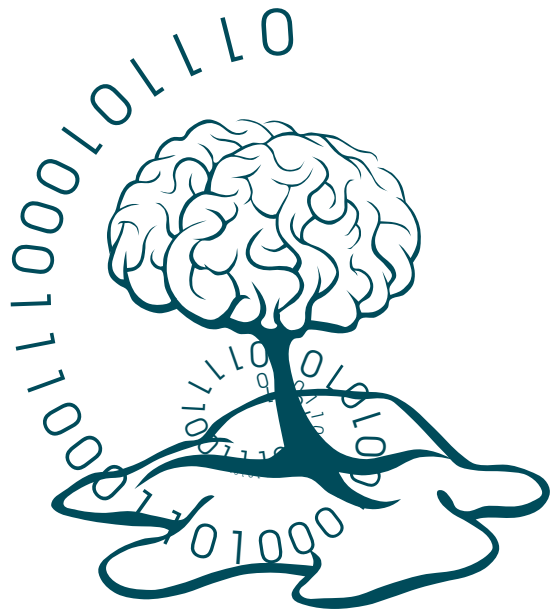
From the Institute of Medical Informatics
of the University of Lübeck
Director: Prof. Dr. rer. nat. habil. Heinz Handels

Decision Forest Variants for Brain Lesion Segmentation



Dissertation
for Fulfillment of
Requirements
for the Doctoral Degree
of the University of Lübeck
from the Department of Computer Sciences

Submitted by
Oskar Maier
from Berlin, Germany

Lübeck 2016

UNIVERSITÄT ZU LÜBECK
INSTITUTE OF MEDICAL INFORMATICS

First referee: Prof. Dr. rer. nat. habil. Heinz Handels

Second referee: Prof. Dr. ing. habil. Alfred Mertins

Date of oral examination: 04/28/2017

Approved for printing. Lübeck, 10/05/2017

Abstract

In this thesis, a cutting-edge framework for brain lesion segmentation in 3D multi-spectral magnetic resonance imaging (MRI) volumes is developed, evaluated, and discussed. It can be readily employed in the diagnosis, treatment, and research of numerous brain lesion causing diseases.

The detection and assessment of brain lesions is usually conducted qualitatively on images of the brain. These are mostly 3D volumes acquired with MRI, the imaging modality that currently offers the most sensitive way to image the human brain in-vivo.

Quantitative values, such as the lesion volume or lesion count, could be used to improve the existing treatments. One expected effect is an increase in reliability through a reduction of subjectivity. By introducing reproducibility, treatment decisions could be reviewed retrospectively. Moreover, the additional data can be employed in expert systems for decision making and safeguarding.

The required quantitative values are computed from exact demarcations of the lesions in 3D MRI sequences. However, manual lesion segmentation is a tedious and time-consuming procedure with a high inter-rater variability, which effectively prevents its use in most clinical and research scenarios. Urgently needed are methods that automatically compute a 3D lesion segmentation based on one or more 3D MRI sequences.

Such automated solutions face a number of challenges. First, clinical MRI sequences have non-standardized gray-values, differ greatly in resolution, and are often corrupted by imaging and movement artifacts. Second, brain lesions can be highly variable in size, locality, and shape. They are furthermore inhomogeneous in appearance, can mimic healthy brain tissue, and change over time. Finally, further difficulties must be addressed, such as a limited training set size and confounding benign lesions.

This thesis proposes a machine learning framework for the automatic segmentation of brain lesions in 3D multi-spectral MRI volumes. The main diseases targeted are cerebral stroke, multiple sclerosis (MS), and gliomas. A careful study of the MRI modality motivates the development of an automatic, state-of-the-art preprocessing pipeline to harmonize the input sequences. From these, a number of image features are extracted that are carefully crafted to account for the brain's anatomy and the diseases' idiosyncrasies. Last, a random forest (RF) based, voxel-wise classifier is extended and optimized to create the lesion segmentation.

Building on this framework, two extensions to the RF model are proposed. The first is local problem forests, which use spectral clustering to address some of the weak points observed in the base model. The second is semi-supervised forests, which are introduced to address the urgent problem of missing longitudinal consistency in MS lesion segmentation.

In a number of experiments the components of the framework and their relative contributions to the overall results are systematically evaluated. Additionally, the impact of each MRI sequence and each proposed feature is carefully assessed and recommendations are derived from the results. Both proposed RF extensions are shown to significantly improve the standard forest's segmentation accuracy.

The presented framework was submitted to six state-of-the-art brain lesion segmentation challenges that were organized in conjunction with the Medical Image Computing and Computer Assisted Intervention (MICCAI) conferences 2015 and 2016 as well as the International Symposium on Biomedical Imaging (ISBI) 2015. In these benchmarks it ranked among the top positions in the leaderboards and won five awards, demonstrating its cutting-edge performance.

Zusammenfassung

In dieser Dissertation wird ein neuartiges Framework zur automatischen Segmentierung von Gehirnläsionen in multispektralen 3D-Magnetresonanztomographieaufnahmen entwickelt, evaluiert und diskutiert. Die ausgearbeiteten Methoden können direkt für die Diagnose, Behandlung und Erforschung läsionsbildender Krankheiten im Gehirn eingesetzt werden.

Die Feststellung und Beurteilung von Gehirnläsionen erfolgt üblicherweise qualitativ auf Grundlage von 3D-Bilddaten des Gehirns. Bevorzugt eingesetzt wird dabei die Magnetresonanztomographie (MRT), ein weit verbreitetes, nicht invasives Bildgebungsverfahren mit exzellentem Weichteilkontrast.

Quantitative Werte, wie z.B. das Läsionsvolumen oder die Läsionsanzahl, können zur Unterstützung der Diagnostik und Verbesserung der Behandlung genutzt werden: Dadurch zu erwarten ist eine weniger subjektive Bewertung und eine entsprechende Verbesserung der Zuverlässigkeit. Durch die Einführung von Reproduzierbarkeit können Behandlungsentscheidungen nachträglich besser nachvollzogen und beurteilt werden. Weiterhin können die gewonnenen Daten zur Entscheidungsfindung und zur Qualitätskontrolle mit Expertensystemen eingesetzt werden.

Die dafür benötigten quantitativen Werte können direkt aus 3D-Segmentierungen der Läsionen berechnet werden. Deren manuelle Erstellung ist allerdings sehr aufwendig, zeitintensiv und weist eine so geringe Interrater-Reliabilität auf, dass dieser Ansatz für die meisten Anwendungen nicht praktikabel ist. Dringend benötigt werden entsprechend automatisierte Methoden, welche die erforderlichen Segmentierungen direkt aus den 3D-MRT-Aufnahmen erstellen.

Doch die Entwicklung solcher automatisierter Verfahren stellt aus verschiedenen Gründen eine Herausforderung dar. Klinische MRT-Bilder folgen weder einer einheitlichen Grauwertskala, noch weisen sie eine einheitliche Auflösung auf. Sie sind häufig von Bildgebungs- und Bewegungsartefakten beeinträchtigt und ihr Erscheinungsbild variiert in Abhängigkeit vom genutzten Gerät. Die gesuchten Läsionen unterscheiden sich beträchtlich in Größe, Form und in ihrer Position im Gehirn. Ihre Erscheinung ist oft inhomogen und kann gesundem Gehirngewebe oder gutartigen Läsionen gleichen. Schließlich müssen noch weitere Schwierigkeiten überwunden werden, wie beispielsweise die eingeschränkte Verfügbarkeit von Trainingsdaten.

In dieser Dissertation wird ein auf Maschinellem Lernen basierendes Framework für die automatische Segmentierung von Gehirnläsionen in multispektralen 3D-MRT-Aufnahmen vorgestellt. Der Fokus liegt dabei auf drei Krankheitsbildern, die durch typische Läsionen gekennzeichnet sind: multiple Sklerose (MS), Schlaganfälle und Gliome. Basierend auf einer sorgfältigen Analyse der Bilddaten und ihrer Eigenschaften wird dabei zunächst eine Vorverarbeitungspipeline auf dem neuesten Stand der Wissenschaft entwickelt, mit der viele der angesprochenen Schwierigkeiten überwunden und vereinheitlichte Bilder erzeugt werden können. Aus diesen multispektralen 3D-Aufnahmen werden daraufhin spezielle Bildmerkmale extrahiert, die in enger Kooperation mit medizinischen Experten entwickelt wurden und sowohl die Anatomie des Gehirns als auch die speziellen Charakteristika von Gehirnläsionen berücksichtigen. Schließlich wird ein grundlegendes Random-Forest-Modell zur voxelweisen Klassifikation erweitert und optimiert, dessen Ausgabe direkt in die gesuchte Segmentierung der Läsionen überführt werden kann.

Aufbauend auf diesem Grundsystem werden anschließend zwei methodische Erweiterungen der Random-Forest-Methode vorgeschlagen: Zum einen die Anwendung von Local Problem Forests, die spektrale Clusteranalyseverfahren nutzen, um gezielt verbleibende Schwachstellen auszugleichen. Zum anderen der Einsatz von Semi-Supervised Forests, die entwickelt wurden, um die longitudinale Konsistenz bei der MS-Läsionssegmentierung zu verbessern, eine der Grundvoraussetzungen für eine Verwendung der Segmentierungen als Surrogatmarker für den Krankheitsverlauf.

In Experimenten wurden die einzelnen Komponenten des Frameworks systematisch evaluiert und ihr relativer Beitrag zum Gesamtergebnis ermittelt. Weiterhin wurden die Einflüsse jeder einzelnen MRT-Sequenz sowie jedes vorgeschlagenen Bildmerkmals analyisiert und auf dieser Grundlage Empfehlungen ausgesprochen. Außerdem wurde gezeigt, dass die vorgeschlagenen Erweiterungen der Random-Forest-Methode das Segmentierungsergebnis beide noch einmal signifikant verbessern.

Schließlich wurde das entwickelte Verfahren bei sechs aktuellen, internationalen Läsionssegmentierungs-Wettbewerben eingereicht, die im Rahmen der Medical Image Computing and Computer Assisted Intervention (MICCAI) Konferenzen 2015 und 2016 sowie des International Symposium on Biomedical Imaging (ISBI) 2015 organisiert wurden. Dort erreichte die ausgearbeitete Methode jeweils hohe Platzierungen und wurde insgesamt fünfmal für ihre gute Segmentierungsleistung ausgezeichnet.

Contents

Chapter 1

Introduction

Magnetic resonance imaging (MRI) is a highly versatile imaging technique used to visualize the anatomy and physiological processes of the body based on nuclear spin echo. Since its introduction towards the end of the twentieth century, the possibility to non-invasively acquire images with excellent soft tissue contrast has substantially advanced health care and biomedical research. Our understanding of the brain in general and neural disorders in particular has subsequently benefited greatly and has lead to the establishment of *neuroimaging* as a separate discipline. Visualizing the pathological changes occurring in cerebral tissue has a multitude of uses: The clinical pictures of diseases are continuously updated and refined, which in turn facilitates the identification of potential leverage points for biochemical substances. The effects of newly developed drugs on an illness's symptoms and causes can be evaluated in-vivo. MRI scans support the clinical diagnosis, rendering it faster, more sensitive and more specific. The ability to regularly monitor the brain for tissue changes results in improved disease management and earlier responses. Surgical interventions can be carefully planed beforehand and executed with higher precision, effectively increasing the success rate. In the field of cognitive neuroscience, the function of the human brain is investigated by relating neurological deficits to damaged regions of the brain as depicted in MRI scans.

In all of these contexts, the foremost prerequisite is the quantification of the territorial damage to the brain, commonly referred to as *brain lesion*. Only then, reliable diagnosis, standardized assessment, and reproducible findings can be achieved. The exact delineation of said lesions is termed *lesion segmentation*. The manual creation of lesion masks is a tedious, time consuming and error prone process with marked inter- and intra-observer differences. The robustness of derived treatment decisions and statistical tests is accordingly limited. Methods for automatic brain lesion segmentation strive to overcome these shortcomings by providing accurate and, above all, reproducible segmentations. Furthermore, the required manpower is significantly reduced, enabling larger subject groups and thus increasing the statistical power of studies.

Automatic image segmentation rests on the principal idea of using image based clues, which can be optionally enriched by constraints derived from knowledge about the pathology at hand, to accurately detect and outline lesions in brain images. For this purpose, the discipline draws on methods from medical image analysis, computer vision, and machine learning, among others. The input to such algorithms usually consists of one or multiple 3D MRI sequences of a patient, the output of a fitting binary mask denoting the detected lesions.

Brain lesion is a general term denoting a pathological territorial alteration of brain tissue. Accordingly, the underlying causes are manifold, including demyelination, bleeding, cancer and necrosis. The exact challenges faced in lesion segmentation vary from disease to disease, each

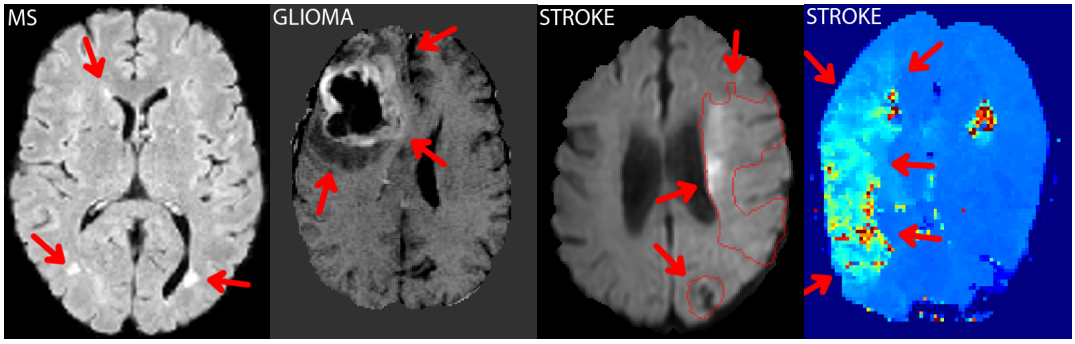


Figure 1.1: Four examples of brain lesions of varying shape, appearance, location and homogeneity. From left to right: Multiple Sclerosis (MS), glioma, sub-acute stroke and acute stroke in different MRI sequences.

producing lesions with different characteristics in terms of shape, appearance, location, homogeneity and count (see See Fig. 1.1). Furthermore, the lesions caused by a single disease can differ greatly depending on severity, age, variant and phenotype. Finally, the MRI sequences themselves are not standardized and their appearance depends on the scanner, exact acquisition parameters and other factors. Thus, automatic segmentation of brain lesions in 3D MRI images is challenging. Nevertheless, the high relevance and significant impact on numerous domains has given rise to a steady stream of publications and led recently to the organization of multiple public brain lesion segmentation benchmarks [Menze *et al.*, 2015a; Carass *et al.*, 2016; Maier *et al.*, 2017].

1.1 Objectives

On a general disease independent level, the task of brain lesion segmentation can be described as 'the detection of abnormal tissue in images of the human brain'. Most published methods target a single type of lesion, drawing to a greater or lesser extend on domain specific knowledge of the disease's characteristics.

In contrast, the first goal of this work is the development of a **general brain lesion segmentation method**. Any differentiation between pathological and healthy tissue must therefore be based exclusively on disease unspecific notions of abnormality. Only generally shared characteristics, such as boundedness, and knowledge related to the brain's anatomy can be utilized. However, due to the marked differences between lesion types, competitive results cannot be expected without at least partial adaptation to the disease at hand. To address this predicament, machine learning is employed. Based on the discriminative decision forest (DF) model for classification, a framework is designed that can be readily trained on data of new diseases with none but the slightest of adaptations. The evaluation is accordingly conducted for multiple pathologies against other state-of-the-art applications in fair and independent benchmarks on multi-spectral 3D MRI lesion segmentation . From the obtained results, the extent of similarity between the different segmentation problems can be estimated. And the developed method can serve as a starting point for pathology specific improvements. The main contributions to the realization of the first goal are a state-of-the-art preprocessing pipeline, a number of features specifically targeting brain lesions without constraint to a single disease, and a novel sub-sampling schema. All were designed in close cooperation with experienced physicians.

The second goal of this work is to extend the basic DF method employed in the first part by specifically targeting the observed limitations and making use of additional information such as unlabeled training samples. To this end, new algorithms are proposed and thoroughly evaluated, while still operating inside in the borders of the general purpose paradigm.

The first contribution to the second goal are **local problem forests**, special purpose classifiers, which target a range of sub-problems at which the classical DFs fail. The idea is to initially separate the training samples into overlapping clusters suitably representing the identified sub-problems. Then, a dedicated forest is trained on each cluster's samples. Since the overall lesion segmentation problem is constituted from all of these sub-problems, the overlapping forests act as a type of guided ensemble classifier voting on the class membership of a new sample.

The second contribution is the development of a **semi-supervised forest variant**. Semi-supervised classification aims to improve classification accuracy by including the unlabeled test samples in the training process. The underlying assumption is that the additional samples allow for a better estimate of the real density distribution in feature space and that the decision boundaries can subsequently be placed more accurately in the areas of low density between the sample clusters. To this end, the forests' node optimization term is extended by an unsupervised term penalizing splits that would pull apart dense clusters. The contributions include a memory friendly, accurate label propagation approach operating on the learned density surface, and a method for fast, iterative updates of large sample co-variance matrices.

1.2 Structure and content

This thesis is structured such that the main methodological chapters 4, 5 and 6 can be treated to some extend as stand-alone segments with individual introductions, literature overviews and conclusions. Readers experienced in the field can therefore concentrate on their subject of interest. However, for a complete picture including all premises, a stringent argumentation and a full comprehension of the implications, the thesis should be treated as a monograph and read from the beginning to the end. Pursuing the above presented objectives, this work is organized as follows.

In **Chapter 2**, the principles of MRI are reviewed in the context of brain lesion segmentation. In particular, the characteristics of various sequences and image artifacts are discussed. The chapter concludes with a summary of the main challenges arising from the modality.

In **Chapter 3**, the medical background is presented with the focus on MS and cerebral stroke as the two main diseases treated in this work. Each part commences with an introduction to the treatment and research scenarios. From these the lesion segmentation applications and their requirements are derived. In the last part, the lesions' appearance in the various MRI sequences, the extend of their variability, and the emerging challenges are discussed.

In **Chapter 4**, a general brain lesion segmentation framework is introduced which is based on voxel-wise classification with DFs. Drawing from the findings of a literature overview and the foundation laid out in the previous chapters, specialized image features for brain lesions and a dedicated training set sampling scheme are derived. The framework is thoroughly evaluated and compared against other state-of-the-art application by means of participation in four public brain lesion segmentation benchmarks, each of which targets a different pathology. A discussion closes the chapter.

In **Chapter 5**, a methodological extension of the framework is proposed that targets a set of sub-problems identified during the evaluation in Chapter 4 by training dedicated forests for each of these. To this end, training samples suitably representing the sub-problems are identified and clustered with a patch-based, non-linear spectral clustering approach resulting in a fuzzy

partitioning of the problem space. These regions, in turn, denote each forest's training as well as application catchment basins. An evaluation on stroke cases and a dedicated discussion follow.

In **Chapter 6**, a semi-supervised variant of DFs is proposed and evaluated on MS cases, which suggest themselves due to their longitudinal properties. This includes the development of multi-dimensional density estimators and a fast label transduction approach. The evaluation additionally comprises an extended hyperparameter analysis and the chapter concludes with a thorough discussion.

In **Chapter 7**, the proposed methods are summarized again, discussed in a general context particularly focusing on their strengths and limitations, reviewed for their clinical relevance, and then assessed for their suitability for practical applications. Finally, an outlook on further developments is given.

1.3 Original contributions

Parts of this work were previously published in journal articles and conference proceedings. The general framework is introduced in a number of publications [Maier *et al.*, 2014a; Maier *et al.*, 2015e; Maier *et al.*, 2014b] and compared against other classifiers [Maier *et al.*, 2015d]. The method participated in multiple public benchmarks [Maier *et al.*, 2016; Maier *et al.*, 2015b; Maier *et al.*, 2015h; Maier *et al.*, 2015g; Maier *et al.*, 2015f], where it won a total of five awards. The local problem forest variant is presented in Maier *et al.*, 2015a. Further works include the organization of two public evaluation challenges on stroke lesion segmentation in the scope of the International Conference on Medical Image Computing and Computer Assisted Intervention 2015 and 2016, resulting in a journal article [Maier *et al.*, 2017] and the publication of a post-proceedings volume [Crimi *et al.*, 2016]

Chapter 2

Magnetic resonance imaging of the brain

Magnetic resonance imaging (MRI) is a non-invasive imaging modality primarily employed to image the anatomy and the physiological processes of the body in both health and disease. The technique is based on strong magnetic fields and radio frequency (RF) waves to trigger faint signals from unbound protons, i.e, principally water molecules. Compared to other imaging techniques, it is highly configurable and can hence be tailored to the specific needs at hand, which makes it especially suitable to visualize the various aspects of brain lesions.

Before attempting segmentation in MRI images, it is favorable to know how the sequences are generated and, subsequently, which physical properties they actually reflect (and which not). Combined with the information about the pathologies presented in the next chapter, this allows to identify and correctly judge the most suitable MRI sequences for each pathology. A sound knowledge of the variability in appearance of MRI sequences and of the imaging artifacts makes it easier to anticipate possible complication and to suitably prepare for them.

In this chapter, the basic mechanisms of MRI are roughly outlined, the sequences employed in this work are introduced and some of the modality's advantages and drawbacks are discussed.

2.1 How it works

MRI is a non-invasive imaging technique based on magnetic fields. In a typical scanner (Fig. 2.1), a static main magnet creates a stable, homogeneous field that is manipulated by gradient and RF coils to trigger faint signals from unbound protons which in turn are recorded by the RF coils to create the image. These coils allow for a deliberate fine control of the magnetic field properties at resolutions of under one millimeter by carefully setting their angles, directions, and power. The interaction with the molecules takes place at the level of the atomic nuclei, whose type, distribution, and structure determines the tissue properties measurable with MRI. What is measured is therefore the abundance of excitable atoms in the currently observed volume, mainly hydrogen, which is why MRI images are often (not quite correctly) described as measuring the tissue's water content. The actual scanning process comprises a number of complex spin manipulation techniques and spin-spin interactions, which to describe is out of the scope of this work. A more detailed account can be found in App. C.

A main characteristic of the MRI modality is its flexibility, which allows for various tissue properties respectively tissue property derived signals to be measured by manipulating the

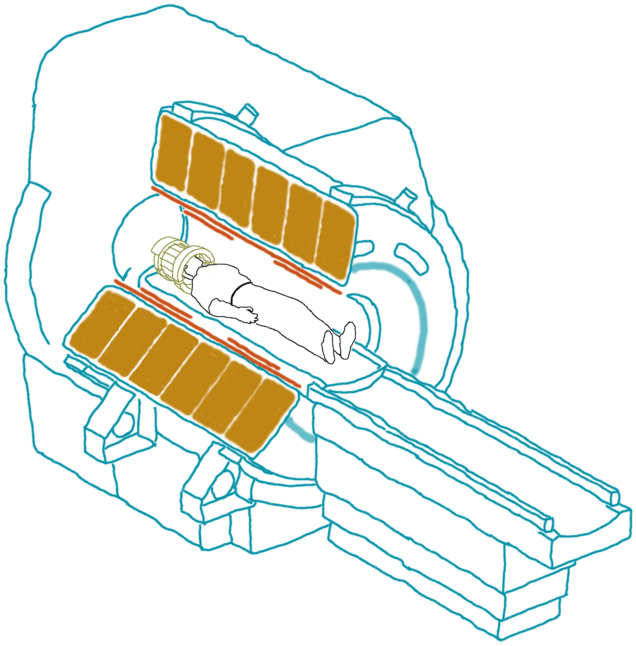


Figure 2.1: Cut-away of an exemplary MRI scanner with main magnet (yellow), gradient coils (orange) and a head RF coil (green). Refer to online version for colors.

scanning parameters. On the downside, this increases the complexity of interpretation.

2.2 MRI sequences

The flexibility MRI offers has led to the definition of a number of sequences, which are essentially collections of scanner parameters suitable to highlight certain tissue properties and/or to discern between tissue types, such as fat, air, white matter (WM), gray matter (GM) and cerebral spinal fluid (CSF). They differ in what they show, how they show it, their tissue contrast, their resolution and their signal-to-noise ratio (SNR). In the following, the sequences that will be employed in this thesis are presented accompanied by image examples. The focus lies hereby on the appearance of normal brains as the lesion appearances differ between pathologies and will be discussed in the next chapter. The section concludes with a short presentation of the challenges arising from the modality's high modularity.

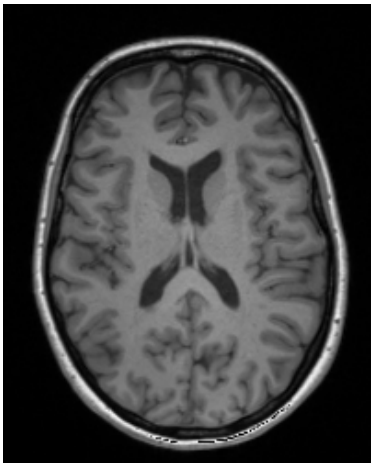


Figure 2.2: **T1**: One of the base MRI sequences whose signal reflects the T1 relaxation times of tissue. It is an often used anatomical image with a good contrast between the dark fluid and brighter brain tissue areas. Furthermore, one can distinguish between darker GM and brighter WM. Many brain lesions appear hypointense in T1 and are difficult to distinguish from the fluid.

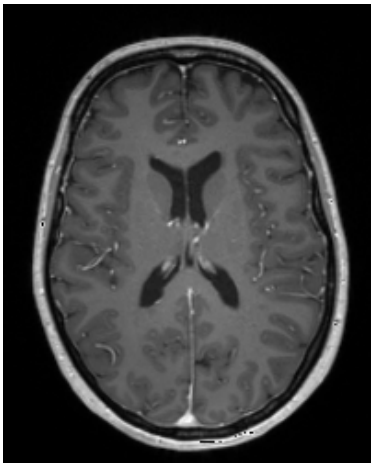


Figure 2.3: **T1c**: Contrast-enhanced imaging with Gadolinium is often applied to T1 images, causing areas with high blood intake to show up hyperintense. This is a useful sequence for many types of brain lesions, e.g., to highlight the actively growing part of a tumor or a Multiple Sclerosis (MS) lesion. Note the enhancement of the cerebral arteries typical for this sequence.

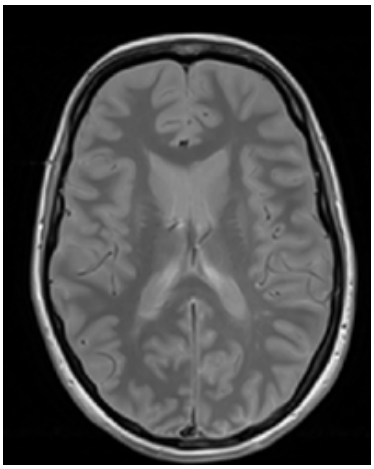


Figure 2.4: **PD**: The positron density sequence reflects the actual positron density and can be placed somewhere between T1 and T2. Fluid and fat appear bright and it possesses a good WM to GM contrast. Cerebral vessels often appear hypointense.

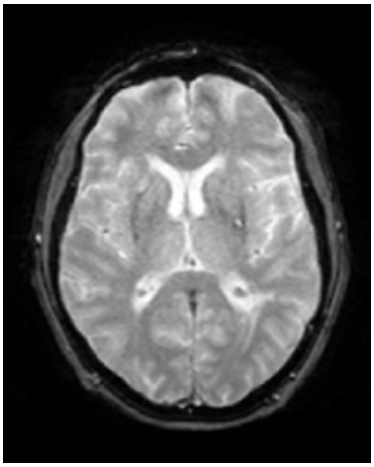


Figure 2.5: **T2**: Another of the base MRI sequences whose signal reflects the T2 relaxation times of tissue. Its gray-scale map is inverse to the T1's, i.e, fluids show bright and brain tissue dark. Many brain lesions appear hyperintense and are difficult to distinguish from the fluid, especially in the cortical and periventricular areas.

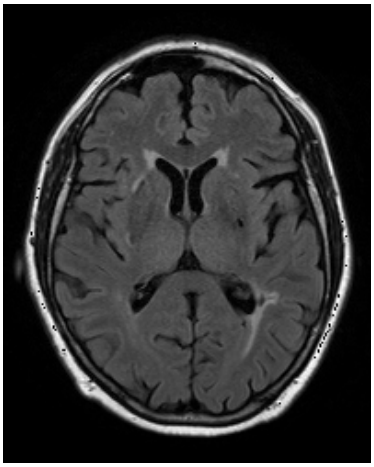


Figure 2.6: **FLAIR**: Fluid attenuation inversion recovery is a T2 image with fluid-signal suppression, i.e., fluids do not longer appear bright, allowing for a better contrast between lesions and CSF (e.g., periventricular). Hence, FLAIR sequences are highly sensitive for lesion detection and used in many settings.

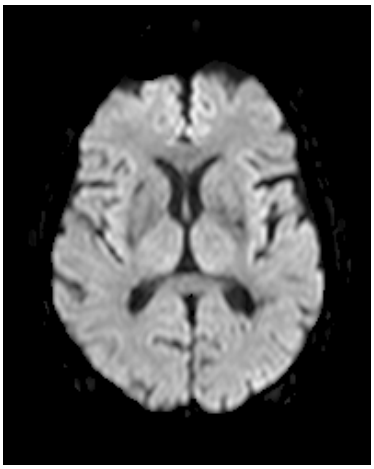


Figure 2.7: **DWI**: Diffusion weighted imaging records the Brownian motion of water molecules, i.e., their diffusion in tissue. Since this effect is measured in three or more directions a DWI image is essentially a tensor image. However, in practice most physicians only work with its mean, the so-called trace image. The sequence has a low SNR and is essentially a mix of T2 relaxation and diffusion signal, controlled by a weighting parameter b (the higher b, the stronger the diffusion weight and the more noise). The here shown image is a DWI trace at $b = 1000$.

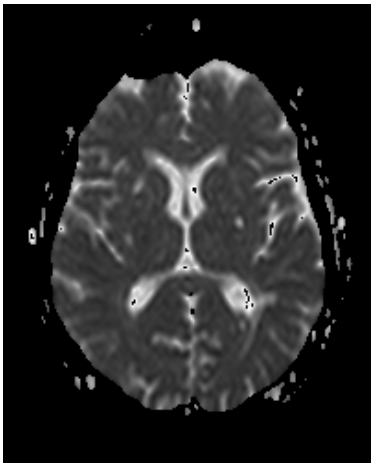


Figure 2.8: **ADC**: The apparent diffusion coefficient is a map computed from two or more DWI trace images recorded at different b-values. While a DWI image is a diffusion weighted T2 image, the ADC map denotes the actual magnitude of diffusion (in $\mathrm{mm}^2\,\mathrm{s}^{-1}$). This allows to overcome the T2 shine-through effect of DWI images as real diffusion restrictions show up hypointense in ADC while T2 hyperintensities show up bright (see Sec. C.6.3 for more details on this subject).

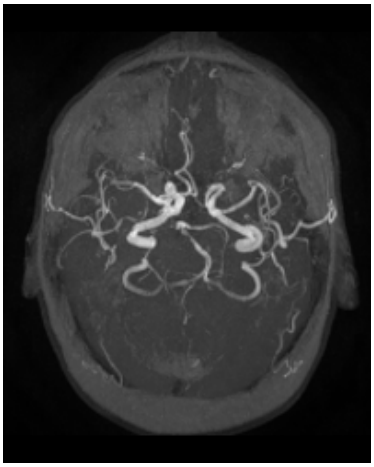


Figure 2.9: **MRA**: Magnetic resonance angiography is a technique to image blood vessels by flow effect or contrast analysis. The resulting volume is usually displayed as maximum intensity projection (see left). While this sequence is not employed in the methods of this thesis, it is of certain value for brain lesion diagnosis and mentioned here for completeness.

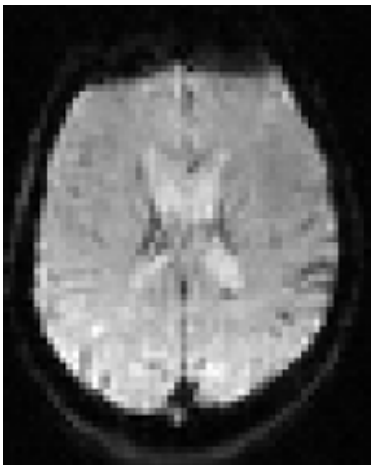


Figure 2.10: **PWI**: Single time point of the multiple T2* images acquired during Dynamic Susceptibility Contrast (DSC) perfusion analysis.

Perfusion weighted imaging A complete family of MRI maps originate from perfusion weighted imaging, which assesses a tissue's perfusion with blood. A common technique is the so-called DSC perfusion MRI, where a gadolinum contrast bolus is injected and passes through the cardiovascular system. During this time, repeated T2* images are shot. A voxel-wise analysis of this time-resolved data allows to plot a DSC intensity curve for each voxel, representing the passage of the bolus over time characterized by a single, sharp maximum. From this curves, a number of semi-quantitative parameters can be obtained, the so-called perfusion maps (see Sec. C.6.4 for a more detailed account).

When interpreting the resulting maps, however, it should be kept in mind that their values are highly dependent on the contrast bolus's compactness and carry little meaning by themselves. Rather, they should be used in a relative sense, e.g., by comparing one hemisphere against the other. Furthermore, the computation of the perfusion maps requires the (manual or automatic) definition of the representative arterial input function (AIF) which can strongly influence the results. Even software claiming to use identical methods may give different results when applied to the same raw data. Caution is therefore advised in relying too heavily on the absolute numbers obtained from such quantitative methods.

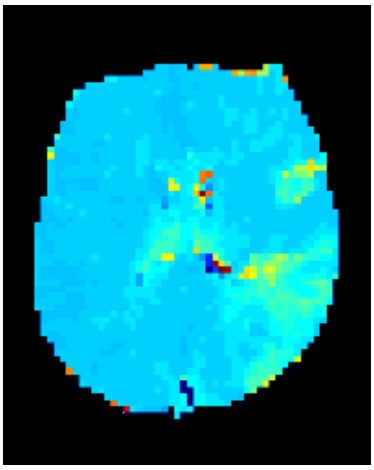


Figure 2.11: **TTP**: Time-to-peak denotes the time in seconds until the maximum contrast is reached in each voxel. Scale omitted. Refer to online version for color.

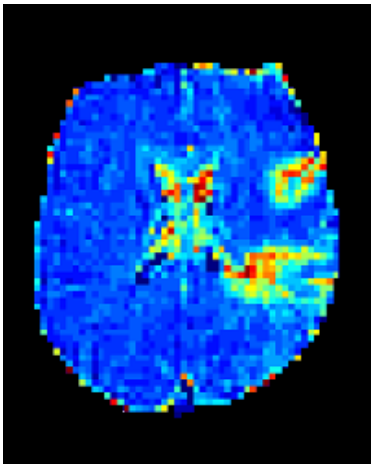


Figure 2.12: **MTT**: The mean transit time denotes the interval between the first notable appearance of the bolus until its last trace in a tissue voxel in seconds. Scale omitted. Refer to online version for color.

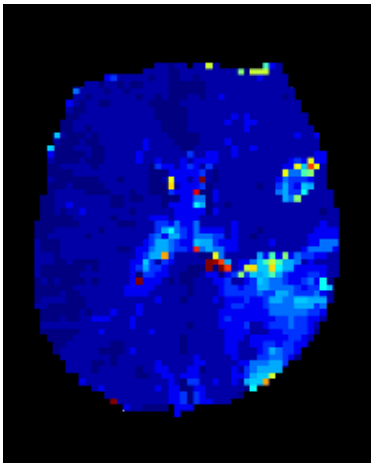


Figure 2.13: **Tmax**: Denotes the time between the peak of the AIF and the peak of the residue function in seconds. It is hence highly dependent on the chosen location of the AIF and its frequent use in stroke analysis has been criticized [Calamante *et al.*, 2010]. All perfusion maps, but especially Tmax, can contain extreme outlier values caused by computational hazards. Scale omitted. Refer to online version for color.

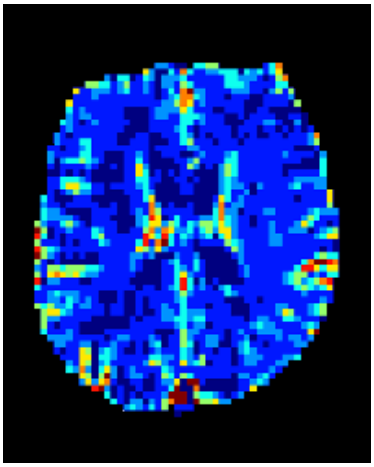


Figure 2.14: **CBV**: The cerebral blood volume denotes the area under each voxels contrast intensity curve in ml/100g. This map is usually computed from CBV=CBF×MTT which makes it a relative measure. Scale omitted. Refer to online version for color.

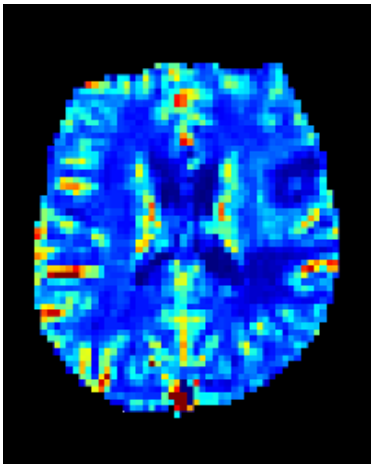


Figure 2.15: **CBF**: The cerebral blood flow is reported in milliliter per minute passing through 100 g of tissue, i.e, in ml/min/100g. It is computed from the rising slope of each voxels contrast intensity curve. Just as the CBV, it is a relative measure. Scale omitted. Refer to online version for color.

2.3 Variability

The many configuration options of MRI are at the same time an advantage, as the scanning process can be easily tailored to the requirements at hand, and a drawback, as one scan can differ greatly from another. This renders it difficult to develop general image processing methods for MRI.

Sequences can differ in terms of resolution at which they are acquired, which in turn determines the SNR as well as the level of image detail. Moreover, the scans often do not constitute full 3D volumes, but rather stacked 2D images, which can be acquired from different directions. The image appearance is additionally influenced by the scanning procedures employed, various of which exist and even the preset procedures vary from vendor to vendor. Furthermore, even the main parameters of the sequences are not reliable: T1 sequences are, for example, defined by a short repetition time (TR) and short echo time (TE), but not by concrete values for these parameters. Finally, there are the image reconstruction algorithms which can differ. Therefore, one T1 weighted image can have little in common with another T1 weighted images, except the direction of the gray-scale map (see Fig. 2.16).

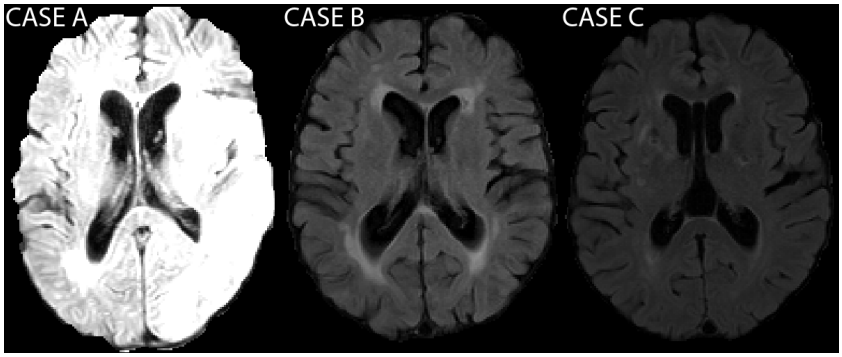


Figure 2.16: Example of three FLAIR sequences of different cases displayed here with the same window settings to highlight the variability in the intensity scales encountered.

To complicate things further, the actual appearance of an MRI sequence depends not only on the scanning settings chosen, but also on the actual machine used, it's calibration, the patient's build, even on the type of objects sharing the room with the scanner. Solutions to deal with this high variability have to be found when aiming at segmentation from MRI images.

2.4 Image artifacts

Beside their natural variability, MRI scans are burdened by a range of imaging artifacts that can manifest as local or global distortions, histology unrelated intensity variations, blurring, and ghostly shadow images. Some are tissue, some motion and other technique related. Some only affect specific sequences and others are the consequence of certain acquisition techniques. Furthermore, high field strength scanners such as the 3 T generation are known to be more prone to artifacts [Dietrich *et al.*, 2008].

For a correct interpretation of MRI sequences, it is essential to be aware of the appearance of these artifacts. Although numerous artifact reduction techniques exist, image processing applications in particular cannot presume that these have been applied and must find ways to cope with their potentially confounding effects.

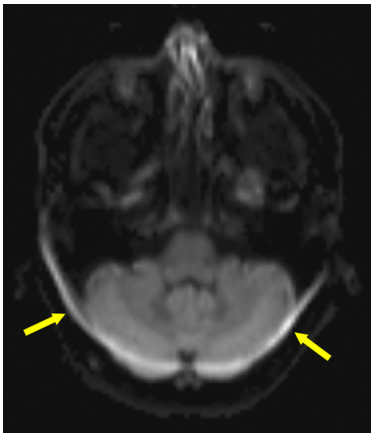


Figure 2.17: Example of chemical shift introduced hyperintensities at the base of the skull of an echo planar imaging recorded DWI trace image. Image source: `http://mri-q.com/chemical-shift-in-phase.html`

Figure 2.18: Example of an inversion recovery bounce point artifact visible as thin black outline at the tissue to CSF interface in a FLAIR sequence. Image source: `http://mri-q.com/ir-bounce-point.html`

In the following, some of the more typical MRI artifacts that affect the sequences employed in this work are described shortly. The focus lies on brain imaging. A reader unfamiliar with the workings of MRI might prefer to read the detailed description of MRI acquisition in App. C first to better understand the underlying effects.

Chemical shift of the first kind Protons precess at different frequencies according to their molecular settings. Since precession determines the protons' resonance frequency, the resulting RF signal after excitation will differ. And since voxel locations are frequency coded, an erroneous shift along the frequency gradient's direction occurs, leading to dark and bright borders displayed on opposite sides of affected structures. This effect is most prominent at (but not restricted to) water to fat transitions, which can be found in the brain at the optical nerves and adjacent to the skull (Fig. 2.17). When an image is recorded with fat-suppression enabled, such as the FLAIR sequence, the effects of this artifact are minimized.

Inversion recovery bounce point artifact At some inversion time (TI) values chosen for FLAIR imaging, the CSF and brain signal cancel each other out. Mixed CSF and brain tissue voxels can therefore appear black, resulting in a thin black outline around the anatomical structure. Contrary to the chemical shift of the second kind, which it resembles in effect, this artifact is common in brain MRI (Fig. 2.18).

Susceptibility Artifact Diamagnetic and paramagnetic tissue properties can lead to dispersion resp. concentration of the magnetic field, resulting in variations of the precessional frequency.

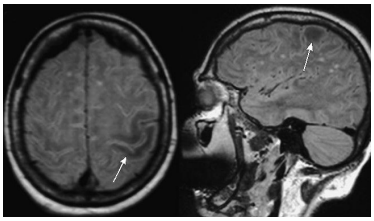
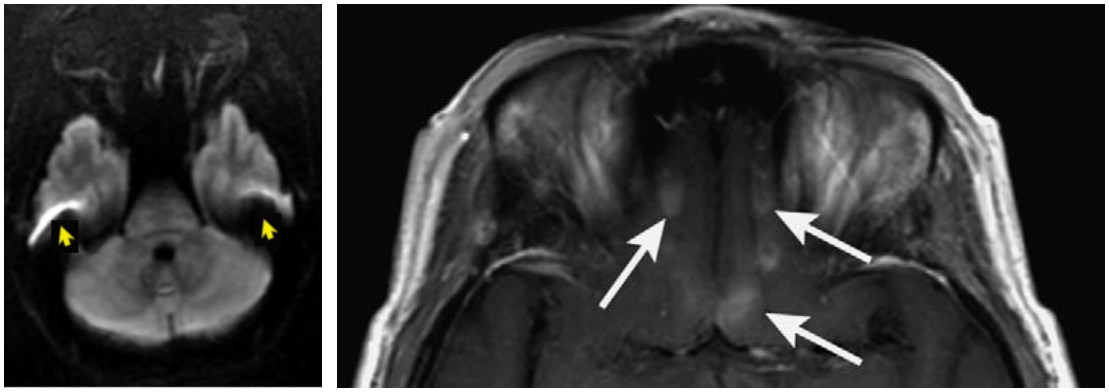


Figure 2.19: Example of susceptibility artifact caused by ferromagnetic objects often encountered in brain imaging. Image source: Vargas *et al.*, 2009



(a) Magnetic susceptibility at the skull base of a DWI trace image.

(b) Pseudoenhancement at air to tissue interface in T1 image.

Figure 2.20: Examples of susceptibility artifacts caused at natural air to tissue interfaces. Image sources: (a) `http://mri-q.com/susceptibility-artifact.html` (b) Vargas *et al.*, 2009

These, in turn, lead to signal loss and spatial mismapping. Alien ferromagnetic objects, such as screws or braces, cause strong signal distortions (Fig. 2.19). But also natural interfaces (e.g., skull base, cortical bone, air to tissue) can cause, albeit weaker, susceptibility artifacts (Fig. 2.20). They are known to increase in higher magnetic fields and are predominately affecting sequences such as DWI and gradient-echo [Dietrich *et al.*, 2008].

Bias-field A number of factors contribute to smooth, often radial, inhomogeneities in the magnetic fields. Some of the effect is attributed to electric currents introduced by the electric field (termed standing-wave effects), others by the distance dependent sensitivity of the employed surface coils (termed surface coil flare). While caused by various mechanisms, the image processing community tends to refer to the resulting smooth intensity inhomogeneity of the image as the *bias-field* (Fig. 2.21).

Patient motion artifacts Patient movement is a common source of imaging artifacts. The exact effects depend on the sequence and recording technique, but often they blur the image strongly enough that it becomes of little therapeutical use (Fig. 2.22).

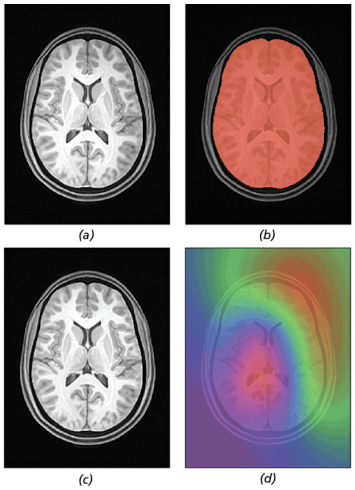


Figure 2.21: Example of a bias-field corrupting a T1 image (a), the brain area used for field computation (b), its estimated strength as color map (d) and the postprocessing cleaned original image (c). Refer to online version for color. Image source: `www.kitware.com/media/html/N3ITKImplementationForMRIBiasFieldCorrection.html`

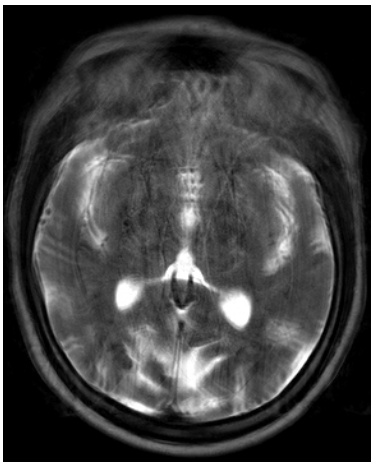
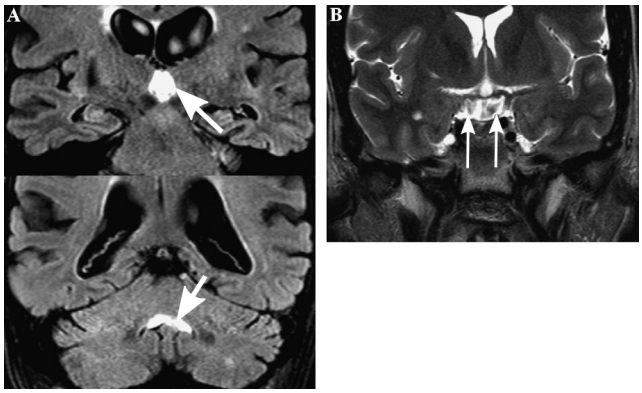
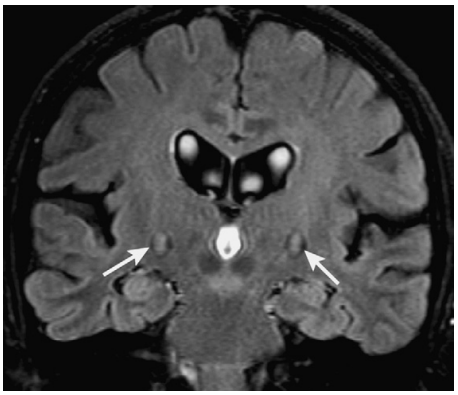


Figure 2.22: Example of a T2 scan effected by patient movement during the scanning session. Image source: `http://mri-q.com/propellerblade.html`

Periodic movement ghosts Contrary to irregular movement, periodical movement, such as caused by breathing, pulsatile flow of blood or CSF and cardiac motion, can cause ghost artifacts, i.e., afterimages of the moving object along the phase encoded direction (due to longer sampling times). Their number and intensity strength depend on the movement strength and the signal intensity of the moving tissue. In the case of brain images, the main source is the pulsative motion of the CSF, resulting in multiple artifacts that may appear hypointense on T2 and hyperintense on FLAIR images (Fig. 2.23).



(a) Hyperintensities in the third and fourth ventricles (FLAIR, left) and the suprasellar cistern (T2, right).



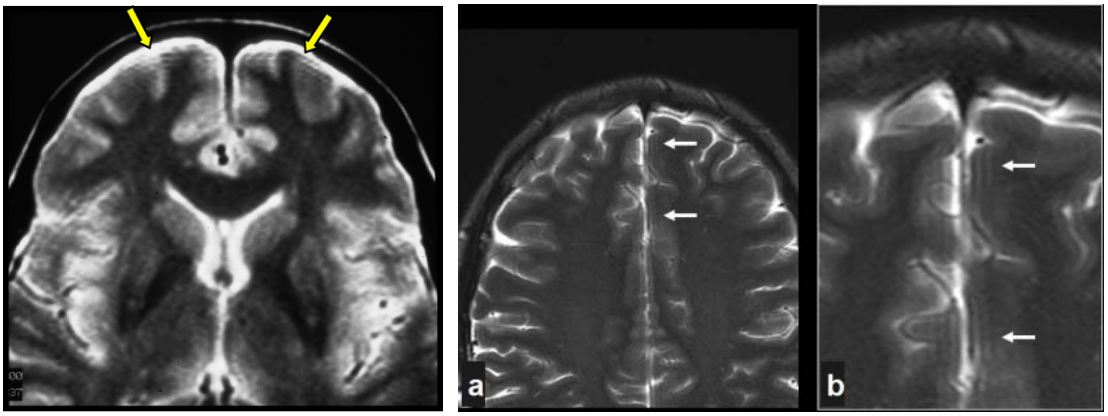
(b) Two ghost image on both sides of the third ventricle caused by its periodic motion.

Figure 2.23: Examples of CSF pulsation caused artifacts. Image source: Vargas *et al.*, 2009

Partial volume effect Since MRI intensities are based on the collective response of protons inside a tissue voxel, each signal constitutes a mixture of the materials present in this volume. Inside homogeneous areas, the RF intensity correctly responds to the tissue properties. But at the borders between materials, the response is a superposition of different characteristics. This is termed the partial volume effect.

Gibbs or truncation artifacts These artifacts are caused by truncation of the Fourier series during reconstruction and manifest themselves as gradually diminishing ripple lines parallel to high contrast interfaces (Fig. 2.24).

Nyquist N/2 Ghosts or three brains artifact This artifact affects only images acquired with the echo planar imaging (EPI) method (e.g., most DWI scans). It shows as three overlapping images, the outer ones with ghostlike appearance at half the image width respectively height in both directions and is caused by slight shifts in the time-reversed zig-zag acquisition used for EPI.



(a) Repeated ghosts of the brain border.

(b) Repeated ghosts of the fissures.

Figure 2.24: Examples of Gibbs artifacts. Image sources: (a) `http://mri-q.com/gibbs-artifact.html` (b) Dietrich *et al.*, 2008

2.5 Summary and conclusion

In this chapter, the MRI modality was shortly introduced, the sequences appearing in the subsequent parts of this thesis presented and examples of the possible imaging artifacts given. Being familiar with the image material at hand allows to anticipate the difficulties of the segmentation task, to identify presented opportunities and to prevent the application of measures that would unnecessarily restrict an algorithm's scope of application.

An important observation is that MRI is highly configurable, which holds advantages as well as disadvantages. On the one hand, the various sequences constitute a rich source of complementary information that can be tapped to develop powerful segmentation algorithms. On the downside, the methods have to cope with different resolutions, varying intensities for the same type of tissue, various image qualities of the same sequences, and more. Furthermore, a number of possibly occurring imaging artifacts have to be anticipated and taken into account.

It is too early to draw any definite conclusions, as this would require knowledge about the concrete pathology regarded. With ischemic stroke and MS, the next chapter will introduce two brain lesion inducing diseases and discuss their appearance in MRI sequences in the light of the observations made in this chapter. The gained insight will furthermore be used to motivate some of the architectural design decisions made in this thesis and to justify some of the measures not undertaken.

Chapter 3

Medical background

Medical image computing as a research field draws its right to exist from the aspiration to facilitate and safeguard clinical analysis, medical intervention and research processes. Before attempting to develop any segmentation solution, it is vitally important to have a complete and profound understanding of the pathology in question: among others its symptoms, management, progression over time, its phenotypes, variability in disabilities and all possible changes in medical imaging lesion appearance. Only then does an informed review of existing solutions, and the careful design of new ones, become feasible. Furthermore, a review of the treatment options and study designs helps to establish a well founded clinical and research motivation for the task. This in turn allows to embed the own work into the overall context without the danger of overestimating its impact.

This chapter details the medical background of ischemic stroke and Multiple Sclerosis (MS), the two main pathologies addressed in this thesis, commencing with an overview of the disease under a medical perspective, including its causes, courses and management. Then, a detailed account of the imaging possibilities is given, discussing the available options, their advantages as well as weaknesses, and typical usage patterns in clinical routine to date. Building on these accounts, an extended motivation for the respective lesion segmentation is established.

3.1 Ischemic stroke

With an estimated number of 15 million incidents annually, ischemic stroke is the most common cerebrovascular disease and one of the most common causes of death and disability worldwide [Mackay *et al.*, 2004]. The identified risk factors for stroke are manifold, including hypertension, obesity, tobacco, etc. It is more frequent among the elderly (40+), while young people are mainly affected due to drug abuse [Mackay *et al.*, 2004].

The human brain tissue is highly dependent on a steady influx of oxygen and other nutritions. Ischemic stroke is triggered by an obstruction in the cerebral blood supply and the ensuing hypoperfusion of brain tissue. Fig. 3.1 schematically depicts such an incidence and the resulting perfusion map. Areas where the blood perfusion drops below a certain threshold (estimated around 10 ml/min/100g [Hakim, 1998]) are ill fated and tissue death is unavoidable. They form the so-called *ischemic core* of diseased tissue. The directly adjacent region with a perfusion between 10 ml and 20 ml/min/100g [Hakim, 1998] is considered tissue at risk, which, if re-perfusion is not established soon, will eventually infarct. The neurons in this so-called *penumbral area* cease to function, but the partial blood supply through proximal arteries keeps them alive for

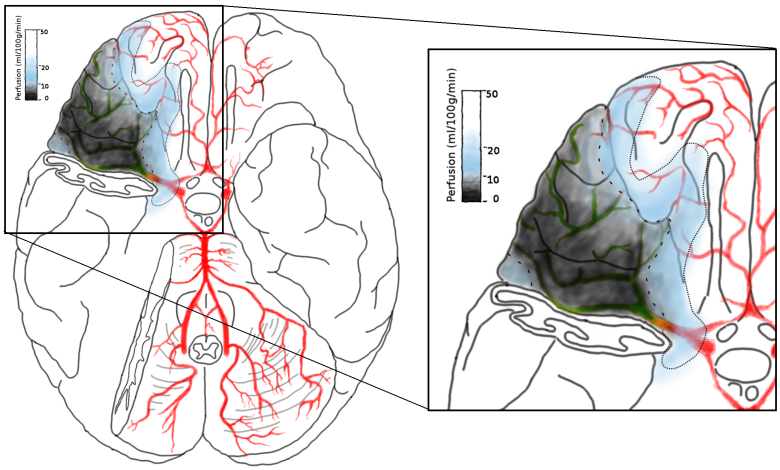


Figure 3.1: Schema of brain tissue perfusion in the case of a middle cerebral artery (MCA) occlusion: The site of the occlusion is marked by a yellow clot and the vessels without blood flow are depicted in green. Hypoperfusion is most severe at the location of the occluded vessels, gradually decreasing towards the areas of non-occluded proximal vessels (e.g., anterior cerebral artery). The dotted line denotes the ischemic core's extend and the broken line the tissue at risk of infarction commonly termed the penumbra. Refer to online version for color.

some hours. This specific area is the target of most treatments and its lifetime determines the available time window.

Default treatment is thrombolysis, the dissolution of an occluding blood clot (embolus) or thrombus by means of intravenously administered drugs. On the downside, this can cause haemorrhage, a risk that increases with the time passed since stroke onset [Clark *et al.*, 1999; Reed *et al.*, 2001]. Official guidelines therefore recommend thrombolysis only up to 3-4.5 hours from onset, which requires a fast assessment of a stroke's cause, location, extend and especially state of infarction, as the risks have to be weighted against the possible gains in terms of rescued penumbral tissue and improved clinical outcome. For an emerging alternative, thrombectomy, i.e., the surgical removal of an occluding thrombus, recent trials [Goyal *et al.*, 2015; Berkhemer *et al.*, 2015] suggest improved functional and mortality outcome in the case of large artery occlusions in the anterior circulation with a wider time window than approved for thrombolysis. But this surgical intervention is associated with similar risks and requires an equally careful weighting of the benefits against the hazards.

For both treatment procedures, non-invasive brain imaging modalities are the physicians most important tools. The acquired images allow for a quick diagnosis and assessment of the current state. But manually interpreting medical images is not trivial, error prone and does not provide quantitative measures on which to base further treatment decisions. The process would benefit greatly from a fast, fully-automatic stroke lesion segmentation method to rule out haemorrhages, quickly assess the current extend and location of both, core and penumbra, and to provide quantitative numbers. At a later stage, a comparison to a follow-up segmentation would allow for a quantitative assessment of the intervention's success.

Unfortunately, the lesions are currently assessed largely visually as the available time is severely limited and manual lesion segmentation is a tedious and time consuming task with segmentation times around 15 minutes per case [Martel *et al.*, 1999], requiring furthermore

sound domain knowledge. Reported inter-rater agreement on ischemic stroke is low [Neumann *et al.*, 2009], rendering derived findings unreliable and irreproducible.

Stroke lesion segmentation is furthermore used as surrogate endpoints in clinical trials evaluating new stroke treatments and as the base component of voxel-to-lesion mapping based brain functionality studies in the field of cognitive neuroscience. In both settings, the required manpower is the main limiting factor of the study size, while the poor inter-rater agreement influences the significance of the findings.

Hence, reliable algorithms for automated ischemic stroke lesion segmentation are in high demand with a broad impact. They are ideally accurate, reproducible, and, at least if intended for the clinical setting, fast. But the task is associated with a number of challenges that are detailed in the next section.

3.1.1 Challenges of stroke lesion segmentation

It is usual to categorize the evolution of ischemic stroke into a number of phases (Fig. 3.2, last row), which are defined by the time passed since onset and roughly associated with the pathological development: The *acute phase* (0 to 24 hours) coincides with the growing of the core until in encompasses the whole penumbra. During the *sub-acute phase* (24 hours to 2 weeks) the swelling persists and many secondary effects and transformations take place. With the *chronic phase* ($>$ 2 weeks), the final lesion outcome is reached. The time course of ischemic stroke lesions is complex and involves a number of not fully understood spatiotemporal interacting processes [Gonzalez *et al.*, 2006], which influence the stroke lesion's appearance in magnetic resonance imaging (MRI) substantially.

Not only between, but also within the phases the lesion appearance changes, especially in the sub-acute phase with its many interacting processes.

The concept of an ischemic stroke lesion does not encompass a single homogeneous area, but rather regions of different molecular composition and with possibly differing tissue fate. Between them there is a gradual transition and they are interacting with each other. This is reflected in the stroke lesion's appearance in MRI sequences (see Fig.3.2, first row), which is inhomogeneous in space as well as in time.

An elaborated arterial system ensures the continuous supply of oxygen, centered at the Circle of Willis at the base of the skull (Fig. 3.3a). The outgoing arteries supply blood through a widely ramified arterial network to their respective areas of the brain, termed their vascular territories (Fig. 3.3b). Arterial occlusion can occur in any of the main and sub-branches, which in turn means that the lesions can appear at any location, and, since furthermore influenced by the layout of the collateral blood supply, in any size and shape.

To complicate things, other types of stroke exist beside the single thrombus cause, such as the embolic shower, where debris from the blood stream lodges in many arterial ends, causing multi-focal lesions, or low-flow stroke in small vessels, which results in multi-focal or stripe-shaped lesions at the borders between the vascular territories.

As intracellular water accumulation increases a few hours after onset, a cytotoxic edema forms in the core's border regions (see Fig. 3.2, middle row). The central core is usually exempt from this due to insufficient blood influx [Quast *et al.*, 1993]. Later on, vasogenetic overtake the cytotoxic causes of swelling, as the disruption of the brain-blood-barrier causes water molecules to seep into the intercellular space [Krieger *et al.*, 1999]. The swelling usually peaks around 24 to 72 hours of onset and resolves within a week [Rosenberg, 1999], which means it mainly affects the sub-acute phase. In the chronic phase, in turn, cavities can form when the diseased tissue is disassembled. The transformations can effect the layout of the whole brain, rendering the application of registration approaches difficult.

