Variability-Aware and Weakly Supervised Learning for Semantic Tissue Segmentation

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

der KIT-Fakultät für Informatik
des Karlsruher Institut für Technologie (KIT)

genehmigte

Dissertation

von

Michael Götz

aus Schwäbisch Hall

Tag der mündlichen Prüfung:	08. Juni 2017
Erster Gutachter:	Prof. Dr.-Ing. Rüdiger Dillmann
Zweiter Gutachter:	PD. Dr. rer. nat. Klaus H. Maier-Hein

Contents

Abstract

Fully automatic methods for semantic tissue characterization play an increasingly important role in analyzing medical images due to the numerous and highly resolved images, which are increasingly generated. These procedures are particular useful in oncological radiology and radiation therapy where they can support physicians in everyday clinical practice with different tasks, such as radiation therapy planning or quantification of therapy outcome. In addition, they are a prerequisite in the implementation of large-scale image-based medical studies, where an automatic evaluation and quantification of the image content is indispensable.

In general, two major questions are investigated: 1) How to handle the variability of training data and 2) how to reduce the annotation effort? The influence of these difficulties is examined, and new methods are presented to reduce the impact of these challenges. The main contributions are 1) two methods to reduce the influence of the variability in the training data and 2) three methods to learn from weakly annotated (image) data.

While the analyses on the normalization of MR data and on imaging modalities presented in this work are strongly focused on applications in the medical field, this is not the case for the proposed methods. All developed methods have the potential to be applied to other questions in the area of computer vision or more general questions in the field of machine learning. Although the proposed methods are only evaluated using medical data, no assumptions are made about the underlying image modality or the entity under consideration. It is also verified that all the algorithms presented can handle multi-class problems.

Kurzfassung

Angesichts der zahlreichen und hofaufgelösten medizinischen Bilddaten, die zunehmend anfallen, spielen Verfahren zur vollautomatischen Gewebecharakterisierung bzw. -segmentierung eine immer wichtigere Rolle. Dabei sind vielfältige Einsatzzwecke möglich: Vereinfachung der Bestrahlungsplanung, bessere Quantifizierung des Therapieerfolges sowie bei der Durchführung von großen bildbasierten medizinischen Studien. Dabei profitieren alle Anwendungsfälle sowohl von der Reduktion der manuellen Arbeit als auch von der höheren Reproduzierbarkeit der Ergebnisse.

Stand der Forschung

In der Forschung wird aktuell vor allem an lernbasierten Methoden zur semantischen Gewebesegmentierung geforscht. Diese liefern auf der einen Seite gute Ergebnisse, auf der anderen ist die Hoffnung, diese relativ einfach an andere Fragestellungen oder eine aktualisierte Bildgebung anpassen zu können. Doch obwohl in der Literatur bereits mehrere solcher Verfahren präsentiert wurden, sind bisher nur einzelne Systeme tatsächlich im klinischen Einsatz. In dieser Arbeit werden zwei hierfür wesentliche Gründe betrachtet. Zum einen weisen medizinische Daten eine hohe Variabilität auf, was sowohl durch die oft qualitative Bildgebung als auch durch die Varianzen der Erkrankungen und Physiologie bedingt ist (Abb. 1a,b,c). Zum anderen ist die Annotation der notwendigen Trainingsdaten meist sehr zeitaufwendig und in der Regel mehrdeutig, müssen aber an die Bildgebung vor Ort angepasst werden.

Um die Varianz der Daten zu verringern wird aktuell häufig auf Intensität-Normalisierung zurückgegriffen. Allerdings ist bekannt, dass diese Verfahren die erzeugte Varianz nur teilweise ausgleichen können. Andere Verfahren versuchen, die Unterschiede zwischen verschiedenen Scannern mithilfe von Transfer Lernen auszugleichen, sind dabei aber auf spezifische Fragestellungen limitiert.

Um keine Trainingsdaten annotieren zu müssen, werden viele Verfahren mit Wettbewerbsdaten evaluiert, können dann aber nicht unbedingt in die klinis-

che Routine eingebaut werden. Andere Verfahren versuchen die Menge der notwendigen Trainingsdaten zu reduzieren, indem entitätsspezifisches Vorwissen eingebracht wird. Allerdings sind die Verfahren häufig nur schwer mit anderen Verfahren zu kombinieren oder auf andere Fragestellungen übertragbar.
In dieser Arbeit werden deshalb zwei wesentliche Fragestellungen untersucht:
Frage 1: Wie kann trotz Varianz gelernt werden?
Frage 2: Wie kann der Annotationsaufwand reduziert werden?

Lernen mit variablen medizinischen Bilddaten

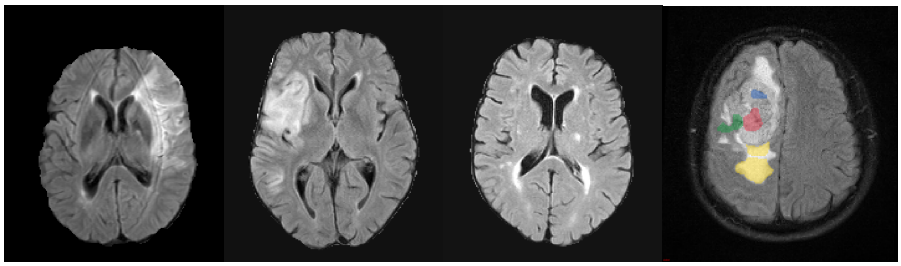
Der erste Schritt dieser Arbeit besteht darin, zu untersuchen inwieweit bestehende Verfahren in der Lage sind die Varianz der medizinischen Bildgebung auszugleichen. In einer Studie wird deshalb untersucht, welchen Einfluss verschiedene Verfahren zur Intensitätsnormalisierung in MRI Bildern bei einem lernbasierten Vorgehen haben. Eine zweite Studie, die ein konventionelles Random Forest Training mit einem neuen, erstmalig für die Segmentierung von Gehirntumoren eingesetzten Verfahren vergleicht, erlaubt die Ergebnisse der ersten Studie besser einzuordnen. Dabei wird festgestellt, dass einfache Normalisierungsverfahren meist besser funktioniere, aber letztlich nicht ausreichen, um alle Unterschiede ausgleichen zu können.

Aufgrund dieses Ergebnisses werden zwei Algorithmen vorgestellt um besser mit dieser Varianz umgehen zu können. Das erste Verfahren kombiniert dabei erstmalig Methoden aus dem Bereich Metric-Learning, Supervised Learning und Atlas-Based Segmentation um für jedes Bild einen individuellen Klassifikator zu trainieren. Dabei werden nur die Trainingsdaten genutzt, die wahrscheinlich am besten geeignet sind. Dadurch wird das Verfahren robuster und es kann gezeigt werden, dass die erreichten Ergebnisse deutlich besser werden. Im Gegensatz dazu wurde das zweite Verfahren entwickelt um nachträgliche Korrekturen von automatisch erzeugten Segmentierungen zu vereinfachen. Hier werden vortrainierte Klassifikatoren durch manuelle Interaktionen an den jeweiligen Patienten angepasst. Durch das so eingesetzte Vorwissen konvergiert das Ergebnis deutlich schneller als dies bei traditionellen manuellen Verfahren der Fall ist.

Die bisher vorgestellten Techniken und Versuchen zeigen auf welch großen Einfluss die Variabilität der Trainingsdaten auf die Genauigkeit von Segmentierungsalgorithmen für medizinische Daten hat. Mit Hilfe der beiden vorgestellten neu entwickelten Algorithmen lässt sich der negative Einfluss dieser Variabilität jedoch stark verringern.

Reduktion des Annotationsaufwandes

Um das Erstellen der notwendigen Trainingsannotationen zu vereinfachen wurden drei Verfahren entwickelt um mit unterschiedlichen Arten der Annotationen zu lernen. Für das erste Verfahren wurde auf eine vollständige Annotation der Bilddaten verzichtet, statt dessen werden nur wenige, dafür aber repräsentative und eindeutig identifizierbare Gebiete in den Trainingsdaten annotiert (Abb. 1d). Ein vorgestelltes Korrekturverfahren bedingt, das die Reduktion zu keiner signifikanten Verschlechterung des Segmentierungsergebnisses führt.



(a) (b) (c) (d)

Figure 1: Beispiel für Varianz in medizinischen Bilddaten: Abbildung (a), (b), und (c) zeigen unterschiedliche Patienten mit der gleichen Erkrankung. Die Erscheinungen sind klar unterschiedlich. Abbildung (d) zeigt beispielhaft die spärliche Annotation eines Patienten. Diese sind häufig in nur einer oder zwei Schichten des 3D-Bildes.

Das zweite Verfahren, dass auf dem vorherigen aufbaut, erlaubt zusätzlich ohne eine Verschlechterung der Ergebnisse auf die Annotation einer Gewebeklasse zu verzichten und so auf bis zu 45 % der Annotationen zu verzichten. Dazu werden zusätzlich Methoden aus dem Bereich des PU-Learnings benutzt. Mit beiden Verfahren kann die Annotationsdauer eines Patienten für die Trainingsdatenbank von mehreren Stunden auf wenige Minuten reduziert werden. Der dritte Ansatz verzichtet vollständig auf eine räumliche Annotation. Stattdessen wird versucht, über das klinisch bestimmte Tumorvolumen einen Klassifikator zu trainieren. Dafür wird der erste Random-Forest-basierte 'Learning from Label Proportions (LLP)' –Algorithmus vorgestellt und evaluiert.

Die Vorgeschlagenen Verfahren erlauben eine schnellere Annotation der Daten und ermöglichen dadurch die Adaptation eines Lernalgorithmus an die aktuellen Gegebenheiten. Der praktische Einsatz der vorgeschlagenen Algorithmen wird zudem in zwei Untersuchungen

Fazit

Insgesamt werden in dieser Arbeit zwei große Fragestellungen untersucht: 1) Wie kann mit der Variabilität von Trainingsdaten umgegangen werden und 2) wie kann der Annotationsaufwand reduziert werden. Dazu wird zuerst der jeweilige Einfluss dieser Schwierigkeiten untersucht, anschließend werden neue Methoden zur Reduktion der negativen Einflüsse dieser beiden Aspekte vorgestellt. Die wesentlichen Beiträge umfassen dabei 1) zwei Verfahren zur Reduktion des Einflusses der Variabilität in den Trainingsdaten und 2) drei Methoden, die es ermöglichen aus schwach annotierten (Bild)Daten zu lernen. Dabei wurde während der Entwicklung der Algorithmen darauf geachtet, dass keine Entitätsspezifischen Annahmen getroffen werden und alle Algorithmen mit Mehrklassen-Fragestellungen umgehen können. Dadurch sollte es möglich sein, die vorgestellten Verfahren auch auf andere Fragestellungen anzuwenden.

Acronyms

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
BRI	Best Ratio based Impurity
CRI	Class-dependent, Ratio based Impurity
CSF	Cerebrospinal Fluid
CSS	Classification Similarity Score
CT	Computed Tomography
DALSA	Domain Adaptation for Learning from Sparse Annotations
DA-PEPE	Domain Adapted Pearson Divergence Prior Estimation
DECT	Dual Energy CT
DWI	Diffusion Weighted Imaging
EM	Expectation - Maximization
ExtraTrees	Extremely Randomized Trees
FLAIR	Fluid-attenuated Inversion Recovery
GTV	Gross Tumour Volume
ICH	Intracerebral Haemorrhages
IDAL	Input Data Adaptive Learning
i.i.d.	independent and identically distributed

ITK	Insight Segmentation and Registration Toolkit
kNN	k-Nearest Neighbours
LCA	Learning from Complete Annotations
LP-Forest	Label Proportion Forest
LLP	Learning form Label Proportions
LRC	Logistic Regression Classifier
LSA	Learning from Sparse Annotations
MCMC	Markov Chain Monte Carlo
MIL	Multi-Instance Learning
MITK	Medical Imaging Interaction Toolkit
MRI	Magnetic Resonance Imaging
NAF	Neighbourhood Approximation Forest
PCA	Principal Component Analysis
PE	Pearson Divergence
PEPE	Pearson Divergence Prior Estimation
PU-learning	Positive and Unlabelled Learning
RBM	Restricted Boltzmann Machine
RDF	Random Decision Forest
ROI	Region of Interest
SC	Similarity Classifier
STAPLE	Simultaneous Truth and Performance Level Estimation
SUR	Sparse and Unambiguous Region
SVM	Support Vector Machine
TPR	True Positive Rate
VC	Voxel Classifier

1

Introduction

1.1 Motivation

Medical imaging allows physicians to obtain a better understanding of their patients. Obtaining a 3D image in oncologic-, radiation-, or chemo-therapy provides insight about the structure of the affected tissue, as well as about the size and location of the pathology. The fast acquisition of relevant information as well as the low invasivity of most methods make imaging an important source of information which are otherwise difficult or impossible to obtain (Kurland et al., 2012; Buckler et al., 2011a; Buckler et al., 2011b). For this reason, the number of medical images is steadily increasing (BfS, 2016).

A common approach of assessing a medical image is the qualitative description of its relevant content. For example, malignant tissue might be characterized on basis of its texture using terms such as 'smooth', 'grizzly', or 'inhomogeneous'. Such descriptions are hard to compare to each other. Not only that different words are used by different radiologist, but also the corresponding meaning differs. For example what defines a 'smooth' image surface and when is it no longer smooth? Having quantitative measurements would allow a better comparison between findings and subsequently a better characterization of diseases. Although grading schemes like PIRADS (Weinreb et al., 2016) or BIRADS (Liberman et al., 1998) giving a more standardized way of reporting the findings but still rely on simply measurements or word-based descriptions of the images. The ambiguity of the description or the measurement errors of simple measurements like largest diameter can lead to unclear or false gradings (Marten et al., 2006). Aerts et al. (2014) showed that using quantitative descriptions like surface to volume ratio or co-occurrence matrix based texture descriptions gives additional information that can improve therapy decisions. They therefore proposed a new approach called radiomics based on the idea to have prognostic feature representation of malignant tissue.

But to obtain quantitative measurements of malign tissue, it is inevitable to have a full segmentation of the affected area. For example, to have a numeric

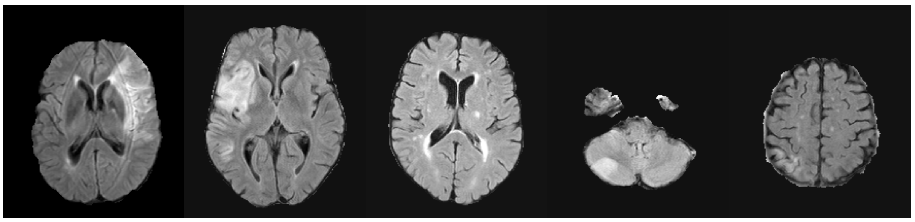
description of the shape, it is necessary to have access to the shape at first. This requirement is a serious limitation of such approaches, as the full annotation of tumorous tissue is often a time consuming and error prone process. Mazzara et al. (2004) showed, that clinical outlines of brain tumours take usually more than 20 minutes per patient and still have an observer-variability of 28 %. So even if manual annotations are available, their quality is often not sufficient to be used for the calculation of quantitative tissue descriptions. This variability in the findings can have severe effects if therapy decisions are made on these findings. For example, a slightly smaller segmentation might indicate a therapy success while in fact the tumour grew and the therapy failed. Another problem might be in radiotherapy, where the radiation is applied according to the previous annotations.

For these reasons automatic tissue characterization systems are currently an open research question. These systems overcome the limitation of manual annotations by estimating the type of tissue at each location, thus resulting in a segmentation of regions of interests like tumours. Such systems do not only reduce the manual effort and therefore cost but are also more consistent in their results (S. Wang and Summers, 2012). Even though the decisions of different systems may be different, the same system will always produce the same segmentation for any given patient. This is an important property for the robust calculation of quantitative image features based on these segmentations.

Nowadays, most of the proposed tissue characterization systems are based on machine learning methods (Bauer, Wiest, et al., 2013). Such systems can achieve high accuracy as it was also shown in general computer vision problems and come with other interesting properties. Since the appearance is learned based on a feature vector or intensity patch only few changes in the algorithm are necessary in theory if new modalities or diseases are targeted. At the same time it is straight forward to include different modalities in a training, i.e. allowing for multi-spectral systems. As different modalities often show complementary information, using more than one modality can significantly improve the performance. As a positive side-effect, learning based systems that make use of more than one modality might even be used to detect connections between different modalities that were previously unknown (Rongjian Li et al., 2014).

1.2 Challenges

The ongoing research in the field of learning-based tissue characterization led to a significant amount of different approaches. But as most techniques are inspired by methods from the general computer vision community, there is a strong assumption that enough training samples are readily available. It can be considered that the necessary effort for creating new training data is the main reason that most clinics do not use a knowledge based system for tissue characterization. The training data set needs not only to be created for every type of questions, like brain tumour or liver tumour segmentation. The training set then needs to capture all variability of the human physiology, imaging artefact and variability in the appearance of the given disease. These variability can lead to significant differences in the appearances. For example, Figure 1.1 shows possible variations of the same disease. In addition, medical imaging, especially Magnetic Resonance Imaging (MRI) are highly variable and



(a) (b) (c) (d) (e)

Figure 1.1: All images are showing MRI Flair images of patients with subacute strokes. The variations within the images are due to imaging differences, different location of the stroke, stroke age, etc.. All images are taken from the ISLES training data set.

non-quantitative. Potential sources of variations include varying internal values ranges, slightly varying sequence implementations, or varying noise characteristics between scanners or different manufacturers. Therefore, an individual training set need to be created or adapted for every MRI-scanner, and each hardware change (Opbroek, Ikram, et al., 2015). Furthermore, MRI sequences are currently subject to research and might change, which also require an adaptation of the training data. Without these adapted training set the quality of the classifier might significantly drop or become even impractical if non-matching sequences are used (Opbroek, Ikram, et al., 2015).

Adapting or recreating new training datasets requires must therefore be fast and cheap. But creating full semantic segmentations of medical images is often time-consuming and error prone. Having three-dimensional (3D) images increases the number of voxels that needs to be labelled. In addition, the necessary information is often distributed over different modalities, making it necessary to continuously switch between different images. The annotation becomes even more complicated since medical images are often ambiguous. Partial volume effects and partial invaded tissue often lead to blurry borders between two tissue types as it can be seen in Figure 1.1. In addition, some tissue types, like blood, inflammation, or edema in brain images, have a common appearance. These uncertainties do not only add to the long annotation time of such images, but also lead to low inter- and intra-observer variabilities. Consequently, Mazzara et al. (2004) report an intra- and inter-observer variability of 20 % and 28 % respectively for gross tumour segmentation. Similar B. H. Menze, Jakab, et al. (2015) reported interrater-variances of 85.8 % for whole tumour segmentation and 74.1 % for active tumour segmentation and an annotation time of four hours for a single subject.

1.3 Objectives

This thesis focuses on the development of approaches that allow to overcome the current limitations and bring learning-based approaches into the clinical routine. Based on the assumption that existing approaches have sufficient accuracy, it therefore evaluates the question how to cope with the uncertainties and availability of medical training data. To make this possible, the following

objectives were followed:

- **Fast training data annotation**: The fast annotation of new training data allows to adapt the system to the individual conditions. This is necessary to adapt the training base to the given questions at hand as well as the imaging routine within the given clinical routine.

- **Handle ambiguity**: It is often impossible to unambiguously assign a tissue class to each voxel in a medical image. The underlying tissue might be only partially infiltrated by tumorous tissue, be affected by partial volume effects or it might be impossible to distinguish between similar looking tissue types. Therefore a new training method should be developed that is able to reduce the effect of such ambiguous areas during the training.

- **Handle variability**: Medical images from the daily routine can be very variable. Not only because of variances between different scanners and sequences but also because of artefact. The developed algorithms should therefore provide a mechanism to adapt for the variability within the training data.

- **General applicability**: The developed algorithms should be generally applicable as much as possible. Methods that included domain-specific knowledge, as for example atlas-based segmentation methods, are therefore excluded.

- **Compatible with state-of-the-art**: There is a vivid research community looking at tissue characterization. It is therefore an important objective that the new developed methods can be easily combined with other learning-based approaches in order to benefit from existing and future solutions.

1.4 Summary of contributions

Instead of a single contribution that meets all the objective, multiple small contributions which are more specific to single challenges, are made. This allow to focus on specific challenges.

- **Evaluation of normalization for variability reduction**: The traditional way to handle MRI-variability is using intensity normalization methods. The performance of these algorithms is evaluated in a learning-based setting. This study provides an important contribution to the understanding of MR normalization and shows that simpler normalization procedures often achieve better results in the area of tumor segmentation.

- **Evaluation of classifier influence**: Two learning algorithms for 'Random Decision Forests (RDF)' are compared and a new training method is used for the first time for the segmentation of tumors. The associated experiments proof that the learning algorithm have an influence on the performance of tumor segmentation systems, but the found influence is lower than the influence of the normalization algorithm.

- **Variability-Aware Learning with 'Input Data Adaptive Learning (IDAL)'**: To deal with the variability of the training data, an individual classifier is trained for each test image from those training images which are most likely to lead to a strong classifier. For this, the new approach combines for the first time techniques from the area of metrics learning, machine learning and atlas-based annotation transfer. The experiments show that a significant improvement can be achieved with this method compared to conventional approaches.

- **Variability-Aware Learning with Semi-Automatic Classifier Adaption**: A second method is presented, which allows a semiautomatic correction of an automatically generated segmentation with few manual interactions. For this, an existing classifier is corrected, which allows including the knowledge from the already annotated data. The experiments show that the presented algorithm significantly fewer interactions that conventional approaches.

- **Annotation reduction with Learning from Sparse Annotation**: A sparse and unambiguous annotation (Typically only 0.5 % are annotated) reduces annotation time. It is demonstrated how the reduction leads to a sampling bias and reducing segmentation performance. A method is presented that reduced the influence of the sampling bias by using domain adaptation techniques. It is demonstrated that the method reduces the annotation time per training patient from four hours to only five minutes without sacrificing classifier accuracy.

- **Annotation reduction with Learning from Positive and Unlabeled Data**: Through the first-time use of 'Positive and Unlabeled Learning (PU Learning)' methods and a domain-adapted version in the field of medical image processing, it is possible to avoid the annotation of a complete tissue class and still use sparse and unambiguous annotations. Experiments show that additional 45 % of the annotation time can be saved if healthy tissue does not have to be annotated while quality of the segmentations is not significantly different from conventional classifiers.

- **Annotation reduction with Learning from Label Proportions**: The third method can make use of already known ratios of individual tissue classes, and does not require their spatial position to be known. The 'Learning from Label Proportions (LLP)' setting is used in this setting a RDF-based training algorithm is proposed for the first time. This new learning method is evaluated using various synthetic as well as artificially grouped data. It is shown that the results of the method are of the same order of magnitude as the classical methods.

- **Application to Image Information Evaluation**: It is additionally evaluated how to assess the information content of different imaging modalities using sparse annotations. For this a learning-based comparison of the different modalities is proposed. By this, it is for the possible to include texture properties and intensity values into a single comparison as previous comparisons are using only a single information source. In addition, a method is introduced that allows combining several sparse annotations. The presented approach is robust against faulty annotation.

1.5 Outline

The first chapter after this introduction, chapter 2 gives the background information that are necessary to understand the presented work. It is not necessary to read the sections of chapter, if one is already familiar with the presented topic. In chapter 3, an overview of the current state-of-the-art is given. Similar to the following sections, this chapter is organized to reflect the four steps of the approach used for this thesis. The methods that are used to solve the given challenge are described in chapter 4, while chapter 5 explains the experiments that are conducted to evaluate the proposed methods. To allow a better understanding, the results are always reported directly after the description of the corresponding experiments. The discussion of the experiments and the findings is then given in chapter 6. Chapter 7 then gives a short outlook on possible further research and concludes the work by summarizing the findings.

2

Background

This thesis is affected by two different research area. State-of-the-art methods from the field of machine learning are used to solve questions within the field of medical imaging. Both fields do have their own techniques, approaches, and notations. Since methods of both fields are used, this chapter provides a basic background and clarification of the used terminology within this work. Hopefully, this will allow reader from both fields to understand the presented work. Based on this objective, this chapter is more generic, avoiding in-depth discussion within the different topics as a complete explanation is out-of-scope for this work.

2.1 Medical Background

The techniques developed for this thesis are not specific to any diseases. Since the aim of the proposed methods is to train a model for tissue appearance based on training data, it can be applied to different questions. Not only in regard to the modality but also with respect to disease which is reason for imaging. Therefore there is no specific medical knowledge required to understand the technical parts of this work. Nevertheless, a short description of the diseases that I worked with for the presented work will be given. Beside showing the importance, it will allow the reader to have a better understanding of the actual problem and terms which are specific to the medical domain.

2.1.1 Brain anatomy

Most tissue characterization algorithms are developed for the brain, mainly because it is less affected by breath motions. Due to that the obtained images are less affected by motion artefacts and contrast registration can be executed rigid, as tissue deformations are rare. Most of the experiments in this thesis are therefore conducted using brain images, therefore the brain anatomy will be described first before discussion each disease.

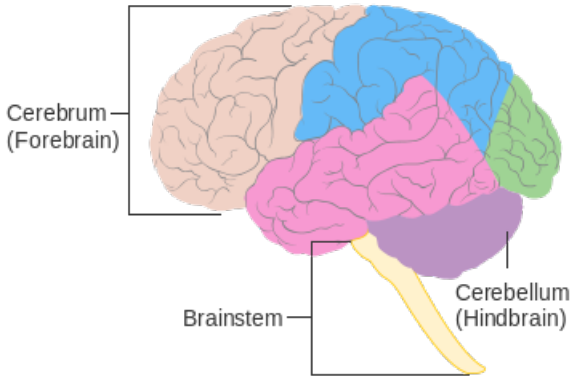


Figure 2.1: Regions of the lateral surface of the brain, and particularly the lobes of the fore-brain: Beige: frontal lobe, Blue: parietal lobe, Green: occipital lobe, Pink: temporal lobe. Taken from Various (2016a)

The human brain weights around 1.2 to 1.4 kg and is divided into three different regions (Figure 2.1). The Cerebrum is the largest part of the brain. Here, higher functions like vision, hearing, speech, and emotions are performed. and is connected to the cerebellum by the brain-stem. The much smaller Cerebellum is responsible for muscle movement and balance. Both are connected by the brain-stem, which is also responsible for many automatic functions like hearth rate, breathing, blinking etc. .

The surface of the Cerebrum is folded and called cortex. The folds are called gyrus, and the space between two gyrus is called sulcus (Figure 2.2). The outer part of the cortex consists of neurons, which form the grey matter. The next tissue type, the white matter, is responsible for connecting different neurons.

The volume not covered by the brain itself is filled with Cerebrospinal Fluid (CSF), a clear and colourless liquid. It provides basic mechanical and immunological protection to the brain. It is also used for the volume regularization and the cerebral auto-regulation. CSF is produced in the ventricular system in the brain, four interconnected cavities within the brain (Figure 2.2).

2.1.2 Brain Tumour (Gliomas)

Gliomas are the most common primary brain tumours. As it is arising from glial cells it affect mainly white matter. Different forms of Gliomas are differentiated by a rating from one to four, with four being the most severe and aggressive form which is also called Glioblastoma multiforme Louis et al. (2007). For this thesis, only data which shows this form are used and thus Glioma, Glioblastoma, and Glioblastoma multiforme are used as synonyms from now on.

Glioblastoma are known to affect mainly white matter, and do usually not spread into grey matter, ventricles, or the Cerebellum. Due to their infiltrate growth, tumorous cells are usually spread over the complete brain after the outbreak of the disease, which prevent a clear location of the tumour. In consequence, the treatment is difficult, and usually not successful, with an average

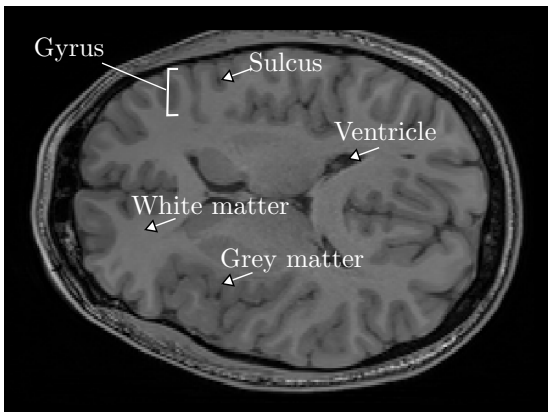


Figure 2.2: Axial slice of an MRI scan. The slice shows a part of the Cerebrum, different anatomical areas are indicated.

survival time around 15 months. The treatment includes usually a surgical intervention, which main purpose is to release the symptoms induced by additional mass.

The most common imaging during the therapy of glioblastoma is MRI. Common sequences that are used are $T2_w$, FLAIR, Diffusion, and $T1_w$ with contrasts. $T1_w$ indicates areas with high amount of active tumour cells, due to the increase uptake of contrast agent at these locations. These areas are often titled as 'active tumour core'. It often surrounds an necrotic area – previous tumorous areas that lost access to the blood system due to the rapid growth of the tumour. Due to the mass effects, the tissue around the active core is usually swollen, leading to an extended edema which is usually also infiltrated by tumour cells. It is best made visible using an $T2$-weighted sequence. Within this work, I am using the Gross Tumour Volume (GTV) definition of Radiation Therapy Oncology Group (RTOG[1], RTOG 0825), which defines GTV as the hyper-intense region within a native T2 or FLAIR sequence. This includes necrotic elements, the active core, and edema.

2.1.3 Intracerebral Haemorrhages

The term Intracerebral Haemorrhages (ICH) refers to a special form of strokes. It is defined as an abrupt, non-traumatic burst of brain vessels and in consequence bleeding into the inner brain. Beside the sudden damage done by the loss of blood supply, additional damage can be caused by space-consuming volume and toxic parts of died blood cells. Different risk for increased probability of ICH are known, like blood thinner therapy, head trauma, tumours, and drug usage, but there are also spontaneous ICH which are not affected to any known risk factor.

ICH causes severe symptoms, including confusion, loss of consciousness, and loss of vision. It is also assigned with a high mortality rate depending on the location of 35 - 52 % within the first 30 days Elijovich, Patel, and Hemphill (2008).

[1] http://www.rtog.org

Two main possible treatments are known. Using a medical treatment, the blood pressure is reduced so it is just high enough to supply the brain. In addition, CSF might be be removed to give space to the haematoma so it can expand without damaging the brain. Sometimes, artificial coma, may be induced. The main target of these therapies is the reduction of the intracranial pressure. The second method a surgical intervention to remove blood. This also reduced the intracranial pressure. It further removes dying blood cells, reducing the risk of damage caused by toxis oddments.

It is not clear which treatment is suited best. Different indication parameter has been evaluated, but no final conclusion is found. A common rule is to measured the blood volume, usually using a fast, diameter-based method, and caring out a surgical intervention if the volume exceeds a given threshold LoPresti et al. (2014). Nevertheless, it is known that this method does ignore important aspects like the CSF volume.

2.1.4 Ischemic Stroke

While ICH based strokes are caused by the loss of blood inside the brain, Ischemic strokes are cause by a block of the blood supply for parts of the brain. This leads to shortage of oxygen and energy, which results finally in the dead of the brain cells previously supplied by the now blocked vessel.

Ischemic strokes are responsible for around 85 % of all strokes. The most common causes for this type of strokes are thrombotic blockades and embolic blockade. The first is caused by plaque within the vessels, which reduce the blood flow and finally blocking the whole transport. The second one is caused if a blood cloth or debris from other parts of the body is swept into the brain and then blocking the vessel.

Ischemic strokes can be treated by medical treatment, using medicals that resolve the blocking. But only a limited number of different medicals is available for this. Another option is an endovascular procedure to remove the blockage manually. The right choice of treatment is not always clear. Since time is critical for the outcome, there is clear need for fast and reliable methods. (Rekik et al., 2012) therefore urges for more automatic methods for the diagnosis, segmentation, and prediction of ischemic strokes.

2.1.5 Pancreatic Tumour

The pancreas is a glandular organ, and is located in the abdomen behind the stomach. It is responsible for the production of several important hormones, for example insulin and glucagon, which are circulated through the blood system. Further, a digestive function is part of the functions of the pancreas. It produces the pancreatic juice, which assists digestion in the small intestine. The pancreas is divided into four parts, the head, the neck, the body, and the tail of pancreas. The whole form of the pancreas is wedge-shaped, with the head being surrounded by two blood vessels and being connected to the small intestine (Figure 2.3).

Worldwide, pancreatic cancer is the seventh most common cause of cancer deaths, with pancreatic adenocarcinoma being the most common type (85 %) (Research on Cancer, 2014). The prevalence for this disease is much higher in the developed world, which accounts for around 70 % of new cases in 2012. It

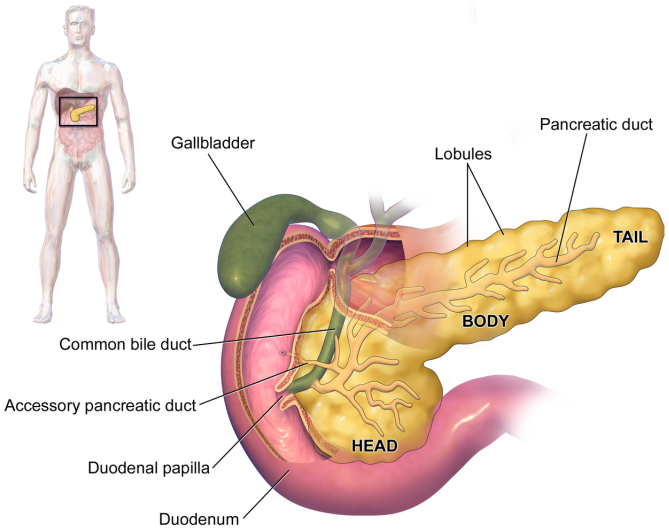


Figure 2.3: Schematic visualization of the anatomy and location of the pancreas. The stomach is not shown for a better visualization of the pancreas. From (Various, 2016b)

does have a poor prognosis with a survival rate of 25 % and 5 % for one and five years, respectively (Research on Cancer, 2014).

The treatment of pancreatic tumours is difficult, with surgical resection being the only curative treatment option today. This interventions does have a high complication rate up to 40 %, which makes accurate characterization of the pancreatic tumour very important (Gouma et al., 2000).However, 'precise diagnosis of pancreatic adenocarcinoma is not always straightforward because they frequently show atypical imaging features and many other diseases may mimic pancreatic' tumours M.-J. Yang et al. (2013) and Coakley et al. (2012). Computer aided diagnosis might help to improve the diagnosis and following the therapy.

2.2 Imaging Modalities

3D imaging is important for the assessment of tissue distributions which are used during diagnosis and therapy. Two of the most common used techniques for this tasks are MRI and Computed Tomography (CT) imaging – both allowing the imaging of an 3D volume.

Although the physical background of these two methods is different, both images are presented using a similar geometric model. Each images is seen as a stack of two-dimensional (2D) images – the so called slices. Depending on the orientation of the axis in relation to the patient, these slices are either axial, coronal or sagittal oriented. Axial slices are stacked from the bottom to the top, coronal from back to the front and sagittal from the patients right side to his left (Figure 2.4).

It is common not to image the whole 3D volume to save imaging time and

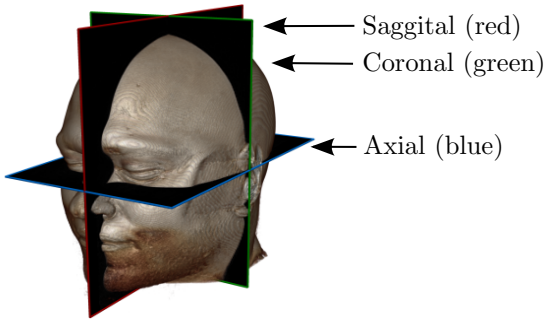


Figure 2.4: Visualization of the orientation name saggital, coronal, and axial used for the description of orientation in radiologic images.

