince, and M. Gabbouj. Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks. *IEEE Trans Biomed Eng*, 63(3):664–675, 2016. `doi:10.1109/TBME.2015.2468589`.

[55] A. Gupta, S. Mishra, S. C. Sahu, and et al. Application of Convolutional Neural Networks for COVID-19 Detection in X-ray Images Using InceptionV3 and U-Net. *New Gener. Comput.*, 41:475–502, 2023. `doi:10.1007/s00354-023-00217-2`.

[56] A. Dumakude and A. E. Ezugwu. Automated COVID-19 detection with convolutional neural networks. *Sci Rep*, 13:10607, 2023. `doi:10.1038/s41598-023-37743-4`.

[57] Sz. Lefkovits, L. Lefkovits, and L. Szilágyi. Applications of Different CNN Architectures for Palm Vein Identification. In *Modeling Decisions for Artificial Intelligence*, pages 295–306, 2019. `doi:10.1007/978-3-030-26773-5_26`.

[58] S. Pereira, A. Pinto, V. Alves, and C. A. Silva. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imag.*, 35:1240–1251, 2016. `doi:10.1109/TMI.2016.2538465`.

[59] X. M. Zhao, Y. H. Wu, G. D. Song, Z. Y. Li, Y. Z. Zhang, and Y. Fan. A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. *Med. Image Anal.*, 43:98–111, 2018. `doi:10.1016/j.media.2017.10.002`.

[60] D. Y. Wu, Y. Ding, M. F. Zhang, Q. Q. Yang, and Z. G. Qin. Multi-features refinement and aggregation for medical brain segmentation. *IEEE Access*, 8:57483–57496, 2020. `doi:10.1109/ACCESS.2020.2981380`.

[61] K. Kamnitsas, C. Ledig, V. F. J. 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. *Med. Image Anal.*, 36:61–78, 2017. `doi:10.1016/j.media.2016.10.004`.

[62] Y. Ding, C. Li, Q. Q. Yang, Z. Qin, and Z. G. Qin. How to improve the deep residual network to segment multi-modal brain tumor images. *IEEE Access*, 7:152821–152831, 2019. `doi:10.1109/ACCESS.2019.2948120`.

[63] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. L. Huang. SegAN: adversarial network with multi-scale L_1 loss for medical image segmentation. *Neuroinform.*, 16:383–392, 2018. `doi:10.1007/s12021-018-9377-x`.

[64] S. C. Chen, C. X. Ding, and M. F. Liu. Dual-force convolutional neural networks for accurate brain tumor segmentation. *Patt. Recogn.*, 88:90–100, 2019. `doi:10.1016/j.patcog.2018.11.009`.

[65] Z. Q. Zhu, X. Y. He, G. Q. Qi, Y. Y. Li, B. S. Cong, and Y. Liu. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. *Information Fusion*, 91:376–387, 2023. `doi:10.1016/j.inffus.2022.10.022`.

[66] Y. Cao, W. F. Zhou, M. Zang, D. L. An, Y. Feng, and B. Yu. MBANet: A 3D convolutional neural network with multi-branch attention for brain tumor segmentation from MRI images. *Biomedical Signal Processing and Control*, 80:104296, 2023. `doi:10.1016/j.bspc.2022.104296`.

[67] Y. K. Chang, Z. Z. Zheng, Y. W. Sun, M. M. Zhao, Y. Lu, and Y. Zhang. DPAFNet: A Residual Dual-Path Attention-Fusion Convolutional Neural Network for Multimodal Brain Tumor Segmentation. *Biomedical Signal Processing and Control*, 79:104037, 2023. `doi:10.1016/j.bspc.2022.104037`.

[68] H. X. Liu, G. Q. Huo, Q. Li, X. Guan, and M. L. Tseng. Multiscale lightweight 3D segmentation algorithm with attention mechanism: Brain tumor image segmentation. *Expert Systems with Applications*, 214:119166, 2023. `doi:10.1016/j.eswa.2022.119166`.