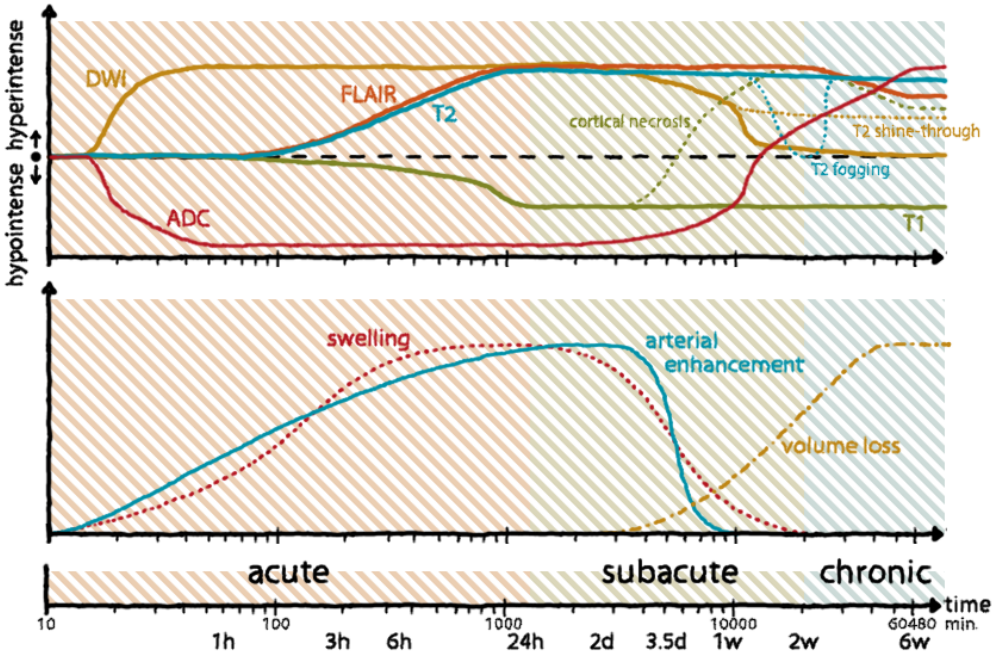


Figure 3.2: Schematic graph of typical development in untreated stroke over time. First row: Intensity changes of stroke lesions in various MRI sequences. Second row: Effects of tissue displacement over time. Last row: Log timeline and the three principal phases of stroke development. Concrete cases might differ substantially. Refer to online version for color.

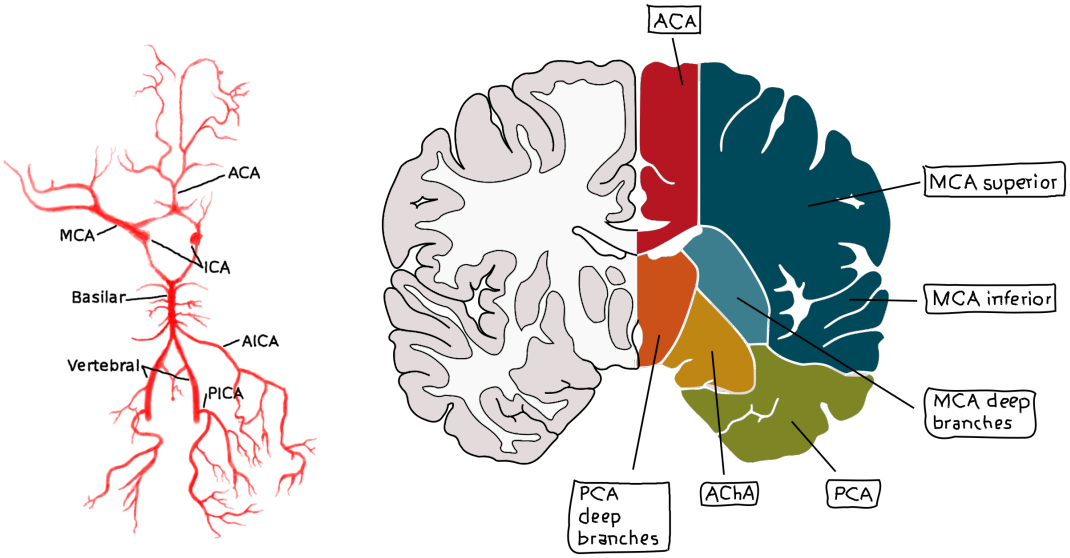
Ischemic stroke segmentation from brain MRI can be considered a challenging problem: The lesion changes over time in composition and appearance, has no fixed location, shape or size; its parts might be undergoing different evolutionary stages and hence appear differently; and, furthermore, secondary effects like swelling and ventricular enhancement can occur. It is usual to concentrate on a single evolutionary phase to control the great variability, but of course, the transition between them is equally gradual. The probably most diverse appearance can be observed in the sub-acute phase, where most changes take place and many processes interact with each other.

In light of the highly variable nature of stroke lesions, envisioned solutions must be kept flexible and general, which opens the possibility of a method transfer to other brain lesion areas. This opportunity is investigated in this thesis.

3.1.2 Imaging of stroke

There is an ongoing debate in the medical community on whether computed tomography (CT) or MRI is the most suited modality for clinical stroke evaluation. Factors often mentioned by authors in favor of CT are the lower costs, the faster acquisition speed and the greater prevalence. Nevertheless, MRI is the modality of choice in many clinical centers since its flexibility in terms of acquisition parameters, as discussed in the previous chapter, can reveal a wider range of details on the stroke lesions. For the same reason it is used in the majority of stroke trials and cognitive neuroscience studies.

Which of the various MRI sequences is most suitable to depict the stroke lesion cannot be



(a) The circle of Willis: Fed by the internal carotid arteries and the vertebral arteries, drained by the brain's major supply arteries.

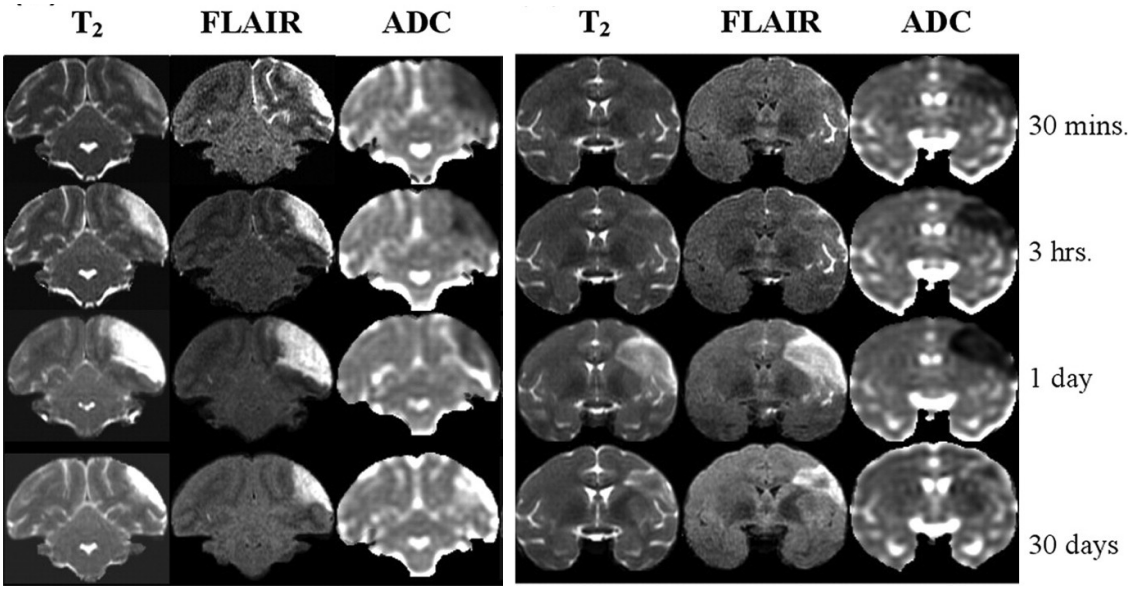
(b) The cerebral vascular territories of some of the brain's major arteries. An individual's layout might differ substantially.

Figure 3.3: Overview of the cerebrovascular system.

readily answered, but rather depends on the lesion's age and the desired information, as each sequence highlights different aspects. As a rule of thumb, PWI and DWI are most informative for lesions in the acute; FLAIR, DWI and possibly T2 in the sub-acute; and T1 as well as FLAIR in the chronic phase. Often, additional sequences can provide further details. Important is that seldom a single sequence suffices, but rather a multi-spectral approach is advisable.

The so-called conventional MRI sequences, T1, T2, FLAIR and PD, are valuable for stroke assessment starting from the sub-acute phase. However, during the acute phase of greatest therapeutic opportunity, these methods do not adequately assess the extent and severity of ischemia [Baird *et al.*, 1998]. Early changes in water molecule diffusion show up on DWI scans and their derived ADC maps, where the visible stroke area roughly denotes cell swelling and the necrotic tissue of the lesion core. Areas of restricted blood perfusion can be visualized with PWI imaging, where reduced values are a signature of hypoperfusion denoting both, the core and the penumbra. This lead to the definition of the diffusion-perfusion mismatch [Albers *et al.*, 2006; Kidwell *et al.*, 2003; Schlaug *et al.*, 1999] to quantify the potentially salvageable tissue, a concept regarded with criticism by some experts [Kidwell *et al.*, 2003]. Finally, MRA is often employed to detect the exact occlusion site.

MRI examples of stroke lesion development over time are given in Fig. 3.4 in accordance with the graphs in Fig. 3.2. Figures 3.5 through 3.8 depict examples of sub-acute stroke lesions, highlighting their diversity in appearance. The concepts of core and penumbra in acute settings and under treatment are detailed by the acute MRI examples in figures 3.9 and 3.10. Chronic stroke is omitted as it is not treated in this thesis. For more details on MRI usage in stroke, see Baird *et al.*, 1998 and Gonzalez *et al.*, 2006.



(a) Stroke over time after 3h MCA occlusion. (b) Stroke over time after persistent MCA occlusion.

Figure 3.4: Stroke appearance in selected MRI sequences over time using the example of macaque primates. Note the different moments of hypo- respectively hyperintensity peaks between the sequences. Images reproduced after Liu *et al.*, 2007

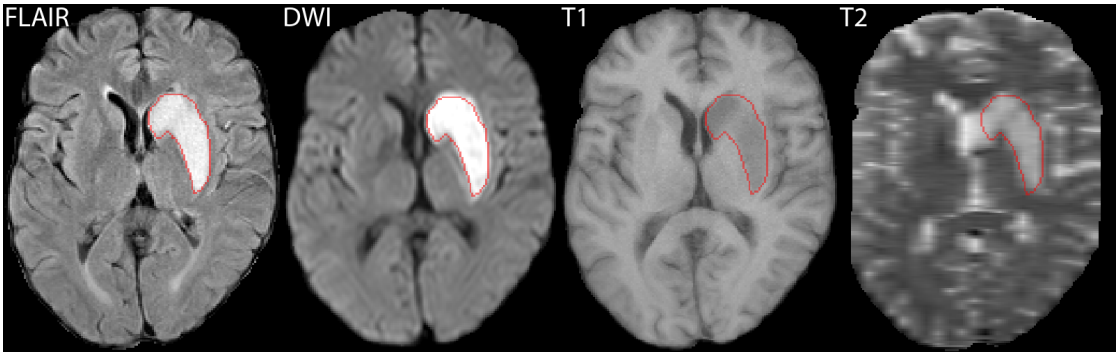


Figure 3.5: Medium sized mono-focal ischemic stroke clearly outlined in all MRI sequences and tightly fitting into the vascular territory of the lateral lenticulostriate artery. Note the compression of the ventricle caused by swelling. According to the image evidence, the lesion can be placed at the early sub-acute phase ranging form approximately one to two days in conformance with the graphs of Fig. 3.2.

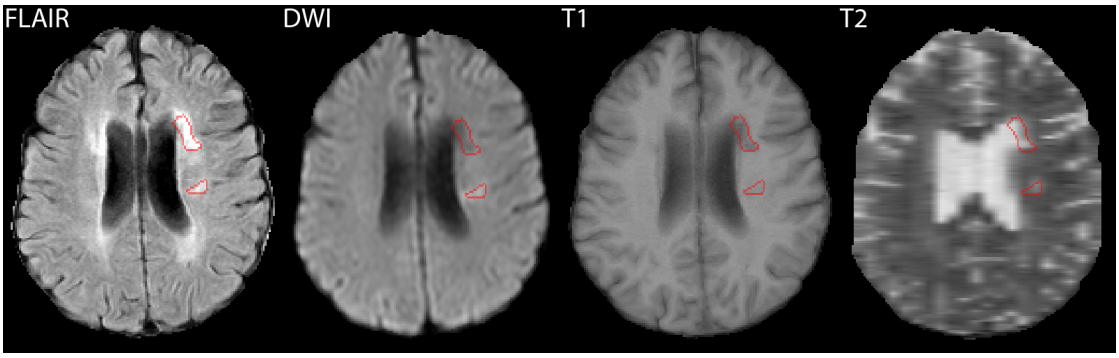


Figure 3.6: Two small lesions of which one can be identified by a hyper- the other by a hypointensity in the DWI sequence. Looking at the T2 image, a T2 shine-through effect in the DWI can be ruled out, which means that the more posterior of the two lesion is most likely younger, maybe caused by dislodged debris from the anterior lesion's thrombus. Both lesions still belong to the sub-acute phase, although to opposite ends of the range.

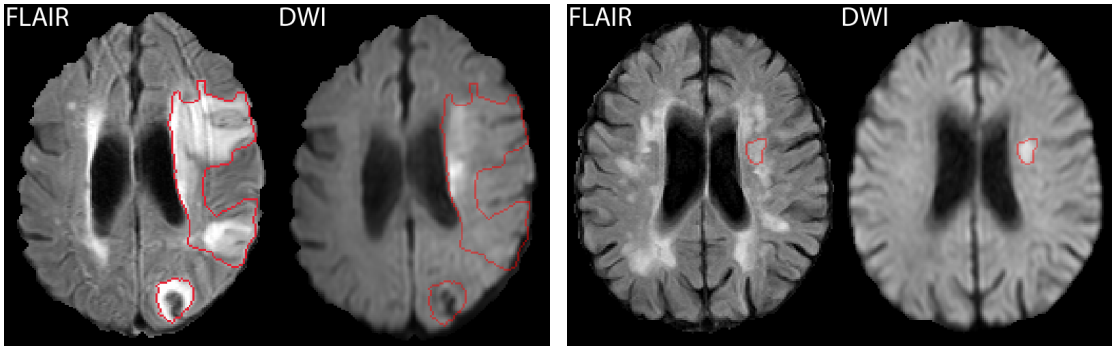


Figure 3.7: Left: Large ischemic stroke lesion with severe spatial intensity inhomogeneities: Only a small part appears clearly hyperintense in the DWI sequence and the intensity levels in the FLAIR differ greatly. Despite the lesion size, only minimal displacement can be observed, hinting towards an age at the sub-acute phase's end around one week. Note also the old haemorrhages (bleedings) in the anterior and most posterior parts of the lesion visible as cavity-like hypointensities. The FLAIR sequence is additionally riddled by image artifacts. Right: Small lesion embedded into an area of many age-induced non-stroke white matter lesions (WMLs). By consulting the DWI sequence, the real lesion localization can be readily established.

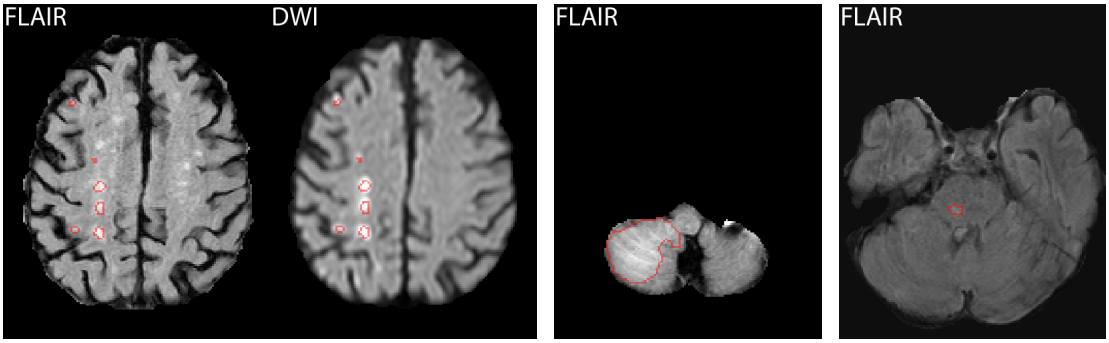


Figure 3.8: Left: Embolic shower, i.e, multi-focal stroke lesion in the dorsal parts of the frontal as well as parietal lobes. Since the area encompasses multiple vascular territories, the causing broken-up thrombus most likely originated from outside of the cerebral vascular system. Middle: Large cerebellar ischemic stroke lesion. Right: Brainstem (pons) lesion. These, despite their small size, are often associated with poor clinical outcome.

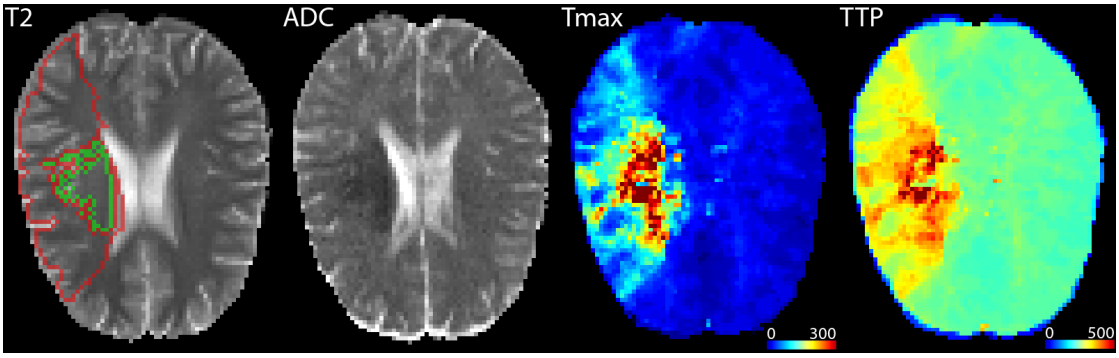


Figure 3.9: Acute stroke with penumbral (red, estimated from Tmax) and core (green, estimated from ADC) outlines. The lesion is not yet discernible in T2, hinting towards a time since stroke of under three hours. The core shows up hypointense in the ADC map, the penumbra and core cause clear perfusion restrictions in the Tmax and TTP images.

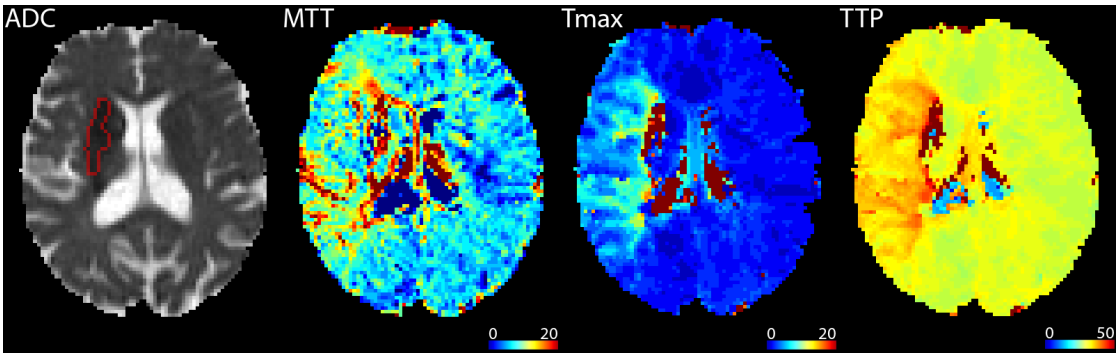


Figure 3.10: ADC and PWI maps of an ischemic stroke at the acute phase (1.5 hours). The outline denotes the final lesion outcome after a successful thrombectomy intervention as delineated in a 90 days follow-up FLAIR scan. While the already necrotic tissue of the core, as depicted hypointense in the ADC, could not be salvaged, the large markedly underperfused region visible in all three PWI maps could be salvaged by re-establishing the blood flow. The expert segmentation does not cover the whole ADC lesion, as early diffusion lesions are in part reversible and often include both irreversibly infarcted tissue and penumbra [Kidwell *et al.*, 2003]

3.2 Multiple Sclerosis

MS is an inflammatory, neurodegenerative and demyelinating disorder of the central nervous system (CNS). The disease is diagnosed in 0.2% of the population [Stüve *et al.*, 2010] and constitutes the leading cause of disability among young adults [Olek, 2015b; Olek, 2015a]. Clinical signs and symptoms include a wide range of sensory disturbances and motor dysfunctions. Peak onset is between 20 and 40 years [Kurtzke *et al.*, 1992; Liguori *et al.*, 2000], with women being affected approximately twice as [Sadovnick *et al.*, 1982] or even more [Debouverie, 2009] often as men. The disease is connected with severe disabilities and a reduced life expectancy of 10 to 12 years [BrønnumHansen *et al.*, 2004]. Annual treatment costs average at $47,000/year [Kobelt *et al.*, 2006].

Clinical characteristics As an inflammatory, demyelinating disease potentially affecting all parts of the CNS, MS displays an extreme variability of clinical signs and symptoms. Some of the most common are a sensory disturbance of the limbs ($\approx$ 30%), partial or complete visual loss ($\approx$ 16%), acute and sub-acute motor dysfunction of the limbs ($\approx$ 13%), diplopia ($\approx$ 7%) and gait dysfunction ($\approx$ 5%) [Stüve *et al.*, 2010]. A *clinical attack* is defined as one or more of these signs occurring for a period of at least 24 hours.

Clinical phenotypes The course of MS is very variable. It can be severe or mild; relapsing-remitting or progressive; and effect mainly the spinal cord and optical nerve or the entire neuroaxis. Most of the gadolinium-enhanced lesions occur early in the disease course, while later stages are less inflammatory but more degenerative. Accordingly, the temporal accumulation of disability accelerates over time.

Multiple phenotypes are defined by the course of the disease. The most common and largely agreed upon definitions of clinical courses of patients with MS were assembled by Lublin *et al.*, 1996 and Lublin *et al.*, 2014 as follows (see also Fig. 3.11):

Relapsing-remitting MS (RRMS) describes a disease course, where clearly defined relapses interchange with full or partial recoveries. Occurs in more than 80% of individuals.

Primary progressive MS (PPMS) describes a largely continuous disease progression from onset, with occasional plateaus or even minor improvements. Occurs in 10% to 20% of individuals. Affects men and women equally, as opposed to RRMS [Weinshenker *et al.*, 1989], which is more prevalent in men.

Secondary progressive MS (SPMS) describes a PPMS-similar second stage of RRMS, which approximately half of all RRMS patients enter within a decade.

Some studies identify further, rare phenotypes (e.g., progressive relapsing MS (PRMS) [Stüve *et al.*, 2010]).

The typical course of the disease starts with the patient entering the RRMS phase. During a period of 10 years on average [Lublin *et al.*, 1996], relapses (clinical attacks) and fast recoveries occur at irregular intervals of approximately one year. These attacks are usually accompanied by a new cerebral or spinal MRI lesion and clinical signs, such as sensory disturbances. During the subsequent recovery phase the clinical symptoms diminish partially or fully, while the associated MRI lesions largely cease their inflammatory activities and either disappear or turn chronic. Such an exemplary course of MS is depicted in Fig. 3.11, left.

Most patients then enter the subsequent SPMS phase of disease progression, which is characterized by a continuous worsening of the clinical symptoms and neurological deficits without any

Figure 3.11: Typical progression of three MS phenotypes as graph of time vs. disability.

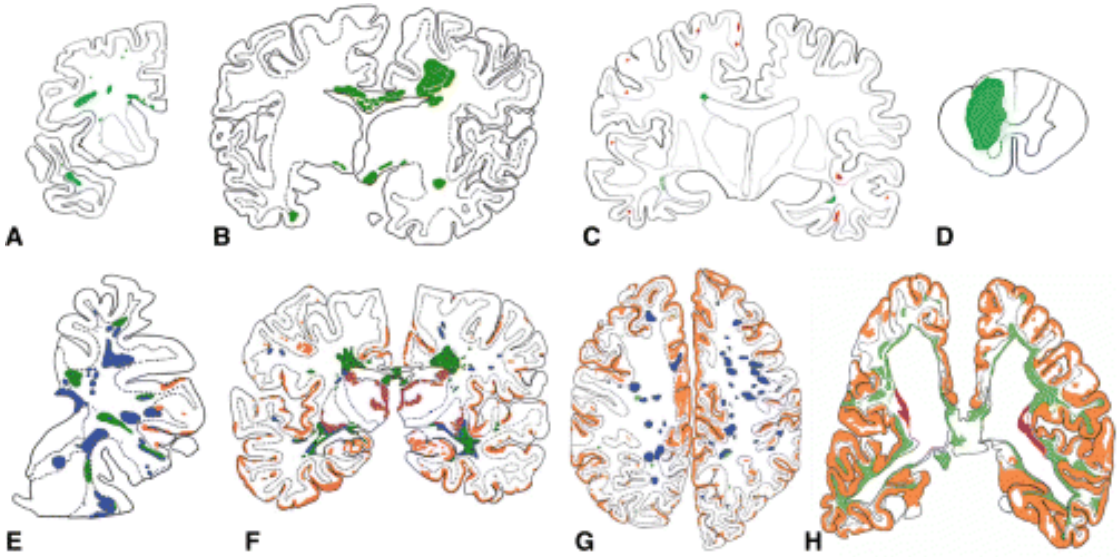


Figure 3.12: MS pathology in the CNS for different phenotypes. Image reproduced after Lassmann *et al.*, 2007.

or only short phases of recovery. In this stage, the MRI lesion load does not necessarily correlate with the clinical impairment.

PPMS constitutes a rarer phenotype of MS with a course similar to SPMS but occurring right from disease onset, i.e., without a preliminary RRMS phase.

Histopathology Histopathologically, MS manifests in two ways: (1) a loss of myelin within the plaque and (2) axonal damage. The former is associated with impaired propagation of action potentials along axons, the latter is by some experts considered the major cause for clinical disability. Both are believed to contribute to brain atrophy, while their relationship, causes and effects remain largely unclear. No clear histological distinction between RRMS and PPMS could be established to date.

Damage to the CNS The MS associated damage to the nervous system tissue has a number of manifestations as shown in Fig. 3.12. For the acute stages of MS (a+b), perivenous inflammatory demyelinating lesions (green) are most typical. In the relapsing/remitting course (e) additional remyelinated shadow plaques (blue) and smaller cortical lesions (red) show up. Multiple cortical

lesions, large demyelinated and few remyelinated plaques are a distinctive manifestation of the secondary progressive stage (f). Since for the primary progressive course (g) no large lesion generating first phase exists, this course is characterized by small, remyelinated focal WMLs, an extensive cortical demyelination and widespread diffuse injury in normal-appearing white matter (WM) (not denoted in the images). (h) shows another examples of the secondary progressive course, with extensive damage to all areas, including the deep gray matter (pink).

3.2.1 Assessment

Periodical neurological examinations are conducted to track changes in disease manifestation. The recommended frequencies range from one month to two years, depending on disease severity and expert consulted. Additionally, brain and spinal cord MRI is undertaken to monitor the disease activity. No consent exists regarding frequency and imaging protocol.

Monitoring options

Clinical trials as well as change monitoring require standardized and reliable methods to assess MS severity. Most methods base on assessing the accumulation of neurological disability (and hence clinical outcome) with different interpretations of the last term.

The Expanded Disability Status Scale (EDSS) proposed by Kurtzke, 1983 was one of the earliest method to quantify MS induced disability. It is still widely used to date [RamoTello *et al.*, 2014; Stellmann *et al.*, 2014], despite heavy criticism for its poor inter- and intrarater reliability, ceiling and floor effects, nonlinearity [Hobart, 2000] and disregard of cognitive impairment.

Various other scores have been proposed and employed in medicament studies, such as the Multiple Sclerosis Severity Score (MSSS) [Roxburgh *et al.*, 2005] and the Multiple Sclerosis Functional Composite (MSFC) [Rudick *et al.*, 2002; Cohen, 2001]. The latter was developed by the US National Multiple Sclerosis Society in reaction to the problems associated with the EDSS. A standardized set of tests of manual (arms and legs) and cognitive function derived from studying changes along longitudinal datasets, the MSFC is largely considered to outperform the EDSS while only moderately correlating with it [Cohen *et al.*, 2000; Kalkers *et al.*, 2000]. A list of further measures can be found in Joy *et al.*, 2001, Chapter 4.

Among researchers, there is no consensus on the usefulness of the various disability measures for MS. But since the measurement of functional status and impairment is central to all aspects of clinical MS research, the development and validation of acceptable measures must remain a priority.

Surrogate outcome measures and their reliability

Above described scores are time-intensive (e.g., MSFC takes around 30 minutes/patient). Hence, considerable effort has gone into developing surrogate biomarkers, which correlate with these scores respectively MS disability, but are faster and easier to obtain.

A surrogate is an outcome measure other than disability that correlates with the latter and is hence suitable to predict clinical outcome. A discussion on validated vs. unvalidated surrogates for MS can be found in Joy *et al.*, 2001, Chapter 6. The search for biomarkers that are suitable surrogate measures is ongoing with few applicable results to date [Davis *et al.*, 2008, Chapter 4]. This may be attributed to a lack of standardization in the field as well as the inherent complexity of MS and the lack of understanding of its underlying processes.

MRI lesion assessment as surrogate outcome measure

MRI scans provide an objective, sensitive, and quantitative measure of change in MS, which make MRI an attractive tool for measuring the outcome of new therapies. In this they differ from clinical outcome measures, which are characteristically insensitive, often poor at reflecting disease activity, and inconsistently defined [Joy *et al.*, 2001, Chapter 6]. Despite being sometimes considered an invalidate surrogate measure, changes on MRI clearly reflect some of the underlying pathological process in MS, implying that MRI is a reasonable surrogate.

MRI could serve as a primary outcome measure in Phase II or exploratory clinical trials, while it should be used only as a secondary outcome measure in Phase III or pivotal trials [Joy *et al.*, 2001, Chapter 6]. In both cases, care has to be taken to exactly define what is being measured to avoid potential errors and difficulties beforehand. McFarland *et al.*, 2002 have shown MRI to be a suitable surrogate for assessing relapses, but not for the progressive phase. Brain and spinal cord atrophy measures are currently under evaluation for this second phase.

Doubts have been raised that MRI reflects the pathology of the disease [McFarland *et al.*, 2002], i.e., a change in the pathological process can be observed that has no or little effect on clinical disease. Furthermore, drugs could act through a mechanism not captured by MRI measures, resulting in false-negative outcomes. Finally, various MRI measures to assess MS progression exist, all of which capture different components and have to be carefully balanced.

Concluding, the use of MRI as unvalidated surrogate measure is wide-spread but remains under discussion. Care has to be taken in the design of any study aiming to employ MRI assessment as outcome measure.

3.2.2 Imaging of MS

The pathological features of MS are highly diverse, which, on the one hand, opens up a great number of monitoring options for disease assessment, but, on the other hand, limits each feature's diagnostic and prognostic value if considered isolated. In this section, the imaging of MS and its derived surrogate measures are discussed.

Visibility

MRI is the modality of choice for MS assessment, as this non-invasive imaging technique offers with his various sequences the tools to make a wide range of MS pathological features visible.

Conventional MRI Different MRI sequences can be used to display different effects of MS on the CNS. On the lesion level, mainly conventional MRI is employed. See figures 3.13 and 3.14 for examples.

T1 Most MS lesions appear hypointense, but in the case of strong inflammation, slight hyperintensities can be observed [Ge, 2006]. If a lesion fails to recover and becomes chronic, it shows up as hypointensity, a so-called 'black hole', which reflects severe tissue damage including both, demyelination and axonal loss [Filippi *et al.*, 2011].

T1c Contrast enhancement with gadolinium reveals blood brain barrier permeability and thus acute inflammation in active lesions. T1c hyperintensities are the first signs pointing toward a newly developing lesion and can last from days to weeks, usually slowly evolving from a nodule to a ring-like shape [Ge, 2006]. In all other aspects, this sequence equals T1.

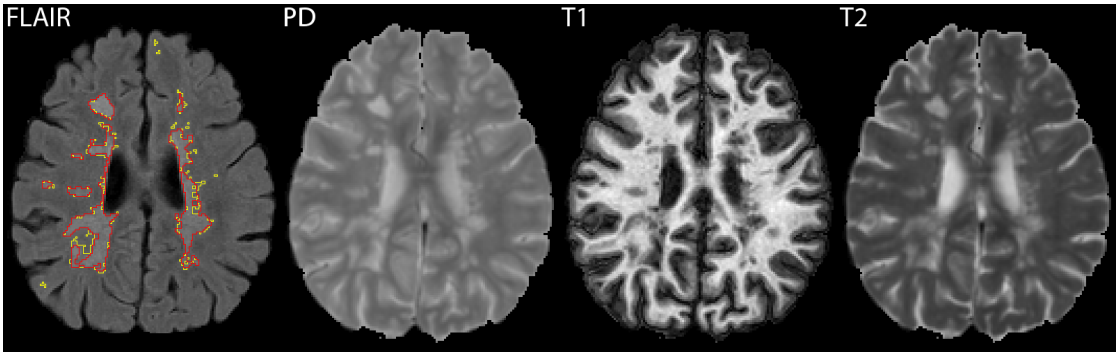


Figure 3.13: Conventional MRI sequences of a patient with heavy MS lesion load: The hyperintense cerebral spinal fluid (CSF) in PD and T2 render it difficult to identify the periventriuclar and fissure-near lesions. In the T1, only parts of the lesions show up as hypointense areas. The best visibility is provided by the FLAIR sequence with its fluid suppression. The two sets of expert segmentations (red and yellow) depicting the lesions in the FLAIR image are largely overlapping but differ at the voxel level. Images taken from the ISBIMS training set. Refer to online version for color.

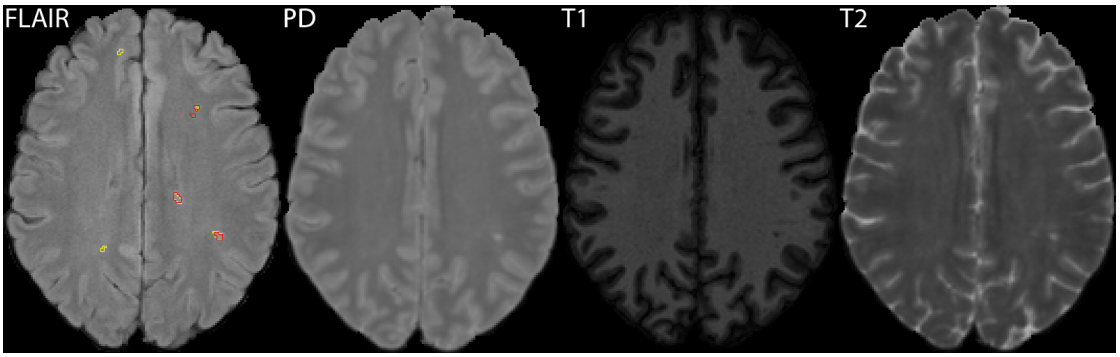


Figure 3.14: Conventional MRI sequences of a patient with light MS lesion load. Only some of the lesions can be identified in the T2 sequence and the expert raters disagree on their quantity. Images taken from the ISBIMS training set. Refer to online version for color.

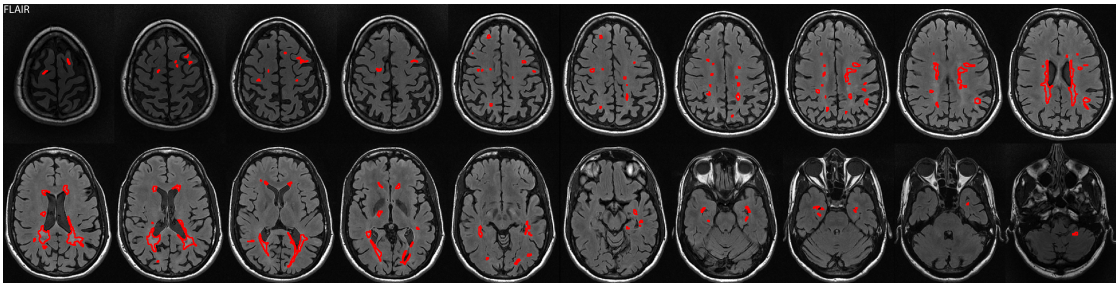


Figure 3.15: Multi-slice example of an MS patient with heavy lesion load. Note the ovoid, finger-like lesion tendrils originating from the ventricular surface and the prevalence in all regions of the brain. Images taken from a previous MS challenge's training set [Styner *et al.*, 2008]. Refer to online version for color.

T2 MS usually manifests in T2 as strongly (in WM) and slightly (in gray matter (GM)) hyperintense focal lesions. They sequence is highly sensitive for lesion detection, but lacks specificity [Filippi *et al.*, 2011] as can be seen in Fig. 3.13.

FLAIR FLAIR, as a T2 derivative sequence, displays similar properties. Differences are a higher sensitivity to GM lesions and a better differentiation of periventricular plaques due to CSF signal suppression (see figures 3.13 and 3.14).

Non-conventional MRI Conventional MRI sequences are sufficient to display MS pathologies on a lesion level. Although sometimes histopathological changes can be derived from a lesion's appearance, only non-conventional MRI allows for the direct in-vivo assessment of the heterogeneity of MS pathological features. Suitable sequences can reveal, e.g., macrophage infiltration, abnormal iron depositing, normal appearing WM, edema, inflammation, demyelination, remyelination or axonal loss [Filippi *et al.*, 2011]. Alas, no standards for these sequences has been established to date, nor has their suitability as surrogate measures been confirmed. Thus, their practical use is limited and they are seldom employed for diagnosis, prognosis or outcome prediction.

Lesion appearances and location MS lesions can appear throughout the brain, take on any shape and appear in WM as well as GM. Nevertheless, a prevalence for the periventricular WM regions has been observed. In their initial stages, they appear thin and linear, but soon take on their typical ovoid shape with the major axis often perpendicular to the ventricular surface [Horowitz *et al.*, 1989]. Other common areas of appearance include the corpus callosum, optic nerve, subcortical region, U-fibers, visual pathways and brain-stem [Ge, 2006]. Fig. 3.15 denotes a typical MS examples illustrating most of these concepts.

Assessment

The good visibility of MS pathological features in MRI renders the modality an ideal candidate for MS assessment. It is generally acknowledged that conventional MRI provides objective measures to monitor disease activity and progression [Filippi *et al.*, 2011] and it has been formally included as a diagnostic criteria (see, e.g., McDonald criteria, revised version [Polman *et al.*, 2005; Polman *et al.*, 2011]), as it reveals disease dissemination in time and space. Current recommendations state that any MRI protocol for MS assessment should include T2 (dual-echo), FLAIR and

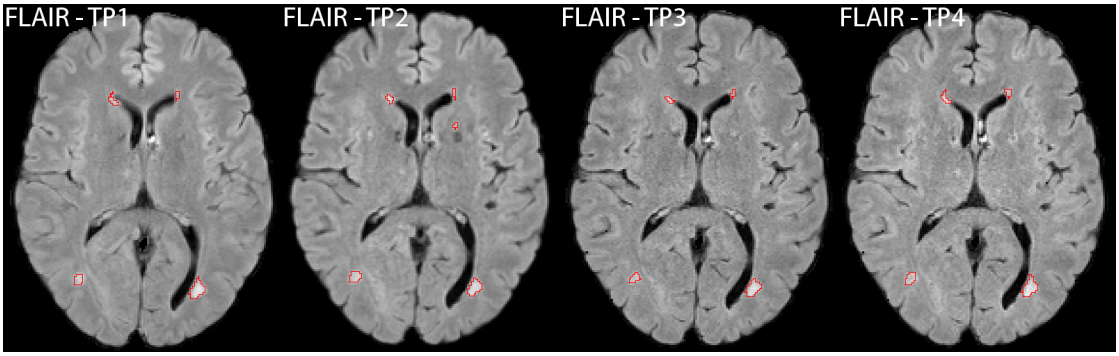


Figure 3.16: Example of MS lesion development over multiple time points (TPs): In the second TP a new lesion appears, probably triggered by a clinical attack. Since it disappears again before the next scan the patient is most likely in the RRMS phase. The other lesions grow and/or shrink slightly, altering their shapes and appearances. Image taken from the ISBIMS training set. Refer to online version for color.

T1c [Filippi *et al.*, 2011]. In the prognosis of the clinical outcome of MS, MRI is only sparsely employed, since only few studies have claimed, albeit limited, prognostic roles [Fisniku *et al.*, 2008; Brex *et al.*, 2002]. The situation differs for treatment trials, where T2 (dual-echo), FLAIR and T1c are established surrogate measures of clinical outcome and commonly used in phase II and III trials [Filippi *et al.*, 2011].

Surrogate measures in use

Various MRI based measurements have been suggested as surrogates for disability respectively clinical outcome. These include, among others [McDonald *et al.*, 2001], lesion count, lesion location, lesion load, new/disappeared lesions, active lesions, enhancing/shrinking lesions and brain atrophy (Fig. 3.16). The suitability of these measures and their correlation with the clinical outcome have been investigated in a number of studies without any conclusive results. The development of a reliable, significant and standardized surrogate measure is of foremost importance in the field and the discussion is ongoing. It is out of the scope of this work to provide an overview over the different positions and results obtained to date.

Concluding, it can be said that MRI is the imaging modality of choice for MS assessment and that lesion-based surrogate measures are widely used in diagnostic and outcome prediction. This underlines the emerging need for reliable, robust and reproducible MS lesion segmentation methods.

3.2.3 Motivation for MS lesion segmentation

Building on the detailed account of MS and its management as given above, this section aims at developing a strong motivation for MS lesion segmentation.

Clinical setting A strong need for automatic MS lesion segmentation exists in the clinical setting. Already the diagnosis of the disease includes an MRI imaging protocol and the count, distribution and load of existing, enhancing and chronic plaques serves as marker for the dissemination in time and space. Manual lesion segmentation and counting is subject to intra- and

inter-observer variabilities [Grimaud *et al.*, 1996; Styner *et al.*, 2008], which renders the diagnosis unreliable and unreproducible. Hence, to allow for a standardized, reproducible and robust quantitative assessment of the lesion measures, an automatic method for MS lesion segmentation from MRI sequences is of the foremost importance.

Research setting But also from a research setting stems a well-founded need for automatic MS lesion segmentation methods. Various MRI based markers are employed in clinical phase II and III trials, where longitudinal studies aim to assess and quantify lesion development over time. Manual approaches can lead to variations in the lesion segmentations between two different TPs, even if the task is performed by the same rater, rendering all derived findings unreliable. Furthermore, the delineation of MS plaques is a very time consuming and tedious task [Grimaud *et al.*, 1996], especially in large studies where hundreds of cases have to be processed. An automatic lesion segmentation method would allow for a more reliable tracking of lesion development over time and largely reduce the workload required for large trials. Furthermore, the obtained results would be rendered reproducible.

Method transfer Beside the above detailed motivations, any findings obtained in MS lesion segmentation can, with some adaptation, be transferred to other MRI segmentation tasks, such as stroke lesion or tumor delineation. This thesis supports this claim by employing the developed methods equally in tumor and stroke lesion segmentation.

Chapter 4

Brain lesion segmentation from multi-spectral MRI with decision forests

The term brain lesion segmentation denotes the delineation of pathological tissue in images of the brain, here specifically magnetic resonance imaging (MRI) sequences. The preceding introductory chapters present two of many pathologies causing brain lesions and ample motivation for automatic segmentation frameworks to support diagnosis, treatment and clinical research.

A brain lesion segmentation application should be fully automatic to impose minimal demands on the user and to allow for an execution in the background; have an execution time below a few minutes to support decision making in time-critical environments, such as acute stroke treatment; be robust against variation in the input data, which can vary strongly in clinical settings; produce a meaningful output that satisfies the clinical requirements; and be able to process multi-spectral images, as each MRI sequence provides a complementary view on the regarded pathology.

Developing a suitable solution faces a number of challenges, which can be divided into two categories. One set is pathology based, i.e., problems arising from the lesion's diversity in shape, locality, number and appearance (see Chapter 3 on stroke and Multiple Sclerosis (MS) lesions). The second set stems from the MRI modality with its varying intensity distributions, numerous artifacts and non-standardized scanning protocols (see Chapter 2). All of these have to be carefully studied and taken into account when aiming for a general brain lesion segmentation method.

The past has seen the publication of numerous brain lesion segmentation methods with incomparable evaluation results obtained on private datasets and with various metrics. Fortunately, the recent surge in so-called medical image processing challenges[1] has led to the release of various benchmark sets for brain lesion related problems with public training and hidden, third-party evaluated testing datasets. The latest at the time of writing include BRATS 2015 for glioma, ISLES 2015 for ischemic stroke and ISBIMS 2015 for MS segmentation, in all of which the method proposed in this chapter is evaluated (see Sec. 4.5). The methods participating in these ongoing competitions [Menze *et al.*, 2015b; Maier *et al.*, 2015c; Pham, 2015] can be considered the current state-of-the-art and, excluding semi-automatic approaches, are summarized here.

A popular generative approach to brain lesion segmentation is to look for outliers in Gaussian Mixture Models (GMMs), which is successfully employed for tissue segmentation in healthy

[1] http://grand-challenge.org

brains [Van Leemput *et al.*, 1999]. This path was chosen by Sudre *et al.*, 2015, who developed a hierarchical model; by Agn *et al.*, 2016, who enhanced the approach by an additional shape prior based on a convolution restricted Boltzmann machine output; by Haeck *et al.*, 2016, who followed up with a level-set for spatial regularization; by TomasFernandez *et al.*, 2015, who combined the global with local atlas based GMMs; and by Catanese *et al.*, 2015, who employed an additional graph-cut step. Unfortunately, these methods do not rank high in the benchmark hierarchies, which can be attributed to the oversimplified lesions models. Brain lesions are known to show considerable intensity overlap with healthy tissue and a strongly varying appearance, in particular for stroke. A similar, but more successful approach is fuzzy C-means tissue segmentation with an explicitly modeled lesion class as taken by Feng *et al.*, 2016.

Alternatively, segmentation in images can be regarded as a classification problem, where each voxel is assigned to a class. This point of view allows for the application of machine learning techniques, powerful algorithms, which can automatically learn complex decision functions from training data. This approach is particularly useful in applications where the causal relation between the input (here: MRI sequences) and the output (here: lesion membership) is not clearly defined but rather based on fuzzy notions of experience, as is the case in brain lesion segmentation.

While a number of methods, such as dictionary learning [Hoogi *et al.*, 2015; Deshpande *et al.*, 2015], have been suggested, two types of approaches dominate the submissions and the challenges' upper ranks.

Popular and successful recent approaches have used deep learning techniques [Dvorak *et al.*, 2015; Dutil *et al.*, 2016; Pereira *et al.*, 2016; Rao *et al.*, 2015; Vaidhya *et al.*, 2016; Kamnitsas *et al.*, 2016; Vaidya *et al.*, 2015], i.e., multi-layer neural networks, which are usually preceded by a convolutional tier for image processing, resulting in the so-called convolutional neural networks (CNNs) [Krizhevsky *et al.*, 2012]. A considerable advantage of these methods is that they are able to learn the image feature descriptors simultaneously with the classification, removing the need to manually craft meaningful features. This in turn means that no special knowledge about the pathology or image modality is required, rendering image segmentation a software architecture task (see Appx. B for a short discussion of the implications). Drawbacks are the high hardware requirements, especially in terms of graphics processing units (GPUs) and memory, the dependency on large and diverse training datasets, and the long training and application times. This might be the reason why only few CNN methods employ actual 3D patches [Kamnitsas *et al.*, 2016; Vaidya *et al.*, 2015].

Among the feature-based classifiers, decision forests (DFs) [Breiman, 2001] are the most used, best performing approaches [Malmi *et al.*, 2015; Meier *et al.*, 2016; Halme *et al.*, 2016; McKinley *et al.*, 2016; Robben *et al.*, 2016; Jesson *et al.*, 2015] and the method of choice for their training and application speed, parallelizability, low hardware requirements and robustness against the choice of hyperparameters. For a successful application, the input data has to be suitably preprocessed and normalized. To this end, adequate knowledge about MRI imaging and lesion appearances as detailed in the previous chapters is employed. Most crucially, suitable features have to be designed that provide the DF with information representative for the problem while not lending themselves to classifier overfitting. The training data must be carefully chosen to suitably represent the problem space. Furthermore, a balance has to be found between big data and training times as well as memory requirements.

Malmi *et al.*, 2015 approach the last problem via a cascading forest architecture which they furthermore claim to better the multi-class prediction. But the DF framework is powerful enough to handle class-imbalances itself as shown in Maier *et al.*, 2015e and inherently solves multi-class prediction. This and other [Robben *et al.*, 2016] cascading approaches could often be substituted by allowing the trees to grow deeper.

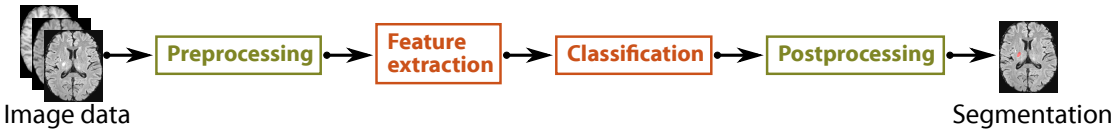


Figure 4.1: Main building blocks of the proposed brain lesion segmentation framework.

The features used by Meier *et al.*, 2016 include first-order texture features. It might be argued that these are unlikely to convey discriminative information about brain lesions, as their texture is usually at a smaller scale than the image noise. Another drawback is their usage of an atlas of tumor probabilities constructed from the training data, an approach that requires large amounts of training data to avoid overfitting, which are not available for all tasks. The challenge rankings show that simply employing as many features as possible [Reza *et al.*, 2015] does not hold any advantages over employing a few, well-designed features, as ,e.g., [Halme *et al.*, 2016] have successfully done. Noteworthy is the well performing method of McKinley *et al.*, 2016, who employ expert knowledge to suitably preprocess the input sequences according to their physical properties.

The challenge of the large problem space is approached by Goetz *et al.*, 2016, who propose a DF that is suited to reveal similarity between images, which in tun is used to select a number of cases similar to the test image from the training set and then subsequently train an online forest on these for the actual segmentation. Unfortunately, this approach did not lead to the best results, denoting the complexity of substantiating an ill-defined concept such as image similarity.

Many proposed DF methods employ a posterior spatial regularization step, such as Markov random fields (MRFs) [Malmi *et al.*, 2015; Jesson *et al.*, 2015] or conditional random fields (CRFs) [Meier *et al.*, 2016], which are useful to correct errors in the forest posteriori prediction. Seen from another perspective, the need to resort to an additional procedure can hint towards an ill-designed forest framework. The assumption is made that the DF classifier should be able to handle most problems directly and therefore only minor morphological operations are used in the proposed approach to reveal the raw DF performance.

Chapter content In this chapter, a DF based, fully-automatic framework for general brain lesion segmentation from multi-spectral 3D MRI is proposed. The key contribution is a set of carefully motivated image features inspired by anatomical, medical and imaging factors. The other contributions include a thoroughly evaluated processing pipeline and a training subset sampling strategy. Fig. 4.1 provides an overview over the main components, which will be discussed in this chapter.

The method is carefully evaluated on four different applications in fair, direct and independent comparisons against current state-of-the-art approaches to show its generalizability to different pathologies. A comparison of the DFs against other standard classifiers confirms my choice and the DF's general popularity. Finally, a thorough investigation of the method's hyperparameters and each feature's contribution is conducted.

In the next section, DFs are shortly introduced. Then the developed features are presented in detail in Sec. 4.2, followed by a description of the chosen pre- and postprocessing methods. The chapter concludes with the evaluation (Sec. 4.4 and Sec. 4.5) and a discussion (Sec. 4.6).

4.1 Decision forest classifier

Machine learning for classification describes the process of automatically inferring general decision rules from a set of manually labeled training samples. Once trained, the classifier can then be applied to formerly unseen samples, ideally assigning the correct class. Each sample is an object to be classified, described by a number of its attributes, commonly named features. Together, these constitute the information available to the classifier for decision making. Interpreting image segmentation as a classification problem entails the assignment of a class (background or foreground in the binary case) to each of the image's voxels. Hence, each voxel constitutes a sample and it's gray value the most simple feature.

Decision trees

The concept of decision tree (DT) classifiers roots in graph theory and was proposed in the early days of machine learning [Breiman *et al.*, 1984]. In a divide and conquer approach they learn simple rules to split a training set along the samples' features according to their class labels and can subsequently be used to classify formerly unseen samples. In this section, their training and application procedures are introduced.

Definition of sample sets Classification describes the process of assigning a class c to an object. Let a single object be described through a set of D_F selected features whose responses form a sample $\mathbf{s} = \{s_1, s_2, \ldots, s_{D_F}\} \in \mathbb{R}^{D_F}$. The type of observation or the features employed are for now of no interest. A sample set $S = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{|S|}\}$ is a collection of such samples describing $|S|$ different objects.

Training Training a DT requires a training set S_{train} with an associated ground truth set $C_{train} = \{c_1, c_2, \ldots, c_{|S|}\}$, denoting the class membership of each sample $s \in S_{train}$. Each $c \in C_{train}$ is from the set of all known classes $C = \{c_a, c_b, \ldots\}$. Ideally, S_{train} is representative for the classification problem at hand, as the algorithm derives its general decision rules from this set.

A DT is constructed top-down, node-by-node from the root node to the leaf nodes as depicted in Fig. 4.2. Let each node be identified by an index n with the root-node receiving the number 0. Training commences at the root node with the set $S_0 = S_{train}$. At each node n the set of incoming samples S_n is split according to a split optimization criterion C_{opt} into two disjunct subsets $S_{n,L}$ and $S_{n,R}$. These are passed on to its left respectively right child nodes. A split function $\phi : S \to \{0, 1\}$ decides to which child node an incoming sample is assigned. It is defined as

$$\phi(\mathbf{s}) := \phi(\mathbf{s}; d, t) = \begin{cases} 0 & \text{if } s_d < t \\ 1 & \text{otherwise} \end{cases}, \tag{4.1}$$

where $d \in \{1, \ldots, D_F\}$ is a feature selector and $t \in \mathbb{R}$ a threshold. Accordingly

$$S_{n,L} = \{\, \mathbf{s} \mid \mathbf{s} \in S_n \wedge \phi(\mathbf{s}) = 0 \,\} \tag{4.2}$$

$$S_{n,R} = \{\, \mathbf{s} \mid \mathbf{s} \in S_n \wedge \phi(\mathbf{s}) = 1 \,\}. \tag{4.3}$$

The training procedure comprises an exhaustive search for the optimal split function

$$\phi_n(\mathbf{s}) = \arg\max_{d,t} C_{opt}(S_n, \phi(\mathbf{s}; d, t)), \tag{4.4}$$

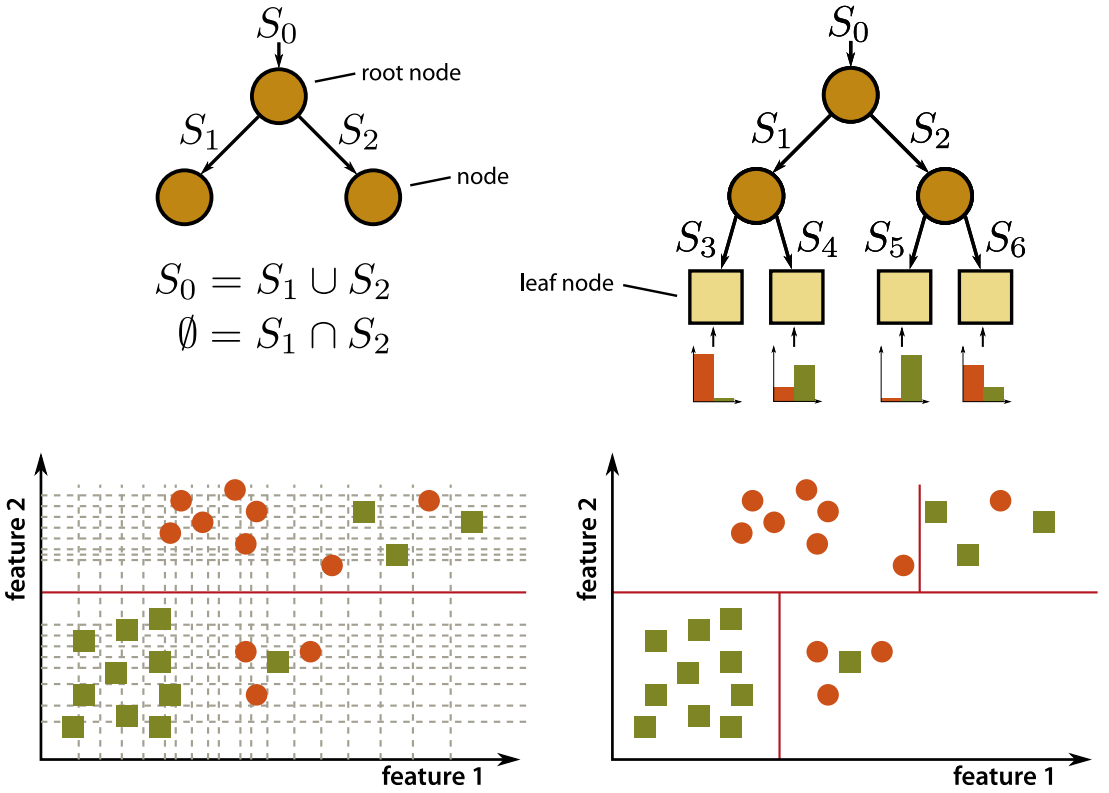


Figure 4.2: Decision tree training toy example. During the construction of a decision tree, the first split (top left) at the root node is determined by an exhaustive search over the feature space (dashed lines, bottom left). From these, the most informative is selected (drawn through red line, bottom left). At the next level, each sub-compartment is split again (bottom right), resulting finally into the full-grown tree (top right) with leaf nodes and associated class histograms. Effectively, the tree partitions the feature space according to the samples' classes.

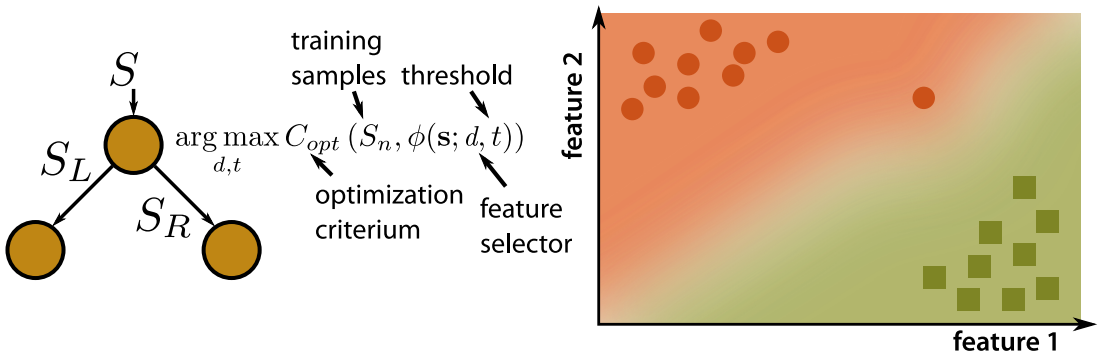


Figure 4.3: Left: Split term minimization during training to find the best split. Right: Smooth class posteriori probabilities of a forest painted onto the feature space. Note the slight bulge in the otherwise straight gradient denoting the soft outlier treatment distinctive for DFs.

for the node n over all combinations of d and t. This process is depicted in the left part of Fig. 4.3, where the incoming set S is split into S_L and S_R according to the optimal $\phi_n(\mathbf{s})$. Once $\phi_n(\mathbf{s})$ is found, its parameters d and t are stored in the node and the tree construction continues with the child nodes.

If a set S_n reaching a node n meets a pre-defined purity criterion, the new node is converted into a leaf node, terminating this branch of the tree. In each leaf node, the class histogram associated with the incoming S_n is stored (see Fig. 4.2, right).

The quality of a split C_{opt} is measures with a gain function IG, which compares the class distribution before the split with the proportionally weighted distributions after the split

$$C_{opt}(S, \phi(\mathbf{s}; d, t)) = IG(S, S_L, S_R) \tag{4.5}$$

$$IG(S, S_L, S_R) = g(S) - \frac{|S_L|}{|S|} g(S_L) - \frac{|S_R|}{|S|} g(S_R). \tag{4.6}$$

One option to quantify the sets' purity is the information gain, obtained when setting $g(S)$ to denote the Shannon entropy

$$g(x) = E(S) = -\sum_{c \in C} p(c|S) \log_2 p(c|S), \tag{4.7}$$

which is monotonically decreasing (as a set of samples cannot get more unordered than before by splitting it up) and hence the information gain IG is non-strictly positive.

Another popular option is the gini impurity, in which case

$$g(x) = 1 - \sum_{c \in C} p(c|S)^2. \tag{4.8}$$

It represents the measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. Other measures were proposed but these two split quality measures are the most popular and the only ones employed in this work.

Application When applying a classifier to a formerly unseen test sample $\mathbf{s}$, it assigns it to a class $c \in C$ according to the decision rules inferred during training. In the case of a DT, the test sample is first passed to the root node. The stored decision function $\phi_n(\mathbf{s})$ is applied to the sample $\mathbf{s}$ and it is sent to the left or right child node according to the functions return value (see Fig. 4.4). The assumption behind this is that what was found to be a good split for the training samples should be an equally good decision for the testing phase. This process is repeated node for node until the sample reaches a leaf node. Then the class membership histogram associated with the leaf node is assigned to the sample s, denoting its class membership. Thus, DTs produce posteriori class probabilities rather than crisp class memberships.

Decision forests

The exhaustive search for the optimal split and the unrestricted growth render DTs prone to overfitting and sensitive to noisy training data. To overcome these drawbacks, DFs were proposed. They constitute a family of ensemble methods, where a large number of weak classifiers (here: DTs) are trained in parallel and then vote on the final sample classification.