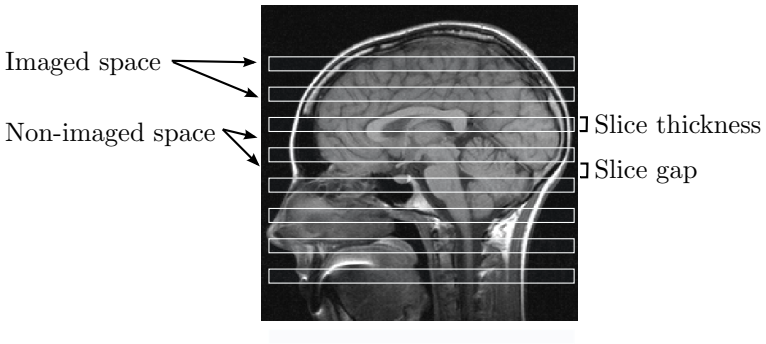


Figure 2.5: Example of an MRI slice, schematically illustrating the meaning of slice thickness and slice gap.

reduce the radiation load. This is achieved by leaving some space between each slice – the slice gap (Figure 2.5). To be still able to apply 3D methods, the imaging volume is usually resampled by combining slice thickness and the slice gap. Due to this and the fact that usually only the slices of a single orientation are analysed, there is often one slice direction offering a good in-plane resolution while the other two do have a more coarse resolution.

2.2.1 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) allows 3D imaging without ionizing radiation with excellent soft tissue contrasts. Using different sequences during a single scan allows further to obtain different contrasts which makes MRI an important imaging modality in clinical routine.

The spins of electrons can be oriented along a magnetic field if the field is strong enough. This orientation can be flipped for specific frequencies[2], using electromagnetic excitation pulses. If this pulse is switched off, the spins re-orientate along the applied magnetic field and during this process, electronic energy is emitted. Beside the strength of the field, the signal depends on the properties of the underlying tissue, for example the proton density and tissue

[2]In medical imaging, typically the frequency or multiple of hydrogen is used.

relaxation time. A more detailed technical description of MRI is given by Bernstein, King, and Zhou (2004)

Contrasts and sequences

An important advantage of MRI is the possibility to measure different properties of tissue. For example, the longitudinal magnetic relaxation time T_1 or the transverse magnetic relaxation time T_2, can be assessed by varying the applied electromagnetic excitation pulses. This variation of excitation pulses is known as 'sequence'. The flexibility of these sequences allows to measure specific properties, suppress specific tissue types like fat, or obtain functional information beside pure anatomical images.

Beside T_1- and T_2-weighted images and the corresponding sequences, Fluid-attenuated Inversion Recovery (FLAIR) is commonly used for brain imaging. By nulling signals from fluids, the resulting images are well-suited to localize brain abnormalities like multiple sclerosis lesions, or brain tumour edema.

Diffusion weighted imaging

Diffusion Weighted Imaging (DWI) allows to estimate the grade of diffusion within tissue. diffusion refers to the natural and random movement of elements that is present in all types of tissues due to intra- and extra-cellular liquid.

If the excitation pule is negated, no signal is emitted for unmoved tissue, but electrons that have been moved due to diffusion will still emit an signal which allows to estimate the diffusion. By using special designed sequences the sensitivity to diffusion can be limited with regard to direction and diffusion range.

To combine the information about the different directions and diffusion ranges, DWI images are often converted into other representation. A common one is the use of tensor matrix that represent the diffusion within each main direction. Compared to the pure representation it has the advantage of allowing more complex features and is easier to visualize. Based on these tensor representation, multiple tissue parameters can be calculated that help to describe the underlying tissue. A more detailed description of underlying physical process, the techniques, and parameters calculated from DWI are given by Basser and Pierpaoli (2011) and Laun et al. (2011).

MRI variability

While different sequences give huge control to the actual measurement of the data, it also adds to the variability of the obtained intensities. Small differences in the implementations of the actual sequence can lead to differences in the final intensity distributions even if the most of the sequences are the same. As there is also no fixed sequence, there are small variations between similar sequences of different vendors and even between different version of the same vendor.

Another source of variability is the magnetic field variability. MRI depends on measuring small changes within magnetic fields. The small signal, the non-linearity of magnetic fields, and the high variability of used materials add to another source of variability within the obtained signal. Other sources of variability are variations of external parameters, like temperature or small magnetic

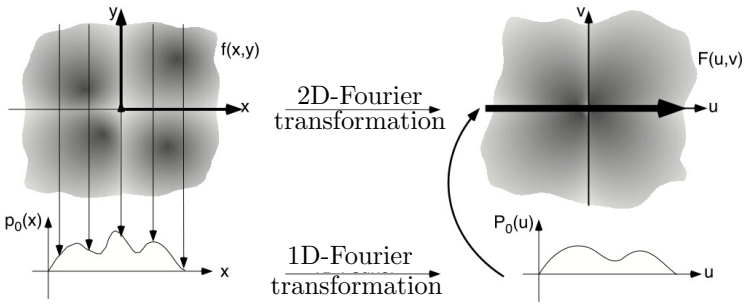


Figure 2.6: Correlation between the Radon transformation, i.e. the projection of a 2D-volume to 1D ($p_o(x)$), and the 2D Fourier transformation $F(u, v)$. Edited reprint from Dössel (2000) with kind permission from Springer International Publishing AG.

sinks, that do influence the magnetic field. There are also variability with each scan, beside the variability between scanners. Variability within the main magnetic field, caused by the magnetic setting, patients position or other factors, can lead to a varying of intensity values within a single image.

Due to that uncertainties, MRI is usually considered as an qualitative imaging method, e.g. the obtained intensity values are non-quantitative. Even if the same subject is scanned, the obtained images usually differ. The differences are usually bigger between scanner of different manufacturer, but even two scanner of the same type will give slightly different images. There are some technical methods to reduce the variability. The devices are usually calibrated, but as this takes relative long and is difficult to do, it is not done regularly. Therefore only really large differences are reduced, smaller changes or changes introduced by a changing environment are not cancelled out.

2.2.2 Computed Tomography

Even though MRI uses tomography techniques, the term Computed Tomography (CT) is used only for 3D imaging with conjunction with X-ray. Using these rays, the projections of X-ray attenuation coefficients μ along all possible orientations are measured. Radon (1917) had shown that these projects can be transformed into the Fourier-transformation of the underlying function of μ. It is therefore possible to obtain the tissue distribution by using the inverse Fourier-transformation (Figure 2.6).

Since CT is using x-rays to measure μ, it inherits some of the properties of conventional x-ray imaging. As the attenuation of X-rays is proportional to μ, CT intensity values are highly correlated between different scanners. To allow a fast comparison of scans that using different energy levels, the intensity values are usually given in Hounsfield units (HU), using the following conversion:

$$[\mathrm{HU}] := \frac{\mu - \mu_{\mathrm{Water}}}{\mu_{\mathrm{Water}}} \cdot 1000 \tag{2.1}$$

By this definition, water is shown with an Hounsfield value of 0, while Air does have -1000 HU. It is possible to give a typical distribution of Hounsfield

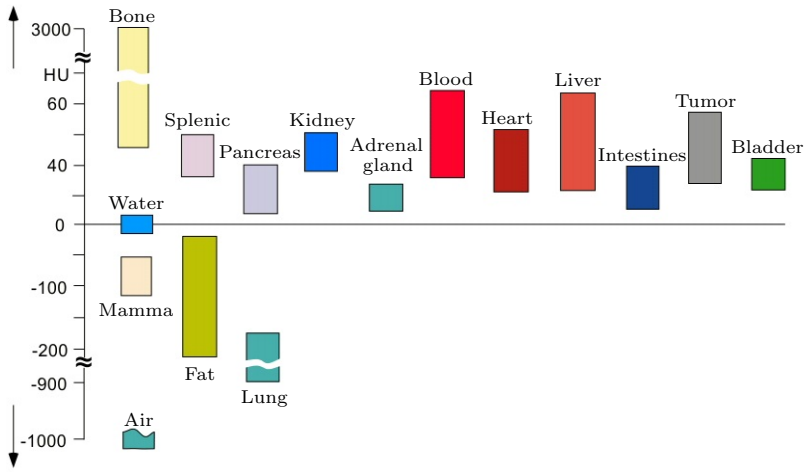


Figure 2.7: Typical HU values for different types of tissues. Edited reprint from Dössel (2000) with kind permission from Springer International Publishing AG.

units per organ, as the distribution of μ is usually known and fix (2.7) which eliminates the need to normalize the resulting images.

X-ray attenuation coefficients are highly correlated with specific tissue types, pure CT is considered an anatomical imaging modality, i.e. pure CT is not specific to any functions like iodine uptake. Different techniques had been developed to increase the capabilities of the CT-technique, some of which are described in the following part. For a more detailed description of CT refer to Hsieh (2009).

Contrast enhanced imaging

While CT contrast are excellent between bones, fat, air, and soft tissue the contrast between different types of soft tissue, like liver, tumorous tissue, spleen, is rather low. It is therefore sometimes necessary to inject contrast agents to increase the contrast between different tissue type, like healthy and tumorous liver tissue.

The actual contrast agent depends on the question at hand, and either increases or decreases the intensity of the up-taking tissue. These, often iodine-based substances, are usually injected in blood vessels as a single bolus.The actual contrast depends therefore often on the time between injection and imaging time.

Perfusion CT

One of the earliest methods that allowed additional function imaging using CT is perfusion CT. It allows to asses blood-flow based parameters, using the idea to measure the intensity change introduced by a contrast agent over time.

In a first step, a reference CT image is taken. After this, the contrast agent is injected and after multiple images are taken at fixed time intervals. The contrast change introduced by the contrast agent is calculated by subtracting the reference image from each time point. This gives a time series of contrast

changes for each voxel. Different models can then be used to extract relevant parameters from these time series.

Two commonly used models are the slope method and Patlak two-compartment model. For the first one, the maximum change between two time points is measured. Diving the maximum slope by the maximum intensity change allow to calculate the blood perfusion (Miles and Griffiths, 2003). For the second method a two-compartment model is fitted to data series. Solving the corresponding differential equations allows to estimate the blood volume and the blood volume at each point (Patlak, Blasberg, and Fenstermacher, 1983). Both models are rather simple compared to other approaches but have proven to be robust in the presence of noise.

Perfusion CT has proven to be useful for different medical applications. Beside the diagnosis of strokes, it is nowadays considered as gold standard for the diagnosis of different types of tumours, like pancreatic carcinoma (Klauss, Stiller, et al., 2012) since the blood uptake of tumorous tissue is usually higher than those of healthy tissue. Beside these advantages, perfusion CT does suffer under an increased radiation dose due to the multiple acquisition of CT images.

More detailed descriptions of the models, the perfusion CT method, and its applications in the medical field are given by Miles and Griffiths (2003) and Sahani (2012).

Dual Energy CT

Other than perfusion CT, Dual Energy CT (DECT) depends on the physical properties of CT. The X-rays used for CT images are usually created using an X-ray tube. The spectrum of the resulting X-rays depends on the tube potential (kV_p) that is used, so it is possible to create different photon energy spectra by using different tube potentials. As the X-ray attenuation coefficients μ depends on the photon energy, the intensity of tissue also depends on the tube potential. It is possible to use these differences by taking two images of the same subject with different potentials.

It is possibility to calculate blended images which resemble image behaviour at tube potentials that were not used during scan time. This can be useful for diagnosis as different tube potentials show different contrast behaviour and noise characteristics. It is further possible to fit models to data to estimate the fraction of specific elements. A common approach is to fit a three compartment model to distinguish between fat, soft tissue and a third material – which is usually the main element of the contrast agent.

Different techniques might be used to obtain these two images. To most simple approach, which does not require specific hardware, is to take sequential acquisitions with different potentials. But this comes with the disadvantage of long acquisition times which can leads to motion artefacts. this can be reduced by using rapid voltage switching. For this, the tube potential is switched for each position during the acquisition. While this allows a speed up of the imaging, it still requires more time than a traditional imaging approach since a small waiting time is required during each switch. So the best approach in term of imaging speed is using two tubes. This allows to take images at the same speed as conventional CT images, but comes at higher hardware cost.

Although the dose of a DECT scan is higher than for an conventional CT, it is still lower than the dose applied during a perfusion CT scan. This makes

it a promising alternative for existing applications, for example for pancreatic tumours. A more detailed description of DECT and potential applications is given by T. Johnson et al. (2011) and T. R. Johnson (2012).

2.3 Software Framework

This work would have not be possible without the availability of public software which allowed to concentrate on the main challenges and avoid the need to implement more basic parts. While this software helped a lot, it also influenced the research, for example in terms of the used technologies. Therefore this section will give a short overview over the frameworks that are mainly used to implement the presented work.

2.3.1 Medical Imaging Interaction Toolkit (MITK)

The Medical Imaging Interaction Toolkit (MITK) Toolkit is an open-source C++ toolkit for the development of medical imaging application. It provides an easy-to-extend graphical user-interface that allows the rapid development of new applications, but command line applications are also supported.

Different platforms are supported by MITK. This includes different Linux distributions, Windows, and OS X. To ensure the compatibility to each platform and to maintain a high code quality, MITK is automatically and regularly build and tested on each of these systems. For this, the infrastructure of the projects allows that own, small tests are written which are automatic executed. Since this thesis is developed as part of the MITK project, these methods were used to increase and monitor the quality of the algorithms.

The core of MITK is the integration of other, widely used C++ toolkits. The two most important toolkits are the Insight Toolkit (ITK)[3] and the Visualization Toolkit (VTK)[4]. Both frameworks are extended to simplify input and output of data, conversion between data formats of different frameworks, and data-type independent algorithm. As most algorithms for medical image processing are implemented either in ITK or VTK they can be easily used within MITK. Beside these two main frameworks, others are also included which are either less general or not specific tailored for medical image processing, like OpenCV[5]. Also, as a part of this thesis, Vigra had been included to support the development of RDF based algorithm.

MITK is using a pipeline concept inherited from ITK. Similar to system theory concept, each algorithm is considered as a 'filter' that is applied to one or more images. The actions carried out by a filter might be very simple, like a threshold segmentation, but can also be very complex and include the execution of different other filter. As this is key concept of MITK, single, clearly separated steps are often refered to as 'filter' even though the corresponding actions are not usually considered as such.

Further descriptions of MITK, the development process and design principles can be found in (Wolf et al., 2005; Nolden et al., 2013).

[3] https://itk.org/
[4] http://www.vtk.org/
[5] http://opencv.org/

2.3.2 Vigra

Vigra[6], short for 'Vision with Generic Algorithms' is a C++ image processing and analysis library. Considering it as an library rather than an framework or toolkit, it does include only few other libraries for input / output operations. In contrast to MITK, there is also no specific target application domain, the whole library is kept generic.

Although Vigra does offer a wide range of different imaging algorithm, the main reason to include it in MITK during this thesis was the RDF support included in this library. This library supports the use of custom splitting function and criteria using template functions. It is therefore possible to adapt the machine learning algorithm and implement new features without the need to change the whole algorithm. Most of the experiments that are described in this work make use of this RDF implementation.

2.4 Machine Learning

Following the definition of Arthur Samuel, machine learning is considered as a 'Field of study [or technique] that gives computers the ability to learn without being explicitly programmed' (P. Simon, 2013). This includes algorithm that can deduce models from data and use these models for predictions on new data. As the algorithms are not problem-tailored but generic, the resulting model depends mainly on the provided data and the used training algorithm but does not include static program instructions. Due to this, machine learning techniques are often considered as 'data driven' approaches.

Depending on the annotation of the training data, machine learning tasks are divided into either unsupervised, semi-supervised, and supervised. Unsupervised machine learning tries to detect hidden structures in the provided data, for example by clustering. It allows insight into the data and might be used to group similar observations. In contrast, supervised learning algorithm train models that are used to predict values based on new observations. This requires that these data are available during the training process – the training data therefore consists of the training observations $\mathbf{X}$ and the corresponding labels $\mathbf{Y}$. Beside supervised and unsupervised learning, it is also possible to improve the training process of supervised algorithm by providing additional, unlabelled data. This scenario is considered as semi-supervised learning.

The data within machine learning tasks consists usually of multiple different observations. Each observations – which might be a single event, a voxel, an element, etc.. – is described by the same features, which are represented usually as an observation-specific vector $\mathbf{x}$. In the case of a traditional supervised learning setting, each observation of the training data does have a specific label y. But beside this traditional way of annotating the data, there are also algorithm that are able to learn models based on data that are annotated in different ways, for example bag-wise.

[6] https://ukoethe.github.io/vigra/

2.4.1 Support Vector Machine (SVM)

Developed by different researches around Vladimir Vapnik (Vladimir N Vapnik and Chervonenkis, 1964; Boser, Guyon, and Vladimir N Vapnik, 1992; Cortes and Vladimir N Vapnik, 1995), Support Vector Machines (SVMs) are now commonly used classifiers. The core idea behind SVMs is the transformation of the data into an higher dimensional feature space which then allows to use a simple linear classifier. This is done by using different Kernels, which allow an efficient calculation of the necessary transformation between these spaces. One advantage of this approach is that it is possible to describe the underlying algorithm within the linear space while still being able to train powerful, non-linear classifier. This simplifies the mathematical analysis of the properties SVMs and supported the wide use and development of SVMs based algorithm.

As already mentioned, SVMs use a linear classifier to predict the the label $\hat{y}$ of an new observation, i.e.

$$\hat{y} = \operatorname{sgn}\left(\mathbf{w} \cdot \mathbf{x} + \mathbf{b}\right) \tag{2.2}$$

with $\mathbf{w}$ being the direction vector and $\mathbf{b}$ the support vector of the decision boundary. Using a loss function that measures that penalize wrong predictions, a regulation term, and the parameter C, the decision boundary is obtained by solving

$$\mathbf{w}, \mathbf{b} = \operatorname{argmin}_{\mathbf{w},\mathbf{b}} \left(\frac{1}{2}\|\mathbf{w}\|^2 + C \cdot \sum_i \operatorname{loss}\left(y_i, \mathbf{x}_i\right) \right) \quad . \tag{2.3}$$

A common used definition of the loss function is the so-called 'hinge loss', which gives an error that is proportional to the distance of the wrong prediction:

$$\operatorname{loss}_{\text{hinge}}\left(y_i, \mathbf{x}_i\right) = \max\left(0, 1 - y_i\left(\mathbf{w} \cdot \mathbf{x}_i + \mathbf{b}\right)\right) \quad . \tag{2.4}$$

Being able to replace the hinge loss function with other loss functions allows to adapt the training algorithm fast to other settings. This is a main reason why the SVM is commonly used to propose training algorithm for non-traditional learning settings like bag-wise annotated data.

2.4.2 Random Decision Forest (RDF)

Based on the observation that the combined prediction of multiple, weak classifier results in a strong classifier, Random Decision Forests (RDFs) are a combination of multiple decision trees. While a single decision tree often fails to capture all variations of complex classification problems, an ensemble of weak classifiers produces stable results that comparable to those of more complex classifiers like SVMs if the correlation between the weak classifiers is low enough.

First introduced by Tin Kam Ho (T. K. Ho, 1995; T. K. Ho, 1998) they were made popular later under the name 'Random Forests' by Leo Breiman (Breiman, 2001). While the mathematical assessment is more difficult, RDFs does have some significant advantages. The rule-based classification scheme of these classifiers allows further insight into the data – including the estimation of feature importance. Further, RDF are known to have an internal feature selection and therefore are less prone to meaningless or noisy features. It is therefore possible to omit a previous feature selection stage and also use features which are only weak indicator.

To obtain good classification results it is important to have non-correlated decision trees within an RDF. Beside the variations in the training of the trees, which are explained later, this is achieved by training each tree of the ensemble only on randomly selected parts of the available training data, i.e. to use bagging. If this is done, the prediction of each sample is simply obtained by taking the majority vote of all trees.

Although other methods exist, decision trees used in RDFs are usually trained using a bottom-up training method (Figure 2.8). Starting from a root node with all training sample the same training method is iteratively used for all nodes. The data S are split into two groups S_1, S_2 and then send to the left and right child node. To find the best split function the gain function G is optimized over the split:

$$G = I_G(S) - \frac{|S_1|}{|S|} \cdot I_G(S_1) + \frac{|S_2|}{|S|} \cdot I_G(S_2) \ . \tag{2.5}$$

The most common impurity function I_G that is used to calculate the gain is the Gini impurity:

$$I_G(S) = \sum_c f_c(1 - f_c) \tag{2.6}$$

with f_c being the fraction of elements with the label c within the dataset S.

This iterative process is repeated until a stopping criterion is fulfilled. These stopping criteria are usually chosen to prevent an over-fitting of the trained classifier, e.g. to prevent a pure learning of the given data, as this would lead to noise sensitive classifier. Typical stopping criteria for decision trees include a minimum gain improvement, a minimum number of observations at a leaf, only one class left, or a maximum tree depth.

While traditional decision trees use an extensive search for the best split function this is usually omitted during the training of random forests. Instead of evaluating all possible splits, usually a fixed number of random splits for a limited number of features is evaluated. This reduces the training time and at the same time introduces randomness, reducing the correlation between the trees. The choice of number of possible splits, and the number of features that is evaluated gives a trade-off between tree correlation and tree classification power.

Even if always the best split is selected at a node with regard to the data at this node, there is no guarantee that this split is the optimum for the overall decision. It might be more beneficial to select a less favourable split in order to have a better second split later. The training of decision trees is therefore considered as a greedy optimization of the chosen gain function. It makes it also impossible to use gain functions that depend on properties of the whole data set, since each node does only have access to limit data.

2.5 Notation and conventions

Within this work a common way of mathematical notations is used. One-dimensional variables are written in italic: a,b. Data, that represents vectors or matrices is named using bold letters, for example $\mathbf{x}, \mathbf{k}$. Sets are typeset using calligraphic font: $\mathcal{R}, \mathcal{X}$.

Classifiers are commonly trained using a set of training data. The set consists of different elements, the name ‘observation’ or ‘sample’ for them are used

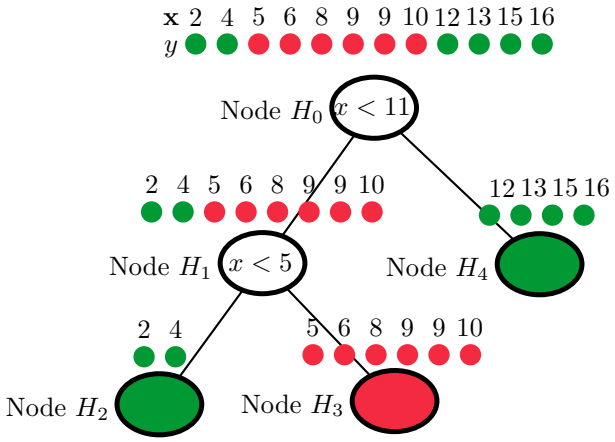


Figure 2.8: Example of a decision tree. During the training, the splitting function, in this case a simple threshold, is determined for each node. For node H_0 the threshold is set to 11. All observation with x below this value send to the left child node H_1 and the other to the right one. This is repeated until a leaf node is reached. For the prediction, the same process is repeated. Based on the thresholds, the observation is send either to the left or right child until a leaf is reached which then determines the label.

interchangeable. As this work is about voxel-based classification, each observation represents represents usually a single voxel of an image.

During the experiments, mostly 3D-images are used. Due to that the term 'voxel' is used instead of pixel. This does not imply that this work is limited to 3D images. Similar, the terms 'voxel classification', 'segmentation', and 'semantic segmentation' are interchangeable. They reflect the same task, but are used in different scientific communities.

3

State of the Art

3.1 Introduction

Multispectral imaging becomes more and more common within the clinical routine as MRI is used more often. Being able to acquire different contrast during a single scan allows to gather more information. The ability of computers to combine information from different sources early led to the idea to let computers learn to characterize tissue for clinical applications. One of the first works based on these ideas was published by Vannier et al. (1985) who applied methods known from satellite image processing. They showed that these systems are able to differentiate between multiple classes and found typical values for different tissue classes.

The complementary information that is present in the different spectres of multispectral systems allows to make better predictions. Chan et al. (2003) compared the power of single-spectrum classifier trained versus multispectral classifier and showed a large improvement in the classifier accuracy. They concluded that multispectral systems are superior to single-spectrum systems. Similar findings of Ampeliotis et al. (2008) and later Ozer et al. (2010) further support these conclusions.

Due to this reason multispectral tissue characterization is a lively research subject. As Figure 3.1 shows, the number of publications with this subject are still increasing year after year. This research covers a wide field of medical questions. For example, Keller et al. (2011) presented an algorithm to detect the breast density in digital mammograms showing multispectral tissues, also including non-imaging data. Similar Jacobs et al. (2003) used clustering methods but combined it with semi-automatic segmentation methods to detected and classified breast lesions.

Other applications are developed for further organs, like prostate X. Liu and Yetik (2011), pancreas (Klauss, Lemke, et al., 2011), or liver Z. Wang et al. (2015). While there is no limitation regarding the area, there is less research in areas that is affected by respiration movement. Overall, the most approaches

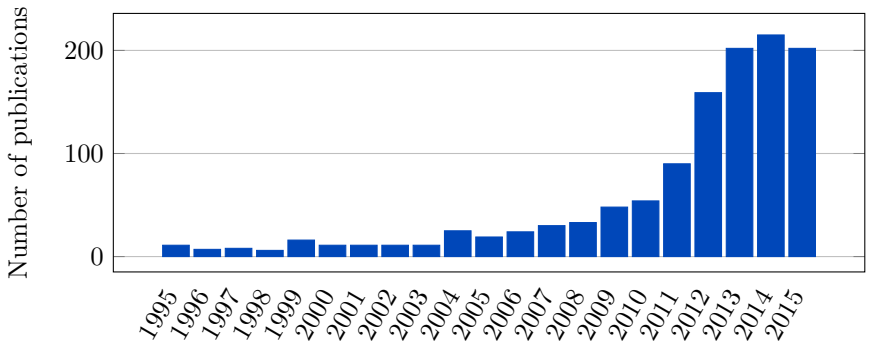


Figure 3.1: Number of publications about multispectral tissue characterization as found by medline (For full search term see *Medline Trend* (2016)). Note that these numbers give only a rough overview. Many publications, like the BraTS contributions, are not indexed by Pubmed.

are presented for brain imaging. Main reasons for this are the availability of large data collections, as well as the reduced and rigid movement between two spectres Angelini et al. (2007). This is further improved by presence of large and readily processed challenge dataset like the brain tumour segmentation challenge BraTS (BraTS, 2016). Most of the methods that will be discussed within this work are therefore dealing with problems from this area.

Due to the vast amount of work that is done in this area it would exceed this work if every work is discussed in detail. Therefore the focus will be on the general aspect and give some examples or name the most import papers. A more detailed review of the current literature is given in topic-specific review papers, for example by Clarke et al. (1995), Bauer, Wiest, et al. (2013), and Lladó et al. (2012).

To simplify the following, the focus will be on systems that work locally, i.e. systems that predict the spacial (voxel-wise) distribution of tissue characterizations. Nevertheless, most of the following is also true for global systems, i.e. systems that predict patient- or organ-wise characterization. These systems will then be discussed in more detail in section 3.5.

The methods that are used for learning based multispectral tissue characterization methods can be split into two groups (Havaei, Davy, et al., 2016). On the one side there are approaches using generative models that allow to sample different tissue classes. These methods often make use of domain-specific knowledge and are therefore usually problem tailored. A common approach for such systems is to learn the appearance of healthy tissue and then to detect abnormal tissue. Clark et al. (1998) used unsupervised clustering to group the voxels of brain MRI scans. Based on rules learned from training patients, the clusters are then assigned to different tissue classes. Prastawa, Bullitt, S. Ho, et al. (2004) used outlier detection to find brain tumours. They learned the appearance of healthy brain tissue and labelled outliers from these models as tumorous areas. These areas are then further divided using a rule-based approach. Another approach using data of healthy volunteers was proposed by Parisot et al. (2014) who proposed a framework for concurrent registration and tumour segmentation for brain images. The basic idea of their approach was to

iteratively register a brain atlas to the given image and marking mismatching areas as tumour. A similar approach was taken by Kwon et al. (2014) who registered pre- and postoperative brain images to identify areas of reoccurring tumour. Other approaches, like B. H. Menze, Van Leemput, et al. (2010), combined tumour growth models and tumour segmentations.

On the other side there are discriminative approaches. Usually, such approaches incorporate less prior knowledge into the algorithm and rely on low-level features. This makes these approaches more general and simplifies the adaptation to new organs or tumour types. For example, D'Addabbo et al. (2003) trained a SVM to detect Multiple Sclerosis Lesions in MRI images and Ruan et al. (2007) trained a similar classifier to segment brain tumours.

These systems often follow the same pattern (Figure 3.2, Bauer, Wiest, et al. (2013)). The core of such algorithms usually consists of classification algorithms, and a wide range of different algorithms are used. Beside SVM, RDF and Neuronal Networks are especially common for multispectral tissue characterization. The use of RDF was particularly boosted by the good results that were obtained on the 2012 BraTS challenge using these algorithms. Especially the work of Zikic, Glocker, Konukoglu, et al. (2012) led to a great interest in this technique. They proposed to use a nearly infinite feature space by calculating the features at each node using random parameters. But other approaches, that used RDF based algorithms, like Kleesiek, Biller, et al. (2014) and Meier et al. (2013) obtained very good results with these classifiers. It is therefore often assumed that these classifiers are especially well suited for the task of tissue characterization in medical images. Although there are different implementations of RDF classification algorithms, most approaches make use of traditional, canonical implementations. For example, Meier et al. (2013) and Kleesiek, Biller, et al. (2014) who scored best or second on the BraTS challenge, both used canonical implementations. The approach of Zikic, Glocker, Konukoglu, et al. (2012) has been slightly different. Dealing with nearly infinite features they implemented a structure that is closer to Extremely Randomized Trees (Geurts, Ernst, and Wehenkel, 2006). This algorithm proved to achieve superior results compared to the canonical implementation. But neither Zikic, Glocker, Konukoglu, et al. (2012) nor others evaluated the influence of their algorithmic choice to the final result.

In the last years, Deep Neuronal Networks became more and more common in tissue characterization, for example (Havaei, Davy, et al., 2016; Pereira et al., 2016; Kleesiek, Urban, et al., 2016). Beside the good results that are achieved using these methods, this trend is also influenced by the recent progress in computer vision community (Krizhevsky, Sutskever, and Hinton, 2012; K. He et al., 2015). Other, less common classifier include Logistic Regression, Gaussian Mixture Models, and Hidden Markov Models (Lachmann and Barillot, 1992; Tian et al., 2011; Solomon, Butman, and Sood, 2006). To give an impression of the different algorithms employed, the methods that have been used in the BraTS challenge from 2012 to 2015 are listed in table 3.1.

There is no fixed set of features, but rather a wide range of different image features are used. Beside the pure intensity values that were used, for example by Havaei, Jodoin, and Larochelle (2014). Other common features include local histogram (Kleesiek, Biller, et al., 2014; Meier et al., 2014), texture filter (N. K. Subbanna et al., 2013; N. Subbanna, Precup, and Arbel, 2014), Haar-like features (Zikic, Glocker, Konukoglu, et al., 2012), and alignment features

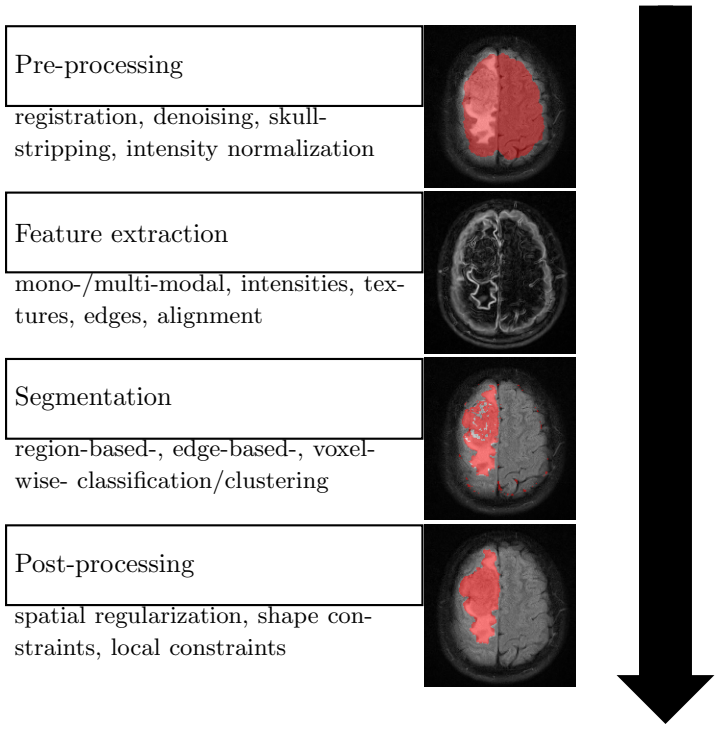


Figure 3.2: Illustration of the main blocks used for building up the segmentation pipeline of most algorithms used for tissue characterization as defined by Bauer, Wiest, et al. (2013). (Figure adapted from (Bauer, Wiest, et al., 2013))

Schmidt et al. (2005). In addition to the calculation of the features with the original image resolution, these features are sometimes also calculated after denoising or downscaling the images Kleesiek, Biller, et al. (2014) and Schmidt et al. (2005). With the advent of deep neuronal networks, the feature design and selection is more and more replaced by feature learning with deep neuronal networks Krizhevsky, Sutskever, and Hinton (2012).

Several challenges dealing with multispectral tissue characterization have been started during the last year, for example the brain tumour segmentation challenge BraTS (2016), the ischemic stroke lesion segmentation Isles (2016) or MR brain image segmentation Neat (2016). These challenges have increased the visibility of different approaches, and allow a comparison of different algorithms for the same task. But as the challenge datasets are often preprocessed and already labelled, the main focus of research is now focused on the feature generation and the following steps. Other difficulties are therefore less investigated.

3.2 Classification pipeline

A special challenge for algorithms that use MRI data is the qualitative imaging obtained with this technique. There is no fix correlation between intensity values

TABLE 3.1
Approaches used at yearly Brats-Challenge

Method	2012	2013	2014	2015
Generative model based	6	2	1	2
Random Forest	3	4	4	2
Neuronal Networks	0	0	3	6
Other Classifier	4	3	0	1

Table with number of different approaches per year using a specific technique. The numbers are based on the corresponding proceedings for these years (B. Menze, Jakab, et al., 2012; B. Menze, Reyes, Jakab, et al., 2013; B. Menze, Reyes, Farahani, and Kalpathy-Cramer, 2014; B. Menze, Reyes, Farahani, Kalpathy-Cramer, and Kwon, 2015).

of an image and the underlying tissue. Therefore, the normalization of these intensity values is an ongoing area of research. But as most algorithms for tissue characterization depend on fix intensity values, the applied MRI normalization does have a huge effect on the final result. This is also shown by the work of Collewet, Strzelecki, and Mariette (2004), who studied the relation between normalization methods and different texture features. They found that not only the final feature values, but also the stability of a feature are heavily dependent on the used type of normalization. Therefore, they conclude that care should be taken which normalization method is chosen.

Mainly two types of intensity normalization are used. First, a one-to-one or global mapping of intensity values, i.e. $I' = f(I)$. The new intensity value I' solely depends on the original intensity value I. These methods have proven to be able to remove most of the variance between different image. The second, commonly used technique consists of a local mapping, $I' = f(I, \mathbf{x})$ where the new intensity value additionally depends on the position $\mathbf{x}$ in the image. This is commonly used to remove bias field effects. For more details, refer to chapter 2.2.1. More complete reviews of normalization techniques are given by Vovk, Pernus, and Likar (2007), Balafar (2012), and Alizadeh et al. (2015).

In the field of tissue characterization histogram based approaches are mainly used for the global normalization. Common approaches that are used are those of L. Wang et al. (1998), Nyúl, Udupa, and X. Zhang (2000), and Hellier (2003). Common to these approaches is the first calculation of an intensity histogram and the matching of these histograms. The main difference is the degree of freedom and type of transfer function that is used to match the histogram to a common value space (Sun et al., 2015).

Beside the global normalization, an additional bias field correction is commonly applied. The most common algorithm for this purpose are the N3 or N4 algorithms (Nicholas J Tustison et al., 2010; Sled, Zijdenbos, and Evans, 1998). But the choice of this algorithm is usually not justified nor is any effect reported. It further seems that the effect also depends on the dataset that has been used. For example, most algorithms that were evaluated on the BraTS data used a bias field correction, while it is far less common for other algorithms.

These combinations lead to a huge variation of normalization algorithms. Zikic, Glocker, Konukoglu, et al. (2012) used features that are based on the differences of intensity values. They argued that it is therefore not necessary to normalize the intensity values in advance. Although they obtained significant results, they provided no further data to prove this assumption.

Nicholas J. Tustison et al. (2015) and based on this Havaei, Davy, et al. (2016) used a simple interval normalization, i.e. matching the $[0.01, 0.99]$ data interval to the value range $[0, 1]$ linearly. After this step an additional bias field correction was applied. Nicholas J. Tustison et al. (2015) reported that they evaluated the more sophisticated normalization technique proposed by Nyúl, Udupa, and X. Zhang (2000) but obtained slightly better results with the proposed algorithm. They did not mention the difference between the two normalization techniques, if they tried other normalization algorithms as well or if the differences were statistically significant.