[69] K. Hu, Q. H. Gan, Y. Zhang, S. H. Deng, F. Xiao, W. Huang, C. H. Cao, and X. P. Gao. Brain tumor segmentation using multi-cascaded convolutional neural networks and conditional random field. *IEEE Access*, 7:92615–92629, 2019. `doi:10.1109/ACCESS.2019.2927433`.

[70] S. Abut, H. Okut, and K. James Kallail. Paradigm shift from artificial neural networks (anns) to deep convolutional neural networks (dcnns) in the field of medical image processing. *Expert Systems with Applications*, page 122983, 2023. `doi:10.1016/j.eswa.2023.122983`.

[71] R. Ranjbarzadeh, A. Caputo, E. B. Tirkolaee, S. Jafarzadeh Ghoushchi, and M. Bendechache. Brain tumor segmentation of MRI images: A comprehensive review on the application of artificial intelligence tools. *Computers in Biology and Medicine*, 152:106405, 2023. `doi:10.1016/j.compbiomed.2022.106405`.

[72] L Szilágyi, D. Iclănzan, Z. Kapás, Zs. Szabó, Á. Györfi, and L. Lefkovits. Low and high grade glioma segmentation in multispectral brain MRI data. *Acta Universitatis Sapientiae – Informatica*, 10(1):110–132, 2018. `doi:10.2478/ausi-2018-0007`.

[73] L. Szilágyi, L. Lefkovits, B. Iantovics, D. Iclănzan, and B. Benyó. Automatic Brain Tumor Segmentation in Multispectral MRI Volumetric Records. In *Neural Information*

Processing, pages 174–181. Springer International Publishing, 2015. `doi:10.1007/978-3-319-26561-2_21`.

[74] Z. Kapás, L. Lefkovits, and L. Szilágyi. Automatic Detection and Segmentation of Brain Tumor Using Random Forest Approach. In *Modeling Decisions for Artificial Intelligence*, pages 301–312. Springer International Publishing, 2016. `doi:10.1007/978-3-319-45656-0_25`.

[75] Z. Kapás, L. Lefkovits, D. Iclănzan, Á. Györfi, B. L. Iantovics, Sz. Lefkovits, S. M. Szilágyi, and L. Szilágyi. Automatic Brain Tumor Segmentation in Multispectral MRI Volumes Using a Random Forest Approach. In *Image and Video Technology*, pages 137–149. Springer International Publishing, 2018. `doi:10.1007/978-3-319-75786-5_12`.

[76] N. J. Tustison, B. B. Avants, P. A. Cook, Y. J. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee. N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imag.*, 29:1310–1320, 2010. `doi:10.1109/TMI.2010.2046908`.

[77] Cancer Stat Facts: Brain and Other Nervous System Cancer. `http://www.miccai.org/special-interest-groups/challenges/miccai-registered-challenges/`. Accessed: 2023-12-10.

[78] L. Szilágyi, S. M. Szilágyi, and B. Benyó. Efficient inhomogeneity compensation using fuzzy *c*-means clustering models. *Comput. Meth. Prog. Biomed.*, 108:80–89, 2012. `doi:10.1016/j.cmpb.2012.01.005`.

[79] I. M. Pisak-Lukáts and L. Szilágyi. Markov Clustering Based Feature Selection for Brain Tumor Segmentation from Multi-Spectral MRI Records. In *2022 IEEE 20th Jubilee World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000165–000170, 2022. `doi:10.1109/SAMI54271.2022.9780855`.

[80] L. Wang, D. Nie, G. N. Li, É. Puybareau, J. Dolz, Q. Zhang, F. Wang, and et al. Benchmark on Automatic Six-Month-Old Infant Brain Segmentation Algorithms: The iSeg-2017 Challenge. *IEEE Transactions on Medical Imaging*, 38(9):2219–2230, 2019. `doi:10.1109/TMI.2019.2901712`.