The deterministic nature of DTs requires the introduction of randomness to grow differing trees in the forest. Various such notions have been proposed over the years. The most popular

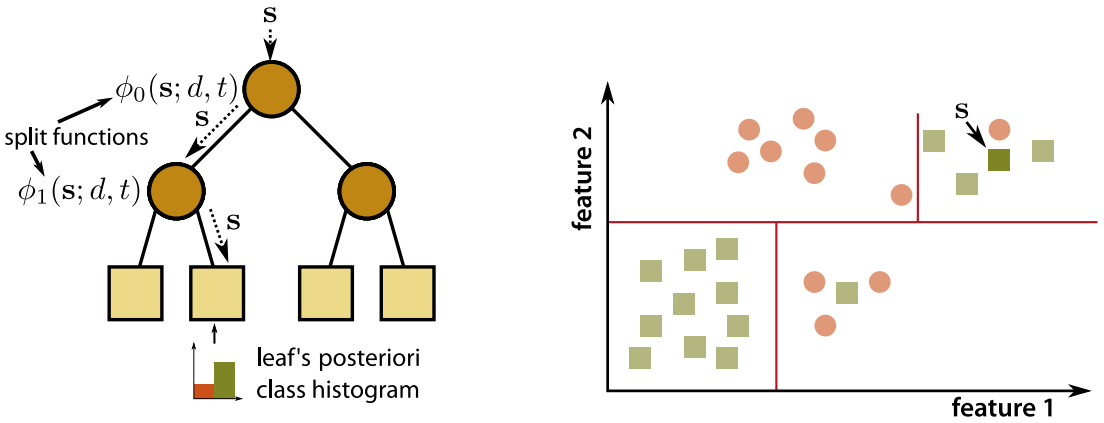


Figure 4.4: Application example of the tree learned in Fig. 4.2 to a new sample **s**. At each node, it is sent either left or right, depending on its value of the specific feature tested at that node. The sample's class membership is determined by the class histogram associated with the leaf it falls into.

ones are: (1) bagging (bootstrapping the training data) by Breiman, 2001 resulting in bagging forests (BFs); (2) random feature selection by Ho, 1998, which in combination with bagging form the classical DFs; and (3) random thresholds by Geurts *et al.*, 2006, which were termed extra forests (EFs) by their authors. The amount of randomness required depends strongly on the problem at hand. As a rule of thumb, the higher the randomness, the more general the learned classification function and the fuzzier the decision boundaries. The controlling parameters are the number of features F_{node} considered at each node when searching to the optimal split, the number T of trees in the forest and the maximum depth t_{depth} allowed. A more detailed account of the inner workings of DFs and alternative approaches can be found in Criminisi *et al.*, 2011.

An important property of DFs is that they return posteriori class probabilities rather than crisp memberships for test samples. Where a single DT subdivides the feature space into rough compartments, each of which triggers the same response (compare Fig. 4.4), the tree voting mechanism of a DF results in gradual transitions better reflecting the classifiers confidence (see Fig. 4.3, right).

DFs are known to generalize well, to be robust against noisy data, to train and classify fast, to be robust against the choice of hyperparameters and to lead to state-of-the-art results for many problems from a range of domains [Mitra *et al.*, 2014; Lombaert *et al.*, 2014; Criminisi *et al.*, 2013]. These attributes make them the ideal choice for brain lesion segmentation where the training data base might be sparse, the samples are very noisy and runtime might be an issue, such as for acute stroke.

A drawback of DFs is that they require carefully selected features which form the samples and hence the information available to the classifier. This puts them at a disadvantage with methods learning their own feature extractors, such as CNNs, in the sense that a considerable effort has to go into designing and selecting the image features. The features designed for this work are described in the next section.

4.2 Image features

A classifier is only ever as good as the information encoded in the features provided to it, as they constitute the data available for decision making. Therefore, the design of the sample features employed in a work is a very important factor. A good set of features provides the classifier with all the necessary information to reliably discriminate between the target classes while concealing information that might lead to overfitting due to underrepresented samples in the training set. E.g., knowing about the possible appearance of stroke lesions in all areas of the brain and working with a training set not covering all regions of the brain forbids the use of exact voxel localizers as features, as the classifier would learn to exclude areas that might well contain lesions in the testing data.

For this work, a number of features are deduced from knowledge about the imaging modality, the clinical decision finding process and the anatomically layout of the head. They are designed to provide the maximum brain lesion segmentation relevant information while being general enough to impede overfitting. The highly variable shape and appearance of most types of brain lesions render texture and shape features impractical, therefore the proposed approach focuses on intensity and location based information extracted at different levels.

4.2.1 Foundations

This section lays out the foundations required for the subsequent feature definitions.

Multi-spectral image Let $I : \Omega \rightarrow \mathbb{R}^{D_{Ch}}$ be a multi-spectral image of D_{Ch} channels, dimensionality D_I and image size $\mathbf{m} = [m_1, m_2 \ldots, m_{D_I}] \in \mathbb{N}^{D_I}$. Its domain $\Omega = \{\, \mathbf{x} \mid 0 \leq x_1 < m_1 \wedge 0 \leq x_2 < m_2 \wedge \ldots \wedge 0 \leq x_{D_I} < m_{D_I} \,\} \subset \mathbb{N}_0^{D_I}$ is furthermore denoted as image space. Each element $\mathbf{x} = [x_1, x_2, \ldots, x_{D_I}] \in \Omega$ of the image space denotes a voxel position. The size of a voxel is defined by $\mathbf{v} = [v_1, v_2, \ldots, v_{D_I}] \in \mathbb{R}^{D_I}$ in mm.

Abstract feature Let $\mathbf{f}(I, \mathbf{x})$ be a vector function taking an image I and a position $\mathbf{x}$ inside this image's image space as input and returning a real valued vector. This function is furthermore denoted as image feature and its return value as the feature's response. Various *types* of features can be defined, which additionally may take configuration *parameters*. Let therefore $\mathbf{f}_m(I, \mathbf{x}) := \mathbf{f}_{type}(I, \mathbf{x}; \textit{parameters}) \in \mathbb{R}^{D_m}$, i.e., the index m conveniently abbreviates a unique combination of a feature *type* and a set of *parameters*.

To maximize the information provided to the classifier, more than one feature is usually extracted. When selecting M different features, then

$$\tilde{\mathbf{F}}(I, \mathbf{x}) = \begin{bmatrix} \mathbf{f}_1(I, \mathbf{x}) \\ \mathbf{f}_2(I, \mathbf{x}) \\ \vdots \\ \mathbf{f}_M(I, \mathbf{x}) \end{bmatrix}^{\sum_1^M D_m} \tag{4.9}$$

defines the concatenated vectors of the M feature responses at an image position $\mathbf{x}$. $\tilde{\mathbf{F}}$ will be furthermore denoted as feature vector and constitutes the description of a voxel passed to the classifier.

Sample sets For image segmentation through voxel-wise classification, the responses of all chosen features are computed at each position of a test image I_{test} and collected into a sample

set S_{test} to be passed to the classifier for classification. Let

$$\mathbf{F}(I) = \left[\tilde{\mathbf{F}}(I, \mathbf{x}_1), \tilde{\mathbf{F}}(I, \mathbf{x}_2), \cdots, \tilde{\mathbf{F}}(I, \mathbf{x}_{|\Omega|})\right]^{|\Omega| \times \sum_1^M D_m} \tag{4.10}$$

be the matrix constructed from the feature vectors obtained from each position $\mathbf{x}$ in the image I. Then $S_{test} = \mathbf{F}(I_{test})$ denotes the testing samples, furthermore referred to as testing set.

Usually, multiple images are used to train a classifier. Let $\mathbf{I}_{train} = \{I_1, I_2, \ldots, I_{N_T}\}$ be a set of N_T training images. Then the matrix

$$S_{train} = [\mathbf{F}(I_1), \mathbf{F}(I_2), \cdots, \mathbf{F}(I_{N_T})] \tag{4.11}$$

constitutes the training set.

The sample sets effectively constitute the interface between the feature extraction procedure and any classifier. These definitions of training and testing set hence correspond to the already introduced notation in Sec. 4.1, which represents a classifier's perspective. The dimensionality of a sample $\mathbf{s}$ is equivalent to the sum of its constituting feature responses' dimensionalities, i.e., $D_F = \sum_1^M D_m$

Single-channel feature A multi-spectral image consists of multiple channels $I = [I_{C1}, I_{C2}, \ldots, I_{CD_{Ch}}]$, with $I_C : \Omega \to \mathbb{R}$. In most cases, each feature is extracted independently from each channel, hence

$$\mathbf{f}_m(I, \mathbf{x}) = \begin{bmatrix} \hat{\mathbf{f}}_m(I_{C1}, \mathbf{x}) \\ \hat{\mathbf{f}}_m(I_{C2}, \mathbf{x}) \\ \vdots \\ \hat{\mathbf{f}}_m(I_{CD_{Ch}}, \mathbf{x}) \end{bmatrix} \tag{4.12}$$

is the vector concatenation of the different channels' feature responses.

4.2.2 Feature definitions

Drawing from above definitions, the following features are defined, all of which are implemented as part of the MedPy library [Maier, 2016b]. Note that in this section the notation $\mathbf{f}_{type}(I, \mathbf{x}; \textit{parameters})$ respectively $\hat{\mathbf{f}}_{type}(I_C, \mathbf{x}; \textit{parameters})$ is used, as the concrete features and their configuration parameters are to be described.

Image intensity The intensity values of medical images of the brain relate to various physical, molecular and histological properties of the portrayed tissue. Albeit corrupted by noise, artifacts and other contaminations, each voxels intensity value is the base unit on which an interpretation is founded (see Fig. 4.5, left). Hence, the image's gray values are employed as low-level feature.

$$\hat{\mathbf{f}}_{int}(I_C, \mathbf{x}) = I_C(\mathbf{x}), \tag{4.13}$$

which is equivalent to

$$\mathbf{f}_{int}(I, \mathbf{x}) = I(\mathbf{x}). \tag{4.14}$$

Weighted local means Corruption with noise causes intrinsically homogeneous areas of tissue to appear with varying degrees of random intensity variations in the images. The most straight forward approach to counter this corruption is to smooth the image with a suitable filter. To this end, Gaussian blurred versions of the original images computed for different σ values are

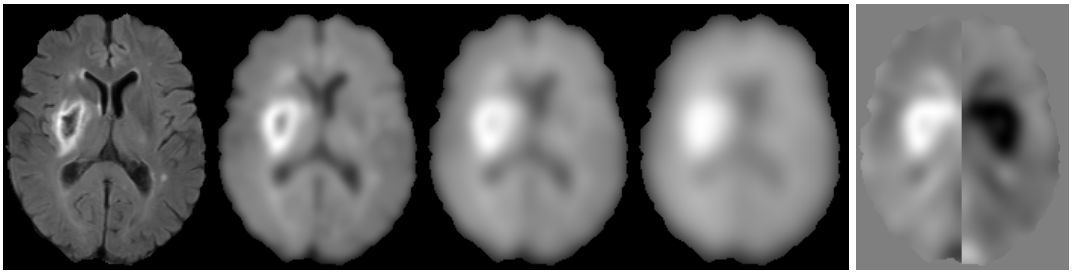


Figure 4.5: Example visualizations of some of the features employed in this thesis. From left to right: Intensity feature, weighted local means ($\sigma = \{3, 5, 7\}$ mm), and hemispheric difference ($\sigma = 5$ mm).

used as feature. Drawback of this approach is a blurring of the structures' edges (see Fig. 4.5, middle).

$$\hat{\mathbf{f}}_{wlm}(I_C, \mathbf{x}; \Sigma) = G_\Sigma(I_C, \mathbf{x}), \tag{4.15}$$

with

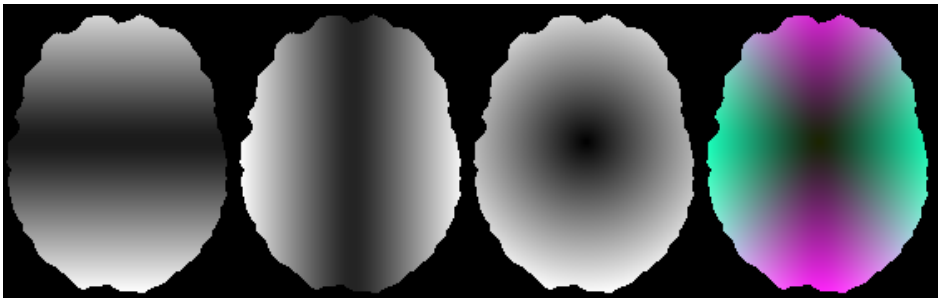
$$G_\Sigma(I_C, \mathbf{x}) = \mathcal{N}(\mathbf{x}, \Sigma) * I_C(\mathbf{x}), \tag{4.16}$$

where Σ is the co-variance matrix defining the size and shape of the Gaussian filter's kernel in mm. Note that the application of a world coordinate kernel to a discrete image space channel is solved with nearest neighbor interpolation. Choosing a diagonal form for the co-variance matrix Σ, the convolution with the multi-variate Gaussian kernel can be realized as a sequence of discrete one-dimensional convolution filters. For isotropic Gaussian kernels, only a single σ has to be supplied from which the Σ is constructed.

Center distance When performing voxel-wise classification, each voxel is treated as an independent sample void of all spatial information. It might be tempting to employ the position of each voxel as a feature to exploit, e.g., local accumulations. But such an explicit manifestation of localization entails the danger of overfitting the classifier, excluding all areas not present in the training data cases. One solution to this would be to transform all training cases to a common taxonomic space, e.g, through registration to an atlas brain. But such an approach is prone to failed registrations, errors introduced through resampling, etc. To provide a certain degree of spatial information without the danger of overfitting, the center distance is employed as feature in this work, which comprises of each voxel's euclidean distance in mm to the assumed brain center independently computed for each dimension. Since the exact center of the brain is unknown, the image center is used as approximation. By using this feature, the classifier can roughly distinguish between voxels by their position inside the brain in the sense of medial-lateral, cortical-periventricular as well as caudal/rostral-midcoronal.

$$\mathbf{f}_{cd}(I, \mathbf{x}; d) = |(\mathbf{m}/2 - \mathbf{x})\mathbf{e}_d - 1/2|\mathbf{v}\mathbf{e}_d, \tag{4.17}$$

where d denotes the dimension and $\mathbf{e}_d$ the corresponding unit vector (e.g., $\mathbf{e}_1 = [1, 0, \ldots]$) (see Fig. 4.6). Note that since $\mathbf{f}_{cd}(I, \mathbf{x}; d)$ only depends on the image size $\mathbf{m}$ and not on the intensity values, a single feature response is returned for the whole image independent of its channels. Accordingly, no hat version ($\hat{\mathbf{f}}_{cd}$) of this feature exists.



(a) X. (b) Y. (c) Z. (d) Combined (RGB).

Figure 4.6: The x, y and z-direction (i.e., $d = 1, 2, 3$) center distances for previous example case (Fig. 4.5). Note the regular repetition of colors in the (pseudo)-RGB image, which combines all three directions: The center distance feature allows only for a rough localization of each voxel's position to avoid overfitting to the training data. Refer to online version for color.

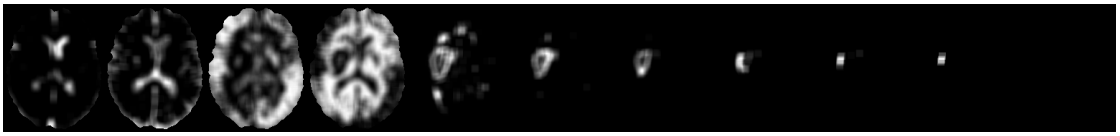


Figure 4.7: The local histogram feature of previous example case (Fig. 4.5) with $b = 11$ bins (from left to right) over a $N = 10\,\mathrm{mm}^3$ neighborhood. Note the clearly distinguishable lesion and ventricular areas. The larger the chosen neighborhood, the stronger the smoothing effect.

Local histogram Lesion segmentation aims to find continuous areas of pathological tissue. It would therefore be advantageous to equip each voxel sample with a certain amount of information about its immediate neighborhood. For this purpose, the local histogram feature was developed. It is constructed by computing the normalized histogram over a small area around a voxel with the histogram's width deduced from the whole image's intensity range and then treating each bin's value as a separate feature.

$$\hat{\mathbf{f}}_{lh}(I_C, \mathbf{x}; b, NL) = [a_1, a_2, \ldots, a_b]^T, \tag{4.18}$$

with

$$a_i = \frac{\sum_{x_n \in NL(\mathbf{x})} \mathbf{1}_{[l_i, r_i]}(I_C(\mathbf{x}_n))}{|NL(\mathbf{x})|}, \tag{4.19}$$

where $NL(\mathbf{x})$ is a local neighborhood of $\mathbf{x}$ that returns the voxel positions in a predefined area around $\mathbf{x}$, usually a rectangular region. The scalar b determines the number of histogram bins and $[l_i, r_i]$ denotes the interval borders of the i-th bin. The borders of the bins are equidistantly distributed over the range $[min(I_C), max(I_C)]$. See Fig. 4.7 for an example.

Hemispheric difference The human brain displays a certain degree of hemispheric symmetry. Since most brain lesions do not appear symmetrically, a hemispheric difference feature is employed. After smoothing the image with a Gaussian filter at a relatively large σ (e.g., 7 mm) kernel size to account for small anatomical differences, the medial longitudinal fissure is approximated by the midsaggital image plane and each hemisphere subtracted from its counterpart, resulting in a difference image.

$$\hat{\mathbf{f}}_{hd}(I_C, \mathbf{x}; \Sigma, d) = G_\Sigma(I_C, \mathbf{x}) - G_\Sigma(I_C, \tilde{\mathbf{x}}), \tag{4.20}$$

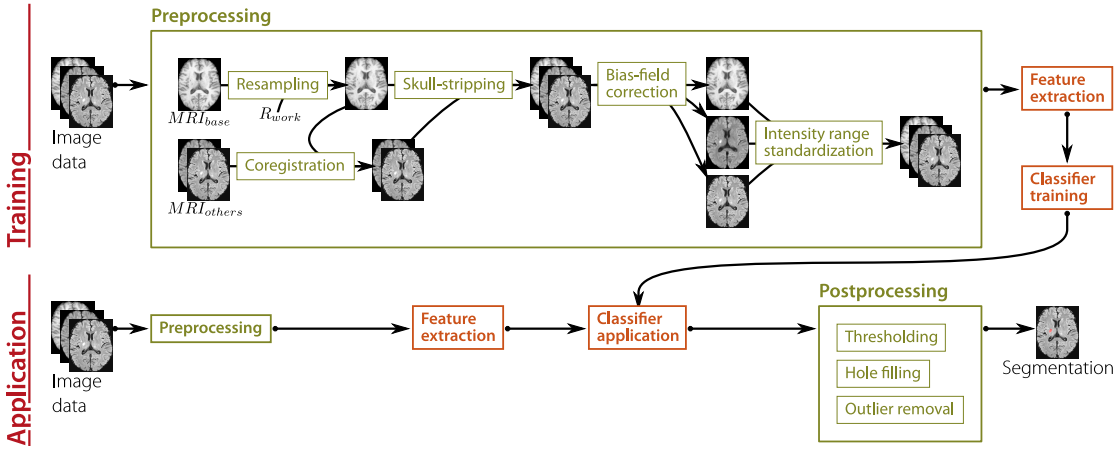


Figure 4.8: Components of the offline training and the online application phases.

with

$$\tilde{\mathbf{x}} = [x_1, \ldots, x_{d-1}, m_d - x_d - 1, x_{d+1}, \ldots, x_{D_I}], \tag{4.21}$$

where G_Σ is defined as above in Eq. 4.16 and the scalar d denotes the axial dimension of the brain image. Thus, $\tilde{\mathbf{x}}$ is the position of $\mathbf{x}$ mirrored along the dimension d, which effectively constitutes a flip along the approximated medial longitudinal fissure.

Effectively, this feature compares the means over large regions between the hemispheres and can hence provide a rough estimation of larger lesions' locations. Using a smaller smoothing filter to enable it to reveal smaller lesions or lesion borders would be overshadowed by the anatomical differences between the hemispheres. Since such large smoothing filters are applied, the center-line approximation through the midsaggital image plane was found to be sufficient (see Fig. 4.5, right). In the special case of MRI perfusion maps (see Sec. 2.2), the hemispheric difference feature serves to render the quantitative maps relative and hence more comparable.

Most presented features are brain lesion segmentation specific, exploiting the anatomical layout and lesion specific idiosyncrasies to reveal subtle intensity anomalies. But none of them is pathology specific, i.e., no assumption other than that a lesion is a hypo- or hyperintense area in the normal brain tissue is made. This allows for a ready application to different pathologies such as MS, stroke, tumors, etc.

4.3 Segmentation framework

As most machine learning methods, the framework is dividable into an offline training and an online application phase (Fig. 4.8). This section presents the remaining components of the overall segmentation framework and how they work together. Particularly, the learning-based intensity range standardization and the training set subsampling strategy are described in detail, as they constitute own contributions.

4.3.1 Preprocessing

The images passed to the framework have to be assumed to originate from clinical routine, i.e., they suffer all the machine, sequence and scanning related problems described in Chapter 2.

Thus, a number of preprocessing steps should be performed to correct the larger deviations, with the aim to render the different cases comparable and processable by the classifier.

Presumed input of the preprocessing module is a single case containing different MRI sequences of a single patient recorded during a single (or at least two subsequent) session.

Resampling and coregistration No application intended to process clinical data can make any assumptions about image resolution or orientation. But the voxel-wise classification scheme requires a fixed working resolution R_{work} and a common stereotactic space for all MRI sequences of the same case. To this end, a chosen base sequence MRI_{base} is resampled to the working resolution with third-order B-spline interpolation. Subsequently, all remaining sequences are rigidly coregistered to the resampled base sequence. A robust and reliable registration with sufficient accuracy is obtained using the elastix toolbox [Klein *et al.*, 2010; Shamonin *et al.*, 2013], which is based on the ITK framework. A rigid four-resolutions approach with advanced mattes mutual information and final third-order B-spline interpolation is employed. Only a rigid registration step is required, as all sequences display the same deformation-free anatomical region. Any anatomical sequence (e.g., T1, T2, FLAIR) of sufficiently high resolution can be chosen as MRI_{base}. The working resolution R_{work} should be set low enough to minimize the upsampling of the remaining sequences and high enough to provide sufficient detail for the segmentation. After execution of this step, all sequences of the case have the same resolution R_{work} and lie in a common stereotactic space.

Skull-stripping To speed up the method and to avoid false-positives, it is desirable to remove the skull and neck from the MRI scans [Maier *et al.*, 2015e]. Various publications have proposed solutions for skull-stripping, of which FSL-BET2 [Jenkinson *et al.*, 2005; Smith, 2002] was found to be reliable and easy to use. It works best on anatomical scans, i.e., T1, T2 and FLAIR, in this order [Maier *et al.*, 2015d]. Once the brain mask is obtained, it is simply applied to the other already coregistered sequences.

Bias-field correction As described in Sec. 2.4, MRI images can suffer from inhomogeneity fields. The CMTK mrbias tool, which performs intensity bias-field correction based on minimization of image entropy [Likar *et al.*, 2001], is therefore applied to each image independently.

Intensity range standardization The intensity values of MRI sequences do not relate directly to physical properties as does the Hounsfield scale of computed tomography (CT). In fact, even the intensity values of the same tissue scanned with the same settings on the same machine in two subsequent sessions do not necessarily correspond (see Sec. 2.3). Accordingly, the image's histograms have to be shifted and stretched suitably to identify the tissues by their intensity values. Under the assumption that the distribution of the intensity values is at least similar between images of the same sequence, different intensity range standardization methods can be applied. The most straightforward approach is setting a zero mean and unit variance, but this can cause unwanted distortions in the presence of pathologies and information loss due to histogram compression. A slightly more advanced method is the histogram matching method, where the target image's histogram is normalized based on a reference image. This framework implements a population based version of histogram matching originally proposed by Nyúl *et al.*, 2000. This training based method first learns a common intensity range space from all training images' intensity percentiles and subsequently transforms them to that space using piece-wise linear transformation between the anchor points. The advantages are that, first, no reference image has to be selected and, second, information preservation can be guaranteed. Such a range model

is learned once for each MRI sequence and formerly unseen images are transformed accordingly during the application phase. It should be noted that for some, usually function related, sequences such as PWI maps, intensity range standardization is not necessary, as their gray-values relate directly to quantitative properties. The implementation created for this work is available as part of the MedPy library [Maier, 2016b].

4.3.2 Classifier training

After the preprocessing steps, the MRI images are ready to be fed to the classifier for training. Many possible confusions, such as image artifacts, patient movement or incomplete scans (see Sec. 2.4) are still present and have to be handled through machine learning. This makes it necessary to train the classifier on a sufficiently large and diverse training set.

Each training case consists of a multi-spectral image formed by the combined MRI sequences and an associated manual ground truth. The selected features are extracted from the images and form the training set $S_{train} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{|S|}\}$ as defined in Eq. 4.11. From the expert segmentations, the associated ground truth set $C_{train} = \{c_1, c_2, \ldots, c_{|S|}\}$ is constructed. Hence, each multi-dimensional sample $\mathbf{s} = \{s_1, s_2, \ldots, s_{D_F}\}$ forms a description of a single voxel with an associated class $c \in C$.

Training set subsampling 3D MRI images are large and even at low resolution and with a moderate amount of training cases, tens or hundreds of millions of training samples become available. The training set size is usually directly proportional to the classifier's training time. For most applications, a representative subset can help to lower the required training time without any loss in classification accuracy.

I therefore propose a dedicated stratified sampling scheme to subsample a representative subset of the training set. Since pathological diversity is represented by the different cases rather than by different regions of a single case, care is taken that each case contributes equally. The fore- and background-voxels inside a single case, on the other hand, often represent redundant information.

Considering a number of N_S desired training samples and N_T available training cases, then N_S/N_T are drawn with uniform stratified random sampling from each case, resulting in the subsampled training set $\tilde{S}_{train} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{N_S}\}$ with $N_S << |S|$. Stratified random sampling means that the case-specific class ratio is kept intact, i.e., usually the background label is over-represented.

This takes into account the fact that the lesion samples are assumed to be comparatively homogeneous and therefore separably, while the background contains numerous anatomical structure which should be sufficiently represented in the training data. Preliminary experiments have shown this approach to lead to superior results than the usually taken balanced class representation [Maier *et al.*, 2015e; Maier *et al.*, 2015d].

4.3.3 Postprocessing

Postprocessing describes the modifications made to the resulting segmentation mask. It can include erosion or dilation to counter the methods tendency to over- respectively undersegment; morphological opening or closing to smooth the result; and hole closing and/or small object removal to avoid anatomically implausible results. The taken measures are highly application dependent and can be optimized in a leave-one-out fashion on the training data. In this framework, they are used in various combinations and the concrete settings chosen are described in the evaluation section.

4.3.4 Evaluation metrics

During the evaluation, a number of metrics are employed that are suited to reveal different aspects of the resulting segmentation. The most important ones are (1) Dice's coefficient (DC), which describes the volume overlap between two segmentations and is sensitive to the lesion size; (2) the average symetric surface distance (ASSD), which denotes the average surface distance between two segmentations; and (3) the Hausdorff distance (HD), which is a measure of the maximum surface distance and hence especially sensitive to outliers. Additionally, precision (or positive prediction value (PPV)) and recall (or sensitivity (SE)) values are reported to assess over- and undersegmentation, respectively.

The DC is defined as

$$DC = \frac{2|A \cap B|}{|A| + |B|} \tag{4.22}$$

with A and B denoting the set of all voxels of ground truth and segmentation respectively. To compute the ASSD, first the average surface distance (ASD), a directed measure, is defined as

$$ASD(A_S, B_S) = \frac{\sum_{a \in A_S} min_{b \in B_S} d(a, b)}{|A_S|}, \tag{4.23}$$

and then averaged over both directions to obtain the ASSD

$$ASSD(A_S, B_S) = \frac{ASD(A_S, B_S) + ASD(B_S, A_S)}{2}, \tag{4.24}$$

Here A_S and B_S denote the surface voxels of ground truth and segmentation respectively. Similar, the HD is defined as the maximum of all surface distances with

$$HD(A_S, B_S) = \max\{\max_{a \in A_S} \min_{b \in B_S} d(a, b), \max_{b \in B_S} \min_{a \in A_S} d(b, a)\}. \tag{4.25}$$

The distance measure $d(\cdot)$ employed in both cases is the Euclidean distance, computed taking the voxel size into account. Finally, precision and recall are calculated as

$$precision, PPV = \frac{|A \cap B|}{|B|} \tag{4.26}$$

and

$$recall, SE = \frac{|A \cap B|}{|A|} \tag{4.27}$$

4.4 Evaluation

The proposed framework is composed of many components, all of which may have an influence on the final segmentation results. Their individual contributions are investigated in the first part of the evaluation, where all features, hyperparameters and component choices are thoroughly examined.

At the heart of any machine learning framework stands the classifier. The second part of the evaluation is therefore dedicated to compare the DFs against a number of alternatives, confirming my choice as well as the DFs general popularity.

My framework aspires to be applicable to many brain lesion segmentation tasks, largely independent of the particular pathology. To substantiate this claim, the last evaluation sections 4.5 presents results obtained on four medical image processing challenges, i.e., public segmentation

benchmarks with hidden ground truth and third party evaluation. The framework's placement in the rankings reveals its standing among the current state-of-the-art approaches, obtained in a direct and fair comparison.

The results presented in these sections were previously published in a range of articles and proceedings [Maier *et al.*, 2015d; Maier *et al.*, 2015f; Maier *et al.*, 2015h; Maier *et al.*, 2015g; Maier *et al.*, 2015b; Maier *et al.*, 2015e; Maier *et al.*, 2016; Maier *et al.*, 2017].

4.4.1 Hyperparameter analysis

Classifier (Extra Forest)	
T	200
F_{node}	$\sqrt{F}$
t_{depth}	unlimited
C_{opt}	Entropy
Subsampling	
N_S	$250,000$
Features	
int	
wlm	$\sigma = 3, 5, 7$ mm
cd	$d = x, y, z$
lh	$b = 11, NL = 5, 10, 15$ mm^3

Preprocessing	
Resampling	$R_{work} = 3$ mm^3
Coregistration	$MRI_{base} = FLAIR$
Skull-stripping	$MRI_{skull} = FLAIR$
Bias-field	yes
Intensity range std	yes
Postprocessing	
Thresholding	0.5
Object threshold	1.5 ml
Hole filling	yes

Table 4.1: Hyperparameter analysis experimental configuration.

The presented framework relies on a number of components working smoothly together, most of which allow for fine tuning by changing their parameters. This evaluation section presents an exhaustive hyperparameter analysis conducted on a representative set of stroke cases in leave-one-out fashion, where one parameter is varied and the others are kept at their default values as denoted in Table 4.1. The obtained results show the method's robustness against its parameter settings and serve as guidance for future improvements.

The image database consists of 35 sub-acute ischemic stroke cases acquired routinely for two clinical studies on spatial neglect [Machner *et al.*, 2014; Machner *et al.*, 2012; Gablentz, 2012]. Each dataset was manually segmented (as filled volume) in axial FLAIR images by an observer with several years of dedicated experience in stroke imaging. If required and available, other MRI sequences were used to resolve ambiguities. In case of a previous ischemic stroke history, only the newest ischemic stroke lesions were segmented. Hemorrhages were only included in the manual lesion segmentations if completely encircled by ischemic tissue. From these cases, two sets are formed: First, a mono-spectral set consisting of all 35 cases' FLAIR sequences, which are considered to be the most sensitive for sub-acute stroke detection (see Sec. 3.1.2). And second, a multi-spectral set containing the subset of 14 cases for which additionally the DWI, ADC, T1 and T2 sequences were available. More information about the patients, lesion characteristics, imaging parameters, and image quality are detailed in Maier *et al.*, 2015e, where parts of the results were also published previously.

Results

Mono-spectral Applied to the 35 mono-spectral cases, the proposed ischemic lesion segmentation method led to the segmentation results presented in the first row of Table 4.3, which serve as reference values for further experiments.

	DC[0,1]	HD(mm)	ASSD(mm)
Skull-stripping→Bias correction→Int. range std.→Postprocessing	0.67	28	4.82
Skull-stripping→Bias correction→Int. range std.→~~Postprocessing~~	0.65	39	5.18
~~Skull-stripping~~→Bias correction→Int. range std.→Postprocessing	0.57	*34	*5.60
Skull-stripping→~~Bias correction~~→Int. range std.→Postprocessing	0.63	32	6.70
Skull-stripping→Bias correction→~~Int. range std.~~→Postprocessing	0.53	**38	**7.12
Skull-stripping→~~Bias correction~~→~~Int. range std.~~→Postprocessing	0.52	**36	**7.17
~~Skull-stripping~~→~~Bias correction~~→~~Int. range std.~~→Postprocessing	0.49	**36	**6.68

Table 4.3: Average results obtained using the default mono-spectral configuration for different pre- and postprocessing combinations at a threshold of 0.5. Note that some settings lead to cases with empty results for which some evaluation measured could not be computed. These are accordingly marked with a star ($*$) for every failed segmentation not contributing to the displayed value, which are therefore not directly comparable to the others.

Multi-spectral Different MRI sequences can contain complementary information about the stroke lesions, e.g., hyperintensities in the DWI sequence can help to differentiate between stroke and other white matter lesions (WMLs) (see Sec. 3.1.2 for details). To assess the presumed superiority of multi- over mono-spectral approaches for stroke lesion segmentation [Agam *et al.*, 2006; Forbes *et al.*, 2010; Mitra *et al.*, 2014] in the sub-acute phase, EF classifiers with all possible sequence combination of FLAIR, DWI, ADC, T1 and T2 are trained. From each sequence the complete range of features described in Table 4.1 was extracted. The results are shown in Fig 4.9. FLAIR is the most discriminative sequence, while the T1 provides important complimentary information. To achieve optimal results, multi-spectral data should be employed and the presented results can be used as guideline to identify the most beneficial sequences for stroke lesion segmentation. If available, additional sequences can be readily added without having to apprehend a drop in classifier performance.

Feature combinations Although forest ensemble methods are not known to be sensitive to redundant or irrelevant features, it is useful to perform a feature analysis to improve the model interpretability and possibly reduce the feature vector's size. Fig. 4.10 displays the DC results obtained with all possible combinations of the feature types using the mono-spectral set-up. The most prominent feature is the proposed local histogram with the center distance as its most informative complement. To achieve optimal results, all features should be combined.

Influence of the postprocessing Omitting the postprocessing, the mean results presented in Table 4.3, second row, are obtained. Its main function is the removal of small outliers, which leads to improved HD and ASSD values, while the DC does not change greatly.

Influence of the preprocessing Preprocessing the images is considered an important part of any segmentation attempt in MRI images. Disabling different components of the preprocessing pipeline leads to the results presented in Table 4.3, third to seventh row. While the different preprocessing measures have different amounts of influence on the evaluation scores and affect each of the three metrics differently, only the combination of all three measures results in the highest score, independent of the metric regarded. This shows the importance of preprocessing when dealing with MRI images and confirms the methods chosen for this pipeline.

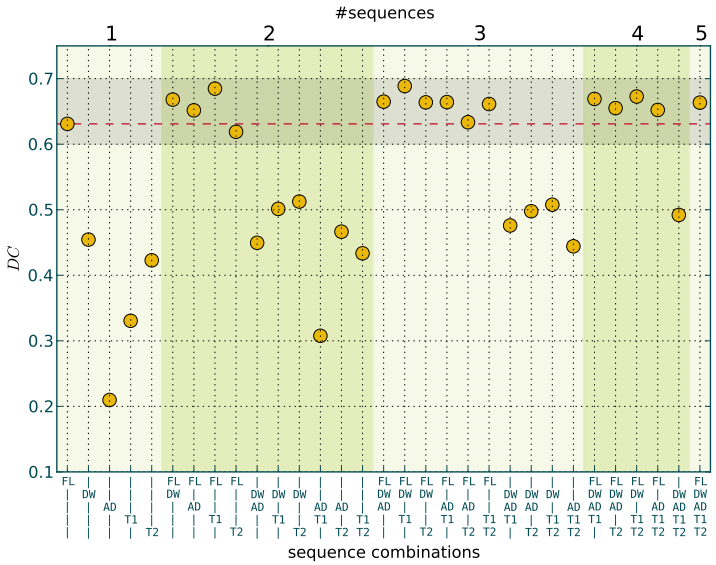


Figure 4.9: Average DC values obtained for all possible sequence combinations. The lower x-axis denotes the combination of sequences used in the corresponding results, where `FL`=FLAIR, `DW`=DWI and `AD`=ADC, while the upper x-axis as well as the alternating background colors denote the changes in the number of sequences used. With the mono-spectral data, the method obtained a DC of ≈ 0.63 over the 14 cases, marked by the horizontal dotted line in the graph. The shaded horizontal band in the upper end of the graph is a visual aid to highlight the group of best combination, which all contain the FLAIR sequence.

Training set subsampling To investigate the variance introduced by the randomized training set subsampling (Sec. 4.3.2) and to examine the robustness of the chosen sampling method, the process of sampling, training and evaluation is repeated 10 times. The low standard deviations observed over all runs (DC-standard deviation (STD) 0.00, HD-STD 1 mm and ASSD-STD 0.08 mm) show the sampling scheme to be robust. Furthermore, the size of the sampled training set was varied. Fig. 4.11 displays the resulting mean DC scores. The similar DC values for different training set sizes supports the claim that the sampling scheme captures all important variations while speeding up the training through the removal of redundant data.

Forest parameters The choice of the forest parameters is known to influence the classification outcome [Criminisi *et al.*, 2011; Breiman *et al.*, 1984]. Although it has been shown that EFs often yield good results over a large range of parameters [Geurts *et al.*, 2006], the optimal choice still depends on the classification problem at hand. It is therefore worthwhile to investigate each parameters influence on the classification accuracy. The results can furthermore provide an insight into the working of the obtained classifier as well as disclosing properties of the classification problem at hand. For these experiments a single parameter is varied over a large range while keeping all others at their default values. The results are summarized in Fig. 4.12.

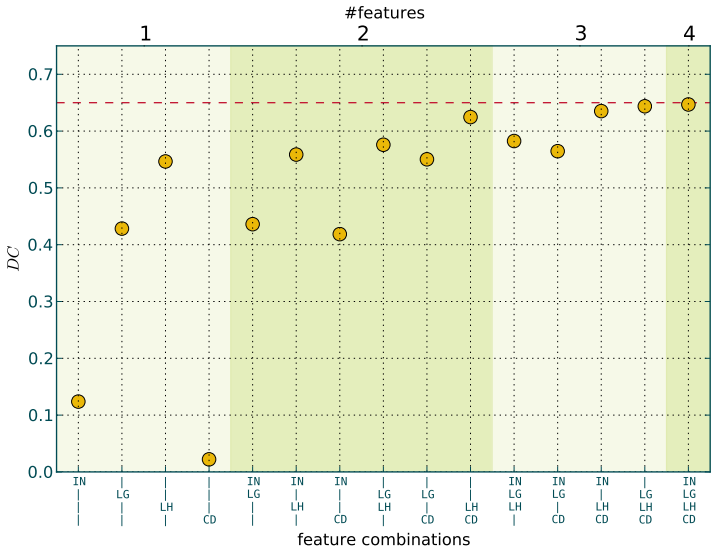


Figure 4.10: Average DC values obtained for all possible feature combinations. The lower x-axis denotes the combination of feature types used in the corresponding results, where `IN`=Intensity, `LG`=Weighted local mean, `LH`=Local histogram and `CD`=center distance, while the upper x-axis as well as the alternating background colors denote the changes in the number of features used. With the default settings a DC of ≈ 0.65 is obtained, marked by the horizontal dotted line in the graph.

Conclusion

The main conclusion that can be drawn from the hyperparameter analysis results is that DFs are very robust against the choice of their parameters and, at least for the problem at hand, show no tendency to overfit (Fig. 4.12b and Fig. 4.12d). The default parameters prove to be well chosen and only increasing the number of features to consider at each node during training would lead to a slightly increased classification accuracy (Fig. 4.12c), which is consistent with the original definition of the EFs by Geurts *et al.*, 2006. The experience gather during this analysis will be employed throughout this thesis.

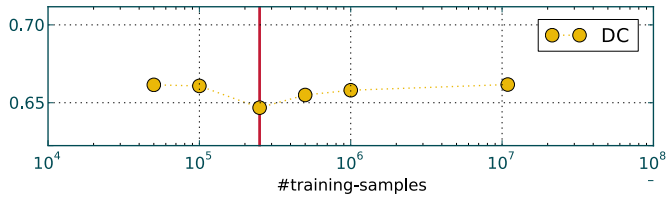
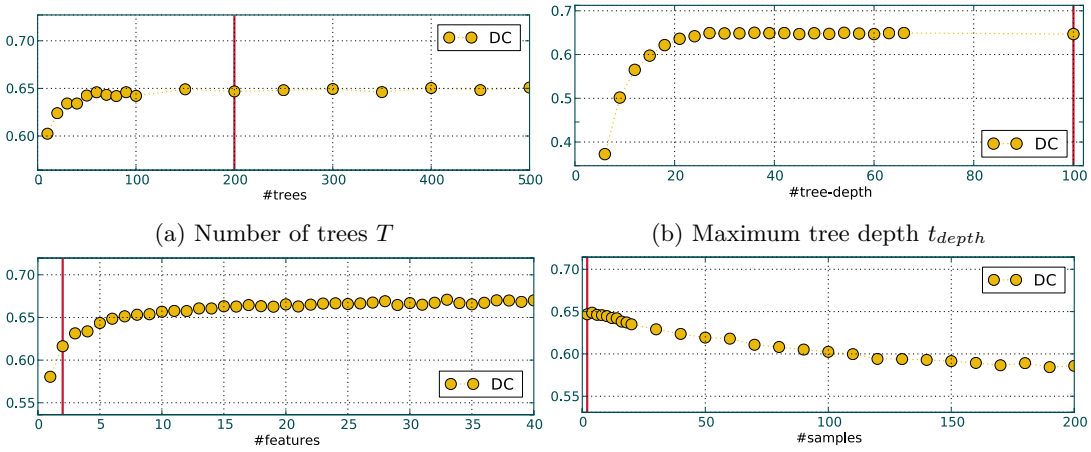


Figure 4.11: Influence of the training set size on the segmentation quality. The vertical line denotes the default value as used in the hyperparameter analysis.



(a) Number of trees T

(b) Maximum tree depth t_{depth}

(c) Feature considered at each node F_{node}

(d) Minimal number of samples in leaf

Figure 4.12: Influence of various forest parameters on the classification quality. The vertical lines denote the default values for each parameter. Note the different DC/x-axis scales.

4.4.2 Classifier comparison

Subsampling		
	N_S	$500,000$
Features		
	int	
	wlm	$\sigma = 3, 5, 7$ mm
	cd	$d = x, y, z$
	lh	$b = 11, NL = 5, 10, 15$ mm

Preprocessing		
	Resampling	$R_{work} = 3$ mm^3
	Coregistration	$MRI_{base} = FLAIR$
	Skull-stripping	$MRI_{skull} = FLAIR$
	Bias-field	yes
	Intensity range std	yes
Postprocessing		
	Thresholding	0.4
	Closing	1 mm^3
	Object threshold	1.5 ml
	Hole filling	yes

Table 4.4: Classifier comparison experimental configuration.

The proposed framework employs forests as classifiers, but other options are available and might be a more suitable choice for brain lesion segmentation. To substantiate the decision for the DF as classifier, a range of popular classifiers is plugged into the framework and all are evaluated on a common dataset. The results presented in this section were previously published in Maier *et al.*, 2015d.

Materials and methods

The mono-spectral sub-acute stroke cases employed in this classifier comparison are the same 35 FLAIR cases already introduced for the hyperparameter analysis in Sec. 4.4.1.

A total of 9 classification methods, whose function and set-up are described in this section, are evaluated and compared with each other in this study. If not noted otherwise, no effort has been undertaken to optimize their parameters for this segmentation problem. Instead, they were executed with their best-practice parameter values, i.e., the default parameters of the *scikit-learn* [Pedregosa *et al.*, 2011] toolkit.

Decision Forests The framework's default classifier is described in Sec. 4.1 in detail. For this application, $T = 100$ trees with a maximum depth of $t_{depth} = 20$ are trained.

Extra Forests As a variant of the DF, the EFs are equally described in Sec. 4.1 and trained with the same parameter settings.

Gaussian Naive Bayes Gaussian Naive Bayes (GNB) approach the classification task with the "naive" assumption of independence between each pair of features and the likelihood of the features to be Gaussian. Even though GNB oversimplifies the reality, they were found to perform surprising well in a number of real-world problems. Furthermore, GNB classifiers require only a small amount of training data, are parameter-free and train very fast. They are well researched, both from a theoretical [Zhang, 2004] and empirical [Rish, 2001] point of view.

k-Nearest-Neighbors The supervised k-Nearest-Neighbors (kNN) [Cover *et al.*, 1967] approach classifies testing samples by transferring the majority label of the k nearest training neighbors to the corresponding test case. The Euclidean distance is most commonly employed and also the choice in this study. KNN classifiers do not generalize from the training set, but simply store the training data. Similar as GNB classifiers, kNN models were found to perform well for many real-world classification problems. Besides the the distance metric, the choice of k is crucial. Higher values for k reduce the influence of noise, whereas lower values lead to more distinct class boundaries. As an additional parameter, the training samples' votes can be weighted by their distance. However, this feature was not used in this study to keep the method as simple as possible.

Generalized Linear Models In a Generalized Linear Model, tissue infarction probability can, for example, be represented by the logistic function as typically used for biological applications:

$$F(t) = \frac{e^t}{e^t + 1} \tag{4.28}$$

with t being a linear function of the input sample $\mathbf{s}$,

$$t = \beta_0 + \beta_1 s_1 + \ldots + \beta_{D_F} s_{D_F} \tag{4.29}$$

The main advantages of the algorithm are its simplicity, the comparably high speed for training and testing, and the possibility to investigate the effects of the multiple input parameters on the outcome probability in terms of the β parameters. However, logistic regression models are also known to be unsuitable for inherently nonlinear problems.

Gradient Boosting classifier Gradient Boosting (GB) classifiers describe a generalized boosting method for arbitrary differentiable loss functions. In the case of the GB classifier implementation used in this study, this method is similar to DFs in the sense that a large number of DTs are trained. These weak classifiers are optimized at each stage to fit the negative gradient of the deviance (twice binomial negative log-likelihood) loss function, i.e., the steepest gradient descent. The learning rate regularization strategy proposed by Friedman, 2000 is employed in this work, but not the bootstrapping strategy described in Friedman, 2002, which would result in stochastic GB. GB classifiers are known to achieve a high predictive power and to be robust against outliers in output space. A severe drawback is their sequential nature, which leads to long training times. They can be considered a predecessor to DFs. GB classifiers require the definition of a number of hyperparameters. In general, there is a trade-off between the learning rate and the number of estimators, while the maximum tree depth should be kept small to allow for faster training. For this comparison, 100 trees with a maximum depth of 20 are trained.

AdaBoost AdaBoost [Freund *et al.*, 1997] represents another well-known boosting method, where a sequence of weak learners is fitted to repeatedly modified versions of the training data. A weighted majority vote at application time is used to achieve the final class prediction. In contrast to DFs, which utilize bootstrapping for this purpose, AdaBoost assigns individual weights to the training samples: The first weak classifier is trained on the uniformly weighted samples, then the weights are iteratively increased for training samples wrongly predicted in previous steps. Hence, difficult and complex training samples obtain a greater weight for later weak classifiers. AdaBoost is often considered as one of the best out-of-the-box classifiers. Nonetheless, it is also known to be sensitive to noise and outliers, as it explicitly increases their influence. The implementation used in this study employs decision tree stumps as weak classifiers. Important additional parameters are the number of estimators and the learning rate, which penalizes later classifiers. The first value was set to 100, the latter kept at its default value of 1.0.

Convolutional Neural Networks In recent benchmarks, neural networks present the winning solutions for various computer vision tasks like object detection, street number recognition and mitosis detection [Szegedy *et al.*, 2015; Goodfellow *et al.*, 2013; Ciresan *et al.*, 2013]. CNNs [LeCun *et al.*, 1989] are a special form of neural networks that transform the input by repeated steps of convolution followed by pooling. The output of this feature extraction step forms the input to a classical fully connected neural network. The whole network including the kernels of the convolution is trained using backpropagation.

By training their own feature extractors, CNNs can be easily applied to new problems. Their classification speed is comparable to other methods. However, their training time is considerably longer. Furthermore, the network's architecture and multiple hyperparameters need to be chosen carefully to obtain good results. In order to achieve a good generalization, a high training sample count, the convolutional architecture [Fukushima, 1980] and dropout layers [Hinton *et al.*, 2012] are recommended.

Contrary to the other methods presented in this section, the CNN uses the raw image input instead of the manually designed features. Therefore, 10^7 overlapping patches of $37 \times 37 \times 3$ voxels are sampled from the training data in a uniform random manner and labeled according to the center voxel's classification in the ground truth. For the experiments, the Caffe [Jia *et al.*, 2014] framework is used. The network is built with three convolution steps, each with rectified linear activation units (RELUs) [Nair *et al.*, 2010] and pooling, followed by one fully connected layer with RELUs and one with softmax activation. The precise network architecture is described in Table 4.6. Learning was performed in a fully supervised manner using a batch size of 500, a learning rate of 0.0001, a weight decay of 0.004, and a momentum of 0.9.

Layer	Type	Maps and neurones	Kernel size
0	input	3 maps of 37× 37 neurons	
1	convolution	100 maps of 35× 35 neurons	3× 3
2	pooling	100 maps of 18× 18 neurons	2× 2
3	convolution	150 maps of 16× 16 neurons	3× 3
4	pooling	150 maps of 8× 8 neurons	2× 2
5	convolution	150 maps of 6× 6 neurons	3× 3
6	pooling	150 maps of 3× 3 neurons	2× 2
7	fully connected	300 neurons	1× 1
8	fully connected	2 neurons	1× 1

Table 4.6: Convolutional neural network architecture. The input is processed from the top to the bottom, where the two output neurons each represent one class. Rectified linear activation units are used after each convolution and after the first fully connected layer. The two final neurons are activated by a softmax function and can be interpreted as the probability of a particular input to belong to the respective class.

Results

For the experiments, all methods were trained and evaluated with the leave-one-out evaluation schema. The results obtained for all classifiers are displayed in Table 4.7. The best-performing method for each evaluation measure is marked bold. Significant differences to this best-performing method computed with the Student's paired t-test are marked with a star ($*$) for a confidence interval of 95% ($p < 0.05$) and two stars ($**$) for a confidence interval of 99% ($p < 0.01$). Nominal p-values are reported without correction for multiplicity. It should be noted that some methods failed completely for certain cases (i.e., achieved a DC of 0). None of the best performing methods were among these. The corresponding cases were excluded from the calculation of the average values for all methods to enable a direct and fair comparison.

The DF, their EF variant and the CNN performed best, confirming the choice of classifier and establishing an idea of possible alternatives. The DF vs. CNN competition will play a role in the next section and receives a longer discussion in App. B.

Classifier	**DM [0, 1]**	**HD (mm)**	**ASSD (mm)**	**Prec. [0,1]**	**Rec. [0,1]**	**Cases**	**Traintime**
100 Nearest Neighbors	0.54±0.20**	36.52±22.4	07.07±4.25**	0.82	0.45	34/35	5s
10 Nearest Neighbors	0.56±0.20**	36.47±25.1	06.58±4.01*	0.82	0.46	35/35	5s
5 Nearest Neighbors	0.58±0.18**	39.72±27.4*	06.80±4.35*	0.79	0.51	35/35	5s
AdaBoost	0.60±0.19*	39.28±27.3*	07.42±6.77*	0.70	0.61	35/35	7m
Extra Forest	0.64±0.19**	29.49±18.5	05.29±3.94	0.84	0.57	35/35	3m
Gaussian Naive Bayes	0.48±0.22**	69.86±26.7**	14.82±8.16**	0.44	0.78	35/35	1s
Generalized Linear Model	0.44±0.25**	38.77±21.3*	08.54±5.76**	0.87	0.34	32/35	2m
Gradient Boosting	0.63±0.18**	32.72±23.2	05.93±5.28	0.72	0.62	35/35	12h
Decision Forest	**0.67**±0.18	**28.16**±20.7	**04.89**±3.63	0.82	0.62	35/35	6m
Convolutional Neural Network	0.67±0.18	29.64±24.6	05.04±5.28	0.77	0.64	35/35	2h

Table 4.7: Classifier comparison results. Average computed over 31/35 cases, stars denote significant difference to best-performing method (in **bold**) with $** = p < 0.01$ and $* = p < 0.05$, train-times given for a single training round, value after ± denotes the standard deviation

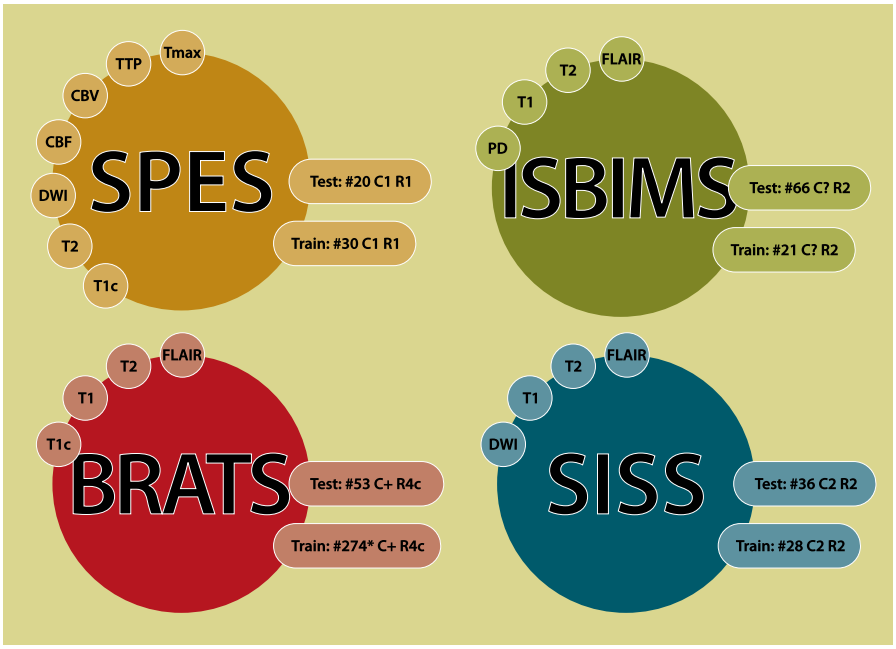


Figure 4.13: Summary of the challenges' configurations, including the provided MRI sequences, training and testing set sizes (#), number of medical centers (C) and number of raters (R). A '+' denotes two or more centers with the exact number unknown. A '?' denotes an unknown number of centers. A '*c*' denotes consensus segmentation of the raters.

4.5 Challenge evaluations

In the previous evaluation section, the proposed framework was thoroughly examined. As a result, the choice of DFs as classifier could be justified in a comparison study. Furthermore, all proposed features were analyzed for their individual and combined contributions. Then, all individual framework components were checked and compared to popular alternatives. Finally, the proposed framework was shown to be robust against the choice of its hyperparameters. But all of these results were obtained on a private dataset in a leave-one-out evaluation fashion and, hence, serve little to shed light on the method's actual performance in clinical situations.

Therefore, this section presents the results obtained in four medical image processing challenges: SPES [Maier *et al.*, 2017]), ISBIMS [Carass *et al.*, 2016]), BRATS [Menze *et al.*, 2015a]), and SISS [Maier *et al.*, 2017]). Their main attributes are summarized in Fig. 4.13. For all of these, a training and a testing set of carefully selected and diverse cases representative of the problem at hand is provided. Numerous research groups participate with their self-tuned state-of-the-art approaches and a set of suitable evaluation measures are employed to evaluate each of them against a hidden testing set ground truth. This enables the community to compare the participating methods under different aspects as revealed by the different evaluation measures, to recognize emerging trends and new advances, and to fairly evaluate their own propositions. Usually, the organizers additionally compile a ranking of all participating methods, which ideally reflects the methods' performance in the different metrics. These leaderboards are suitable to declare winners but should be treated with care as it is unclear what makes a *good* method and suitable ranking algorithms are an unsolved problem (more on this in App. A).

4.5.1 Acute stroke penumbra estimation (SPES)

Classifier (Decision Forest)	
T	100
F_{node}	$\sqrt{F}$
t_{depth}	unlimited
C_{opt}	Gini
Subsampling	
N_S	$1,000,000$
Features	
int	
wlm	$\sigma = 3, 5, 7$ mm
cd	$d = x, y, z$
hd	$\sigma = 1, 3, 5$ mm

Preprocessing	
Resampling	$R_{work} = 2$ mm^3 (*)
Coregistration	$MRI_{base} = T1c$ (*)
Skull-stripping	$MRI_{skull} = T1c$ (*)
Bias-field	no
Intensity std	yes (T1c, T2, DWI)
Postprocessing	
Thresholding	0.35
Others	Keep only largest object

Table 4.8: SPES experimental configuration. Starred steps were performed by the challenge's organizers.

SPES was organized as part of ISLES[2] [Maier *et al.*, 2017]), a medical image segmentation challenge at the International Conference on Medical Image Computing and Computer Assisted Intervention 2015 (October 5-9th). The task is the estimation of the joined penumbra and core (see Sec. 3.1) regions from multi-spectral MRI cases acquired in the acute (see Sec. 3.1.1) development phase of middle cerebral artery (MCA) strokes (see Figures 3.1 and 3.3). To segment the penumbral region in acute stroke images, already minute changes in the low resolution perfusion maps have to be detected and interpreted correctly. To this end, the classifier has to learn to distinguish between areas still sufficiently supplied by collateral arteries and tissue at risk of infarction.

Image data A total of 30 training and 20 testing MCA cases are provided by the organizers via the SICAS Medical Image Repository[3]. All data originates from the Institute for Neuroradiology, Universitätsspital Bern, Switzerland and was semi-automatically segmented by a single expert rater. The MRI sequences available are T1c, T2, DWI, CBF, CBV, TTP and Tmax. See Chapter 2 for a detailed discussion of these sequences. All images are provided skull-stripped, coregistered and resampled to an isotropic resolution of 2 mm^3.

Evaluation schema ISLES employs a specialized ranking scheme, suitable to combine a number of unrelated evaluation measures, which is described in App. A. The evaluation measures combined for the SPES challenge are DC and ASSD, where the first denotes the segmentation's overlap with the ground truth and the second the average surface distance error.

Framework configuration The settings employed in the challenge are summarized in Table 4.8. No bias-field correction is applied as the SPES data was found to be of superior quality and preliminary experiments revealed no benefits. Since the perfusion maps are assumed to reflect physical quantities derived from the cerebral blood flow (see Sec. 2.2), no intensity range standardization is applied to them. Finally, the Tmax image's intensity values are capped at a value of 100, since larger values are known to constitute computational errors. The posteriori threshold is set to $t = 0.35$ to counter a slight undersegmentation tendency and, since the challenge treats MCA strokes only, all but the largest connected binary component are removed.

[2] www.isles-challenge.org/ISLES2015/
[3] www.smir.ch

rank	method	cases	ASSD (mm)	DC [0,1]
2.02	CH-Insel	20/20	1.65 ± 1.40	0.82 ± 0.08
2.20	**DE-UzL**	20/20	1.36 ± 0.74	0.81 ± 0.09
3.92	BE-Kul2	20/20	2.77 ± 3.27	0.78 ± 0.09
4.05	CN-Neu	20/20	2.29 ± 1.76	0.76 ± 0.09
4.60	DE-Ukf	20/20	2.44 ± 1.93	0.73 ± 0.13
5.15	BE-Kul1	20/20	4.00 ± 3.39	0.67 ± 0.24
6.05	CA-USher	20/20	5.53 ± 7.59	0.54 ± 0.26

Table 4.10: SPES challenge leaderboard after evaluating the 7 participating methods on the testing dataset with the proposed framework highlighted. The *rank* is the final measure for ordering the algorithms' performances relative to each other (see App. A). The *cases* column denotes the number of successfully (DC>0) segmented cases. All evaluation measures are given in mean±STD. This data was published in Maier *et al.*, 2017.

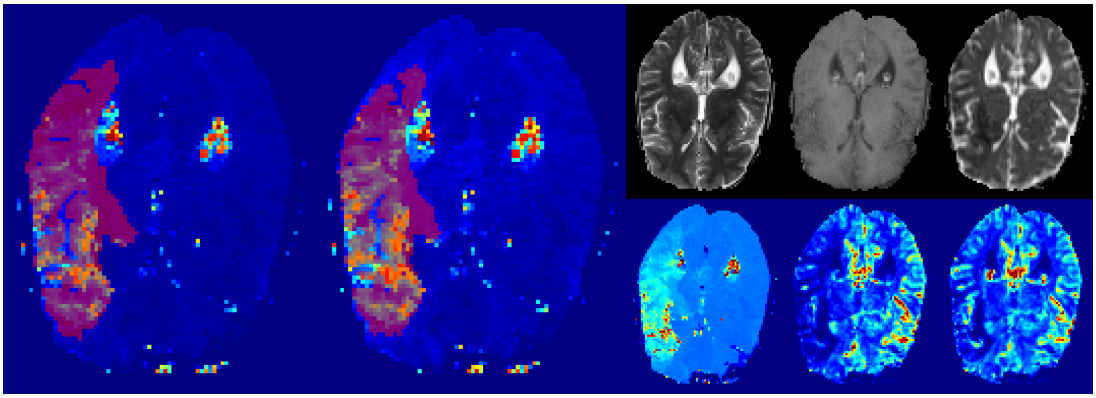


Figure 4.14: Exemplary SPES result (training set, case 16). Left: Ground truth on MTT; Middle: The proposed framework's segmentation on MTT; Right, from top-left clockwise: T2, T1c, DWI, TTP, CBV, CBF. Refer to online version for color.

Results The results of the testing set evaluation conducted by the challenge organizers is given in Table 4.10, where the proposed method reached the second place. Additionally, a statistical investigation was conducted, which revealed no significant difference between the proposed framework and the first ranked method [Maier *et al.*, 2017]. Both first placed methods were additionally tested for their runtime, which are reported with 6 minutes for CH-Insel and 20 seconds for the proposed framework (DE-UzL), including all pre- and postprocessing steps. A visual example of one of the training cases is shown in Fig. 4.14.

Discussion In the SPES challenge, the proposed framework obtained a, statistically spoken, shared first place. The high DC and low ASSD values show the approaches high accuracy, while the low standard deviation denotes its robustness.

Despite its name, SPES targets a combined penumbra and core rather than a pure penumbra segmentation. The resulting segmentation masks are clinically relevant to compute the so-called diffusion-perfusion mismatch, which is assumed to quantify the potentially salvageable tissue and hence supports treatment decisions. In the acute settings, time is an important factor. The framework's speed places it well below the upper bound of acceptable execution times for a clinical application, which is set to 5 minutes (see Sec. 3.1).

Containing only MCA occlusion cases, the challenge cannot be considered representative for acute stroke in general and the task is facilitated by the absence of small, multi-focal, frontal or cerebellar lesions. On the other hand, these large MCA lesions are the main target for thrombectomy interventions and the most common type of ischemic stroke.