Contrary to these approaches, Kleesiek, Biller, et al. (2014) used a complex normalization process. They first used histogram matching to ensure similar value areas and the classified CSF areas using a simple, intensity based RDF classifier. Based on this result, the mean intensity values of CSF are matched to 0 and the overall mean to 1 using a linear transformation. Finally, N4 bias field correction was applied to the data. The proposed algorithm scored second on the 2014 BraTS challenge, and the authors claimed that a main reason for this was the extended normalization process. But they provided no further research to support their claim.

Based on the information provided by these approaches, it is difficult to judge whether a complex or a simple approach is better suited for the normalization. Even though the three previously mentioned normalization approaches are used for the same task, they are quite different. The authors agree that normalization is important but fail to evaluate these more closely as they were focusing more on the feature design process. This is not uncommon as most paper are only shortly reporting which normalization was used and even often fail to cite the given method (Shinohara et al., 2014).

Verma et al. (2008) reported that they used histogram matching to normalize the intensity values, without any additional bias field correction. But they did not mention which algorithm they used for the histogram matching, how they created the template histogram or if they tried any other algorithm.

These examples show the huge difference in MRI normalization that are currently used. Although it is widely accepted that the normalization is important, little research is done to investigate the best combination. Results from normalization of healthy subjects can not easily be transferred to patients with pathologies. For example, Corso et al. (2008) reports that normalization failed for some patients with severe glioblastoma due to the violation of some assumptions made for the normalization algorithm. Although they realized these points they haven't investigated this any further and also proposed no solution.

An exception is the work of Shah et al. (2011). They evaluate the influence of different normalization techniques for the segmentation of Multiple Sclerosis. For this, they trained three different classifiers with three different normalization techniques and reported their results. But they only use Bayesian classifiers and an outlier detection approaches. These classifiers are weak compared to SVM or RDF based approaches. Furthermore, their findings cannot easily applied to other tasks, like brain tumour segmentation, because the size of the lesions is

very different in both cases.

3.3 Pre- and post-training data selection

It is well known in the field of machine learning that the quality of the training data directly relates to the detection accuracy of the trained classifier (Hastie, Tibshirani, and Friedman, 2013). Cortes, Jackel, and Chiang (1995) showed a mathematical proof that the performance of a classifier is bounded not only by the amount of training data but also its quality. This was validated by multiple experiments, for example, by Sheng, Provost, and Ipeirotis (2008). This includes not only correctly labelled training data but also training data that is representative for the labelled data.

As medical image data is often affected by labelling noise this is an important fact for medical data. Nevertheless, there are only few approaches to increase the quality of the data. Probablly the most common approach is the fusion of labels from multiple raters. Warfield, Zou, and W. M. Wells (2004) proposed Simultaneous Truth and Performance Level Estimation (STAPLE) which allows the combination of multiple segmentations with regard to the accuracy of each segmentation. Even though this increases the quality of the labeling it does not ensure correctness and adds to the labelling cost.

Kleesiek, Biller, et al. (2014) used a similar approach for their BraTS challenge contribution. They claimed that the label quality of the provided training data was not sufficient enough and therefore manually relabelled all provided training data. Using this new data and a rather simple algorithm they achieved high scores during the challenge. As this shows the importance of the training data, it does not solve the problem of false labelling, or incorrect training data.

Only few algorithmic solutions beside the (re)labelling of the data has been employed in the field of medical imaging. Zikic, Glocker, and Criminisi (2013) proposed to train decision trees on a single image and then combine them later, similar to the decision trees in the canonical RDF algorithm. But as they combine the predictions of each tree by majority voting, the effect of falsely labelled data is only reduced and does not vanish.

Using simulated data for the training is another solution. Prastawa, Bullitt, and Gerig (2009) proposed an algorithm to simulate MRI images of brain tumours and suggested to use these images for the training and validation of segmentation algorithms. But since the datasets were not similar enough, no simulated brain images were used for any training.

Similar to this, Heimann et al. (2014) showed that he was able to train a classifier for ultrasound transducer using synthetic training data. But this approach is not easily adaptable to the field of tissue characterization where data is often more difficult to be labeled.

While most of these approaches are suitable to reduce the label noise, they do not indicate which training images are suited best. There are some approaches in the general computer vision community that deal with the challenge of inhomogeneous training data. Most of these approaches depend on a high level description of the image which is obtained by calculating different features of the image. For example, C. Liu, Yuen, and Torralba (2011b) proposed an extension of the well known SIFT operator (Lowe, 1999) which is well suited to detect similar frames within a video stream. Also very well known are histograms of

oriented gradients (HOGs) as proposed by Dalal and Triggs (2005) for the detection of humans. Other approaches make use of the mutual information that is also common for image registration (W. Wells et al., 1996).

These features are then used to asses the similarity between two images. C. Liu, Yuen, and Torralba (2011a) proposed to find similar images based on the euclidean distance between the features. A previously unseen image is then labelled by transferring the labels from similar images. The whole system is developed for the fast labelling of video streams. As the system focuses on the labelling process with many different labels, the main task mainly was to find images which contain similar labels, rather than the ones that look similar. Hays and Efros (2008) used a similar idea as C. Liu, Yuen, and Torralba (2011a) but used clustering instead of finding a fix number of closest neighbours. They identified then the similarity of unlabelled images with each cluster. The main aim of their work was also not to find the best training images but rather to identify the location where the image was originally taken.

One of the first papers that pre-selected the most similar training images before training a new classifier was proposed by Russell et al. (2007). Based on the L_1-norm the closest images where identified using different high level features. Based on these similar images they decided which pre-trained SVM classifier was used. Therefore, the image similarity was mainly used to predict possible labels in the images.

The approach of Tighe and Lazebnik (2013) contained the combination of multiple features and a more complex rule-based retrieval scheme to find the closest images. The individual set of training images was then used to segment the new image. This is done by dividing the image in superpixels and label each superpixel based on a local feature set.

While these approaches are used to find similar images, the similarity is usually defined only as a fixed metric – e.g. the euclidean distance between the two feature vectors. It is not evaluated if some features are more important than others. Also, non of these approaches are using the closest neighbours to train a new classifier. It is therefore difficult to combine them with other approaches that are more focused on a single classifier. This is not surprising, as most similarity based approaches are more focused on reducing the labels rather than finding similar training images or reduce the label noise.

3.4 Reduced annotation effort

As a consequence of the annotation difficulties and uncertainties, the annotation times that are necessary to create a training base are very high. For example, Mazzara et al. (2004) report that it takes on average about 30 minutes to create a complete, clinical annotation of a brain tumour, but the annotation is then associated with a high degree on uncertainty, with an inter-operator variability of around 28 %. Creating more accurate segmentations does require even longer annotation time. It took, for example, on average four hours to create the annotation of a single patient of the BraTS challenge but still achieved only an average Dice score of 80 % (B. H. Menze, Jakab, et al., 2015). Other studies report similar rater variabilities and labelling times (Deeley et al., 2011; Weltens et al., 2001; Porz et al., 2014).

The long labelling times make it time intensive and expensive to create a

new training database. A common approach is therefore to use the freely available data of a challenge. For brain tumour segmentation, the BraTS challenge became popular (BraTS, 2016), and with the recently started Isles challenge a collective for ischemic stroke segmentation is publicly available (Isles, 2016). Other challenges, like the Neat challenge on healthy brain tissue segmentation (Neat, 2016), are listed on the Grand Challenge Website (Website, 2016).

But while challenges allow to verify and compare algorithms, the proposed datasets usually reflect only a very specific setting. As clinics can have their own MRI protocols, the available contrasts might be different from those present in the training data. There is also a high variability between images of different scanners (Braithwaite et al., 2009). It is therefore still necessary to create a specific training set for each clinical setting.

Another solution is using support tools during the annotation process. Beside semi-automatic segmentation approaches that allow to reduce the annotation time by incorporating previous knowledge, these tools are usually designed to increase the labelling quality. For example, Pedoia et al. (2012) published a guideline and a tool how to create annotations of brain tumour and showed that their combination helps to reduce the inter-operator variability. Another approach published by Warfield, Zou, and W. M. Wells (2004) is the combination of annotations from different raters with an Expectation - Maximization (EM) algorithm. But the use of such tools incorporates some kind of modal assumption on training data. For example, if the labelling was done with an interactive, RDF-based tool, it is very likely that other classifiers might perform worse, even if they would be better suited.

Other solutions try to reduce the amount of necessary training data. This is especially common for approaches using generative models, where training data can be replaced using problem specific data. Prastawa, Bullitt, S. Ho, et al. (2004) used atlas-based information in combination with an outlier detection system which allowed to use only the data of healthy patients during the initial training state and only use a minimum of training samples for tumour. Kaus et al. (2001) compared the unlabelled images with an atlas and labelled differences between both as tumours. Similar, Parisot et al. (2014) combined registration to an atlas and segmentation of brain tumours within a single algorithmic step, and therefore no longer required labelled brain tumour data. B. H. Menze, Van Leemput, et al. (2010) developed an algorithm that made use of simulated tumour growth patterns. But with the incorporation of task-specific prior knowledge, these approaches cannot easily be applied to other settings, for example liver tumour segmentation. For example, registration to a liver atlas would be very difficult due to the high variability of shape and vessel structure within the livers.

Few discriminative approaches have been published that tried to reduce the annotation time and uncertainty. Verma et al. (2008) trained a Bayesian classifier for intra-patient segmentation and a SVM for inter-patient segmentation using partial annotations. They marked areas between two tissue classes or ambiguous areas as unlabelled. But they did not investigate if this labelling had any influence on the classifier performance or if they should correct for any errors made by this labelling scheme. And as their labelling scheme still required the labelling of nearly all the data, the labelling time is still significant.

Another solution would be the use of simulated data as, for example, done by Heimann et al. (2014). But as previously mentioned, this is difficult to do for

tissue characterization as it requires the capability to create an artificial image of tissue which is not completely possible by now (Prastawa, Bullitt, and Gerig, 2009).

These solutions do not solve the problem of data annotation for tissue characterization. It is still either time-consuming and error prone to generate the necessary training data or domain-specific knowledge is necessary. This increases the amount work that is necessary before an approach can be used for other tasks as well.

From a more technical point of view there are some approaches to this problem that should be discussed in the following.

It is well known that it can be difficult to acquire a representative training set, especially if the data involve humans. Studies – often conducted on universities – are more likely to include students or people working at universities than other people. The results therefore need to be adapted to the general distribution of people. Based on this findings, James J Heckman (1977) and James J. Heckman (1979) proposed an new idea, the so called 'sampling bias'. This idea, which was later awarded a nobel price, concludes that it is possible to estimate the real results, even if the dataset that is used does not reflect the true data distribution.

This idea has later been followed by Shimodaira (2000) in a much-noticed work. He proposed to compensate the sampling bias of a training set by weighting each observation with an unique weight. He showed that it is possible to estimate the weights so that the sampling bias is compensated. This allowed to train basically all classifiers using biased datasets. This technique was then successfully used in many different projects of different research fields as it allowed to omit the annotation of new training data if training data from other areas could be used. Examples that make use of this technique are the work of Gopalan, Ruonan Li, and Chellappa (2011) who trained classifiers to categorize images based on content. Instead of creating a new labelling approach, they used an existing dataset and adapted it to the new image source.

Similarly Bickel, Brückner, and Scheffer (2007) showed that spam detection can be improved by adapting the existing and labelled spam database to the currently used emails. If this technique is not used, the changes made by the spammer would otherwise reduce the effectiveness of the spam detection algorithm. Similarly, this technique is used for natural language processing to adapt the text recognition to an individual speaker and therefore improve the detection quality (Pan, Tsang, et al., 2011). A more complete review of this technique, the various fields of applications and variations of this technique can be found in a the reviews of Pan and Q. Yang (2010) and A. Margolis (2011). But this technique is mainly used to adapt existing training data to new data. It still requires training data of similar appearance, which makes it difficult to use it, for example, if the imaging modality changed.

A different approach is the partial annotation of training data. Denis (1998) introduce the idea of learning from only positive data annotations. He discussed that this is especially helpful if one label class is hard to identify within the training data – either due to similarities between two label classes or the rare occurrence of this type of observations. Other reasons might be that it is technically not possible to identify such labels. He suggested the name 'Positive and Unlabelled Learning (PU-learning)' for this kind of training setting.

This idea was used within several applications from different fields. Portnoy,

Eskin, and Stolfo (2001) used PU-learning to train a classifier for intrusion detection in networks. Since intrusions are difficult to differentiate from regular network traffic, the labelling of such events is very time-consuming. Using only positive samples still allowed to train a classifier. Similar Cerulo, Elkan, and Ceccarelli (2010) used the idea of PU-learning to train a classifier to detect gene regulation effects. Labelling all genes would be impossible, so he only labelled genes that were known to be responsible for gene regulation and left the other genes unlabelled. Other fields that benefited from the reduced annotation effect are text classification Peng, Zuo, and F. He (2008) or remote sensing image processing W. Li, Guo, and Elkan (2011). A more complete review of the different fields of application and different techniques used for PU-learning can be found in De Comité et al. (1999) and Letouzey, Denis, and Rémi Gilleron (2000). Most of these applications use PU-learning if the ratio between two classes is large, e.g. if one class is rare and therefore difficult to be labeled. Within the field of medical imaging, this technique is still not very common, and rarely used.

Not surprisingly these widespread possibilities of application led to the development of further algorithms beside the originally proposed techniques of Denis (1998). These techniques differ not only in the assumptions made but also in the underlying classification algorithms.

B. Liu et al. (2002), for example, proposed an algorithm for PU-learning using an EM based algorithm to find the parameter of a data model based on positive labeling only. Lee and B. Liu (2003) used an adapted version of logistic regression for classification. The main idea of their algorithm was the introduction of a weighting factor, adapted to the setting of PU-learning and logistic regression. Adaptations of RDF-based algorithms were proposed by Liang et al. (2012) and De Comité et al. (1999). Even more common are approaches that are based on SVMs as this technique is easier to be described mathematically. Algorithms are published by X. Li and B. Liu (2003), Hwanjo Yu, Han, and Chang (2002), and Denis, Remi Gilleron, and Tommasi (2002) for example.

The downside of these algorithms is the limitation to a specific type of classifier. It can therefore be challenging to include this technique in an existing approach. Therefore Elkan and Noto (2008) proposed a technique that is based on cost-sensitive learning which allows to use PU-learning with basically all classifiers.

PU-learning has so far not been used for multi-modal tissue characterization. Prastawa, Bullitt, S. Ho, et al. (2004) proposed an approach using an outlier detection technique to find differences between healthy and tumorous tissue. But as shown by Manevitz and Yousef (2001), outlier detection or one-class learning tends to be weaker than PU-learning algorithms due to the reduced amount of information that is used. The difference between one-class learning and PU-learning became even bigger with the results obtained with the algorithm proposed by Hailong Yu, Zuo, and Peng (2005).

An even further reduction of the labelling process can be obtained by using Learning form Label Proportions (LLP). In this setting, the training data is grouped, for example by geographic locations or by different measurement runs Patrini et al. (2014). This rather recent learning setting can be seen as an extension of the Multi-Instance Learning (MIL) setting where the data is also grouped, but it is just known if a group contains positive samples or not. Kuck and Freitas (2005) introduced the LLP setting and proposed to use a hierarchical

model that is sampled with an Markov Chain Monte Carlo (MCMC) algorithm to find the final parameters.

Later, B.-C. Chen et al. (2006) adapted the idea of Kuck and Freitas (2005). They used a slightly different model by enforcing self-consistent labels. Furthermore, they did not used a MCMC algorithm but instead evaluated different classification algorithms and obtained improved results compared to the original work of Kuck and Freitas (2005). This work was then carried on by the work of Musicant, Christensen, and Olson (2007). In 2009, Quadrianto et al. (2009) proposed a new approach. Instead of modelling the data structure they proposed to estimate the mean of each class and incorporate this information in the loss function. They incorporated this loss function into a SVM algorithm and showed that they were able to improve the results of previous LLP algorithm. A similar idea was proposed by Rueping (2010) who proposed to use soft labels by using the bag-wise label probability as labelling information. By this, they transformed the LLP to a regression problem which they solved using SVM classifiers.

Other SVM based solutions were proposed by F. Yu et al. (2013) and later Patrini et al. (2014). F. Yu et al. (2013) created a theoretically founded approach which outperformed other solutions at that time. But later Patrini et al. (2014) extended the idea of Quadrianto et al. (2009) by estimating the mean operator using a manifold regularization technique. This led to further improvements of these techniques.

While approaches using SVM algorithms still seem more powerful than other approaches, there are also other approaches to solve the LLP problem. K. Fan et al. (2014) used a Bayesian interpretation of LLP and solved it using Restricted Boltzmann Machine (RBM). J. Hernández and Inza (2011) and Hernández-González, Inza, and Jose A Lozano (2013) used structural learning of a Bayesian network with missing data for the same purpose. Beside these ideas, S. Chen et al. (2009) and later Stolpe and Morik (2011) proposed to cluster the data in advance. The clusters are then assigned to the labels such that the proportions of labels match the known true label proportions. Using these simple labels they can then train traditional classifiers. But it was later shown by K. Fan et al. (2014) that some assumptions they made can lead to completely false results.

None of the better working algorithms can be used independently from the classifier. While B.-C. Chen et al. (2006) evaluated some decision tree based classifiers, recent results significantly outperform these results. Aside from these results, no approach that could be used in conjunction with RDF based algorithm was proposed. This makes it difficult to use this technique for multi-modal tissue characterization algorithms that are based on these classifiers.

3.5 Assessment of image acquisition

A common question in medicine is the suitability of a specific modality or combination of modalities for the assessment of information. Knowing which modality can be used allows to avoid unnecessary imaging.

A common approach to measure the information within a medical image are Region of Interest (ROI) based approaches (Ali et al., 2008). For example, regions covering healthy and tumorous tissues are marked. Comparing the contrast between both areas then allows to estimate how well different tissue classes

can be separated. Depending on the question that is answered using ROI-based comparison the ROIs are placed either by anatomical or by shape-based borders (Froeling, Pullens, and Leemans, 2016).

Beside the benefits of this kind of analysis there are also well-known problems assigned to it. The alignment of the ROIs may contain errors and therefore wrong type of tissue might be compared (Stieltjes et al., 2006). In their Nature Neuroscience paper, Kriegeskorte et al. (2009) showed that the manual placement of ROIs can significantly influence the outcome. It might even be possible that entirely wrong conclusions are drawn. The so caused problems are not only theoretical. Heye et al. (2013) showed that the reproducibility of ROI-based results can be very low. They found differences up to 28.5 % between different runs. An even worse result is reported by Goh et al. (2008), who found a reproducibility of only 50 %. Based on this result, they suggested to have the ROIs drawn by different raters and later use a consent. But they did not suggest a specific method, or proved that this would improve the results.

These findings also correlate with the results of Lambregts et al. (2011) who also found a high inter- and intra-rater variability. As a solution, he suggests to use complete segmentations instead of small regions. While this might be a solution for some cases, it can be difficult in others. Creating a whole segmentation is not only time consuming but also error prone, as previously discussed and confirmed by multiple studies (Weltens et al., 2001; Mazzara et al., 2004; Deeley et al., 2011; Porz et al., 2014; B. H. Menze, Jakab, et al., 2015)

A different idea is to improve the quality of the annotation by making the analysis more independent from the ROI placement. Based on the idea of Laidlaw, Fleischer, and Barr (1998) a Bayesian model is adapted to differentiate voxels that separate three types of tissue: two pure tissue classes and voxels containing both tissue types, i.e. that are affected by partial volume effect. This model is fitted using an EM algorithm, and analysis is then performed based on this distribution (Noe and Gee, 2001; D. Simon et al., 2012). While this approach does reduce the influence of the individual rater, it is still subject to the rater where the ROIs are placed.

To reduce this influence from a rater, other methods make use of model-based approaches that allow an automatic selection of the regions that are evaluated. Ortiz et al. (2014) used a model of unhealthy tissue which was then compared to normal Gray- and White-matter. Using this model, they found areas that are affected by Alzheimer disease and used these areas for further evaluation. But their main aim was the identification of new feature sets rather than comparing the contrast of tissue types.

Similar to this approach, Lim et al. (2013) proposed the use of Brain atlases for the automatic ROI placement and selection. But like other model-based approaches, this only works for organs that allow the creation of atlases. The same is true for other model-based methods, for example also for Tract-based spatial statistics (TBSS) (Smith et al., 2006). While this method allows to define a ROI with respect to the individual variations of each patient, it is tailored to specific applications – namely in the field of Diffusion MRI of the brain.

It is therefore not surprising that the most common approach is still the manual placement of ROIs. Either by one or multiple raters, as for example done by Davenport et al. (2013). While the latter option allows to produce more stable results and deliver inter-rater variabilities, the question which results to choose still remains open if the results of the individual raters differ from each

other.

3.6 Summary of State of the Art

Multimodal tissue characterization is an important area of research. Currently most approaches make use of learning based algorithms to validate the corresponding segmentation algorithms. But due to the availability of public datasets, like the BraTS challenge, the focus is mainly on the latter steps of such algorithms, like the feature selection. Even though it is acknowledge in literature that normalization is important, it is not systematically evaluated which normalization algorithm is suited best for such systems. Further, it is not evaluated if existing approaches could be improved by using more current classifiers like ExtraTrees instead of canonical RDF.

A further aspect that is often neglected is the variability of the training data. While there are some general approaches from computer vision to group data based on similarity this is not common for medical data. The approaches from the computer vision are often difficult to apply to medical images as they do make assumptions like many-class-problems that are not typical for medical images. Algorithms that allow to handle the variability of medical images are therefore still missing.

It is known that annotating the training data is often time consuming and error prone. But while there are some algorithms that incorporate model assumptions there is no general applicable algorithm for the training of new learning algorithms. But such algorithms are difficult to translate to different organs or modalities and also difficult to combine with other solutions from the state-of-the-art. A general approach that allows to reduce the annotation time while at the same time being compatible with the typical learning based approach is unknown to me.

Similar, the question of assessing the right imaging modality is often neglected. While it is known that annotation that are used do contain errors, these are only few methods to reduce these errors for small annotations, and to my knowledge they all make model assumptions about the underlying tissue. There are further no comparison methods that take texture information into account. Existing comparisons are therefore often limited to the pure contrast even though human vision does also incorporate texture information to distinguish various areas.

4

Methods

This chapter describes the methods that are used to validate the assumptions made for this thesis and describe the new developed methods. Following the overall structure of this work, it consists of four main sections. The first section, section 4.1, is about the evaluation of different pipeline steps. Here, the methods that are used to evalaute the influence of each step are described. In section 4.2 then describes two different methods that aim to reduce the effect of data variability. For this, the training data are either limited before the training or after the training. In addition to this, section 4.3 is used to describe three new methods to reduce the amount of labeling that is necessary to create the training data. The is achieved by proposing methods that allow the training of an classifier from weakly annotations. In the last section, section 4.4 the task of tissue characterization is evaluated within the scope of finding the right imaging modality. For this, a new combination of different sparse annotation is propose and an evaluation based on techniques from section 4.3 is described.

4.1 Classification pipeline

In the last time a wide variety of learning based algorithms for automatic tissue characterization has been proposed. Although there are some differences, the pipeline of most algorithm does contain similar steps (Figure 4.1). The first step is usually finding the structure of interest, for example by the use of a skull extraction algorithm in brain images or shape models for organ segmentation. Tje algorithms that are used for this step are highly dependend on the target.

The second step is only necessary if qualitative imaging methods – like MRI – are used. Most features which are calculated from the images are sensitive to intensity shifts. Similar, most classification algorithms correlate the actual value of a feature with its meaning and in return are sensitive to feature shifts introduced by varying image intensities. Therefore a normalization step is crucial to adapt for the variability of the imaging technique which is evaluated in (section 4.1.1).

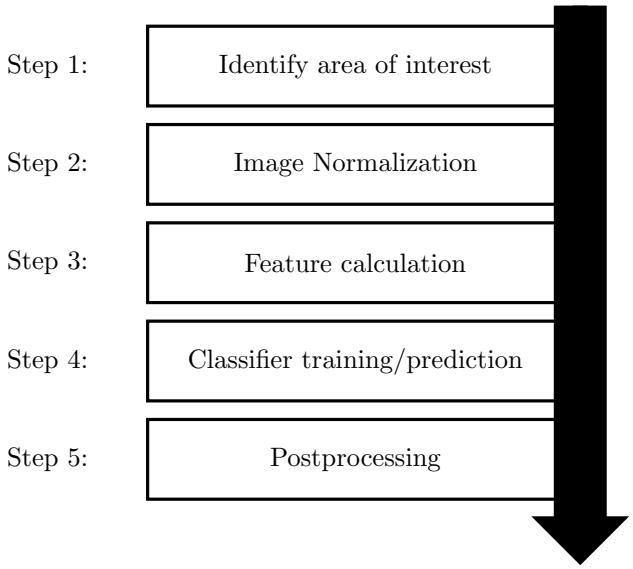


Figure 4.1: Algorithm scheme as it is usually used for the learning part of tissue characterization algorithm. The chosen steps are – different to Figure 3.2 – concrete for the use-case of machine learning based voxel segmentation for MRI data.

The third and fourth step consist of calculating the features and training of a classifier or the prediction of the new data. Depending on the learning algorithm this is sometimes done within one step or in two different steps. For this work, these two steps were condidered as two separate tasks, since the features are usually problem-specific. The choice of classifier is described in section 4.1.2.

Similar to the annotation of the region of interest, the post-processing is again depending on the actual task. It usually incorporates some kind of knowledge about the data, the question at hand, and algorithm. A throuough analysis of the postprocessing step is therefore beyond the scope of this work.

4.1.1 Evaluation of MR Normalization

Learning based methods require comparable ranges of values within the training data. This imposes a challenge if MRI is used as imaging modality, since the resulting images are only qualitative and not quantitative (see section 2.2.1). A range of methods have been proposed to make the intensity values of different images comparable. Each of these methods are based on different assumptions, therefore there is no single best solution but the optimal method is rather problem depended.

Obtaining a ground truth for the true meaning of different constrasts is usually not possible. The true tissue proportion at each image localization is unknown. Although it might be estimated, it is not possible to define a reference intensity value. I therefore decided to use a learning based approach to evaluate and quantify the influence of the normalization methods.

The experiments for this work are limited to the area of brain tumour segmentation. The main reason for this is availability of these data as public datasets as well as the vivid research within this area. While the findings give some indication for similar use-cases, they might not be directly applicable to all settings.

Classification set-up

A simple comparison of the histogram is insufficient for the comparison of different normalization methods. Due to the different tissue distributions in each image, two similar distributions do not necessarily reflect a good normalization (Shah et al., 2011). Therefore the success is measured using a k-Nearest Neighbours (kNN) classifier. These classifiers transfer the most common abel from the k most similar observations to the query observations. In the setting of brain tumour segmentation if the right label is predicted for an observation, this observation is similar to other observations (e.g. voxels) of the same type within the training data. The euclidean distance was chosen for the same reason as distance measure, i.e. it measures the absolute difference between the appearance of two voxels.

A single contrast is not enough for a successful brain tumour segmentation. There are some tissue types that – although very different from a physiological point of view – have a similar appearance. For example, active tumour and edema appears similar in MR Flair images, but can be separated if $T1_w$ MR are also used. As suggested by Verma et al. (2008) multiple contrasts were combined to allow an clear separability. No additional features were used besides the intensity of all image values to maximize the influence of correct or wrong normalized intensity values.

The lazy learning[1] of kNN classifier leads to a high dependency between prediction time and the number of training observations. Depending on the algorithm, a complexity up to $\mathcal{O}(n)$ is possible (Hastie, Tibshirani, and Friedman, 2013). To reduce the prediction times the training data where therefore randomly sampled until a fixed number of samples of each tissue class was drawn.

Bias field correction

A common assumption in computer vision is that same intensity values represent the same meaning, regardless of the location within the image. But this is not true for MRI images due to non-uniformity effects caused by inhomogeneities in the magnetic field. While the effects are too small to disturb a human observer, the changes might be large enough to reduce the detection quality of a learning based algorithm.

It is therefore a common approach to use algorithms that try to eliminate the influence of the inhomogeneities caused by the magnetic field. This is called bias field correction. Probably the most commonly used algorithm for this task is the so called 'N3'-algorithm proposed by Sled, Zijdenbos, and Evans (1998). Assuming that the bias field contains only small changes, and that the true image information is contained in higher frequencies, the algorithms estimates the bias field and subtracts it from the image. Based on this, Nicholas J Tustison

[1] Lazy learning: A basic model is learned very fast, but the computational effort for a prediction is therefore higher.

et al. (2010) proposed an openly available extension called 'N4ITK' which is also commonly used.

Because bias field correction algorithms impose further assumptions on the data, the benefit of these algorithms is not always clear. While some algorithms make use of bias field correction it is omitted by others. An aim was therefore ot validated the use of this algorithm during experiments.

Intensity normalization

The methods for MRI normalization can be classified into two groups, linear and non-linear methods. Both impose different assumption on the data. Therefore methods from both groups were selected for the experiments.

Linear methods Linear MRI normalization methods are based on the assumption that the differences in intensity values are only caused by an offset Δ and a scaling factor α. If these parameters are known, the intensity I of each voxel can be normalized as

$$I_{\text{normalized}} = \frac{I - \Delta}{\alpha} \tag{4.1}$$

Although it is well known that this imposes an over-simplification of the whole process, these methods are more stable than more complex, non-linear methods Shinohara et al. (2014).

Different strategies are used to estimate Δ and α. A common approach is to calculate basic statistic values, like mean and standard deviation and use them. Wihthin this work, this approach is refered to as **statistical normalization**.

Another method to estimate the parameter is to fit a single Gaussian function to the data using an expectation maximization algorithm. The mean and standard deviation of the fitted Gaussian function are then used to parametrize the normalization. This **peak normalization** is more robust to different tissue distributions than the statistic normalization, since only the main peak is considered. This method is similar to using a statistic normalization but using the mode instead of the mean.

Non-linear methods Multiple algorithms can be used for non-linear MRI normalization. To limit the number of test during the experiments, two of the more common approaches which represent the main ideas are selected.

Nyúl, Udupa, and X. Zhang (2000) proposed to use piecewise linear normalization. Based on the histogram different intensity percentiles are identified. Each percentile is then mapped to a fixed range (Figure 4.2). This is done by mapping the borders of each percentile to a fixed intensity value and linearly interpolating for points between. This **percentile normalization** assumes that the tissue distribution is similar within each image and that the borders of the percentiles are always at similar tissue types.

A different approach was proposed by Hellier (2003), who suggests to fit a fixed number n of Gaussian functions g_i with the maxima at m_i to the data using an expectation maximization algorithm. The same is done for an atlas image thus obtaining g_i^a and m_i^a. Based on this, the polynomial function p that is minimizing the difference

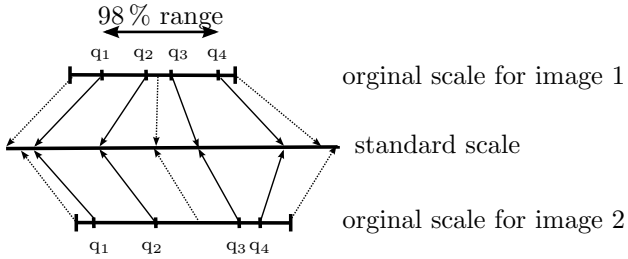


Figure 4.2: Schematic concept of the non-linear proposed by Nyúl, Udupa, and X. Zhang (2000). Within each quantile the intensity values are mapped linearly to a standard histogram, but due to the division into different quantiles, the whole process is non-linear. (Adapted from Nyúl, Udupa, and X. Zhang (2000))

$$\underset{p(\cdot)}{\operatorname{argmin}} \left(\sum_{i=0}^{n} \mid p(m_i) - m_i^a \mid \right) \tag{4.2}$$

is used to transfer each intensity value in the original image. This **polynomial normalization** does not assume a fixed distribution of tissue classes but a fixed number of different tissue classes and the clear separation of each class.

4.1.2 Evaluation of Classification Algorithms

Systems for learning based tissue characterization are often based on Random Decision Forest (RDF) algorithms. For example 50 % of the submissions for the 2014 BraTS challenge are based on this classifier family. It seems that this type of classifier is especially well-suited for this data. Most systems make use of the canonical form of random forest and ignore the substantial progress that was made in this area. Small changes in the training algorithms can already lead to visible improvement of the classification result.

To evaluate the effect of those changes and help to improve existing algorithms the influence of the classification algorithm is assesed with experiments. For this, the segmentation quality of a brain tumour segmentation system once using canonical random forests as suggested by Breiman (2001) and Extremely Randomized Trees (ExtraTrees) as suggested by Geurts, Ernst, and Wehenkel (2006) are compared. ExtraTrees were chosen because this algorithm is similar to random forests. It is therefore possible to adapt existing algorithms with only few changes. Other algorithms – like Rotation Forests (Rodriguez, Kuncheva, and Alonso, 2006) – seems to be even more powerful, but require more changes to the underlying tree structure.

Preprocessing

Based on previous findings and own experience the bias field error is corrected using ITK-N4 algorithm (Sled, Zijdenbos, and Evans, 1998; Nicholas J Tustison et al., 2010). Although there is some controversy about the benefit of this algorithm (M. d. C. V. Hernández et al., 2016), it was decided to apply it since it is commonly used for systems of automatic brain tumour segmentation. This gives a clear indication that its use is benefical for this specific dataset.

To account for the qualitative measurement of MRI, each image was normalized separately. Due to the presence of tumours with different sizes, no histogram based normalization algorithms were used. Instead, the mode[2] is chosen as reference point. The reasoning behind this is that this peak is within the same, healthy, type of tissue as usually more healthy tissue than tumorous tissue is present. The mode should therefore always represent the same type of tissue. As discussed in the previous section, the intensity of the reference value was subtracted from all intensities. The remaining then divided by the standard deviation within each image.

Features

The features are kept simple, to allow a simple reproducibility of the experiments. It has also been shown, for example by Kleesiek, Biller, et al. (2014), that simple features are strong enough for brain tumour segmentation.

Each feature was calculated by calculating a complete new feature image based on the original input image. This allowed to add, remove, or parametrizied each feature independently, and to vary the used feature combination without recalculating each feature. It also allows to use efficient implementation for image-wise operations, like smoothing etc.. . The final feature vector for each observation is then obtained by combining the intensity of the feature images of all used features.

Overall 54 feature were calculated for each modality:

(Smoothed) intensity value The intensity value of the original is used, as well as the intensities of images smoothed with a Gaussian filter. For the filtering a kernel sigma of 3 and 7 image steps was used.

Local histogram A local histogram was calculated for each image by using a $11 \times 11 \times 11$ neighbourhood. The range of the histograms was always from the minimum to the maximum intensity value within each image. The bin count – and therefore the number of features – is set to 11.

First order statistic Some first order statistics of the intensity distribution are calculated for a $7 \times 7 \times 7$ neighbourhood of each voxel. These are: mean, variance, skewness, kurtosis, minimum, and maximum of all intensity values.

Second order statistic A co-occurrence matrix filled with all values within a radius of 3 was used to calculate the second order statistics for the three main directions (Haralick, Shanmugam, et al., 1973). The features extracted from the co-occurrence matrix were energy, entropy, correlation, inertia, clustershade, clusterprominence, harralick feature, and the difference of moments.

Histogram based segmentation The remaining class labels of some automatic, histogram based threshold methods are used. The main idea of these methods is to find the best global threshold value based on some histogram based criteria. While these methods are not specially designed

[2] The most common value within the histogram

for brain tumour segmentation, they do often give good results. For the experiments the methods that are implemented in Insight Segmentation and Registration Toolkit (ITK)[3], namely Huang, Intermode, Isodata, Kittler, Li, Entropy, Moments and Otsu (Sharkey and Beare, 2016) are chosen. For all except the Otsu-threshold a two-class problem is assumed. For the Otsu, a two-, a three- and a four-class problem was assumed.

Training and prediction

The training data are randomly sub-sampled for the training of the classifier. Of all available and labelled brain tumour voxels that are used for a training run, only 0.5 % are actually used. This down-sampling affects the performance of the classifier only slightly since most of the training observations are very similar. But at the same time the necessary training time is significantly reduced, allowing for more detailed tuning of the final algorithm.