[81] N. L. Weisenfeld and S. K. Wartfeld. Normalization of joint image-intensity statistics in MRI using the Kullback-Leibler divergence. In *IEEE Int. Symp. on Biomedical Imaging (ISBI, Arlington VA, USA)*, pages 101–104, 2004. `doi:10.1109/ISBI.2004.1398484`.

[82] F Jäger, Y. Deuerling-Zheng, B. Frericks, F. Wacker, and J. Hornegger. A new method for MRI intensity standardization with application to lesion detection in the

brain. In *Vision Model, Visualization (AKA GmbH, Köln, Germany)*, pages 269–276, 2006. URL: `https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=73f4342c6aff53692e311c447e26d9b9916db495`.

[83] K. K. Leung, M. J. Clarkson, J. W. Bartlett, S. Clegg, C. R. Jr. Jack, M. W. Weiner, N. C. Fox, and S Ourselin. Robust atrophy rate measurement in Alzheimer's disease using multi-site serial MRI: Tissue-specific intensity normalization and parameter selection. *Neuroimage*, 50:516–523, 2010. URL: `https://pubmed.ncbi.nlm.nih.gov/20034579/`.

[84] R. T. Shinohara, C. M. Crainiceanu, B. S. Caffo, M. I. Gaitán, and D. S. Reich. Population-wide principal component-based quantification of blood-brain-barrier dynamics in multiple sclerosis. *Neuroimage*, 57:1430–1446, 2011. `doi:10.1016/j.neuroimage.2011.05.038`.

[85] V. Stimper, S. Bauer, R. Emstorfer, B. Schölkopf, and R. P. Xian. Multidimensional contrast limited adaptive histogram equalization. *IEEE Access*, 7:165437–165447, 2019. `doi:10.1109/ACCESS.2019.2952899`.

[86] A. M. Gab Allah, A. M. Sarhan, and N. M. Elshennawy. Edge U-Net: Brain tumor segmentation using MRI based on deep U-Net model with boundary information. *Expert Systems with Applications*, 213:118833, 2023. `doi:10.1016/j.eswa.2022.118833`.

[87] S. Tankala, G. Pavani, B. Biswal, G. Siddartha, G. Sahu, N.B. Subrahmanyam, and S.S Aakash. A novel depth search based light weight CAR network for the segmentation of brain tumour from MR images. *Neuroscience Informatics*, 2(4):100105, 2022. `doi:10.1016/j.neuri.2022.100105`.

[88] J. D. Christensen. Normalization of brain magnetic resonance images using histogram even-order derivative analysis. *Magn. Reson. Imaging*, 21(7):817–820, 2003. `doi:10.1016/S0730-725X(03)00102-4`.

[89] R. Meier, U. Knecht, T. Loosli, and et al. Clinical Evaluation of a Fully-automatic Segmentation Method for Longitudinal Brain Tumor Volumetry. *Scientific Reports*, 6:23376, 2015. `doi:10.1038/srep23376`.

[90] A. Ellwaa, A. Hussein, E. AlNaggar, M. Zidan, M. Zaki, M. A. Ismail, and N. M. Ghanem. Brain Tumor Segmantation Using Random Forest Trained on Iteratively Selected Patients. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 129–137, 2016. `doi:10.1007/978-3-319-55524-9_13`.

[91] M. Rezaei, K. Harmuth, W. Gierke, T. Kellermeier, M. Fischer, H. J. Yang, and C. Meinel. Conditional Adversarial Network for Semantic Segmentation of Brain Tumor. *CoRR*, abs/1708.05227, 2017. URL: `http://arxiv.org/abs/1708.05227`.

[92] L. Fidon, W. Q. Li, L. C. García-Peraza-Herrera, J. Ekanayake, N. Kitchen, S. Ourselin, and T. Vercauteren. Scalable multimodal convolutional networks for brain tumour segmentation. *CoRR*, abs/1706.08124, 2017. `doi:10.1007/978-3-319-66179-7_33`.