Another simplification made by the organizers is the exclusion of cases with imaging artifacts and the silent assumption that all sequence types will be available, both of which cannot be assured in the clinical setting.

From a clinical point of view, SPES showed that the current state-of-the-art methods are advanced enough to support stroke diagnosis and treatment decisions through fast, accurate, reliable, and reproducible acute stoke lesion segmentation. And for clinical studies the proposed framework could supply credible estimates of the acute diffusion lesion's extend and location.

4.5.2 Longitudinal MS lesion segmentation (ISBIMS)

Classifier (Extra Forest)	
T	200
F_{node}	all
t_{depth}	unlimited
C_{opt}	Gini
Subsampling	
N_S	$1,000,000$
Features	
int	
wlm	$\sigma = 3, 5, 7$ mm
cd	$d = x, y, z$
lh	$b = 11, NL = 5, 10, 15$ mm^3
tissue	$\sigma = 1, 3, 7, 15, 31$ mm

Preprocessing	
Resampling	$R_{work} = 1$ mm^3
Coregistration	$MRI_{base} =$? (*)
Skull-stripping	$MRI_{skull} =$? (*)
Bias-field	yes
Intensity std	yes
Postprocessing	
Thresholding	0.4
Closing	1 mm^3
Object threshold	1 mm^3
Hole filling	yes

Table 4.11: ISBIMS experimental configuration. Starred steps were performed by the challenge's organizers. Undisclosed settings are denoted by a question mark (?).

ISBIMS[4] [Carass *et al.*, 2016] took place at the 2015 International Symposium on Biomedical Imaging in New York, NY, April 16-19. It is concerned with the automatic segmentation of MS lesions from multi-spectral MRI cases from multiple time points (TPs) (Sec. 3.2). MS lesions are multi-focal regions of varying appearance (Sec. 3.2.2). They can grow, shrink, appear or disappear over time. A suitable segmentation method has to capture their change in appearances and morphology.

Image data Longitudinal data of 5 patients (21 cases in total) was provided by the organizers for algorithm tuning and training. The testing data was separated into two sets of 10 and 5 patients (66 cases in total) released at two different dates before the day the challenge took place. On the second testing set, the time between the data download and the upload of the results was assessed to determine the methods' computational efficiency. Two expert raters segmented all cases manually. For each patient, between 4 and 5 TPs were released. The MRI sequences available are T1, T2, PD and FLAIR. See Chapter 2 for a detailed discussion of these sequences. All images are provided skull-stripped, coregistered and sampled to an isotropic resolution of 1 mm^3.

[4] http://iacl.ece.jhu.edu/MSChallenge

Rank	Method	Score	Nm-Dice	Nm-PPV	Nm-TPR	Nm1-lFPR	Nm-lTPR	Long-Corr	Total-Corr	min
1	IIT Madras	0.72	0.94	1.25	0.74	0.59	0.67	0.55*	0.88*	1748
2	PVG_1	0.70	1.06	1.27	0.89	0.85	0.52	0.25	0.85*	1488
3	**IMI**	0.70	1.01	1.32	0.84	0.73	0.60	0.25	0.86*	198
4	CMIC	0.65	0.94	1.07	0.82	0.61	0.47	0.33	0.85*	243
5	MS metrix	0.65	0.94	1.20	0.75	0.62	0.53	0.33	0.86*	305
6	VISAGES Deux	0.64	1.02	1.22	0.89	0.69	0.68	0.06	0.80*	354
7	DIAG	0.61	0.85	0.87	0.88	0.42	0.74	0.21	0.80*	183
8	CRL	0.56	0.71	1.11	0.51	0.59	0.35	0.33	0.85*	359
9	TIG-UCL	0.55	0.60	1.11	0.40	0.43	0.62	0.18	0.81*	1916
10	VISAGES Trois	0.52	0.68	1.01	0.56	0.56	0.46	0.17	0.65*	3159

Table 4.13: ISBIMS challenge leaderboard after evaluating the 10 participating methods on both testing datasets with the proposed framework highlighted. A star (*) behind a correlation value denotes its statistical significance at $p < 0.05$. The *score* is the final measure for ordering the algorithms' performances relative to each other and is based on a weighted mean of the metrics (see text). Most metrics are shown as inter-rater normalized mean values and the time for processing the second testing set (21 cases in total) is given in minutes.

Evaluation schema For evaluation, the organizers combine lesion detection false positive rate (lFPR), lesion detection true positive rate (lTPR), longitudinal volume change correlation (longCorr), general volume correlation (totalCorr), DC, PPV (equals precision) and voxel based true positive rate (TPR) (equals recall). All results with an Nm- prefix are reported in a normalized version. To this end, the interrater results are set as 100% reference point and the teams' average scores reported as numbers relative to this fixed point. For the final rank, these measures are combined into a score using the following formula: 20% Nm-1-LFPR, 20% Nm-LTPR, 20% LongCorr, 20% TotalCorr and 20% average of Nm-Dice, Nm-PPV, and Nm-TPR. No information was provided regarding the motivation behind this ranking scheme. Beside the segmentation accuracy ranking, a second rank is assigned to each participating method for its processing speed. Combined with the accuracy rank, this was used to award an efficiency price.

Framework configuration The settings employed in the challenge are summarized in Table 4.11. Of the default features, the hemispheric difference is discarded, as especially periventricular MS lesion often appear symmetric over both hemispheres. Instead, tissue probability maps are introduced to account for the fact that the target type of MS lesions only appear in the WM. To this end, WM, GM and CSF tissue maps are computed with FSL-Fast [Zhang *et al.*, 2001] and the weighted local means feature extracted from all three at different scales. Effectively, these three maps are treated as additional image channels. The posteriori class probabilities produced by the forest are thresholded at a value of $t = 0.4$ to counter a slight undersegementation tendency. MS lesions can be very small, hence only single-voxel results are removed as potential outliers. On the remaining binary objects, a 3D morphological closing of size 1 mm and a 3D hole closing is performed, since MS lesions are solid objects.

Results The results of the testing sets evaluation conducted by the challenge organizers is given in Table 4.13, where the method reached the third place in accuracy and won the efficiency price. The non-normalized evaluation results obtained by the proposed framework are depicted in Table 4.14, together with selected inter-rater results. A more complete evaluation can be found in the challenge article Carass *et al.*, 2016. A visual example of one of the training cases is shown in Fig. 4.15.

Method	Dice	PPV	TPR	lTPR	lFPR	Long-Corr	LLTPR	LLFPR
IMI	0.57	0.73	0.51	0.35	0.31	0.25	0.11	0.51
R1 as truth	0.56	0.73	0.50	0.35	0.18	0.18	0.06	0.30
R2 as truth	0.56	0.50	0.73	0.83	0.65	0.18	0.94	0.70

Table 4.14: Non-normalized results achieved by the proposed framework in the ISBIMS challenge and the inter-rater scores for direct comparison.

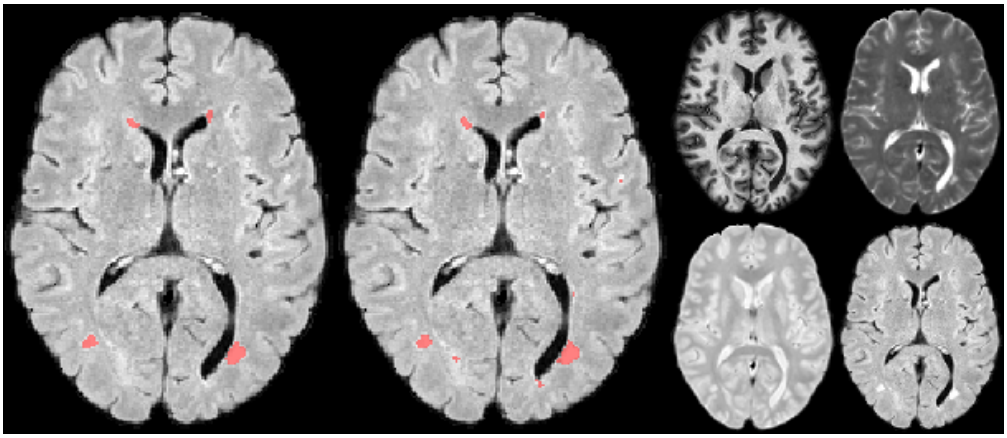


Figure 4.15: Exemplary ISBIMS result (training set, case 04_03). Left: Ground truth on FLAIR; Middle: The proposed framework's segmentation on FLAIR; Right, from top-left clockwise: T1 (MPRAGE), T2, FLAIR, PD. Refer to online version for color.

Discussion The proposed method obtained a good third place with a score abreast with the second and close to the first placed team. No statistical test was performed by the organizers to establish the significance between the positions.

Thanks to a high accuracy and excellent runtime the method obtained the special efficiency price. But it should be noted that MS lesion segmentation is not a time-critical task and any method fast enough to process the daily influx of new cases can be readily employed in practice. When conducting large studies, on the other hand, a faster method can be beneficial.

Considering the small size and multiplicity of MS lesions, competitive average metric values are obtained by the proposed framework, especially when compared to the inter-rater results. However, the two expert raters' segmentations correlate poorly with each other, especially for the longitudinal lesion change, which is an important surrogate measure for disease burden (see Sec. 3.2.2). This shows once again that manual lesion segmentation is highly unreliable and underlines the need for automatic solutions.

When investigating the results in detail, the organizers observed clear performance differences between patients, but not between TPs. Consecutive scans of the same patient can therefore be considered similar in appearance and the longitudinal consistency should be higher than observed.

The main criticism of the challenge is the poor and non-significant longitudinal correlation for the raters, which partially invalidates the teams' longitudinal correlation result. Except in one case, the teams' correlation coefficients are furthermore non-significant. Under these circumstances, it is difficult to make a statement regarding longitudinal consistency, except that it is a pressing problem and apparently hard to obtain.

Another drawback is the lack of an exact definition of the lesion based measures (lTPR, lFPR), as lesion correspondence is a complex and by no means straight forward topic as discussed in Styner *et al.*, 2008.

Disease diagnosis, assessment and research requires the quantitative evaluation of MS lesions and the longitudinal tracking of their changes. Not one method reached a suitably high longitudinal lesion correlation to be employed for lesion tracking in practical scenarios. On the other hand, since the raters correlations are equally low, it can be doubted whether the reported correlation results are meaningful. In general, the poor interrater scores put in doubt the whole concept of MRI MS lesion assessment as a surrogate measure for disease burden.

The low human longitudinal correlation might also explain the surprising observation that not a single participating method made use of the information of which cases are longitudinal scans of the same patient to, e.g., ensure longitudinal consistency. Some teams reported to have tried and dismissed the idea. One can conclude that either MS lesion load, appearance and distribution changes considerably over time or, alternatively, that the human experts are introducing a variation great enough to shadow the actual disease induced changes. Both interpretations bring into question the correct usage of quantitative MRI lesion surrogate measure in clinical studies and treatment of MS.

4.5.3 Multimodal brain tumor segmentation (BRATS)

The BRATS challenge[5] approaches the task of segmenting high- as well as low-grade glioma. The ground truth consists of four different classes: the edema, the non-enhancing solid core, the necrotic/cystic core and the enhancing core. In set-up, the 2015 version of the challenge was equal to the previous years [Menze *et al.*, 2015a].

[5] http://braintumorsegmentation.org

Classifier (Decision Forest)	
T	100
F_{node}	all
t_{depth}	unlimited
C_{opt}	Gini
Subsampling	
N_S	$1,000,000$ (min 500/class)
Features	
int	
wlm	$\sigma = 3, 5, 7$ mm
cd	$d = x, y, z$
hd	$\sigma = 1, 3, 5$
lh	$b = 11, NL = 5, 10, 15$ mm^3

Preprocessing	
Resampling	$R_{work} = 1$ mm^3 (*)
Coregistration	MRI_{base} =? (*)
Skull-stripping	MRI_{skull} =? (*)
Bias-field	yes
Intensity std	yes
Postprocessing	
Object threshold	5 ml
Others	see text

Table 4.15: BRATS experimental configuration. Starred steps were performed by the challenge's organizers. Undisclosed settings are denoted by a question mark (?).

Image data For BRATS, three structures are evaluated: the whole tumor (all 4 labels), the tumor core (non-enhancing solid core, necrotic/cystic core, enhancing core) and the active tumor (enhancing core). The training as well as the testing set contain a mix of low grade (LG) and high grade (HG) glioma cases. Further details of the set-up are shown in Fig. 4.13.

Evaluation schema Employed evaluation metrics were DC, PPV, SE and the kappa value. The exact definitions of these metrics have not been provided by the organizers. The employed ranking scheme to determine the winners has only been explained orally at the challenge's workshop.

Framework configuration The settings employed in the challenge are summarized in Table 4.15. Main deviation from the default approach is that a minimum of 500 samples were extracted for each class per case. This became necessary due to the multi-class problem posed by BRATS. Note that still an overall number of approximately $1,000,000$ training samples is sampled. For postprocessing, the edema was allowed to grow morphological with a size of 1 mm^3 into the background and then the inner non-enhancing solid core to perform similar, but only at the expense of the edema label. This corresponds roughly to a slight inflation of the non-enhancing solid core and a subsequent adaptation of the surrounding edema.

Results On the day of the challenge, the framework reached a good mid-field place. Table 4.17 details the results obtained on the BRATS training dataset, as the results on the testing dataset have not been made public. A visual example of one of the training cases is shown in Fig. 4.16.

Cases	DC			PPV / precision			SE / recall			Kappa
	comp	core	enha	comp	core	enha	comp	core	enha	
HG 252/274	0.75	0.60	0.56	0.71	0.56	0.59	0.88	0.81	0.64	0.98
LG 053/053	0.79	0.68	0.59	0.77	0.70	0.64	0.84	0.72	0.61	0.99

Table 4.17: BRATS training dataset results for the proposed framework. Some of the resulting segmentation masks were empty and hence did not count towards the average values presented here. On the challenge, the proposed method reached an upper mid-field position, but the leaderboard has never been made public.

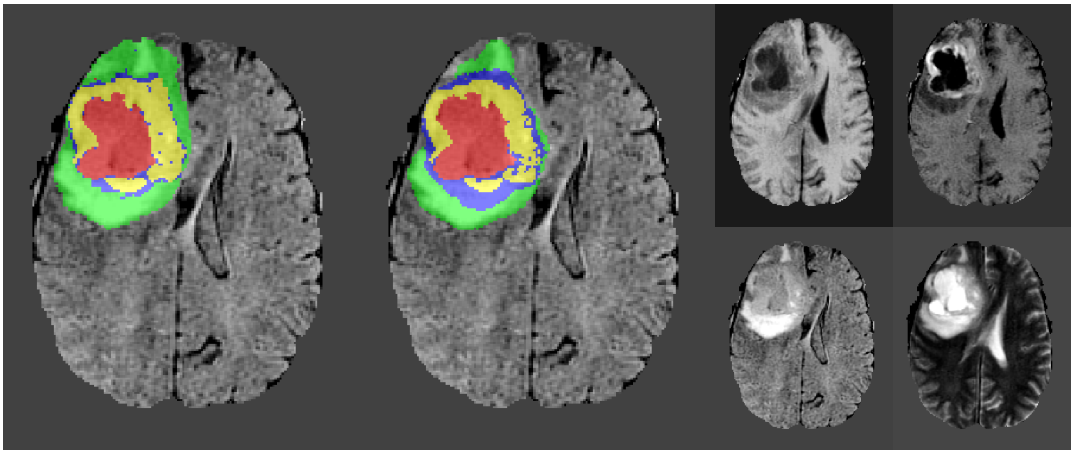


Figure 4.16: Exemplary BRATS result (training set, case brats_tcia_pat374_0001). Left: Ground truth on FLAIR; Middle: The proposed framework's segmentation on FLAIR; Right, from top-left clockwise: T1, T1c, FLAIR, T2. Refer to online version for color.

Discussion Unfortunately, the organizers of the BRATS challenge released only sparse information about the datasets, the results, and the tasks motivation, such that only few conclusions can be drawn. The proposed framework obtained a good mid-field place and produced similar results for both LG and HG cases. On average, the complete tumor segmentation is better than the core results, which in turn is better than the enhancing tumor only.

Glioma segmentation from MRI is a complex problem with diverse appearances. Nevertheless, better results were expected considering the very large and hence presumably representative database. Part of this deficit might be attributable to the poor data quality and the sometimes questionable ground truth. But it can equally mean that the proposed framework overgeneralizes in the case of very complex and diverse segmentation problems, since other approaches, such as the BRATS challenge winners, obtained somewhat better results.

4.5.4 Sub-acute ischemic stroke lesion segmentation (SISS)

Classifier (Decision Forest)	
T	100
F_{node}	$\sqrt{F}$
t_{depth}	unlimited
C_{opt}	Gini
Subsampling	
N_S	$1,000,000$
Features	
int	
wlm	$\sigma = 3, 5, 7$ mm
cd	$d = x, y, z$
hd	$\sigma = 1, 3, 5$ mm
lh	$b = 11, NL = 5, 10, 15$ mm^3

Preprocessing	
Resampling	$R_{work} = 1$ mm^3 (*)
Coregistration	$MRI_{base} = FLAIR$ (*)
Skull-stripping	$MRI_{skull} = T1$ (*)
Bias-field	yes
Intensity std	yes
Postprocessing	
Thresholding	0.4
Closing	1 mm^3
Object threshold	1 ml
Hole filling	yes

Table 4.18: SISS experimental configuration. Starred steps were performed by the challenge's organizers.

SISS was, just as SPES, organized as part of the ISLES challenge [Maier *et al.*, 2017]. The

Rank	Method	Cases	ASSD (mm)	DC [0,1]	HD (mm)
3.25	UK-Imp2	34/36	05.96 ± 09.38	0.59 ± 0.31	37.88 ± 30.06
3.82	CN-Neu	32/36	03.27 ± 03.62	0.55 ± 0.30	19.78 ± 15.65
5.63	FI-Hus	31/36	08.05 ± 09.57	0.47 ± 0.32	40.23 ± 33.17
6.40	US-Odu	33/36	06.24 ± 05.21	0.43 ± 0.27	41.76 ± 25.11
6.67	BE-Kul2	33/36	11.27 ± 10.17	0.43 ± 0.30	60.79 ± 31.14
6.70	**DE-UzL**	31/36	10.21 ± 09.44	0.42 ± 0.33	49.17 ± 29.6
7.07	US-Jhu	33/36	11.54 ± 11.14	0.42 ± 0.32	62.43 ± 28.64
7.54	UK-Imp1	34/36	11.71 ± 10.12	0.44 ± 0.30	70.61 ± 24.59
7.66	CA-USher	27/36	09.25 ± 09.79	0.35 ± 0.32	44.91 ± 32.53
7.92	BE-Kul1	30/36	12.24 ± 13.49	0.37 ± 0.33	58.65 ± 29.99
7.97	CA-McGill	31/36	11.04 ± 13.68	0.32 ± 0.26	40.42 ± 26.98
9.18	SE-Cth	30/36	10.00 ± 06.61	0.38 ± 0.28	72.16 ± 17.32
9.21	DE-Dkfz	35/36	14.20 ± 10.41	0.33 ± 0.28	77.95 ± 22.13
10.99	TW-Ntust	15/36	07.59 ± 06.24	0.16 ± 0.26	38.54 ± 20.36
	inter-observer	36/36	02.02 ± 02.17	0.70 ± 0.20	15.46 ± 13.56

Table 4.20: SISS challenge leaderboard after evaluating the 14 participating methods on the testing dataset with the proposed framework highlighted. The *rank* is the final measure for ordering the algorithms' performances relative to each other (see App. A). The *cases* column denotes the number of successfully (DC$>$ 0) segmented cases. All evaluation measures are given in mean±STD. Note that the average ASSD and HD values were only computed over the successfully segmented cases, but all contributed to the ranking. This data was published in Maier *et al.*, 2017.

task was the segmentation of sub-acute stroke lesions from multi-spectral MRI images. In its sub-acute evolution phase, the stroke lesion appearance undergoes a range of changes (Sec. 3.1.1) that have to be captured by a suitable segmentation approach.

Image data Various MRI sequences are typically utilized in the clinical routine for the assessment of ischemic stroke lesions, as they provide insights into different aspects of the disease. FLAIR MRI is probably the most prominent technique for imaging in sub-acute ischemic stroke patients, followed by DWI and T1 datasets. In the sub-acute phase (here: $>$ 24 hours and $<$ 2 weeks), the lesion usually appears hyperintense in FLAIR and DWI and hypointense in T1 datasets. The cases display a wide range of stroke lesions: Large ($>$ 300 ml) single and small (1 ml) multi-focal (up to 14) lesions; various degrees of non-lesion white matter hyperintensities (WMH) load; localization all over the brain; and different effected arteries and laterals. Some cases show additional bleeding (haemorrhages) inside the lesion. Of all the benchmarks, this task can be considered the most challenging.

Evaluation schema For evaluation, the same ranking procedure as for SPES is employed (App. A), only that beside the DC and ASSD, the HD is also considered.

Framework configuration The settings employed in the challenge are summarized in Table 4.18. The postprocessing includes the filling of holes in the segmentation, a morphological closing operation of size 1 mm and a removal of all unconnected components smaller than 1 ml as presumed outliers.

Results The results of the testing set evaluation conducted by the challenge organizers is given in Table 4.20, where the method ranked favorably. Additionally, a statistical investigation was

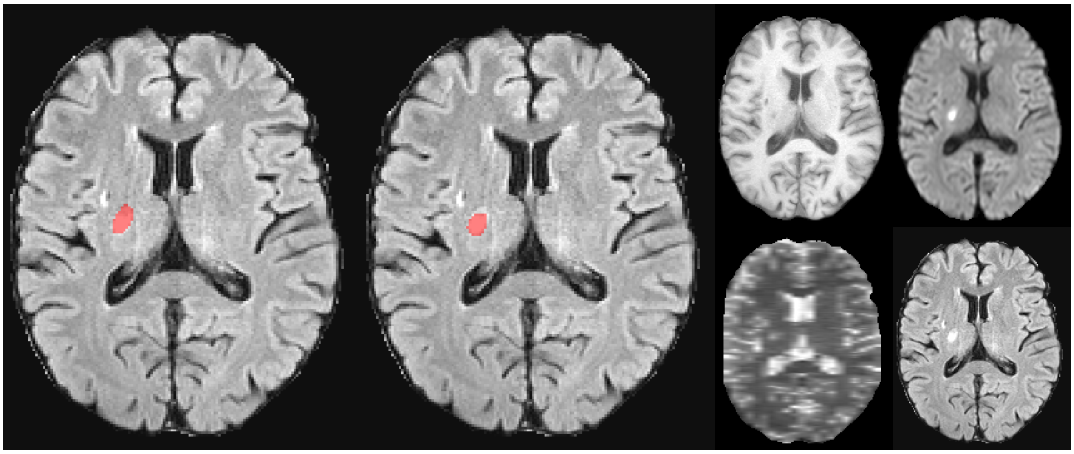


Figure 4.17: Exemplary SISS result (training set, case 23). Left: Ground truth on FLAIR; Middle: The proposed framework's segmentation on FLAIR; Right, from top-left clockwise: T1, T1c, FLAIR, T2. Refer to online version for color.

conducted, which revealed no significant difference between the fourth, fifth and sixth ranked methods, effectively placing the framework in a shared fourth place. A visual example of one of the training cases is shown in Fig. 4.17.

Discussion For sub-acute stroke, the framework reached a good midfield placement and, statistically speaking, a shared fourth rank. The SISS challenge proved very difficult, as can be seen by the overall low average scores. Partially this can be attributed to the diverse appearance of stroke lesions in the sub-acute phase (see Fig. 3.2); partially to the imaging artifacts ridden, clinically realistic cases; and partially to scanning protocol differences between the medical centers. In fact, some of the lesions were missed entirely by all participating methods.

Nevertheless, the inter-observer accuracy is considerably higher. From this observation two conclusions can be drawn: First, human observers agree largely on the concept of a sub-acute stroke lesion, despite the fact that the automated methods seem to be unable to detect them accurately. And, subsequently, that there is much room for improvement until the accuracy of a human observer is reached.

As already observed with the BRATS challenge, the proposed framework shows again difficulties with highly complex segmentation tasks. Noisy samples and diverse appearances combined with few training cases seem to pose a challenge to the method and keeps it from the winning places. For such segmentation tasks, one might have to consider methodological improvements, data augmentation approaches or even feature learning classifiers such as CNNs (see App. B for a short discussion on the topic).

From a clinical point of view, sub-acute stroke lesion segmentation is of little interest. Clinical and neuroscientific studies, on the other hand, often require the segmentation of large amounts of sub-acute stroke cases, a task which could be significantly sped up by automatic methods. Beside enabling larger patient groups, the reproducible and reliable results would additionally render the statistical evaluation and hence reported findings more reliable.

As it is now, neither the proposed framework nor any other participating method can be safely recommended for a completely automated use. Instead, they could be employed in a supportive, semi-supervised fashion. Alternatively, a usage in study candidate selection is imaginable.

4.6 Conclusion

A framework for automatic general brain lesion segmentation from multi-spectral MRI acquired in clinical routine was presented. Main contributions are the image features, the sampling schema and the framework's architecture.

The results from four different brain lesion segmentation challenges place the method among the top-ranking state-of-the-art approaches, underlining its applicability to a range of pathologies with only minor adaptations. This in turn signifies that the proposed features fulfill their intended function of providing discriminative information for general brain lesion identification and that the assembled architecture is capable of overcoming the numerous challenges posed by MRI images.

Another strong point of the method is its speed, both in application and training. The former has been attested by an efficiency award (at ISBIMS) and a special mention for its short runtime (at SPES). The latter can be attributed to the intelligent training set subsampling schema proposed, which was shown to successfully represent each task's diversity while avoiding redundant information.

Designed as a general purpose method, the framework nevertheless obtained top results in two (SPES and ISBIMS) of the independently evaluated challenges. Conclusively, the proposed framework can be recommended for all MRI brain lesion segmentation tasks of low to medium complexity. In the case of problems with high diversity and noise level (BRATS and SISS), the obtained segmentation accuracy is robust but surpassed by more specialized approaches. This in turn means that there is room for improvement by incorporating pathology specific knowledge, such as longitudinal features for MS, brain tumor models for glioma or vascular territory maps for stroke segmentation.

Following a pipeline pattern, the framework possesses a number of hyperparameters which have to be set. It was found that most of these relate to the segmentation accuracy in a logarithmic form, i.e., optimal values can be chosen from a (potentially infinitely) wide plateau. Parameter optimization is hence only necessary when the training and application speed are to be exhaustively minimized.

During the introduction to MRI, the modalities' high variability in appearance were emphasized (Sec. 2.3) and a number of dedicated preprocessing steps introduced to equalize the images before classification (Sec. 4.3.1). The results obtained in the hyperparameter analysis (Sec. 4.4.1, Table 4.3) reveal a definite drop in segmentation performance when disabling any of these measures. This confirms that the inter-sequence variability of MRI indeed poses a problem to the images' computational processing. It reveals furthermore that the selected counter-measures do alleviate the overall segmentation problem.

How much of the encountered variability they actually manage to remove is difficult to assess, as no perfect MRI scan exists to be used as gold standard for comparison. But some visually confirmed failures of, e.g., the skull-stripping, show their fallibility and indicate that there is room for improvement.

This highlights one of the major problems when comparing published MRI lesion segmentation methods: They are mostly frameworks of inter-dependent components whose individual contributions to the final segmentation result cannot be readily established. The perceptible influence of the selected preprocessing as shown in this work is likely to account for many of the reported differences between methods rather than the proposed changes to the segmentation algorithms to which they are usually attributed. It is therefore vital to describe the employed preprocessing steps in detail and to analyze their influence on the final results in order to provide as complete a picture as possible.

Choosing the best preprocessing method requires a sound domain knowledge (e.g., exclusion

of the PWI maps from the intensity standardization, selecting an anatomical sequence for the skull-stripping, etc.). Indeed, selecting an unsuitable arrangement might have easily cost the method its favorable placement in the challenges.

Interestingly, the three SISS winners include a CNN, a DF and even a generative modeling, non-machine learning solution. Other methods from the same three families can be found all over the leaderboard. This indicates strongly that it's not the type of method that matters, but rather a careful and, above all, informed adaptation to the task at hand, where all involved components are carefully tuned and synchronized. Therefore, the conducted hyperparameter analysis (Sec. 4.4.1) is highly justified.

Chapter 5

Local problem forests for sub-acute stroke lesion segmentation

In the previous chapter, a framework for brain lesion segmentation based on voxel-wise classification with decision forests (DFs) was presented. A thorough evaluation including independently rated results from multiple challenges demonstrated the robustness and accuracy of the proposed method.

Of the three covered brain lesion segmentation tasks, sub-acute stroke lesions proved the most challenging. The investigation into the influences of the different magnetic resonance imaging (MRI) sequences (see Fig. 4.9 in Sec. 4.4.1) showed that multi-spectral data is required to obtain competitive results. For example, the skull-stripping is more reliable on T1 or T2 sequences [Maier *et al.*, 2015e; Maier *et al.*, 2015d] and the hyperintense respectively hypointense appearance of periventricular white matter lesions (WMLs) on DWI respectively ADC maps allows to distinguish them from stroke lesions (see also the discussion in Sec. 3.1.2 on stroke appearance in MRI). Unfortunately, often just a single FLAIR follow-up scan is acquired for sub-acute stroke assessment due to time and financial restrictions.

Hence, it would be desirable to improve the method's accuracy on mono-spectral FLAIR data. A close inspection of the segmentation results obtained with the proposed framework reveals some areas of frequent misclassification highlighted in Fig. 5.1: (1) scalp fat residues from imperfect skull-stripping, (2) periventricular WMLs, (3) remains of ocular bulbs, (4) lesions reaching right up to the brain border, and (5) larger fissures.

For the remainder of this chapter, these will be referred to as *sub-problems* of the overall classification problem, as each of them poses a unique challenge to the classifier. Some are associated with oversegmentation (1, 2, 3), others with undersegementation (4, 5). Furthermore, they can share common characteristics, e.g., the periventricular WMLs and the lesions reaching up to the brain border are both defined by a close proximity of hyperintense tissue and CSF, but require opposing treatment, i.e., the WMLs conservative and the brain border proximity greedy classification.

All sub-problems have in common that they display lesion-like intensities locally and that they are distinguishable by their non-local anatomical neighborhood context, e.g., the proximity of the skull or the presence of ocular bulb remains. They may or may not affect a case, they have no fixed global anatomical location, and they are encountered with moderate frequency.

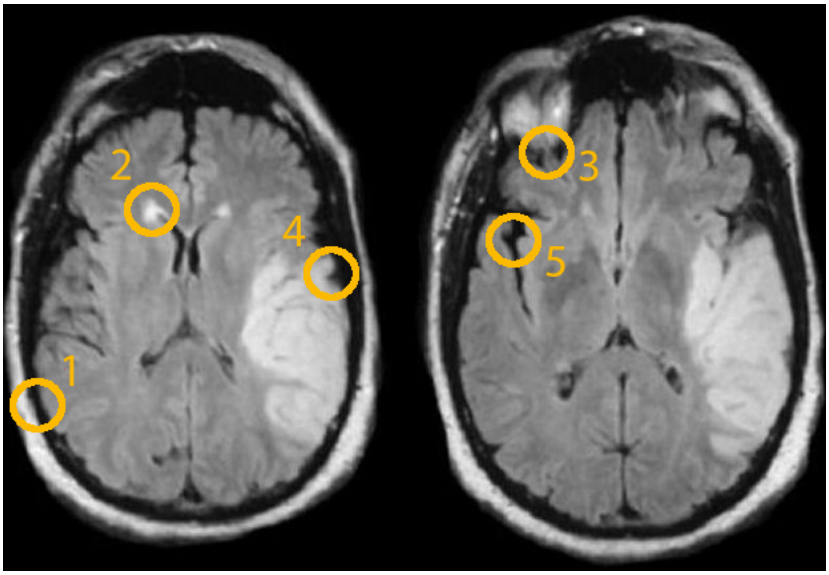


Figure 5.1: Sub-problems identified in the case of sub-acute stroke lesion segmentation from mono-spectral FLAIR data. Refer to the text for a detailed description.

Providing the DF with non-local information through suitable features might seem like a straight forward solution to treat these sub-problems. But despite the local histogram and the weighted local means features (as introduced in Sec. 4.2) the segmentation errors persist.

This poor behavior might be attributable to the forests' divide & conquer training approach (see Fig. 5.2): Since the sub-problems are not very numerous and share intensity characteristics with the true lesions as well as each other, they are not addressed in the upper nodes of the trees. At deeper nodes, the samples representing a sub-problem are already widely spread all along the width of the tree, very likely tightly bundled with samples representing the real lesions and the other sub-problems. Now the overfitting countermeasures implemented (e.g., depth restriction, minimal number of samples per node) prevent that small portions of the sets are shaved off, effectively treating the few remaining sub-problem samples as noise. Loosening the depth and other similar restrictions would possibly allow to correctly treat the sub-problems but only at the cost of an undesired overfitting.

An alternative way to improve the DF's classification behavior is to encourage the trees to specialize on a specific sub-problem. To this end, Lombaert *et al.*, 2014 recently proposed to replace the default bagging/bootstrap aggregation [Breiman, 1996] scheme, which ensures the growing of subtly different trees through random sampling from the training set, by a *guided bagging* approach.

In their work, they address the problem of multi-organ segmentation in abdominal CT scans. Their training cases consist of various clippings of the abdominal area, each of which poses slightly distinct difficulties, together forming the whole multi-organ segmentation task. Before training, Lombaert *et al.*, 2014 automatically sort the training cases into clusters displaying similar cut-outs of the abdomen. These thus form the bags of training data. For each, a single voxel-level tree is subsequently trained on the samples extracted from all contained cases. The result is a forest with trees specialized on the problems represented by their respective clusters.

With the default bagging scheme, all trees receive an equal vote on a new sample's class membership at application time. This is sensible, as all trees are randomly trained and there is

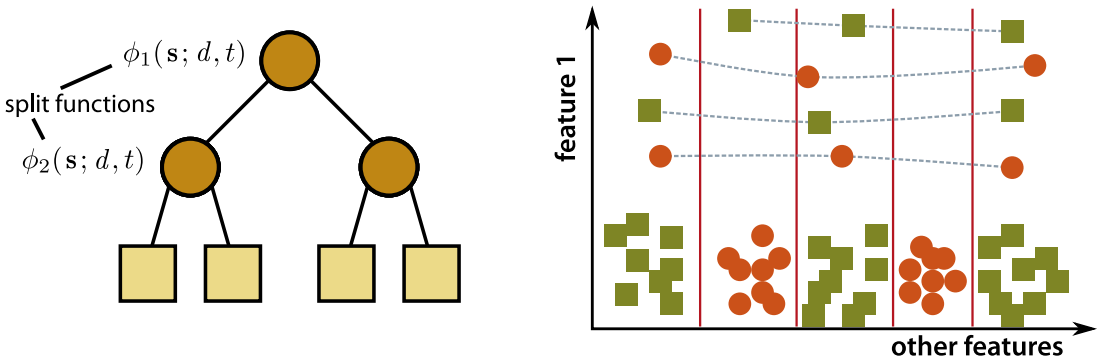


Figure 5.2: Schema of how the DF's divide & conquer approach might fail to handle the sub-problems. Right side shows the feature space with classes denoted by squares respectively circles and the sub-problem samples with a dotted gray line. Feature 1, a regional feature, allows to distinguish between them. The other features are all highly correlated, e.g., because they are intensity based. During training of the first nodes, a vertical split will always be preferred over a horizontal split. The random sub-space mechanism, which randomly selects the features visible to the node training, fails to avoid this in the case of many correlated features. After all vertical splits (red lines) are set, the expected gain of the potential horizontal splits is, seen by themselves, insufficient to be judged desirable by the DF training procedure. If the DF would be able to see all data at each node, the situation would be different. As it is, the Feature 1 will never by select at any node.

no way of knowing which is most suited to make the decision. But with guided bagging the trees are specialized and a sample should be processed by the ones especially fitted to the problems it represents. Therefore, weighted voting is employed. To this end, a formerly unseen case is first compared to the training cases in each cluster and the average similarity over all of them is subsequently computed. This similarity value then forms the vote weight of the tree associated with the regarded cluster. Thus, a new cases is mainly processed by the trees trained on similar cases.

Inspired by their approach, *local problem forests (LPF)* are proposed, a method to separate the above identified sub-problems of mono-spectral sub-acute lesion segmentation and train specialized classifiers on each resulting cluster. The following paragraphs are dedicated to a generalized descriptions of the method and motivate the concrete implementation decisions made for the specific problem at hand.

The targeted sub-problems are defined at a local neighborhood, rather than the whole image level as in Lombaert *et al.*, 2014. That means a suitable representation for the identified sub-problems and an apt distance measure, which allows to distinguish between them, have to be designed. Together, these define a space in which a clustering algorithm can separate the sub-problems from each other. For the remainder of this chapter, this sought-after space will be referred to as the *problem space*, i.e., a concrete, multi-dimensional space with a topology defined by the selected representation and distance measure, which is not to be confused with the feature space. In a suitable problem space, all samples not belonging to any sub-problem but similar in representation appearance are clustered together with their respective nearest sub-problem. All other samples not belonging to and not similar to any sub-problem will be situated all over the space in an unknown distribution. Fig. 5.3 schematically denotes this concept.

Which problem space is most suitable for the task depends on the sub-problems identified.

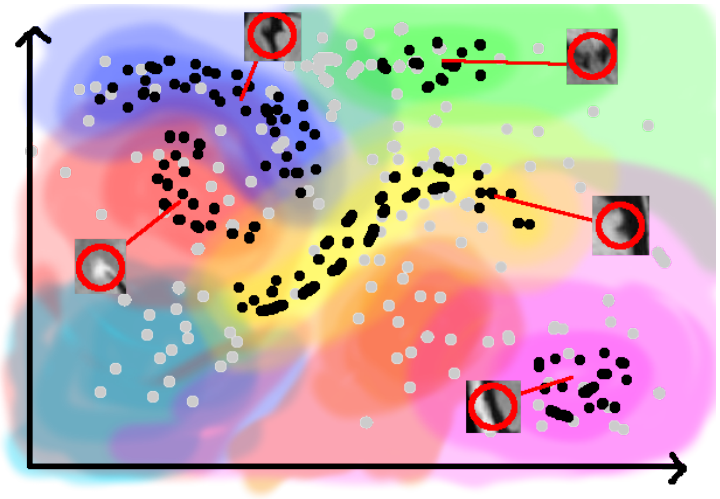


Figure 5.3: Schema of an ideal problem space with black dots representing specific sub-problems as identified in Fig. 5.1. Unrelated samples are depicted as gray dots and the specialized forests' catchment areas in different colors. Refer to online version for color.

The proposed method attempts to differentiate between the ones depicted in Fig. 5.1, which are characterizable by their non-local neighborhood context. Therefore, patches are chosen as representative entities for the sub-problems. Hence, each training image is divided into suitably sized patches.

After clustering the sub-problem representing patches in the problem space with a suitable clustering algorithm, a classifier is trained on each resulting bag. Owing to the complexity of the addressed segmentation problem and not to loose the advantages of classical bagging, small DFs are used rather than single trees as proposed by Lombaert *et al.*, 2014. The LPF method can therefore be considered as *guided super-bagging*. The hypothesis behind the proposition is that, since the other sub-problems will ideally not appear in the training data, the sub-problem targeted by the associated forest will be suitably represented and sufficiently distinguishable to be solved by the DF algorithm. As samples similar in the sense of distance in the problem space will also be present, the DF simultaneously learns to distinguish these from the real sub-problem.

In the application described by Lombaert *et al.*, 2014, each case belongs to one specific anatomical region (i.e., their sub-problems). In the present case, most of the patches will not be associated with any of the targeted sub-problems (see gray dots in Fig. 5.3). First, this means that a number of cluster higher than the count of sub-problems has to be chosen to allow for additional clusters of mutually similar, but not sub-problem related patches. Second, the training samples contained in a single cluster might be too specific to allow for successful forest generalization. To address this last problem and to take into account that proximal clusters should share some of their training data, *fuzzy stratified random sampling* is introduced, such that the forests have overlapping catchment areas in the problem space from which they draw their training data. These are denoted by the colored regions in Fig. 5.3.

At application time, a formerly unseen sample is placed in the problem space, the distance to each forest established and they subsequently perform a weighted vote on the sample's class membership. Ideally, the forests responsible for the specific sub-problem posed by the sample are getting the highest votes.

A detailed description of the clustering algorithm and the distance measure that together span

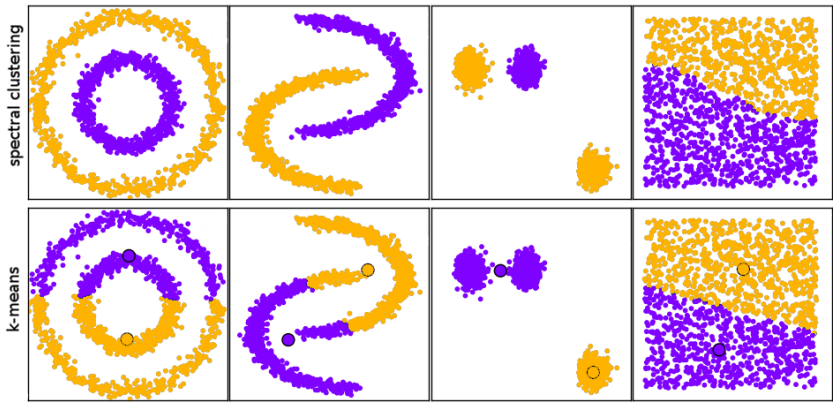


Figure 5.4: Clustering behavior of k-means and spectral clustering on different example distribution. Note the ability of the latter to capture even complex cluster shapes by following the local connectivity pattern. The k-means results are depicted with cluster centers denoted as black outlined dots. Image adapted after `http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html`

the problem space has since now been omitted. Based on the patch representations, two clustering algorithms are investigate in the remainder of this chapter. The first is a k-means approach for which the cumulative histograms of the patches are computed and the clustering performed directly on them with k-means using Euclidean distances. The second proposed approach is based on spectral clustering [Von Luxburg, 2007], for which the normalized histograms of the patches are constructed. Building on the relative bin derivation (as used, e.g., in Caicedo *et al.*, 2007 and implemented in Maier, 2016b) as distance measure, spectral embedding is carried out and the clustering performed in the resulting low-dimensional representation of the problem space. In both cases, the distance measures are selected to be scale and rotation invariant. The main difference is that spectral clustering uses a local distance preserving dimensionality reduction technique that allows to circumvent the non-trivial clustering in high dimensional space and can follow non-linear connectivity. Hence, spectral clustering is able to capture more complex cluster shapes than k-means as depicted in the toy examples of Fig. 5.4. Lombaert *et al.*, 2014 based their clustering on spectral embedding with the sum of squared differences as distance measure but did not investigate whether the simpler k-means algorithms would have sufficed.

By isolating the sub-problems from each other, the LPF method effectively raises their prevalence in the associated forest's training set above the noise level. It should therefore be able to circumvent the inability of the classical forest to treat each sub-problem independently.

Three design choices can be expected to have a substantial influence on the segmentation outcome and can be considered the main challenges connected with the proposed method: First, the selection of a suitable representation of the sub-problems. Second, the associated distance measure to distinguish between them. And third, a clustering algorithm able to capture the true cluster shapes in the problem space defined by the first two together. If these three are suitably chosen, the proposed method should lead to a robust improvement in quantitative segmentation accuracy compared to the standard approach. Furthermore, it should be possible to confirm this qualitatively through visual inspection of the areas representing the identified sub-problems.

The details of the implementation, both clustering algorithms, and the rationale behind them are described in detail in the following section. A thorough evaluation (Sec. 5.2) on the 35 sub-acute ischemic stroke cases that were already employed in the previous chapter for the

hyperparameter analysis and the classifier comparison will show if the LPF approach improves the segmentation of stroke lesions and whether the non-linear spectral clustering is necessary to capture the clusters' shapes or if the linear k-means suffices. The findings are then discussed in Sec. 5.3.

5.1 Method

The proposed LPF method introduces a number of modification to the brain lesion processing pipeline previously introduced in Chapter 4. These changes are depicted in Fig. 5.5 and explained throughout this section. The base mechanism is to split the training samples into different fuzzy overlapping subsets according to a patch-wise clustering scheme. On the voxel-wise features of each of these subsets a DF is trained, presumed to specialize on the specific sub-problem represented by the associated cluster. At application time, a formerly unseen sample is passed to all trained DFs, which decide on its class membership via a weighted vote. An important concept is the problem space, i.e., the space in which (1) the training patches are placed, where (2) their clustering takes place, in which (3) the trained DF are anchored, and in which (4) the test patches have to be embedded.

The more generic components of the LPFs are described first, while the last sub-sections detail the two compared implementations of the clustering step: Spectral clustering in Sec. 5.1.5 and k-means clustering in Sec. 5.1.4.

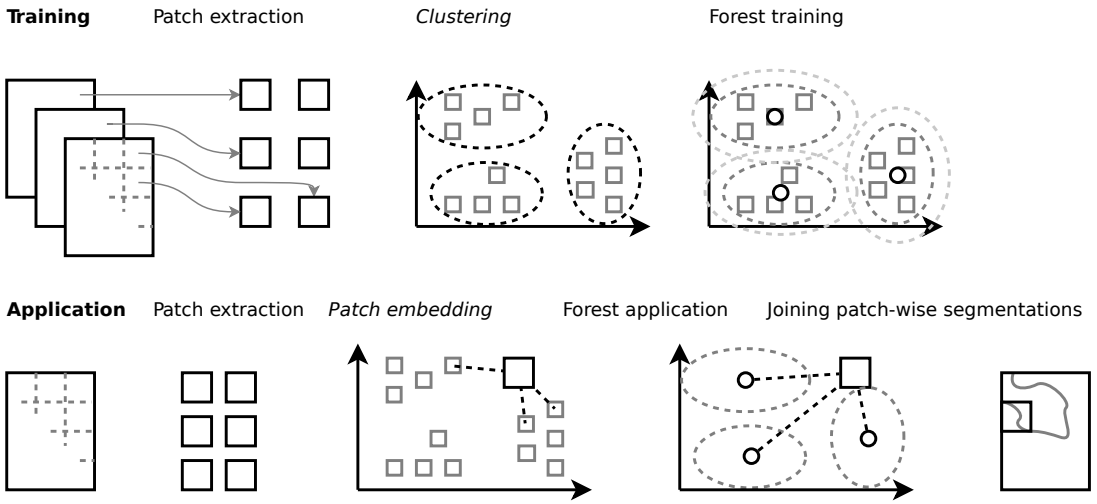


Figure 5.5: Schematic overview of the LPF method and how it fits into the already proposed framework. With k-means and spectral clustering, two variants of the clustering/patch embedding steps are proposed and compared.

5.1.1 Preprocessing and patch extraction

First, the images are preprocessed as described previously (Sec. 4.3.1), including resampling, skull-stripping and intensity range normalization.

The sub-problems of stroke lesion segmentation identified in the introduction are distinguishable by their non-local neighborhood context. To access this information, each of the N_T training

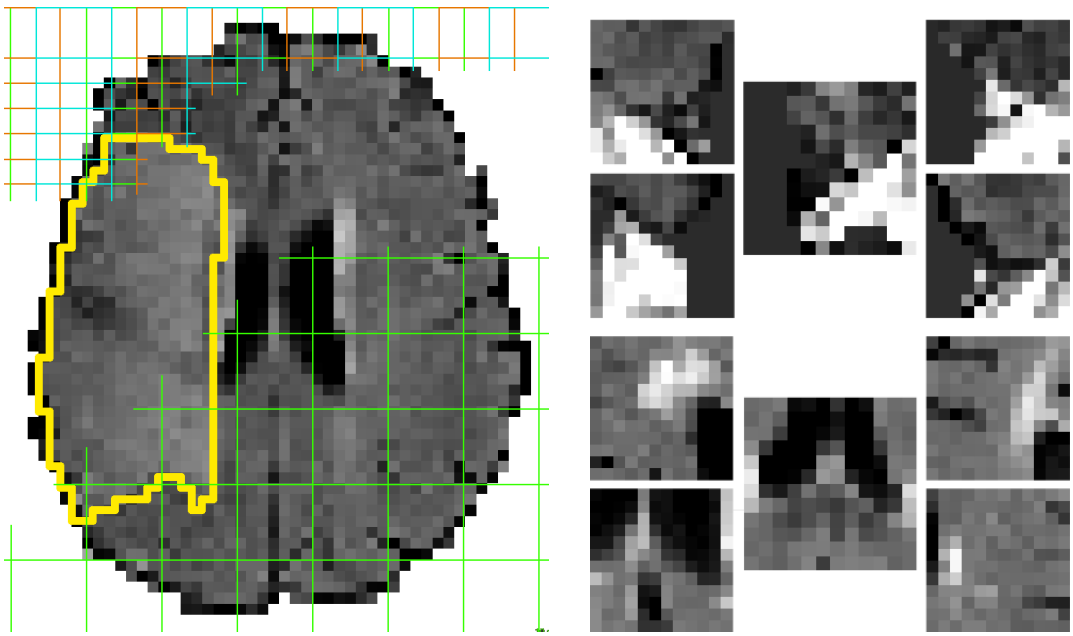


Figure 5.6: *Left*: Example case with ground truth as yellow outline, partially overlapping training patch grid in upper left and non-overlapping testing patch grid in lower right corner. *Right*: Two example patches and their nearest neighbors. The upper group displays parts of the ocular bulbs, the lower one ventricles with adjacent white matter hyperintensitiess (WMHs).

images I_i is broken down into N_{P_i} patches $P_{i,j}$ of cubic size ρmm^3 with an overlap of $\frac{1}{3}$. All together then form the set of training patches $P = \{P_{1,1}, ..., P_{N_I,N_{P_i}}\}$ (see Fig. 5.6, left). The partial overlap accounts for small shifts between similar patches and augments the training samples, as a single voxel sample can contribute to multiple forests. Patches prove particularly suitable to represent the identified sub-problems (see Fig. 5.6, right). This first step results in a large number of training patches that are next passed to the clustering algorithm (see schema in Fig. 5.5).

5.1.2 Fuzzy sampling and forest training

The application of one of the two clustering implementation described later in this chapter leads to a partition of all patches into different clusters in the problem space. In the case of spectral clustering, this can result, e.g., in the situation depicted in Fig. 5.9, where the partition is denoted by the different colors of the patch dots. The clusters are furthermore represented by ellipses with small crosses in the center.

In the ideal case, none of the cluster holds samples from more than one sub-problem, but a single sub-problem may be distributed over multiple clusters. The straight forward approach would be to use the patches in each cluster to train a local forest as done by Lombaert *et al.*, 2014. But the modeled sub-problems cannot be assumed to have crisp borders, rather do neighboring sub-problems share common characteristics. Furthermore, there will be other clusters beside the ones defining sub-problems, since most samples belong to none of the sub-problems. More suitable would be an approach with diffuse cluster borders, such that the fuzzy catchment areas of the local forests cover the whole problem space in an overlapping fashion. To this end, *fuzzy*

stratified random sampling is proposed to collect the training set for each LPF.

The idea is that every patch participates in the training of each local forest and that the number of contributed voxel samples depends on the distance to the respective cluster's center. To take the cluster shape into account, the Mahalanobis rather than the Euclidean distance is employed. And to ensure that at least all the samples actually belonging to a cluster contribute substantially to the associated local forest, the distances are normalized.

First, the centers c_k and the sample co-variance matrices Σ_k of the N_C clusters obtained are computed. The Mahalanobis distance between a cluster $k \in \{1, \ldots, N_C\}$ and a patch $l \in \{1, \ldots, |P|\}$ represented by its embedded position v_l is defined as

$$d_{M,c_k}(v_l) = \sqrt{(c_k - v_l)^\top \Sigma_k^{-1} (c_k - v_l)}. \tag{5.1}$$

To obtain the contribution of each patch to a tree trained on a cluster k, the distances of all patches to this cluster center are calculated, forming the set of distances $U_k = \{d_{M,c_k}(v_l) | l \in 1, \ldots, |P|\}$.

The distances are subsequently shifted by subtracting the minimal distance over the set $d_{M,c_k}^{min} = \min D_k$, such that the nearest cluster member patch has a distance of 0. Now, the distances are inverted to similarities and simultaneously normalized using the Gaussian function

$$r_{c_k,l} = \exp -(\frac{d_{M,c_k}(v_l)^2}{2\sigma_r^2}), \tag{5.2}$$

where the value for σ_r is the distance of the patch most distant to c_k that is still a member of the cluster k. Thus, for all cluster member patches $r_{c_k,l} \geq \frac{1}{\sqrt{e}}$ and for all others $r_{c_k,l} \in [0, \frac{1}{\sqrt{e}}]$. This ensures that the member patches contribute to the LPF of a cluster independent of its actual shape and extend.

The value of $r_{c_k,l}$ is directly used to determine the proportion of voxels of patch P_l that are added to the training set of cluster k. Stratified random sampling is employed to keep the background-to-foreground ratio of each patch intact (see Sec. 4.3.2 for the rationale behind this).

For each of the N_C clusters a local forest is trained, each ideally representing a sub-problem and additionally part of the global classification task.

5.1.3 Application

At application time, the formerly unseen test image is partitioned into non-overlapping patches which are processed individually. Each test patch P_t is first embedded into the local problem space at a position v_t. Next, the Mahalanobis distance $d_{M,c_k}(v_t)$ to each trained local forest is computed and normalized with a Gaussian to form the forest weight

$$w_{k,t} = \exp -(\frac{d_{M,c_k}(v_t)^2}{2\sigma_c^2}), \tag{5.3}$$

with $\sigma_c = \tilde{d}_{M,c_i}(v_t)$, where $\tilde{d}$ denotes the median value over all distances, which is more robust against outlier patches than the mean of the distances employed by Lombaert *et al.*, 2014.

Next, the features of the voxels in patch P_t are passed through all forests and their class probability response is weighted according to $w_{k,t}$. Local forests close to a test patch in problem space have thus a stronger vote on the final class membership of the voxels in the patch. The final segmentation is obtained by thresholding the class probabilities at a suitable value and assembling the patches back into the test image shape.

5.1.4 K-means clustering

K-means is a method to automatically partition multi-dimensional data into k clusters of equal variance. Starting form an initial guess, data points are assigned to clusters and the cluster centers subsequently updated until the algorithm converges to a local minimum. Its main advantages are its speed and simplicity. Fig. 5.7 denotes schematically how the clustering and the patch embedding steps are performed, all of which are described in the following two paragraphs.

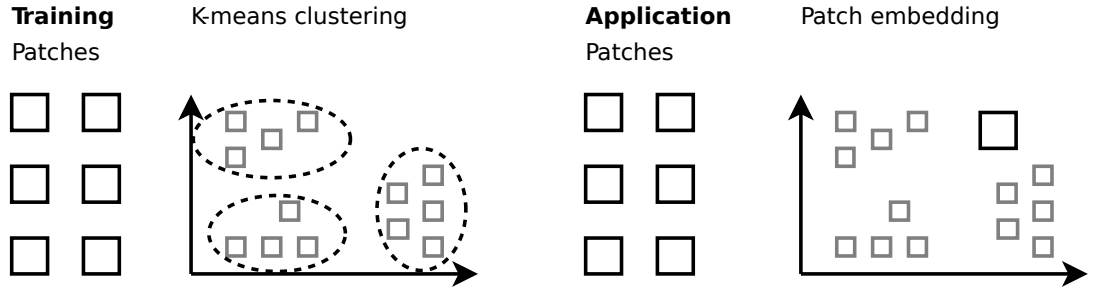


Figure 5.7: Schematic overview of the clustering and patch embedding steps of the k-means clustering variant. Note that here the clustering and embedding is performed directly in the problem space.

Clustering Since k-means works with Euclidean distances, it is not advisable to perform the clustering on the patches directly, but rather on a suitable representation. Especially in the case of the considered sub-problems, scale and rotation invariance would be desirable. To this end, the *cumulative histograms* of the patches are constructed and the k-means executed over these. Cumulative histograms reveal the intensity distribution of the patches, which is assumed to be a suitable representation to distinguish between the different sub-problems. Alternatives, such as local or globally normalized histograms, were tried and found inferior. Hence, the problem space is the space spanned by the cumulative histograms with the Euclidean distance as metric.

Patch embedding Since the problem space is defined by the patches cumulative histogram, the embedding of a new test patch in this space is straightforward: The cumulative histogram representation of each test patch directly denotes its position in the problem space.

Disadvantages The k-means variant of the LPFs has a number of drawbacks. First, since k-means does not explicitly employ the pairwise distances between the data points, but rather iteratively converges towards the minimum within-cluster sum-of-squares, it is limited to the Euclidean distance. Second, it is known to perform worse on high dimensional data due to inflation of the Euclidean distances. This second point is largely circumvented by limiting the number of bins in the cumulative histograms ($M = 10$). Third, k-means makes the implicit assumption that the sought clusters are convex and isotropic, which means it performs poorly for elongated clusters and manifolds with irregular shapes (see examples in Fig. 5.4). In the present case, nothing is known about the shape of the clusters, nor is it possible to visualize them suitably.

5.1.5 Spectral clustering

To address the limitations of k-means clustering, a second LPF variant based on spectral clustering [Von Luxburg, 2007] is proposed. This approach performs a dimensionality reduction, which preserves local but not global distances, before executing the clustering step. Thus, it is able to correctly partition elongated clusters and manifolds with irregular shapes (see examples in Fig. 5.4). Furthermore, since it performs its calculation on the pairwise distance matrix between the data points, any type of metric can be used.

The method is described in the next paragraphs and follows to some extend the ideas of Lombaert *et al.*, 2014. A schema can be found in Fig. 5.8.

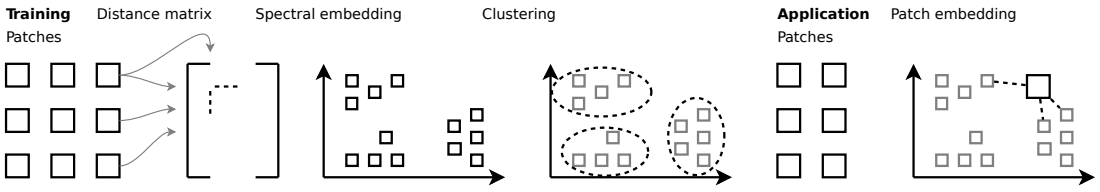


Figure 5.8: Schematic overview of the clustering and patch embedding steps of the spectral clustering variant. Note that all operation (clustering, patch embedding, distance computation, etc.) are performed in a low dimensional manifold of the actual problem space, obtained through the spectral embedding step. The distance matrix is the representation of the actual, high-dimensional problem space.

Clustering Spectral clustering is performed in three steps: First, the distance matrix between the patches is constructed. Then, the spectral embedding, i.e., the dimensionality reducing transformation, is performed. Finally, k-means is employed for clustering in the low dimensional representation.

To suitably represent the sub-problems contained in the training patches, their histograms are constructed under the assumption that the intensity distribution reveals their sub-problem memberships. As distance measure the scale and rotation invariant *relative bin deviation* is chosen, which is defined as

$$d_{rbd}(H_{l1}, H_{l2}) = \sum_{m=1}^{M} \frac{|H_{l1}(m) - H_{l2}(m)|}{\frac{1}{2}\left(|H_{l1}(m)| + |H_{l2}(m)|\right)} \tag{5.4}$$

and computed between the normalized patch histograms (H) of $M = 10$ bins (see Cha *et al.*, 2002 for a discussion on histogram distances). The measure penalizes large derivations in the bin sizes more severely, while differences small in comparison to the bin size are of limited influence. Essentially, it is a measure to compare the patches' intensity distributions, which is a logical choice considering what defines and distinguishes the identified sub-problems. If chosen correctly, the metric will enable the clustering algorithm to assign the different sub-problems into distinct clusters. The *pairwise distances* $d_{rbd}(P_{l1}, P_{l2})$ between all pairs of patches $(P_{l1}, P_{l2})_{l1 \neq l2} \in P \times P$ form then a complete simple undirected weighted graph defining the patch neighborhood in the sense of the distance measure.

Since the transformation should preserve only local distances, a sparse version of the pairwise distance matrix is constructed. To this end, *mutual k_m-Nearest-Neighbors* (mutual k-Nearest-Neighbors (kNN)) [Von Luxburg, 2007] is applied. It only allows for connections that are mutual, introducing a regularization effect which increases the focus on close together patches. This also

means that some patches might be disconnected from the graph and subsequently removed from P under the assumption that they constitute isolated samples. Fig. 5.6 (right) gives two examples of patch groupings resulting from this step, which suitably capture the targeted sub-problems and therefore confirm the choice of distance metric.

The embedding requires a similarity, rather than a distance matrix. Hence, a sparse symmetric *adjacency matrix* $A^{|P|\times|P|}$ is formed from the graph by applying a Gaussian to the remaining vertexes weights, i.e.

$$a_{l1,l2} = \exp -(\frac{d_{rbd}(P_{l1}, P_{l2})^2}{2\sigma_a^2}). \tag{5.5}$$

In contrast to Lombaert *et al.*, 2014, the median ($\sigma_a = \tilde{d}_{rbd}$) instead of the mean over all remaining vertex weights is chosen to render the normalization step in Eq. 5.5 more robust against outliers.

Next step is the graph Laplacian construction $L = D - A$, where D is a diagonal matrix with $d_{l1} = \sum_{l2} a_{l1,l2}$, and a subsequent partial eigenvalue decomposition, retaining only the two first dimensions. This step corresponds to the method of *Spectral Embedding / Laplacian Eigenmaps*, which generates a graph that can be considered a discrete approximation of the low dimensional manifold in the original space. The mapping is a locality preserving projection i.e., patches near to each other in the distance space are also near to each other on the Laplacian Eigenmap and vice-versa [Von Luxburg, 2007]. By building the Laplacian Eigenmap from a sparse (computed with mutual kNN) rather than a fully connected graph, a bias towards similar patches is introduced. The choice to map the data into a two-dimensional space was made heuristically. The transformed data points now reside in the lower dimensional manifold of the problem space. Hence, each patch is represented by a two-dimensional spectral coordinate v_l as visualized by the dots in Fig. 5.9.

Finally, under the assumption that close together patches are accurately represented in this space, an Euclidean k-means is performed to group similar patches into N_C clusters. These are represented by ellipses with crosses of corresponding color in their centers in Fig. 5.9.