Using the down-sampled training data two different classifiers are trained. The structure of the model that is trained from both algorithms is identical and consists of multiple decision trees which use a one-dimensional (1D) decision threshold at each node. So the main difference between both algorithms is the training process. The first algorithm does the training according to the random forest (Breiman, 2001) implementation of Vigra[4]. The second algorithm is using the learning scheme of ExtraTrees proposed by Geurts, Ernst, and Wehenkel (2006). It was implemented by adapting the random forest implementation of Vigra accordingly.

Applying these two algorithms in the same way allows to compare the results directly and allows an estimation of the improvement obtained by using ExtraTrees instead of the canonical RDF. No post-processing was used to cancel out any further influence. Nevertheless, it should be noted that the segmentation results can be improved with post-processing steps.

4.2 Pre- and post-training data selection

The quality of a classification depends heavily on the quality of the training data. The labelling of the training data needs to be correct, the most common variances needs to be covered, and the correct type of data need to be annotated. These requirements are often not met for training data for medical imaging problems. Imaging artefacts, diffuse tissue appearances, and other problems may reduce the quality of single training images. I therefore developed two approaches to improve the quality of classifiers for voxel-wise tissue characterization. The first one, described in section 4.2.1, is used to find the best-suited training data for each image and improve the overall obtained prediction quality. The second one, described in section 4.2.2 can be used to improve the quality of an already trained model with few interactions.

[3] https://itk.org/

[4] https://ukoethe.github.io/vigra/ , see also chapter 2.3.2

4.2.1 Input Data Adaptive Learning (IDAL)

The traditional approach for learning voxel-wise classifiers is to train a single classifier that is then used for all images. While this approach is successful for general computer vision problems, it posses some challenges for medical images. These images can be very different, for example due to imaging artefacts, small differences in the pathology, or differences in the physiology of the patient (Adams et al., 1993; García-Gómez et al., 2008).

Instead of training a single classifier that is used to predict all unseen images a new approach is proposed to adaptively train a new classifier for every new image (Figure 4.3 and 4.4). This allows to use only few, but similar images during training. While such an approach makes each classifier less general, it is expected that the so-trained classifier is better suited to deal with the afore mentioned heterogeneity. This allows to adapt for multiple effects, compared to the approach of Opbroek, Vernooij, et al. (2015) or similar approaches which usually adapts only for one source of differences.

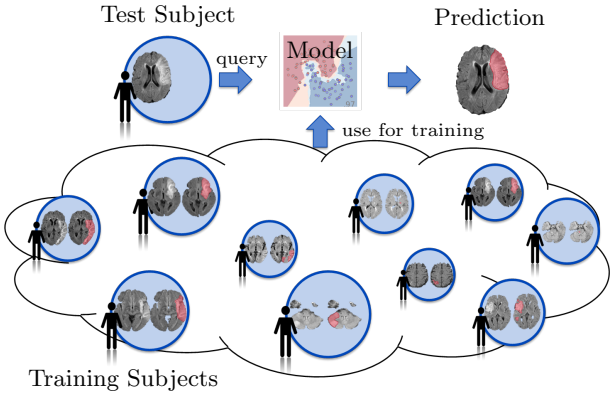


Figure 4.3: Traditional learning scheme. A classification model is trained based on a set of training subjects. Each subject contributes to the final model. This model is then used to predict each test subject.

The key idea of this algorithm is to learn the similarity between images. This is done in a first step where a Similarity Classifier (SC) is trained. Given an input image, this SC is trained to find those images within the training data base that are suited best as training base for the input image. The training and prediction of the SC is described in more detail in this section on page 45. This step can be performed off-line since is is independent of the current image that needed to be labelled.

To determine the segmentation of different tissue areas in a new test image or (i.e. a 'query image') the previously trained SC is used to find the most suitable training images. An individual Voxel Classifier (VC) is then trained using these training images. The combination of training images therefore depends only on the query image and all possible combinations of training images can be used, the VC is trained online. More details of the VC that is used are given in this section on page 47.

An overview of this work-flow is given in Algorithm 1. Beside the fact that this scheme allows to correct for most sources of differences it can also

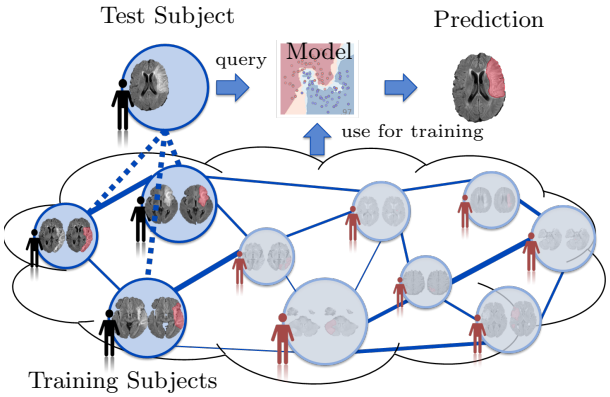


Figure 4.4: Proposed scheme. Instead of creating a single model from all training subjects, a small subset of training subjects is selected for each test subject individually. Only these training subjects, which are most similar to the test subject, are then used to train a specific model that is used to segment the given subject. The other training subjects are ignored during this run.

be combined with other approaches. It does not depend on a special type of VC, preprocessing, or post-processing. This allows to use it for most types of learning-based tissue characterization problems and combine it with most existing learning based approaches.

Algorithm 1 IDAL algorithm

$I_k := \{\mathbf{X}_k, Y_k\}$
$\mathcal{I} := \{I_0 ... I_n\}$ ▷ Training set with n subjects

$I_T := \{\mathbf{X}_T\}$ ▷ Test subject

procedure GLOBAL LEARNING($\mathcal{I}$) ▷ Off-line Training
    SM ← Calculate Similarity Matrix($\mathcal{I}$)
    global SC ← Train Similarity Classifier($\mathcal{I}$, **SM**)
end procedure

function PREDICTION(SC, I_T)
    SR ← Estimate Similarity Ranks(SC, I_T)
    $\mathcal{I}'$ ← Select s most similar training subjects($SR, \mathcal{I}$)
    VC ← Train Voxel Classifier($\mathcal{I}'$) ▷ Online Training
    Y'_T ← Predict Segmentation (VC, I_T)
    return Y'_T
end function

Similarity Classifier (SC)

The main goal of the similarity is to identify those images, that are suited best to train a classifier for a new query image. Therefore the Classification

Similarity Score (CSS) between two images or subjects I_0 and I_1 is defined as the segmentation accuracy that is achieved segmenting image I_1 using a classifier that is trained using only image I_0 (Figure 4.5). In this use-case, the accuracy is measured using the Dice score.

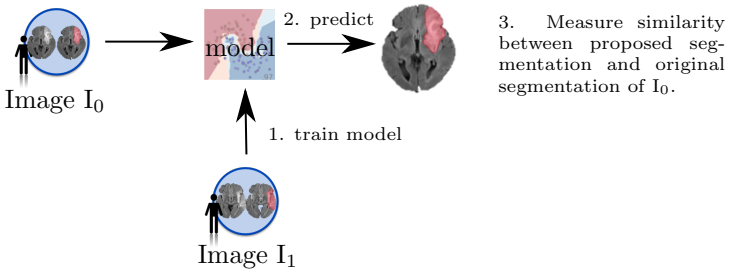


Figure 4.5: Visualization of the defined subject or image similarity. To asses the similarity of I_0 to I_1 a classifier is trained using I_1. Using this classifier, a segmentation is predicted for I_0. The similarity is then the overlap of the true segmentation and the predicted segmentation, measured using the Dice score.

This definition of similarity is directly depending on the segmentation approach that is used. Using two different segmentation algorithms – that might differ in the classification algorithm, the preprocessing, or the features that are used – might lead to completely different CSS for the same images. So CSS does not only take into account the images but also accounts for the algorithm that is used. This is an important difference compared to other approaches, like the one of Tighe and Lazebnik (2013). It omits the need for manually adapting a global feature set and gives a direct answer to the main questions: which are the best images to learn from for a specific query image?

A challenge using CSS is the calculation if the ground truth for one of the images is missing as it is usually the case if a image should be segmented. It is then either not possible to train a classifier or it is not possible to estimate the classifier accuracy and therefore needs to be estimated. Since the idea of the proposed algorithm is to use CSS to find the best training image it is necessary to estimate it.

Neighbourhood Approximation Forests (NAFs) proposed by Konukoglu et al. (2012) and Konukoglu et al. (2013) are used for the estimation of CSS. A NAF consists of many decision trees that are trained by grouping similar images together. During prediction, each tree indicates the training images that are similar to the query image and the training images with the most votes are those which are most similar. They were specifically developed having images in mind. Rather than estimating the similarity it estimated the order of similarity between the training image and the query image. Further, having a decision tree like structure they offer in-build feature selection, allowing to use all features and do not take care of the feature importance.

In the first step CSS $\rho(\cdot,\cdot)$ is calculated between all training images and the NAF is trained to be useed as a SC. After this a feature vector is calculated from all training images and a NAF is trained. There is no limitation regarding the features that are selected. The feature vector used during the training and test of the NAF consists of the first order statistics (intensity minimum, maximum,

range, mean, variance, sum, median, standard deviation, mean absolute deviation, root means square, uniformity, entropy,energy, kurtosis, skewness, and the the number of voxels) of the preprocessed images. Although more complex features can be used only those simple features are used to keep the whole process are simple as possible. It would be further possible to use non-image based information, like the patient age, or the diagnosis. But these information can not be obtained from the images, and were not provided with the data.

The NAF is trained with 100 trees, a minimum of two samples at each leaf, 30 random tests for best split at each node during the training and a maximum tree depth of 12. After predicting a new patient (Online training state, see Algorithm 1) the s highest ranked training images are chosen to train the new VC. For this work, the NAF implementation provided by Konukoglu (2016) which is based on distance rather than similarity is used. Based on the similarity, it is calculated as '$1000 - 1000 \cdot \rho(\cdot, \cdot)$'.

Voxel Classifier (VC)

The actual estimation of the labels for each voxel is done by a separate classifier. The proposed approach is independent of the algorithm that is used for this task. A RDF algorithm is used because it is commonly used for tissue characterization. Based on the previous experiments (See chapter 4.1.2) it was decided to use ExtraTrees, which usually perform slightly better than canonical RDF. This was further supported by the fact that this algorithm was already successfully used for the segmentation of ischemic strokes (Maier et al., 2015), the same task that was chosen to evaluate the proposed approach.

The classification is kept as simple as possible to emphasis on the main novelty – the selection of the training data. Therefore only basic features were calculated to describe each voxel. Beside the intensity and the difference of intensities for each modality, the Gaussian, difference of Gaussian, Laplacian of Gaussian (in three directions) and the Hessian of Gaussian were calculated. If applicable with Gaussian Sigmas of 1 mm, 3 mm, and 5 mm. This leads to a total number of 82 features per voxel if 4 different modalities are used.

Each ExtraTrees classifier was trained with 50 trees and the Gini purity as optimization measurement. The maximum tree depth was not limited. During each training (during similarity calculation and final VC training) the best class weights and minimum samples at leaf nodes were independently estimated using cross validation.

The evaluate of the proposed approach is based on the problem of sub-acute ischemic stroke segmentation. This disease shows a high variability within the data, and additional a public dataset that allowed the comparison with other techniques is available. To show that the proposed approach can be combined with multiple learning methods the DALSA-learning scheme is also incporporated, which is described in more detail in section 4.3.1. This was done to show that this algorithm can be used with complex approaches. The necessary relabelling of the data was done in less than $2\frac{1}{2}$ h for the complete training set. A more detailed description of the dataset is given within the experiment sections (section 5.1).

4.2.2 Pre-trained semi-automatic tissue characterization

The final decisions made during clinical routine needs to be made by physicians. It is therefore necessary to incorporate methods that allow physicians to correct the results obtained with automatic algorithms. But at the same time it is also important to enable them to do this with as less effort as possible.

Weighted forest

Therefore a new method was that allows to label data in an interactive way, while at the same time makes use of previous labelled data. It is further possible to improve the quality of the trained classifier using the corrections made during the semi-automatic annotation process. Here, the focus is on RDF based classification algorithms.

One important part of the proposed algorithm is an already trained RDF classifier. This classifier might be used for an automatic classification pipeline as previously suggested or may be trained only for this purpose. There are no limitations regarding the set of features, the actual used RDF algorithm, the set of parameter, the pre-, or the the post-processing.

The workflow of the proposed algorithm is shown in Algorithm 2. The user marks some small areas of each label to start the annotation process. This can be done, for example, by labelling only some voxels of each class by clicking on them. While this is done, the features for the whole image are calculated.

Algorithm 2 Interactive labelling algorithm (Weighted update)

data: $\mathbf{X}$:= Input image
data: RF := pre-trained random decision forest

training points $p \leftarrow \{\}$
$\mathbf{X}' \leftarrow$ Calculate additional features($\mathbf{X}$)
do
    $p \leftarrow p \cap$ new training points ▷ User interaction
    $w_t \leftarrow$ calculate weights($p, RF, \mathbf{X}$)
    $RF \leftarrow$ Update RF with weights(w_t, RF)
    $Y' \leftarrow$ Predict Segmentation($RF, \mathbf{X}$)
while Y' quality is good enough ▷ User interaction

result: Y', RF

After this, the classifier is used to predict a label for the already manually labelled voxels. The prediction accuracy of each tree of the used RDF can be estimated by comparing the obtained labels with the known ground truth. This allows to estimate the quality of each tree for the given classification problem. With this, it is possible to assign the prediction of each tree an individual weight $w_t = \text{TPR}$ based on the True Positive Rate (TPR).

Instead of using simple majority voting to combine the prediction probability of a class C with a forest, these weights are now used for a weighted majority voting of each tree t, i.e. the contribution of each tree to the final vote is determined by the weight:

$$P(C \mid \mathbf{x}) = \frac{1}{\sum w_t} \sum_t w_t P_t(C \mid \mathbf{x}) \tag{4.3}$$

The so modified classifier is then used to predict the complete image, and the result is shown to the user. The user can then accept the result or can again label some points to further improve the classification results. In this case, the additional points are also used to estimate the weights for the tree. These steps are repeated until the result is accepted by the user.

Reference method

The state of the art approach for interactive segmentation is the new training of each tree (Algorithm 3). The same interaction scheme as for the previous described algorithm is used, i.e. some points are labelled by the user, a first proposal is shown and the user can then add some label information to yield a better segmentation. This approach does not make any use of previous annotated data and reflects the current state-of-the-art for interactive segmentation.

Algorithm 3 Interactive labelling algorithm (Relearning)

data: $\mathbf{X}$:= Input image

training points $p \leftarrow \{\}$
$\mathbf{X}' \leftarrow$ Calculate additional features($\mathbf{X}$)
do
    $p \leftarrow p \cap$ new training points ▷ User interaction
    $RF \leftarrow$ train random decision forest($p, \mathbf{X}$)
    $Y' \leftarrow$ Predict Segmentation($RF, \mathbf{X}$)
while Y' quality is not sufficient ▷ User interaction

result: Y', RF

Classifier setup

Both approaches are based on the same classifier setup to make the results comparable. ExtraTrees were used as training algorithm for the individual trees, and the same set of features are used.

To allow a smooth user experience the used features should be calculated very fast. Therefore, only features that are fast to calculate were used.

Intensity value The intensity value of each voxel and the intensity value of twelve neighbours according to Figure 4.6 are used as feature

Smoothed intensity value The intensity value of Gaussian smoothed images.

Laplacian of Gaussian (LOG) The Laplacian operator is used to enhance edge-like structures in the images. A Gaussian smoothing is applied before to reduce the noise.

Difference of Gaussian (DOG) The difference of two Gaussian smoothed images are used. As with LOG, this highlights border areas.

First order statistics First order statics are calculated over a local neighbourhood, namely mean, minimum, maximum, variance, skewness, and kurtosis.

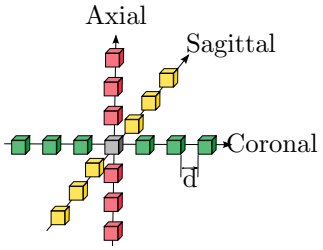


Figure 4.6: For each voxel (gray) 18 neighbours are selected, six for each direction. The neighbours are sampled with a distance d.

4.3 Methods for Reduced Annotation Effort

There are two obvious methods to reduce the time necessary to create annotations for learning based methods. The first one is to use software programs during the annotation process that aids the process and lead to a speed up. These are usually semi-automatic segmentation algorithms like Region Growing, Graph-Cut, or learning based approaches (Deng et al., 2010; Zhao, Wu, and Corso, 2013; Sommer et al., 2011). While these methods are suitable to reduce the annotation time, they include some model assumption into the final training data. This could make it difficult or even impossible to adapt these solutation to other task and further complicates the combination with other approaches.

Therefore the focus is on the second methods, namely the reduction of necessary annotations. There are different methods to reduce the necessary training material. This can be done by either reducing the amount of annotated voxels or by using non-voxel based annotations. An example for non-voxel based annotations are information about tissue ratios. Since the size of pathological tissue is often required during clinical routine, the tissue ratios are often already known. This allows to use data from the clinical routine without further annotations. As shown in this chapter, it is also possible to combine the different annotation methods. The evaluated combination for the reduction of the training data are given in table 4.1.

4.3.1 Learning from Sparse Annotations

To reduce the labelling time necessary for creating training data for automatic tissue characterization, I propose the annotation of Sparse and Unambiguous Regions (SURs) instead of the segmentation of the complete image. Unlike Learning from Complete Annotations (LCA), Learning from Sparse Annotations (LSA) introduces a sampling bias. I propose to correct this error with domain adaptation, which I refer to as Domain Adaptation for Learning from Sparse Annotations (DALSA). The different methods used for annotating, sampling, and using training data that are described in this section are summarized in Figure 4.7.

TABLE 4.1
METHODS FOR REDUCED ANNOTATION

Name	Annotations	Section
Traditional	Full annotations of the images. All voxels need to be annotated	
Sparse Annotations	Sparse and Unambiguous Regions are annotated. No complete annotation is necessary.	4.3.1
Positive Sparse A.	Sparse and Unambiguous Regions that contain only example of one class. The ratio of this tissue type is also known.	4.3.2
Image-wise A.	For every image only the ratio of tissue types is known. No voxel-based annotation is required.	4.3.3

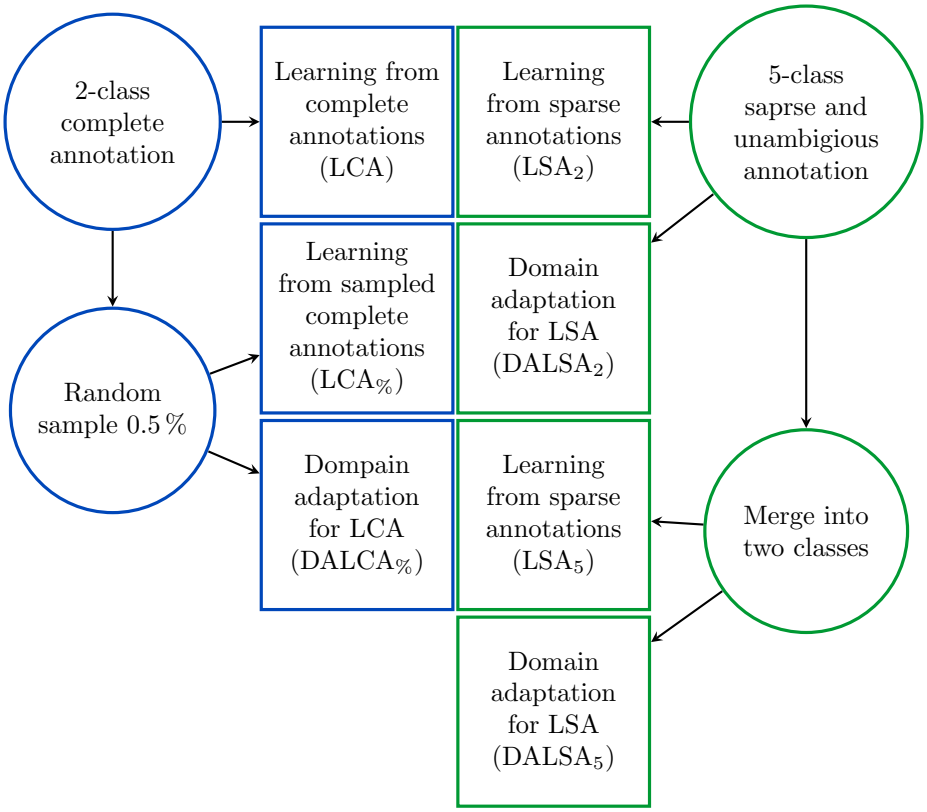


Figure 4.7: Different methods used for annotating, sampling, and using training data for supervised learning. Most state-of-the-art approaches make use of LCA or $LCA_{\%}$, which require a complete annotation of the data and differ in their sampling strategy. The use of sparsely annotated training data (LSA and DALSA) is proposed to reduce the annotation time. The sparse annotations of 5 tissue classes were either treated separately or merged to two classes ('healthy' and 'fluid' were merged to 'healthy', 'edema', 'active' and 'necrosis' were merged to 'tumorous')

Sparse and Unambiguous Regions

Different levels of incomplete annotations are possible. Both – labelling all voxels expect one, or labelling only one voxel – will give sparse annotations. But this would either reduce the positive effect of the sparse annotation or make learning from the data extreme difficult. It is therefore important to have annotations in between those two extremes. To avoid this, a set of basic rules for the creation of incomplete annotations is identified and the resulting regions are named Sparse and Unambiguous Regions (SURs):

Representative The annotation should cover representative areas of the annotated tissue type and include most possible variations of this tissue type.

Sparse The annotations should be sparse, i.e. contain not too many information. The definition of sparse depends on the experts who is is performing the annotation.

Unambiguous The annotation should not contain any area that is ambiguous, including, but is not limited to, areas with unclear pathological areas, or voxels affected from partial volume effect. These areas should be left out during the annotation process.

Free placed There is no limitation about the size of a SUR, the location within an image of the number of connected areas that are used to annotate a single tissue type.

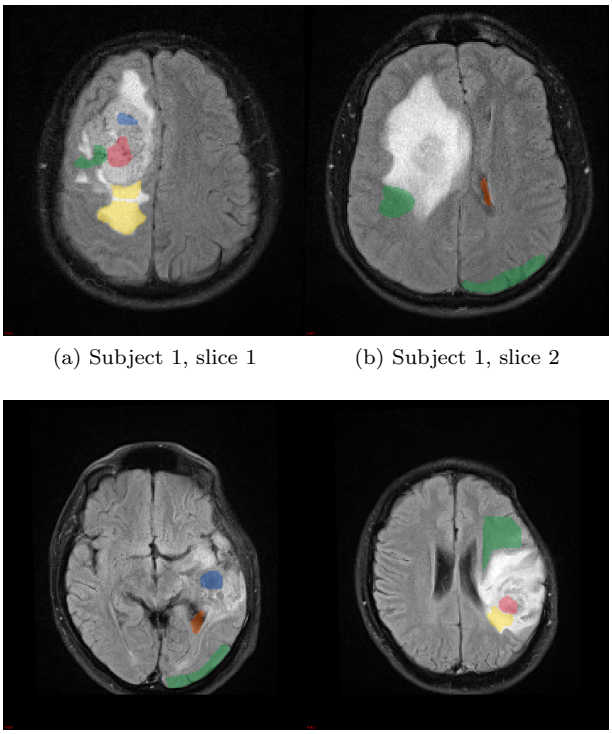
The best degree of sparsity depends on the actual data. To make the annotation process as simple as possible, and to avoid to make a solution that is only tailor-fitted to a single use-case, the rules for SURs were kept as generic as possible.

Using these rules, typical annotations for brain tumour segmentation covered about 1 mm of the brain volume and are usually located in one or two slices of the brain. Example images of such annotations are given in image 4.8.

To be able to evaluate the effect of different annotation schemes further some more specific annotation rules were defined which are listed in table 4.2. The scheme 'Main' corresponds to the previous given rules, while the other annotation types are more specific about the location, size, and tissue border areas.

Domain adaptation

A basic assumption in machine learning is that training are independent and identically distributed (i.i.d.) (Vladimir N. Vapnik, 1998; Duda, Hart, and Stork, 1999; Hastie, Tibshirani, and Friedman, 2013). This assumption is fulfilled if a classifier is trained for voxel-wise classification of tissue characterization based on fully annotated data. It even holds if the already labelled data are randomly reduced to save training time like it is commonly done. However, if only small parts on an image are annotated by experts and used for the training the i.i.d.-assumption is violated and a sampling selection bias occurs (James J. Heckman, 1979). The distribution of features $\mathbf{x}$ and labels y will be different in the observations $(\mathbf{x}, y)$ processed during training and prediction, i.e. $P_{\text{Train}}(\mathbf{x}, y) \neq P_{\text{Predict}}(\mathbf{x}, y)$. This will lead to classifier with non-optimal



(a) Subject 1, slice 1 (b) Subject 1, slice 2

(c) Subject 2, slice 1 (d) Subject 2, slice 2

Figure 4.8: Examples of SUR annotation for two subjects. The annotation for both subject is done on two slices of the whole image space. The colour coding is 'green': healthy, 'orange': CSF, 'yellow': edemea, 'red': active tumour, 'blue': necrosis.

decision boundaries – some features may be over-represented while others are under-represented.

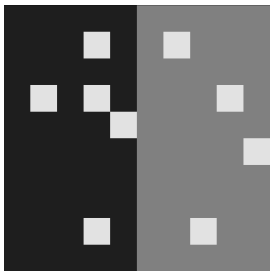
Figure 4.9 shows a simplified example to demonstrate the effect of sampling selection error and domain adaptation. The probability for a combination of feature vector and label $P(\mathbf{x}, y)$ is affected by a sampling bias. This probability can be written as:

$$P(\mathbf{x}, y) = P(y \mid \mathbf{x}) \cdot P(\mathbf{x}) \tag{4.4}$$

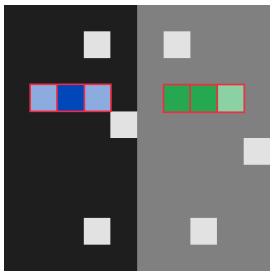
A theoretical assumption often made in domain adaptation is that the meaning of a feature is the same in the training and prediction domain, i.e.

$$P_{\text{Train}}(y \mid \mathbf{x}) = P_{\text{Predict}}(y \mid \mathbf{x}) \tag{4.5}$$

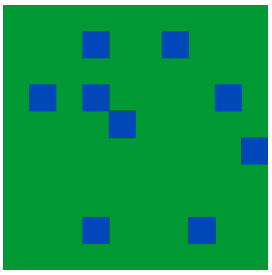
It is save to make this assumption for thus application. If the annotated areas are representative for all tissue classes, the meaning of all features should be the same, regardless if the full images are annotated or only parts of them. Further Huang et al. (2007) showed that techniques, that depend on this assumptions, are useful even if it only partially fulfilled. The remaining difference between



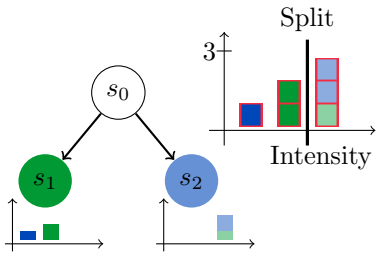
(a) Original image. Bright spots are noise.



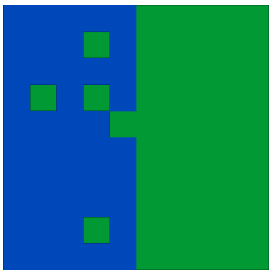
(b) SUR annotation of training voxels (red border) for left (blue), and right (green)



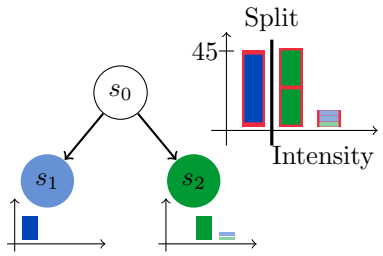
(c) Segmentation result if no domain adaptation is used.



(d) Tree created when no domain adaptation is used. Sampling selection error dominates results.



(e) Segmentation result if domain adaptation is used.



(f) Tree created when domain adaptation is used.

Figure 4.9: Simplified example to demonstrate the effect of sampling selection error and domain adaptation. (**a**) The given image with 100 pixels is classified into left and right using intensity as feature. On both sides salt noise (bright pixels) simulates noise in the multidimensional data. (**b**) For training data, SURs are used. A single tree with a tree depth $T = 1$ is used as classifier. (**c**) shows the segmentation result with the original image if no domain adaptation is used; (**d**) shows the resulting tree that has a false split due to the noise pixels. (**e**) gives the segmentation result if domain adaptation is used, (**f**) gives the tree from the corrected data. The number of pixels at each node differs from the number of the pixels within the SURs because the classifier uses the number of pixels multiplied by a weight. For example, there are nine bright pixels in the given image and the SURs cover three of them. Therefore the weight for bright features is $w(\text{BRIGHT}) = 9 \div 3 = 3$.

TABLE 4.2
SUR Annotation Strategies

Type	Description	Diameter	Location
Main	$1-3$ SUR per class	rater depended	covering bordering as well as central tissue areas
Type 1	1 SUR per class	$6-14\,\mathrm{mm}$	arbitrarily varying
Type 2	3 SURs per class (different slices)	$6-14\,\mathrm{mm}$	covering bordering as well as central tissue areas
Type 3	3 SURs per class (different slices)	$6-14\,\mathrm{mm}$	covering central tissue areas only
Type 4	3 SURs per class (different slices)	$6-14\,\mathrm{mm}$	covering bordering tissue areas only

Description of different SUR labelling strategies. A complete set of SURs was created for each strategy.

the distribution in the training and prediction data is the probability of a given feature vector

$$P_{\text{Train}}(\mathbf{x}) \neq P_{\text{Predict}}(\mathbf{x}) \tag{4.6}$$

Shimodaira (2000) calls this situation covariate shift. He suggests compensating the difference by weighting each observation with the density ratio of the feature vectors during training

$$w(\mathbf{x}) = \left(\frac{P_{\text{Predict}}(\mathbf{x})}{P_{\text{Train}}(\mathbf{x})}\right)^{\lambda} . \tag{4.7}$$

The ration $w(\mathbf{x})$ is high for observations occurring often within prediction data and seldom within training data, while $w(\mathbf{x})$ is low for observations that are rare in the prediction data but frequent in the training data. In this case it means that labelled voxels (training data) that are typical for the entire image (prediction data) receive more emphasis than a-typical voxels. Annotated areas of healthy tissue are usually assigned a high weighting factor than areas of tumorous tissue because there is usually more healthy tissue within a brain while the SURs are usually about the same size.

The relaxation coefficient $\lambda \in [0...1]$ was introduced by Shimodaira (2000) to control the effect of the weights. The weights have no effect if $\lambda = 0$ and for $\lambda = 1$ the effect of the weights is maximized. The best value for the relaxation coefficient depends on the used classifier; in general λ needs to be smaller for small training sets. I set λ to 1 because, being voxel based, the training base is rather large. This choice was further evaluated with an experiment.

Since the distributions of features within the training and prediction data are usually unknown, $w(\mathbf{x})$ is usually estimated. There are several ways to do this and Sugiyama and Kawanabe (2012) give an overview for the most common methods. The approach of assessing $w(\mathbf{x})$ by estimating the probability of whether an observation with feature vector $\mathbf{x}$ belongs to the training or prediction data (Bickel, Brückner, and Scheffer, 2007) was chosen. If observations

that are used for the training data are labelled $z = 1$ and observations that are predicted $z = 0$ then $\mathbf{x}$ can be estimated by

$$\begin{aligned} \hat{w}(x) &= \left(c \cdot \frac{\hat{p}\left(z = 0 \mid \mathbf{x}\right)}{\hat{p}\left(z = 1 \mid \mathbf{x}\right)} \right)^{\lambda} \\ &= \left(c \cdot \frac{1 - \hat{p}\left(z = 1 \mid \mathbf{x}\right)}{\hat{p}\left(z = 1 \mid \mathbf{x}\right)} \right)^{\lambda} \quad . \end{aligned} \tag{4.8}$$

The estimation of the probability $\hat{p}(z = 1 \mid \mathbf{x})$ is done by training a Logistic Regression Classifier (LRC). For this purpose, each voxel within a SUR is labelled as training data, i.e. $z = 1$. Additionally, all voxels that belong to the brain are labelled as prediction data, i.e. $z = 0$. Thus the voxels that belong to the SURs appear twice: once within the training data and once within the test data. These data are then used to train the parameter function of the logistic regression $\theta(\mathbf{x})$, which can then be used to estimate the required probability by:

$$\hat{p}\left(z = 1 \mid \mathbf{x}\right) = \frac{1}{1 + \exp\left(-\theta(\mathbf{x})\right)} \tag{4.9}$$

The estimation of $w(\mathbf{x})$ can be further simplified by incorporating equation 4.9 in equation 4.8:

$$\hat{w}(\mathbf{x}) = \left(c \cdot \exp\left(-\theta(\mathbf{x})\right)\right)^{\lambda} \quad . \tag{4.10}$$

This approach allows a fast calculations of the weights and eliminates the need for division, which increases the numerical stability. A generalized linear model with a logit function as link function and a binomial distribution to fit the logistic regressor (Dobson and Barnett, 2002) was used for this. The advantage of this method is the parameter-free training of the logistic regressor, which makes the whole algorithm more robust and easier to use.

The weights are calculated patient-wise, i.e. for each patient the SURs are created and then the weights for this patient are calculated. Therefore the weights for a patient are independent of other patients and new patients can be added to the training data without recalculating the weights within the existing training data. Also, no full tumour segmentation is necessary.

The constant c can be used to control the influence of each image volume during the training without changing the relations of voxels which belong to the same image. The sum of weights of all SUR voxels is $c \cdot n_{\text{Predict}}$ (n_{Predict} is the total number of voxels in the brain mask, see appendix A.1 for a mathematical derivation). A common approach is to set $c = \frac{n_{\text{Train}}}{n_{\text{Predict}}}$, with n_{Train} being the number of voxels in the SURs (Sugiyama and Kawanabe, 2012). This normalizes the sum of all weights to the number of training points. However, in this case it would mean that the contribution of an image to the overall training depends on the size of the SURs, i.e. an image with large SURs would have more impact on the final classifier than an image with small SURs, although the latter might contain more valuable information. Therefore c is set to $c = 1$. In this case, the impact of an image is determined by n_{Predict}, as it would be in a standard classifier training scenario.

Observation weighted Classifier

The advantage of using observation weights to correct for error made by the reduced annotation scheme. This allows to combine this approach with all classification algorithm that allow observation-based weights which is true for most algorithms. It is even possible to use observation-based weighting if this is not supported natively by the learning algorithm. Zadrozny, Langford, and Abe (2003) showed that it is possible to insert the weights for all algorithms by creating ensembles of classifier and including the weights into the ensemble building process.

The classification in the main experiments is mainly based on Random Decision Forest (Amit and Geman, 1997; Breiman, 2001). RDF-based methods have previously achieved promising results in brain tumour segmentation Zikic, Glocker, Konukoglu, et al., 2012; Bauer, Wiest, et al., 2013. Since the canonical RDF definition does not support observation weights, a variant is used that is similar to the RDF implementation in the python scikit module (Pedregosa et al., 2011). Here, the prediction algorithm itself is not modified. An unseen observation is passed down the decision trees based on binary tests within each node of the tree until it reaches a leaf node. The prediction is then obtained by majority voting of all trees. The training of observation-weighted random forests is also similar to the original version. At each node, the best split within a random set of features is determined based on an impurity measurement. All data are then split into two groups, which are used to train the child nodes. This is repeated until the maximum tree depth is reached or only one label type is left.

The major difference to the canonical RDF implementation is how the impurity is calculated. For this the Gini Impurity I (Breiman et al., 1984) is used:

$$I(V) := 1 - \sum_{y \in Y} P_V(y)^2 \tag{4.11}$$

The class probability $P_V(y)$ for the data V at each node is usually calculated from the number of observations with this label divided by the overall number of observations. If observation weights are used, $P_V(y)$ is calculated using the sum of the weights instead of the number of observations i.e.

$$P_V(y) := \frac{1}{\sum_{\mathbf{x}_i \in V} w(\mathbf{x}_i)} \cdot \sum_{\mathbf{x}_j \in \{V | y_j = y\}} w(\mathbf{x}_j) \ . \tag{4.12}$$

The forests in this experiments consisted of 1000 trees; the number of features at each node was set to the square-root of the number of all features (i.e. 4). The minimum sample size at each leaf node was set to 1, the noise reduction being instead achieved by limiting the maximum tree depth. To account for the different levels of noise and amount of data, multiple runs of the experiments were ran with different tree depths and used the optimal tree depth for each approach.