[93] X. Chen, B. P. Nguyen, C. K. Chui, and S. H. Ong. An automated framework for multi-label brain tumor segmentation based on kernel sparse representation. *Acta Polytech. Hung.*, 14(1):25–43, 2017. URL: `http://acta.uni-obuda.hu//Chen_Nguyen_Chui_Ong_72.pdf`.

[94] W. Chen, B. Q. Liu, S. T. Peng, J. W. Sun, and X. Qiao. Computer-Aided Grading of Gliomas Combining Automatic Segmentation and Radiomics. *International Journal of Biomedical Imaging*, 2018(2512037):11, 2018. `doi:10.1155/2018/2512037`.

[95] Y. Wu, B. Liu, W. Wu, and et al. Grading glioma by radiomics with feature selection based on mutual information. *J Ambient Intell Human Comput*, 9:1671–1682, 2018. `doi:10.1007/s12652-018-0883-3`.

[96] J. Chang, L. M. Zhang, N. J. Gu, X. C. Zhang, M. Q. Ye, R. Z. Yin, and Q. Q. Meng. A mix-pooling CNN architecture with FCRF for brain tumor segmentation. *Journal of Visual Communication and Image Representation*, 58:316–322, 2019. `doi:10.1016/j.jvcir.2018.11.047`.

[97] L. Lefkovits, Sz. Lefkovits, and L. Szilágyi. Brain tumor segmentation with optimized random forest. In *Modeling Decisions for Artificial Intelligence*, volume 10154, pages 88–99, 2017. `doi:10.1007/978-3-319-55524-9_9`.

[98] Sz. Lefkovits, L. Szilágyi, and L. Lefkovits. Brain Tumor Segmentation and Survival Prediction Using a Cascade of Random Forests. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 334–345, 2019. `doi:10.1007/978-3-030-11726-9_30`.

[99] S. X. Bao, C. Bermudez, Y. K. Huo, P. Parvathaneni, W. Rodriguez, S. M. Resnick, P. F. D'Haese, M. McHugo, S. Heckers, B. M. Dawant, I. Lyu, and B. A. Landman. Registration-based image enhancement improves multi-atlas segmentation of the thalamic nuclei and hippocampal subfields. *Magnetic Resonance Imaging*, 59:143–152, 2019. `doi:10.1016/j.mri.2019.03.014`.

[100] J. Huo, J. Wu, J. W. Cao, and G. H. Wang. Supervoxel based method for multi-atlas segmentation of brain MR images. *NeuroImage*, 175:201–214, 2018. `doi:10.1016/j.neuroimage.2018.04.001`.

[101] H. Z. Jia, Y. Xia, Y. Song, W. D. Cai, M. Fulham, and D. D. Feng. Atlas registration and ensemble deep convolutional neural network-based prostate segmentation using

magnetic resonance imaging. *Neurocomputing*, 275:1358–1369, 2018. `doi:10.1016/j.neucom.2017.09.084`.

[102] J. H. Zhou, Z. N. Yan, G. Lasio, J. Z. Huang, B. S. Zhang, N. Sharma, K Prado, and W. D'Souza. Automated compromised right lung segmentation method using a robust atlas-based active volume model with sparse shape composition prior in CT. *Computerized Medical Imaging and Graphics*, 46:47–55, 2015. `doi:10.1016/j.compmedimag.2015.07.003`.

[103] Y. M. Niu, L. Qin, and X. C. Wang. Structured graph regularized shape prior and cross-entropy induced active contour model for myocardium segmentation in CTA images. *Neurocomputing*, 357:215–230, 2019. `doi:10.1016/j.neucom.2019.04.052`.

[104] S. J. W. Kim, S. Seo, H. S. Kim, D. Y. Kim, K. W. Kang, J. J. Min, and J. S. Lee. Multi-atlas cardiac PET segmentation. *European Journal of Medical Physics*, 58:32–39, 2019. `doi:10.1016/j.ejmp.2019.01.003`.