Patch embedding The dimensionality reducing transformation of spectral embedding is not explicitly computed. Hence, new test patches cannot be readily placed on the Laplacian Eigenmap. Instead, the position of a new test patch P_t has to be indirectly inferred from the training patches [Von Luxburg, 2007]. To this end, the test patch's distances to all training patches in the problem space is computed with $d_{t,l} = d_{rbd}(P_t, P_l)$. From these the k_n nearest neighbors are selected and their respective spectral coordinates v_l retrieved. Based on them, the test patch's position is obtained using weighted linear interpolation, i.e., $v_t = \frac{1}{Z}\sum_{l\in kNN(t)} d_{t,l} v_l$, where $Z = \sum_{l\in kNN(t)} d_{t,l}$.

It should be noted that this position is an approximation as the real embedded location is unknown. But since local distances are preserved during the transformation from the problem space to the Laplacian Eigenmap, the chosen interpolation technique can be assumed to lead to valid results, as long as the k_n nearest neighbors are situated near to each other in both spaces. The triangles in Fig. 5.9 denote the embedding of six real test patches and it can be readily observed that the nearest neighbors are indeed close together.

Disadvantages On the downside, spectral clustering requires considerable amounts of memory and computing resources to construct the pairwise distance matrix. Furthermore, the patch embedding procedure is only accurate if the nearest neighbors of a patch are sufficiently close together. And finally, the embedding step makes it necessary to preserve all training patches for the application.

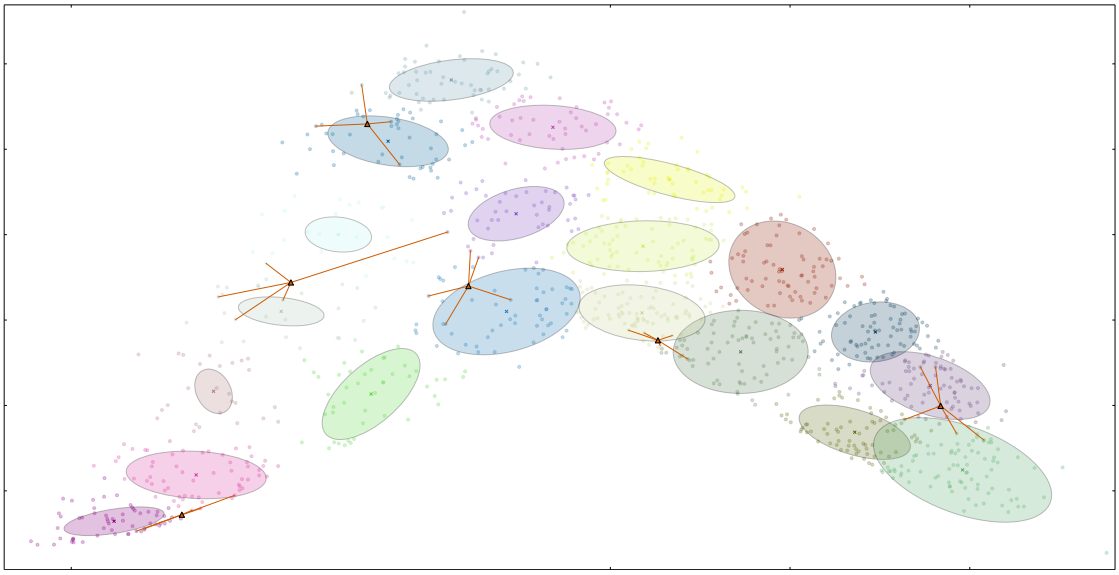


Figure 5.9: Patch distribution (dots) and clusters (ellipses with central crosses) on the two dimensional Laplacian Eigenmap, obtained through spectral embedding from the problem space, which in turn is defined by the patches' histograms and the relative bin derivation metric. Six exemplary test patches are represented by triangles and their nearest $k_n = 5$ neighbors used for coordinate interpolation connected by lines. This data represents a real case from the leave-one-out cross evaluation. Refer to online version for colors.

5.2 Experiments and results

Classifier (Decision Forest)	
T	$400/20 \times 20$
F_{node}	$\sqrt{F}$
t_{depth}	unlimited
C_{opt}	Gini
Subsampling	
N_S	$250,000$
Features	
int	
wlm	$\sigma = 3, 5, 7$ mm
cd	$d = x, y, z$

Preprocessing	
Resampling	$R_{work} = 3$ mm^3
Skull-stripping	$MRI_{skull} = FLAIR$
Bias-field	yes
Intensity range std	yes
Postprocessing	
Thresholding	0.5
Object threshold	1.5 ml
Hole filling	yes

Table 5.1: Experimental configuration.

In this section, results obtained with the two proposed LPF variants are compared against each other and the previously introduced DF framework. This will reveal if the proposed method leads to an improvement and, if yes, whether the more powerful but equally more complex spectral clustering holds any advantage over the k-means variant. To ensure that any potential improvement is attributable to a successful clustering of the sub-problems rather than other design decisions (e.g., the data augmentation through overlapping patches), an additional local forest variant with random distances and subsequently random spectral clustering is furthermore applied to the data.

Method	DM[0, 1]	HD(mm)	ASSD(mm)	Prec.[0, 1]	Recall[0, 1]
baseline DF	0.55	76	9.86	0.70	0.53
LPF: spectral clustering	0.59^{*krb}	65^{*rb}	8.46^{*krb}	0.72^{*krb}	0.57^{*krb}
LPF: k-means clustering	0.56^{*r}	57^{*rb}	8.38^{*r}	0.75^{*rb}	0.51
LPF: random distances (avg. of 10 runs)	0.53	79	9.88	0.69	0.50

Table 5.3: Mean Dice's coefficient (DC), average symetric surface distance (ASSD), Hausdorff distance (HD) as well as precision and recall of the variants. A star (*) with additional character index indicates a significant difference against the variants *s*pectral, *k*-means, *r*andom or the *b*aseline with $p < 0.01$ according to a two-sided paired t-test. Refer to Sec. 4.3.4 for details on the evaluation metrics.

Data description The experimental data consist of the 35 preprocessed FLAIR images displaying ischemic stroke lesions already used in Sec. 4.4 and their associated expert ground truth. Their resolution is isotropic $3mm^3$. Leave-one-out cross-validation is employed for evaluation. Independent of the variant tested, the forests were all trained with the same settings as denoted in Table 5.1 to ensure comparability. Please refer to Chapter 4 for details on how to interpret the table entries.

Experimental set-up The LPF variant with spectral clustering is trained with the following parameter values: $n_c = 20$ clusters, heuristically determined; $n_t = 20$ trees per cluster, which was previously found to provide a good balance between forest size and segmentation results (see Fig. 4.12a in the hyperparameter analysis); $k_m = 200$ mutual neighbors to keep the number of outlier patches to exclude form the training set low; a small number of $k_n = 5$ neighbors for the patch embedding's linear interpolation to take into account that only local distances are preserved in the Laplacian Eigenmap, such that more distant neighbors would lead to an incorrect placement of the test patch; and, finally, $\rho = 35$mm patch side length with 23.3mm overlap stride, which results in patches clearly encompassing the identified sub-problems (see grid in Fig. 5.6, left).

The k-means variant shares this configuration where applicable, i.e., $n_c = 20$ clusters, $n_t = 20$ trees and $\rho = 35$mm patch side length.

In the previous chapter, decision forests were employed for medical image segmentation using voxel-wise classification, which can be considered a standard approach to image segmentation with classifiers [Mitra *et al.*, 2014; Lombaert *et al.*, 2014; Maier *et al.*, 2015e; Maier *et al.*, 2015d]. As a baseline method against which to compare the proposed method, a single DF is trained with equally 400 trees and the same features. In each case the *sklearn* [Pedregosa *et al.*, 2011] forest implementation and the *medpy* [Maier, 2016b] feature extractors are employed.

Finally, a spectral clustering variant of LPFs with random patch distances is tested. This will show if the employed patch similarity measures are responsible for possibly better segmentation results, or if the gain has to be attributed to secondary effects, such as the training data augmentation of the method. Since this variant contains considerable random effects, the mean results over ten runs are reported.

Results All obtained results are displayed in Table 5.3. The spectral clustering variant of LPFs performs significantly better than the baseline in all evaluation measures, the k-means variant shows significant improvement in HD and precision. Training the spectral clustering variant took

$10h$ on a 4-core computer, application a few minutes. The baseline trained over half an hour, the k-means forests even faster and in both cases the application took a few seconds.

The main purpose of the LPF is the better treatment of the sub-problems identified in the introduction. Fig. 5.10 depicts examples of particularly difficult sub-problems where the spectral clustering variant lead to improvements. The first row gives an example of the baseline methods

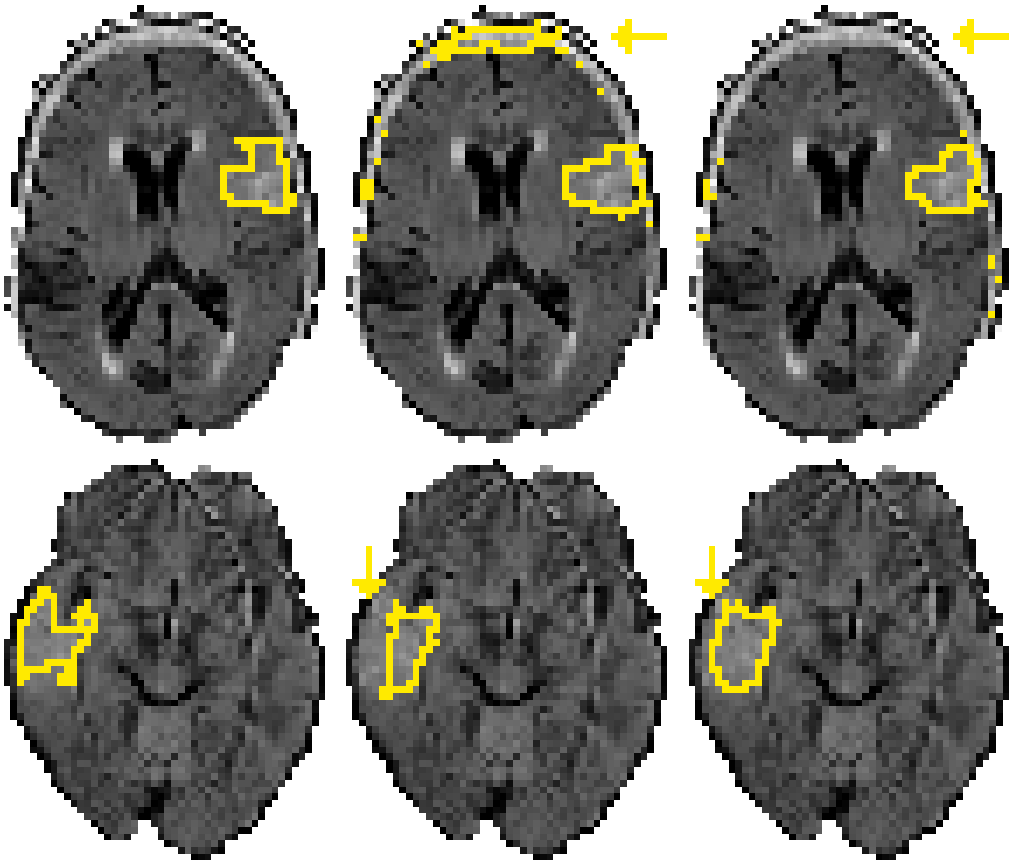


Figure 5.10: Juxtaposition of ground truth (left), baseline (middle) and LPF (right) results. The arrows denote the sub-problem areas.

problems with failed skull-stripping, while the second row shows how the LPFs with spectral clustering overcome the already noted tendency of the baseline method to keep a distance from the brain surface (see Maier *et al.*, 2015e).

A preliminary investigation into the spectral clustering's parameters found them to exercise only a limited influence on the evaluation results, with the notable exception of k_m, which should be chosen sufficiently large (e.g., $\frac{|P|}{2}$) to ensure a single connected component graph. Using crisp cluster membership instead of fuzzy sampling leads to a drop in DC of 0.011*, especially unbalancing the harmonic mean between precision and recall at the cost of the latter. Finally, using completely overlapping patches for training significantly increases the segmentation accuracy, but raises memory consumption by the power of three and strongly increases the runtime.

It could be argued that DFs already contain the ability to partition the problem space. But

first, they cannot readily be adapted to work with multi-dimensional distance measures and second, the normalization imposed by the distance matrix's sparsity could not be achieved.

5.3 Discussion

The results presented in the previous section show a significant and stable improvement of the spectral clustering LPF variant over the baseline method for all evaluation metrics. This indicates a meaningful partitioning of the problem space by the proposed LPF method and supports the hypothesis that some sub-problems should be handled independently. Clearly, the proposed LPFs outperform the standard DFs.

The k-means variant, on the other hand, only leads to significant improvements for HD and Precision. Comparing the variants, spectral clustering excels significantly in terms of DC and Recall, k-means in HD and Precision. This constellation indicates a partitioning of the problem space in a way that leads to over-specialization (i.e., increased HD, ASSD and Precision) at the cost of generalization (i.e., decreased Recall), amplifying the baseline method's undesired tendency to undersegment (i.e., Precision > Recall). Furthermore, only a few selected images seem to benefit from the k-means variant, as can be seen by its considerably, but not significantly lower ASSD score. Compared to spectral clustering, this approach is considerably faster and does not require to keep the original training patches to embed a test patch in the Laplacian Eigenmap, as the distances to the nearest forest can be computed directly through their associate cluster's center.

A consistent although not significant drop in segmentation accuracy compared to the baseline can be observed for the random metric variant, which effectively partitions the problem space into random compartments. This result serves as a strong indicator that the various secondary mechanisms of the proposed LPF method, such as data augmentation through overlapping patches and the forest ensembles, do not improve the evaluation results autonomously. This in turn signifies that the clustering has a beneficial influence and fulfills its intended purpose of identifying the sub-problems to handle them separately in a divide and conquer fashion.

The evaluation results themselves do not disclose whether the spectral clustering actually captured the targeted sub-problems or if other beneficial groups of training samples are formed in the process. Since the multi-dimensional problem space cannot be readily visualized and no ground truth of which patches represent which sub-problem exists, only indirect evidence can be presented. To this end, first a clear improvement in some of the areas associated with the identified sub-problems could be observed during the visual examination of the segmentation results, indicating that the chosen problem space formulation successfully captures these sub-problem. And second, when investigating the groups formed after the mutual kNN application to the patches during the spectral embedding process (see the two examples in Fig. 5.6, right side), it can be observed that some of these suitably represent the targeted sub-problems.

It should be kept in mind that most patches do not belong to any sub-problem at all, some only partially to one or multiple of them and others might form an unidentified but reasonable additional sub-problem. Therefore, more cluster than targeted sub-problems are and should be selected.

In the context of DFs, the proposed method can be regarded as a type of guided super-bagging, which allows for the training of specialized classifier ensembles. The main challenge in successfully applying the proposed LPFs consists in the identification of the sub-problems, the subsequent choosing of a suitable representation and finally the definition of a distinguishing distance metric. The modeling of local sub-problems is hence not limited to spectral location, such as the vicinity of the ventricles, but can capture also other relations, e.g., the presence of

periventricular WMHs, as can be readily observed in the achieved patch groupings (Fig. 5.6, right side).

Thanks to the flexibility of the proposed method, the procedure can be generalized to any group of sub-problems that can be suitable represented and distinguished. E.g., Lombaert *et al.*, 2014 have applied a similar approach to cluster whole training cases, which corresponds to an application at image rather than patch level.

Owing to its modular set-up, the approach is not limited to decision forests, but could be used with any voxel-wise, or, slightly modified, even patch-based classifier. These include support vectors machines (SVMs) and convolutional neural networks (CNNs).

In the proposed version, the LPF method can only be applied to handle a single group of sub-problems. Theoretically, a clustering based on the definition of multiple groups of sub-problems simultaneously is imaginable, but would require to conveniently join the different problem spaces without loss in expressiveness. If such a method exists remains to be shown.

In the case of classification problems where no suitable sub-problems can be identified, the proposed method is unlikely to lead to better results. This is, e.g., the case when multi-spectral MRI images are employed, which seem to provide sufficient information in themselves to overcome the identified challenges.

5.4 Conclusion

In this chapter, a new method to train sub-problem specialized classifiers is presented that can act as an ensemble classifier termed LPFs to improve segmentation through classification. At the example of sub-acute stroke lesion segmentation from mono-spectral FLAIR images the approach is shown to produce significantly better results than traditional voxel-wise DFs. A visual evaluation (Fig. 5.10) demonstrated that performance indeed improved for the targeted sub-problems.

A comparison to an alternative, k-means based version showed that the form of the sought clusters in problem space, albeit unknown, is better captured by spectral clustering than by the linear k-means. The decision to use fuzzy forest catchment areas for training and application allows for a more efficient handling of classification characteristics shared over multiple sub-problems. The inferior results of the random variant indicate that an unguided partitioning of the problem space has no beneficial effects on the segmentation results.

Essential for the success of the method is the careful selection of the distance function, as it must be able to disclose the affiliations within a sub-problem. Here, the relative bin derivation proved suitable for stroke lesion segmentation, which might be attributable to its scale and rotation invariance.

Notable disadvantages compared to the standard approach are the additional hyperparameters and the increased training and testing times, owed to the enlarged training set and the required patch distance computations. The memory and time restriction impeded an application to the public challenge data (as presented in Sec. 4.5). Furthermore, many of the specific sub-problems targeted with the chosen configuration are already solved employing multi-spectral data. And finally, while theoretically adaptable to many different types of sub-problems, the method can always just model one category of these as the clustering is performed according to a single characteristic. To model two or more sub-problem types, the clustering would have to be performed according to multiple characteristics simultaneously.

For the future, it would be interesting to use overlapping patches during the application phase, to apply the method to other segmentation tasks and with other classifiers. To speed up the Laplacian Eigenmap computation and test patch embedding, manifold forests [Gray *et*

al., 2013] might be employed, which automatically learn patch similarity from different features through clustering trees [Moosmann *et al.*, 2008].

Chapter 6

Semi-supervised forests for longitudinal MS lesion segmentation

Multiple Sclerosis (MS) is an increasingly, but irregularly worsening disease, which usually progresses over many years. As detailed in the introduction to the pathology in Sec. 3.2, its treatment involves the regular acquisition of brain magnetic resonance imaging (MRI) scans to monitor the lesion activity. Depending on the local treatment protocol, the cycle can span between half and a full year, with additional scans performed after every clinical attack.

From time point (TP) to TP, lesions can appear and disappear, grow and shrink, and alter their appearance. The lesion morphology over time is an important clinical parameter to assess disease progression and treatment effectiveness (see Sec. 3.2.3 for a motivation of MS lesion segmentation). In practice, this means that large amounts of longitudinal data is available and that the desired lesion segmentations should be consistent over the TPs to draw the correct conclusions.

In Chapter 4, MS lesion segmentation is addressed with the brain lesion segmentation framework developed and described in the same chapter, obtaining competitive results in the independent ISBIMS evaluation benchmark. But that method did not make use of the longitudinal conformity of the data nor did it include mechanisms to ensure temporal consistency.

The same segmentation challenge provided multiple TPs for each patient and longitudinal coherence formed part of the evaluation scheme. Surprisingly, not a single participating approach made use of this additional information, a fact that was intensively discussed during the workshop. Multiple teams stated that they undertook an effort to incorporate longitudinal consistency, but failed to observe a beneficial effect and hence discarded this course of action. It remained unclear whether this failure originated from their algorithmic designs or from the wide range of morphological and appearance changes the MS lesions are subject to.

This chapter introduces a semi-supervised forest (SSF) method to assess if the longitudinal information can have a positive effect on the segmentation accuracy and which combination of segmented and un-segmented TPs leads to the temporally most consistent segmentation.

Literature review on semi-supervised classification

Labeling training samples is a tedious and time consuming process, while unlabeled training data is often available in abundance. The idea of semi-supervised classifier training is to make use of the unlabeled data's distribution in guiding the otherwise supervised model learning process.

This approach may or may not have a beneficial effect, depending on whether the smoothness and cluster assumptions underlying the semi-supervised classification model hold for the problem at hand [Chapelle *et al.*, 2010, Chapter 1]. If not, its application can even lead to inferior results by misguiding the model learning.

Various methods for semi-supervised classifier training have been proposed over the years, most of which are nicely summarized and discussed in an overview by Zhu, 2005. The most basic are the heuristic approaches like self-learning [Triguero *et al.*, 2015] and co-training [Didaci *et al.*, 2012]. A second class is formed by graph-based methods, such as graph cuts [Boykov *et al.*, 2001] or random walks [Grady, 2006] and yet another by generative models, e.g., based on Bayes' theorem. Finally, a last class comprises of the low-density separation approaches, which aim to find boundaries in areas with few samples. Transductive Support Vector Machines (TSVM) [Joachims, 1999] belong to this last group.

Attempting semi-supervised classification with decision forests (DFs) has received comparatively little attention. An exception are self- and co-training methods, which are both techniques combining multiple supervised classifiers to mutually amplify and correct each other. Hence, they are independent of the concrete classifiers employed and can also be used with DFs. E.g., Chandna *et al.*, 2010 showed that a recent approach [Driessens *et al.*, 2006], which is similar to self-learning, improves the classification accuracy compared to classical DFs trained on the labeled data only.

But these approaches do not make use of the DFs training procedure. Badrinarayanan *et al.*, 2013 claim to employ SSFs, but a closer inspection of their description reveals that they simply use soft instead of crisp class labels, which can be solved with classical DFs based on the information gain with Shannon entropy split criterion.

A first serious attempt was made by Leistner *et al.*, 2009. They describe a method which exploits the maximum margin optimization properties of the DFs. The labeled data optimization term is reformulated as maximum margin optimizer and a second term added, which simultaneously optimizes the unlabeled data maximum margin taking into account the labels it would get assigned. To solve this complex problem, deterministic annealing, implemented in the form of a temperature term, is used which results in a complex optimization and large runtime.

Another approach is taken by Liu *et al.*, 2013, who observed that supervised classification is guided by labels while unsupervised classification is usually guided by data density, and hence attempt to combine both. For the supervised part the traditional information gain is used, while for the unsupervised part they employ kernel density estimation. Since this tends to be unfeasible for and unstable in high dimensional feature space (curse of dimensionality), they first transform the sample at a node into a 1-D space, defined by the axis perpendicular to the separating hyperplane investigated. In this space they then estimate the density through a Gaussian and try to find a split separating along the lowest density and simultaneously according to the best class separation. Drawback is that their approach requires the use of oblique, rather than the simpler univariate (i.e., axis-aligned) split function to allow for a multidimensional meaningful density estimation.

This is picked up, generalized and improved on by Criminisi *et al.*, 2013, where the density is estimated in the higher dimensional space directly, allowing for the utilization of simple, axis-aligned splitting functions. Unfortunately, they describe their ideas only on a mainly theoretical level, giving only simple toy examples. This chapter first describes their method, which can be

classified as low density separation just as TSVMs. Next, a number of substantial improvements are proposed, which enable an application to large sets of high dimensional data, as often encountered in medical image processing. The resulting new method is freely available as ready-to-use Python package [Maier, 2016c]. Finally, the proposed SSF is applied in various MS lesion segmentation scenarios to demonstrate its performance and to establish whether MS segmentation benefits from a semi-supervised approach.

6.1 Method

The SSFs used in this work are based on Criminisi *et al.*, 2013, Part I, Chapter 8, where they are explained in detail. Two modifications to the supervised DF model (see Sec. 4.1) are necessary to enable the handling of partially labeled data. First, a new information gain term for the node splitting is defined, which consists of the weighted sum of the classical entropy used for classification forests and the upper bound of the differential entropy employed for density trees. Second, a label propagation mechanism is required following the tree training to label the unlabeled portion of the training data a posteriori. These two steps and how they enable semi-supervised classification are described in the next two sub-sections 6.1.1 and 6.1.2.

Criminisi *et al.*, 2013 provide an example implementation for SSFs applicable only for 2D feature spaces, which severely limits the practical application of the algorithm. In parts, their decision might be attributable to the complexity of implementing the proposed semi-supervised model for arbitrary dimensionality and the numerous pitfalls of numeric instability. To overcome some of these problems, a method for dynamic statistical co-variance matrix update is proposed in Sec. 6.1.3 that is publicly available as independent implementation [Maier, 2016a]. Then, an alternative method for label propagation based on label diffusion is introduced in Sec. 6.1.2, which simplifies the costly label transduction process at minimal loss of precision. Next, further implementation details are discussed and, finally, the Cython and Python based, ready-to use extension *sklearnef* [Maier, 2016c] for the *sklearn* toolbox [Pedregosa *et al.*, 2011] is presented in Sec. 6.1.4.

6.1.1 The node split optimization term

A detailed account of growing a DF was previously given in Sec. 4.1. In short, a tree is trained from the root upwards node by node, at each of which all possible splits of the set of incoming samples S into two distinct subsets S_L and S_R are explored and the best one according to an optimization criterion is selected (see Fig. 4.2). Upon reaching a stop criterion, a terminal leaf node is formed.

Short review on the supervised term A node is trained by maximizing the optimization criterion defined in Eq. 4.5, which is based on the information gain, here repeated from Eq. 4.6

$$IG(S, S_L, S_R) = g(S) - \frac{|S_L|}{|S|} g(S_L) - \frac{|S_R|}{|S|} g(S_R), \tag{6.1}$$

with $g(S)$ denoting a measure of entropy (disorder or uncertainty), which ideally is decreasing and positive. This definition of information gain differers from the usual usage as a synonym for the Kullback–Leibler divergence [Kullback *et al.*, 1951], but rather resembles the mutual information.

For the supervised version IG^{sup}, to which Sec. 4.1 simply refers to as IG, the (discrete) Shannon entropy is employed to measure the uncertainty in a set according to the distribution

of all classes C in a set of samples S

$$g(S) = E(S) = -\sum_{c \in C} p(c|S) \log_2 p(c|S), \tag{6.2}$$

i.e., the more evenly filled the class histogram bins the higher the entropy. Subsequently

$$IG^{sup}(S, S_L, S_R) = E(S) - \frac{|S_L|}{|S|} E(S_L) - \frac{|S_R|}{|S|} E(R). \tag{6.3}$$

Un-supervised term Importantly, semi-supervised classification signifies that only a (usually comparatively small) part of the training samples are labeled and the Shannon entropy cannot be computed over the unlabeled majority. Criminisi *et al.*, 2013, Part I, Chapter 6 propose instead to employ a measure of the subsets density to determine the goodness of a split. The appropriate continuous version of the (discrete) Shannon entropy is the limiting density of discrete points (LDDP) as introduced in Jaynes, 1957. Criminisi *et al.*, 2013 instead fall back to the notion of differential (continuous) entropy as proposed by Shannon *et al.*, 1964

$$E_d(S) = -\int_S PDF_S(s) \log PDF_S(s) ds. \tag{6.4}$$

In comparison with the LDDP, this definition lacks some of the convenient properties of the (discrete) Shannon entropy [Marsh, 2013]. Namely, it is variant under change of coordinates ($E_d(S+a) \neq E_d(S)$), variant under change of variable scale ($E_d(S*a) = E_d(S) + \log|a|$) and not restricted to positive values (e.g., in the case of a uniform distribution in $[0, \frac{1}{2}]$). Nevertheless, it is easier to compute and serves its purpose when used to compare entropy between distributions.

Considering a set of incoming samples S, the real probability distribution denoted by its PDF, PDF_S, from which these samples originate, is unknown. Hence, the working assumption that S is normally distributed is introduced. The differential entropy of a multivariate normal distribution is defined as

$$E_d(S) = \frac{1}{2} \log \left[(2\pi e)^{D_F} |\Sigma(S)| \right], \tag{6.5}$$

where $\Sigma(S)$ is the $D_F \times D_F$ statistical co-variance matrix and $|\cdot|$ its determinant [McEliece, 1977]. Note how the differential entropy of a multivariate normal distribution only depends on the Gaussian's co-variance Σ but not on its mean $\boldsymbol{\mu}$, i.e., it's translation invariant unlike Eq. 6.4, removing one of the undesired attributes of the differential entropy. Furthermore, $E_d(S)$ is monotonically related to the determinant of the co-variance matrix. And since the determinant is a form of measuring the spread of dispersion of the underlying data S, the entropy captures the density of a set of samples in feature space. In other words, a "good" split is a split after which the resulting subsets form a dense, Gaussian-like shape in the feature space. This leads to compact clusters of points, which in turn are assumed to share the same labels (see Fig. 6.1).

The introduced working assumption of a Gaussian distribution becomes possible in this application case since, for a given statistical variance, the probability density with the greatest differential entropy is the Gaussian density (demonstrated, e.g., in Goldman, 1953 and Conrad, 2013). Hence, the optimization criterion operates with the upper bound or maximum error.

During node optimization, the goal is to determine which of all regarded splits divides a set of incoming samples S into its two distinct subsets S_L and S_R with the maximum proportional density respectively minimum proportional differential entropy. Hence, for the un-supervised version of the information gain IG^{usup}, the differential entropy of a multivariate normal distribution E_d is chosen as measure of disorder, leading to

$$IG^{usup}(S, S_L, S_R) = \log|\Sigma(S)| - \frac{|S_L|}{|S|} \log|\Sigma(S_L)| - \frac{|S_R|}{|S|} \log|\Sigma(S_R)|. \tag{6.6}$$

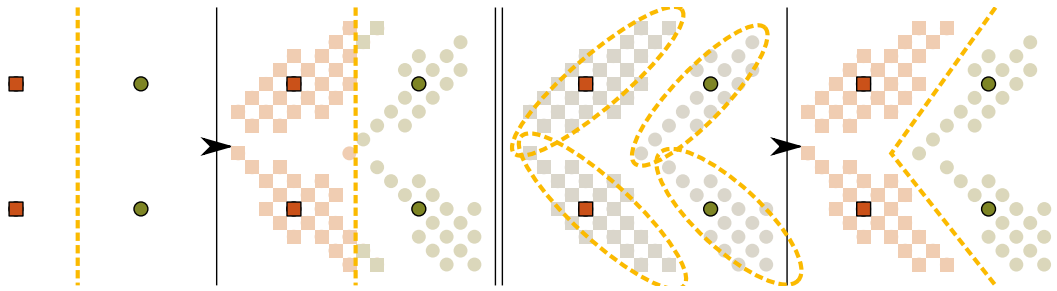


Figure 6.1: The left side denotes the difficulties of the supervised DF to capture the real class boundaries when only few labeled training samples (black outlined) are available. The right side denotes the possible improvement when additionally employing unlabeled samples to place the splits in areas of low sample density. A sample's real class memberships is denotes by its shape, the classification result by its color. Refer to online version for colors.

Semi-supervised term For un-supervised tree training, the supervised and un-supervised information gain terms are combined according to Criminisi *et al.*, 2013 into

$$IG^{ssup}(S, S_L, S_R) = \alpha IG^{sup}(S^{sup}, S_L^{sup}, S_R^{sup}) + (1-\alpha) IG^{usup}(S, S_L, S_R), \tag{6.7}$$

where IG^{sup} is only computed over the labeled subset S^{sup} of the incoming sample set S. This formulation leads effectively to trees that split the feature space into compartments containing dense clusters of training data with uniform labels.

So far, this section presented the theoretical work of Criminisi *et al.*, 2013, extended by an explanation of the theoretical background and implications.

Limiting the codomain of the un-supervised term As discussed above, the differential entropy E_d is variant under change of variable scale and not restricted to positive values, even under the assumption of a Gaussian distribution. Since $\Sigma(S)$ is positive semi-definite it follows that $|\Sigma(S)|$ in $[0, +\inf]$ and hence $E_d(S)$ in $[-\inf, +\inf]$. The Shannon Entropy E, on the other hand, is bound by $[0, 1]$ and monotonically decreasing with the number of samples, thus IG^{sup} in $[0, 1]$. Unfortunately, Criminisi *et al.*, 2013 fail to discuss the imbalance between the terms of the weighted information gain in their work.

This setting renders it highly difficult to find a suitable weight α for Eq. (6.7) to balance the supervised against the un-supervised term IG^{ssup}. Even more inconveniently, the ideal value for α can change while descending the tree if a feature contains information at multiple scales. To overcome this problem, IG^{usup} has to be adapted to the same codomain as IG^{sup}.

Three changes to the unsupervised term are proposed to shift and scale its codomain. First, a lower constant limit $\hat{C}$ is imposed on the determinant

$$E_d(S) = \log\left(\max(|\Sigma(S)|, \hat{C})\right). \tag{6.8}$$

Since the determinant is effectively a measure of the volume of the Gaussian distribution represented by $\Sigma(S)$, it takes on small values only for very dense clusters of samples. By choosing $\hat{C}$ sufficiently small, this condition only applies when the density of S anyway suggest itself for forming a tree leaf. In practice, it acts as an additional stop criteria for tree growth. $\hat{C}$ is fixed to 10^{-6}, such that E_d in $[log(10^{-6}), +\inf]$.

Second, a constant is added

$$E_d(S) = \log\left(\max(|\Sigma(S)|, \hat{C})\right) + \log(\hat{C}), \tag{6.9}$$

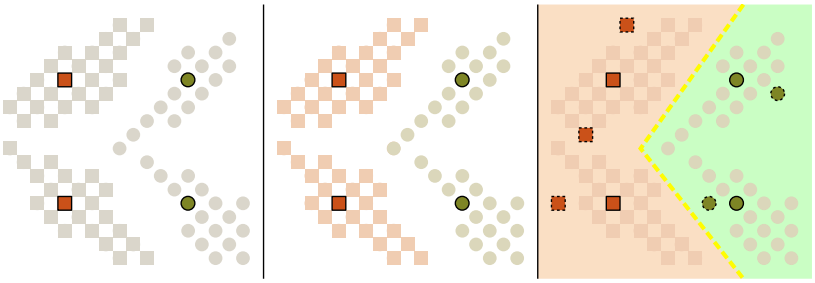


Figure 6.2: Left: Training data with labeled (black outlined) and unlabeled (gray) samples. The shapes denote their real class memberships. Middle: Label assignment after transduction step (pale colors). Right: Labeling formerly unseen samples (dashed outline) through induction with the dashed yellow line and background colors denoting the forests decision boundary. Refer to online version for color.

which effectively shifts the codomain of E_d to $[0, +\inf]$. This step is possible since IG^{usup} compares entropy values which are all effected by the change.

Third, a normalization step is added

$$IG^{usup}(S, S_L, S_R) = 1 - \frac{|S_L|}{|S|}\frac{\log\left(|\Delta(S_L)|\right)}{\log\left(|\Delta(S)|\right)} - \frac{|S_R|}{|S|}\frac{\log\left(|\Delta(S_R)|\right)}{\log\left(|\Delta(S)|\right)}, \tag{6.10}$$

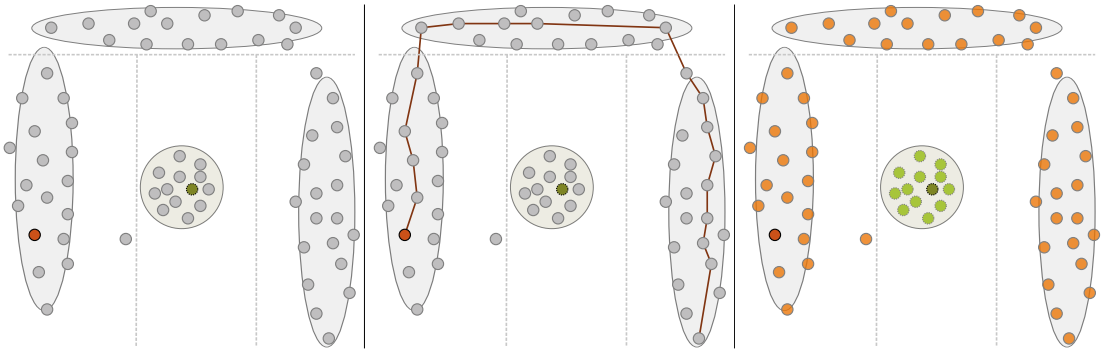
such that finally IG^{usup} in $[0, 1]$. This step is possible since $det(S) \leq det(S_L)$ if $S_L \subseteq S$. Since this normalization is unique to each node, the scale of the data reaching the node does not exercise any influence anymore.

6.1.2 Transduction

After fully growing the tree using the described semi-supervised splitting function, a mix of labeled and unlabeled samples end up in the tree leafs. It remains the task of labeling the unlabeled data. Two types of labeling are associated with semi-supervised classification: Transduction describes the propagation of labels from the labeled part of the training data to the unlabeled part. Induction, on the other hand, denotes the subsequent labeling of previously unseen data (see Fig. 6.2). The implementation of induction is trivial for SSF with the induction through transduction approach described in Criminisi *et al.*, 2013. This work concentrates therefore on transduction, of which three different strategies are introduced in this section

Transduction through label propagation on a dense geodesic surface During training, a tree splits the feature space into small compartments as depicted in Fig. 6.3a, where the dashed split lines are the learned decision borders, which ideally run along areas of low density. The compartments represent the tree leaves, each containing any combination of labeled and unlabeled samples. These leaf samples denote a density, which, under the working assumption of a Gaussian distribution, can be represented by their sample co-variance matrix and sample means. In the figure, these are denoted by gray ellipses.

Combined, these Gaussians form a piece-wise multivariate Gaussian distribution in feature space that constitutes the learned, non-normalized probability density function (PDF) of the training data. Transduction is performed by searching the shortest path from each unlabeled to each labeled sample along the surface defined by this PDF and subsequently transferring the label.



(a) The feature space fragmentation after tree training. Each compartment represents a leaf.

(b) Shortest path from the gray sample in the lower right corner to its nearest labeled sample.

(c) The resulting labeling of the unlabeled samples with full dense label propagation.

Figure 6.3: Transduction on a dense geodesic surface. Gray points denote the unlabeled samples, red with solid outline and green with dashed outline the respectively labeled ones. Gray ellipses represent the multivariate Gaussian sample distributions, dashed gray lines the learned decision boundaries. Refer to online version for color.

Deriving a geodesic distance on a surface formed by piece-wise multivariate Gaussian distributions is a complex task and likely to be connected with large computational costs. Criminisi *et al.*, 2013 propose an approximation scheme instead, which makes use of the fact that not only the PDF, but also the training samples from which it was derived are known.

First, a semi-metric is defined between any two samples $\mathbf{x}$ and $\mathbf{y}$ based on the Mahalanobis distance

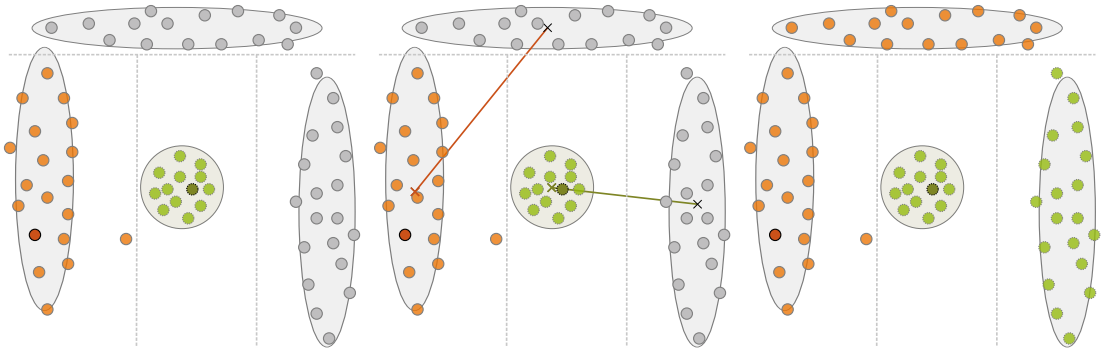
$$d_m(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\left(\sqrt{(\mathbf{x}-\mathbf{y})^\top \Sigma_{\mathbf{x}} (\mathbf{x}-\mathbf{y})} + \sqrt{(\mathbf{y}-\mathbf{x})^\top \Sigma_{\mathbf{y}} (\mathbf{y}-\mathbf{x})}\right) \tag{6.11}$$

where $\Sigma_{\mathbf{x}}$ denotes the sample co-variance matrix associate with the leaf to which the sample $\mathbf{x}$ belongs. As the Mahalanobis metric is a measure of the distance between a point and a distribution, the given definition of $d_m(\mathbf{x}, \mathbf{y})$ denotes the semi-distance between two samples taking into account their source distributions.

Next, the semi-distance is computed pair-wise between all training samples, resulting in a fully connected, undirected distance graph. By resolving the shortest path from each unlabeled sample to each labeled sample, the label transduction is achieved and the labels for the unlabeled portion of the training data are returned.

Since the given semi-distance does not satisfy the triangle inequality, a path from $\mathbf{x}$ to $\mathbf{y}$ can be shortened by taking a route via other samples lying in areas of high sample density. This can be seen in Fig. 6.3b, where the red line denotes the shortest geodesic path from an unlabeled sample to the nearest labeled sample.

This approach leads to very accurate results (see theoretical schema in Fig. 6.3c and experimental results in Sec. 6.2.1), but suffers from a high computational complexity of $\mathcal{O}(MN^2)$ depending on the number of training samples N and feature space dimensionality M. In preliminary experiments, the full, dense label propagation was found to run for approximately an hour for as few as 10.000 samples with 5 dimensions. In the domain of medical image processing one easily faces millions of samples with high feature dimensionality, thus preventing an application of this approach.



(a) The sample labels after the first step of the approximate label propagation.

(b) Shortest path from the unlabeled leaf clusters to the nearest labeled cluster.

(c) The resulting labeling of the unlabeled samples with approximated label propagation.

Figure 6.4: Transduction on an approximated geodesic surface. Gray points denote the unlabeled samples, red with solid outline and green with dashed outline the respectively labeled ones. Refer to online version for color.

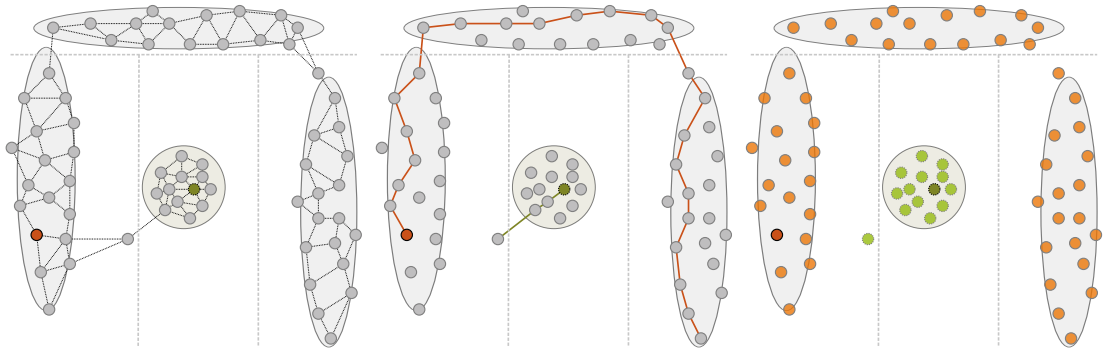
Transduction through approximated label propagation Criminisi *et al.*, 2013 shortly discuss and use a highly simplified approximation of the above described method. Under the assumption that samples in a leaf share the same class membership, all leaves containing at least one labeled sample are processed first. A class occurrence histogram (i.e., counting the label occurrences into a class histogram) is constructed from all labeled samples present in a leaf node and simply transfered to all unlabeled samples as shown in Fig. 6.4a.

Next, the semi-distance as defined in Eq. (6.11) is computed between all cluster centers, again assuming that samples in the same leaf share the same class membership. By searching for the shortest path from each unlabeled cluster to the nearest labeled cluster, the label propagation is performed (see Fig. 6.4b).

This variant leads has a complexity of $\mathcal{O}(L^2)$, with L denoting the number of tree leaves. Since nearly always $L << N$, it can be computed much faster and therefore be applied to huge training sets and high dimensional feature space. On the downside, it does not follow the curved density distributions (see Fig. 6.4c) and includes some very rough approximations.

Transduction through label diffusion To overcome the problems of the two above presented label propagation methods, label diffusion is proposed. The idea is based on the work of Zhu *et al.*, 2003, who propose a graph solver with similarities to random walks, electric networks and normalized cuts [Boykov *et al.*, 2001]. However, unlike these alternatives, their proposition uses Gaussian fields over a continuous state space rather than random fields over the discrete label set. The proposed energy formulation can be solved globally uniquely for multi-class problems and has a closed form solution computable with matrix methods, two characteristics which make it a ideal candidate for the label propagation task.

Effectively, the proposed algorithm is a semi-supervised classifier in itself, which determines for a number of unlabeled samples S^{usup} the probability to belonging to any of the known classes $c \in C$ by their distance to a set of labeled samples S^{sup}. Required input is a weight matrix $W = [w_{i,j}]$ denoting the weights between any two samples in $S = S^{usup} \cup S^{sup}$. This weight matrix W specifies the data's manifold structure, which is honored by the solver. Note that W can be sparse.



(a) The sample neighborhood graph connecting all samples after kNN.

(b) Shortest path from two unlabeled samples to their nearest labeled sample in the graph.

(c) The resulting labeling of the unlabeled samples portion with label diffusion.

Figure 6.5: Schema of transduction on an sparse geodesic surface with label diffusion. Gray points denote the unlabeled samples, red with solid outline and green with dashed outline the respectively labeled ones. Refer to online version for color.

Let $D = diag(d_i)$ be a diagonal matrix with $d_i = \sum_j w_{i,j}$. Let further D_{uu} and W_{uu} be the blocks of their respective base matrices D and W which denote the relationships between the unlabeled samples only. Equally, let W_{ul} be the block of W holding the weights from the unlabeled to the labeled samples.

For each $c \in C$ the vector $\mathbf{p}_{u,c}$ denoting the probability of each unlabeled sample in S^{usup} to belong to the class c can be obtained by solving

$$\mathbf{p}_{u,c} = (D_{uu} - W_{uu})^{-1} W_{ul} \mathbf{b}_{l,c}, \tag{6.12}$$

where $\mathbf{b}_{l,c}$ is a binary vector with ones for each sample in S^{sup} belonging to c and zeros otherwise. This equation can be solved efficiently with matrix solvers and, executed once for each label, provides a globally optimal solution for the label propagation problem.

To employ the solver of Zhu *et al.*, 2003, the data manifold structure, as expressed through the weight matrix W must be known. This information is extracted by the forest from the data directly by learning the PDF from which the training data presumably originated. Thus, the graph of all samples S embedded into the full geodesic surface as described in above paragraph on transduction through label propagation on a dense geodesic surface provides the data manifold structure and hence the weight matrix W.

As mentioned before, the computation of all pairwise geodesic distances between the samples is computationally costly. In this thesis, the following approximation procedure to obtain a sparse version of W is therefore proposed. Starting point are the fully trained trees with each leaf containing a set of samples and corresponding sample Gaussian distribution. The piece-wise conjunction of these Gaussian distributions covers the whole feature space and, appropriately normalized, constitutes the learned PDF. As first step, a k-Nearest-Neighbors (kNN) is applied to all samples under the assumption that local semi-distances on the geodesic surface can be approximated by the Euclidean distance. This leads to a sample connectivity as displayed in Fig. 6.5a. The Euclidean distance computation is by magnitudes cheaper than the Mahalanobis distance and efficient approximative algorithms such as the KD-Trees [Bentley, 1975] exist.

To obtain the weights between all neighbors, the semi-distance previously defined in Eq. (6.11) is used, i.e., $w_{i,j} = d_m(\mathbf{s}_i, \mathbf{s}_j)$, resulting in the desired sparse, undirected weight matrix W. This

is now plugged into Eq. (6.12) and solved once for each label. The result is equivalent to the shortest path to the nearest labeled sample as shown in Fig. 6.5b.

The proposed label diffusion approach leads to results very similar to the full, dense distance graph method (see Fig. 6.5c for a schema and Sec. 6.2.1 for experimental results) at a fraction of the computational costs. The complexity $\mathcal{O}(MNk)$ depends only linearly on the number of samples N and the efficient usage of optimized and iterative solvers allows for an application to large, high-dimensional datasets as found in medical image processing.

6.1.3 Dynamic statistical co-variance matrix update

The estimation of the differential entropy, as defined in Eq. (6.4), requires to compute the statistical co-variance matrix of the sample sets for each split investigated at every node. Considering training sets of many thousand samples, this soon becomes computational inefficient. To overcome this problem, the Dynamic Statistical Co-Variance (DynStatCov) method is proposed.

DynStatCov is a Cython Library for fast dynamic statistical co-variance update. It is intended for usage in applications, where a statistical co-variance matrix has to be computed from observations and periodically updated. The naïve approach would require to re-compute the co-variances every time from all samples and to hold them in memory. By rewriting the equations, the computation can be realized using intermediate sums and thus achieving a higher computational speed.

Consider the case of node optimization during DF training and recall the associated procedure (see Fig. 4.2). At a given time, all possible splits t of the incoming set S according to a feature d are explored. For an effective processing, S is sorted according to d and split into the two disjunct subsets S_R and S_L, presumably starting out with $|S_L| = 1$ and $|S_R| = |S| - 1$. Investigating the next split according to the subsequent thresholds is now replaced by simply passing the left-most sample from S_R to S_L.

Following this scheme, the DynStatCov method allows to compute the statistical co-variance matrices initially once for S_L and S_R and then update them iteratively according to the added respectively removed samples. The procedure proposed is as follows.

Reformation The procedure described in this paragraph is a standalone method. All variable definition used here should therefore be treated independent from the remainder of the thesis.

Given a sample of n independent observations $\mathbf{x}_1, ..., \mathbf{x}_n$ of length m, the sample (statistical) co-variance matrix of $X \in R^{n \times m}$ is given by

$$Q = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mathbf{x}})(\mathbf{x}_i - \hat{\mathbf{x}})^T \tag{6.13}$$

where $\mathbf{x}_i$ denotes the i-th observation and

$$\hat{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \tag{6.14}$$

is the sample mean. Rewriting the first equation leads to

$$Q = \frac{1}{n-1} \left[\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T - \hat{\mathbf{x}} \left(\sum_{i=1}^{n} \mathbf{x}_i \right)^T - \left(\sum_{i=1}^{n} \mathbf{x}_i \right) \hat{\mathbf{x}}^T + n \hat{\mathbf{x}} \hat{\mathbf{x}}^T \right] \tag{6.15}$$

which is essentially the application of $(a-b)^2 = 2a^2 - 2ab - b^2$. Substituting the sums with A resp. $\mathbf{b}$ gives raise to

$$Q = \frac{1}{n-1}\left[A - \hat{\mathbf{x}}\mathbf{b}^T - \mathbf{b}\hat{\mathbf{x}}^T + n\hat{\mathbf{x}}\hat{\mathbf{x}}^T\right] \tag{6.16}$$

Updating From this form, an efficient update of Q_{n+1} when a new sample $\mathbf{x}_{n+1}$ becomes available is derived. First A and $\mathbf{b}$ are updated

$$A_{n+1} = A_n + \mathbf{x}_{n+1}\mathbf{x}_{n+1}^T \tag{6.17}$$
$$\mathbf{b}_{n+1} = \mathbf{b} + \mathbf{x}_{n+1} \tag{6.18}$$

then the new sample mean is calculated

$$\hat{\mathbf{x}}_{n+1} = \frac{1}{n+1}\mathbf{b}_{n+1} \tag{6.19}$$

and finally the updated co-variance matrix computed

$$Q_{n+1} = \frac{1}{n}\left[A_{n+1} - \hat{\mathbf{x}}_{n+1}\mathbf{b}_{n+1}^T - \mathbf{b}_{n+1}\hat{\mathbf{x}}_{n+1}^T + (n+1)\hat{\mathbf{x}}_{n+1}\hat{\mathbf{x}}_{n+1}^T\right] \tag{6.20}$$

Complexity and speed With the default formulation in (6.13), the complexity of each threshold's evaluation amounts to $\mathcal{O}(nm^2)$, with the DynStatCov in (6.20) only $\mathcal{O}(m^2)$. A fast Cython implementation is available for free online [Maier, 2016a]. A simple speed test using 10000 initial observations of 3 features each and then computing an update upon arrival of a new observation lead to average runtimes of:

- complete re-computation with numpy.cov: **553us**
- dynamic update implemented in Python: **66us**
- DynStatCov dynamic update implemented in Cython: **0.804us**

6.1.4 Measures taken against numerical instability

The computation of the differential entropy is not trivial and numerically unstable, especially in a higher dimensional feature space or when only few samples remain. Criminisi *et al.*, 2013 restricted their exemplary implementation, the Sherwood library, to 2D, for which these problems do not arise. To allow an application to higher dimensional data, various numerical problems have to be overcome.

Training data scaling Features with varying variances do not pose a problem for the method, especially after the proposed un-supervised information gain normalization step. But the magnitude of the features can influence the training process. This is caused by numerical instabilities in the co-variance matrix, which cause a zero valued determinant for features with absolute values below one. To counter this effect, data scaling is advisable. *Sklearnef* therefore implements an independent scaling of all training data dimensions to unit variance, followed by a multiplication with a high value (e.g., 10^8) to counter the zero determinant effect.

Minimal samples per leaf Another issue with the sample co-variance computation is that it requires a minimal amount of samples to be computed. The stop criterion is therefore adapted to enforce that each leaf contains at least as many samples as there are feature dimensions.

Regularization of the co-variance matrix Co-variance matrices are positive semi-definite, i.e., they may contain zeros on the diagonal. In the chosen computation, some diagonal values might even be slightly negative due to floating point errors. Since the computation of the determinant requires a positive definite matrix, each co-variance matrix is first regularized by adding a very small value (10^{-6}) to the diagonal to ensure a successful computation of the determinant.

In this section, a SSF method is introduced. To this end, the DF's supervised node optimization criterion is extended by an un-supervised term derived from the differential entropy of Gaussian distributions. Thus, the tree's splits are guided by both, the label purity and the sample density. Next, a diffusion based transduction scheme was proposed to propagate class labels over an approximation of the geodesic surface the learned PDF spans in feature space. Finally, a fast method to update a statistical co-variance matrix upon the arrival or removal of new samples was described. In the next section, the various components of the SSF are evaluated in a number of toy and real world examples.

6.2 Experiments and results

A new SSF variant was presented in the previous section. The following three part evaluation investigates the method in detail and determines whether it can improve longitudinal MS lesion segmentation. The first part (Sec. 6.2.1) compares the three discussed transduction (i.e., full, approximate, and diffusion) methods in terms of memory requirement, runtime and accuracy. The second part (Sec. 6.2.2) investigates the SSF's hyperparameters in an extensive analysis and compares its properties to the classic DF's. The third and final part (Sec. 6.2.3) presents a complete and extensive evaluation on the ISBIMS challenge's longitudinal MS data including a direct face to face comparisons to DFs.

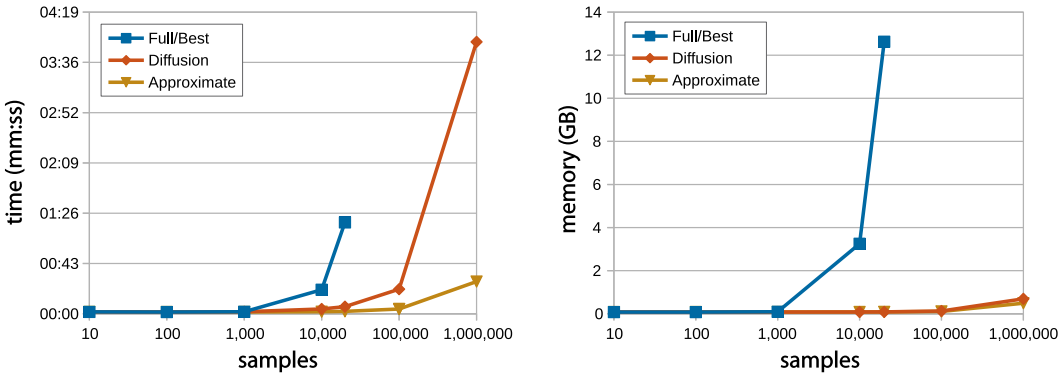
6.2.1 Comparison of transduction methods

A new transduction method based on label diffusion was introduced that is fast enough to handle large amounts of training data, while presumably producing results nearly as accurate as the full nearest label search on the dense geodesic distance graph. To support this claim, a number of toy experiments are conducted.

First, the three methods are compared regarding their runtimes and memory requirements on the same machine for various training sample sizes. The results are displayed in Fig. 6.6. All three approaches were moderately optimized for runtime and memory. Both the full and the diffusion methods increase exponential in runtime, but the latter can still process millions of samples in a short and higher numbers in acceptable time[1], an essential requirement for a real life application to cases easily comprising ten million samples. The approximation runtimes are negligible in comparison. In terms of memory, the diffusion and approximation algorithms demonstrate only moderate requirements that can be fulfilled by any modern computer. The full graph search, on the other hand, requires prohibitively large amounts of memory, preventing its application to medical images.

In a second experiment, the proposed diffusion method is compared to Criminisi *et al.*, 2013's approximation approach in terms of classification accuracy by means of increasingly difficult toy datasets as presented in figures 6.7 to 6.11. Classes are denoted by different colors, labeled samples are depicted as larger circles. The top left graph denotes the ground truth, the top right

[1]The minimal requirement is to process cases at least as fast as they are recorded, which, depending on the medical center, can be more than one or two cases a day.



(a) Runtimes.

(b) Memory requirements.

Figure 6.6: Comparison of the different transduction methods' resource requirements for performing the label propagation in a single tree. Note the log-scales on the x-axis. In practice, all three approaches can be readily parallelized for forests of trees.

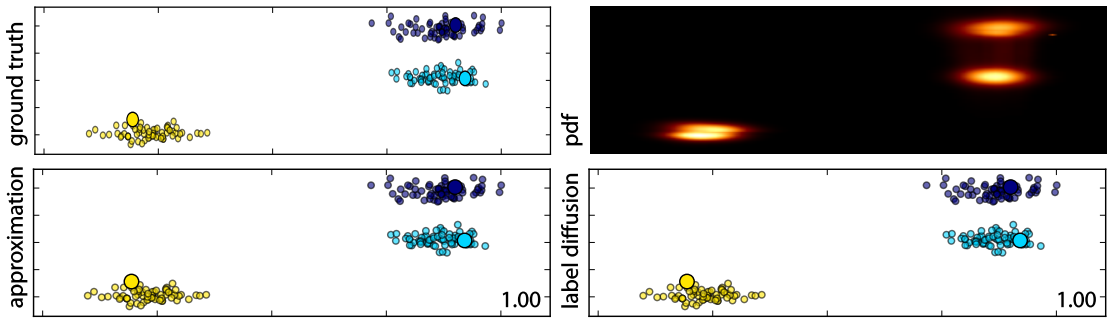


Figure 6.7: Transduction methods comparison: Blobs example. Both methods reached maximum accuracy on these easily separable, Gaussian-shaped blobs. Refer to online version for colors.

graph the PDF learned by the forest in the training process. The bottom row displays the two transduction methods' results with the achieved classification accuracy written in the corner. Please refer to the high-resolution online version of these graphs for a detailed view.

The proposed diffusion approach clearly outperforms the approximation method in all experiments and demonstrated its ability to trace even complex shapes by following the areas of highest sample density. Hence, by employing the label diffusion transduction, it becomes feasible to apply the SSF to multi-dimensional, large-scale problems without loosing the ability to solve complex shaped semi-supervised problems.

It remains to be seen if the MS lesion segmentation poses a) a difficult problem which justifies the application of the label diffusion, b) a simple blob-like shape, equally solvable by the approximation approach or c) if the inherent sample distribution violates the smoothness and/or cluster assumption, such that none of the semi-supervised methods will lead to improved segmentation results.

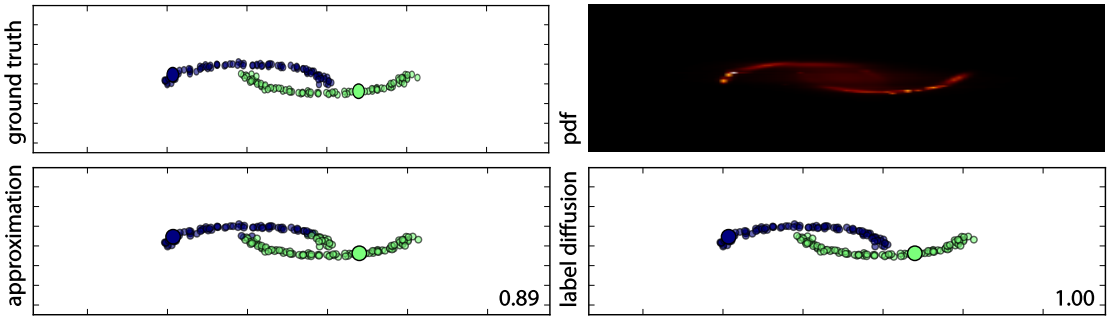


Figure 6.8: Transduction methods comparison: Moons example. The slight curvature causes problems for the approximation based transduction method, while the label diffusion reaches optimal result despite the only labeled blue sample being situated at the outer edges of the structure. Refer to online version for colors.

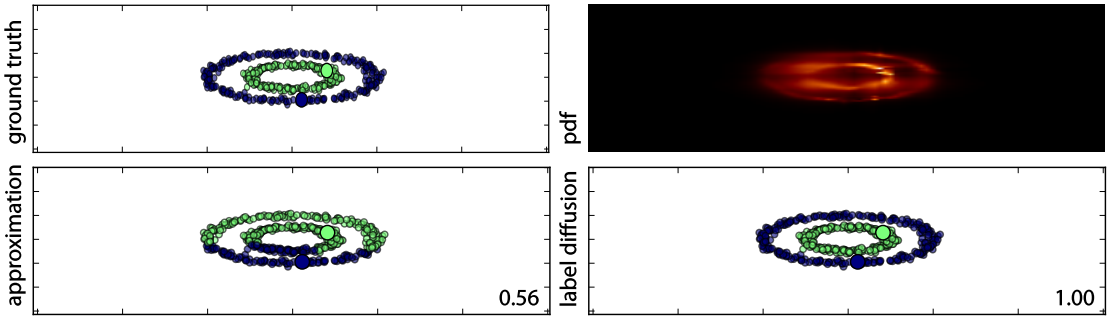


Figure 6.9: Transduction methods comparison: Circles example. The approximations method's difficulties with curvature become apparent in this example, where nearly 180 degrees of curvature have to be traced. The label diffusion, on the other hand, follows the density distribution perfectly. Refer to online version for colors.