While random forests were used in the main experiments, the proposed method can be used with any classification algorithm that allows for the incorporation of observation weights. This is demonstrated in an additional set of experiments on basis of weighted SVMs as described by X. Yang, Song, and

Y. Wang (2007). The experiments are based on a non-linear radial basis function kernel (Gaussian) and the Karush-Kuhn-Tucker stop criterion (Andreani, Martinez, and Schuverdt, 2005). The noise sensitivity is regularized using the cost parameter c.

4.3.2 Learning from Only Positive Annotations

Learning from SURs already reduces the number of voxels that needs to be labelled. But creating Sparse and Unambiguous Regions (SURs) for tissue that is highly diverse – like healthy brain tissue – can still be time-consuming. The annotation with SURs needs to cover all subtypes within a tissue class. For example, this includes, but is not limited to, Gray matter-, White matter-, CSF-, and cerebellum-tissue for healthy brain tissue. Annotating all these subtypes might not only increase the necessary time but also increases the chance that the created SURs are not representative for every tissue type.

Another challenge arises from the new possibilities to annotate medical images using crowd-based approaches (Maier-Hein, Mersmann, Kondermann, et al., 2014). The need to annotate more than one class requires either a switch of the annotation class or multiple annotation runs. Both options significantly increase the time required by the crowd workers and therefore the annotation cost.

Therefore a new method that allows to learn having only type of tissue annotated, for example tumorous tissue only, is developed. By using the information about the tissue ratio within the images, it is now possible to train a classifier with only one annotated tissue type.

Tissue Ratio estimation

Additional information are necessary to be able to train a classifier for separating two tissue types using annotations that contain only one class. The proposed algorithm is based on an algorithm that makes use of the ratio π of the annotated tissue type within the unlabelled tissue $\mathcal{U}$, i.e.

$$\pi := P(y = 1 \mid \mathcal{U}) \tag{4.13}$$

For some data these ratio is already estimated. For example, during tumour therapy, the volume of the tumour is an important measure for therapy success and therefore commonly measured (Suzuki et al., 2008). If this is not case for the data at hand, it is still possible to obtain a fast estimation of π either by during a fast manual measurement or by using an automatic estimation.

Manual estimation If the volume V of the annotated tissue type is known for each patient, this ratio can be calculated based the number of annotated n, the number of unlabelled n' voxels, and the voxel volume v_{Voxel}:

$$\pi = \frac{V - n \cdot v_{\text{Voxel}}}{(n + n') \cdot v_{\text{Voxel}}} \tag{4.14}$$

While the number of voxels and the voxel volume can be easily calculated using the image data, the tumour size needs to be estimated if no complete annotation is given. Inspired by the method of A. B. Miller et al. (1981) and

Galanis et al. (2006), which is commonly used within clinical routine, evaluated two methods to manually asses the volume of specific tissue type like a lesion were evaluated. For the first method the area $_c$ of the largest circle fully within the tissue area is multiplied by the height $h_{\text{TissueArea}}$ of the tissue area. For the second method the largest possible diameter l_a and the largest perpendicular diameter within the same slice l_b where measured. The tissue volume was then estimated by $V = 0.5 \cdot l_a \cdot l_b \cdot h_{\text{TissueArea}}$.

Algorithmic estimation Being able to automatically estimate the tissue class ration π allows further reduction of manual labour, which is especially important for images where the size of the annotated class was not estimated during the clinical routine. Another reason might be an highly non-regular shape of lesions, for example. In such cases simple estimations of the lesion volume using only 2D measurements are highly unreliable.

Assuming that the observations of both classes are different enough and the annotated examples are drawn i.i.d., π might be estimated from the data. According to Marthinus Christoffel du Plessis and Sugiyama (2014) this can be done by finding the overall class prior $\theta = P(y = 1)$ that reduces the differences between the two distribution $\theta \cdot P(\mathbf{x} \mid s = 1)$ and $P(\mathbf{x})$ if the label s indicates whether an observation is labelled ($s = 1$) or not ($s = 0$). Figure 4.10 schematically illustrates this concept. The assumption behind this technique is that both classes are different enough the labelled data will cover only area of one class. By minimizing the difference between all data and labelled data in this area the ratio between the two classes could be found. They suggest to use the Pearson Divergence (PE) to measure the similarity between both distributions. Based on this idea they then derive an analytic solution for θ. Since an estimation of π is needed it can be calculate from θ by

$$\pi = \frac{\theta - P(s = 1)}{1 - P(s = 1)} \tag{4.15}$$

and call this method Pearson Divergence Prior Estimation (PEPE).

Marthinus Christoffel du Plessis and Sugiyama (2014) made the assumption that the labelled observations are drawn i.i.d.. Since in the given setting the labels are manually selected I expect a sampling bias. This might affect the performance of the estimation algorithm similar to the affection of a classification algorithm (Figure 4.10). Therefore a correction of the sampling bias by importance weighting (Shimodaira, 2000) is proposed using the same technique that was already used for DALSA. Again, each observation that is used for the calculation of θ is weighted with the following ratio:

$$w(\mathbf{x}) = \frac{P_{\text{Prediction}}(\mathbf{x})}{P_{\text{Training}}(\mathbf{x})} \tag{4.16}$$

by correcting the class estimation according to

$$\hat{P}(\mathbf{x} \mid s = 1) = \frac{w(\mathbf{x})}{\mid w \mid} \cdot n \cdot P(\mathbf{x} \mid s = 1) \tag{4.17}$$

$$\hat{P}(\mathbf{x}) = \frac{w(\mathbf{x})}{\mid w \mid} \cdot n \cdot P(\mathbf{x}) \tag{4.18}$$

Neither $P_{\text{Prediction}}(\mathbf{x})$ nor $P_{\text{Training}}(\mathbf{x})$ are known and therefore $w(\mathbf{x})$ must be estimated. For this the estimation technique that is also used for the non-i.i.d. correction in sec. 4.3.1 is used and this method called Domain Adapted Pearson Divergence Prior Estimation (DA-PEPE).

Learning from positive annotations only

Traditional classifier train models that distinguish between two or more classes by finding some rules. To be able to determine these rules all classes needs to be contained and labelled in the training data. But in the given setting, only observations from a single class are labelled. This class is defined as the positive class, and refers to all labelled observations with $\mathcal{P}$. Beside $\mathcal{P}$ the training data base consists of unlabelled data $\mathcal{U}$ containing observations of both classes, positive and negative class. This setting is known as PU-learning (Lee and B. Liu, 2003).

Nevertheless, it is still possible to train classifiers within these settings. For this, an approach proposed by Elkan and Noto (2008) was chosen. They showed that it is possible to obtain a classifier for the label y by training a classifier that estimates if an observation is labelled ($s = 1$) or not ($s = 0$) by introducing class-costs during the training.

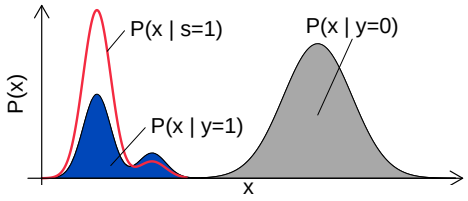
Cost-based classification training was originally developed to be able to include information about the cost of a false classification into the training process (Elkan, 2001). A typical use-case is the estimation if a customer will be able to pay back a loan. Granting the loan and loosing all the money will be more expensive than not granting the loan. Therefore the classifier should be restrictive with his predictions. These cost are usually incorporated by class-wide weights, but it is also possible to import them with observations-based weights. This allows to use these technique with all classification algorithms.

For PU-learning, the costs $c_{\mathcal{P}}$ and $c_{\mathcal{U}}$ are chosen based on the class prior $P(y = 1 \mid \mathcal{U})$ and the positive sampling rate $\eta = \frac{n}{n+n'}$, with n and n' being the numbers of labelled and unlabelled voxels respectively. They are chosen in a way that the classification error for the classifier is minimized for the estimation of the observation label y. The costs used in the experiments to this algorithm are similar to the ones proposed by Marthinus C du Plessis, Niu, and Sugiyama (2014), accounting for the greedy optimization strategy of random forests, and the convex loss function:

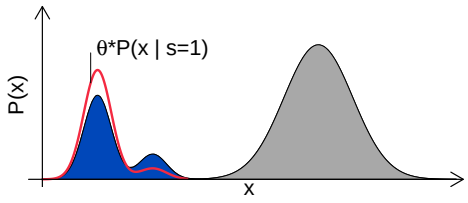
$$c_{\mathcal{P}} = \frac{4 \cdot \pi}{\eta} \tag{4.19}$$

$$c_{\mathcal{U}} = \frac{1}{1 - \eta} \tag{4.20}$$

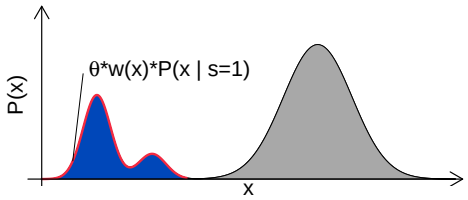
These costs are usually calculated globally, i.e. the costs are the same for all samples of one class, to which is further refered by the term ‘global mode’. Compared to other scenarios, voxel-wise learning offers the advantage of dealing with image-wise grouped (or batched) data. Since π depends heavily on the stage of the pathology, as well as the physiology of a patient it can vary within a wide range, even within a single training data base. Instead of using the mean class prior, as it usually done, a ‘batched mode’ it proposed. For this, the cost assigned with every observation is chosen based on π of the originating image.



(a) Original distributions



(b) Distribution with global weight Θ



(c) Distributions with observation weight $w(x)$ and global weight Θ

Figure 4.10: Visualization of the scheme of PEPE and DA-PEPE. (a) shows the distribution of the unlabeled data with blue indicating positive samples. The red curve shows the distribution of the positive annotated examples. (b) PEPE: The difference between both distributions is minimized by introducing a global weight factor Θ. A complete matching is not possible due to the sampling bias. (c) DA-PEPE: The sampling bias is corrected by weighting the labelled samples with $w(x)$ prior to the estimation of Θ.

4.3.3 Learning from Bag-Wise-Annotations

The third evaluated method was training a voxel-wise classifier based solely on the ratio of tissue classes for each patient and the unlabelled images. Compared to partially annotated images with SURs this offers a even faster method for annotation. Especially if the ratio of tissues are already known, for example because the tumour load was determined after surgery. It allows further to train classifiers for characteristics that cannot be seen images because the information is hidden in the multi-spectral dimensionality of the data.

Training a classifier from class ratios of bagged observations is called LLP Kuck and Freitas (2005). As already discussed in chapter 2.4.1 and 3.4, most of these methods are SVM-based and are only formulated for two-class problems. Therefore a new algorithm, Label Proportion Forest (LP-Forest), that is based on RDF was proposed, inheriting all the positive aspects of these algorithms. This includes the multi-class capability, the strong performance on medical image segmentation, and the possibility to deduce feature importance.

For this algorithm, the training data $\mathcal{T}$ consist of m images. Each of these images is seen as a bag of data $\mathcal{B}_i$ which consists of n_i different observations (voxels). Beside the observations, the probability of each class c is known for each bag:

$$\pi_i^c := P\left(y = c \mid \mathcal{B}_i\right) \tag{4.21}$$

Similar to other RDF-based classification algorithms, the model that is trained by the classification algorithm consists of multiple decision trees (Amit and Geman, 1997; Breiman, 2001).

LP-Forest prediction

To predict a new observations $\mathbf{x}_{\text{new}}$ it is passed down each tree. Starting from the root nodes N_0, $\mathbf{x}_{\text{new}}$ is passed from each node N_i either to the left or right child node $N_{i'}$, depending on the result of the splitting rule τ_i. This is repeated until a leaf node is reached. The vote that is saved within this leaf node is then used as vote for $\mathbf{x}_{\text{new}}$. The votes of all trees within the forest are then combined. The final classification decision of LP-Forests are obtained by majority voting, i.e. the class with the most votes is the predicted class. A estimation of the classification sureness $P\left(y \mid \mathbf{x}_{\text{new}}\right)$ of a specific class y can be estimated by the percentage of trees that vote for y.

LP-Forest training

The first step in the training of a LP-Forests is the bagging of the training data (Breiman, 1996). Each decision tree of the forest is trained using a sub-sample of training $\mathcal{T}_i$ randomly drawn from all available training data $\mathcal{T}$. Instead of drawing single observation it is also possible to draw bags to avoid the introduction of false class ratios into the bags. This is especially useful it the size of the bags is relatively small.

The training of the decision trees DT_k is done bottom-up in an iterative way. Starting from the root node N_0 the training data $\mathcal{T}_{k,i}$ of each node N_i are split in two groups and passed to the child nodes based on the splitting rule τ_i.

This is repeated until the stopping criteria is fulfilled in which case the node is converted into a leaf node.

Each splitting rule τ_i is chosen to maximize the purity of the child nodes – or equally to minimize the impurity. To increase the diversity of the decision trees the space of all possible splitting rules is randomly reduced. For LP-Forests this is done in two ways: first the number of features ν_{Features} that are evaluated is limited and the specific features are randomly chosen at each node. Following the arguments of Geurts, Ernst, and Wehenkel (2006) the full search on the selected features is also omited and only a fix number ν_{Splits} of random splits using these features are tested. Beside the slightly increased classification accuracy induced by this sparse optimization, Criminisi, Shotton, and Konukoglu (2011) showed that these optimization strategies allows to use a infinitive feature space.

Impurity definition The main difference between LP-Forests and other RDF-based algorithms is the definition of the impurity. Most impurity definitions are based on properties of single observations. For classification tasks usually the observation label y is used, which are not available the context of LLP. Using loss functions for SVM-based algorithms for LLP that were previously proposed is also not working with the canonical training process. The greedy and local optimization that used for these algorithms prevents a loss function that depends on the global distribution of data (J Ross Quinlan, 2014). There are some algorithms that allow to train decision trees without greedy optimization, like the work of Norouzi, M. Collins, et al. (2015). But these algorithm are usually less time efficient. Further, they increase the dependency between trees in a forest which reduced the classification accuracy Geurts, Ernst, and Wehenkel (2006).

Therefore a new new measure for the impurity of a node for this learning setting is proposed. This impurity is based on two assumptions:

Assumption 1: The distribution of a feature depends only on the label and is independent of the bag to which the corresponding observation belongs to, e.g.

$$P_{\mathcal{B}_i}(\mathbf{x} \mid y) = P(\mathbf{x} \mid y) \tag{4.22}$$

This assumption basically means that the elements of each class are the same independent of the bag.

Assumption 2: The data of all bags are sampled i.i.d. .

Combining assumption 1 and 2 allows to conclude that the ratio of observations belonging to given class and a given bag should be independent of the features that are evalauted. It is possible to estimate this ratio, based on the class prior π for each bag which is named the bag class contribution ratio ρ_i^c of bag i for class c:

$$\rho_i^c = \frac{n_i \cdot \pi_i^c}{\sum_{k=m} n_k \cdot \pi_k^c} \tag{4.23}$$

Similar to ρ it is also possible to calculate the contribution σ_i^k of a bag i to the training data $\mathcal{T}_k$at node k

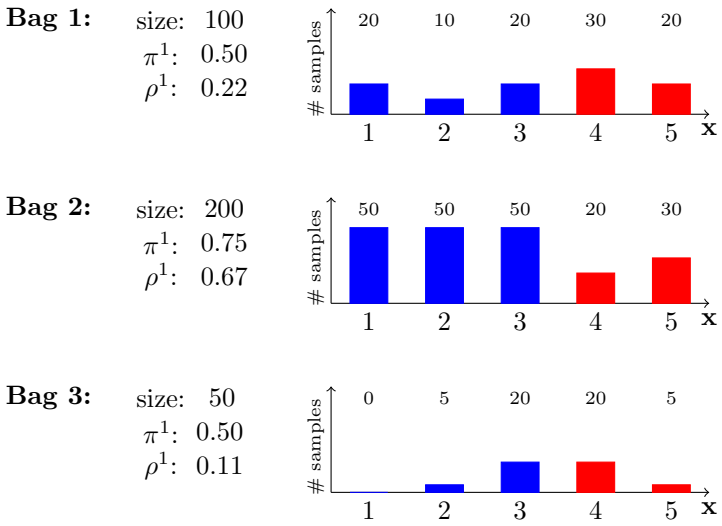


Figure 4.11: Exemplar set of a two-class training data set, consists of three bags with different number of samples (size), class ratios (π^1), and therefore contribution ratios ρ^1. Also the data distributions are given. The class border is $\tau = 3.5$. Calculations of the absolute error are given in Table 4.3.

$$\sigma_i^k = \frac{\mid \mathcal{T}_k \cap \mathcal{B}_i \mid}{\mid \mathcal{T}_k \mid} \tag{4.24}$$

If a bag contains only observations of a single class both ratios should be equal if both assumptions are fulfilled. It is therefore possible to measure the impurity I of a node by calculating the difference between those two ratios:

$$\Delta(i,k,c) = \sigma_i^k - \rho_i^c \tag{4.25}$$

$$I_k^c = \sum_{i=0}^{m} \Delta^2(i,k.c) \tag{4.26}$$

This impurity is called Class-dependent, Ratio based Impurity (CRI). To allow small bags that cannot cover the complete feature space, the impurity is calculated by always using only those bags that contain observations. Following the previous definition the obtained impurity is class-depended. To determine the best threshold for a node the smallest CRI of all classes is chosen(Best Ratio based Impurity (BRI)), e.g.

$$I_k = \min_c I_k^c \tag{4.27}$$

This behaviour is quite different from the behaviour of other impurity measurements like the Gini Impurity (Breiman et al., 1984) or Entropy Impurity (J. Ross Quinlan, 1986). Other than these impurities, CRI is always optimized for a single class. This can be be disadvantageous in a multi-class setting. In this case it might be preferable not to separate a single class from the other

TABLE 4.3
EXAMPLE OF LLP IMPURITY

τ	bag	samples (left of split)	σ^1	ρ^1	$\mid \Delta^1 \mid$
2.5	1	30	0.22	0.22	0.00
	2	100	0.74	0.67	0.07
	3	5	0.04	0.11	0.07
	sum	135			**0.14**
3.5	1	50	0.22	0.22	0.00
	2	150	0.67	0.67	0.00
	3	25	0.11	0.11	0.00
	sum	225			**0.00**
4.5	1	80	0.27	0.22	0.05
	2	170	0.58	0.67	0.09
	3	45	0.15	0.11	0.04
	sum	295			**0.18**

Exemplar calculation of the ratio error for the example given in Figure 4.11. For three different thresholds τ the number of samples in the left child node, the bag size ratio, the expected ratio for class ρ^1, and the absolute error $\mid \Delta \mid$ are given for the observations below the threshold. The best overall error is achieved if $\tau = 3.5$.

data but trying to obtain two groups of nearly similar size (Breiman, 2001). To avoid such problems, artificial classes might be introduced into the search for a best split. This can be easily done by a linear combination of the corresponding ratios. For example. $\rho^{12} = \rho^1 + \rho^2$ would be used to calculate a superclass based from class 1 and 2. This allows a fast adaptation to multi-class problems.

Beside this, CRI offers the benefit of allowing the estimate of the vote for a leaf node by selecting the vote that minimize the impurity:

$$c_{\text{vote}} = \underset{c}{\operatorname{argmin}}\, I_k^c \tag{4.28}$$

In Figure 4.11 and table 4.3 an example is given. In this case the training data consist of three bag with different number of observations and different class ratios. Within the data are two classes, the border of these two classes is the feature value 3.5. From the table it can be clearly seen that the absolute error Δ is minimized if the threshold is set to the class border.

Stopping Criteria The stopping criteria is needed to prevent an over-fitting of the classifier. Typically this is achieved by checking for one or more of the following criteria:

- No split was found that further decreased the impurity.

- A pre-defined, maximum tree depth was reached.
- A further split would lead to nodes with less than a pre-defined minimum number of samples.

Each of these stopping criteria can be used for the training of LP-Forests. Additionally, a new stopping criteria is proposed and included:

- The number of bags that contain no observation within the current node exceeds a pre-defined limit.

Allowing some bags to be empty is especially important if the training data contains many bags with only few training samples. There is a high probability that not all bags cover every area of the feature space, even if the bags are sampled i.i.d. . Allowing some bags to contain no samples after splitting increases therefore the chance to find a meaningful threshold. But empty bags also decrease the certainty of the proposed impurity. Due to that the number of empty bags that is allowed at most is limited to a certain threshold.

Parameter Most of the parameters for the configuration of LP-Forests are shared with other RDF algorithms, but some new and unique parameters are also used. For example, the **number of trees**, **maximum tree depth**, and **minimum number of observations at a leaf node** allow to adapt the learning algorithm to the training data. This is also true for **features per node** and **evaluated splits per node**. Although this implementations makes use of the partly optimization strategy suggest by Geurts, Ernst, and Wehenkel (2006), it is possible to implement any other optimization strategy, since the main difference between LP-Forests and other RDF algorithms is the impurity definition. It is therefore also possible to use a wide range of **splitting function**. Furhter, this implementation is based on a linear split on a single feature at each node to keep the algorithm simple and comparable to the other results reported in this thesis. But using other splitting functions, like combinations of features (Criminisi, Shotton, and Konukoglu, 2011), boosted combinations (Hastie, Tibshirani, and Friedman, 2013), or Principal Component Analysis (PCA)-based features (Rodriguez, Kuncheva, and Alonso, 2006).

Different from other features is the usage of the **bagging ratio**. It is differently interpreted, although the influence is likely the same, i.e. determining the coupling of different trees. For this reason, in the proposed algorithm the bagging is done on the bag-level instead on observation base to avoid wrong ratios for different bags. Another different parameter is the previous discussed **maximum number of empty bags** which correlated to the minimum number of observations at a leaf node. But it also allows a relaxation of the made assumptions.

4.4 Assessment of image acquisition

Not only local information – like tumour probability – are important during the clinical routine, but also general or global information. Examples for such information are the expected survival rate, the expected response rate for a specific treatment, but also meta-information like the suitability of a specific imaging modality for a given diagnosis (Stupp et al., 2006).

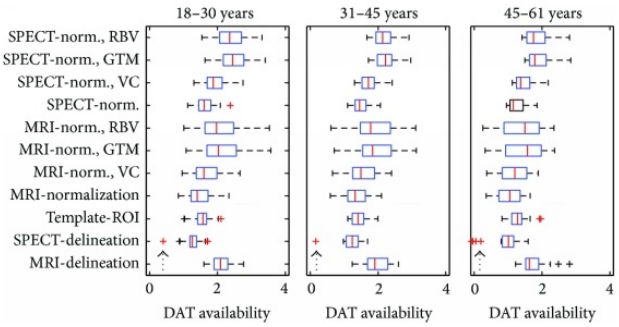


Figure 4.12: Example of ROI-based comparison. Taken from Yin et al. (2014) and adapted, it shows the influence of the creation method of the used ROIs by comparing the results obtained by different annotation schemes.

4.4.1 Multi-rater region of interest comparison

Answering the question if and how good two types of tissue can be separated is important to chose the right imaging modality during clinical routine. This information is often assessed by Region of Interest (ROI) based comparisons. The intensity distributions of ROIs placed in both types of tissue are compared and allow to asses how large the contrast between those two types of tissue is. An example of such a comparison is given in Figure 4.12.

ROIs of multiple observers are often used to avoid the dependency on a single observer and to improve the reliability and reproducibility of a study (Hallgren, 2012). This allows to report mean results which are not based on a single observer and also allows the assessment of inter-rater reliability as done by Davenport et al. (2013), for example.

If multiple ROIs are within a single image return different results, the remaining question is which one reflects the true mean distribution. A well-established solution is to report the mean value of all ROIs within a single image. But this weights all ROIs equally, even though some might be placed by more experienced observer that other. It is also possible that some ROIs are placed in a wrong place and should be considered as outlier. Therefore, using the mean might not be the best possible solution as it has already been shown in the case of image segmentations (Warfield, Zou, and W. M. Wells, 2004).

Therefore a new method to combine multiple ROIs is proposed. Instead of fusing already reduced data, to calculate a mean value the distribution of intensity values within each ROI is used to calculate a new and optimal common ROI. The effect of outlier is reduced by weighting the results of each observer based on the agreement between all data.

The method takes n different ROIs $\mathcal{R}_i$ and estimates a single target ROI $\mathcal{R}_t$ by weighting each ROI:

$$\mathcal{R}_t = \sum_{i}^{n} w_i \cdot \mathcal{R}_i \quad . \tag{4.29}$$

The summing of the different regions is achieved by combining the voxels of all ROIs and weighting each voxel with the weight w_i that belongs to the ROI

of its origin. The weights are estimated in an iterative procedure, starting with an mean initialization, e.g.

$$w_i = \frac{1}{n} \tag{4.30}$$

Taking this as starting condition the weights are iteratively updated. For this, a density function D_i is estimated from each ROI and the distance d_i between $\mathcal{R}_i$ and $\mathcal{R}_t$ is calculated by summing the difference of the distributions:

$$d_i = \int_{-\infty}^{\infty} \mid D_i(x) - D_t(x) \mid \mathrm{d}x \tag{4.31}$$

Assuming that the image data are binned, the integral can be replaced with a sum. It is further possible to ignore the bin-size ($\mathrm{d}x$) as this would be cancelled out in later step:

$$d_i = \sum_x |D_i(x) - D_t(x)| \tag{4.32}$$

The distance is calculated for a fixed range of points, which are equally distributed over the complete observation range. After calculating all d_i the weights are updated according to

$$w_i = \frac{\frac{1}{d_i}}{\sum \frac{1}{d_i}} \quad . \tag{4.33}$$

ROIs that are more similar to the target distribution are therefore given more weight which increases the overall similarity as the influence of outlier is reduced. The reweighing is repeated until the change of all weights is below a given threshold. Figure 4.13 visualizes the whole algorithm for fusion.

4.4.2 Classifier-based information assessment

The traditional way to estimate if a imaging modality can be used to differentiate between two tissue types are ROI-based comparison. The intensity values of all voxels within the ROIs are compared allowing to asses intensity differences between tissue classes.

A drawback of this method is the fact that only differences in the intensity values are captured while the result is not affected by texture differences. But these differences are important for the human vision (Mayhew and Frisby, 1978), as indicated by Figure 4.14. It is therefore important to asses not only the intensity differences by also the differences in the texture.

I therefore propose to asses the differences between two tissue classes using machine learning methods. Instead of comparing only the intensity values of each modality and tissue classes, for each modality classifiers are trained that separate the tissue classes. Since this allows to include multiple features and texture descriptions, this approach allows to simulate the decision process made during the diagnosis more closely.

As an additional advantage, the so trained classifiers can not only be used for the assessment of information content in each modality but also for the computer aided tissue characterization. Depending on the type of classifier that is used either full segmentations or probability maps can be generated that then

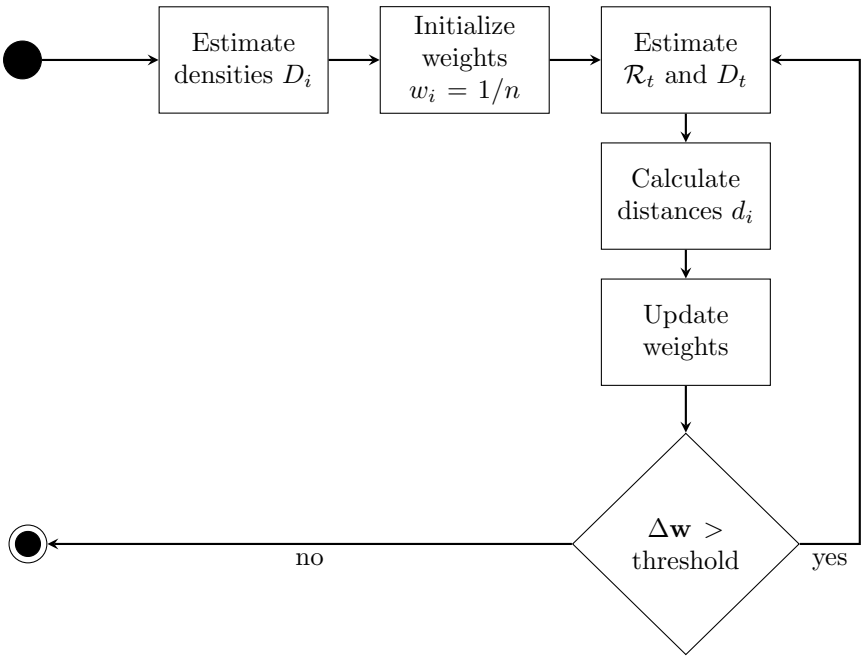


Figure 4.13: Simple work-flow for the proposed algorithm for fusing multiple ROIs. The weights are updated until the sum of differences caused in the current run are lower than a pre-defined threshold.

might be used to indicate areas of specific tissue types. This might be used as an additional source of information during the diagnosis, indicating areas where for example tumour might be while leaving the final decision to the physician.

Since complete tumour segmentations are error-prone and would include further bias depending on the modality that was used for labelling, it was decided to train and evaluate the classifiers using only small ROIs. This has the additional benefit that the necessary effort and the results are more comparable to those of traditional ROI-based studies.

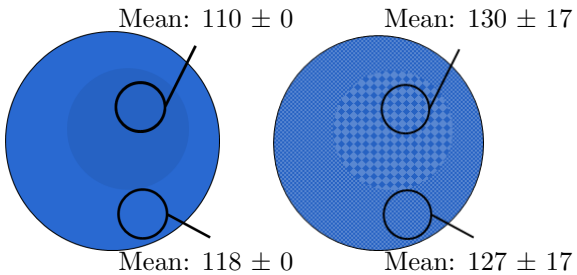
To evaluate the quality of the proposed automatic tumour segmentations the overlap of the segmentation with the ROIs is measured. This raises again the problem of sampling bias. A score might not represent the result that would be obtained on the whole image if all voxels within the ROI are treated equal because some voxels might be over- while others are under-represented. To correct this error a weighted Dice Score is used:

$$wDice = \frac{2 \cdot \sum_{i \in A \cap B} w(x_i)}{\sum_{i \in A} w(x_i) + \sum_{i \in B} w(x_i)} \quad . \tag{4.34}$$

With A, B being the sets of indices of the voxels with the corresponding label in both images and the weighting

$$w(x_i) = \frac{\text{probability of } x_i \text{ in whole image}}{\text{probability of } x_i \text{ in ROIs}} \quad . \tag{4.35}$$

The weighting used for wDice is equal to the weighting factor that is used to train the classifier; therefore the same estimation is used in both cases.



(a) High contrast (b) With texture

(c) Dark in middle (d) Dark at border

Figure 4.14: Human vision incorporates more information than only contrast. For example, it is easier to separate the two circles in (b) than in (a) although the pure contrast is higher in the latter one. Also, the ROI-placement influences the result as shown in (c) and (d)

Classifier set-up

To compensate for the error that was made due to the incomplete annotations, the technique described in section 4.3.1 was used. The ROIs were therefore treated as SURs, and the classifier were trained using the DALSA scheme, i.e. weighting each observation with a specific weighting vector w that corrects the sampling bias made during the annotation.

Feature selection is especially important as some textures might be irrelevant for the separation between two different tissue types. Therefore, if all features contribute with the same importance, noisy texture features might reduce the classification accuracy. It was therefore decided to use RDF-based algorithms as these include feature selection (Genuer, Poggi, and Tuleau-Malot, 2010; Qi, 2012). Beside this, these algorithms has the advantage of offering some insight into the importance of different features. Thus, it is possible to estimate which features are relevant.

Although this approach allows to use a wide range of features it was decided to use only basic features so far. The main reason for this decision is to make the results comparable to previous findings from the literature. Therefore only intensity values of the original images, of gaussian smoothed images, and the difference of gaussian smoothed images were used as features.

4.5 Summary of Methods

Within this chapter, the methods that were used to address the limitations described in the State-of-the-Art chapter, are described.

First a pipeline for the segmentation of a brain tumour based on kNN classifier is described, with respect to the used features and the whole training process. This pipeline is then combined with different normalization algorithms which are also described to assess the influence of different normalization algorithms. Similar to this, two pipelines based on canonical RDF and ExtraTrees are described, which allow to quantify the improvement by using ExtraTrees.

After this, two algorithms are described which can be used to reduce the influence of variability. The first algorithm, called IDAL, is designed to find the best suited training images for each specific image which is segmented. With these specific images, a classifier is then trained which is specific for the image that is segmented. In addition to this, an algorithm for semi-automatic segmentation is proposed, that allows to incorporate a pre-trained RDF-based classifier. Instead of training a new classifier, the existing classifier is update with each interaction. This is done in order to include the knowledge from the training data into semi-automatic segmentation process.

Three different methods are proposed to solve the challenge of creating a new training dataset. In order to reduce the amount of necessary annotations, each methods allows to train a classifier with only partly annotated data. The first method, DALSA, allows the training from sparse annotation of each tissue class. This leads to a sampling bias, which is corrected during the training according to DALSA using domain adaptation techniques. The second method adapts the DALSA scheme, but extends it by incorporating methods to learn only from positive annotations. The so developed methods allows to train a classifier using only annotations of one class, leaving most of the image without

annotation. In order to achieve this, a method to estimate the volume of each class is proposed. The third method can then be used to train a classifier based solely on the information about the volume of each tissue class. For this, the first random forest based training algorithm for this scenario, called Learning form Label Proportions has been proposed.

In the last section two methods are described that helps to improve to assess the information content of imaging modalities. The first method allows to combine the sparse annotations of multiple rater. Similar to the STAPLE algorithm, the combination is more similar to all, but contrary to it, it does not require a full annotation of the tissue. After this, a pipeline based on the previous described DALSA algorithm is proposed that is used to assess the information in two different CT-based imaging modalities by measuring it using the classification score. This allows to incorporate information about the intensity value as well as textures.

5

Experiments and Results

In the previous chapter different methods have been proposed to either validate a starting hypothesis or to overcome a certain challenge within the field of medical tissue characterziation. Within this chapter, the previosu described methods are evaluated with multiple experiments.

A variety of different datasets are used for the experiments to avoid over-fitting to a single dataset. At the same time, some of the datasets are used in more than one experiment. Therefore, all datasets are described in the first section, section 5.1. The following sections reflect the sections from the 'Methods' chapter, i.e. section 5.2 is about the influence of different algorithms, section 5.3 is about algorithms for variability control, section 5.4 is about learning from weakly annotated data, and section 5.5 is about image modality selection.

Each subsection corresponds to the experiments of a single algorithm or hypothesis. It starts with a description of the experiments that were carried out and after this the results of these experiments are reported, in order to keep the description of the experiments and results for a single algorithm close together.

5.1 Data collections

Different datasets are used during the experiments. Each of this datasets does have different properties. Beside the organ that is shown, which is mostly the brain, another important question is the availability of the dataset. While public datasets allow a comparison of different algorithms, the use of such public datasets might lead to an additional bias within the results. During an invited keynote atthe 2014 NIPS conference[1] Dwrok (2014) cautioned that the constant testing and retesting on public datasets can lead to an over-fitting of methods to a particular dataset. It is therefore important to a) use different public datasets and b) validate or even better develop new methods using private data (Dwork

[1]The yearly NIPS conference (Neural Information Processing Systems conference) is one of the biggest and renowest machine learning conferences.

TABLE 5.1
DATASETS

Abb.	Section	Modalities	Public	Organ / Disease
DS-1	5.1.1	MRI + DWI	No	Brain with High Grade Gliomas
DS-2	5.1.2	Perfusion CT + Dual Enegery CT	No	Pancreas with pancreatic tumours
DS-3	5.1.3	MRI	BraTS[a]	Brain with High Grade Gliomas
DS-4	5.1.4	MRI	ISLES[b]	Brain with ischemic stroke lesions
DS-5	5.1.5	mixed	modified public	no medical data

Overview of the different datasets that are used during the experiments.

and Roth, 2014). This allows an unbiased comparison and reduces the chance of false findings if the new method or new thesis is tested on public data.

Table 5.1 summarizes the different datasets that were used during the experiments for training and testing. A more detailed description of each set, including the origin and applied preprocessing, is then given in the following sections.

5.1.1 In house dataset - High Grade Glioma (DS-1)

The first private dataset was acquired in cooperation with the Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology and the Institute of Automatic Control at the Silesian University of Technology, both located in Gliwice, Poland. The cohort included 19 patients with high-grade glioma, and all images were acquired during clinical routine. This makes this dataset realistic and relevant for the transformation of findings into the clinical routine.