[105] K. Karasawa, M. Oda, T. Kitasaka, Misawa K., M. Fujiwara, C. W. Chu, G. Y. Zheng, D. Rueckert, and K. Mori. Multi-atlas pancreas segmentation: Atlas selection based on vessel structure. *Medical Image Analysis*, 39:18–28, 2017. `doi:10.1016/j.media.2017.03.006`.

[106] O. Acosta, E. Mylona, M. Le Dain, C. Voisin, T. Lizee, B. Rigaud, C. Lafond, K. Gnep, and R. De Crevoisier. Multi-atlas-based segmentation of prostatic urethra from planning CT imaging to quantify dose distribution in prostate cancer radiotherapy. *Radiotherapy and Oncology*, 125(3):492–499, 2017. `doi:10.1016/j.radonc.2017.09.015`.

[107] H. Arabi and H. Zaidi. Comparison of atlas-based techniques for whole-body bone segmentation. *Medical Image Analysis*, 36:98–112, 2017. `doi:10.1016/j.media.2016.11.003`.

[108] L. Shan, C. Zach, C. Charles, and M. Niethammer. Automatic atlas-based three-label cartilage segmentation from MR knee images. *Medical Image Analysis*, 18(7):1233–1246, 2014. `doi:10.1016/j.media.2014.05.008`.

[109] B. Oliveira, S. Queirós, P. Morais, H. R. Torres, J. Gomes-Fonseca, J. C. Fonseca, and J. L. Vilaça. A novel multi-atlas strategy with dense deformation field reconstruction for abdominal and thoracic multi-organ segmentation from computed tomography. *Medical Image Analysis*, 45:108–120, 2018. `doi:10.1016/j.media.2018.02.001`.

[110] D. C. T. Nguyen, S. Benameur, M. Mignotte, and F. Lavoie. Superpixel and multi-atlas based fusion entropic model for the segmentation of X-ray images. *Medical Image Analysis*, 48:58–74, 2018. `doi:10.1016/j.media.2018.05.006`.

[111] M. K. Sharma, M. Jas, V. Karale, A. Sadhu, and S. Mukhopadhyay. Mammogram segmentation using multi-atlas deformable registration. *Computers in Biology and Medicine*, 110:244–253, 2019. `doi:10.1016/j.compbiomed.2019.06.001`.

[112] L. Sun, L. Zhang, and D. Q. Zhang. Multi-Atlas Based Methods in Brain MR Image Segmentation. *Chinese Medical Sciences Journal*, 34(2):110–119, 2019. `doi:10.24920/003576`.

[113] L. Sun, W. Shao, M. L. Wang, D. Q. Zhang, and M. X. Liu. High-order feature learning for multi-atlas based label fusion: application to brain segmentation with MRI. *IEEE Trans. Image Proc.*, 29:2702–2713, 2019. `doi:10.1109/TIP.2019.2952079`.

[114] S. Q. Bao and A. C. S. Chung. Feature sensitive label fusion with random walker for atlas-based image segmentation. *IEEE Trans. Image Proc.*, 26:2797–2810, 2017. `doi:10.1109/TIP.2017.2691799`.

[115] Q. Zheng, Y. H. Wu, and Y. Fan. Integrating semi-supervised and supervised learning methods for label fusion in multi-atlas based image segmentation. *Front. Neuroinform.*, 12:69, 2018. `doi:10.3389/fninf.2018.00069`.

[116] B. Surányi, L. Kovács, and L. Szilágyi. Segmentation of brain tissues from infant MRI records using machine learning techniques. In *Proc. 19th IEEE World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 455–460, 2021. `doi:10.1109/SAMI50585.2021.9378653`.