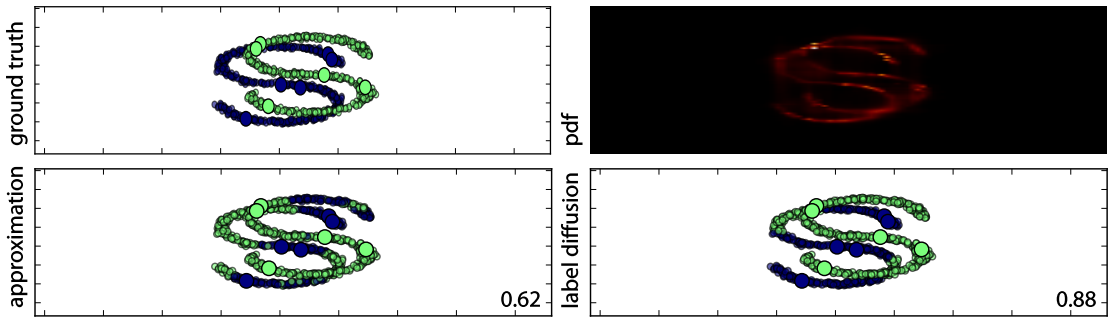


Figure 6.10: Transduction methods comparison: S-curve example. This challenging toy example of two overlapping s-curves cannot be solved perfectly, as the class membership distribution of the samples in the intersecting areas violates the smoothness and cluster assumptions underlying the semi-supervised classification idea. Nevertheless, with at least one labeled sample in the homogeneous compartments between the intersections, the label diffusion approaches the best possible classification result, demonstrating its ability to correctly detect the principal direction of the sample clusters and discouraging sharp turns. The approximation method fails to follow even the homogeneous areas' curvature. Refer to online version for colors.

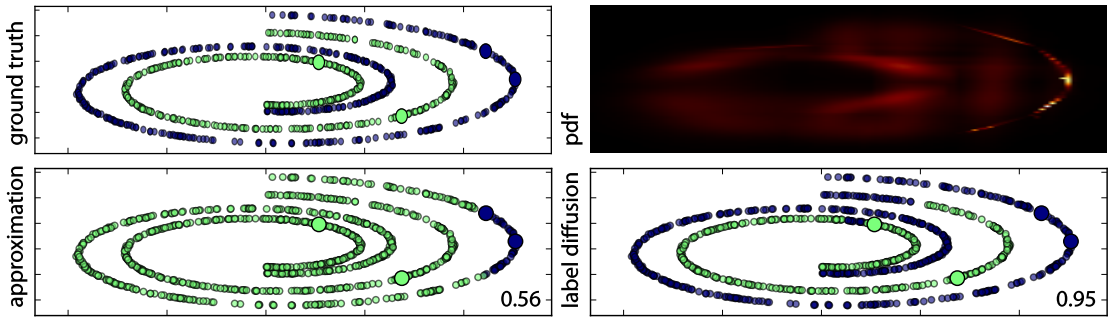


Figure 6.11: Transduction methods comparison: Swiss roll example. In this very difficult example, the label propagation methods have to follow the density shape along multiple iterations of a medium dense spiral without leaping to the proximate second spiral. While the approximate method fails near to completely, the label diffusion classifies most samples correctly. Only a portion separated by a gap from the rest of the spiral is not successfully assigned to the associated class. More unlabeled samples would presumably bridge this gap and reduce the error. Refer to online version for colors.

6.2.2 Hyperparameter analysis

Classifier (Semi-supervised Forest)	
T	10
F_{node}	$\sqrt{F}$
t_{depth}	7
C_{opt}	Gini
Semi-supervised parameters	
N_L	$10,000$
α	0.75
Data parameters	
$lratio$	0.05
$scale$	10^5
Features	
int	
wlm	$\sigma = 3, 5, 7$ mm

Preprocessing	
Resampling	$R_{work} = 1$ mm^3 (*)
Coregistration	MRI_{base} =? (*)
Skull-stripping	MRI_{skull} =? (*)
Bias-field	yes
Intensity range std	yes
Postprocessing	
Thresholding	optimal
Object threshold	none
Hole filling	no

Table 6.1: SSF hyperparameter analysis experimental configuration. Starred steps were performed by the ISBIMS challenge's organizers.

The proposed SSF mechanism differs substantially from the classic DFs, such that the observations made during the latter's hyperparameter analysis (Sec. 4.4.1) are unlikely to apply. This section presents the results of a dedicated SSF hyperparameter analysis performed on the ISBIMS challenge's MS cases to update previous findings and investigate the influence of the newly introduced hyperparameters.

Beside the classic DF parameters, such as the number of trees T and the maximum tree depth t_{depth}, this experiment investigates a number of further parameters. One of them is the number of labeled samples N_L required for a successful semi-supervised classification. The number of unlabeled samples is given by and depends on the test case and ranges between three and four million. Another SSF specific parameter is the supervised weight α as introduced in Eq. 6.7, which balances the supervised against the unsupervised information gain term. Finally, two training data related parameters are investigated: First, compared to DFs, the SSFs might not favor an inherent class balance in the training data, therefore different lesion ratio ($lratio$) values are compared. Second, the effect of data scaling (see Sec. 6.1.4) is explored for different values for $scale$.

Experimental setup and results

In short, a segmented version of a patient's first TP is used to segment all consecutive TPs. To allow for a direct comparison, a classic DF, as described in Sec. 4.1, is likewise trained on the same N_L labeled samples for each investigated set of hyperparameters.

The first TPs from the 5 patients of the ISBIMS challenge's training data serve as training cases, the second and fourth TPs respectively as testing cases, i.e., the presented results are the average over 10 real multi-spectral (T1, T2, PD, FLAIR) MS cases. It is assumed that all observations will be equally true for larger datasets. For a detailed presentation of the data see Sec. 4.5.2. All cases are prepared with the same preprocessing steps as for the main challenge evaluation (Sec. 4.5.2). No postprocessing is employed.

Each TP of each patient is processed individually: The formerly unseen testing cases (e.g., P1TP4) samples constitute the SSF classifier's unlabeled data to be labeled through transduction. Additionally, a number of N_L labeled samples is drawn randomly from the respective patients first TP (e.g., P1TP1) with a fixed lesion-to-background ratio $lratio$ to guide the classification.

To facilitate a straightforward interpretation, only the Dice's coefficient (DC) metric as previously introduced in Sec. 4.3.4 is employed. As the harmonic mean of precision and recall it arguably constitutes one of the best stand-alone measures for binary segmentation assessment.

Thresholding the forest's posteriori class probability map is a major regulatory parameter to balance over- against undersegmentation. Unfortunately, the ideal threshold value can vary considerably depending on the forest's training parameters selected. To ensure a meaningful comparison between the different hyperparameter sets independent of the probability threshold, the latter is set to its optimal value as determined with a step size of 0.1. In other words, every time after a DF or a SSF is trained with a set of hyperparameters and subsequently applied to the 10 testing cases, a single, DC optimizing threshold is determined for these 10 cases as a set. I.e., the optimization is performed per experimental run, not per case.

For the experiments, always one parameter is varied and all others kept at their default value as listed in Table 6.1. The average DC obtained by the SSF respectively DF classifiers over the 10 testing cases are summarized in Fig. 6.12. The SSF were trained with the proposed label diffusion method for transduction.

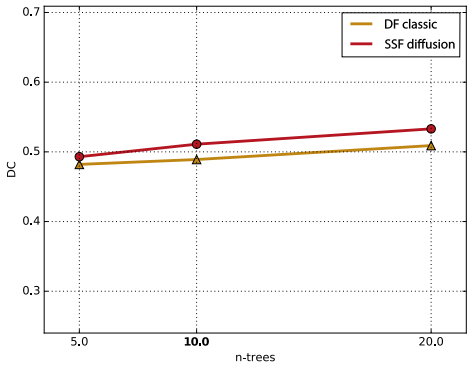
Discussion

Number of trees T Increasing the number of trees improves the results slightly for both forest methods. But considering the only marginal gain and the manifold increased runtime for the diffusion SSF it is preferable to keep the number of trees low.

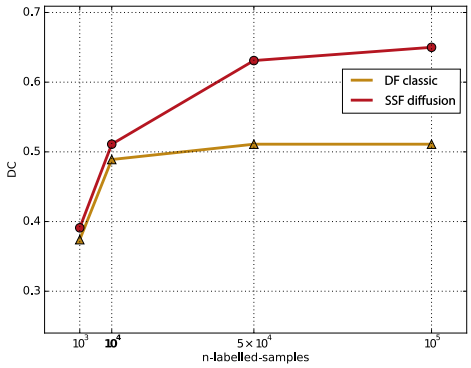
Number of labeled samples N_L More labeled sample increase the results for both forest variants. Especially the SSF benefits from more samples to base its split decisions on. Starting from $N_L = 50,000$, the beginning of a plateau is reached and more samples lead to only marginal additional gain. In terms of manual segmentation effort, $50,000$ labeled samples signify that only little over one percent of a patients first TP has to be segmented. However, the random labels are drawn from completely segmented cases in this experiment. Simply segmenting a single slice is unlikely to lead to the same segmentation accuracy. Which manual segmentation protocol for the first TP is the most suited for a subsequent SSF classification remains to be shown.

Maximum tree depth t_{depth} As observed before (see Fig. 4.12), the DFs prove again robust against overfitting: A certain minimal depth should be allowed to achieve optimal result, but an unrestricted growth holds no disadvantage. A quite different situation presents itself for the SSFs: Here a restricted growth is crucial for optimal classification results. Considering the nature of its Gaussian-density based split and label propagation mechanisms, this comes as no surprise since leaves with too few samples are unlikely to form expressive Gaussian blobs to guide the diffusion process. Hence, the tree depth is a critical parameter for SSFs.

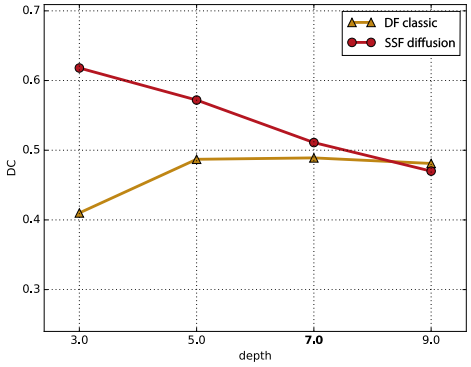
Data lesion ratio $lratio$ The previously proposed stratified random sampling scheme (Sec. 4.3.2) is based on the observation that DFs obtain best results when trained on data with the inherent background-to-lesion class ratio kept intact. The obtained results confirm this, as the DFs reach higher DC scores as the $lratio$ approaches the natural ratio of ≈ 0.05. For SSFs the situation is reversed: An artificially increased ratio of lesion samples in the training data distinctively improves the results and a balanced $1:1$ class ratio can be recommended. This behavior might be attributed to the SSFs necessity to correctly label dense clusters of unlabeled samples based only on the sparsely labeled samples present. With a low $lratio$, the few correct lesion samples are likely to be overruled by noisy background samples, leading to misclassification.



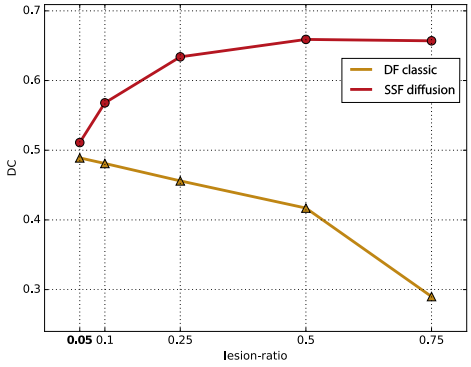
(a) Number of trees T



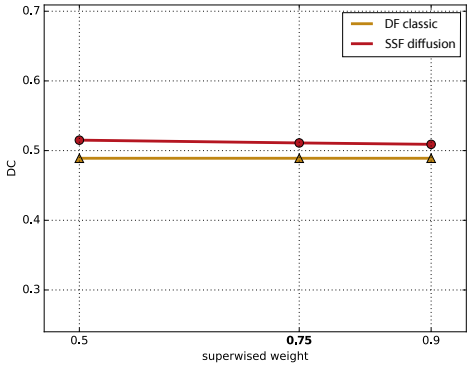
(b) Number of labeled samples N_L



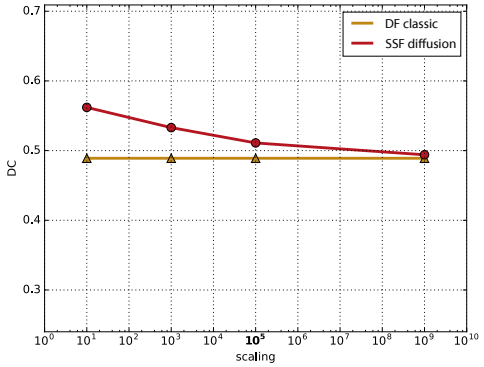
(c) Maximum tree depth t_{depth}



(d) Lesion class ratio *lratio*



(e) Supervised weight α



(f) Data scaling factor *scale*

Figure 6.12: Hyperparameter analysis results. Each graph shows the average DC values obtained using SSF diffusion based transduction when varying the associated parameter and keeping all others fixed at their default values (denoted by bold x-labels). Furthermore, for comparison, the results as obtained with a classic DF trained with the same parameters are depicted. Note the nearly constant superiority of the semi-supervised over the classification approach.

Supervised weight α As an exclusive SSF parameter, α does not affect the DF results. Surprisingly, neither does it influence the SSF's DC scores significantly. Presumably, the supervised term dominates the splitting decision at each node and the unsupervised term only determines at which location in the unlabeled samples filled, far apart space between the few labeled samples the split is placed. In such a situation the weight of the terms is irrelevant. If this is true, it would also mean that a desirable split into two dense clusters of samples cannot overrule an unfavorable distribution of labeled samples. For the practice this term can be considered irrelevant and set to its default value of 0.5.

Data scaling *scale* Contrary to prediction, the SSF does not benefit from upscaled training data. In the method section it is argued that insufficiently upscaled data might theoretically lead to a disappearance of the determinant value in the unsupervised term. This in turn would cause purely supervised split decisions in the lower levels of the tree. Hence, *scale* acts as a growth restriction term similar in effect to the t_{depth} parameter, which could explain the beneficial effects of less data scaling.

Features Previous experiments have shown the DF to be robust against redundant and correlated features (see Figures 4.9 and 4.10). While not extensively investigated here, preliminary experiments revealed that the SSF does not cope well with the local histogram feature, which might be attributable to its regularly spaced, often sparse, and multi-dimensional nature. Most likely the assumption of Gaussian distribution in feature space made for the unsupervised information gain term limits the type of features that can be employed in SSFs.

Conclusion

An detailed hyperparameter analysis for SSFs was conducted and recommendations for choosing the best parameter settings offered. The SSFs prove more sensitive to the parameter choices than DFs. For some parameters, they even differ in the direction of their response's slope. When employing SSFs, it is recommended to 1) carefully select the features, 2) ensure a balance class ratio in the training data, 3) restrict the tree depth to avoid overfitting and 4) provide at least one percent labeled samples. The remaining parameters seem to exercise only a limited influence on the segmentation quality.

Informative is the observation that the SSF leads for all except one set of settings to better segmentation results than the DF. And the best DC value is improved from ≈ 0.50 to ≈ 0.65 using SSF. This demonstrates that employing the unsupervised samples in the classification process improves the segmentation results for MS lesion segmentation. Whether this holds true for a larger dataset will be investigated in the next section.

The hyperparameter analysis results presented in Fig. 6.12 should be interpreted carefully, as the parameters might exhibit correlation effects. E.g, the gain in accuracy observed for a larger lesion ratio cannot be expected to simply add to the gain induced by a restricted tree depth, as both parameters might trigger the same underlying effects or even neutralize each other. The results presented in the next section will show how the ideal settings derived from this hyperparameter analysis perform in clinical scenarios.

6.2.3 Semi-supervised MS segmentation

Classifier (Semi-supervised Forest)	
T	20
F_{node}	$\sqrt{F}$
t_{depth}	DF/SSF approx 7, SSF diffusion 4
C_{opt}	Gini
Semi-supervised parameters	
N_L	$100,000$
α	0.5
Data parameters	
lratio	DF/SSF approx 0.05, SSF diffusion 0.5
scale	10^1
Features	
int	
wlm	$\sigma = 3, 5, 7$ mm

Preprocessing	
Resampling	$R_{work} = 1$ mm^3 (*)
Coregistration	MRI_{base} =? (*)
Skull-stripping	MRI_{skull} =? (*)
Bias-field	yes
Intensity range std	yes
Postprocessing	
Thresholding	optimal
Object threshold	no/5 mm^3
Hole filling	yes

Table 6.3: Semi-supervised MS segmentation experimental configuration. Starred steps were performed by the ISBIMS challenge's organizers.

Building on the observations made in the hyperparameter analysis of the previous section, a number of clinically plausible automatic lesion segmentation scenarios are investigated to study the potential benefits of a semi-supervised approach. As detailed in the introduction to MS (see Sec. 3.2.1), its treatment and diagnosis involves the manual lesion segmentation of multiple TPs of the same patient, which are regularly acquired at varying time intervals of one year on average. At the time a new TP is scanned, manual segmentation of previous TPs and/or other patients are likely to be available. Starting from this observation; the following three scenarios are derived:

Scenario I: ■ **P***x***TP**1→**P***x***TP**2$^+$ In this scenario, a full or partial manual segmentation of the patient's first TP is assumed to be available and all following TPs are to be segmented automatically. It can be considered the most basic of situations and holds two main advantages: First, the personal segmentation style employed in the segmentation of the first TP is honored for the following TPs, increasing the longitudinal segmentation consistency and hence comparability between TPs; second, only a partial, rough segmentation of the first time point is required judging from the low number of labeled samples necessary for a high segmentation accuracy (see the results of the hyperparameter analysis in Fig. 6.12). A possible disadvantage is that the classifier might be unable to deal with formerly unseen types of MS lesions or other new segmentation problems, both of which are likely to occur over the disease's lifetime.

Scenario II: ● **P***all***TP**1→**P***x***TP**2$^+$ In this scenario, the classifier's training data is augmented by the first TP of other patients to increase its knowledge base and to overcome the shortcomings of the first scenario without losing its advantages. This settings comes at no additional cost, since an MS clinic is unlikely to treat only a single patient and partial manual segmentations of the first TP are anyway required.

Scenario III: ▲ **P***others***TP**1→**P***x***TP**2$^+$ Both of the above scenarios require the, at least partial, segmentation of each patient's first TP. The third scenario therefore tests the situation where a number of manual segmentation from various patients are once made available and then used to guide the classification of all further scans, independent of the patient they belong to.

In most aspects, this scenario corresponds to the classic DF approach where a classifier is once trained on specially prepared training data and then used repeatedly for unseen cases. By employing SSFs, an attempt is made to see if the additional information provided by the test case's unlabeled samples improves the classification results.

Other scenarios are conceivable (e.g., employing the automatic results of a TP2 to segment TP3 and so on), but the above ones are deemed the most suited for actual clinical use and judged most informative to evaluate the proposed SSF method: The third scenario will reveal if a semi-supervised approach is beneficial in the classic DF domain. The first will show how the SSF performs in the case of minimal, specialized training data. And the comparison of these results against the second scenario will reveal if additional knowledge is required for an optimal segmentation result. In total, three forest variants are evaluated in the three scenarios: The classic classification DF (denoted as 'Cla'), the introduced SSF with approximation based transduction (denoted as 'Approx') and the introduced SSF with the proposed diffusion based transduction (denoted as 'Diff'). The final results are hence suited to reveal which of the nine method-scenario combinations is best suited for longitudinal MS lesion segmentation.

Experiments

Data The data employed in the following experiments is the training dataset of the ISBIMS lesion segmentation challenge as introduced in Sec. 4.5.2. Since a segmentation of the first TP is required for two of the three scenarios the testing dataset could not be employed. Thus, the data of 5 patients with 4 to 5 TPs each is used, making a total of 16 test cases when excluding the first TPs, which are required for training.

Evaluation The segmentation results are evaluated with a subset of the ISBIMS challenge's evaluation metrics as introduced in Sec. 4.5.2: the DC to denote the overlap; precision and recall (also known as positive prediction value (PPV) and true positive rate (TPR)) to reveal under- and oversegmentation, respectively; the lesion detection true positive rate (lTPR) and lesion detection false positive rate (lFPR) to denote how well newly appearing and vanishing lesions are tracked over the TPs; and the relative absolute volume difference (rAVD) to examine the method's suitability for MS lesion load based disease assessment. Compared to the challenge's results, the lTPR and lFPR implementation employed in this chapter might differ. As discussed in detail in Styner *et al.*, 2008, establishing true correspondence between binary objects is not a trivial task. Since the organizers of ISBIMS did not reveal their implementation details in their article [Carass *et al.*, 2016], the Styner *et al.*, 2008 definitions of these metrics are employed, which are likely to be stricter.

Preprocessing All images are prepared as previously for the challenge contribution (Sec.4.5.2) and denoted in Table 6.3.

Thresholding and postprocessing Both, DFs and SSFs produce posteriori lesion probability maps, which have to be thresholded to obtain the binary segmentation mask. In this evaluation, different methods are compared whose ideal threshold values are likely to differ. Fixing a single threshold for all nine method-scenario combinations would not result in a fair comparison, as some of the methods would be unjustly favored over the others. The ideal threshold for each method-scenario combination is therefore determined by an exhaustive search with a step size of 0.01 over all possible threshold values, selecting the one leading to the highest mean DC score.

		DC[0,1]	Prec.	Rec.	lTPR	1-lFPR	rAVD
■ Px	DF classic (thr=0.67)	0.64	0.67	0.64	0.60	0.09	0.27
	SSF approx (thr=0.61)	0.63	0.65	0.65	0.53	0.09	0.28
	SSF diffusion (thr=0.40)	0.65	0.62	0.70	0.63	0.28	0.23
● Pall	DF classic (thr=0.74)	0.53	0.65	0.55	0.61	0.09	0.50
	SSF approx (thr=0.70)	0.49	0.61	0.53	0.55	0.08	0.59
	SSF diffusion (thr=0.61)	0.68	0.66	0.73	0.64	0.19	0.25
▲ P$others$	DF classic (thr=0.72)	0.48	0.62	0.50	0.65	0.08	0.57
	SSF approx (thr=0.68)	0.45	0.57	0.50	0.60	0.07	0.66
	SSF diffusion (thr=0.42)	0.43	0.58	0.48	0.53	0.09	0.72

Table 6.5: Results for classic DF classification, SSF with diffusion and SSF with approximation based transduction applied to all three scenarios (denoted by colored symbols). The number in brackets denote the DC optimized posteriori probability map thresholds used to obtain the results.

This procedure corresponds to leave-one-patient-out parameter tuning. All presented results thus reflect the methods' peak performances.

The only postprocessing measure taken is the removal of small connected components from the binary segmentation mask, which are unlikely to constitute MS lesions. In this evaluation, most results are presented without the postprocessing measure applied to provide a clear picture of the methods' performances.

Hyperparameters All methods are trained with their respectively optimal hyperparameters as determined in the previous section and denoted in Table 6.3. The approximation SSFs use the same settings as the classical DFs, as this configuration leads to better results than using the diffusion SSFs' settings.

Results

For all three scenarios, three different types of forests are trained: First the classic DFs, second the proposed SSFs with Criminisi *et al.*, 2013's approximated transduction method, and third the proposed SSFs with the novel diffusion transduction. The obtained results are summarized in Table 6.5. A two-sided paired t-test is employed to reveal whether the differences between the nine method-scenario combinations are statistically significant at $p < 0.05$. The results are visualized as a network graph in Fig. 6.13. Enabling the postprocessing scheme, the results presented in Table 6.6 are obtained. As noted in the introduction to this chapter, MS lesion segmentation is not an intrinsically time critical operation, but it would nevertheless be advantageous to achieve speeds that allow to process the daily influx of data. Table 6.7 therefore lists the approximated average runtimes of the three methods.

Discussion

The conducted evaluation on the MS data is suited to highlight a number of aspects of the compared methods. In the following, the results are discussed under different points of view and the questions posed in the introduction of this section are answered.

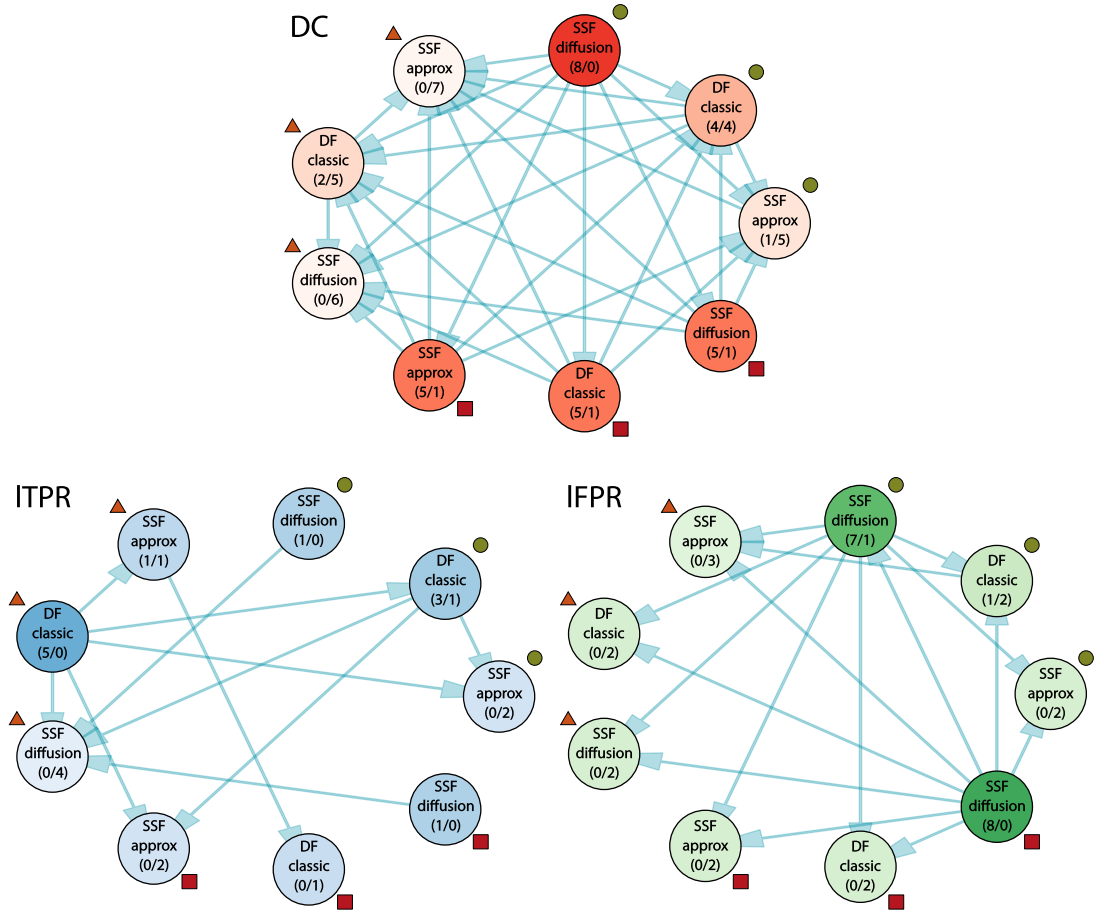


Figure 6.13: Significant differences between the nine evaluated method-scenario combinations (Table 6.5) according to a two-sided paired t-test ($p < 0.05$) in, from left to right, DC, lTPR and lFPR. Each node represents a combination of scenario and SSF variant, each edge a significant difference of the tail side team over the head side team. Therefore, the less outgoing and the more incoming edges a node has (denoted by numbers in brackets (*out/in*) for easier interpretation), the weaker its method compared to the others. The saturation of the node colors indicates the strength of a method, where better methods are highlighted with more saturated colors. Note that all methods with the same number of incoming and outgoing edges perform, statistically spoken, equally well. A higher importance of incoming over outgoing edges or vice-versa cannot be readily established.

		DC[0,1]	Prec.	Rec.	lTPR	1-lFPR	rAVD
■ Px	DF classic (thr=0.61)	0.66	0.65	0.69	0.54	0.28	0.30
	SSF approx (thr=0.61)	0.65	0.68	0.65	0.47	0.28	0.25
	SSF diffusion (thr=0.40)	0.65	0.63	0.70	0.59	0.69	0.23
● Pall	DF classic (thr=0.74)	0.55	0.70	0.55	0.52	0.27	0.41
	SSF approx (thr=0.70)	0.51	0.65	0.53	0.49	0.26	0.50
	SSF diffusion (thr=0.61)	0.69	0.68	0.73	0.60	0.58	0.23
▲ P$others$	DF classic (thr=0.72)	0.50	0.67	0.50	0.57	0.25	0.47
	SSF approx (thr=0.68)	0.47	0.61	0.50	0.53	0.24	0.56
	SSF diffusion (thr=0.42)	0.44	0.60	0.48	0.45	0.32	0.64

Table 6.6: Results for classic DF classification, SSF with diffusion and SSF with approximation based transduction applied to all three scenarios (denoted by colored symbols) after applying the postprocessing. Compare to values in Table 6.5.

Method	Classification <	Approximation <	Diffusion
Runtime	$\approx 15min$	$\approx 1h$	$\approx 7h$

Table 6.7: Runtimes for the different forest methods for a single segmentation case.

Method most suited for longitudinal MS segmentation The experimental results provide a clear picture of which method is most suited for longitudinal MS lesion segmentation: Diffusion transduction based SSFs in the Pall scenario. A significant better mean DC than all eight alternatives, an excellent lFPR surpassing the other forest variants, a lTPR abreast with the best scores and a rAVD more than twice as good as the other methods in this scenario. The DC's standard deviation (STD) is low with 0.09 and the worst of the 16 test cases still reached a DC of 0.48. With postprocessing enabled, most measures could be further improved, the lFPR even nearly doubled. Hence, the proposed SSF outperforms the alternative methods. The remainder of this section discusses the results under different point of views, together striving to answer which mechanism might underly this success.

Interpreting ■ Px In this scenario, where the first TP is used to segment the following TPs, all methods perform similar with no significant differences in DC or lTPR. But for the lFPR metric, the diffusion based SSF is significant better than the two alternatives with a mean value nearly thrice as good. Semi-supervised classification strives to improve the classification accuracy by using the unlabeled testing data to guide the classifier's training process. The results show that this has a beneficial effect where longitudinal, only moderately changing data is concerned as in scenario Px. Nevertheless, suitable models must be found to exploit the learned PDF's topology: Oversimplification, as for the approximation based SSF, does not improve the results. Required is a suitably complex system able to follow the areas of greatest density in feature space, as provided by the SSFs with diffusion based transduction. Since only the lFPR was improved significantly, the segmentation results in the areas of lesion-similar white matter hyperintensitiess (WMHs) seems to benefit most. This might be explained by the fact that these samples, while forming cluster near to lesion clusters, are reached at lower costs from various background labeled sample groups.

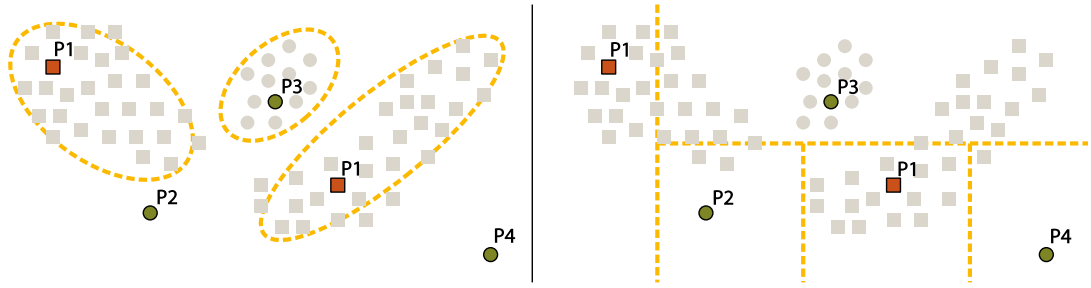


Figure 6.14: Schematic explanation of the possible reason behind the decreased performance for DFs and increased performance for diffusion SSFs when adding labeled samples from other cases as in P*all* compared to Px. The graph shows two times an exemplary two dimensional feature space in which the trees perform their splits. Labeled samples are depicted in color, red and squared for samples from patient 1 (P1), TP 1 and green and circular for samples from other patients (Px). The unlabeled test samples from P1, TP 2 are denoted in gray, squared for sample groups already present in TP 1, circular for a new sample group describing, e.g., a newly appeared lesion of different appearance. On the left side, the ellipses show the clusters detected by the SSF, which guides its feature space splitting to be placed in areas of low density. This way the squared samples of known problems are correctly classified and the circular samples of a new problem are correctly assigned to a labeled sample of another patient (P3), describing its appearance. On the right side, the maximum margin split lines of the classical DF are denoted by the yellow broken lines. Since the DF is not aware of the unlabeled samples, the lines are placed to split the dense clusters of common samples, leading to misclassification. Without the additional and often irrelevant samples from the other patients, a single and ultimately better split would have been placed between the two labeled samples of the first patient (P1, red squared dots). This example might explain the increase performance of the DF when additional patients are added to the training set.

Interpreting ● P*all* For the second scenario, labeled samples from the other patients' first TPs are added to the training set, potentially improving the segmentation accuracy by better treatment of new lesion types and formerly unseen problems. Looking at the results, it becomes obvious that the approximation as well as the classification method do not benefit from the additional training data. On the contrary, the recall and hence equally DC and rAVD scores are lower than for the first scenario. Precision, lTPR and lFPR are not altered greatly. The proposed diffusion label transduction SSFs, on the other hand, benefit from the additional samples, increasing precision, recall and hence also DC significantly. Only the lFPR is slightly reduced, but this effect is countered by the postprocessing. Essentially, for the DF, the additional labeled samples have a confounding effect, since it cannot draw on the guiding dense clusters of the unlabeled data. The SSF, on the other hand, can fall back to the new labeled samples to successfully classify accumulations of formerly unseen types of samples. A schematic explanation of this effects is given in Fig. 6.14. The approximation SSFs are affected even worse than the DF by the additional data. This means that the PDF learned during forest training has a complex topology that cannot be exploited by the simplified label propagation scheme.

Interpreting ▲ P*others* Removing the first TP of the testing patient from the training data leads to the expected drop in performance for all methods compared to the other scenarios. Interestingly, the diffusion SSF perform in this scenario not only worse than the other forest variants, but reach the lowest mean score of all nine method-scenario combinations. This signifies that

in the classical supervised classification domain represented by scenario P*others*, the SSF approach does not improve the results. On the contrary, the segmentation accuracy is significantly diminished.

Interpretation of approximation SSFs Compared to the diffusion SSFs, the approximation variant is unable to cope with complex PDF topographies, as has been shown before with toy examples. In a real case as the evaluated MS lesion segmentation, the difference is not as pronounced and depends on the scenario. On the other hand, the magnitude of difference between the approximation and diffusion SSFs can tell us something about the complexity of the learned PDF: In cases where the split lines are set unsuitably (P*others*), there is no significant difference between the methods. Where the PDF is dominated by dense clusters of similar samples (Px), both perform similar. But where a complex PDF topology is encountered (P*all*), only the diffusion approach is able to correctly propagate the labels to the test samples. Note that these descriptions base on assumptions, as the high-dimensional feature space cannot be readily visualized. Conclusively it can be said that Criminisi *et al.*, 2013's approximated transduction method oversimplifies the label propagation problem and that subsequently the proposed diffusion variant is superior. While the approximation is faster, it never performs better than the classical DF classification variant, which trains even faster than the approximation SSFs.

Interpretation of diffusion SSFs The proposed SSF method with the near-exact diffusion label propagation leads to the overall best results with a significant gap to all other methods. It especially excels in the lFPR, where the other methods proved particularly weak.

Comparison to challenge participation In a previous chapter (see Sec. 4.5.2), the results of the participation in the ISBIMS challenge are presented, obtained with the proposed DF brain lesion segmentation framework which corresponds to the classical DF approach used in this evaluation. While a direct comparison between the previous result (as shown in Table4.14) and the semi-supervised results from this section (Table 6.5) is not possible, as the former were obtained on a hidden testing set with a third-party evaluation, while the latter were computed over 16 cases of the training set, the rough values can be used to provide a general impression. The current method-scenario configuration that corresponds to the challenge contribution is the P*others* with classical DF. Clearly, higher means scores, especially in precision and lFPR, are possible when more training cases (21 cases of 5 patients) and a larger training set ($1,000,000$ samples) are used. But by employing the first TP additionally, as in scenarios Px and P*others*, even higher segmentation accuracy can be achieved on fewer training cases (5) and less samples ($100,000$). That signifies that in the context of longitudinal segmentation, the semi-supervised approach should be preferred over the supervised.

Influence of postprocessing The only postprocessing employed is the removal of small connected components in the binary segmentation falling under a size threshold. Its application especially boost the lFPR, but also other measures. Only the lTPR looses slightly, but not significantly. Since all nine method-scenario combinations are similarly affected by the postprocessing, the significancies of their pairwise differences is largely unaltered.

Runtimes The only real drawback of the proposed diffusion SSFs is their manifold increased runtime. Fortunately, MS segmentation is not time critical application and the process is fast enough to keep up with new incoming scans. In the context of a clinical study, where large

amounts of cases have to be processed in a limited time, a faster method might be preferable or multiple machines must be used.

Manual segmentation effort Whereas scenario P*others* requires the onetime careful segmentation of a sufficiently high number of cases and can then be applied repeatedly, the other two scenarios suited for SSFs need repeated manual intervention. Each time a new patient is admitted or added to the database of cases, the first TP has to be partially segmented. If training with $100,000$ samples from 5 cases with a 0.5 lesion-to-background ratio, as determined to be optimal during the experiments, $10,000$ labeled samples must be available from each class for each case. While the background samples are easily obtained with broad strokes far from the lesions, marking the foreground samples might well involve a certain manual effort. But it will be considerably lower than for a complete segmentation and additionally allow the SSF to adapt to the raters personal style.

Conclusion

In this section, the introduced SSFs variants were compared against each other and the classical DF at the real world examples of MS lesion segmentation in three different scenarios. The proposed diffusion SSF lead to the best overall segmentation results and its usage can be recommended for longitudinal MS lesion segmentation, where it is likely to increase accuracy and longitudinal consistency. The proposed scenario requires the partial segmentation of a patients first TP. In cases where this is not feasible, the classical DF should be preferred over the SSF for optimal results. But wherever a segmented previous time point, a partially labeled testing set or very similar cases are available, the SSF should be given preference. The approximated label transduction method as proposed by Criminisi *et al.*, 2013 proved also in a real case to be insufficiently equipped to follow the the distributions in a complex PDF topology. A downside of the proposed diffusion SSF is its long runtime. Here, some improvements might be necessary for an application to a large number of cases.

6.3 Conclusion

In this chapter, a new SSF method was derived. To this end, the theoretical foundation of Criminisi *et al.*, 2013 was extended and the, to the authors best knowledge, first multi-dimensional version implemented. This entailed a reformulation of the information gain term to incorporate a normalization step and the implementation of multiple measures against numerical instability. Additionally, with DynStatCov, a fast, dynamical statistical co-variance matrix update procedure was introduced and made available as stand-alone Python package [Maier, 2016a]. Finally, a new solution to solve the exhaustive search along the geodesic PDF surface for the purpose of label propagation had to be found. The proposed label diffusion approach uses significantly less memory and has a lower runtime than the full solution, rendering the method applicable to problems with millions of samples. Compared to the approximated solution proposed by Criminisi *et al.*, 2013 and implemented in their Sherwood library, the diffusion based method is able to follow complex paths along the areas of highest density in feature space.

At the example of a number of toy datasets, the superiority of the new label propagation approach was shown. An extensive hyperparameter analysis served to investigate the effects of the various parameters on the SSF's performance and allowed for a juxtaposition against the supervised DF. In the last evaluation section, the proposed SSF was compared against the supervised DF in the real world application of longitudinal MS segmentation, where it was able to improve the segmentation results in all scenarios involving a previous TP of the same patient

in the training data. It can therefore be concluded that the SSF, provided training data similar to the testing set is available, can successfully discover communality between samples and place the nodes' splits in the areas of low density, subsequently leading to better results. In the classical unsupervised scenario the semi-supervised approach was not found to perform better.

The tested MS segmentation scenario in which the SSF obtained significant better scores than all other methods is directly applicable in clinical context. Requiring only partially labeled first TPs, the manual effort is reduced. At the same time, personal segmentation styles of the raters are honored and the longitudinal consistency improved.

The proposed SSF are not limited to brain lesion segmentation, but, as a general semi-supervised classifier, can be applied to a range of other problems where partially labeled data is available. Thanks to their comparatively low memory requirements, they can be used to process datasets reaching millions of samples. A convenient implementation on top of the popular *sklearn* toolbox [Pedregosa *et al.*, 2011] is provided with *sklearnef* [Maier, 2016c].

Whether the method is equally suited for other semi-supervised classification tasks and how it performs compared to other semi-supervised classification approaches remains to be shown. A good starting point would be a direct comparison against the methods described in Zhu, 2005, followed by the alternative SSF methods of Leistner *et al.*, 2009 and Liu *et al.*, 2013.

One of the remaining unknowns of the introduced SSF is its sensitivity to the range, shape and distribution of the provided features. During the experiments, the local histogram feature was found to be unsuited for a use with SSF, probably due to its sparse, integer range nature. How well the utilized density estimator reacts to different types and combinations of features remains to be shown.

Finally, the main bottleneck of the SSFs is their long runtime, which is caused by the label diffusion for which a large, sparse linear system has to be solved. In the case of multiple labels, this process can be readily parallelized, since the label wise computations are independent of each other. Solving a single linear system may be outsourced to the graphics processing unit (GPU): MATLAB™, for example, provides such solvers in its newer versions, which can be up to six times faster for large matrices.

Chapter 7

Summary

The methodological chapters of this thesis contain individual discussions and conclusion. This chapter summarizes the main contributions, discusses the medical perspective and concludes with an outlook on future works.

7.1 Contributions

Three main contributions were made in this work to address the objectives formulated in Sec. 1.1: First, a general brain lesion segmentation framework was designed, which can be easily adapted to various types of lesion causing diseases; second, a spectral clustering based ensemble forest classifier was developed to address a number of remaining segmentation problems; third, a novel semi-supervised decision forest (DF) variant was proposed with the aim of improving segmentation accuracy in scenarios where longitudinal data is available.

The **general purpose brain lesion segmentation framework** introduced, discussed and evaluated in this thesis is principally designed to segment pathological tissue from multi-spectral 3D magnetic resonance imaging (MRI) brain scans. To this end, a number of specialized image features are developed that do not rely on disease specific knowledge, but rather exploit the brain's anatomy and general brain lesion idiosyncrasies. All framework components are designed to function fully automatically. The idea is that, except for some minor adaptations to the specific MRI sequences, the method can be readily trained for and applied to any type of brain lesion causing disease. A successful application on acute stroke, sub-acute stroke, MS and glioma images indicates that this objective could be accomplished. The obtained results show that DFs are able to adapt to the idiosyncrasies of different pathologies and that the supporting elements of the segmentation framework, such as pre- and postprocessing components, are only required to target the modality but not the disease specific problems.

To overcome some of the problems observed with the general method, **local problem forests** were proposed. By combining a non-linear spectral clustering with fuzzy forest catchment areas for training and application, they improve the segmentation accuracy in the targeted areas. As a general method, this forest variant could be readily applied to segmentation tasks other than mono-spectral sub-acute stroke segmentation.

Longitudinal scans over multiple time points (TPs), as acquired for Multiple Sclerosis (MS) treatment, indicate the use of unlabeled data to improve the classification mechanism. To this end, a new **semi-supervised forest** variant was developed and successfully evaluated in multiple MS segmentation scenarios. Its main characteristics are an improved runtime and reduced memory requirements, rendering it suitable for medical image processing. Using this new approach,

the segmentation accuracy could be improved over classical DFs and the required number of labeled training samples reduced. Furthermore, rater-individual segmentation styles are respected, ultimately improving the longitudinal segmentation consistency.

Conclusion Through the participation in multiple benchmarks, it could be shown that the various brain lesion segmentation tasks have enough in common that a general machine learning method such as DFs can obtain a segmentation accuracy with state-of-the-art performance. While reaching consistently high placements in fair and independent comparisons, the absolute evaluation scores obtained for different diseases differ substantially. In particular glioma and sub-acute stroke exhibit severely varying lesions and their segmentation might require an incorporation of disease specific aspects. The treatment of locally limited problems that escape the DF due to low prevalence in the training data can be improved with suitable training data clustering in a well-defined problem space with a subsequently weighted classification vote of multiple problem-specific classifiers. In the case of longitudinal assessed diseases, a semi-supervised approach proved preferable over a supervised one, improving the segmentation accuracy and honoring personal delineation style.

7.2 Medical perspective

The main purpose of medical image computing is to empower, facilitate and safeguard clinical analysis, intervention and research processes. In this section, the introduced methods are reviewed under the medical perspective and discussed to which extend they meet the clinical boundary conditions as described in Chapter 3.

Acute stroke The treatment of acute ischemic stroke follows a well established and largely standardized workflow. The derived main requirements that any aspiring automatic segmentation method must fulfill are accuracy, reliability, reproducibility and speed (Sec. 3.1). The evaluation conducted on the SPES benchmark (Sec. 4.5.1) shows that the proposed framework processes a case in less than five minutes, reaches a high segmentation accuracy (Dice's coefficient (DC)=0.81, average symetric surface distance (ASSD)=1.36), is reliable (DC standard deviation (STD)=$\pm$0.09, ASSD STD=$\pm$0.74) and deterministic. By fulfilling all stated requirements, the method can be considered ready to be used in clinical practice. Its application would allow physicians to make better informed treatment decisions; provide a safeguard against possible over- or underestimation of the lesion, a critical point wherever treatment risks have to be carefully weighted against potential benefits; and make it possible to derive new decision rules based on quantitative measures such as lesion volume or location, rather than a purely visual assessment. Another beneficiary are drug trials and other clinical studies, in which the additional gain of any newly proposed treatment has to be established in terms of rescued quantitative lesion area, e.g., by comparing the acute diffusion volume against the final outcome follow-up lesion volume.

Sub-acute stroke In the sub-acute stroke setting, the treatment window has passed and MRI scans are usually acquired for follow-up assessment only. On the other hand, congnitive neuroscientists often conduct their studies, which aim to relate cognitive deficits to affected brain areas, on lesions in this development state. Yet other clinical studies are interested in the array of interacting changes the lesion undergoes in the sub-acute phase. Subsequently, the requirements an automatic lesion segmentation algorithm has to fulfill differ substantially from the acute setting. Processing speed is only of minor importance, while the need for reproducibility and reliability is greatly enhanced. A suitable automatic approach would relieve researchers from the

time consuming burden of manual lesion segmentation, which is recognized as the main limiting factor in these kind of research studies. Subsequently, larger patient cohorts could be recruited and fully reproducible segmentations created, rendering the results more robust and allowing for a sound statistical evaluation. While favorably placed in a direct comparison against other state-of-the-art methods in the SISS benchmark (Sec. 4.5.4), the proposed method unfortunately does not reach the segmentation accuracy and reliability required for a ready application in above described settings. And while the proposed local problem forests (LPF) approach improves the results again, more work is required to obtain segmentations that can compete with the inter-observer scores¡.

Multiple Sclerosis MS is the second of the two main pathologies addressed in this thesis. Its assessment and treatment, while not a time critical environment, pose high demands on consistency and reproducibility. The proposed and evaluated segmentation method (Sec. 4.5.2) enables reliable and reproducible MS lesion segmentation unaffected by inter-rater variations. The latter is a severely limiting factor of manual MS lesion segmentation, since the expert raters often change from one year's scan to the next. To employ automatic MS lesion segmentation as a surrogate measure in clinical trials or as a factor in quantitative disease burden analysis, changes in lesion volume and new lesion appearances as well as disappearances have to be closely monitored. This requires consistent segmentation over different TPs. With the proposed semi-supervised forest (SSF) approach, high segmentation accuracy can be achieved with only a limited manual segmentation effort. Since the method learns from the partially segmented first TP, personal segmentation styles are honored and the longitudinal consistency increased even further. Both methods compare favorably to the inter-rater scores and could be considered ready to be employed in clinical and research settings.

7.3 Perspectives

The methods developed in the context of this work provide several links to subsequent research projects. Some are summarized here.

7.3.1 Extending the general brain lesion segmentation method

Three ways of extending the general method presented themselves during the investigation conducted in this work.

Adding more features Since the sought after solution should be disease independent, the type of features that can be used is limited. Nevertheless, the list of features investigated in this work is far from complete. One possibility would be to draw from physiological and anatomical knowledge of the healthy brain. An often proposed approach is to employ atlases of the human brain, such as ICBM-152 [Fonov *et al.*, 2009] or MNI-305 [Collins *et al.*, 1994], and register them to the test cases, e.g., with the methods employed by Klein *et al.*, 2009. Subsequently, the difference to the average default value at each voxel can be used as a features that provides a notion of abnormality. Challenges faced with this approach are the difficult registration between pathological and healthy brains and the great inter-sequence variations of MRI. Other possibilities are texture features or other intensity value derived characteristics. Since DFs were found particularly robust against redundant or correlated features, employing more features offers itself as an easy way to improve segmentation accuracy.

Sequential segmentation methods To overcome the weaknesses of one segmentation approach, many researchers decide to employ a second, more or less different segmentation method to improve the results of the first. Promising results have been reported for conditional random fields (CRFs) [Lee *et al.*, 2005; Meier *et al.*, 2016] and Markov random fields (MRFs) [Zhang *et al.*, 2001; Malmi *et al.*, 2015; Jesson *et al.*, 2015], which are useful to correct errors in the forests posteriori prediction. As a spatial regularization step, they provide structured lesion estimation that the voxel-wise processing and neighbor state unaware DF cannot, hence complementing its functionality.

Replacing parts of the framework by a CNN During the investigation of alternative classifiers for the proposed framework in Maier *et al.*, 2015d, the convolutional neural network (CNN) method proved abreast with the DFs, an observation which supports the recent popularity of this approach (as already noted in Maier *et al.*, 2017). As discussed in App. B, the DF method requires carefully hand-crafted features, while the CNN is not subject to such restrictions but, on the contrary, able to learn the most discriminative features automatically. Thus, they promise to be a general, easily adaptable brain lesion segmentation method. It remains to be shown if a single network architecture is indeed capable to cope with various lesion segmentation tasks. It would be especially interesting to investigate in how far the MRI preprocessing steps would become obsolete when using CNNs.

7.3.2 Incorporating disease specific knowledge

One of the purposes of developing a general brain lesion segmentation framework was to build upon it. By incorporating features derived from knowledge about pathological idiosyncrasies, a disease specific solution can be developed that is likely to outperform the general approach. Which information can be useful depends strongly on the pathology. One possibility would be features denoting the distribution or localization of lesions as, e.g., provided by the vascular territories of the brain [Derntl *et al.*, 2016] for ischemic stroke. Another approach could be to model a disease's morphology, e.g., through a multi-spectral stroke evolutionary model or tumor onion skin structure [Menze *et al.*, 2010].

7.3.3 Improving and understanding the semi-supervised forests

The proposed SSF is a new development and, while based on a sound theoretical model, its behavior in practical situations is yet poorly understood. With the application to the large scale MS lesion segmentation problem and the thorough hyperparameter analysis, one use case is covered in this work. It remains to investigate the SSF's behavior for different sets of feature and derive general recommendations. Such findings would furthermore show how the SSFs react to the high dimensional feature space problem affecting so many density based approaches. Next step would be to apply the SSFs to other problems from the medical image processing domain and beyond. The algorithm is a general one, applicable in most if not all scenarios of semi-supervised labeling. Algorithmically, the numerical robustness should be tested further and its shortcomings improved. Finally, the processing time bottleneck of the label propagation should be sped up, e.g., by employing graphics processing unit (GPU) based solvers as already available for, e.g., Matlab®. Of further interest would be a direct comparison to alternative semi-supervised forest methods such as Leistner *et al.*, 2009 and Liu *et al.*, 2013, as well as a range of other semi-supervised approaches as summarized in Zhu, 2005.

7.3.4 Investigating the local problem forests

Just as the SSFs, the LPFs require further investigation. A first step would be an application to other diseases to see if these domains would equally benefit. Next a thorough investigation into various problem spaces (i.e., representation and distance measure combinations) would be desirable to improve the understanding of the problem space concept and to find general recommendations for application. Finally, it would be convenient to develop a faster implementation of the clustering step.

Appendix A

Segmentation benchmark ranking scheme

Public segmentation challenges aspire to provide fair and transparent evaluation benchmarks for researchers around the world. Beside a problem representative dataset, a suitable measure of a segmentations quality has to be provided. When the ground truth answers the *what* question, then the evaluation framework aims at the *how*. For most tasks, no ideal evaluation metric exists and surrogate measures have to be employed. After carefully selecting a set of suitable measures, each of which highlights a different aspect of the evaluated segmentation, they have to be combined suitably such that they together provide a (nearly) complete picture. This appendix describes the ranking scheme developed for and employed in the ISLES 2015 challenge (see Sec. 4.5).

Metrics For the SISS part of the ISLES challenge, the metrics employed are the Dice's coefficient (DC), average symetric surface distance (ASSD) and Hausdorff distance (HD) as defined in Sec. 4.3.4. They denote the volume overlap between two segmentations, the average surface distance between two segmentations, and the maximum surface distance respectively. For the SPES part, only the DC and ASSD are employed due to voxel-sized holes in the expert segmentation, which negate the HD's ability to denote outliers. But below described ranking scheme can be applied to all types and combinations of evaluation metrics as long as they poses order.

Ranking After selecting suitable evaluation metrics, the second problem faced is the establishment of a meaningful ranking for the competing algorithms: The different measures are neither in the same range nor direction. In the simplest case, a single of the above metrics can be chosen and used in the ranking [Menze *et al.*, 2015a]. But this would mean neglecting the aspects revealed by the remaining measures and is hence a bad choice for most challenges.

A second approach taken by some challenges [Styner *et al.*, 2008] is to compare two expert segmentations against each other. The resulting evaluation values are then assumed to indicate the upper limit and hence denote the 100 percent mark of each measure. New segmentations are then evaluated and the values compared to their respective 100 percent mark, resulting in a percentage rating for each measure. Drawback is, that for not range restricted measures, such as the ASSD, one has to define an arbitrary 0 percent mark.

A third approach is chosen, based on the ideas of Murphy *et al.*, 2011 and Murphy, 2011, which establishes that a ranking must only honor the direction of a relationship between two

items (i.e., higher, lower or equal), but not its magnitude. After obtaining from each participating team the segmentation results for each case, the following steps are executed:

1. Compute the DC, ASSD & HD values for each case
2. Establish each team's rank for DC, ASSD & HD separately for each dataset
3. Compute the mean rank over all three evaluation measures/case to obtain the team's rank for the case
4. Compute the mean over all case-specific ranks to obtain the team's final rank

Graphically, the schema looks like displayed in Fig. A.1.

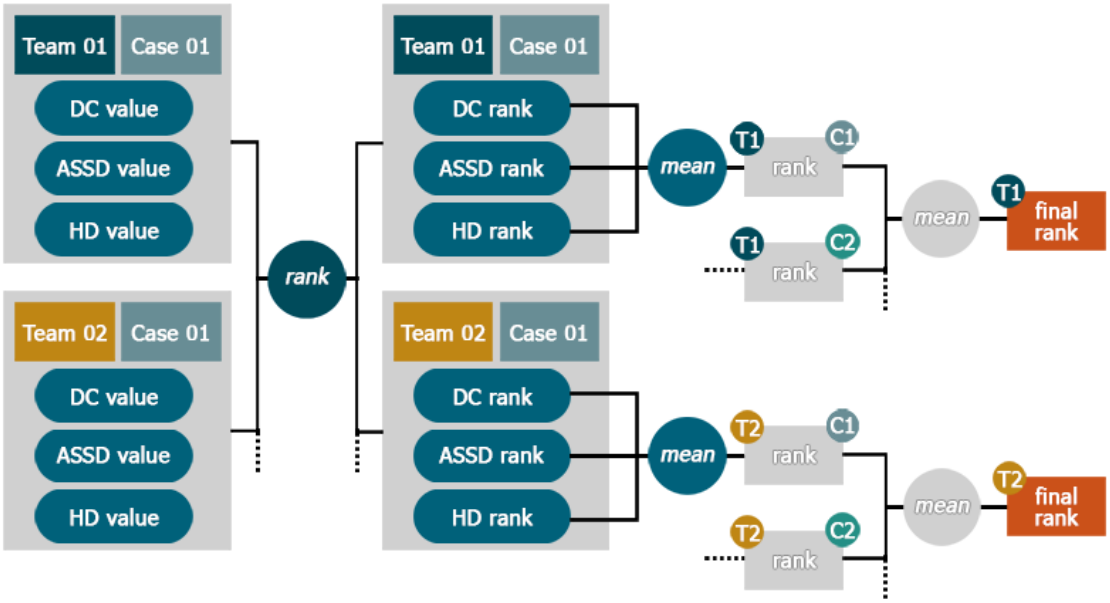


Figure A.1: Ranking schema as employed in the ISLES challenge.

The outcome of the procedure is a final rank (real number) for each participant, which defines its standing in the leaderboard. For SISS, with two ground truth sets for the testing dataset, their respective final ranks are averaged. For SPES, on the other hand, the HD is excluded from the computation of the ranks, as the provided ground truth contains voxel-sized holes, which defeat the purpose of the HD measure, i.e., penalizing the outliers.

This approach can be applied to any number of measures, independent of their range, type or direction. Its outcome denotes only the differences between algorithms and hence serves its purpose. For any interpretation of the results, the distinct evaluation measure values obtained have to be considered, too.

A challenge with winners requires an absolute ranking, an ongoing benchmark not. For the online, ongoing leaderboard, the rank is not computed, rather each user is invited to sort the result table according to their favorite evaluation measure.

Resolving ties In one step of the proposed algorithm, the performance of each team on one case regarding a single evaluation metric have to be ranked. Such a situation can lead to ties, which have to be handled specially. Both tied teams are decorated with the upper rank and leaving the following empty (see Table A.1 for an example).

Team	DC
T-A	0.33
T-B	0.33
T-C	0.50
T-D	0.33
T-E	0.31

(a) Before...

Rank	Team
1	T-C
2	T-A, T-B, T-D
3	
4	
5	T-E

(b) ...after.

Table A.1: Example of resolving ties for ISLES.

This behavior has an interesting effect for very difficult cases, where most teams fail to produce a valid segmentation, as can be seen in the example of Table A.3. Thus, difficult cases

Team	DC
T-A	0.00
T-B	0.00
T-C	0.10
T-D	0.00
T-E	0.00

(a) Before...

Rank	Team
1	T-C
2	T-A, T-B, T-D, T-E
3	
4	
5	

(b) ...after.

Table A.3: Tie resolving for difficult cases.

do not alter the mean, as they would do when simply averaging the DC values over all cases. Instead only the performance relative to all other algorithms is compared, resulting in a more expressive ranking.

Failed cases Beside resolving ties, the concept of a failed case is defined. When faced with (1) a missing segmentation mask or (2) a DC value of 0.00 (i.e., no overlap at all), the concerned case has been declared failed and all metric evaluation values subsequently set to infinity. Combined with the employed ranking approach and above described treatment of ties, this allows to incorporate missing segmentations in the ranking in a natural and fair manner. It could be argued that a DC of 0.00 could well mean that another part of the brain has been segmented. But the case has nevertheless to be considered as a failed one, as the target structure has not been detected. Not declaring the case a failure would lead methods submitting a single random voxel segmentation to be ranked higher than an empty segmentation mask.

Conclusion The presented ranking scheme is applicable to all evaluation measures which poses an order, which includes all real metrics. It provides a fair and direct ranking of the participating methods relative to each other, elegantly handling ties and missing results. Exceptionally easy or difficult cases in the training data do not distort the results, as they would in the case of simple metric averaging. Thus, a segmentation's quality can be assessed based on more than one metric,

increasing the leaderboard's expressiveness. On the downside, as a relative ranking, it has to be re-computed each time an entry is added or removed. This limitation is negligible considering that a clear ranking is only required for the on-site determination of the winners, while the ongoing submission should rather be made sortable by each metric than ranked according to a, possibly in another context meaningless, ranking system.

Appendix B

A short discussion on DNNs, CNNs and DFs

When the work on this thesis began, deep learning was still a relatively new term mainly used in the field of computer vision. Since then, the concept has spread rapidly to all areas of computer science and recently even to the general news. Part of its popularity can be attributed to serious campaigning and funding by large IT companies (including Alphabet, Facebook, Apple and Microsoft), part to the impressive results obtained in various domains.

Deep neural networks (DNNs), in particular their convolutional neural network (CNN) variant, exercise a considerable influence in segmentation from medical images. While an extensive discussion of DNNs and their impact is out of the scope of this thesis, I made an attempt at a short juxtaposition of DNNs and decision forests (DFs) for segmentation. The following text makes no pretense to be complete or comprehensive. Rather, it should be treated as a comment.

DNNs for image segmentation

The term DNN refers to a variant of the well known artificial neural networks (ANNs). The latter have been long considered to be only extensible in width and, until recently, dropped from the view of many researchers. Network architectures with many hidden layers (i.e., DNNs), were deemed untrainable due to gradient vanishing effects, excessive training times and tendencies to overfit. This changed recently with two new developments: 1. a string of publications on what effectively are training regularizers [Srivastava *et al.*, 2014; Ioffe *et al.*, 2015] and 2. the manifold increased computational power provided by graphics processing units (GPUs).

The second part is easily apprehended. Large networks have billions of parameters and train accordingly slow. Modern GPUs render it possible to fit a whole training step into their memory and are highly optimized for the required matrix computations, resulting in faster network training. That said, a typical DNN might still train days or even weeks. But these periods are manageable.

The first part of the new developments is more difficult to fathom. The initial idea making certain types of DNNs feasible was pre-training [Hinton *et al.*, 2006b]. To this end, auto-encoder similar network architectures are formed and trained unsupervised, encouraging the network to learn representative low-level representations of the training images [Hinton *et al.*, 2006a; Vincent *et al.*, 2010]. Then, a short fully connected network intended for classification is attached to the pre-trained, narrowing funnel of the auto-encoder and subsequently trained supervised for a small number of additional epochs. The effect is mainly one of overfitting avoidance and training

acceleration.