All images were taken on a single 1.5 T Siemens Avanto (Siemens Healthcare, Erlangen, Germany) using a standard protocol with a duration of less than 20 minutes per examination. Three different native images were acquired and included in this dataset:

$T1_w$c T1-weighted image with contrast enhancement. In plane resolution is between 0.55×0.55 mm and 0.65×0.65 mm. The slice distance is 6 mm. The applied contrast agent leads to an enhancement of areas with active tumour.

FLAIR The T2-weighted FLAIR image has the same resolution as the $T1_c$c image. Using a FLAIR sequence suppresses free water, therefore showing free fluids dark. This sequence is usually used to assess edema which is hyperintense in FLAIR images.

DTI A diffusion tensor image is constructed from a DWI. The DWI parameters were: single-shot spin echo EPI sequence, echo time = 95 ms, repetition

time = 3.6 s, slice thickness of 4 mm, and pixel spacing of $1.8 \times 1.8 \times 5.2$ mm. Two shells with 48 directions were acquired.

Several commonly used parameter maps are calculated from the DTI: fractional anisotropy (FA), relative anisotropy (RA), axial diffusivity (AD), radial diffusivity (RD), clustering anisotropy (CA), and mean diffusivity (MD). These parameters are described in more detail by Beaulieu (2009). All parameter images were calculated with and without free water elimination (FWE) (Pasternak et al., 2009). Using both versions, with and without free-water estimation, allows to use freewater-information without the danger of dismissing important information. A free-water map (FW) and an extracted b0-image (B0) are also included. All diffusion calculations are conducted using MITK. Example images are shown in Figure 5.1 and Figure 5.2.

In addition, three raters (one expert radiologist and two medical students) segmented independent sets of sparse annotations called SURs (c.f. Table 5.2, 'Main'). Each rater was blindfolded to the complete tumour segmentation and the SURs created by the other raters. SURs were defined for each of five different clinically relevant tissue classes including highly proliferative tumour parts (active tumour, e.g. as potential target for biopsies), necrosis (e.g. as an indicator of tumour grade and poor target for biopsies), and lowly proliferative tumour parts (edema, e.g. as part of the peripheral tumour border) in addition to healthy tissue and cerebrospinal fluid (CSF). The task of the raters was to annotate small areas which are typical for each tissue class. If possible, areas close to tissue borders should be included if they were clearly distinguishable from the neighbouring class. No other restriction in terms of size, number of ROIs per patient, relative location, or number of annotated slices was made. It took less than five minutes to create these small 2D-ROIs, which were usually located in one or two single slices of an image. Figure 4.8, page 53 shows examples of SUR annotations. The mean coverage ratio of segmented voxels to brain voxels for the SURs created by rater 1, 2 and 3 were $0.53\,\% \pm 0.23\,\%$, $0.41\,\% \pm 0.11\,\%$, and $0.18\,\% \pm 0.05\,\%$, respectively. The minimum and maximum coverage ratios were 0.24 %, 0.17 %, and 0.08 % and 1.22 %, 1.16 %, and 0.33 %, respectively. On average, $2.6\% \pm 1.5\%$, $1.6\% \pm 1.1\%$, and $1.0\% \pm 0.5\%$ of the tumorous tissue were covered by SURs.

To analyse the effect of varying SUR placement strategies, a medically trained expert created four additional different sets of SURs using different labelling strategies (c.f. Table 5.2, 'Type 1' - 'Type 4'). The mean coverage ratio of segmented voxels to brain voxels for these SUR sets was $0.2\% \pm 0.1\%$ (maximum: 0.63%, minimum: 0.06%). The SURs of Type 1 were the smallest ($0.12\% \pm 0.03\%$) and those of Type 2 the largest ($0.28\% \pm 0.10\%$). Those of Type 3 and 4 were similar in size scattering around 0.18%. The mean tumour coverage ratio was $1.3\% \pm 0.1\%$ across all types.

For each patient a single time point is selected from all available time points. As most patients underwent surgical treatment either the last available pre-operative image (13 cases) or the latest available time point (6 post-operative cases) if no pre-operative image was available was chosen. This was necessary because it is more difficult to distinguish tumorous tissue from post-operative traumata, swelling, blood, and other non-tumorous tissue classes do have a similar appearance and it is difficult, even for expert radiologists, to differentiate them from active tumour, necrosis, or edema caused by tumour growth.

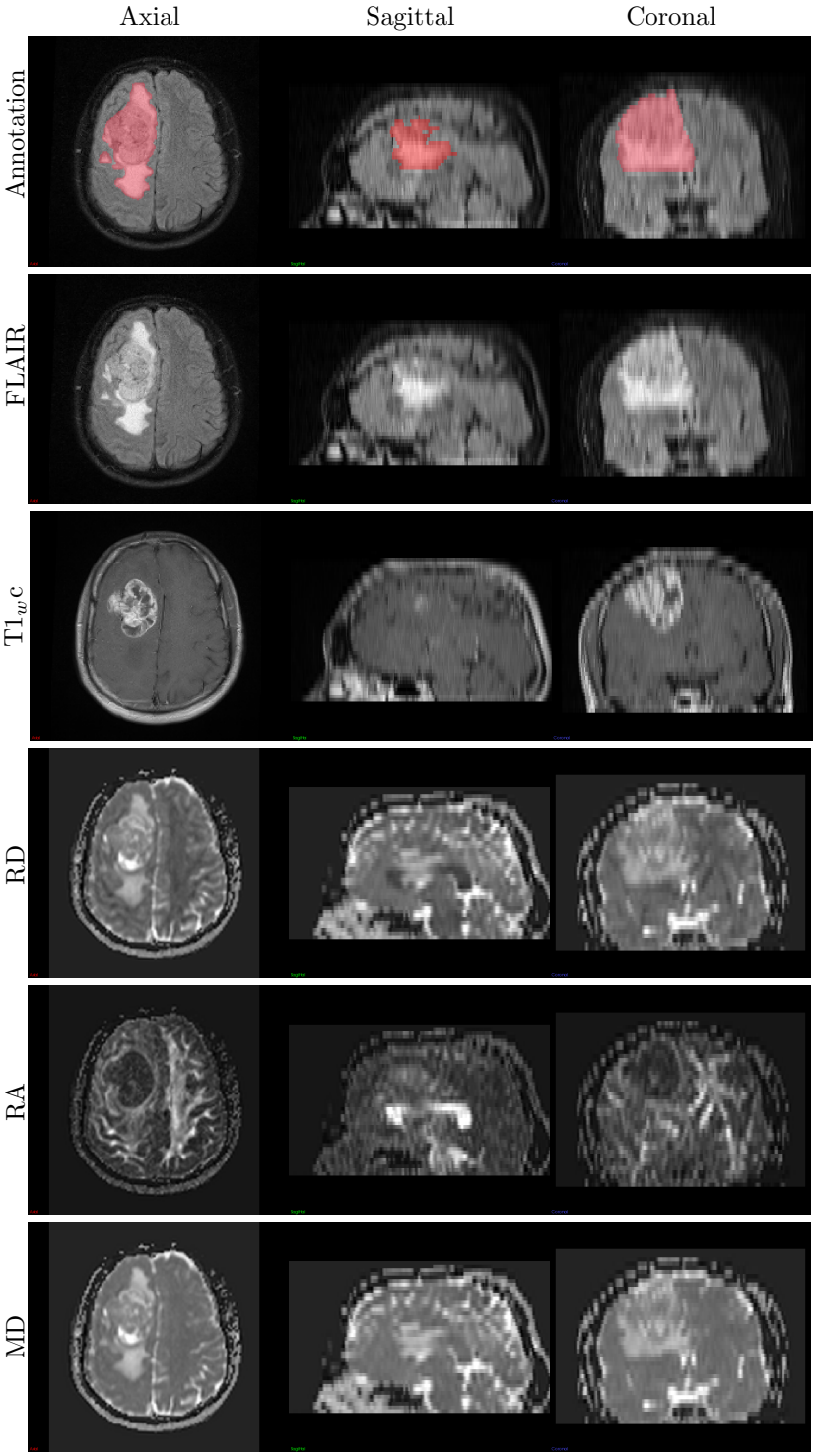


Figure 5.1: Example images of the different contrasts that are provided with the data of the DS-1 dataset. For more detailed description see Figure 5.2.

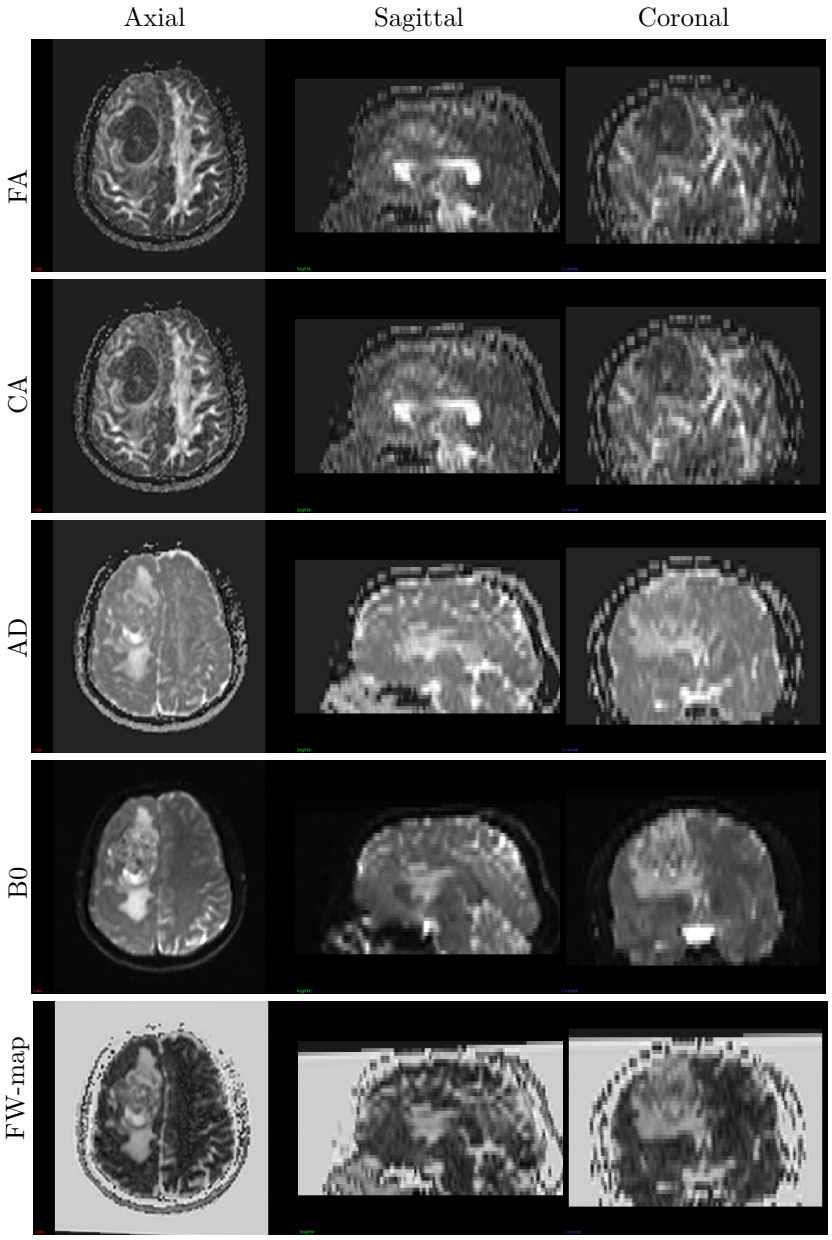


Figure 5.2: Example patient from DS-1. For each contrast three different views (axial, sagittal, and coronal) with the same slices are shown. The annotation is shown as red overlay over the FLAIR image, all DTI-derived parameter are calculated with MITK.

TABLE 5.2
Labelling strategies

Type	Description	Diameter	Location
Main	1-3 SUR per class	rater dependent	covering bordering as well as central tissue areas
Type 1	1 SUR per class	$6-14\,\mathrm{mm}$	arbitrarily varying
Type 2	3 SURs per class (different slices)	$6-14\,\mathrm{mm}$	covering bordering as well as central tissue areas
Type 3	3 SURs per class (different slices)	$6-14\,\mathrm{mm}$	covering central tissue areas only
Type 4	3 SURs per class (different slices)	$6-14\,\mathrm{mm}$	covering bordering tissue areas only

Description of different SUR labelling strategies. A complete set of SURs was created for each strategy.

All images were rigidly co-registered to the FLAIR image and then resampled to a common resolution of 1×1 mm in plane. The slice thickness was set to 3 mm, compromising between resampling artefacts and the number of slices that need to be labelled. A semi-automatic brain-mask is created using by first running the approach of Bauer, Fejes, and Reyes (2012) and later manually correcting the errors. This was necessary because tumours growth significantly affects the performance of the brain segmentation algorithm.

Experts segmented the GTV manually based on $T1_w$c and FLAIR. Within this work, GTV is defined as the area that covers edema, contrast-enhancing areas and necrosis. To obtain consistent annotations and to differentiate between actual tumorous tissue and tissue of similar appearance the segmentation is refined in multiple runs. To help this process, 'Tumor Progression Maps' (TPM) are used (Weber et al., 2014). This technique shows corresponding slides of various times steps next to each other. Also changes in the segmentation of each time step are highlighted. This allows to find differences between two time points and verify if these changes results in biological changes or simply caused by inconsistent segmentation.

5.1.2 In house dataset - CT Images (DS-2)

Pancreatic Tumour

Beside the main dataset which consisted of images from brain tumour patients (section 5.1.1), a second dataset was gathered in cooperation with the University Hospital Heidelberg. This set contained images of patients with pancreatic carcinoma, which were histological validated for all 20 patients.

The imaging was performed with an dual-source DECT scanner (Somatom Definition Flash; Siemens Healthcare, Forchheim, Germany). The acquisitions were started about five seconds after the contrast agent was injected, and 34 evenly time-displaced, axial acquisitions were taken for each patient. This took roughly 50 seconds. Each acquisition was made using tube potentials of 80 kVp and 140 kVp – therefore allowing the calculation of CT-perfusion maps and dual

energy CT maps. To reduce the applied dose, only a single slice was imaged and not a complete 3D volume.

A reference image was selected for each patient out of the 34 images of each series. Based on this information, a medical doctor created a manual segmentation of the pancreas. Furthermore, two ROIs were placed in tumorous and healthy tissue, respectively. The other time-points are then non-rigidly motion corrected based on Demons deformable registration. This is done with a software developed at the University Hospital Heidelberg (Klauss, Stiller, et al., 2012). If the motion correction failed for a time point, this time point was excluded from further calculations.

Perfusion CT

The perfusion CT images were calculated by fitting a model to the previously described acquisitions. For the experiments a maximum-slope and a two compartment Patlak-model are fitted to the data. A detailed description of these models and the fitting method is given by Klauss, Stiller, et al. (2012) and Miles and Griffiths (2003). Using a software package developed at the University Hospital Heidelberg (Stiller et al., 2015; Skornitzke et al., 2015) the perfusion, permeability and blood volume was estimated from the 80 kVp images. Examples of these maps are given in Figure 5.3

Dual-energy CT

Similar to perfusion CT, dual-energy CT allows the estimation of different tissue fraction maps by fitting suitable models to the data. A common model is a three compartment model, which allows, for example, the estimation of iodine maps (T. R. Johnson, 2012). It is further possible to calculate linearly blended mixed images with a varying ratio. A ratio of 0.5 (M 0.5) is then noise-equivalent to a single-energy 120 kVp image. This allows to estimate images that would haven been taken with tube voltages which were not used during the acquisition.

These maps were calculated with software from the University Hospital Heidelberg. The necessary reference values for the iodine enhancement vector and the non-enhanced vector were taken from previous measurements and commercial DECT post-processing software (Syngo.-via, Liver VNC, Siemens Healthcare, Forchheim, Germany). For the calculations, a single time point was selected from all available time points according to the findings by Stiller et al. (2015). If this time point was not available the due to problems with the motion correction, the next available time point is selected instead. Based on this data, the dual-energy data of a single patient contained a 80 kVp, a 140 kVp, a M 0.5, and a iodine map image. Examples of these maps are given in Figure 5.4

Intracerebral Haemorrhages (ICH)

To have a data collective consisting of CT images of the brain, another patient collective is assembled. This dataset consists of 30 subjects suffering from spontaneous ICH. All subjects were treated at neurological and neurosurgical departments of the Heidelberg University Hospital from 2008 to 2015 and gathered retrospective. The inclusion criteria was a spontaneous ICH, patients with ICH associated with vascular malformations, tumour, ischemic stroke, or trauma are excluded.

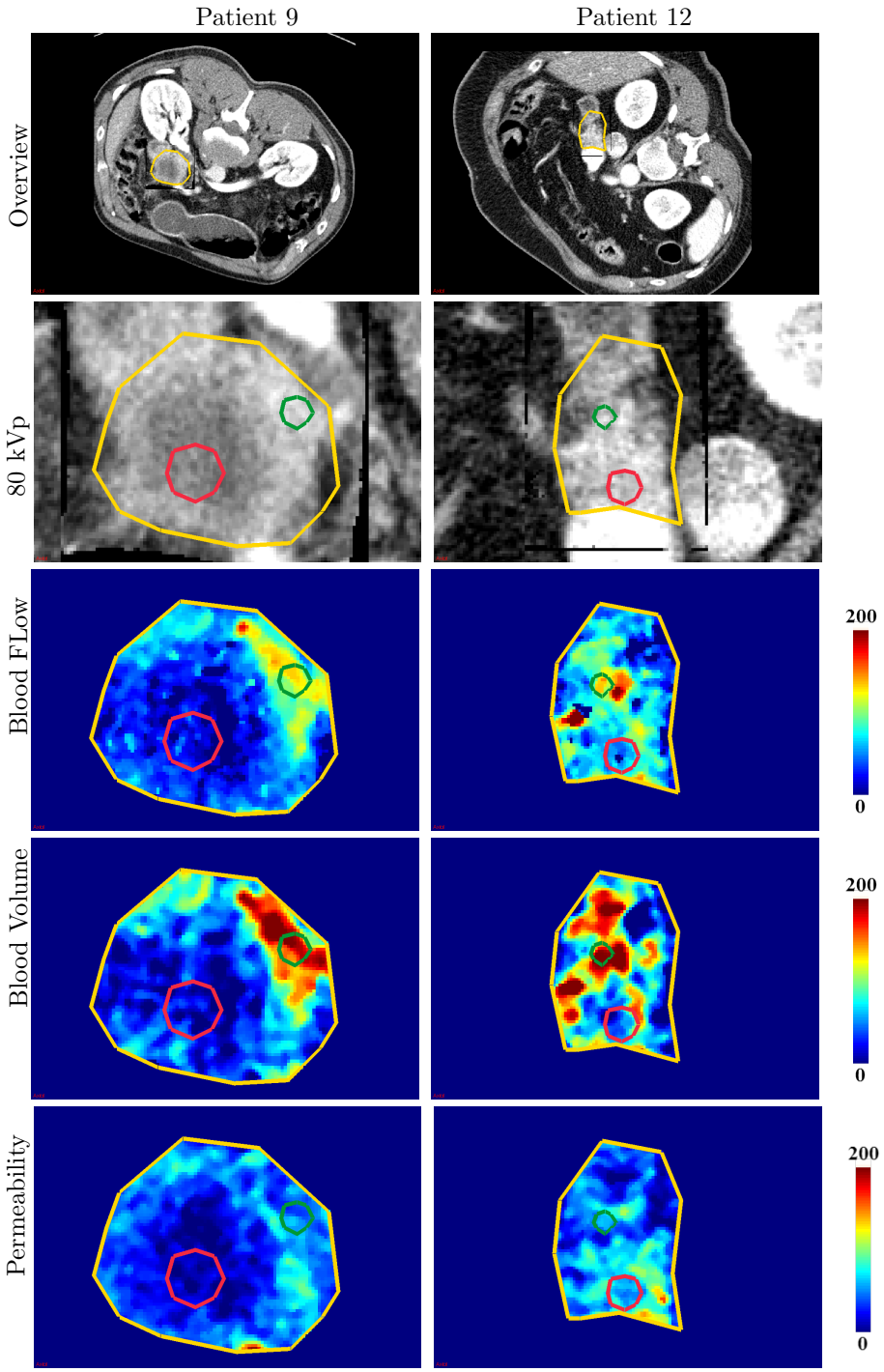


Figure 5.3: Example images of the different perfusion CT based contrasts that are used from DS-2. The first image indicates the area of the following images. The colour coding is as following: 'yellow': whole pancreas, 'red' tumorous, and 'green': healthy.

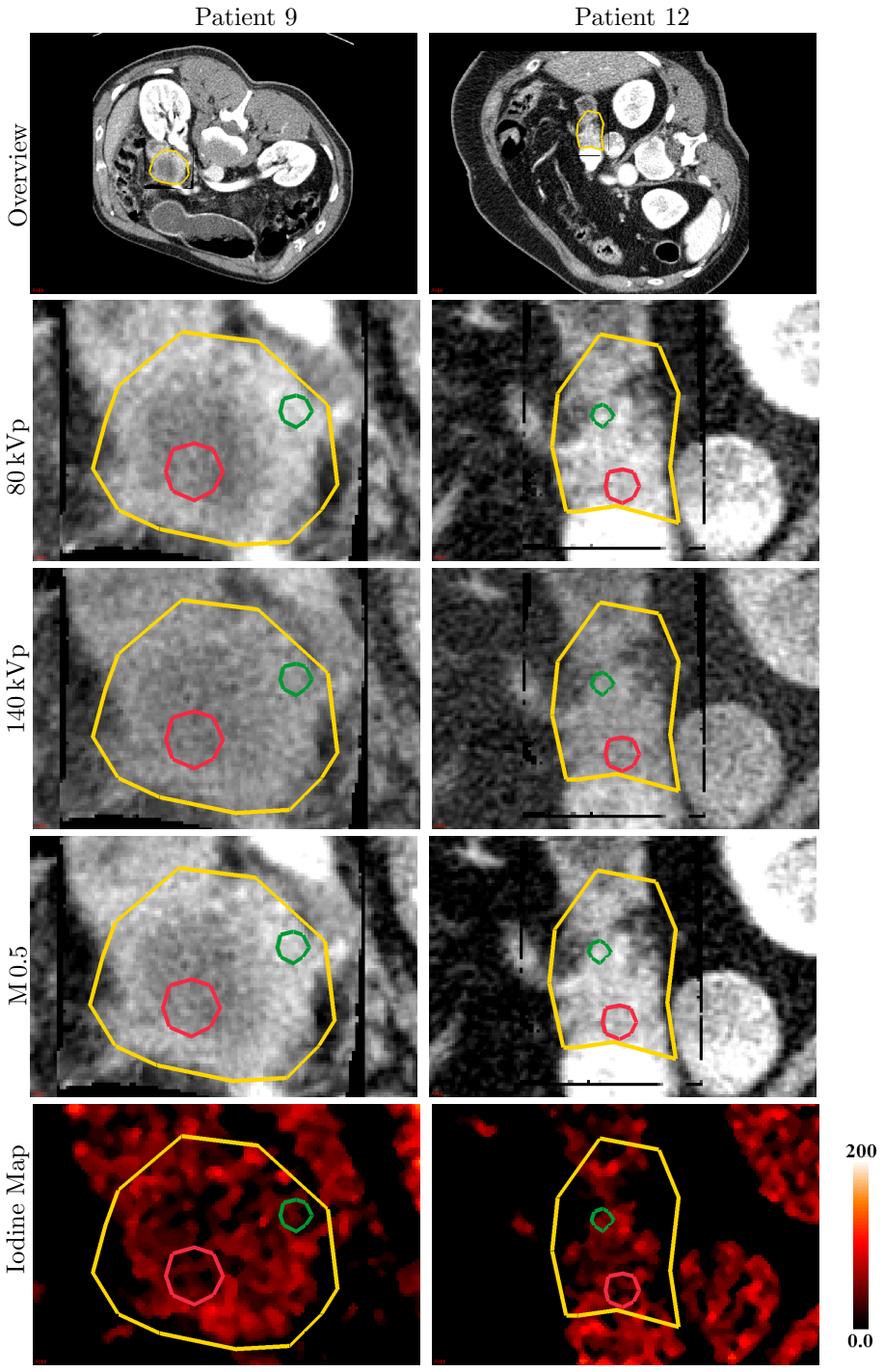


Figure 5.4: Example images of the different dual-energy CT based contrasts that are used from DS-2. The first image indicates the area of the following images. The colour coding is as following: 'yellow': whole pancreas, 'red' tumorous, and 'green': healthy.

One pre-operative CT image per patient without contrast agent is retrieved from the clinical image system. As these scans are usually taken as emergency treatment, there are significant differences between each scan. The slice thickness varies from 2 mm to 10 mm.

A medical doctor with more than five years of experience created a ground truth segmentation of the tissue. This includes annotations of CSF, Lesion, and other brain tissue (labelled as 'brain'). The ground truth was generated semi-automatic using a region-growing approach implemented in MITK and manually corrected if necessary.

5.1.3 Challenge dataset - BraTS challenge (DS-3)

The Brain Tumor Segmentation Challenge (BraTS) (BraTS, 2016) has been conducted as a side-event of the yearly Medical Image Computing and Computer Assisted Interventions (MICCA) conference. Starting 2012, this challenge became very popular within the tumour segmentation community, and the datasets are commonly used to verify new brain tumour segmentation approaches. The datasets of these challenges consist of two parts; a publicly available training set, consisting of several subjects with available labels and a test set with hidden ground truth. To test against the training set, the created segmentation needs to be submitted online and are then evaluated.

As suggested by the name, the main purpose of the challenge is the segmentation of brain tumours, specifically gliomas. MRI is used as imaging modality, and four different contrasts (c.f. Figure 5.5) are provided for every patient (B. H. Menze, Jakab, et al., 2015):

$T1_w$: T1-weighted images. Acquisition made sagittal or axial in 2D; slice thickness between 1 and 6 mm.

$T1_w$c : T1-weighted image with contrast enhancement. 3D acquisition with isotropic voxel size for most patients.

$T2_w$: T2-weighted images, axial acquisitions with slice thickness between 2 and 6 mm.

FLAIR : T2-weighted FLAIR images. Taken as 2D images, either axial, coronal or sagittal. Slice thickness between 2 and 6 mm.

The acquisition, pre-processing, and organisation of the data is described in more detail in (B. H. Menze, Jakab, et al., 2015) and (BraTS, 2016). The images were taken in four different places. The vendors of the scanner varied as well as the field strength (1.5 T and 3 T). To have all images within a single space per patient, they are co-registered to the $T1_w$c image of each subject. Afterwards the images are resampled to an isotropic resolution of 1 mm with linear interpolation. Finally all images are skull-stripped to preserve the privacy of the patients using the work of Bauer, Fejes, and Reyes (2012).

The labels of the data contain five different classes. Healthy or non-tumorous tissue is not marked; the labels are used to differentiate between subtypes of tumorous tissue. These types are 'edema', 'enhancing (solid) core', 'necrotic (or fliod-filled) core', and 'non.-enhancing core'. While these classes do meet some radiological criteria they also include some areas of similar looking tissue and

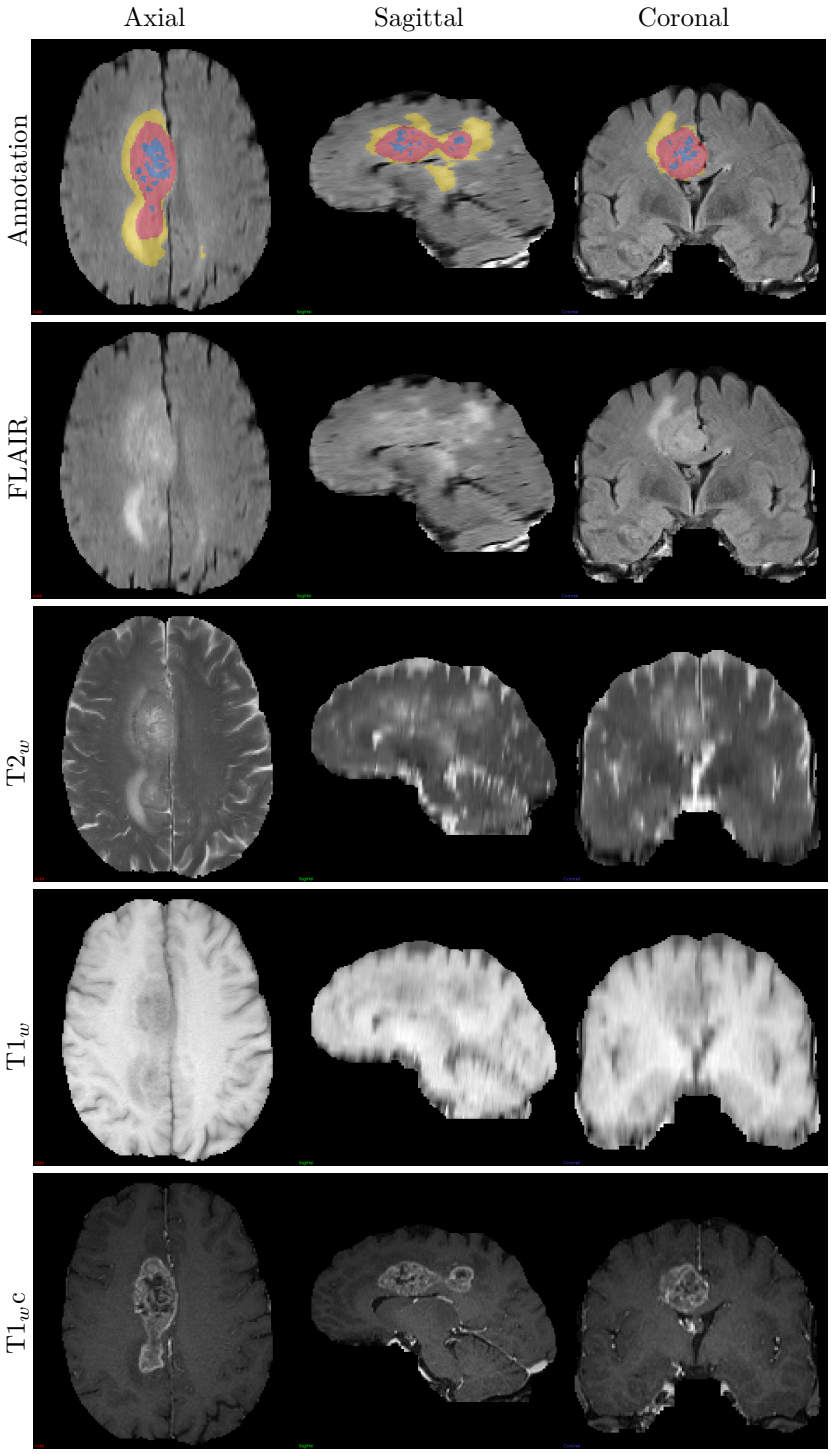


Figure 5.5: Example images of the different contrasts that are included in the dataset of the BraTS challenge. For each contrast three different views (axial, sagittal, and coronal) with the same slices are shown. The colour encoding for the annotation is: 'Yellow': Edema, 'Red': enhancing core, 'Blue': necrotic core, and 'Green': non-enhancing core (Usually very small).

are therefore not completely biological sound. For example, enhancing vessel structures may be contained in the 'non-enhancing core' if they are close to the tumour core (B. H. Menze, Jakab, et al., 2015).

These labels had to be created. For the first dataset, that was used during the 2012 and 2013 challenge, this was done manually. Four different raters created the labels according to a given labelling scheme, including the use of some semi-automatic methods to differentiate between different parts of the core. A detailed description of this scheme is given in B. H. Menze, Jakab, et al. (2015). This changed as the number of subjects was increased for the 2014 and 2015 challenge. Only one rater created the labels for the training dataset. The labels for the test dataset, which are hidden from the challenge participants, were created by fusing the results of the best algorithm from the previous years. No information is available if the labelling was done prior to the resampling of the data, but resampling artefacts in the labels of the 2012 and 2013 data suggest that this is the case.

5.1.4 Challenge data - ISLES challenge (DS-4)

The ISLES challenge (Isles, 2016) was conducted for the first time in 2015 and conducted again in 2016. It is performed in parallel to the BraTS challenge, as part of to the yearly MICCAI meeting. The objective of this challenge is the segmentation of stroke lesions in medical images and is divided into two sub-challenges: SPES which is about acute strokes and SISS which is about sub-acute ischemic strokes.

I used the dataset of the SISS sub-challenge, because the images provided are similar to those of DS-1 as this sub-challenge also used MRI for the imaging. The images of this dataset were also skull-stripped, co-registered to FLAIR contrast and resampled to an isotropic resolution of 1 mm. Overall six different contrasts were provided with the training data:

- FLAIR
- $T2_w$ TSE
- $T1_w$ TFE/TSE
- DWI

So far, no further information about the imaging – like direction of imaging (axial, saggital or coronal) or original resolution – has been released. An example of each contrast is shown in Figure 5.6.

The training dataset consisted of the images of 28 subjects. Each subject was manually annotated by an experienced medical doctor. The 36 testing patients that were used to evaluate the algorithms were annotated twice by two experts. The final ranking was then obtained by evaluating the automatically submitted segmentations against both manually created segmentation. This allowed to gain further insight into the influence of the rater variability to the scores for each training algorithm. Similar to the BraTS challenge, the training data are released including the manually created labels while the test data were released without any labelling.

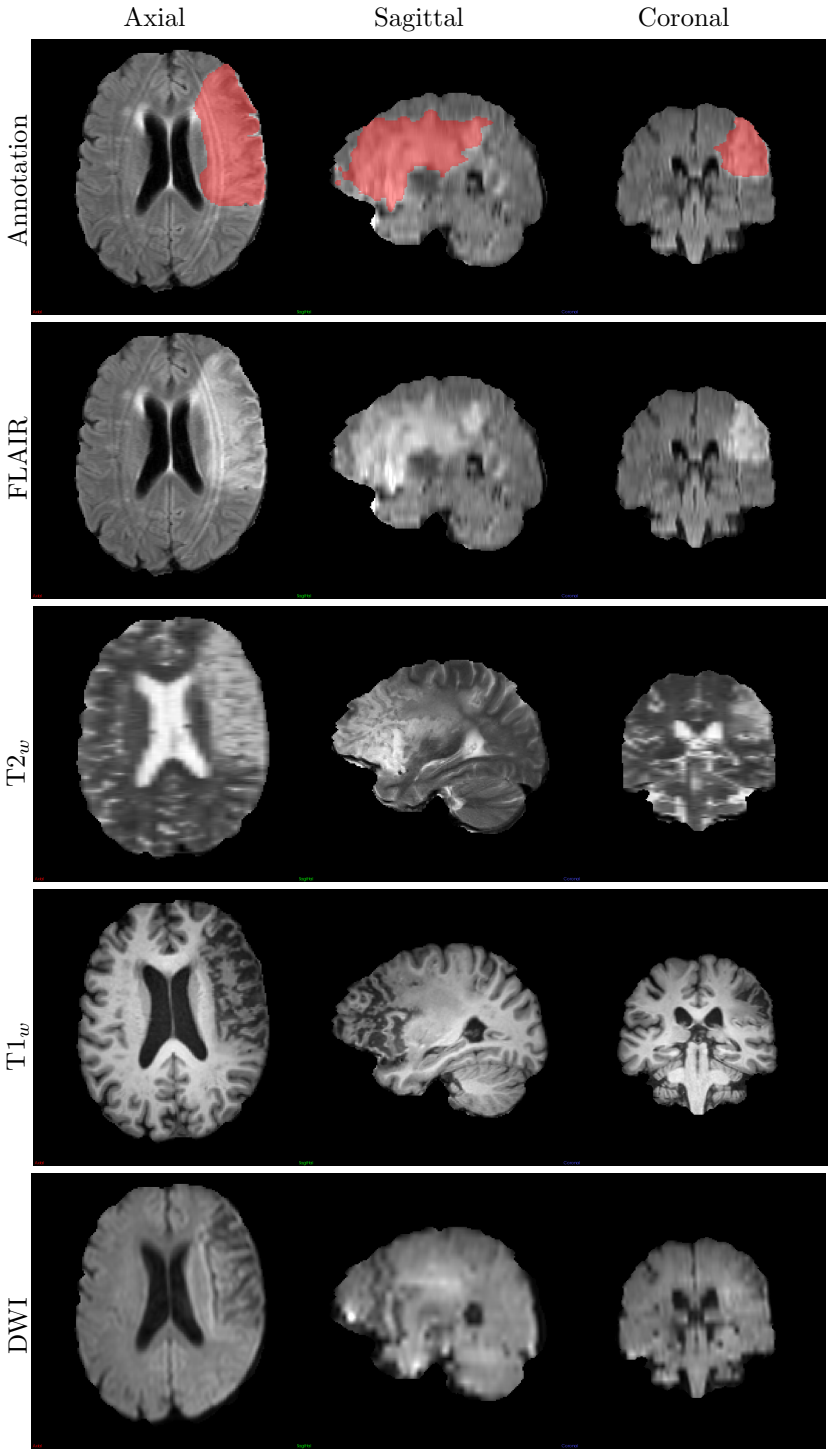


Figure 5.6: Example images of the different contrasts that are provided with the data of the isles challenge. For each contrast three different views (axial, sagittal, and coronal) with the same slices are shown. The annotation was done by a single doctor, and is shown as red overlay over the FLAIR image.