[117] Sz. Csaholczi, L. Kovács, and L. Szilágyi. Automatic Segmentation of Brain Tumor Parts from MRI Data Using a Random Forest Classifier. pages 000471–000476, 2021. `doi:10.1109/SAMI50585.2021.9378618`.

[118] D. Iclănzan, R. I. Lung, Zs. L. Kucsván, B. Surányi, L. Kovács, and L. Szilágyi. The role of atlases and multi-atlases in brain tissue segmentation based on multispectral magnetic resonance image data. pages 55–60, 2021. `doi:10.1109/AFRICON51333.2021.9570952`.

[119] L Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[120] 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. *Med. Image Anal.*, 35:18–31, 2017. `doi:10.1016/j.media.2016.05.004`.

[121] S. Pereira, A. Pinto, J. Amorim, A. Ribeiro, V. Alves, and C. A. Silva. Adaptive feature recombination and recalibration for semantic segmentation with fully convolutional networks. *IEEE Trans. Med. Imag.*, 38:2914–2925, 2019. `doi:10.1109/TMI.2019.2918096`.

[122] M. Bhalerao and S. Thakur. Brain Tumor Segmentation Based on 3D Residual U-Net. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 218–225, 2020. `doi:10.1007/978-3-030-46643-5_21`.

[123] F. Wang, R. Jiang, L. Zheng, C. Meng, and B. Biswal. 3D U-Net Based Brain Tumor Segmentation and Survival Days Prediction. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 131–141, 2020. `doi:10.1007/978-3-030-46640-4_13`.

[124] X. Guo, C. Yang, T. Ma, P. Zhou, S. Lu, N. Ji, D. Li, T. Wang, and H. Lv. Brain Tumor Segmentation Based on Attention Mechanism and Multi-model Fusion. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 50–60, 2020. `doi:10.1007/978-3-030-46643-5_5`.

[125] Sz. Lefkovits, L. Lefkovits, and L. Szilágyi. HGG and LGG Brain Tumor Segmentation in Multi-Modal MRI Using Pretrained Convolutional Neural Networks of Amazon Sagemaker. *Applied Sciences*, 12(7), 2022. `doi:10.3390/app12073620`.

Own publications pertaining to theses

[GyA1] Á. Györfi, L. Kovács, and L. Szilágyi. A Feature Ranking and Selection Algorithm for Brain Tumor Segmentation in Multi-Spectral Magnetic Resonance Image Data. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 804–807, 2019. `doi:10.1109/EMBC.2019.8857794`.

[GyA2] Á. Györfi, T. Fülöp, L. Kovács, and L. Szilágyi. The Effect of Spectral Resolution Upon the Accuracy of Brain Tumor Segmentation from Multi-Spectral MRI Data. In *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 325–328, 2020. `doi:10.1109/SAMI48414.2020.9108770`.

[GyA3] T. Fülöp, Á. Györfi, B. Surányi, L. Kovács, and L. Szilágyi. Brain tumor segmentation from MRI data using ensemble learning and multi-atlas. In *18th IEEE World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 111–116, 2020. `doi:10.1109/SAMI48414.2020.9108752"`.

[GyA4] A. Kőble, Á. Györfi, Sz. Csaholczi, B. Surányi, L. Dénes-Fazakas, L. Kovács, and L. Szilágyi. Identifying the most suitable histogram normalization technique for machine learning based segmentation of multispectral brain MRI data. In *Proc. IEEE AFRICON*, pages 71–76, 2021. `doi:10.1109/AFRICON51333.2021.9570990`.

[GyA5] Á. Györfi, Sz. Csaholczi, I. M. Pisak-Lukáts, L. Dénes-Fazakas, A. Kőble, O. Shvets, Gy. Eigner, L. Kovács, and L. Szilágyi. Effect of spectral resolution on the segmentation quality of magnetic resonance imaging data. In *2022 IEEE 26th International Conference on Intelligent Engineering Systems (INES)*, pages 53–58, 2022. `doi:10.1109/INES56734.2022.9922634`.