Soon, the focus shifted towards CNNs for image processing [Krizhevsky *et al.*, 2012]. This network type is characterized by special layers with shared weights that act as image convolution kernels sweeping the input image. Each convolutional layer is typically succeeded by a pooling layer, which act as additional regularizers (although newer publications claim these additional layers to be obsolete [Springenberg *et al.*, 2014]). Effectively, a CNN exploits the spatial nature of images to restrict the networks training possibilities substantially and, similar to auto-encoders, encourages the network to learn low-level representations of important image features. That means less training weights and a reduced overfitting. CNNs proved very effective for object recognition and segmentation in images [Krizhevsky *et al.*, 2012; Ciresan *et al.*, 2013; Shelhamer *et al.*, 2015]. The range of application is mainly limited by the available GPU memory.

Most recently, fully connected networks became fashionable again, be it in their classical DNN variant or using 1×1 convolutional layers for down- and upsampling [Szegedy *et al.*, 2015]. The current tendency seems to employ various tricks to make them as deep and slim as possible [Simonyan *et al.*, 2014]. Others rely on special architectures [Ronneberger *et al.*, 2015], which again encourage the network to learn in a certain direction instead of giving it a free hand.

The development is still ongoing and new methods are proposed weekly. Some of them, such as mini-batch training, deal with how to fit large networks into the available GPUs and might soon become obsolete. Others are merely variants of existing solutions or supersede their predecessors. At the time of writing, the process of careful sighting, comparing and categorizing the approaches has barely begun and will most likely still take some time

Any attempts on summarizing the current state or making predictions for the future are accordingly difficult. Running the risk of misinterpretation, I venture to state that the major part of the proposed methods constitute regularizers. By intelligently circumventing layers, modifying the network architecture, introducing special purpose layers, and other measures, the DNNs are tailored towards the task at hand and overfitting is avoided.

These networks are no longer universal, i.e., left completely to their own devices to learn the best decision function from the training data, but rather encouraged into certain directions and barred from others. Hence, although often praised as general purpose solutions for a wide range of classification tasks, DNNs that achieve state-of-the-art performance are usually carefully designed and elaborately tuned custom algorithms.

A distinct advantage they do hold over most traditional classifiers is that they can be trained on the raw data (e.g., multi-spectral magnetic resonance imaging (MRI) images) directly without the need to resort to a feature-based representation. If well designed, a DNN automatically learns the low-level representations most discriminative for the classification task at hand. Thus, no manual feature design is necessary and the available training information can be put to maximum use.

Comparing DFs and DNNs

DFs and DNNs are based on quite different concepts. But since they can both be used for classification tasks, they are competitors. Based on above short introduction to DNNs for classification and the description of DFs in Chapter 4, I attempt a short juxtaposition of the two method as far as that is possible.

DNNs learn their own features from the raw input data. This puts them at an advantage over DFs, which require discriminative features to be meticulously crafted. Design decisions for manual features necessarily depend on a string of assumptions and postulate a sound domain knowledge. Fortunately, DFs are robust against correlated and uninformative features. But they have no means to recover information implicit in the training data but not explicitly expressed by

the supplied features. DNNs, on the other hand, are theoretically able to learn the ideal features from appropriate training data and should thus reach at least human inter-rater performance.

In practice, limited learning time, overfitting, numerical instabilities, and other factors keep the networks from optimal performance. Nevertheless, thanks to the increased computational power and clever regularizers, they now reach top performances in a range of domains, including many fields of medical image processing.

One of the best ways to compare methods are public benchmarks. Examining the rankings in the public challenges presented in this thesis (Sec. 4.5) and a number of others carried out in the last year [1], it becomes apparent that many leaderboards are headed by DNN approaches. This seems to be especially true in the case of very challenging tasks (such as SISS), while lighter segmentation problems (such as SPES) are more often headed by DFs. Overall, DF methods are found among the highest ranking submission and the mid-table, while DNNs are encountered in all positions.

It seems that a carefully designed DNNs reaches top performance, while a less ideally planned network easily fails the task completely. DFs reach state-of-the-art performance if well designed and, even if not, one can still expect competitive results. Thus, it is possible that there is a ceiling of complexity the DFs are unable to breach. On the other hand, they are robust and generalize well. Both of these points are likely connected since strong generalization renders a classifier robust against flawed training data while simultaneously limiting its ability to handle underrepresented and overly complex problems.

Both methods require adaptation to the task they are intended for. In the case of DFs, this means crafting features and tuning the hyperparameters. The former is challenging, the latter comparatively simple. The forests are robust against the exact choice of their hyperparameters and most parameter optimization curves have a log-like shape with a wide plateau (see Sec. 4.4.1, in particular Fig. 4.12). Furthermore, since they train comparatively fast (Table 4.7), exhaustive parameter optimization schemes are applicable.

Employing a DNN instead of a DF means trading the complexity of the feature design against the complexity of the architectural design. Where the former requires a sound knowledge of the task's background (e.g., where can the target lesion appear anatomically), the latter requires a good grasp of the image material (e.g., where can the target lesion appear in the image space). Beside the architectural design, other decision have to be made: Which layers to employ, in which combinations, which regularizers to use, what learning rate to set and how to adapt it, and many more. In general, DNNs are harder to tune than DFs as their hyperparameters are highly interdependent and the long training times forbid an exhaustive optimization.

DNNs belong to the family of black box classifiers, i.e., it is near impossible to understand what they are actually learning or to follow their decision process. DFs, on the other hand, are white box classifiers. The decision process can be readily investigated by examining the nodes of each tree. Whether any new insight can be deduced from such an observation is another matter.

Summarizing, DF are a good choice for simpler tasks, proof of concepts, prototypes and training. They are well researched, simple to understand and competitive results are fast and easily obtained. In cases where optimal performance is sought, DNNs seem the better choice. Although requiring more time for understanding, training and tuning, they can be optimized to outperform other classifiers in many tasks. Furthermore, they are still under active development and likely to further improve in the future. Finally, the available processing power increases continuously, constantly lowering one of the main hurdles of DNNs.

[1] `grand-challenge.org`

Appendix C

Magnetic resonance imaging

Magnetic resonance imaging (MRI) is a non-invasive imaging modality primarily employed to image the anatomy and the physiological processes of the body in both health and disease. The technique is based on strong magnetic fields and radio frequency (RF) waves to trigger faint signals from unbound protons, i.e, principally water molecules. Compared to other imaging techniques, it is highly configurable and can hence be tailored to the specific needs at hand. In this appendix, the basic mechanisms of MRI will be described as far as required to understand the relation between anatomical tissue and image intensities. It serves as an addition to the general MRI Chapter 2 for the interested reader, treating the image acquisition process in more detail than there.

C.1 The MRI scanner and its main components

The typical MRI scanner is a large, tubular machine with an attached table to move the patient into the opening during the scanning process (Fig. C.1). Its four main components are a large, superconducting, tubular main magnet, a number of gradient coils, incorporated and/or optional RF modules, and a controlling computer (Fig. C.2).

The powerful main magnet creates a uniform field running through its bore (the tube) and the axial patient plane. Typical field strengths are 1.5 and 3 Tesla, but also 7 Tesla machines are employed. Outside the tube, an external fringe field is generated (Fig. C.3a). Since interferences in the outer field cause inhomogeneities in the internal field, it is usually shielded. This main magnetic field remains activated during the whole scanning process.

The gradient coils, each dedicated to one spatial direction (x, y or z), are used to produce field strength gradients in the otherwise homogeneous main magnetic field (Fig. C.3b). They allow for a deliberate fine control of the magnetic field properties at resolutions of under one millimeter by carefully manipulating their angles, directions, and power.

The RF system provides the communications link with the patient's body for the purpose of producing an image. It serves simultaneously as sender and receiver, able to emit strong RF at a predetermined frequency and to record the faint responses of the excited protons.

Finally, the controlling computer coordinates the sophisticated interaction between above systems and collects the scanning data.

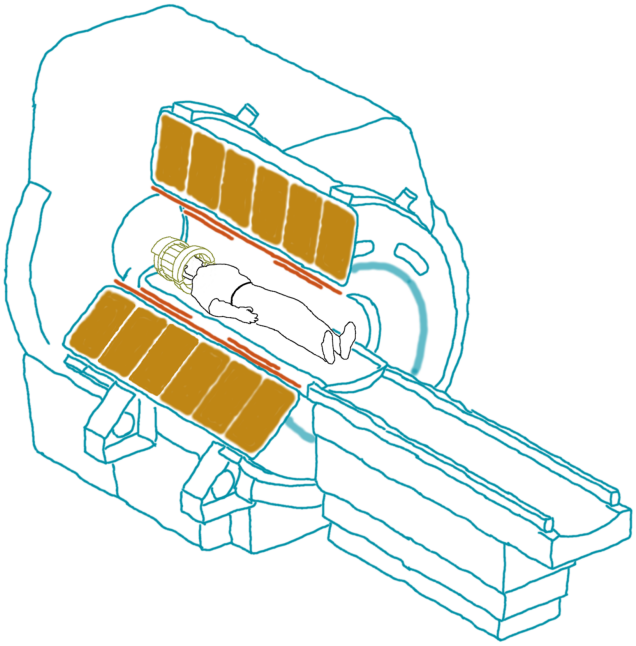


Figure C.1: Cut-away of a exemplary MRI scanner with main magnet (yellow), gradient coils (orange) and a head RF coil (green). Refer to online version for colors.

C.2 Field-Proton interactions

To correctly interpret an MRI image, it is essential to understand the relationship between the actual anatomical tissue and the final intensity values obtained. The MRI modality is very flexible and allows for various tissue properties respectively tissue property derived signals to be measured. On the downside, this increases the complexity of interpretation. Here, the interactions between the magnetic field, RF pulses and the tissue protons will be described in some detail, to later build on these basics when introducing the various sequences and scanning techniques relevant for this work.

MRI is based on nuclear magnetic resonance i.e., an interaction with nuclei that works as follows: If placed in a strong magnetic field, certain materials, such as body tissue, become magnetized and take on a resonant characteristic proportional to the magnetic field strength. If excited with a RF pulse tuned to the resonant frequency, they respond by emitting a signal whose characteristics depend on the tissue properties.

This interaction takes place on the level of the atomic nuclei, whose type, distribution, and structure actually determines the tissue properties measurable with MRI. Nuclei consist of protons and neutrons. These have a property termed spin, an intrinsic form of angular momentum, which is usually illustrated as a piercing arrow (Fig. C.4) denoting the rotation axes. If a nucleus possesses an uneven combination of protons and neutrons, they do no longer cancel out their respective spins and the nucleus will exhibit a net spin characteristic. Together with its electric charge, the spinning produces a magnetic property known as the magnetic moment. Its direction is depicted by the arrow head (Fig. C.4).

To contribute to the measures response signal, nuclei have to posses a strong magnetic moment, be part of a free moving molecule and possess a number of additional characteristics. Inside

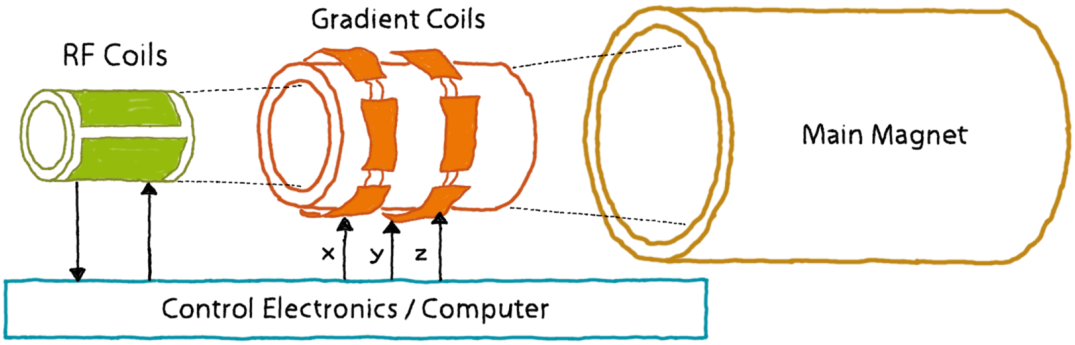
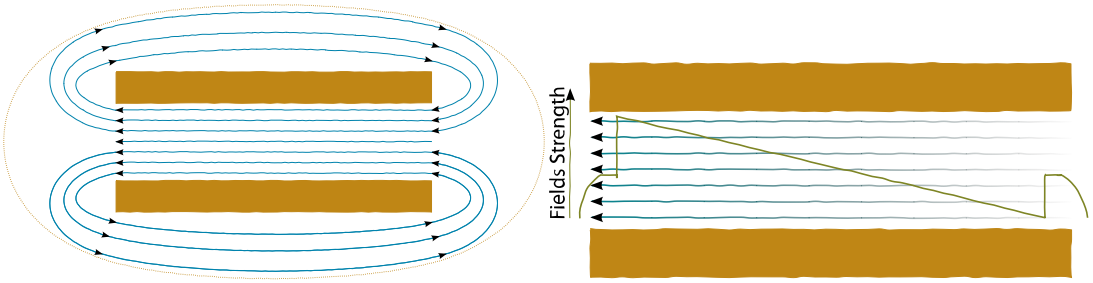


Figure C.2: The main components of an MRI scanner and their nesting. Refer to online version for colors.



(a) Homogeneous inner and fringe field created by the main magnet inside its bore. The dotted line denotes the 1 Gauss border of the outer field component.

(b) Gradient field created by superimposing a gradient onto the main field.

Figure C.3: Magnetic fields of an MRI scanner.

the human body, the only abundant (62% of all atoms) element fulfilling these requirements is hydrogen, which is also know to produce the strongest response. Its isotope H-1 consists of a single proton and it is largely bound in water molecules, which are free moving. Hence, for all practical purposes, what MRI measures is hydrogen density, which in turn is directly related to water content.

The RF response signal of the protons is the results of a careful and purposeful manipulation of their spin directions. When subjected to a strong, uniform magnetic field exerting torque on them, some protons align their spin axes to the main field direction (Fig. C.5, second column).

But the alignment is not absolute. The proton's spin axis 'wobbles' around the main field direction at a frequency directly proportional to the field strength. This effect is called proton precession and renders the proton sensitive and receptive to RF pulses of the same frequency as its rotation (Fig. C.5, third column). Hence, the main magnetic field serves to tune the protons to the required resonance frequency. When summarized over a small volume, the magnetic moments of the protons form the tissue magnetization. Since the direction is aligned to the main magnetic field, it is termed the longitudinal magnetization of the tissue.

But the longitudinal magnetization creates no signal that can be measure. This is where the RF comes in: A RF pulse generated by the RF coils at the targeted proton's resonance frequency

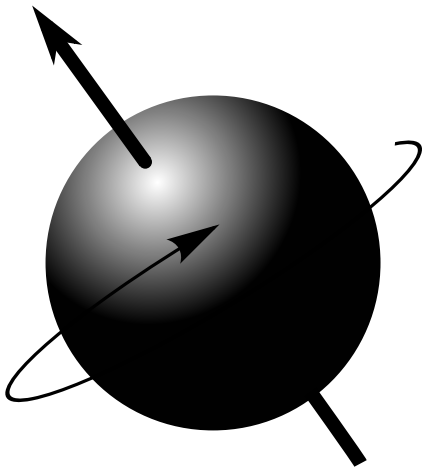


Figure C.4: Simplified schema of a proton spin.

injects energy and causes the proton's spin axes to flip from the longitudinal orientation (i.e., aligned to the main field's direction) to an angle determined by the pulse's strength and duration (Fig. C.5, fourth column). The proton is now in an excited state and rotates in the transverse plane of the magnetic field, causing transverse magnetization of the tissue and simultaneously faint RF signals, which can be picked up by the RF coils and translated to intensity values to form an image.

After disabling the RF pulse, the so-called T1 proton relaxation commences (Fig. C.5, last column). Over a short time, which duration depends on the flip angle and various tissue properties, the proton transfers its excess energy and re-aligns to the magnetic field.

But in practice, the recovery of longitudinal magnetization is much slower than the loss of transversal magnetization. Since MRI records the collective magnetization over a small volumetric element, a voxel, a large number of protons contribute to the longitudinal as well as transversal magnetization. And the total magnetization of the tissue voxel is the sum of its protons' magnetic moments. That means in turn that the transversal magnetization is only kept intact when all protons are rotating largely in phase in the transverse plane of the magnetic field. Directly after applying the RF pulse, this holds true and the magnitude of the transversal magnetization corresponds to the previous longitudinal magnetization. But two effects contribute to fast unphase the proton rotation: The spin-spin interaction (T2 relaxation) and inhomogeneities in the magnetic field. While the first constitutes a intrinsic tissue property and hence a desirable effect, the second is tissue unrelated and undesirable. Unfortunately, the inhomogeneity effect is stronger and masks the spin-spin interaction. The cumulated decay is termed T2* relaxation.

C.3 The magnetization cycle and the three types of spin relaxation (T1, T2 and T2*)

T1 and T2 relaxation are important concepts for understanding MRI imaging. Fig. C.6 serves to illustrate the development of the two measures over time and to related them to the proton behavior.

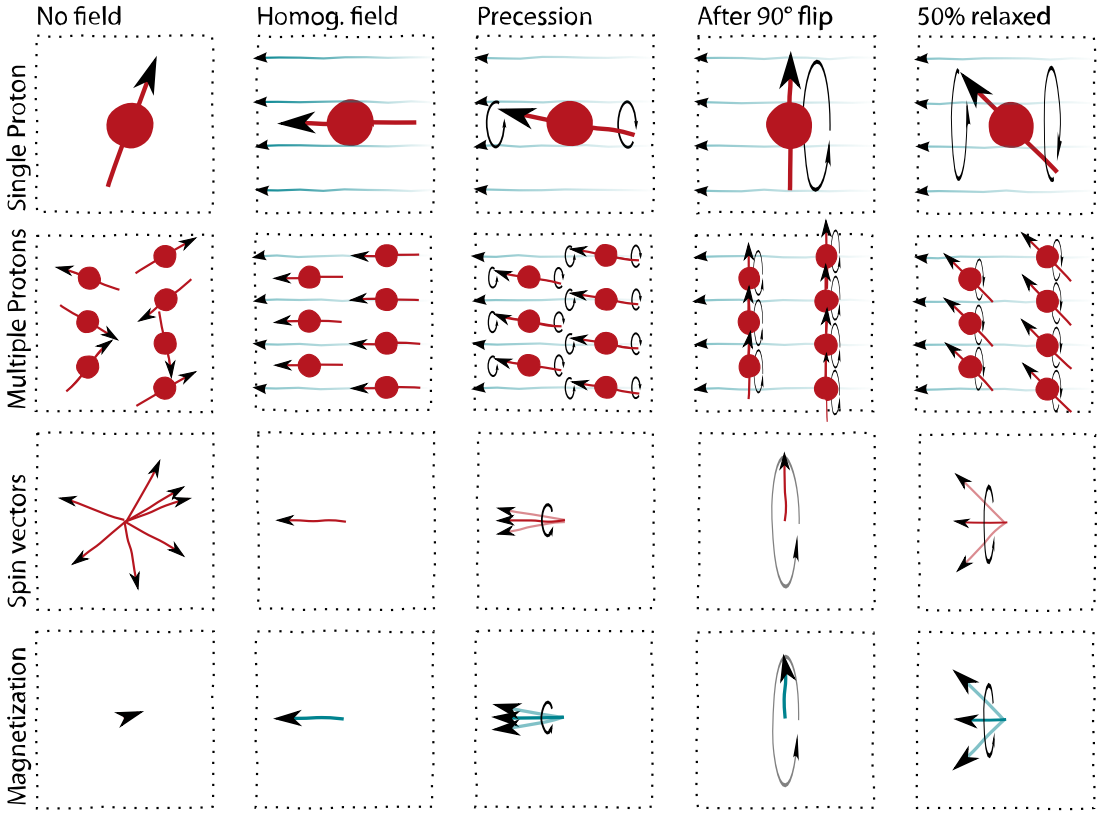
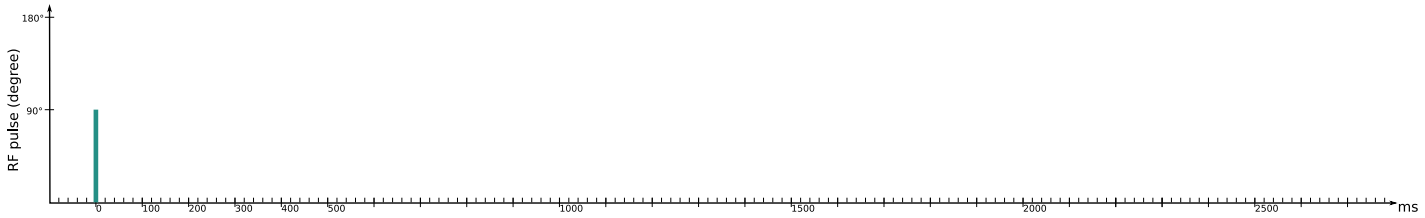
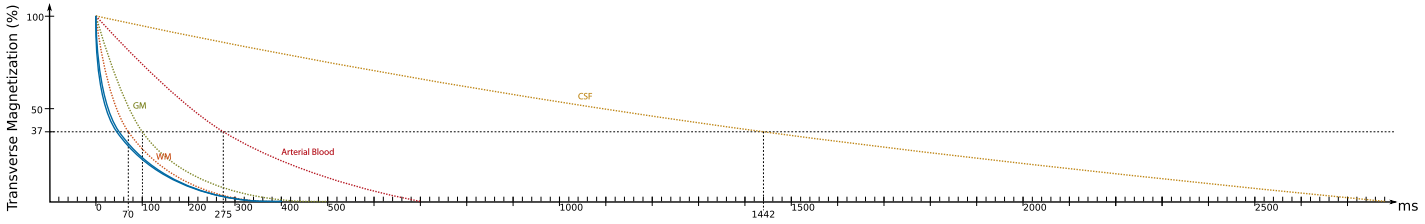


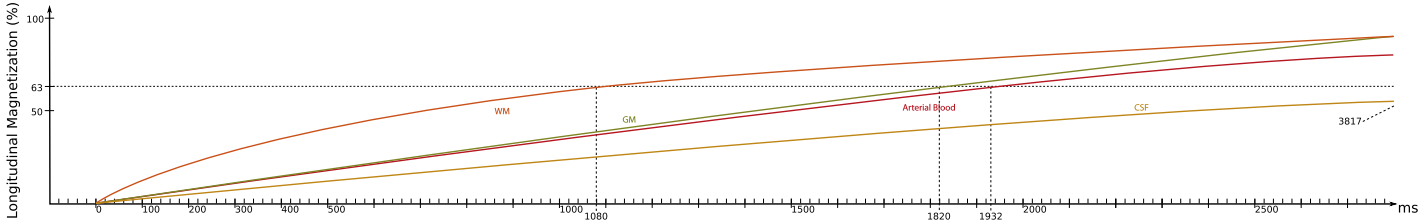
Figure C.5: Simplified schema of proton spin behavior in magnetic fields and under RF impulses. The rows illustrate: (1) a single affected proton, (2) multiple protons inside a tissue voxel, (3) their spin vectors, and (4) their summed magnetization. The columns illustrate: (1) random orientation without magnetic field, (2) idealized alignment to strong uniform magnetic field, (3) precession movement in uniform magnetic field, (4) orientation and rotation immediately after a 90° RF impulse, and (5) situation after 50% of the re-alignment to the uniform magnetic field is realized (T1 relaxation).



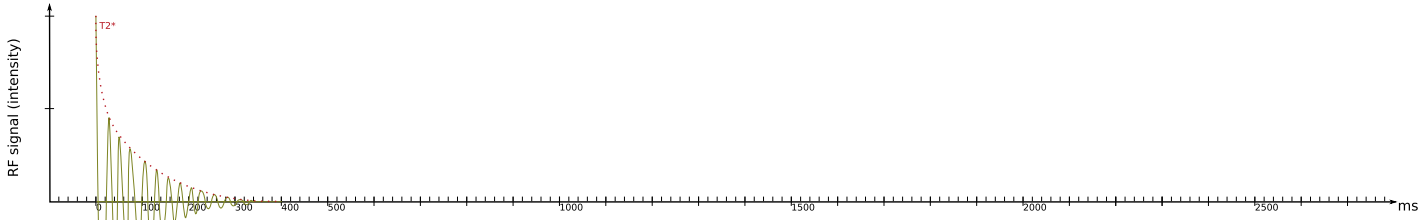
(a) RF pulse: Send to trigger a 90° spin axes flip in the protons.



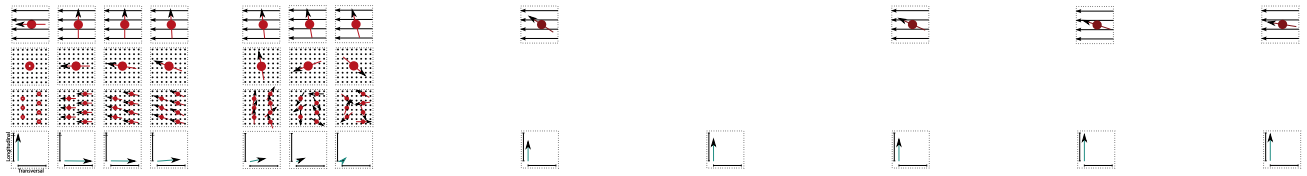
(b) T2 relaxation: Rapidly decaying transversal magnetization.



(c) T1 relaxation: Slowly recovering longitudinal magnetization.



(d) RF signal: Faint RF signal emitted by the tissue during transversal magnetization.



(e) Proton behavior: Single protons, multiple protons and overall magnetization.

Figure C.6: Timeline of a single flip-relaxation cycle. Refer to online version for details.

C.3.1 Longitudinal and transversal magnetization

Before the trigger time point $t = 0ms$, the protons had time to partially align to the main magnetic field and hence to accumulate the maximum longitudinal magnetization for the target tissue. The first graph denotes the RF pulse activated at the time $t = 0ms$ and set to resonance frequency, which causes an immediate 90° flip of all protons spin axes (Fig. C.6a). The longitudinal magnetization is hence lost and converted directly into transversal magnetization.

Now T2 relaxation commences, causing the transversal magnetization to diminish rapidly (Fig. C.6b). The solid lines denote the actually observed magnetization decay (T2* relaxation), which is mainly driven by field inhomogeneities and hence very similar for all tissue types. The desired T2 relaxation are denoted by dotted lines. The exact form of this falling functions depends on the tissue properties and the graph sketches some typical progressions for tissues of interest in brain imaging. How to eliminate the undesired part of the T2* effect and obtain the pure T2 signal will be discussed later in this chapter.

The T1 relaxation is equally triggered by the RF pulse, describing the comparatively slow recovery of the longitudinal magnetization (Fig. C.6c). The depicted lines denote typical recovery patterns for some tissue types.

Note that the maximum longitudinal magnetization of the tissue types equally differs, although not visible here. This factor is largely determined by their PD.

C.3.2 Proton spin change over the magnetization cycle

The orientation and rotation of the single protons and multiple protons inside the observed tissue voxel are depicted below the graphs for a number of time points (Fig. C.6e). Note the rapid dephasing of the protons compared to the comparatively slower re-alignment. The fourth row denotes the total as well as partitioned magnetization of the tissue voxel.

Only the transversal component causes a RF signal, which intensity depends on the transversal magnetization magnitude and the tissue type (Fig. C.6e, third row). I.e., its intensity is directly proportional to the T2* relaxation, as denoted by the dotted line (Fig. C.6d).

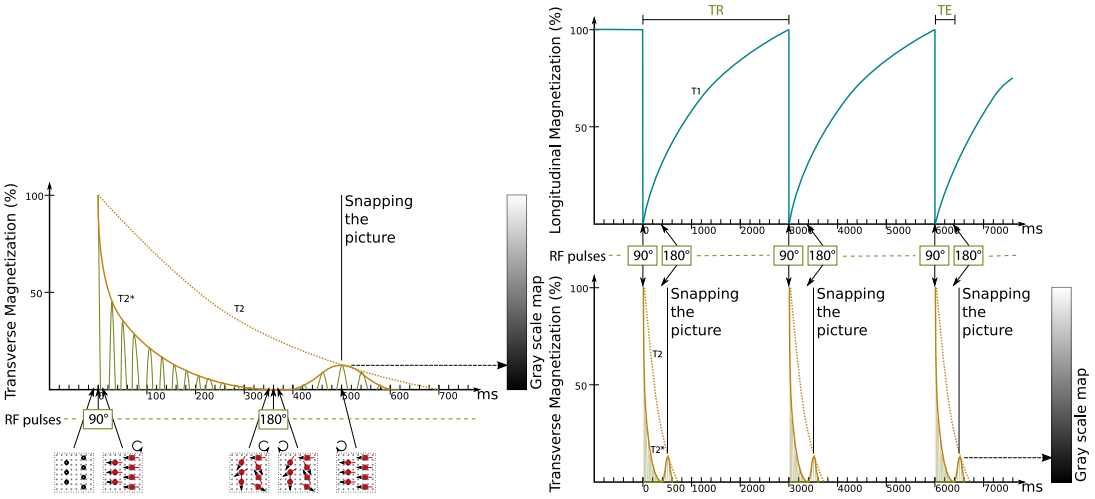
After roughly 2400ms, the longitudinal magnetization of most tissue types has recovered completely and the initial state is reached (Fig. C.6e, first row).

C.4 Measuring the magnetization

Looking at the proton excitation and relaxation cycle (Fig. C.6), three intervals can be observed in which the different tissue types produce distinctly different signals: The T1 relaxation, the T2 relaxation and the maximum longitudinal magnetization (i.e., PD) intervals. A measurement, i.e., taking a picture, corresponds to a vertical cut through the graph and a direct mapping of the values to a gray-scale map. Hence, the vertical distance between the curves of two tissues denotes the contrast between them and the value at the time point of measure the intensity of the image pixel.

C.4.1 Measuring the T2 relaxation

The magnitude of transversal magnetization relates directly to the intensity of the emitted RF signal (Fig. C.6, last row). To measure the T2 relaxation of a tissue voxel, the RF signal can be mapped directly to the intensity scale. But the observed transversal magnetization decay T2* is dominated by the inhomogeneity caused proton dephasing (Fig. C.7a, solid line), which hast to be filtered out.



(a) The recording procedure.

(b) Multiple recordings.

Figure C.7: Recording of a tissue voxel's T2 relaxation properties.

This undesired portion follows a regular pattern, i.e., some protons rotate faster and others slower at a difference determined by the static field inhomogeneities (Fig. C.7a, orientation and rotation of protons in transversal field). When subjected to an 180° RF pulse, the spin direction of the photons is reversed. Since the faster ones still spin faster and the slower ones slower, a rephasing takes place, increasing the transversal magnetization up to its T2 capped values. This approach works, because the spin-spin interaction causing the T2 relaxation is not reseted by reversing the spin rotation. The emitted RF signal (Fig. C.7a, sinusoid curve) shows a clear peak value, which is the desired T2 relaxation value. Mapped to a linear gray value map, an T2 image is obtained, which denotes the T2 relaxations times of the tissues.

All measures are determined by two parameters: The echo time (TE) and the repetition time (TR). The TE represents the time in milliseconds between the application of the 90° flip pulse and the peak of the echo signal, i.e, when the image is snapped. Its value determines the contrast between different tissue types as well as the overall signal intensity. The nature of MRI recording and image reconstruction requires the repeated recording of a signal from each tissue voxel. The duration of such a cycle is set by the TR parameter, given in milliseconds.

Fig. C.7b illustrates these values at the example of repeated T2 relaxation recordings. The TR interval includes a waiting period for the longitudinal magnetization to recover to its maximum magnitude, hence TE<<TR.

C.4.2 Measuring the positron density

Another standard MRI sequence is the PD sequence, which measures the positron density of the tissue (Fig. C.6). Since the longitudinal magnetization cannot be measured directly, a 90° pulse is applied to flip the direction of the magnetization while preserving its magnitude. Then the same measurement as for the T2 sequence is applied, i.e., a 180° pulse. The TE value has to be chosen relatively short to actually capture the PD contrast and not the T2 relaxation times.

C.4.3 Measuring the T1 relaxation

The third default MRI sequence is the T1 sequence, which reflects the T1 relaxation times of tissue. By choosing a smaller TR value than for PD, the magnetization is flipped when the T1 contrast is strongest. The TE value is chosen equally short as for PD to avoid the T2 relaxation influence.

Hence, T2 sequences are characterized by large TR and large TE values; PD by large TR and short TE; and T1 by short TR and short TE.

C.5 Advanced acquisition techniques and acquisition details

Above described recording technique is termed the single spin echo method. Other methods, such as the gradient echo, exists. Furthermore, various improvements have been introduced over the years to shorten the recording time, e.g, by recording multiple echos during one excitation phase, by exciting multiple tissue slices at once or by recording two sequences at once. While all of these also influence the resulting image quality and the frequency of imaging artifacts, they exhibit only a marginal impact on the image's appearance and are therefore not discussed here.

Slice-wise MRI image acquisition is performed by tuning the target slice with a selection gradient to the desired RF frequency and to then emit a RF pulse of that frequency. Only the protons inside the tissue of this single slide with be excited and hence create and echo pulse. Before and during the echo event, two additional gradients (phase encoding and frequency encoding) are applied to allow to distinguish between different rows and columns, and hence the individual voxels. The received tissue echo RF signals are stored in the so-called k-space and finally transformed into the image during image reconstruction. Other techniques allow for the acquisition of multiple slices at ones and even complete 3D volumes, but the descriptions of these techniques are out of the scope of this work.

C.6 Non-standard MRI sequences

The operator can select from many preset protocols for specific clinical procedures or change protocol factors for special applications.

C.6.1 Contrast agents

By injecting paramagnetic material with a high magnetic susceptibility, such as Gadolum, a high contrast for specific tissues or liquids can be achieved. The obtained sequences are T1 contrast enhanced (T1c).

C.6.2 Fluid attenuated inversion recovery (FLAIR)

FLAIR is a recording technique based on the T2 sequence with the fluid signal suppressed. In brain scans, that means that the cerebral spinal fluid (CSF) signal, which is usually normally bright, appears dark in the images. This allows for a better distinction between it and the deep WM adjacent to the ventricles as well as improving the fissures' contrast. To achieve this affect, a 180° RF pulse is applied, resulting in a negative longitudinal magnetization (Fig. C.8). Since the T1 relaxation of fluid is the lowest of all tissues, it crossed the zero-magnetization last. At

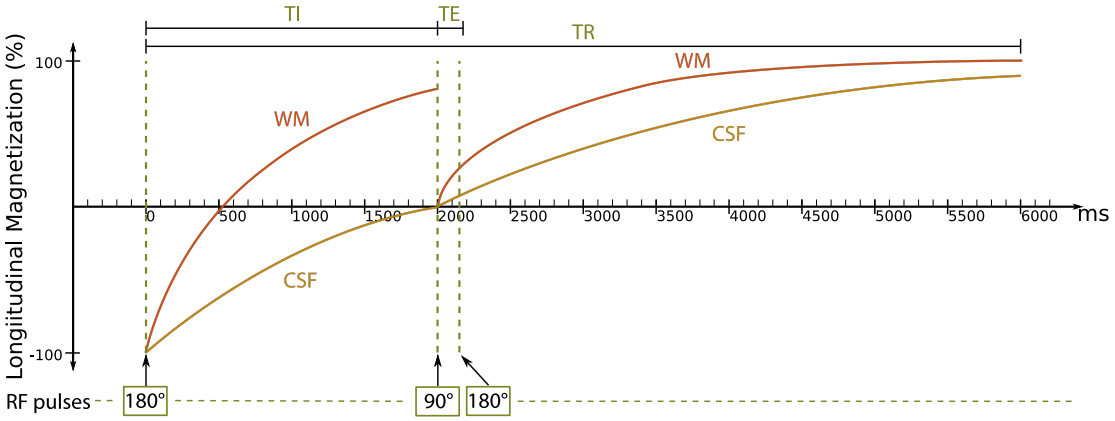


Figure C.8: Recording a FLAIR sequence suppressing the fluid signal. The inversion time (TI) is set to 2000 ms, the time of CSF zero-crossing. The image recorded is T2 weighted, snapped TE=160 ms after the 90° RF flip, capturing the zero signal of the CSF. Since all other tissues recover faster, they show up similar as in a usual T2 scan.

this moment, the 90° RF pulse is triggered, allowing to record a T2 image where the fluid signal is zero, i.e., mapped to black. This requires an additional scanning parameter, the TI.

Fig. C.9 illustrates the results of a FLAIR compared to the standard T2 sequence. The peripheral and periventricular pathological areas that border at CSF are better distinguishable in the FLAIR image.

C.6.3 Diffusion weighted imaging (DWI)

DWI records the random Brownian motion of water molecules, i.e, their diffusion inside the tissue.

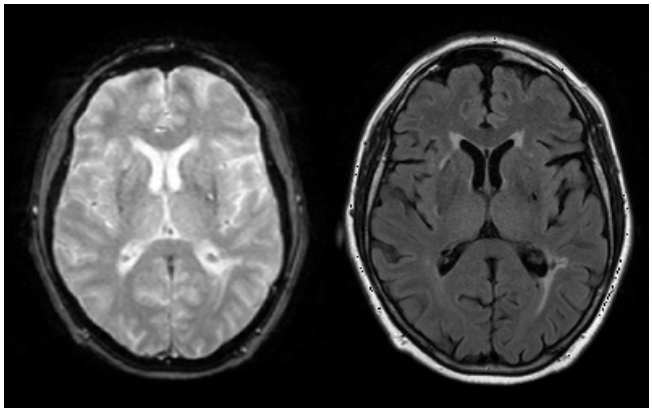
Most fluids and homogeneous solid materials allow for the same diffusion in all directions. These are termed isotropic with a single diffusion coefficient D. In virtually all biological tissue, the diffusion is restricted by its internal molecular structure, primarily by cell membranes. This results in different diffusion properties along different directions, rendering the tissue's diffusion property anisotropic. E.g., white matter is highly anisotropic because of the parallel orientation of its nerve fiber tracts.

To describe diffusion in anisotropic materials, a diffusion tensor is used.

$$\mathcal{D} = \begin{bmatrix} D_{xx} D_{xy} D_{zx} D_{yx} D_{yy} D_{yz} D_{xz} D_{zy} D_{zz} \end{bmatrix} \tag{C.1}$$

Hence, changes in the cellular density of tissue and the amount of intracellular versus extracellular water will impact the degree of diffusion restriction within that volume of tissue. Various pathological conditions, such as ischemic infarcts within the brain, tend to restrict diffusion strongly.

To record the diffusion, two repeated gradient pulses are applied on both sides of the spin-echo's 180° RF pulse (Fig. C.10). Gradient pulses cause a phase shift in proton spins, whose magnitude depends on the induced energy. The first causes a forward phase shift of the protons and the second, exactly equal gradient pulse, shifts them back (since the intermediate 180° RF pulse inverted the spin direction). For stationary spins, the forward and backward shifts equal out perfectly, resulting in a high RF signal. But diffusing spins that meanwhile have moved



(a) T2 (b) FLAIR

Figure C.9: Comparison of T2 and FLAIR. In the latter, fluids do no longer appear bright, allowing for a better contrast between lesions and CSF (e.g., periventricular).

along the gradient direction will receive a different backward shift energy and hence emit a phase shifted RF signal. The magnitude of this shift depends on the distance traveled.

Fig. C.11 illustrates this process by means of some randomly moving protons. During the first gradient pulse, all protons are located in the volume of interest and receive the same amount energy triggering equal forward phase shift. During the intermediate period, they move randomly (here denoted by the yellow path) and their spin is, at some point, inverted by the 180° RF pulse. At the time of the repeated application of the same gradient pulse, the spins' phase shift back according to the pulse's energy at their exact location: Stationary protons receive the same amount of energy as during the first gradient pulse, turn back into their original phase and emit a strong RF signal. Protons with a net movement (pink arrow) perpendicular to the gradient's direction are equally unaffected. But protons that moved at least partially along the gradient direction will now receive a backward phase shift pulse of a different energy (and hence magnitude) than the previous forward phase pulse, resulting in a phase shifted signal compared to the stationary protons and hence a lower RF signal emitted. When recording the image after the second gradient pulse, it will thus contain contrast caused by proton diffusion in the tissue voxel considered.

The recording procedure is repeated in all three gradient directions (x, y, z) to fill the diffusion tensor elements (Eq. C.1).

The signal S measured during DWI is defined as

$$S = S_0 e^{-bD} \tag{C.2}$$

S_0 denotes the signal strength at baseline, i.e., when no gradient pulses are applied (which, for all practical purposes, is a T2 image). D is the diffusion coefficient (given in $\mathrm{mm}^2\,\mathrm{s}^{-1}$), which denotes the average flux of a group of particles inside the observed tissue voxel across a surface. The last parameter is the b-value, which depends on the operator's settings and characterizes the DWI sequence recorded.

The b-value, which is defined as $b \propto q^2 * \Delta\,\mathrm{s\,mm}^{-2}$, i.e., proportional to the gradients' strength (q) and the time between them (Δ), determines the sequences susceptibility to the random water molecule movement (Fig. C.10). The exact relationship depends on the shape of the gradient pulses (rectangular, trapezoid, etc.). The larger the b-value chosen, the darker the image overall

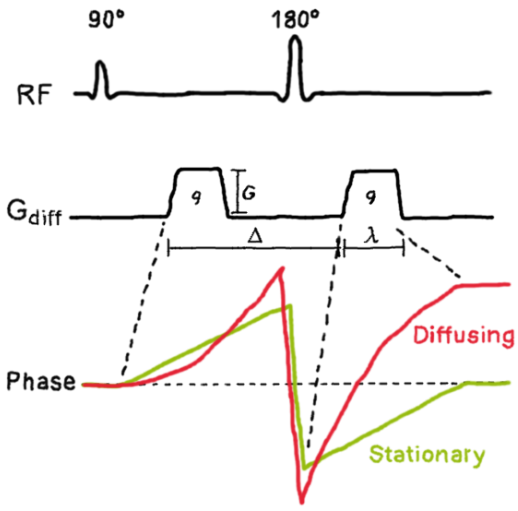


Figure C.10: Activating gradient pulses on both sides of the spin-echo's 180° RF pulse allows to capture water molecule diffusion.

but the brighter the seriously restricted areas. Lower b-values, on the other hand, lead to image contrasts only mildly affected by the diffusion properties of tissues. A b-value of 0 is equivalent to a T2 sequence.

Usually multiple diffusion images with different b-values are recorded as illustrated in Fig. C.12. These typically range from 0 to 2000 $\mathrm{s\,mm^{-2}}$ for brain scans, since larger values lead to a worse signal-to-noise ratio.

A typical DWI scan usually commences with the scanning of a $b0$ image (Fig. C.13a), followed by the acquisition of at least three perpendicularly oriented directed diffusion *source images* at a chosen b-value. E.g., in x-direction this would result in an image with intensity values of $S_x = S_0 e^{-bD_{xx}}$ (Fig. C.14).

These source images are intermediate products, which are usually combined into a single image termed *diffusion-weighted* or *trace image*, usually through the formation of the geometric mean:

$$D_{DWI} = \sqrt[3]{S_x S_y S_z} = S_0 e^{-b(D_{xx}+D_{yy}+D_{zz})/3} = S_0 e^{-bD_{trace}/3} = S_0 e^{-b*ADC}. \tag{C.3}$$

The trace, i.e., the sum of the diagonal elements of the diffusion vector (Eq. C.1) reduces the multi-dimensional flow to a single value, which is termed the ADC. Bright areas in the trace image denotes restricted and dark areas free perfusion (Fig. C.13b).

It is important to keep in mind that the trace images are not a map of the diffusion but rather diffusion weighted image. That means that they possess considerable T2 weighting. This causes an effect termed T2 shine-trough, which can contaminate the trace image: When a pathology appears bright in the T2 sequence, it will equally appear bright in the trace image, despite possibly unhindered diffusion.

To counter this effect, an ADC map can be computed. By combining the information of multiple trace images recorded with different b-values, the pure diffusion effects are sought. Hence, the ADC map denotes the magnitude of diffusion (in $\mathrm{mm^2\,s^{-1}}$).

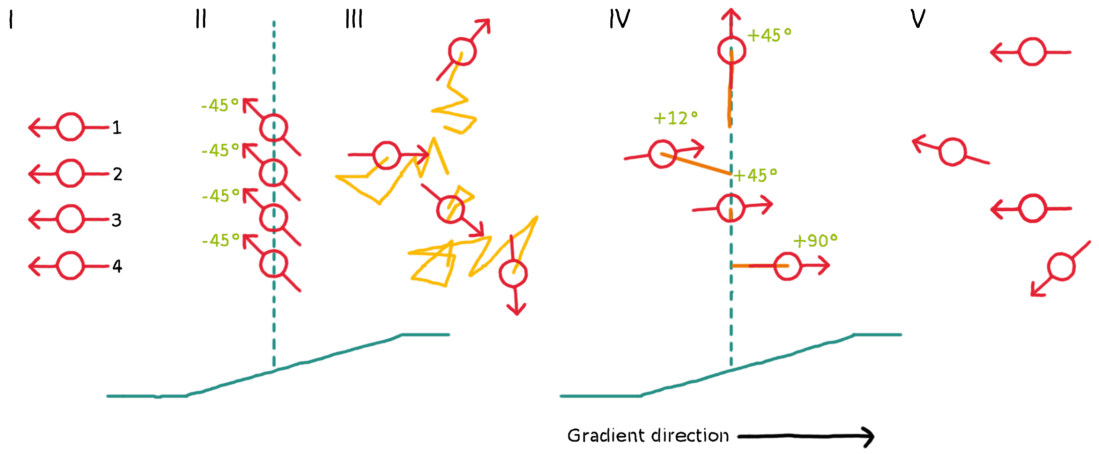


Figure C.11: The gradient pulses' effect on randomly moving protons' phases in a tissue voxel: Initially all protons spin in the same phase. The first gradient pulse shifts their phase forward by the same degree (here: 45°). During the interim period, the protons are subject to random movement (yellow path). As in the standard recording procedures, spin-spin interaction (causing T2 relaxation) desynchronizes their phases and a 180° RF pulse flips their spin directions to eliminate the effects of field inhomogeneities. At the time of the second gradient pulse, the component of along the gradient's axis of the total net movement determines the backward phase shift each proton receives. At the time of image snapping, only the stationary and perpendicularly moving protons (3 and 1) are in phase, the other two (2 and 3) are shifted. The more the protons in the observed tissue voxel move, the lower the final RF signal of this voxel.

white matter	670-800
cortical grey matter	800-1000
deep grey matter	700-850
CSF	3000-3400

Table C.1: Approximate ADC values in $10 \times 10^{-6}\,\mathrm{mm^2\,s^{-1}}$ for brain tissue.

Considering the last part of Eq. C.3, the ADC values are defined as:

$$D_{ADC} = -\frac{1}{b}\ln(\frac{S_{DWI}}{S_o}) \tag{C.4}$$

Bright areas in the ADC maps denotes free and dark areas restricted perfusion, i.e., the gray values are usually inverse to the trace image's. This is illustrated in Fig. C.15.

But if a T2 shine through effect is present, the affected area appears bright in both, trace image and ADC map (Fig. C.16).

Conversely, the T2 blackout appears dark in both or even causes calculation artifacts (Fig. C.17).

Since ADC maps, unlike trace image, are b-value independent representations of the physical apparent diffusion coefficient, some rough values exists to identify different tissue types (Table C.1).

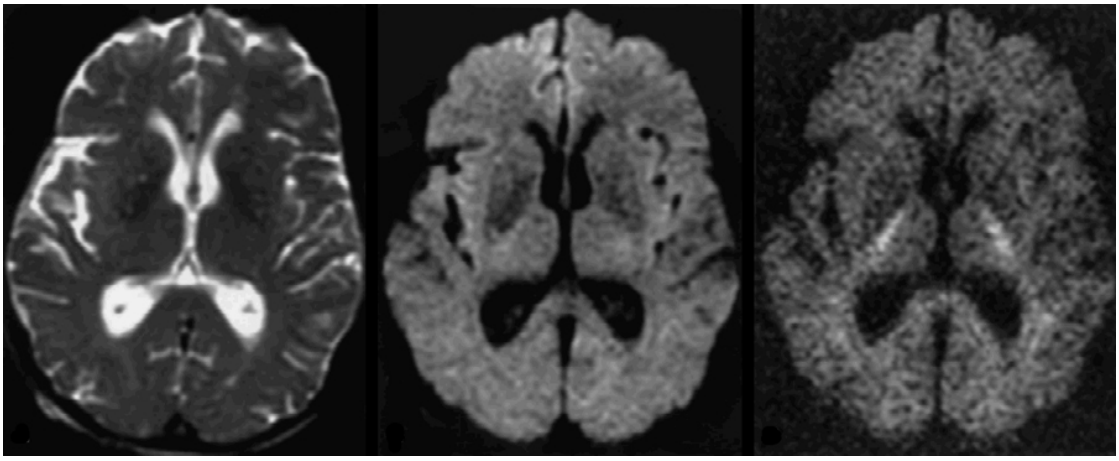
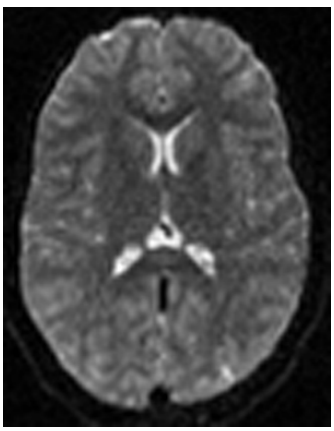
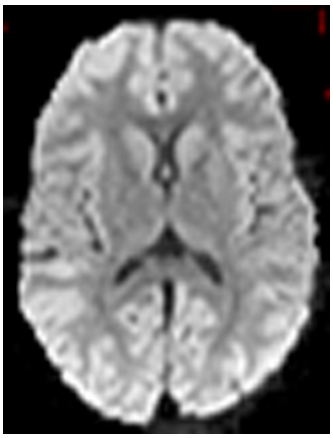


Figure C.12: DWI trace images recorded with different b-values (0, 1000 and 3000 from left to right).



(a) DWI b0 trace image, i.e., just showing T2 contrast without any diffusion weighting.



(b) DWI b 1000 trace image, i.e, the geometric mean of the three source images.

Figure C.13: Examples of trace images

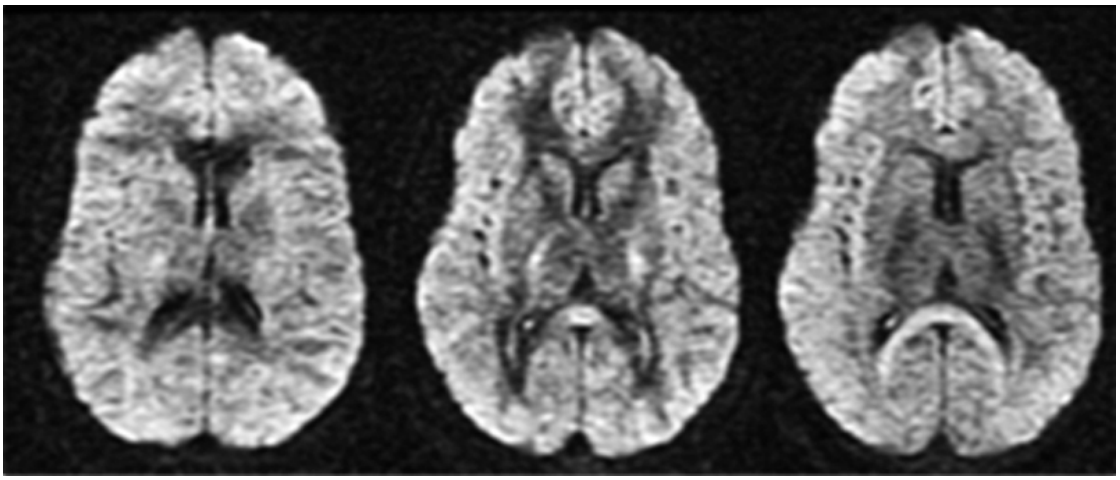
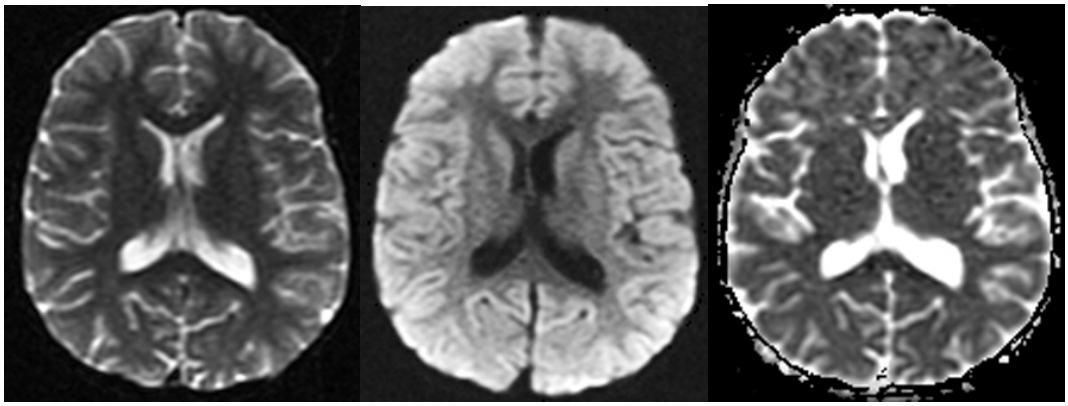


Figure C.14: DWI source images recorded in three perpendicular directions (x, y and z from left to right).

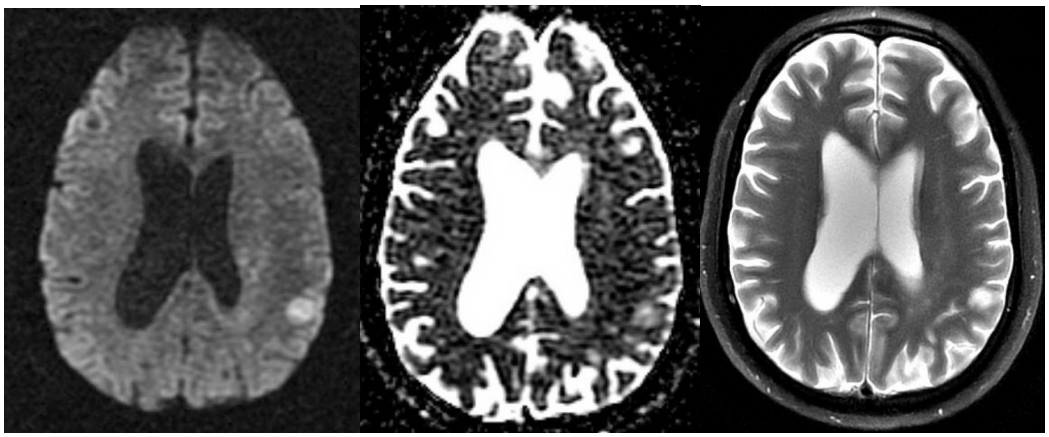


(a) The $b0$ image.

(b) The corresponding trace image recorded with $b = 1000$.

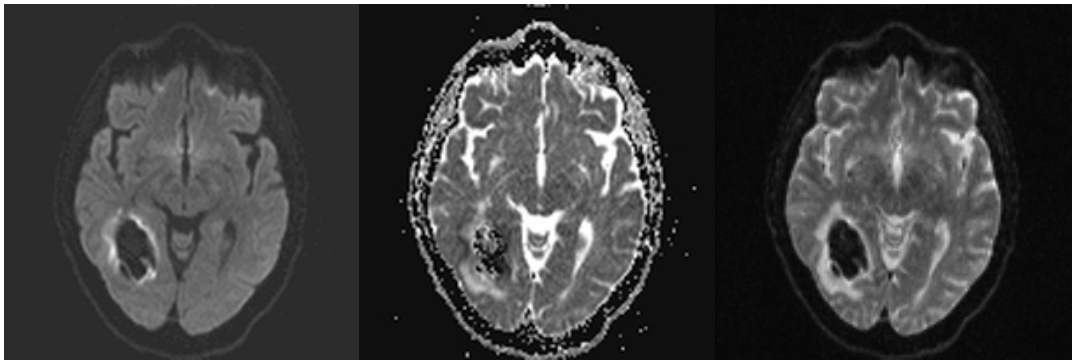
(c) The corresponding ADC map.

Figure C.15: Comparison between $b0$ image, trace image and ADC map.



(a) Trace image. (b) ADC map. (c) T2 image.

Figure C.16: Example of the T2 shine through effect: The hyperintense area in the trace image seems to denote restricted perfusion. But it appears equally hyperintense in the ADC map, hinting towards a T2 shine through effect, which is confirmed by the T2 image.



(a) Trace image. (b) ADC map. (c) T2 image (here: $b0$ image).

Figure C.17: Example of the T2 blackout effect: The hypointense area in the trace image seems to denote free perfusion. But it appears equally hypointense and with calculation artifacts in the ADC map, hinting towards a T2 black out effect, which is confirmed by the T2 image.

C.6.4 Perfusion weighted imaging (PWI)

To record the tissue's perfusion with blood, perfusion weighted imaging can be employed. A common technique is the so-called Dynamic Susceptibility Contrast (DSC) perfusion MRI, where a Gadolinium contrast bolus is injected and passes through the cardiovascular system. Its properties cause a faster T2/T2* decrease in the penetrated tissue. Hence, the more blood passes through a tissue voxel, the lower its T2/T2* signal during the passage of the bolus (Fig. C.18). During this time, repeated T2* images are shot. A voxel-wise analysis of this time-resolved data

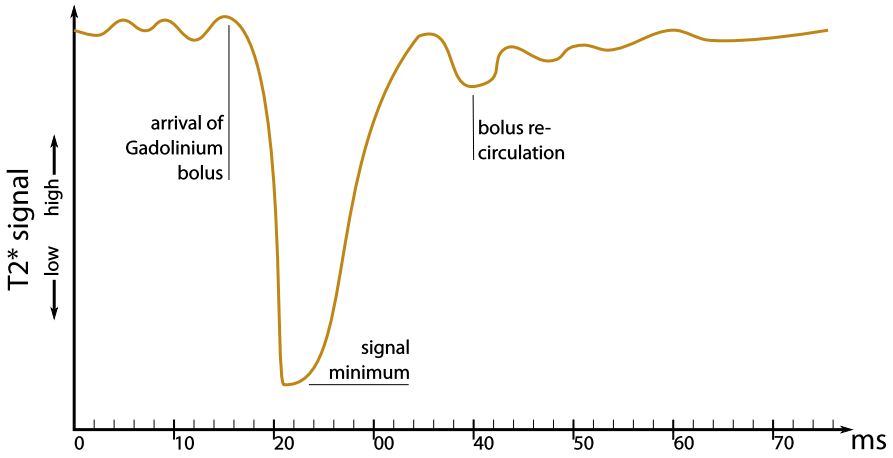


Figure C.18: Variations in the T2* signal emitted by a scanned voxel during the passage of a Gadolinium bolus.

allows to plot a DSC intensity curve for each voxel, representing the passage of the bolus over time characterized by a single, sharp minimum.

From this curve, a number of semi-quantitative parameters can be obtained. Of relevance for this work is only the TTP. When interpreting the resulting maps, however, it should be kept in mind that their values are highly dependent on the efficients/compactness of the contrast bolus and carry little meaning by themselves. Rather, they should be used in a relative sense, e.g., by comparing one hemisphere against the other.

Further analysis of the data allows to compute some additional parameters, such as the CBV, the CBF, the MTT and the Tmax. The procedure requires the estimation of gadolinum concentration from the change in T2*-relaxation rate, a semi-automatic selection determination of the arterial input function (AIF), which denotes the bolus influx, by marking the center of a feeder artery, and mathematical modeling involving Fourier methods for deconvolution with singular value decomposition. Hence, while these quantitative parameters reflect potentially more meaningful physiological information, they have equally to be treated with care, as their actual values depend highly on the choices made during their calculation.

This holds especially true for the Tmax parameter, which in theory denotes the time between the peak of the AIF and the peak of the residue function. It is hence highly dependent on the chosen location of the AIF and its frequent use in stroke analysis has been criticized.

The CBF is reported in milliliter per minute passing through 100 g of tissue, i.e, in ml/min/100g, the CBV in ml/100g, and MTT, TTP as well as Tmax in s. Typical CBV values for the brain are 60 ml/min/100g for gray and 20 ml/min/100g for white matter.

It should be noted that significantly different results may be obtained depending on the choice of the AIF as well as which commercial software product for calculation is used. Even

software claiming to use identical methods may give different results when applied to the same raw data. Caution is advised in relying too heavily on the absolute numbers obtained from such quantitative methods.

Appendix D

Selected publications resulting from this work

Main publications

A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. H. Sudre, M. J. Cardoso, N. Cawley, O. Ciccarelli, C. A. M. Wheeler-Kingshott, S. Ourselin, L. Catanese, H. Deshpande, P. Maurel, O. Commowick, C. Barillot, X. Tomas-Fernandez, S. K. Warfield, S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, A. Jesson, T. Arbel, O. Maier, H. Handels, L. O. Iheme, D. Unay, S. Jain, D. M. Sima, D. Smeets, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, P.-L. Bazin, P. A. Calabresi, C. M. Crainiceanu, L. M. Ellingsen, D. S. Reich, J. L. Prince, and D. L. Pham. "Longitudinal Multiple Sclerosis Lesion Segmentation: Resource & Challenge". Manuscript submitted for publication. 2016.

A. Crimi, B. Menze, O. Maier, M. Reyes, and H. Handels, eds. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Vol. 9556. LNCS. Cham: Springer International Publishing, 2016, p. 298.

O. Maier, C. Schröder, N. D. Forkert, T. Martinetz, and H. Handels. "Classifiers for Ischemic Stroke Lesion Segmentation: A Comparison Study." *PLOS ONE* 10.12 (Jan. 2015), e0145118.

O. Maier, M. Wilms, J. von der Gablentz, U. M. Krämer, T. F. Münte, H. Handels, U. M. Krämer, T. F. Münte, and H. Handels. "Extra Tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences". *Journal of Neuroscience Methods* 240 (Jan. 2015), pp. 89–100.

O. Maier, B. H. Menze, J. von der Gablentz, L. Häni, M. P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen, D. Christiaens, F. Dutil, K. Egger, C. Feng, B. Glocker, M. Götz, T. Haeck, H.-L. Halme, M. Havaei, K. M. Iftekharuddin, P.-M. Jodoin, K. Kamnitsas, E. Kellner, A. Korvenoja, H. Larochelle, C. Ledig, J.-H. Lee, F. Maes, Q. Mahmood, K. H. Maier-Hein, R. McKinley, J. Muschelli, C. Pal, L. Pei, J. R. Rangarajan, S. M. Reza, D. Robben, D. Rueckert, E. Salli, P. Suetens, C.-W. Wang, M. Wilms, J. S. Kirschke, U. M. Krämer, T. F. Münte, P. Schramm, R. Wiest, H. Handels, and M. Reyes. "ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI". *Medical Image Analysis* 35 (2017), pp. 250–269.