TABLE 5.3
Labelling strategies

Dataset	# observations	# features	# classes	Source
a1a	1605	119	2	LibSVM
australian	690	14	2	LibSVM
satimage	4435	36	6	UCL
vote	435	16	2	UCL
Statlog (heart)	270	13	2	UCL

Different datasets that are used for the evaluation of the LLP algorithm. Each dataset is data

5.1.5 Challenge data - Machine learning datasets (DS-5)

To my knowledge there are currently no public datasets which are only bag-wise annotated. It is therefore necessary to create these data, and therefore publicly available data were used for the creation of such datasets. The UCL Machine Learning Repository (MLR, 2016) provides different datasets from various domains. These sets are commonly used to verify the performance of new learning algorithm and is therefore well-suited for this purpose. Based on the selection of datasets in previous publications about Learning form Label Proportions (LLP) (K. Fan et al., 2014; Patrini et al., 2014) i selected some of the datasets from this repository. Table 5.3 lists the selected datasets and gives more information about them. It also list if the data are taken directly from the UCL repository[2] or from LibSVM [3] data repository that provides some of the original datasets in a cleaned version.

These datasets are then converted into bag-wise annotated datasets. Each observation is randomly assigned a bag such that each bag contains a fixed number of observations. Then the ratio for each class, except the last one, is calculated for each bag. The class ratio of the last class is then set, so that the sum of all ratios match one to ensures that rounding errors do not lead to an impossible combination of ratios.

5.2 Classification pipeline

Within this section, the influence of the MRI normalization algorithm is evaluated and two different, random forest based learning algorithms are compared.

5.2.1 Evaluation of MR Normalization

As already discussed (chapter 4.1.1), I chose to evaluate the influence of the normalization algorithm using kNN classifiers as this algorithm make direct use of the underlying features. The results are therefore more affected by errors within the classification. Due to the availability of the data, the experiments are conducted using the DS-3 (2012).

[2] https://archive.ics.uci.edu/ml/datasets.html
[3] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

As a learning-based approach is used for the evaluation of the normalization methods, the best parameters for the classifiers has to be found. For this, the training data are split into a training- and a test-group. Using these two groups a brute-force approach is used to optimize all parameters. Reasonable results are obtained if the number of nearest neighbours (k) was set to 15 and with only 40.000 randomly sampled training data points (voxels). The random sampling of the training data is necessary, as kNN uses lazy learning and the prediction time is thus depended on the number of training samples.

The kNN-algorithm has no included feature selection. It is therefore necessary to monitor the quality of the features, as too many or low-quality features can lead to a degrease in classification accuracy. Therefore it was decided to only use the the intensity values of the four available contrasts after a pre-processing step as features.

In the experiments five different MRI normalization methods were compared, namely:

None : No normalization. Image is unchanged. This is included as reference.

Statistic Linear normalization to common mean and standard deviation.

Peak Matching of fitted Gaussian curves to single point.

Percentile Matching position of percentiles.

Polynomial Match polynomial function to images.

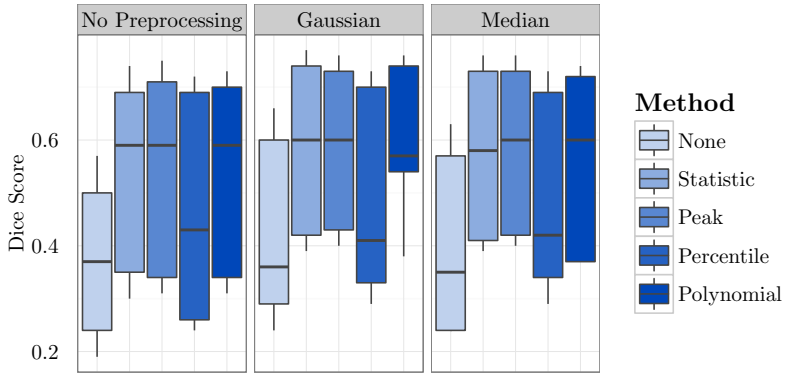
A more detailed description is given in section 4.1.1. A reference histogram is required for the last two methods. To obtain this, the mean distribution of all images after a **Statistic** normalization is used. This process is kept as simple as possible because it was found in experiments that the creation process of the reference histogram has only limited influence to the final result.

Influence of the normalization algorithm

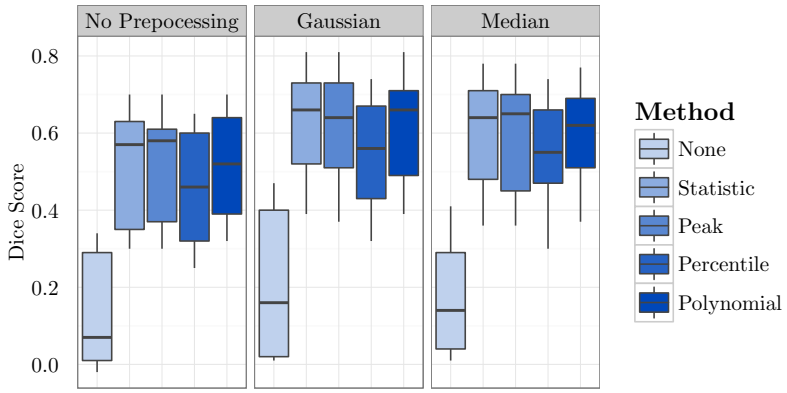
Using the previous listed normalization algorithm and parameters, multiple leave-one-patient-out experiments were ran using the available training set. To incorporate the information of the close neighbours, the images are preprocessed either by using a Gaussian, or a median filter before classification. The results of these experiments are shown in Figure 5.7, splitted according to the different pre-processing steps.

Influence of reference histogram creation

To evaluate the effect of the reference histogram that is used for the last two normalization methods two methods were evaluated. Beside the previously mentioned normalization using a **Statistical** normalization, the images are normalized using **Peak** normalization. The results of the different runs with both algorithm are shown in Figure 5.8.

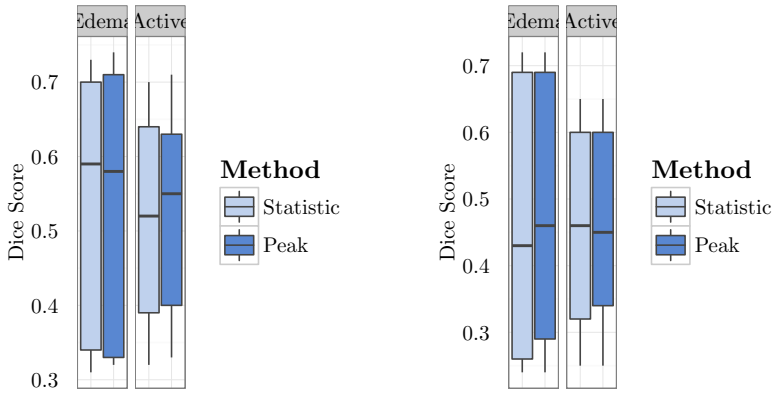


(a) Edema



(b) Active Tumour

Figure 5.7: Dice scores from leave-one-patient-out experiments with the BraTS 2012 training data (DS-3). Three different types of preprocessing are compared, either none, a Gaussian smoothing, or an median filtering. The results obtained by four different normalization methods are compared to non-normalized data. The whiskers of the boxplots in this section indicated the mean ± standard deviation.



(a) Percentile (b) Polynomial

Figure 5.8: Evaluation of the different methods to create a reference histogram. No additional preprocessing was applied beside the MRI normalization.

Influence of bias field correction

The influence of field bias correction is evaluated by running the pipeline once with and once without bias field correction using a leave one out scheme on the BraTS 2012 training data. Figure 5.9 shows the absolute results obtained with the different methods if the images are smoothed using a Gaussian before the classification process.

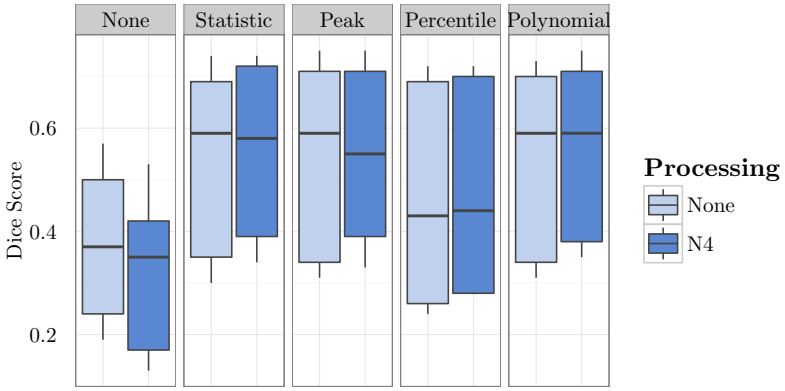
Evaluation using BraTS test set

To validate the main findings, e.g. the performance of the different normalization method by contributing to the challenge. For this, the test dataset of the BraTS challenge was used. The best configuration for each normalization was chosen based on the previous findings. The tests were limited in order to reduce the number of tests on this dataset. The obtained scores for the four different normalization methods are given in Figure 5.10, a bias field correction and Gaussian filter were chosen as preprocessing based on the previous results. The result without normalization were not tested, as previous experiments validated the expectation that a normalization is necessary.

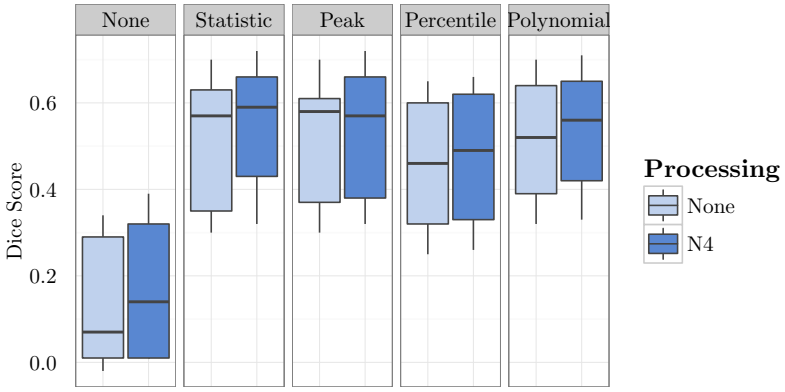
5.2.2 Evaluation of Classification Algorithms

Description of the experiments

The power, i.e. the classification strength, of the traditional RDF and the ExtraTrees are evaluated using dataset DS-3 – the BraTS challenge from the year 2013. A classifier is trained on the 20 training datasets using all available modalities, namely T2 Flair, T1, T1 with contrast agent and T2. With the so-trained classifier the 10 high-grade glioma evaluation datasets are labelled



(a) Edema



(b) Active Tumour

Figure 5.9: Influence of the N4 bias field correction to the final segmentation result. Each normalization method is tested once with and once without a previous bias field correction step. **(a)** shows the dice scores for the 'edema' class and **(b)** for 'active tumour'.

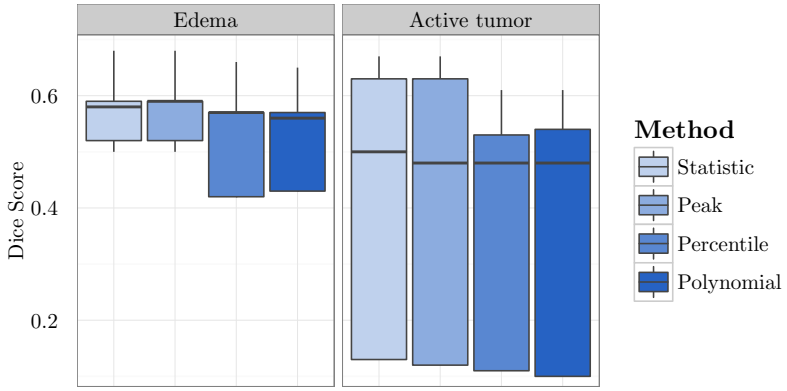


Figure 5.10: Comparison of different normalization methods using the 2012 BraTS test set (DS-3) to validate the previous findings. Therefore only the various approaches are evaluated. Gaussian smoothing and N4 bias correction is applied as additional preprocessing.

and the results are evaluated by the provided online tool.

For the evaluation the overlap with 3 labels is measured using the Dice score. The first label '*complete tumour*' includes necrosis, edema and both enhancing and non-enhancing tumour. The second label, *tumour core*, is the same as the complete tumour but without edema. Finally, the label *enhancing tumour* is evaluated.

Results of the experiments

Figure 5.11 provides the Dice scores for both tested classifiers. The performance of the ExtraTrees is usually higher than the performance of the canonical classifier, although the difference is rather small. Using ExtraTrees, the Dice score of five patients improved by more than 1 % while it dropped only for one patient by more than 1 % with respect to a canonical RDF.

Figure 5.12 shows example segmentations for both classifiers.It can be seen that the results obtained with ExtraTrees are more accurate. They contain usually less false positives. It is also visible that the the labelling seems to be more accurate than those obtained with canonical forests.

5.3 Pre- and post-training data selection

This section describes the experiments and obtained results for the evaluation of the two proposed algorithm for selective learning algorithms.

5.3.1 Input Data Adaptive Learning (IDAL)

Dataset DS-4 (ISLES challenge dataset) is selected as data set for the evaluation of the proposed Input Data Adaptive Learning (IDAL) approach. Beside the

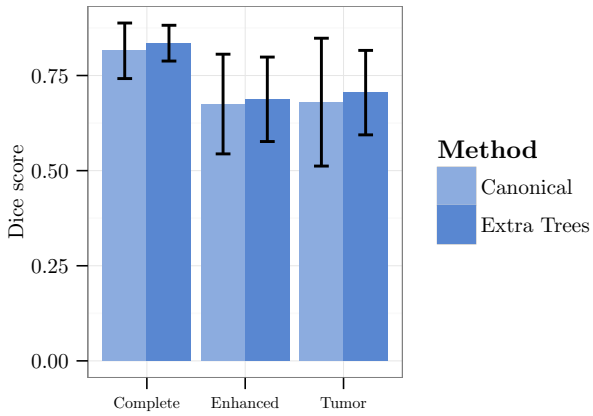


Figure 5.11: Dice scores for both tested classifier. The overall performance of the ExtraTrees is better than the performance of the canonical RDF algorithm.

fact that ischemic strokes are highly variable in appearance, the second reason to chose this dataset is the high labelling quality compared to other public datasets.

If not mentioned otherwise, all experiments in this section are conducted using the previously mentioned (section 4.2.1) pipeline. This includes the described preprocessing and features.

Evaluation of training data similarity

As proposed the similarity between two different images is measured by the ability of a classifier trained on the first image to segment the second image. In a first step the similarity between all images is measured by this method. The necessary parameter tuning is done individually for each image. Figure 5.13 gives a correlation matrix between all images. The so-defined similarity is not symmetric, i.e. the similarity from image A to B might be different from the similarity of image B to A.

Based on the correlation matrix, Figure 5.14 shows the distribution of similarity values as boxplot, both for the overall data set and for all images individually. The results indicate that there are large difference in the similarity and that some images might be more important than others, supporting the proposed idea.

Evaluation of similarity classifier

The previous calculated similarity matrix is used for the evaluation of the proposed similarity classifier. Using a leave-one-out scheme, a similarity classifier is trained for each subject in the dataset and the similarity ranking is calculated.

Figure 5.15 shows the pre-calculated similarity matrix with the three most similar subjects marked for every individual patient. Corresponding, Figure 5.16 shows the similarity ranking obtained for each patient. A more quantitative evaluation is shown in Figure 5.17, that indicates the average similarity rank

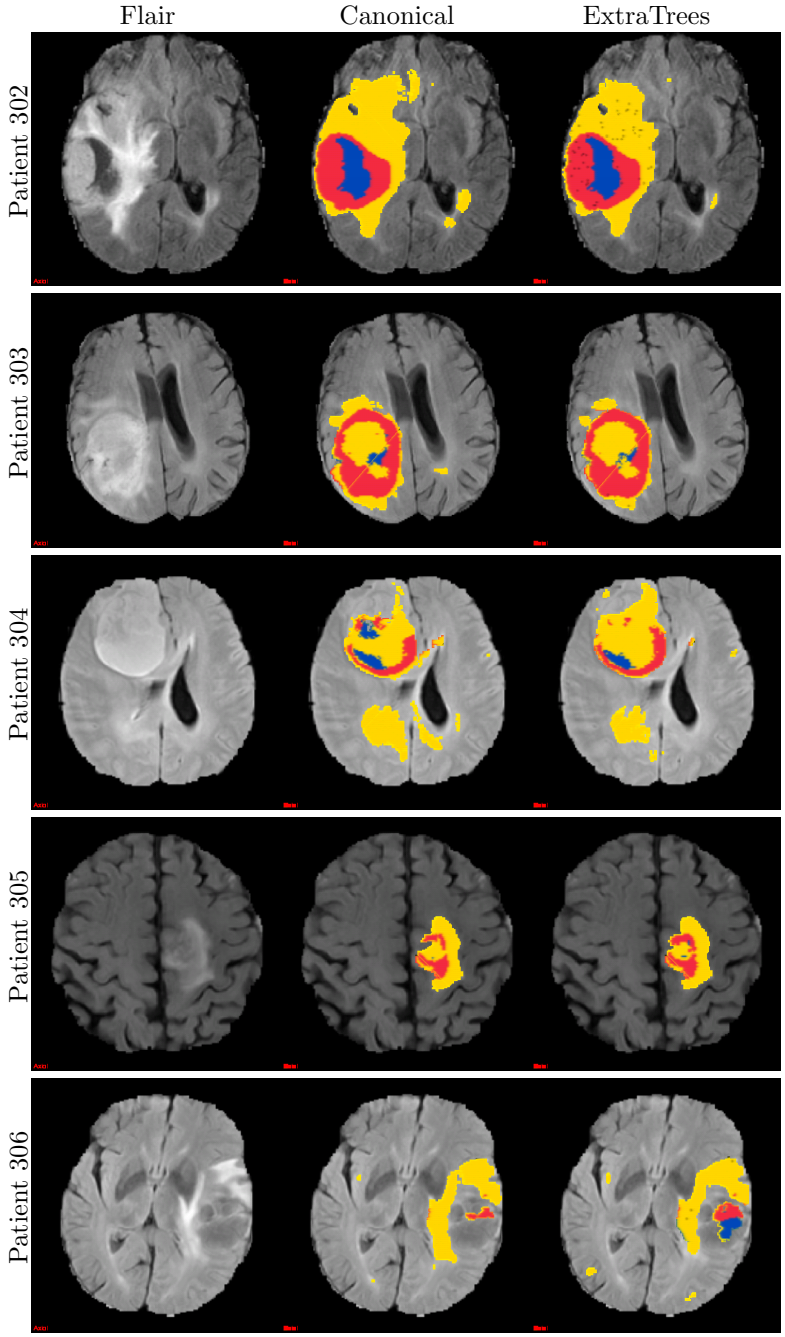


Figure 5.12: Example slices from different patients showing the results obtained from canonical forests and ExtraTrees. The colour encoding for the annotation is: 'Yellow': Edema, 'Red': enhancing core, 'Blue': necrotic core, and 'Green': non-enhancing core. The ground truth is not revealed and therefore not shown.

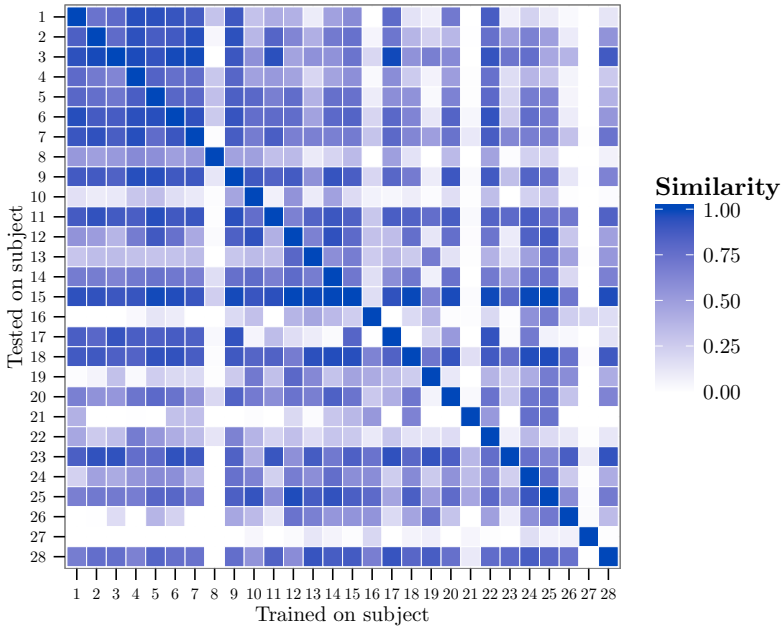


Figure 5.13: Similarity matrix between the patients within the training data of the ISLES challenge dataset. A higher similarity indicates that a classifier trained on the corresponding subject is better suited to predict the labels for the corresponding test subject.

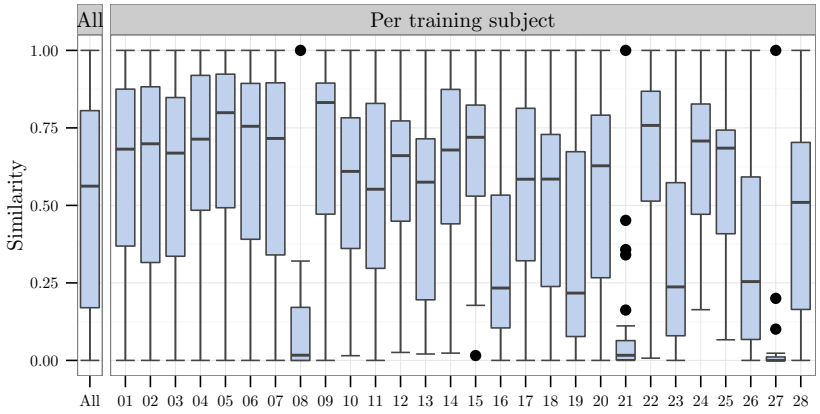


Figure 5.14: Distribution of the similarities obtained per training image. A higher variability within a distribution indicates a higher variability in the suitability for training of the corresponding image.

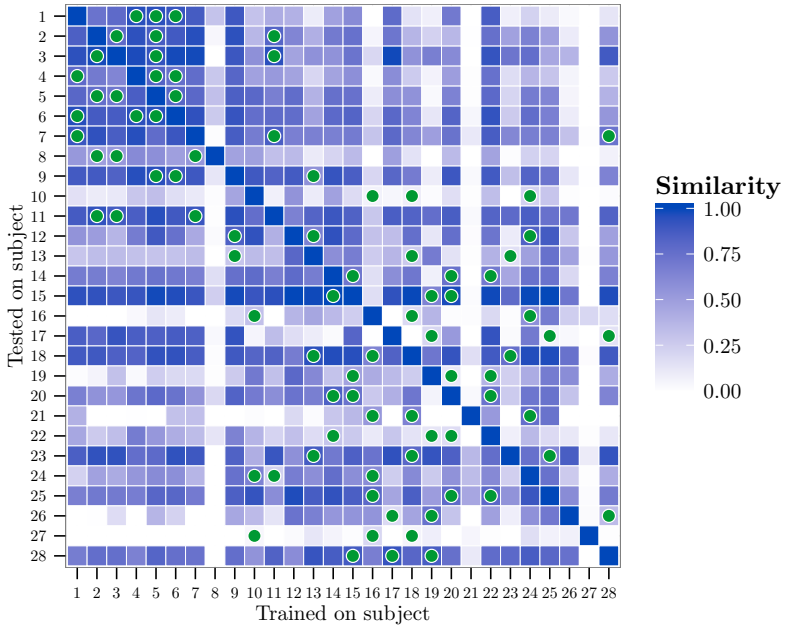


Figure 5.15: Similarity matrix between the patients within the training data of the ISLES challenge dataset. The three highest ranked training subjects per test subject are marked by green points.

for different number of included images. Note that the ranking is absolute and does not reflect the differences. For example, if the similarity of A to B and C is 1 and 0.5, respectively, the ranking of B and C would be 1 and 2, respectively. The same would hold true for similarities of 1 and 0.99. But a wrong estimation of the similarity rank would mean a more significant error in the first case. Due to this, Figure 5.18 gives also a comparison of the possible best average Dice score and the obtained Dice scores.

Sparse annotations are used for these experiments. Due to that, the complete annotation is theoretically not available for the estimation of the complete training process – the similarity matrix for the training data is therefore also only calculated using sparse annotations. A comparison of the similarity matrices obtained from sparse annotations and full annotations is given in the appendix (Appendix B).

Evaluation of complete IDAL pipeline

The complete algorithm is evaluated by estimating the most similar patients and then using the most similar subjects for a trained classifier. Figure 5.19 shows the results of the final runs, both with the estimated best images and the true most similar results.

There are also other options to determine which patients to use instead of using only a fixed number of training patient. Therefore I ran two additional experiments to evaluate two other methods. In the first experiment, each patient

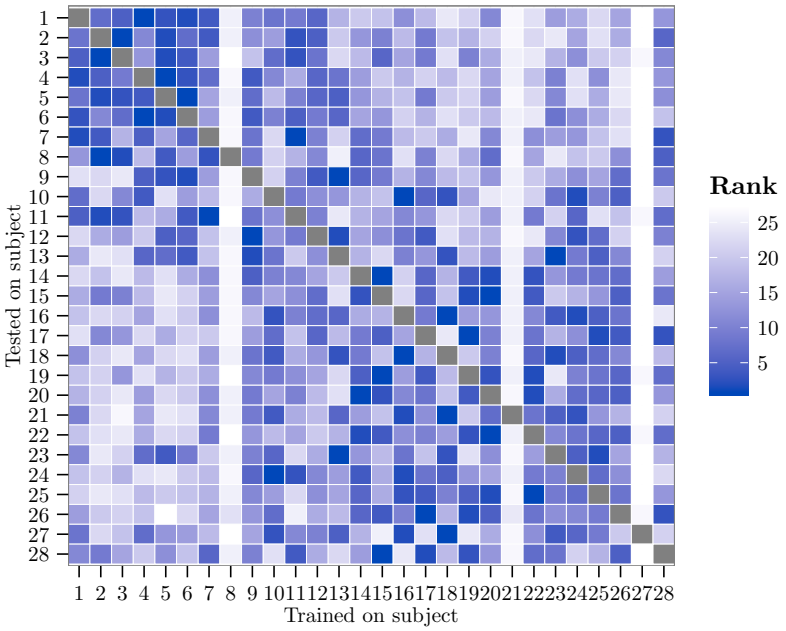


Figure 5.16: Ranking of each training subject for each test subject. The ranking goes from 1 to 27 for each test subject, the distance between two ranks does not reflect the difference between the similarity of these two images.

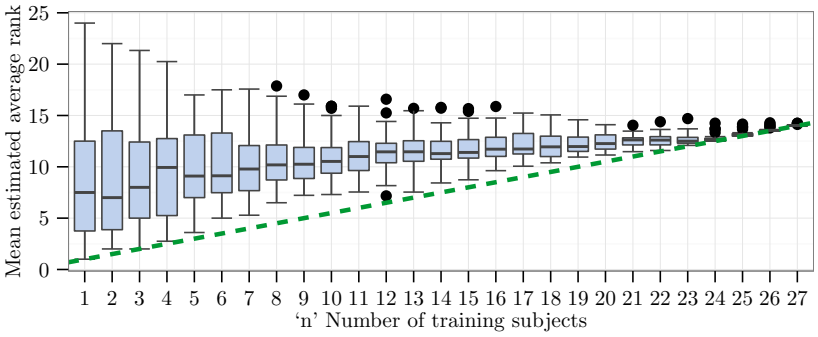


Figure 5.17: Average mean rank obtained for each test subject if the 'n' highest ranked training subjects are taken. The dashed green line indicates the mean rank if the true 'n' best training subjects are selected.

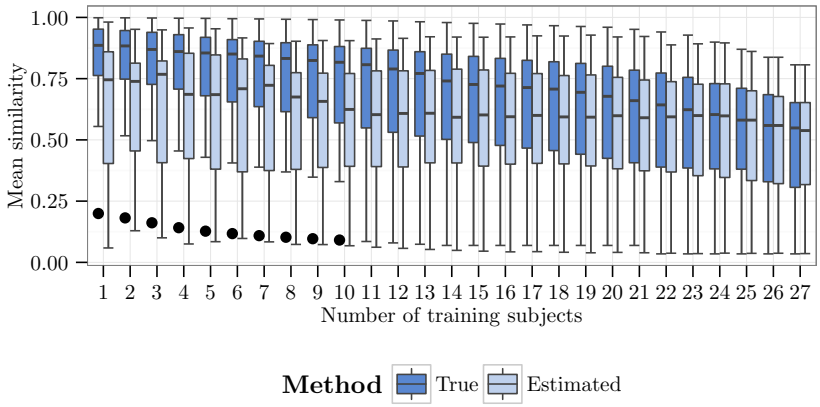


Figure 5.18: Average similarity score for the estimated 'n' best training subjects compared to the average similarity score that is obtained if the true best training subjects are selected.

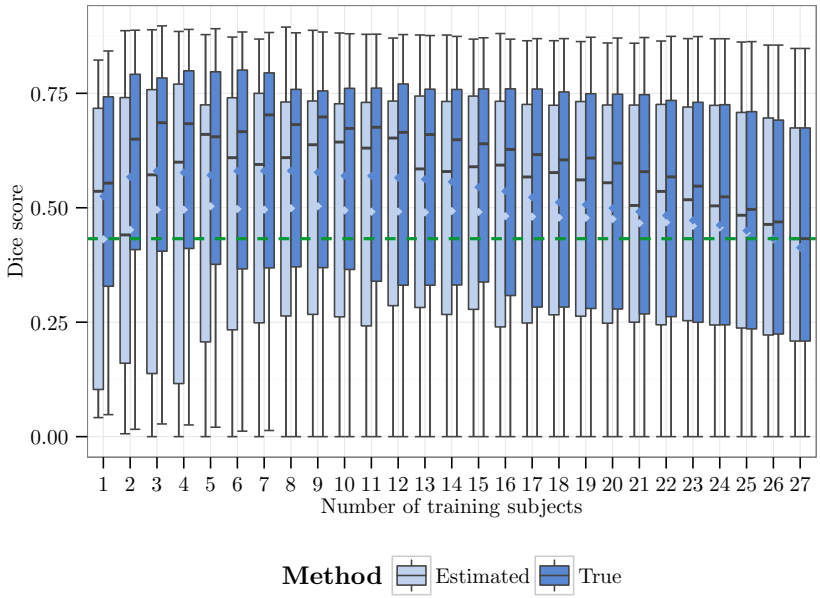


Figure 5.19: Boxplot of the obtained Dice score if the given number of best training samples are given. The results are given if the estimated, or the true, most similar training images are used. The green dashed line indicates the median result if all images are used. The notches indicate the mean value of the corresponding colour.

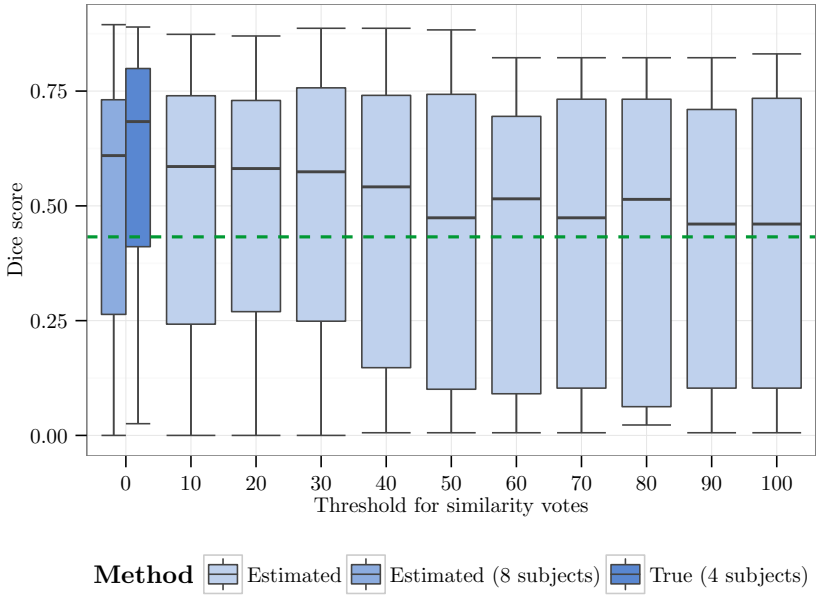


Figure 5.20: Boxplot of the obtained Dice score if only training subjects are used that obtained a similarity score over the given threshold. The green dashed line indicates the median result if all images are used. The two most-left results are the best results obtained from a fixed number of training patients.

with a similarity vote[4] over a certain threshold is used as training subject (Figure 5.20). The best results obtained when limiting the number of training subjects are included – once using the true and once using the estimated similarity. For the run limiting at a threshold of 10 the mean and median number of used training subjects are 15.35 and 16 respectively. For a threshold of 20 these are 7.96 and 8, respectively.

The second method is limiting the number of training subjects by thresholding the sum of their votes, i.e. including as much training subjects as necessary to reach a given number of overall votes (Figure 5.21). There are no clear indication which threshold will lead to the best results, the results are unstable. The mean and median number of used training subjects for some of the best results (number of votes: 100,240,320) are 2.61, 7.43, 12.14 and 3, 7, 12, respectively.

Based on these experiments the optimal number of neighbours was set to seven. With this I contributed to the ISLE challenge using the test set once with a standard approach (i.e. using all subject to train a single classifier) and the IDAL approach. Figure 5.22 shows the obtained final scores of these two runs.

For more qualitative examples, Figure 5.23 and 5.24 gives some example

[4]As mentioned earlier: The similarity estimation is done as a voting process, i.e. several classifier vote for each patient instead of estimating the true similarity. The votes are then accumulated, a higher number indicates that the corresponding subject is estimated as most similar by more similarity classifier.

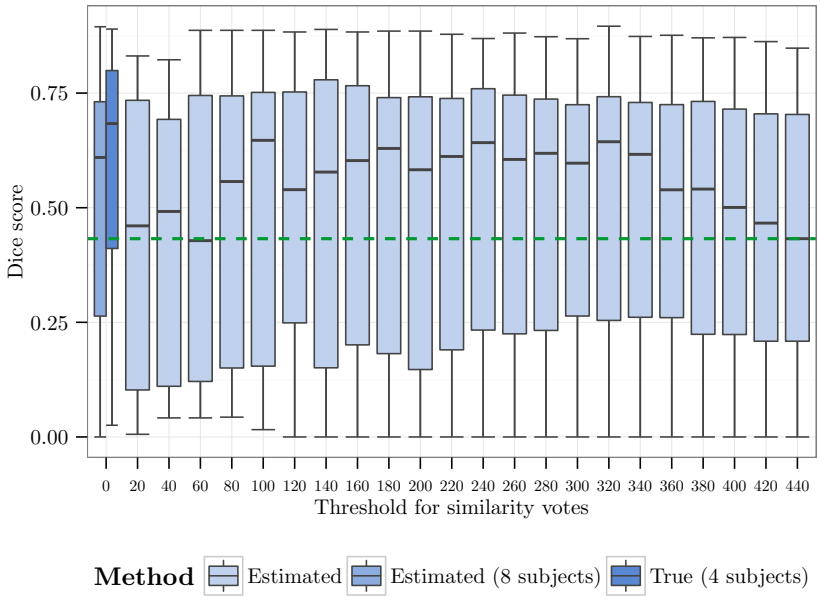


Figure 5.21: Boxplot of the obtained Dice score if the best training subjects are used for which the sum of similarity votes is below the given threshold. The green dashed line indicates the median result if all images are used. The two most-left results are the best results obtained from a fixed number of training patients.