[GyA6] Á. Györfi, L. Szilágyi, and L. Kovács. A Fully Automatic Procedure for Brain Tumor Segmentation from Multi-Spectral MRI Records Using Ensemble Learning and Atlas-Based Data Enhancement. *Applied Sciences*, 11(2):564, 2021. `doi:10.3390/app11020564`.

[GyA7] Á. Györfi, L. Kovács, and L. Szilágyi. Brain Tumor Detection and Segmentation from Magnetic Resonance Image Data Using Ensemble Learning Methods. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 909–914, 2019. `doi:10.1109/SMC.2019.8914463`.

[GyA8] Á. Györfi, L. Kovács, and L. Szilágyi. A two-stage U-net approach to brain tumor segmentation from multi-spectral MRI records. *Acta Universitatis Sapientiae – Informatica*, 14(2):223–247, 2022. `doi:10.2478/ausi-2022-0014`.

[GyA9] Á. Györfi, L. Kovács, and L. Szilágyi. Brain Tumor Segmentation from Multi-Spectral MRI Records Using a U-net Cascade Architecture. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, volume paper no. 623, pages 1317–1322, 2023. `doi:10.1109/SMC53992.2023.10394168`.

[GyA10] Á. Györfi, L. Kovács, and L. Szilágyi. Brain Tumour Segmentation from Multispectral MR Image Data Using Ensemble Learning Methods. In I. Nyström, Y. Hernández Heredia, and V. Milián Núñez, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP)*, pages 326–335. Springer International Publishing, 2019. `doi:10.1007/978-3-030-33904-3_30`.

[GyA11] Á. Györfi, Sz. Csaholczi, T. Fülöp, L. Kovács, and L. Szilágyi. Brain Tumor Segmentation from Multi-Spectral Magnetic Resonance Image Data Using an Ensemble Learning Approach. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1699–1704, 2020. `doi:10.1109/SMC42975.2020.9282987`.

[GyA12] T. Fülöp, Á. Györfi, Sz. Csaholczi, L. Kovács, and L. Szilágyi. Brain Tumor Segmentation from Multi-Spectral MRI Data Using Cascaded Ensemble Learning. In *2020 IEEE 15th International Conference of System of Systems Engineering (SoSE)*, pages 531–536, 2020. `doi:10.1109/SoSE50414.2020.9130550`.

Own publications not pertaining to theses

[N1] Z. Kapás, L. Lefkovits, D. Iclănzan, Á. Györfi, B. L. Iantovics, Sz. Lefkovits, S. M. Szilágyi, and L. Szilágyi. Automatic brain tumor segmentation in multispectral mri volumes using a random forest approach. In P. Manoranjan, C. Hitoshi, and Q.M. Huang, editors, *Image and Video Technology*, pages 137–149. Springer International Publishing, 2018. `doi:10.1007/978-3-319-75786-5_12`.

[N2] Zs. Szabó, Z. Kapás, L. Lefkovits, Á. Györfi, S. M. Szilágyi, and L. Szilágyi. Automatic segmentation of low-grade brain tumor using a random forest classifier and gabor features. In *14th International Conference on Neural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 1106–1113. IEEE, 2018.

[N3] L. Szilágyi, D. Iclănzan, Z. Kapás, Zs. Szabó, Á. Györfi, and L. Lefkovits. Low and high grade glioma segmentation in multispectral brain MRI data. *Acta Universitatis Sapientiae – Informatica*, 10(1):110–132, 2018. `doi:10.2478/ausi-2018-0007`.

[N4] Á. Györfi, Z. Karetka-Mezei, D. Iclănzan, L. Kovács, and L. Szilágyi. A Study on Histogram Normalization for Brain Tumour Segmentation from Multispectral MR Image Data. In I. Nyström, Y. Hernández Heredia, and V. Milián Núñez, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP)*, pages 375–384, 2019. `doi:10.1007/978-3-030-33904-3_35`.