Conference proceedings

O. Maier, M. Wilms, J. von der Gablentz, U. M. Krämer, and H. Handels. "Ischemic stroke lesion segmentation in multi-spectral MR images with support vector machine classifiers". *SPIE Medical Imaging*. Ed. by S. Aylward and L. M. Hadjiiski. Vol. 9035. San Diego: International Society for Optics and Photonics, Mar. 2014, p. 04.

O. Maier, M. Wilms, J. von der Gablentz, U. M. Krämer, and H. Handels. "Segmentierung von ischämischen Schlaganfall-Läsionen in multispektralen MR-Bildern mit Random Decision Forests". *Bildverarbeitung für die Medizin (BVM)*. Ed. by T. M. Deserno, H. Handels, H.-P. Meinzer, and T. Tolxdorff. Informatik aktuell. Springer Berlin Heidelberg, 2014, pp. 156–161.

O. Maier and H. Handels. "Local problem forests: Classifier training for locally limited subproblems using spectral clustering". *International Symposium on Biomedical Imaging (ISBI)*. IEEE, Apr. 2015, pp. 806–809.

O. Maier, M. Wilms, and H. Handels. "Image Features for Brain Lesion Segmentation Using Random Forests". *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Ed. by A. Crimi, O. Maier, B. Menze, M. Reyes, and H. Handels. Vol. 9556. LNCS. Springer International Publishing, 2016, pp. 119–130.

Open source tools

O. Maier. *DynStatCov*. 2016. URL: `https://pypi.python.org/pypi/DynStatCov` (visited on 07/03/2016).

O. Maier. *MedPy - Medical image processing in Python*. 2016. URL: `https://pypi.python.org/pypi/MedPy` (visited on 01/28/2016).

O. Maier. *skleanef*. 2016. URL: `https://github.com/loli/sklearnef` (visited on 07/03/2016).

Awards

First prize in the Ischemic Stroke Lesion Segmentation challenge 2016, Task II: Clinical outcome prediction. Held in conjunction with the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference 2016 in Athens.

Second prize in the Ischemic Stroke Lesion Segmentation challenge 2015, acute stroke penumbra estimation task. Held in conjunction with the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference 2015 in Munich.

Third prize in the Longitudinal MS Lesion Segmentation Challenge 2015. Held in conjunction with the International Symposium on Biomedical Imaging (ISBI) 2015 in New York.

Third prize in the Ischemic Stroke Lesion Segmentation challenge 2016, Task I: Lesion outcome prediction. Held in conjunction with the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference 2016 in Athens.

Efficiency award in the Longitudinal MS Lesion Segmentation Challenge 2015. Held in conjunction with the International Symposium on Biomedical Imaging (ISBI) 2015 in New York.

Bibliography

[Agam *et al.*, 2006] G. Agam, D. Weiss, M. Soman, and K. Arfanakis. "Probabilistic brain lesion segmentation in DT-MRI". *International Conference on Image Processing (ICIP)*. 2006, pp. 89–92.

[Agn *et al.*, 2016] M. Agn, O. Puonti, P. M. af Rosenschöld, I. Law, and K. Van Leemput. "Brain Tumor Segmentation Using a Generative Model with an RBM Prior on Tumor Shape". *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Ed. by A. Crimi, B. Menze, O. Maier, M. Reyes, and H. Handels. Vol. 9556. LNCS. Springer International Publishing, 2016, pp. 168–180.

[Albers *et al.*, 2006] G. W. Albers, V. N. Thijs, L. Wechsler, S. Kemp, G. Schlaug, E. Skalabrin, R. Bammer, W. Kakuda, M. G. Lansberg, A. Shuaib, W. Coplin, S. Hamilton, M. Moseley, M. P. Marks, and DEFUSE Investigators. "Magnetic resonance imaging profiles predict clinical response to early reperfusion: the diffusion and perfusion imaging evaluation for understanding stroke evolution (DEFUSE) study." *Annals of Neurology* 60.5 (Nov. 2006), pp. 508–17.

[Badrinarayanan *et al.*, 2013] V. Badrinarayanan, I. Budvytis, and R. Cipolla. "Semi-supervised video segmentation using tree structured graphical models". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.11 (2013), pp. 2751–2764.

[Baird *et al.*, 1998] A. E. Baird and S. Warach. "Magnetic resonance imaging of acute stroke." *Journal of Cerebral Blood Flow and Metabolism* 18.6 (June 1998), pp. 583–609.

[Bentley, 1975] J. L. Bentley. "Multidimensional binary search trees used for associative searching". *Communications of the ACM* 18.9 (1975), pp. 509–517.

[Berkhemer *et al.*, 2015] O. A. Berkhemer, P. S. S. Fransen, D. Beumer, L. A. van den Berg, H. F. Lingsma, A. J. Yoo, W. J. Schonewille, J. A. Vos, P. J. Nederkoorn, M. J. H. Wermer, M. A. A. van Walderveen, J. Staals, J. Hofmeijer, J. A. van Oostayen, G. J. Lycklama à Nijeholt, J. Boiten, P. A. Brouwer, B. J. Emmer, S. F. de Bruijn, L. C. van Dijk, L. J. Kappelle, R. H. Lo, E. J. van Dijk, J. de Vries, P. L. M. de Kort, W. J. J. van Rooij, J. S. P. van den Berg, B. A. A. M. van Hasselt, L. A. M. Aerden, R. J. Dallinga, M. C. Visser, J. C. J. Bot, P. C. Vroomen, O. Eshghi, T. H. C. M. L. Schreuder, R. J. J. Heijboer, K. Keizer, A. V. Tielbeek, H. M. den Hertog, D. G. Gerrits, R. M. van den Berg-Vos, G. B. Karas, E. W. Steyerberg, H. Z. Flach, H. A. Marquering, M. E. S. Sprengers, S. F. M. Jenniskens, L. F. M. Beenen, R. van den Berg, P. J. Koudstaal, W. H. van Zwam, Y. B. W. E. M. Roos, A. van der Lugt, R. J. van Oostenbrugge, C. B. L. M. Majoie, D. W. J. Dippel, and MR CLEAN Investigators. "A randomized trial of intraarterial treatment for acute ischemic stroke." *The New England Journal of Medicine* 372.1 (Jan. 2015), pp. 11–20.

[Boykov *et al.*, 2001] Y. Boykov, O. Veksler, and R. Zabih. "Fast approximate energy minimization via graph cuts". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.11 (2001), pp. 1222–1239.

[Breiman, 1996] L. Breiman. "Bagging Predictors". *Machine Learning* 24.2 (1996), pp. 123–140.

[Breiman, 2001] L. Breiman. "Random forests". *Machine Learning* 45.1 (2001), pp. 5–32.

[Breiman *et al.*, 1984] L. Breiman, J. H. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984, p. 368.

[Brex *et al.*, 2002] P. A. Brex, O. Ciccarelli, J. I. O'Riordan, M. Sailer, A. J. Thompson, and D. H. Miller. "A longitudinal study of abnormalities on MRI and disability from multiple sclerosis." *The New England journal of medicine* 346.3 (Jan. 2002), pp. 158–64.

[BrønnumHansen *et al.*, 2004] H. Brønnum-Hansen, N. Koch-Henriksen, and E. Stenager. "Trends in survival and cause of death in Danish patients with multiple sclerosis." *Brain* 127.4 (Apr. 2004), pp. 844–50.

[Caicedo *et al.*, 2007] J. C. Caicedo, F. A. Gonzalez, and E. Romero. "Content-Based Medical Image Retrieval Using Low-Level Visual Features and Modality Identification". *Advances in Multilingual and Multimodal Information Retrieval*. Vol. 5152. LNCS. Springer Berlin Heidelberg, 2007, pp. 615–622.

[Calamante *et al.*, 2010] F. Calamante, S. Christensen, P. M. Desmond, L. Østergaard, S. M. Davis, and A. Connelly. "The physiological significance of the time-to-maximum (Tmax) parameter in perfusion MRI". *Stroke* 41.6 (June 2010), pp. 1169–1174.

[Carass *et al.*, 2016] A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. H. Sudre, M. J. Cardoso, N. Cawley, O. Ciccarelli, C. A. M. Wheeler-Kingshott, S. Ourselin, L. Catanese, H. Deshpande, P. Maurel, O. Commowick, C. Barillot, X. Tomas-Fernandez, S. K. Warfield, S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, A. Jesson, T. Arbel, O. Maier, H. Handels, L. O. Iheme, D. Unay, S. Jain, D. M. Sima, D. Smeets, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, P.-L. Bazin, P. A. Calabresi, C. M. Crainiceanu, L. M. Ellingsen, D. S. Reich, J. L. Prince, and D. L. Pham. "Longitudinal Multiple Sclerosis Lesion Segmentation: Resource & Challenge". Manuscript submitted for publication. 2016.

[Catanese *et al.*, 2015] L. Catanese, O. Commowick, and C. Barillot. "Automatic Graph Cut Segmentation of Multiple Sclerosis Lesions". *Longitudinal Multiple Sclerosis Lesion Segmentation Challenge, ISBI*. Ed. by D. L. Pham. online, 2015. URL: `http://iacl.ece.jhu.edu/index.php/MSChallenge`.

[Cha *et al.*, 2002] S.-H. Cha and S. N. Srihari. "On measuring the distance between histograms". *Pattern Recognition* 35.6 (2002), pp. 1355–1370.

[Chandna *et al.*, 2010] P. Chandna, S. Deswal, and M. Pal. "Semi-supervised learning based prediction of musculoskeletal disorder risk". *Journal of Industrial and Systems Engineering* 3.4 (2010), pp. 291–295.

[Chapelle *et al.*, 2010] O. Chapelle, B. Scholkopf, and A. Zien, eds. *Semi-supervised learning*. Cambridge, Mass., United States: MIT Press, 2010, p. 528.

[Ciresan *et al.*, 2013] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. "Mitosis Detection in Breast Cancer Histology Images using Deep Neural Networks". *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 8150. LNCS. Springer Berlin Heidelberg, 2013, pp. 411–418.

[Clark *et al.*, 1999] W. M. Clark, S. Wissman, G. W. Albers, J. H. Jhamandas, K. P. Madden, and S. Hamilton. "Recombinant tissue-type plasminogen activator (Alteplase) for ischemic stroke 3 to 5 hours after symptom onset. The ATLANTIS Study: a randomized controlled trial. Alteplase Thrombolysis for Acute Noninterventional Therapy in Ischemic Stroke." *Journal of the American Medical Association* 282.21 (Dec. 1999), pp. 2019–26.

[Cohen, 2001] J. A. Cohen. "Use of the Multiple Sclerosis Functional Composite as an Outcome Measure in a Phase 3 Clinical Trial". *Archives of Neurology* 58.6 (June 2001), p. 961.

[Cohen *et al.*, 2000] J. A. Cohen, J. S. Fischer, D. M. Bolibrush, A. J. Jak, J. E. Kniker, L. A. Mertz, T. T. Skaramagas, and G. R. Cutter. "Intrarater and interrater reliability of the MS functional composite outcome measure." *Neurology* 54.4 (Feb. 2000), pp. 802–6.

[Collins *et al.*, 1994] D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans. "Automatic 3D Intersubject Registration of MR Volumetric Data in Standardized Talairach Space". *Journal of Computer Assisted Tomography* 18.2 (Mar. 1994), pp. 192–205.

[Conrad, 2013] K. Conrad. *Probability distributions and maximum entropy.* Tech. rep. 2013, online.

[Cover *et al.*, 1967] T. M. Cover and P. E. Hart. "Nearest Neighbor Pattern Classification". *IEEE Transactions on Information Theory* 13.1 (1967), pp. 21–27.

[Crimi *et al.*, 2016] A. Crimi, B. Menze, O. Maier, M. Reyes, and H. Handels, eds. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries.* Vol. 9556. LNCS. Cham: Springer International Publishing, 2016, p. 298.

[Criminisi *et al.*, 2011] A. Criminisi, J. Shotton, and E. Konukoglu. "Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning". *Foundations and Trends® in Computer Graphics and Vision* 7.2-3 (2011), pp. 81–227.

[Criminisi *et al.*, 2013] A. Criminisi and J. Shotton. *Decision Forests for Computer Vision and Medical Image Analysis.* Ed. by A. Criminisi and J. Shotton. 1st ed. Springer London, 2013, p. 368.

[Davis *et al.*, 2008] M. Davis, S. Hanson, and B. Altevogt, eds. *Neuroscience Biomarkers and Biosignatures: Converging Technologies, Emerging Partnerships, Workshop Summary.* National Academies Press (US), 2008. URL: `http://www.ncbi.nlm.nih.gov/books/NBK53112/`.

[Debouverie, 2009] M. Debouverie. "Gender as a prognostic factor and its impact on the incidence of multiple sclerosis in Lorraine, France." *Journal of the Neurological Sciences* 286.1-2 (Nov. 2009), pp. 14–7.

[Derntl *et al.*, 2016] A. Derntl, C. Plant, P. Gruber, S. Wegener, J. S. Bauer, and B. H. Menze. "Stroke Lesion Segmentation Using a Probabilistic Atlas of Cerebral Vascular Territories". *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries.* Ed. by A. Crimi, O. Maier, B. Menze, M. Reyes, and H. Handels. Vol. 9556. LNCS. Springer International Publishing, 2016, pp. 21–32.

[Deshpande *et al.*, 2015] H. Deshpande, P. Maurel, and C. Barillot. "Adaptive dictionary learning for competitive classification of multiple sclerosis lesions". *International Symposium on Biomedical Imaging (ISBI).* Apr. 2015, pp. 136–139.

[Didaci *et al.*, 2012] L. Didaci, G. Fumera, and F. Roli. "Analysis of Co-training Algorithm with Very Small Training Sets". *Structural, Syntactic, and Statistical Pattern Recognition.* Ed. by G. Gimel'farb, E. Hancock, A. Imiya, A. Kuijper, M. Kudo, S. Omachi, T. Windeatt, and K. Yamada. Vol. 7626. LNCS. Springer Berlin Heidelberg, 2012, pp. 719–726.

[Dietrich *et al.*, 2008] O. Dietrich, M. F. Reiser, and S. O. Schoenberg. "Artifacts in 3-T MRI: physical background and reduction strategies." *European Journal of Radiology* 65.1 (Jan. 2008), pp. 29–35.

[Driessens *et al.*, 2006] K. Driessens, P. Reutemann, B. Pfahringer, and C. Leschi. "Using Weighted Nearest Neighbor to Benefit from Unlabeled Data". *Advances in Knowledge Discovery and Data Mining*. Ed. by W.-K. Ng, M. Kitsuregawa, J. Li, and K. Chang. Vol. 3918. LNCS. Springer Berlin Heidelberg, 2006, pp. 60–69.

[Dutil *et al.*, 2016] F. Dutil, M. Havaei, C. Pal, H. Larochelle, and P.-M. Jodoin. "A Convolutional Neural Network Approach to Brain Lesion Segmentation". *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Ed. by A. Crimi, B. Menze, O. Maier, M. Reyes, and H. Handels. Vol. 9556. LNCS. Springer International Publishing, 2016, pp. 53–58.

[Dvorak *et al.*, 2015] P. Dvorak and B. H. Menze. "Structured Prediction with Convolutional Neural Networks for Multimodal Brain Tumor Segmentation". *Multimodal Brain Tumor Image Segmentation (BRATS) Challenge, MICCAI*. 2015.

[Feng *et al.*, 2016] C. Feng, D. Zhao, and M. Huang. "Segmentation of ischemic stroke lesions in multi-spectral MR images using weighting suppressed FCM and three phase level set". *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Ed. by A. Crimi, B. Menze, O. Maier, M. Reyes, and H. Handels. Vol. 9556. LNCS. Springer International Publishing, 2016, pp. 233–245.

[Filippi *et al.*, 2011] M. Filippi and M. A. Rocca. "MR imaging of multiple sclerosis." *Radiology* 259.3 (June 2011), pp. 659–81.

[Fisniku *et al.*, 2008] L. K. Fisniku, P. A. Brex, D. R. Altmann, K. A. Miszkiel, C. E. Benton, R. Lanyon, A. J. Thompson, and D. H. Miller. "Disability and T2 MRI lesions: a 20-year follow-up of patients with relapse onset of multiple sclerosis." *Brain* 131.3 (Mar. 2008), pp. 808–17.

[Fonov *et al.*, 2009] V. Fonov, A. Evans, R. McKinstry, C. Almli, and D. Collins. "Unbiased nonlinear average age-appropriate brain templates from birth to adulthood". *NeuroImage* 47 (2009), S102.

[Forbes *et al.*, 2010] F. Forbes, S. Doyle, D. Garcia-Lorenzo, C. Barillot, and M. Dojat. "Adaptive weighted fusion of multiple MR sequences for brain lesion segmentation". *International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2010, pp. 69–72.

[Freund *et al.*, 1997] Y. Freund and R. E. Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". *Journal of Computer and System Sciences* 55.1 (Aug. 1997), pp. 119–139.

[Friedman, 2000] J. H. Friedman. "Greedy Function Approximation: a Gradient Boosting Machine". *The Annals of Statistics* 29.5 (2000), pp. 1189–1232.

[Friedman, 2002] J. H. Friedman. "Stochastic gradient boosting". *Computational Statistics & Data Analysis* 38.4 (Feb. 2002), pp. 367–378.

[Fukushima, 1980] K. Fukushima. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". *Biological Cybernetics* 36.4 (1980), pp. 193–202.

[Gablentz, 2012] J. von der Gablentz. "Visuelle Exploration von dynamischen Szenen bei Schlaganfallpatienten mit Neglect-Syndrom". PhD thesis. Klink für Neurologie, Universität zu Lübeck, Lübeck, 2012, p. 96.

[Ge, 2006] Y. Ge. "Multiple Sclerosis: The Role of MR Imaging". *American Journal of Neuroradiology* 27.6 (June 2006), pp. 1165–1176.

[Geurts *et al.*, 2006] P. Geurts, D. Ernst, and L. Wehenkel. "Extremely randomized trees". *Machine Learning* 63.1 (Apr. 2006), pp. 3–42.

[Goetz *et al.*, 2016] M. Goetz, C. Weber, C. Kolb, and K. Maier-Hein. "Input Data Adaptive Learning (IDAL) for Sub-acute Ischemic Stroke Lesion Segmentation". *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries.* Ed. by A. Crimi, B. Menze, O. Maier, M. Reyes, and H. Handels. Vol. 9556. LNCS. Springer International Publishing, 2016, pp. 284–295.

[Goldman, 1953] S. Goldman. *Information theory.* Prentice-Hall, 1953, p. 385.

[Gonzalez *et al.*, 2006] R. G. Gonzalez, J. A. Hirsch, W. J. Koroshetz, M. H. Lev, and P. W. Schaefer, eds. *Acute Ischemic Stroke - Imaging and Intervention.* 2nd ed. Springer Berlin Heidelberg, 2006.

[Goodfellow *et al.*, 2013] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet. "Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks" (Dec. 2013). arXiv: `1312.6082`.

[Goyal *et al.*, 2015] M. Goyal, A. M. Demchuk, B. K. Menon, M. Eesa, J. L. Rempel, J. Thornton, D. Roy, T. G. Jovin, R. A. Willinsky, B. L. Sapkota, D. Dowlatshahi, D. F. Frei, N. R. Kamal, W. J. Montanera, A. Y. Poppe, K. J. Ryckborst, F. L. Silver, A. Shuaib, D. Tampieri, D. Williams, O. Y. Bang, B. W. Baxter, P. A. Burns, H. Choe, J.-H. Heo, C. A. Holmstedt, B. Jankowitz, M. Kelly, G. Linares, J. L. Mandzia, J. Shankar, S.-I. Sohn, R. H. Swartz, P. A. Barber, S. B. Coutts, E. E. Smith, W. F. Morrish, A. Weill, S. Subramaniam, A. P. Mitha, J. H. Wong, M. W. Lowerison, T. T. Sajobi, and M. D. Hill. "Randomized assessment of rapid endovascular treatment of ischemic stroke." *The New England Journal of Medicine* 372.11 (Mar. 2015), pp. 1019–30.

[Grady, 2006] L. Grady. "Random Walks for Image Segmentation". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.11 (Nov. 2006), pp. 1768–1783.

[Gray *et al.*, 2013] K. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, and D. Rueckert. "Manifold forests for multi-modality classification of Alzheimer's disease". *Decision Forests for Computer Vision and Medical Image Analysis.* Springer London, 2013, pp. 261–272.

[Grimaud *et al.*, 1996] J. Grimaud, M. Lai, J. Thorpe, P. Adeleine, L. Wang, G. J. Barker, D. L. Plummer, P. S. Tofts, W. I. McDonald, and D. H. Miller. "Quantification of MRI lesion load in multiple sclerosis: A comparison of three computer-assisted techniques". *Magnetic Resonance Imaging* 14.5 (Jan. 1996), pp. 495–505.

[Haeck *et al.*, 2016] T. Haeck, F. Maes, and P. Suetens. "ISLES Challenge 2015: Automated Model-Based Segmentation of Ischemic Stroke in MR Images". *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries.* Ed. by A. Crimi, B. Menze, O. Maier, M. Reyes, and H. Handels. Vol. 9556. LNCS. Cham: Springer International Publishing, 2016, pp. 246–253.

[Hakim, 1998] A. M. Hakim. "Ischemic penumbra: the therapeutic window." *Neurology* 51.Suppl 3 (Sept. 1998), S44–6.

[Halme *et al.*, 2016] H. L. Halme, A. Korvenoja, and E. Salli. "ISLES (SISS) challenge 2015: Segmentation of stroke lesions using spatial normalization, random forest classification and contextual clustering". *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries.* Ed. by A. Crimi, B. Menze, O. Maier, M. Reyes, and H. Handels. Vol. 9556. LNCS. Springer International Publishing, 2016, pp. 211–221.

[Hinton *et al.*, 2006a] G. E. Hinton and R. R. Salakhutdinov. "Reducing the Dimensionality of Data with Neural Networks". *Science* 313.5786 (2006).

[Hinton *et al.*, 2006b] G. E. Hinton, S. Osindero, and Y.-W. Teh. "A Fast Learning Algorithm for Deep Belief Nets". *Neural Computation* 18.7 (July 2006), pp. 1527–1554.

[Hinton *et al.*, 2012] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. "Improving neural networks by preventing co-adaptation of feature detectors" (July 2012). arXiv: `1207.0580`.

[Ho, 1998] T. K. Ho. "The random subspace method for constructing decision forests". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.8 (Aug. 1998), pp. 832–844.

[Hobart, 2000] J. Hobart. "Kurtzke scales revisited: the application of psychometric methods to clinical intuition". *Brain* 123.5 (May 2000), pp. 1027–1040.

[Hoogi *et al.*, 2015] A. Hoogi, A. Lee, V. Bharadwaj, and D. L. Rubin. "Multimodal Brain Tumor Segmentation (BRATS) using Sparse Coding and 2-layer Neural Network". *Multimodal Brain Tumor Image Segmentation (BRATS) Challenge, MICCAI.* 2015.

[Horowitz *et al.*, 1989] A. L. Horowitz, R. D. Kaplan, G. Grewe, R. T. White, and L. M. Salberg. "The ovoid lesion: a new MR observation in patients with multiple sclerosis." *American Journal of Neuroradiology* 10.2 (Mar. 1989), pp. 303–305.

[Ioffe *et al.*, 2015] S. Ioffe and C. Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift" (Feb. 2015). arXiv: `1502.03167`.

[Jaynes, 1957] E. T. Jaynes. "Information theory and statistical mechanics". *Physical review* 106.4 (1957), p. 620.

[Jenkinson *et al.*, 2005] M. Jenkinson, M. Pechaud, and S. Smith. "BET2: MR-Based Estimation of Brain, Skull and Scalp Surfaces". *Organization for Human Brain Mapping (OHBM).* Vol. 17. 3. 2005.

[Jesson *et al.*, 2015] A. Jesson and T. Arbel. "Hierarchical MRF and Random Forest Segmentation of MS Lesions and Healthy Tissues in Brain MRI". *Longitudinal Multiple Sclerosis Lesion Segmentation Challenge, ISBI.* Ed. by D. L. Pham. online, 2015. URL: `http://iacl.ece.jhu.edu/index.php/MSChallenge`.

[Jia *et al.*, 2014] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. "Caffe: Convolutional Architecture for Fast Feature Embedding". *International Conference on Multimedia (MM).* ACM Press, 2014, pp. 675–678. arXiv: `1408.5093`.

[Joachims, 1999] T. Joachims. "Transductive inference for text classification using support vector machines". *International Conference on Machine Learning (ICML).* Vol. 99. 1999, pp. 200–209.

[Joy *et al.*, 2001] J. E. Joy and R. B. Johnston Jr., eds. *Multiple Sclerosis: Current status and strategies for the future.* Washington (DC): National Academies Press (US), 2001, p. 457.

[Kalkers *et al.*, 2000] N. F. Kalkers, V. de Groot, R. H. Lazeron, J. Killestein, H. J. Adèr, F. Barkhof, G. J. Lankhorst, and C. H. Polman. "MS functional composite: relation to disease phenotype and disability strata." *Neurology* 54.6 (Mar. 2000), pp. 1233–9.

[Kamnitsas *et al.*, 2016] 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" (Mar. 2016). arXiv: `1603.05959`.

[Kidwell *et al.*, 2003] C. S. Kidwell, J. R. Alger, and J. L. Saver. "Beyond mismatch: evolving paradigms in imaging the ischemic penumbra with multimodal magnetic resonance imaging." *Stroke* 34.11 (Nov. 2003), pp. 2729–35.

[Klein *et al.*, 2009] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, et al. "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration". *NeuroImage* 46.3 (2009), pp. 786–802.

[Klein *et al.*, 2010] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim. "elastix: a toolbox for intensity-based medical image registration." *IEEE Transactions on Medical Imaging* 29.1 (Jan. 2010), pp. 196–205.

[Kobelt *et al.*, 2006] G. Kobelt, J. Berg, D. Atherly, and O. Hadjimichael. "Costs and quality of life in multiple sclerosis: a cross-sectional study in the United States." *Neurology* 66.11 (June 2006), pp. 1696–702.

[Krieger *et al.*, 1999] D. W. Krieger, A. M. Demchuk, S. E. Kasner, M. Jauss, and L. Hantson. "Early Clinical and Radiological Predictors of Fatal Brain Swelling in Ischemic Stroke". *Stroke* 30.2 (Feb. 1999), pp. 287–292.

[Krizhevsky *et al.*, 2012] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., 2012, pp. 1097–1105.

[Kullback *et al.*, 1951] S. Kullback and R. A. Leibler. "On information and sufficiency". *Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86.

[Kurtzke, 1983] J. F. Kurtzke. "Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS)." *Neurology* 33.11 (Nov. 1983), pp. 1444–52.

[Kurtzke *et al.*, 1992] J. F. Kurtzke, W. F. Page, F. M. Murphy, and J. E. Norman. "Epidemiology of multiple sclerosis in US veterans. 4. Age at onset." *Neuroepidemiology* 11.4-6 (Jan. 1992), pp. 226–35.

[Lassmann *et al.*, 2007] H. Lassmann, W. Brück, and C. F. Lucchinetti. "The immunopathology of multiple sclerosis: an overview." *Brain Pathology* 17.2 (Apr. 2007), pp. 210–8.

[LeCun *et al.*, 1989] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. "Backpropagation Applied to Handwritten Zip Code Recognition". *Neural Computation* 1.4 (1989), pp. 541–551.

[Lee *et al.*, 2005] C.-H. Lee, M. Schmidt, A. Murtha, A. Bistritz, J. Sander, and R. Greiner. "Segmenting brain tumor with conditional random fields and support vector machines". *Computer Vision for Biomedical Image Applications*. Vol. 3765. LNCS. Springer Berlin Heidelberg, 2005, pp. 469–478.

[Leistner *et al.*, 2009] C. Leistner, A. Saffari, J. Santner, and H. Bischof. "Semi-Supervised Random Forests". *International Conference on Computer Vision (ICCV)*. IEEE. 2009, pp. 506–513.

[Liguori *et al.*, 2000] M. Liguori, M. G. Marrosu, M. Pugliatti, F. Giuliani, F. De Robertis, E. Cocco, G. B. Zimatore, P. Livrea, and M. Trojano. "Age at onset in multiple sclerosis." *Neurological sciences* 21.4 Suppl 2 (Jan. 2000), S825–9.

[Likar *et al.*, 2001] B. Likar, M. A. Viergever, and F. Pernuš. "Retrospective correction of MR intensity inhomogeneity by information minimization". *IEEE Transactions on Medical Imaging* 20.12 (2001), pp. 1398–1410.

[Liu *et al.*, 2007] Y. Liu, H. E. D'Arceuil, S. Westmoreland, J. He, M. Duggan, R. G. Gonzalez, J. Pryor, and A. J. De Crespigny. "Serial diffusion tensor MRI after transient and permanent cerebral ischemia in nonhuman primates". *Stroke* 38.1 (Jan. 2007), pp. 138–145.

[Liu *et al.*, 2013] X. Liu, M. Song, D. Tao, Z. Liu, L. Zhang, C. Chen, and J. Bu. "Semi-supervised node splitting for random forest construction". *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013, pp. 492–499.

[Lombaert *et al.*, 2014] H. Lombaert, D. Zikic, A. Criminisi, and N. Ayache. "Laplacian Forests: Semantic Image Segmentation by Guided Bagging". *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 8674. LNCS. Springer Berlin Heidelberg, 2014, pp. 496–504.

[Lublin *et al.*, 1996] F. D. Lublin and S. C. Reingold. "Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis." *Neurology* 46.4 (Apr. 1996), pp. 907–11.

[Lublin *et al.*, 2014] F. D. Lublin, S. C. Reingold, J. A. Cohen, G. R. Cutter, P. S. Sørensen, A. J. Thompson, J. S. Wolinsky, L. J. Balcer, B. Banwell, F. Barkhof, B. Bebo, P. A. Calabresi, M. Clanet, G. Comi, R. J. Fox, M. S. Freedman, A. D. Goodman, M. Inglese, L. Kappos, B. C. Kieseier, J. A. Lincoln, C. Lubetzki, A. E. Miller, X. Montalban, P. W. O'Connor, J. Petkau, C. Pozzilli, R. A. Rudick, M. P. Sormani, O. Stüve, E. Waubant, and C. H. Polman. "Defining the clinical course of multiple sclerosis: The 2013 revisions". *Neurology* 83.3 (July 2014), pp. 278–286.

[Machner *et al.*, 2012] B. Machner, M. Dorr, A. Sprenger, J. von der Gablentz, W. Heide, E. Barth, and C. Helmchen. "Impact of dynamic bottom-up features and top-down control on the visual exploration of moving real-world scenes in hemispatial neglect". *Neuropsychologia* 50.10 (2012), pp. 2415–2425.

[Machner *et al.*, 2014] B. Machner, I. Könemund, A. Sprenger, J. von der Gablentz, and C. Helmchen. "Randomized Controlled Trial on Hemifield Eye Patching and Optokinetic Stimulation in Acute Spatial Neglect". *Stroke* 45.8 (2014), pp. 2465–2468.

[Mackay *et al.*, 2004] J. Mackay and G. A. Mensah. *The atlas of heart disease and stroke*. 1st ed. Geneva: World Health Organization, 2004, p. 112.

[Maier, 2016a] O. Maier. *DynStatCov*. 2016. URL: `https://pypi.python.org/pypi/DynStatCov` (visited on 07/03/2016).

[Maier, 2016b] O. Maier. *MedPy - Medical image processing in Python*. 2016. URL: `https://pypi.python.org/pypi/MedPy` (visited on 01/28/2016).

[Maier, 2016c] O. Maier. *skleanef*. 2016. URL: `https://github.com/loli/sklearnef` (visited on 07/03/2016).

[Maier *et al.*, 2014a] O. Maier, M. Wilms, J. von der Gablentz, U. M. Krämer, and H. Handels. "Ischemic stroke lesion segmentation in multi-spectral MR images with support vector machine classifiers". *SPIE Medical Imaging*. Ed. by S. Aylward and L. M. Hadjiiski. Vol. 9035. San Diego: International Society for Optics and Photonics, Mar. 2014, p. 04.

[Maier *et al.*, 2014b] O. Maier, M. Wilms, J. von der Gablentz, U. M. Krämer, and H. Handels. "Segmentierung von ischämischen Schlaganfall-Läsionen in multispektralen MR-Bildern mit Random Decision Forests". *Bildverarbeitung für die Medizin (BVM)*. Ed. by T. M. Deserno, H. Handels, H.-P. Meinzer, and T. Tolxdorff. Informatik aktuell. Springer Berlin Heidelberg, 2014, pp. 156–161.

[Maier *et al.*, 2015a] O. Maier and H. Handels. "Local problem forests: Classifier training for locally limited sub-problems using spectral clustering". *International Symposium on Biomedical Imaging (ISBI)*. IEEE, Apr. 2015, pp. 806–809.

[Maier *et al.*, 2015b] O. Maier and H. Handels. "MS-Lesion Segmentation in MRI with Random Forests". *Longitudinal Multiple Sclerosis Lesion Segmentation Challenge, ISBI*. Ed. by D. Pham. online, 2015. URL: `http://iacl.ece.jhu.edu/MSChallenge`.

[Maier *et al.*, 2015c] O. Maier, M. Reyes, and B. H. Menze, eds. *Ischemic Stroke Lesion Segmentation (ISLES) Challenge, MICCAI*. online, 2015. URL: `http://www.isles-challenge.org/ISLES2015/`.

[Maier *et al.*, 2015d] O. Maier, C. Schröder, N. D. Forkert, T. Martinetz, and H. Handels. "Classifiers for Ischemic Stroke Lesion Segmentation: A Comparison Study." *PLOS ONE* 10.12 (Jan. 2015), e0145118.

[Maier *et al.*, 2015e] O. Maier, M. Wilms, J. von der Gablentz, U. M. Krämer, T. F. Münte, H. Handels, U. M. Krämer, T. F. Münte, and H. Handels. "Extra Tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences". *Journal of Neuroscience Methods* 240 (Jan. 2015), pp. 89–100.

[Maier *et al.*, 2015f] O. Maier, M. Wilms, and H. Handels. "Highly discriminative features for Glioma Segmentation in MR Volumes with Random Forests". *Multimodal Brain Tumor Image Segmentation (BRATS) Challenge, MICCAI*. Ed. by B. H. Menze, M. Reyes, K. Farahani, J. Kalpathy-Cramer, and D. Kwon. Munich, 2015, p. 38.

[Maier *et al.*, 2015g] O. Maier, M. Wilms, and H. Handels. "Random Forests for Acute Stroke Penumbra Estimation". *Ischemic Stroke Lesion Segmentation (ISLES) Challenge, MICCAI*. Ed. by O. Maier, M. Reyes, and B. H. Menze. Lübeck, 2015, p. 77.

[Maier *et al.*, 2015h] O. Maier, M. Wilms, and H. Handels. "Random Forests with Selected Features for Stroke Lesion Segmentation". *Ischemic Stroke Lesion Segmentation (ISLES) Challenge, MICCAI*. Ed. by O. Maier, M. Reyes, and B. H. Menze. Lübeck, 2015, p. 17.

[Maier *et al.*, 2016] O. Maier, M. Wilms, and H. Handels. "Image Features for Brain Lesion Segmentation Using Random Forests". *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Ed. by A. Crimi, O. Maier, B. Menze, M. Reyes, and H. Handels. Vol. 9556. LNCS. Springer International Publishing, 2016, pp. 119–130.

[Maier *et al.*, 2017] O. Maier, B. H. Menze, J. von der Gablentz, L. Häni, M. P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen, D. Christiaens, F. Dutil, K. Egger, C. Feng, B. Glocker, M. Götz, T. Haeck, H.-L. Halme, M. Havaei, K. M. Iftekharuddin, P.-M. Jodoin, K. Kamnitsas, E. Kellner, A. Korvenoja, H. Larochelle, C. Ledig, J.-H. Lee, F. Maes, Q. Mahmood, K. H. Maier-Hein, R. McKinley, J. Muschelli, C. Pal, L. Pei, J. R. Rangarajan, S. M. Reza, D. Robben, D. Rueckert, E. Salli, P. Suetens, C.-W. Wang, M. Wilms, J. S. Kirschke, U. M. Krämer, T. F. Münte, P. Schramm, R. Wiest, H. Handels, and M. Reyes. "ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI". *Medical Image Analysis* 35 (2017), pp. 250–269.

[Malmi *et al.*, 2015] E. Malmi, S. Parambath, J.-M. Peyrat, J. Abinahed, and S. Chawla. "CaBS: A Cascaded Brain Tumor Segmentation Approach". *Multimodal Brain Tumor Image Segmentation (BRATS) Challenge, MICCAI*. 2015.

[Marsh, 2013] C. Marsh. *Introduction to continuous entropy*. Tech. rep. 2013, online.

[Martel *et al.*, 1999] A. L. Martel, S. J. Allder, G. S. Delay, P. S. Morgan, and A. R. Moody. "Measurement of Infarct Volume in Stroke Patients Using Adaptive Segmentation of Diffusion Weighted MR Images". *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 1679. LNCS. Springer Berlin Heidelberg, 1999, pp. 22–31.

[McDonald *et al.*, 2001] W. I. McDonald, D. A. S. Compston, G. Edan, D. Goodkin, H.-P. Hartung, F. D. Lublin, H. F. McFarland, D. W. Paty, C. H. Polman, S. C. Reingold, M. Sandberg-Wollheim, W. Sibley, A. J. Thompson, S. Van Den Noort, B. G. Weinshenker, and J. S. Wolinsky. "Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis". *Annals of Neurology* 50.1 (July 2001), pp. 121–127.

[McEliece, 1977] R. J. McEliece. *The theory of information and coding: a mathematical framework for communication*. Addison-Wesley Pub. Co., 1977, p. 302.

[McFarland *et al.*, 2002] H. F. McFarland, F. Barkhof, J. Antel, and D. H. Miller. "The role of MRI as a surrogate outcome measure in multiple sclerosis." *Multiple Sclerosis* 8.1 (Feb. 2002), pp. 40–51.

[McKinley *et al.*, 2016] R. McKinley, L. Häni, R. Wiest, and M. Reyes. "Segmenting the Ischemic Penumbra: A Decision Forest Approach with Automatic Threshold Finding". *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Ed. by A. Crimi, B. Menze, O. Maier, M. Reyes, and H. Handels. Vol. 9556. LNCS. Springer International Publishing, 2016, pp. 275–283.

[Meier *et al.*, 2016] R. Meier, V. Karamitsou, S. Habegger, R. Wiest, and M. Reyes. "Parameter learning for CRF-based tissue segmentation of brain tumors". *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Ed. by A. Crimi, B. Menze, O. Maier, M. Reyes, and H. Handels. Vol. 9556. LNCS. Springer International Publishing, 2016, pp. 156–167.

[Menze *et al.*, 2010] B. H. Menze, K. van Leemput, D. Lashkari, M.-A. Weber, N. Ayache, and P. Golland. "A Generative Model for Brain Tumor Segmentation in Multi-Modal Images". *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Vol. 6362. LNCS. Springer Berlin Heidelberg, 2010, pp. 151–159.

[Menze *et al.*, 2015a] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M. A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, C. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H. C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)". *IEEE Transactions on Medical Imaging* 34.10 (Oct. 2015), pp. 1993–2024.

[Menze *et al.*, 2015b] B. H. Menze, M. Reyes, K. Farahani, J. Kalpathy-Cramer, and D. Kwon, eds. *Multimodal Brain Tumor Image Segmentation (BRATS) Challenge, MICCAI*. online, 2015. URL: `http://people.csail.mit.edu/menze/papers/proceedings%7B%5C_%7Dmiccai%7B%5C_%7Dbrats%7B%5C_%7D2015.pdf`.

[Mitra *et al.*, 2014] J. Mitra, P. Bourgeat, J. Fripp, S. Ghose, S. Rose, O. Salvado, A. Connelly, B. Campbell, S. Palmer, G. Sharma, S. Christensen, and L. Carey. "Lesion segmentation from multimodal MRI using random forest following ischemic stroke." *NeuroImage* 98 (Sept. 2014), pp. 324–35.

[Moosmann *et al.*, 2008] F. Moosmann, E. Nowak, and F. Jurie. "Randomized clustering forests for image classification." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.9 (Sept. 2008), pp. 1632–46.

[Murphy, 2011] K. Murphy. "Development and evaluation of automated image analysis techniques in thoracic CT". PhD thesis. Utrecht University, May 2011, p. 203.

[Murphy *et al.*, 2011] K. Murphy, B. Van Ginneken, J. M. Reinhardt, S. Kabus, K. Ding, X. Deng, K. Cao, K. Du, G. E. Christensen, V. Garcia, T. Vercauteren, N. Ayache, O. Commowick, G. Malandain, B. Glocker, N. Paragios, N. Navab, V. Gorbunova, J. Sporring, M. De Bruijne, X. Han, M. P. Heinrich, J. A. Schnabel, M. Jenkinson, C. Lorenz, M. Modat, J. R. McClelland, S. Ourselin, S. E. A. Muenzing, M. A. Viergever, D. De Nigris, D. L. Collins, T. Arbel, M. Peroni, R. Li, G. C. Sharp, A. Schmidt-Richberg, J. Ehrhardt, R. Werner, D. Smeets, D. Loeckx, G. Song, N. Tustison, B. Avants, J. C. Gee, M. Staring, S. Klein, B. C. Stoel, M. Urschler, M. Werlberger, J. Vandemeulebroucke, S. Rit, D. Sarrut, and J. P. W. Pluim. "Evaluation of registration methods on thoracic CT: The EMPIRE10 challenge". *IEEE Transactions on Medical Imaging* 30.11 (Nov. 2011), pp. 1901–1920.

[Nair *et al.*, 2010] V. Nair and G. E. Hinton. "Rectified Linear Units Improve Restricted Boltzmann Machines". *International Conference on Machine Learning (ICML)*. 3. 2010, pp. 807–814.

[Neumann *et al.*, 2009] A. B. Neumann, K. Y. Jonsdottir, K. Mouridsen, N. Hjort, C. Gyldensted, A. Bizzi, J. Fiehler, R. Gasparotti, J. H. Gillard, M. Hermier, T. Kucinski, E.-M. Larsson, L. Sørensen, and L. Ostergaard. "Interrater agreement for final infarct MRI lesion delineation." *Stroke* 40.12 (Dec. 2009), pp. 3768–71.

[Nyúl *et al.*, 2000] L. G. Nyúl, J. K. Udupa, and X. Zhang. "New variants of a method of MRI scale standardization." *IEEE Transactions on Medical Imaging* 19.2 (Feb. 2000), pp. 143–50.

[Olek, 2015a] M. J. Olek. *Symptom management of multiple sclerosis in adults*. 2015. URL: `http://www.uptodate.com/contents/symptom-management-of-multiple-sclerosis-in-adults` (visited on 07/24/2015).

[Olek, 2015b] M. J. Olek. *Treatment of relapsing-remitting multiple sclerosis in adults*. 2015. URL: `http://www.uptodate.com/contents/treatment-of-relapsing-remitting-multiple-sclerosis-in-adults` (visited on 07/24/2015).

[Pedregosa *et al.*, 2011] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830. arXiv: `arXiv:1201.0490v2`.

[Pereira *et al.*, 2016] S. Pereira, A. Pinto, V. Alves, and C. A. Silva. "Deep Convolutional Neural Networks for the Segmentation of Gliomas in Multi-sequence MRI". *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Ed. by A. Crimi, B. Menze, O. Maier, M. Reyes, and H. Handels. Vol. 9556. LNCS. Springer International Publishing, 2016, pp. 131–143.

[Pham, 2015] D. L. Pham, ed. *Proceedings of the Longitudinal Multiple Sclerosis Lesion Segmentation Challenge held in conjunction with ISBI 2015.* online, 2015. URL: `http://iacl.ece.jhu.edu/MSChallenge`.

[Polman *et al.*, 2005] C. H. Polman, S. C. Reingold, G. Edan, M. Filippi, H.-P. Hartung, L. Kappos, F. D. Lublin, L. M. Metz, H. F. McFarland, P. W. O'Connor, M. Sandberg-Wollheim, A. J. Thompson, B. G. Weinshenker, and J. S. Wolinsky. "Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria"." *Annals of Neurology* 58.6 (Dec. 2005), pp. 840–6.

[Polman *et al.*, 2011] C. H. Polman, S. C. Reingold, B. Banwell, M. Clanet, J. A. Cohen, M. Filippi, K. Fujihara, E. Havrdova, M. Hutchinson, L. Kappos, F. D. Lublin, X. Montalbán, P. W. O'Connor, M. Sandberg-Wollheim, A. J. Thompson, E. Waubant, B. G. Weinshenker, and J. S. Wolinsky. "Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria." *Annals of Neurology* 69.2 (Feb. 2011), pp. 292–302.

[Quast *et al.*, 1993] M. J. Quast, N. C. Huang, G. R. Hillman, and T. A. Kent. "The evolution of acute stroke recorded by multimodal magnetic resonance imaging." *Magnetic Resonance Imaging* 11.4 (Jan. 1993), pp. 465–71.

[RamoTello *et al.*, 2014] C. Ramo-Tello, L. Grau-López, M. Tintoré, A. Rovira, L. Ramió i Torrenta, L. Brieva, A. Cano, O. Carmona, A. Saiz, F. Torres, P. Giner, C. Nos, A. Massuet, X. Montalbán, E. Martínez-Cáceres, and J. Costa. "A randomized clinical trial of oral versus intravenous methylprednisolone for relapse of MS." *Multiple Sclerosis* 20.6 (May 2014), pp. 717–25.

[Rao *et al.*, 2015] V. Rao, M. S. Sarabi, and A. Jaiswal. "Brain Tumor Segmentation with Deep Learning". *Multimodal Brain Tumor Image Segmentation (BRATS) Challenge, MICCAI.* 2015, p. 56.

[Reed *et al.*, 2001] S. D. Reed, S. C. Cramer, D. K. Blough, K. Meyer, J. G. Jarvik, and D. Z. Wang. "Treatment With Tissue Plasminogen Activator and Inpatient Mortality Rates for Patients With Ischemic Stroke Treated in Community Hospitals Editorial Comment". *Stroke* 32.8 (Aug. 2001), pp. 1832–1840.

[Reza *et al.*, 2015] S. M. S. Reza, L. Pei, and K. M. Iftekharuddin. "Ischemic Stroke Lesion Segmentation Using Local Gradient and Texture Features". *Ischemic Stroke Lesion Segmentation (ISLES) Challenge, MICCAI.* 2015.

[Rish, 2001] I. Rish. "An empirical study of the naive Bayes classifier". *Empirical methods in artificial intelligence workshop, IJCAI.* Vol. 22230. JANUARY 2001. 2001, pp. 41–46.

[Robben *et al.*, 2016] D. Robben, D. Christiaens, J. R. Rangarajan, J. Gelderblom, P. Joris, F. Maes, and P. Suetens. "A Voxel-wise, Cascaded Classification Approach to Ischemic Stroke Lesion Segmentation". *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries.* Ed. by A. Crimi, O. Maier, B. Menze, M. Reyes, and H. Handels. Vol. 9556. LNCS. Springer International Publishing, 2016, pp. 254–265.

[Ronneberger *et al.*, 2015] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation". *Medical Image Computing and Computer-Assisted Intervention (MICCAI).* Vol. 9351. LNCS. Springer Berlin Heidelberg, 2015, pp. 234–241.

[Rosenberg, 1999] G. A. Rosenberg. "Ischemic brain edema." *Progress in Cardiovascular Diseases* 42.3 (Jan. 1999), pp. 209–16.

[Roxburgh *et al.*, 2005] R. H. S. R. Roxburgh, S. R. Seaman, T. Masterman, A. E. Hensiek, S. J. Sawcer, S. Vukusic, I. Achiti, C. Confavreux, M. Coustans, E. le Page, G. Edan, G. V. McDonnell, S. Hawkins, M. Trojano, M. Liguori, E. Cocco, M. G. Marrosu, F. Tesser, M. A. Leone, A. Weber, F. Zipp, B. Miterski, J. T. Epplen, A. Oturai, P. S. Sørensen, E. G. Celius, N. T. Lara, X. Montalbán, P. Villoslada, A. M. Silva, M. Marta, I. Leite, B. Dubois, J. Rubio, H. Butzkueven, T. Kilpatrick, M. P. Mycko, K. W. Selmaj, M. E. Rio, M. Sá, G. Salemi, G. Savettieri, J. Hillert, and D. A. S. Compston. "Multiple Sclerosis Severity Score: using disability and disease duration to rate disease severity." *Neurology* 64.7 (Apr. 2005), pp. 1144–51.

[Rudick *et al.*, 2002] R. Rudick, G. R. Cutter, and S. C. Reingold. "The Multiple Sclerosis Functional Composite: a new clinical outcome measure for multiple sclerosis trials". *Multiple Sclerosis* 8.5 (Oct. 2002), pp. 359–365.

[Sadovnick *et al.*, 1982] A. D. Sadovnick and P. A. Baird. "Sex ratio in offspring of patients with multiple sclerosis." *The New England Journal of Medicine* 306.18 (May 1982), pp. 1114–5.

[Schlaug *et al.*, 1999] G. Schlaug, A. Benfield, A. E. Baird, B. Siewert, K. O. Lovblad, R. A. Parker, R. R. Edelman, and S. Warach. "The ischemic penumbra: Operationally defined by diffusion and perfusion MRI". *Neurology* 53.7 (Oct. 1999), pp. 1528–1528.

[Shamonin *et al.*, 2013] D. P. Shamonin, E. E. Bron, B. P. F. Lelieveldt, M. Smits, S. Klein, and M. Staring. "Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease." *Frontiers in neuroinformatics* 7.50 (Jan. 2013), p. 50.

[Shannon *et al.*, 1964] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1964.

[Shelhamer *et al.*, 2015] E. Shelhamer, J. Long, and T. Darrell. "Fully Convolutional Networks for Semantic Segmentation". *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3431–3440.

[Simonyan *et al.*, 2014] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition" (Sept. 2014). arXiv: `1409.1556`.

[Smith, 2002] S. M. Smith. "Fast robust automated brain extraction." *Human Brain Mapping* 17.3 (Nov. 2002), pp. 143–55.

[Springenberg *et al.*, 2014] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. "Striving for Simplicity: The All Convolutional Net" (Dec. 2014). arXiv: `1412.6806`.

[Srivastava *et al.*, 2014] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting". *Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958.

[Stellmann *et al.*, 2014] J. P. Stellmann, E. Vettorazzi, J. Poettgen, and C. Heesen. "A 3 meter Timed Tandem Walk is an early marker of motor and cerebellar impairment in fully ambulatory MS patients." *Journal of the Neurological Sciences* 346.1-2 (Nov. 2014), pp. 99–106.

[Stüve *et al.*, 2010] O. Stüve and J. Oksenberg. "Multiple Sclerosis Overview". *Gene Reviews*. Ed. by R. A. Pagon, M. P. Adam, H. H. Ardinger, S. E. Wallace, A. Amemiya, L. J. Bean, T. D. Bird, N. Ledbetter, H. C. Mefford, R. J. Smith, and K. Stephens. May 11, 20. University of Washington, Seattle, May 2010. URL: `http://www.ncbi.nlm.nih.gov/books/NBK1316/`.

[Styner *et al.*, 2008] M. Styner, J. Lee, B. Chin, and M. Chin. "3D segmentation in the clinic: A grand challenge II: MS lesion segmentation". *Midas* (2008), pp. 1–6.

[Sudre *et al.*, 2015] C. H. Sudre, M. J. Cardoso, W. H. Bouvy, G. J. Biessels, J. Barnes, and S. Ourselin. "Bayesian Model Selection for Pathological Neuroimaging Data Applied to White Matter Lesion Segmentation". *IEEE Transactions on Medical Imaging* 34.10 (Oct. 2015), pp. 2079–2102.

[Szegedy *et al.*, 2015] C. Szegedy, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going deeper with convolutions". *Conference on Computer Vision and Pattern Recognition (CVPR)* (Sept. 2015), pp. 1–9. arXiv: `arXiv:1409.4842v1`.

[TomasFernandez *et al.*, 2015] X. Tomas-Fernandez and S. K. Warfield. "A model of population and subject (MOPS) intensities with application to multiple sclerosis lesion segmentation". *IEEE Transactions on Medical Imaging* 34.6 (2015), pp. 1349–1361.

[Triguero *et al.*, 2015] I. Triguero, S. Garcia, and F. Herrera. "Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study". *Knowledge and Information Systems* 42.2 (2015), pp. 245–284.

[Vaidhya *et al.*, 2016] K. Vaidhya, S. Thirunavukkarasu, V. Alex, and G. Krishnamurthi. "Multi-modal Brain Tumor Segmentation Using Stacked Denoising Autoencoders". *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Ed. by A. Crimi, B. Menze, O. Maier, M. Reyes, and H. Handels. Vol. 9556. LNCS. Springer International Publishing, 2016, pp. 181–194.

[Vaidya *et al.*, 2015] S. Vaidya, A. Chunduru, R. Muthuganapathy, and G. Krishnamurthi. "Longitudinal Multiple Sclerosis Lesion Segmentation using 3D Convolutional Neural Networks". *Longitudinal Multiple Sclerosis Lesion Segmentation Challenge, ISBI*. Ed. by D. L. Pham. online, 2015. URL: `http://iacl.ece.jhu.edu/MSChallenge`.

[Van Leemput *et al.*, 1999] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. "Automated model-based tissue classification of MR images of the brain." *IEEE Transactions on Medical Imaging* 18.10 (Oct. 1999), pp. 897–908.

[Vargas *et al.*, 2009] M. I. Vargas, J. Delavelle, R. Kohler, C. D. Becker, and K. Lovblad. "Brain and spine MRI artifacts at 3 Tesla". *Journal of Neuroradiology* 36.2 (May 2009), pp. 74–81.

[Vincent *et al.*, 2010] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion". *Journal of Machine Learning Research* 11 (2010), pp. 3371–3408.

[Von Luxburg, 2007] U. Von Luxburg. "A tutorial on spectral clustering". *Statistics and Computing* 17.4 (2007), pp. 395–416.

[Weinshenker *et al.*, 1989] B. G. Weinshenker, B. Bass, G. P. Rice, J. Noseworthy, W. Carriere, J. Baskerville, and G. C. Ebers. "The natural history of multiple sclerosis: a geographically based study. I. Clinical course and disability." *Brain* 112.1 (Feb. 1989), pp. 133–46.

[Zhang, 2004] H. Zhang. "The Optimality of Naive Bayes". *Florida Artificial Intelligence Research Society Conference (FLAIRS)*. Vol. 1. 2. 2004, p. 97.

[Zhang *et al.*, 2001] Y. Zhang, M. Brady, and S. Smith. "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm". *IEEE Transactions on Medical Imaging* 20.1 (Jan. 2001), pp. 45–57.

[Zhu, 2005] X. Zhu. *Semi-Supervised Learning Literature Survey*. Tech. rep. 1530. Computer Sciences, University of Wisconsin-Madison, 2005, p. 60.

[Zhu *et al.*, 2003] X. Zhu, Z. Ghahramani, J. Lafferty, et al. "Semi-supervised learning using gaussian fields and harmonic functions". *International Conference on Machine Learning (ICML)*. Vol. 3. 2003, pp. 912–919.

Abbreviations

AIF	arterial input function. 10, 11, 151, 173
ANN	artificial neural network. 131
ASD	average surface distance. 51
ASSD	average symetric surface distance. 51, 53, 54, 62, 63, 70, 87, 89, 122, 127, 128
BF	bagging forest. 43
CNN	convolutional neural network. 38, 43, 58, 59, 71, 73, 90, 124, 131, 132
CNS	central nervous system. 28, 29, 31
CRF	conditional random field. 39, 124
CSF	cerebral spinal fluid. 6, 8, 13, 16, 32, 33, 143–145
CT	computed tomography. 22, 49
DC	Dice's coefficient. 51, 53–56, 59, 62, 63, 65, 68, 70, 87–89, 109–111, 113–117, 122, 127–129
DF	decision forest. 2–4, 38, 39, 41–43, 51, 55–59, 61, 73, 75–78, 80, 86–90, 94, 95, 97, 102, 104, 108–111, 113, 114, 116–119, 121–124, 131–133
DNN	deep neural network. 131–133
DSC	Dynamic Susceptibility Contrast. 9, 10, 151, 173
DT	decision tree. 40, 42, 43, 58
DynStatCov	Dynamic Statistical Co-Variance. 102, 103, 119
EDSS	Expanded Disability Status Scale. 30
EF	extra forest. 43, 53–55, 57, 59
EPI	echo planar imaging. 16
GB	Gradient Boosting. 58
GM	gray matter. 6, 7, 33
GMM	Gaussian Mixture Model. 37, 38
GNB	Gaussian Naive Bayes. 57
GPU	graphics processing unit. 38, 120, 124, 131, 132
HD	Hausdorff distance. 51, 53, 54, 70, 87, 89, 127, 128
HG	high grade. 68, 69
kNN	k-Nearest-Neighbors. 57, 84, 85, 89, 101
LDDP	limiting density of discrete points. 96
lFPR	lesion detection false positive rate. 65, 67, 113, 115–118
LG	low grade. 68, 69
longCorr	longitudinal volume change correlation. 65
LPF	local problem forests. 77–80, 82–84, 86–90, 123, 125
lTPR	lesion detection true positive rate. 65, 67, 113, 115–118
MCA	middle cerebral artery. 20, 24, 62, 64
MRF	Markov random field. 39, 124

MRI	magnetic resonance imaging. 1–3, 5–8, 10, 12, 13, 16, 17, 21–24, 28–35, 37–39, 48–50, 52, 53, 61, 62, 64, 67, 69, 70, 72, 75, 90, 93, 121, 122, 124, 132, 135–138, 142, 143, 151, 173
MS	Multiple Sclerosis. 2–4, 7, 17, 19, 28–35, 37, 64, 65, 67, 72, 93, 95, 104, 105, 108, 111–114, 116, 118–121, 123, 124, 176
MSFC	Multiple Sclerosis Functional Composite. 30
MSSS	Multiple Sclerosis Severity Score. 30
PDF	probability density function. 98, 99, 101, 104, 105, 116–119
PPMS	primary progressive MS. 28, 29
PPV	positive prediction value. 51, 65, 68, 113
PRMS	progressive relapsing MS. 28
rAVD	relative absolute volume difference. 113, 116, 117
RELU	rectified linear activation unit. 58
RF	radio frequency. 5, 6, 13, 16, 135–147
RRMS	relapsing-remitting MS. 28, 29, 34
SE	sensitivity. 51, 68
SNR	signal-to-noise ratio. 6, 8, 12
SPMS	secondary progressive MS. 28, 29
SSF	semi-supervised forest. 93–95, 98, 104, 105, 108–111, 113–120, 123–125
STD	standard deviation. 54, 63, 70, 116, 122
SVM	support vectors machine. 90
TE	echo time. 12, 142–144
TI	inversion time. 13, 144
totalCorr	general volume correlation. 65
TP	time point. 34, 35, 64, 67, 93, 108, 109, 112, 113, 116–121, 123
TPR	true positive rate. 65, 113
TR	repetition time. 12, 142, 143
TSVM	Transductive Support Vector Machines. 94, 95
WM	white matter. 6, 7, 30, 33
WMH	white matter hyperintensities. 70, 81, 90, 116
WML	white matter lesion. 25, 30, 53, 75

.

Glossary

ADC	Magnetic resonance imaging (MRI) map computed from multiple DWI sequences denoting the apparent diffusion coefficient. 9, 23, 26, 27, 52–54, 75, 146, 147, 149, 150, 177
BRATS	Multimodal brain tumor segmentation challenge 2015. 37, 61, 67–69, 71, 72
CBF	An MRI PWI map denoting cerebral blood flow. 11, 62, 63, 151
CBV	An MRI PWI map denoting cerebral blood volume. 11, 62, 63, 151
DWI	Diffusion weighted MRI sequence, here usually referring to the its trace image. 8, 9, 13, 14, 16, 23, 25, 52–54, 62, 63, 70, 75, 144–146, 148, 149, 173
FLAIR	T2 weighted MRI sequence with fluid signal suppression. 8, 12, 13, 16, 23, 25, 27, 32–34, 49, 52–54, 57, 64, 66, 69–71, 75, 76, 87, 90, 108, 143–145
ISBIMS	Longitudinal multiple sclerosis lesion segmentation challenge 2015. 32, 34, 37, 61, 64–66, 72, 93, 104, 108, 112, 113, 118
ISLES	Ischemic stroke lesion segmentation challenge 2015, comprised of SISS and SPES. 37, 62, 69, 127–129, 173, 177, 179
MRA	MRI technique for angiography sequence acquisition with a contrast agent. 9, 23
MTT	An MRI PWI map denoting the maximum transit time of the bolus. 10, 11, 63, 151
PD	MRI sequence weighted based on actual positron density. 7, 23, 32, 64, 66, 108, 141–143
PWI	Time-resolved MRI perfusion weighted imaging, here always conducted with the Dynamic Susceptibility Contrast (DSC) method. 9, 23, 27, 50, 73, 173
SISS	Sub-acute ischemic stroke lesion segmentation challenge 2015, part of ISLES. 61, 69–73, 123, 127, 128, 133, 173
SPES	Stroke penumbra segmentation challenge 2015, part of ISLES. 61–64, 69, 70, 72, 122, 127, 128, 133, 173
T1	MRI sequence weighted based on T1 values. 7, 8, 12, 14, 15, 23, 31, 32, 49, 52, 53, 64, 66, 69–71, 75, 108, 138–141, 143, 173
T1c	T1 weighted MRI sequence acquired with contrast agent. 7, 31, 34, 62, 63, 69, 71, 143
T2	MRI sequence weighted based on T2 values. 7–9, 15, 16, 23, 25, 26, 32–34, 49, 52, 53, 62–64, 66, 69, 71, 75, 108, 138, 140–148, 150, 151, 173, 177
T2*	The effective rather than the natural T2 relaxation time. 9, 10, 138, 141, 151
Tmax	An MRI PWI map denoting the time between the peak of the arterial input function (AIF) and the peak of the residue function. 11, 26, 62, 151
TTP	An MRI PWI map denoting the time to the maximal bolus induced intensity. 10, 26, 62, 63, 151

List of Figures

List of Tables