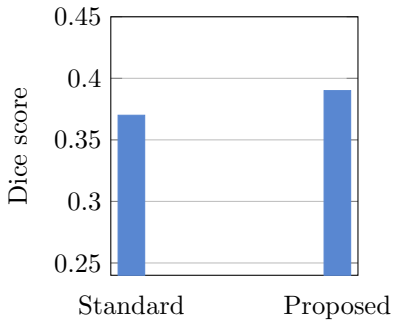


Figure 5.22: Results obtained on the test set of DS-4.

TABLE 5.4
Random Decision Forest Parameter

Parameter	Value
Tree count	100
Maximum allowed tree depth	5
Randomly drawn trainings points per subject	300
Sigma for post-classification result smoothing	0.75 mm

Parameter set for the evaluation post-classification data selection.

results of segmentations obtained both by using standard approach and using IDAL based segmentation.

5.3.2 Pre-trained semi-automatic tissue characterization

The interactive scenario is tested using the in-house CT dataset containing images of Intracerebral Haemorrhages (ICH). A random forest is trained using these data and the best parameter set is estimated using 5-fold cross-validation. The obtained parameters are given in table 5.4. A more detailed description of the used features, and sampling strategy is given in the corresponding method section.

Using the results from the 5-fold experiments the ten subjects with the worst resulting Dice score are selected. By this, the data used in this experiment contains the elements that failed during the automatic segmentation. For each of the ten selected subjects a semi-automatic segmentation is created by manually adding new training points. These training points are then saved for further experiments.

Two different approaches for the creation of a semi-automatic segmentation are evaluated, as described in the corresponding section 4.2.2, namely: .

Weighted approach: A RDF is trained using 24 randomly selected subject, excluding the current one. For each new, manually added point each tree in the existing RDF is weighted.

New learning approach: During each iteration, a new RDF classifier is trained using only the manually added points.

Note, that the manually placed annotation points allow to identify areas that belong to the lesions, as these areas are usually connected. Non-connected lesion areas are removed from the automatic segmentation.

Using the pre-selected ten patients with additional manually annotations points, the experiments are carried out using leave-one-patient-out scheme. For the evaluation of each patient, a single manually placed annotation point is selected and added to the training collection of this patient. After this, the new created segmentation is evaluated. This is repeated for all 40 training points of each subject, and each subject is evaluated ten times to account for the random selection of manual annotation points.

Results

Figure 5.25 shows the obtained Dice scores depending on the number of the used training points. The coloured area indicates the standard deviation while the lines shows the corresponding mean value. The Dice scores for the three different tissue classes are shown using colour coding.

The final results are shown in Figure 5.26 giving the best obtained Dice scores for the brain, lesion and CSF segmentation. Additional the sensitivity and specificity are also given. The results are obtained using the lowest number of samples that lead to stable results.

5.4 Methods for Reduced Annotation Effort

Within in this section, the experiments that were carried out to evaluate the learning from weak are described. The ordering is according to the amount of annotations that are necessary for a successful training.

5.4.1 Learning from Sparse Annotations

Classifiers were trained in leave-one-patient-out experiments. The quality of the obtained segmentation on the left-out patient is evaluated on the basis of the manually annotated ground truth using the well-known Dice score (Dice, 1945) as well as the sensitivity (true positive rate) and specificity (true negative rate). The ground truth contains only the labels 'tumorous' and 'healthy'. Thus, in the five-class automatic segmentation, the labels 'healthy' and 'fluid' are relabelled 'healthy', while the labels 'edema', 'active' or 'necrosis' are relabelled 'tumorous' prior to evaluation. In all experiments – except the generation of Figure 5.32 – the decision threshold of the classifiers is left at 50% for the two class problem. The decision threshold is not affected by adding the class weights used bythe proposed method. Three different setups are used in the experiments. First, each setup and the corresponding experiments are described and following the results are given.

Primary Experiments: DS-1 with RDF

Setup I consists of dataset DS-1 in conjunction with weighted random forests. The feature vector x_i for a voxel v_i of this setup consists of the gray values of all available images, i.e. FLAIR, T1C, B0, AD, CA, FA, MD, RA, RD, AD-FWE, CA-FWE, FA-FWE, MD-FWE, RA-FWE, RD-FWE, and FW.

On the basis of setup I, seven different methods for annotating, sampling, and using training data (cf. Figure 4.7) were assesed. As a reference, three of those methods are trained using the ground truth GTV segmentations as training labels. Two different sampling strategies are applied: sampling of all labelled voxels (Learning from Complete Annotations, LCA) and random sampling of the labelled voxels at 0.5 % ratio (similar to SUR coverage ratio of the expert rater). The randomly sampled training data are used with ($\text{DALCA}_{\%}$) and without ($\text{LCA}_{\%}$) domain adaptation. The reference methods are compared to classifiers trained on SURs created by the expert radiologist, either using (DALSA) or not using (LSA) domain adaptation. The SURs differentiate five different tissue classes while the ground truth segmentations only differentiate tumorous from

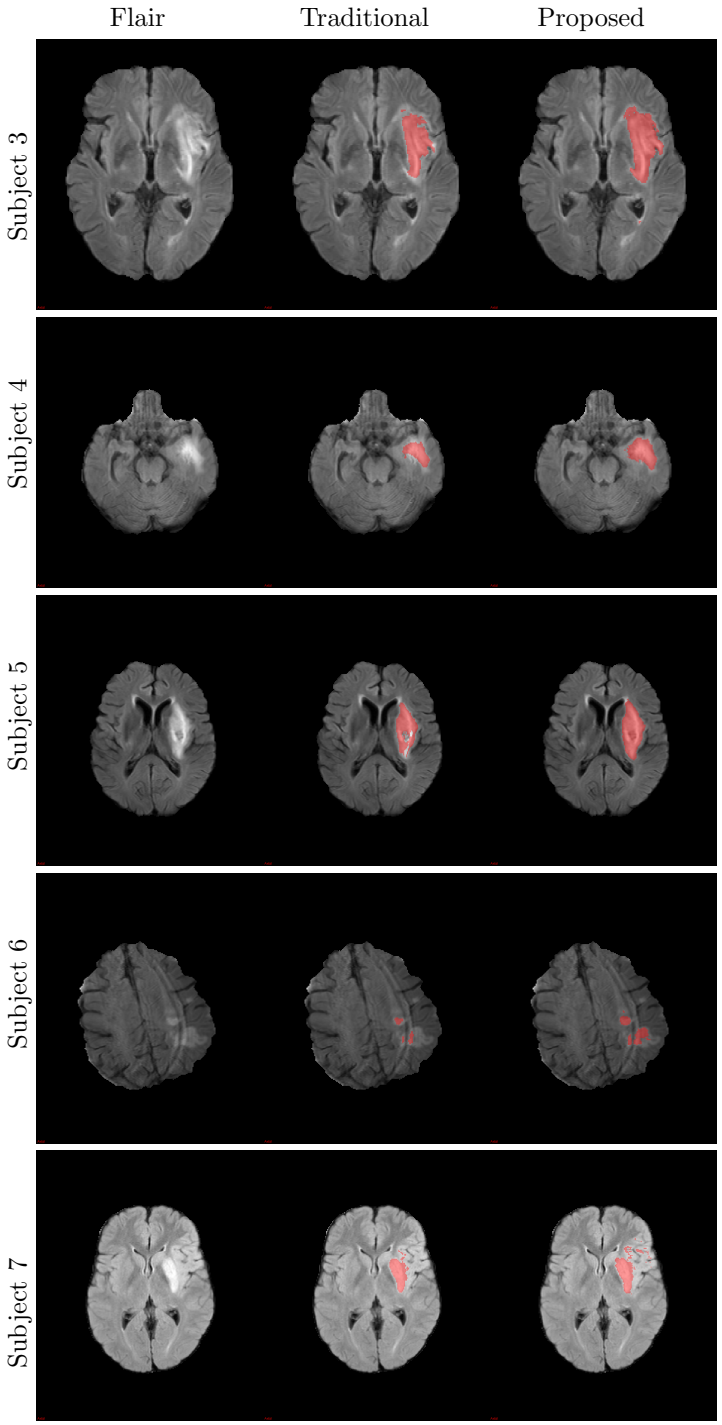


Figure 5.23: Example axial slices from 12 subjects of the ISLES 2015 test-set without provided ground-truth. The red area indicates the automatic segmentation of the lesion. The results are obtained using leave-one-patient-out scheme. For the traditional approach, the used classifier is trained on the remaining 27 training subjects.

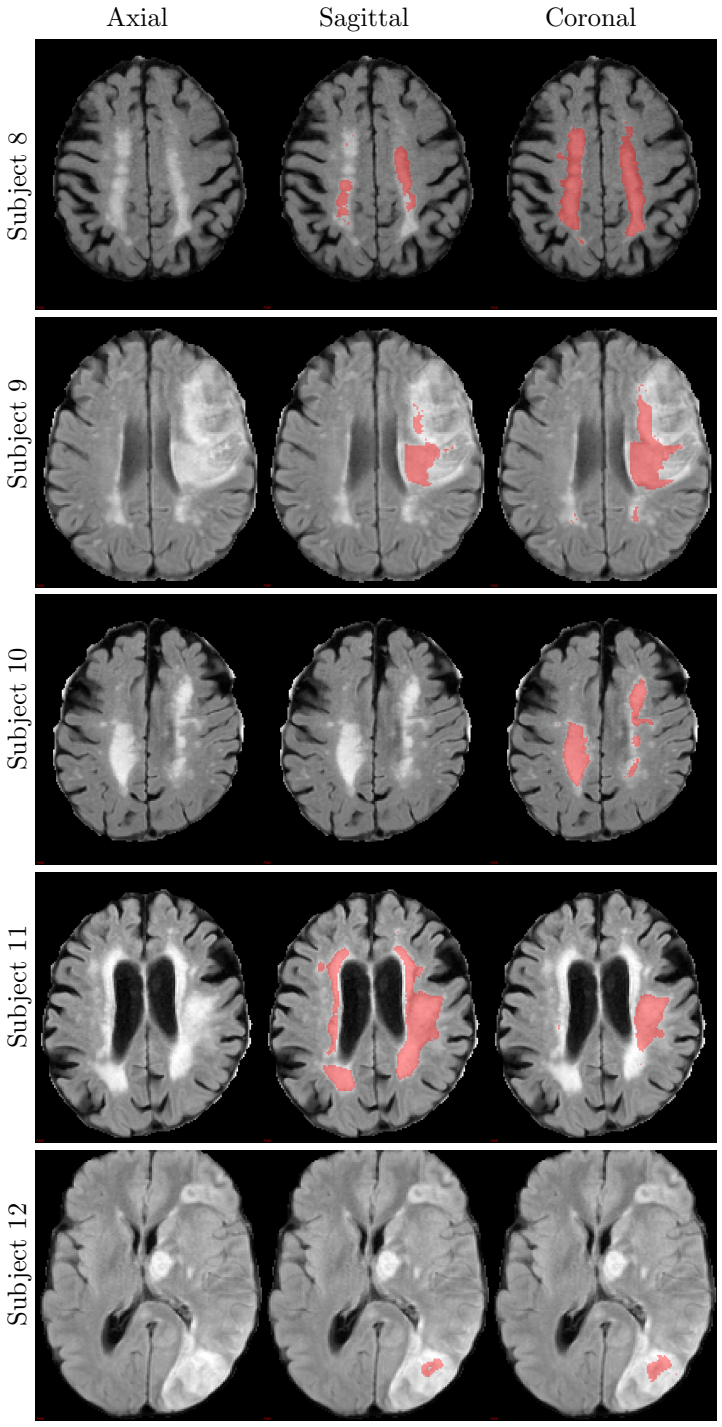
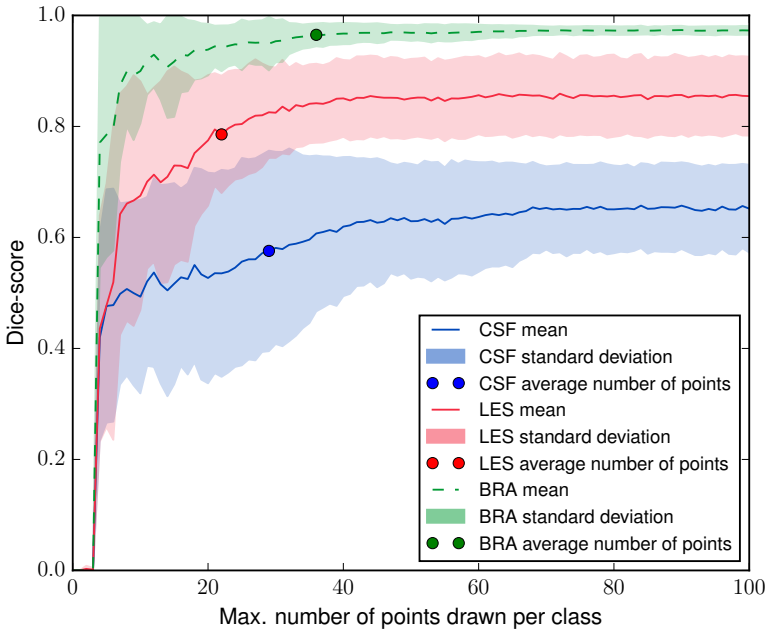
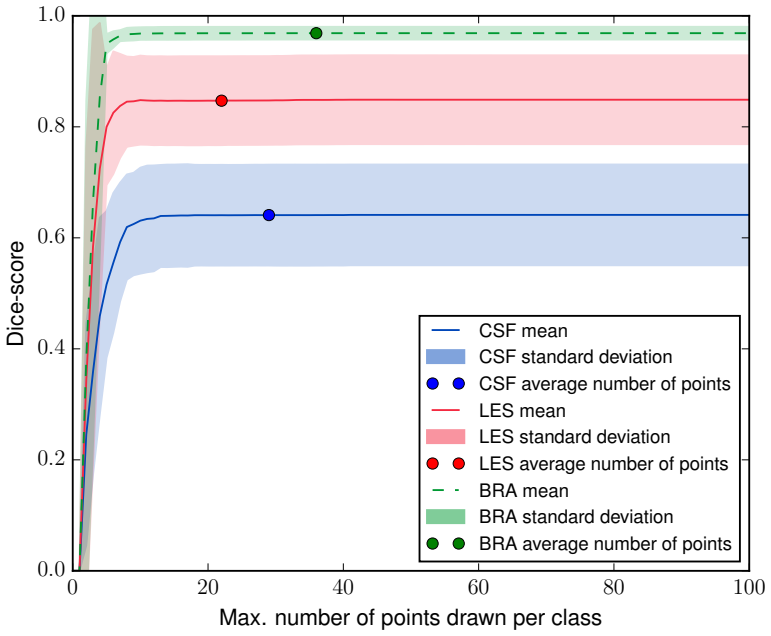


Figure 5.24: Example axial slices from 12 subjects of the ISLES 2015 test-set without provided ground-truth. The red area indicates the automatic segmentation of the lesion. The results are obtained using leave-one-patient-out scheme. For the traditional approach, the used classifier is trained on the remaining 27 training subjects.



(a) New learning (standard) approach



(b) Weighted (proposed) approach

Figure 5.25: Dice scores obtained by two different interactive correction methods. The Figure shows the result of multiple runs with randomly selected annotation points.

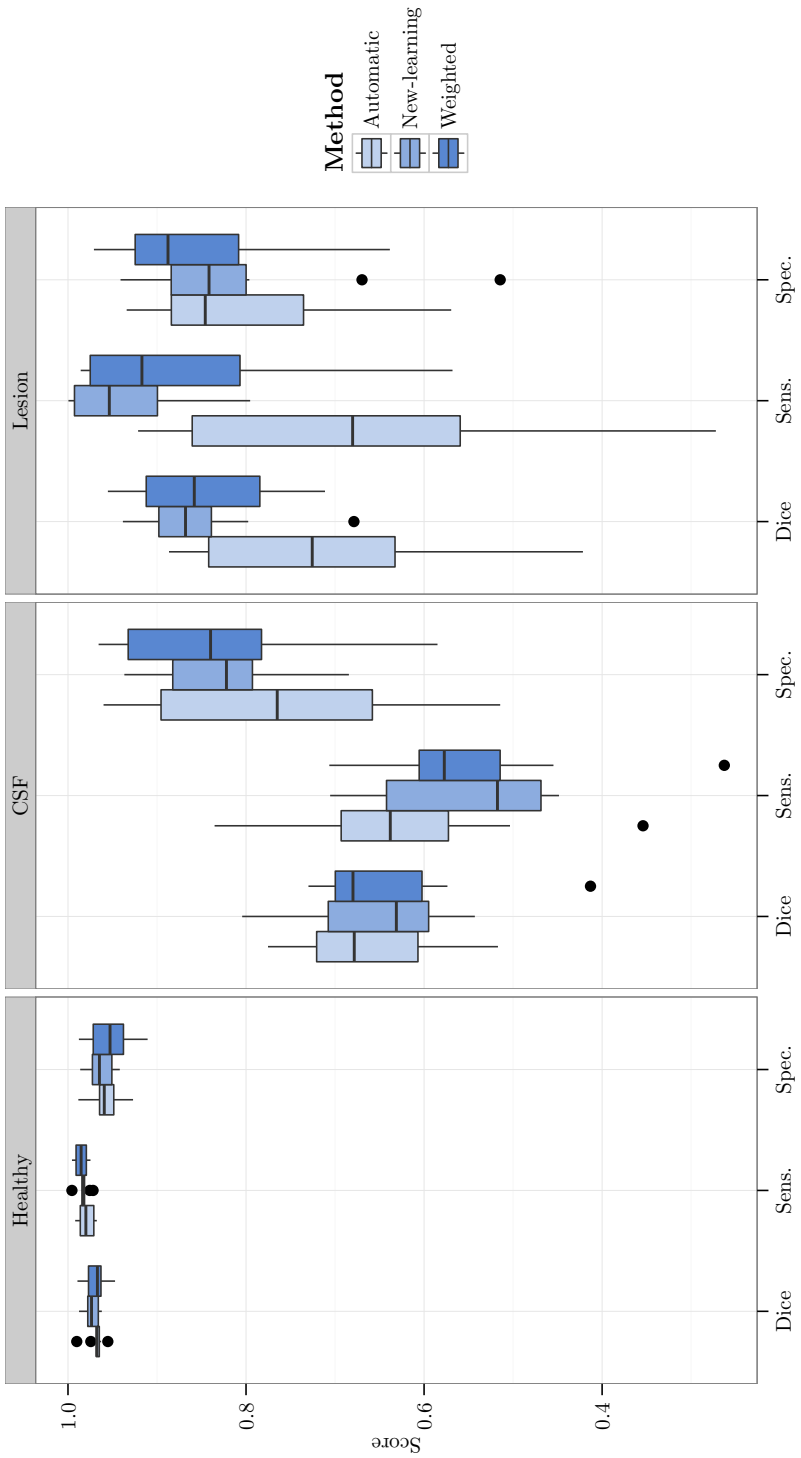


Figure 5.26: Scores obtained by the different methods. The fully automatic reference method is obtained without any manual correction by training a classifier from previous labelled training data. For the two other approaches, the best results of several semiautomatic runs is used.

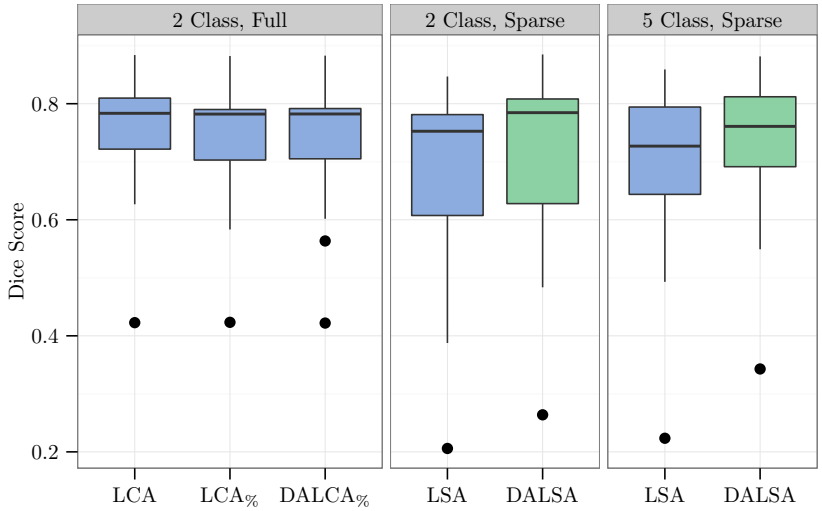


Figure 5.27: The results of the leave-one-patient-out experiments using setup I. Boxplots showing the results grouped by classifier scheme. The left section shows the results for classifiers that are trained on complete segmentations. The middle and right sections show results for classifier trained on SURs using two and five different tissue classes, respectively.

healthy tissue. It is therefore necessary to fuse tissue classes to allow direct comparison of the different methods. Indices indicate the number of classes that are used during training (e.g. LSA_2: fusion of labels before training; LSA_5 fusion of labels after training in predicted images).

The influence of λ is evaluated by conducting leave-one-patient-out experiments for $DALSA_2$. The maximum tree depth was set to 4, the minimum sample size at each leaf node to 1, and the maximum number of evaluated features at each node to 4. Then λ was varied between 0.0 and 1.0. The influence of altering SUR annotations is evaluated in two experiments: First, LSA_2 and $DALSA_2$ classifiers are trained on SURs with varying annotation strategies (cf. Table 5.2, 'Type 1' - 'Type 4'). Second, the expert's influence on the resulting segmentation quality is compared by training LSA_2 and $DALSA_2$ classifiers on SUR sets created by the expert rater and compared these with sets created by two student raters. Also majority voting was applied to compute combined results of all raters.

Primary results: DS 1 with RDF

Figure 5.27 and 5.28 show the obtained Dice scores and ROC analysis for the different methods assessed. Table 5.5 lists the corresponding uncorrected statistical significance values on the basis of the Wilcoxon signed rank test.

Figure 5.29 and 5.30 provides some exemplary qualitative results. The proposed domain adaptation could effectively reduce the drop in segmentation quality caused by learning from sparsely annotated data. DALSA results do

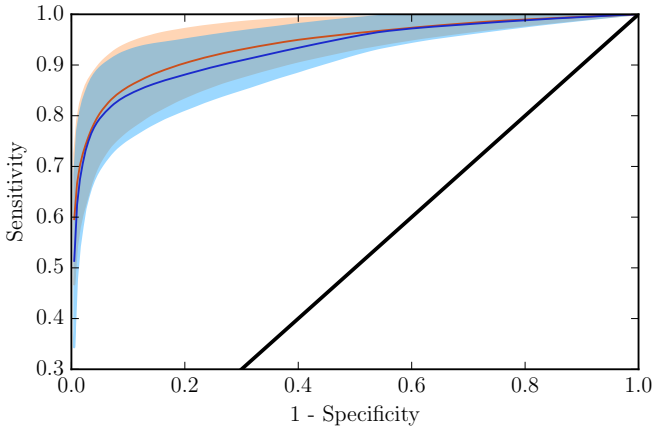


Figure 5.28: The results of the leave-one-patient-out experiments using setup I. ROC-curves for LSA_2 (blue) and $DALSA_2$ (red). The curves are obtained by varying the decision threshold of the classifier and then calculating the mean (solid) and standard deviation (coloured area).

not significantly differ from the results obtained by $LCA_{\%}$, which is a commonly applied sampling strategy in other studies but requires complete annotations. It was possible to further increase the quality of the final segmentation by merging the segmentations obtained from different experts. The merged DALSA results do not significantly differ from LCA ($p = 0.084$). There is also a significant increase in segmentation quality when applying domain adaptation to $LCA_{\%}$. The extent of the effect, however, was very small (median Dice difference 0.0008, mean 6.23×10^{-5}).

Figure 5.31 demonstrates the effect of domain adaptation on the classification results. The LSA_2 Dice scores are plotted over a moving decision threshold (blue curve) and should optimally exhibit a bell-shaped curve with its maximum at 50 %. The SUR-based sampling bias, however, lead to a skewed curve with suboptimal classification results. DALSA corrects for this effect and yields a more bell-shaped curve (Figure fig:dalsa-threshold, red curve). Figure 5.33a shows the performance of DALSA under different SUR labelling strategies. DALSA significantly outperforms LSA ($p \leq 0.001$) in all cases. Similarly, DALSA outperforms LSA regardless of which expert labelled the data (Expert 1: $p = 0.015$, student 1: $p = 0.007$, student 2: $p = 0.001$, c.f. Figure 5.33b). DALSA performance is always comparable to $LCA_{\%}$. Figure 5.32 shows the influence of the relaxation coefficient λ.

Classifier experiments: DS 1 and SVM

Setup II is similar to setup I, but with weighted SVM instead of weighted random forests. On the basis of setup II, is assed whether SVM-based classification can also profit from DALSA. Leave-one-patient-out runs were conducted at varying cost settings between 0.01 and 0.08 and compared the results obtained by LSA_2 and $DALSA_2$.

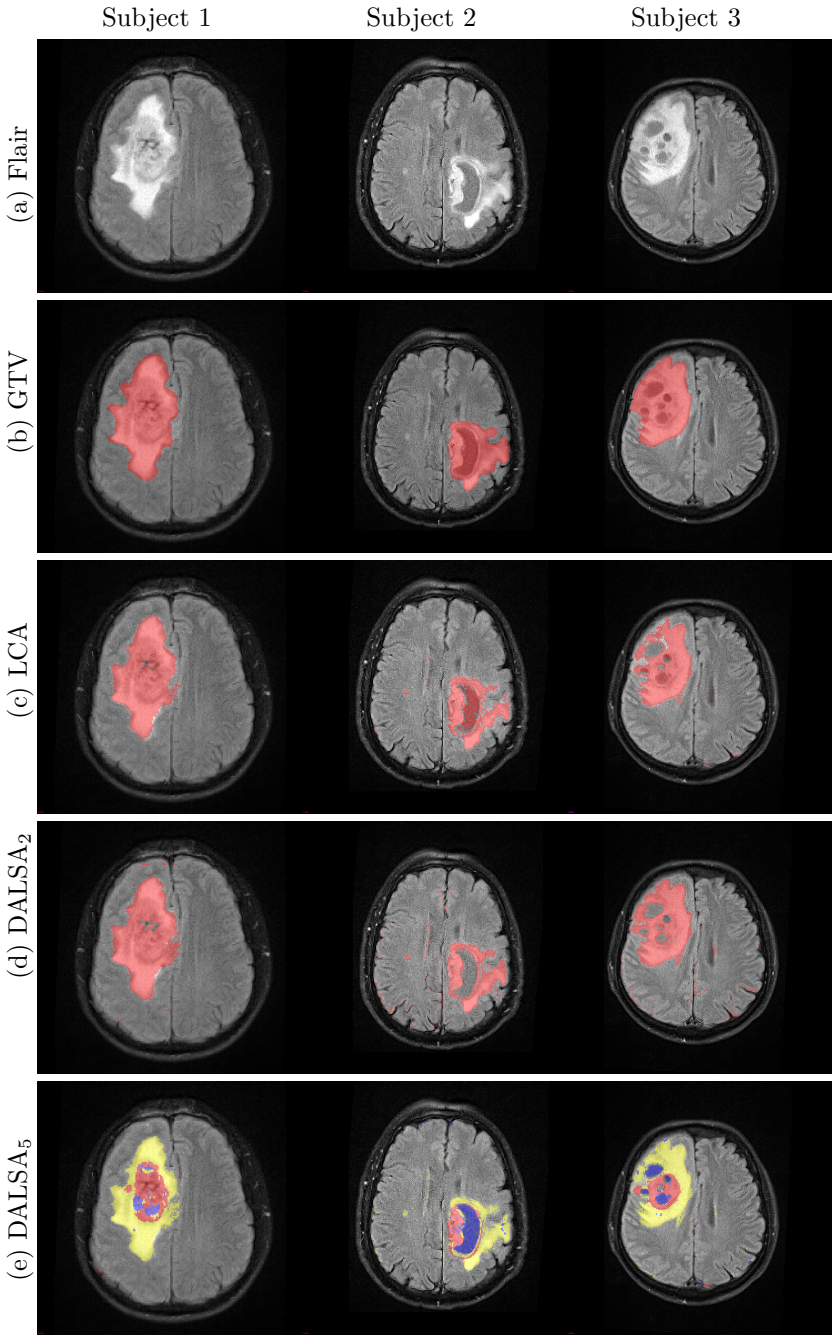


Figure 5.29: Example axial slices from setup I. (a) FLAIR image, (b) gold-standard segmentation (c) result of classifier trained on complete segmentations, (d) DALSA with 2 classes, and (e) DALSA with 5 classes. In (b-d), the red colour indicates 'gross tumour volume'. The colour coding in (e) is: yellow: 'edema', red: 'active tumour', blue: 'necrosis'.

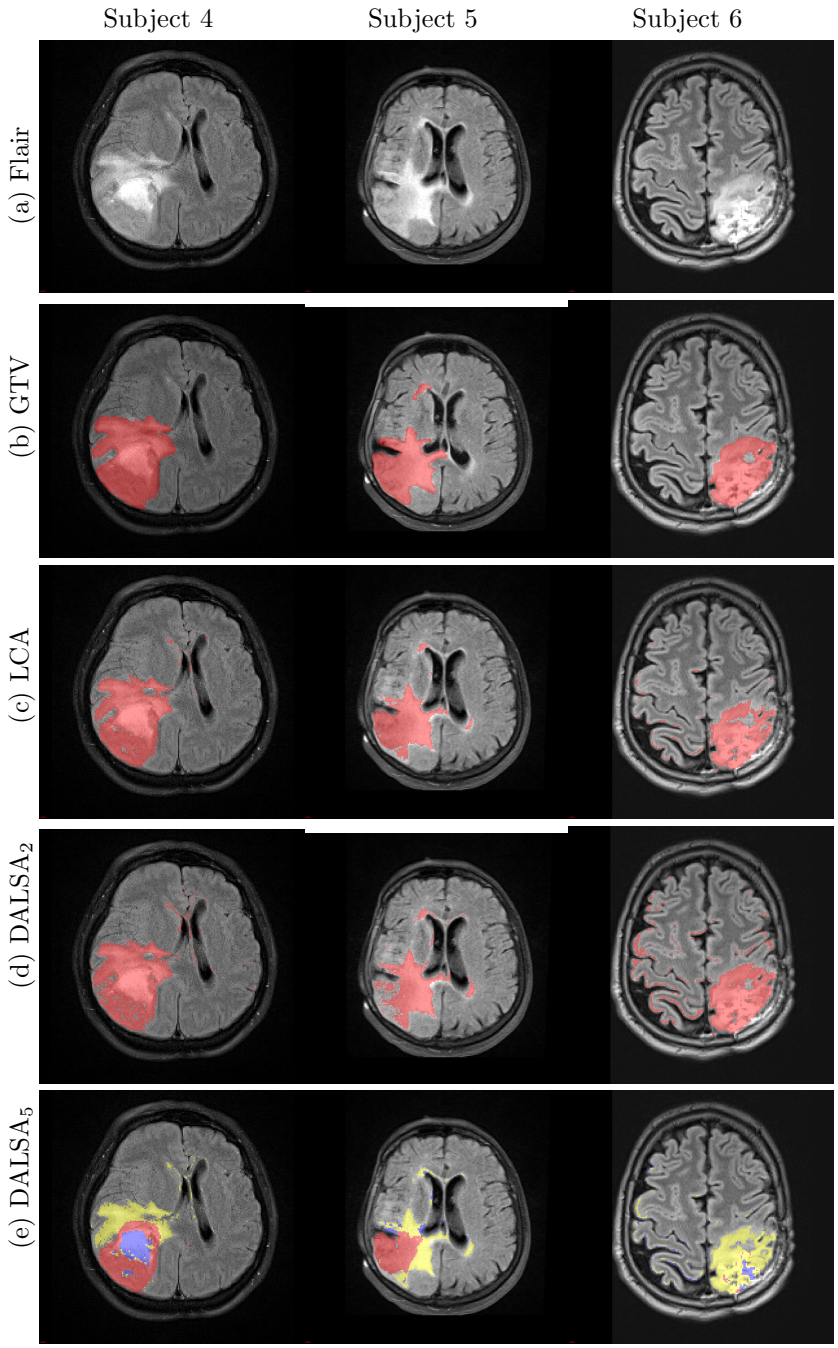


Figure 5.30: Example axial slices from setup I. (a) FLAIR image, (b) gold-standard segmentation (c) result of classifier trained on complete segmentations, (d) DALSA with 2 classes, and (e) DALSA with 5 classes. In (b-d), the red colour indicates 'gross tumour volume'. The colour coding in (e) is: yellow: 'edema', red: 'active tumour', blue: 'necrosis'.

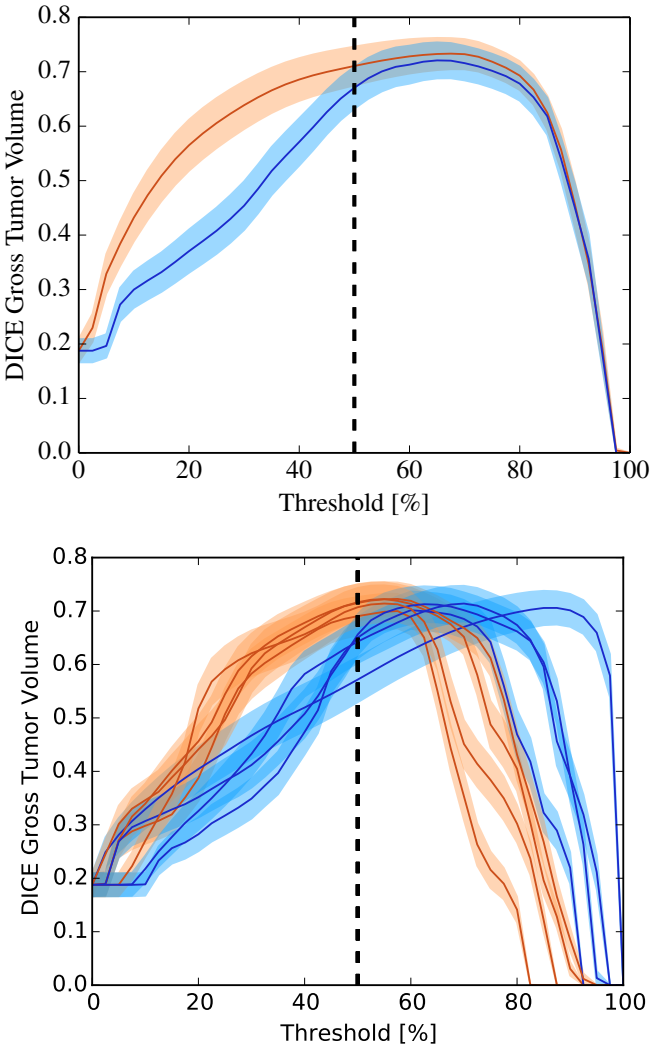


Figure 5.31: Mean Dice score (line) and standard error (area) for LSA (blue) and DALSA (red) at varying decision thresholds. A balanced curve indicates a well-balanced classifier, while a a skewed curve indicates an under- or over-representation of a class. The default threshold is at 50 % for a two-class problem. Top: Curves generated from the set of SURs created by the expert rater. Bottom: Different curves for the different labelling strategies.

TABLE 5.5
STATISTIC SIGNIFICANCE BASED ON WILCOXON SIGNED-RANK TEST FOR DALSA EVALUATION

	LCA	$LCA_{\%}$	$DALCA_{\%}$	LSA_2	$DALSA_2$	LSA_5
$LCA_{\%}$	**.001** ↑ .007	X				
$DALCA_{\%}$	**.001** ↑ .007	**.017** ← .001	X			
LSA_2	**.003** ↑ .087	**.015** ↑ .077	**.012** ↑ .078	X		
$DALSA_2$	**.035** ↑ .019	.409	.387	**.015** ← .038	X	
LSA_5	**.003** ↑ .046	**.028** ↑ .039	**.023** ↑ .040	.121	.191	X
$DALSA_5$	**.017** ↑ .019	.220	.220	**.003** ← .051	.251	**.003** ← .027

Uncorrected p-values of Wilcoxon signed-rank test indicating differences in segmentation results based on the Dice score for the gross tumour volume. $p \leqslant .05$ is shown in bold. The absolute difference of the group median Dice scores is shown below the significant p-values. Arrows point to the group with higher median score. For example, LCA performs significantly better than $LCA_{\%}$.

Classifier results: : DS 1 and SVM

Figure 5.33c shows the results obtained by SVM-based classification. Again, DALSA outperformed LSA in all experiments ($p \leq 0.005$). DALSA results on the basis of SVM did not significantly differ from DALSA results on the basis of random forests (p-values between 0.08 and 0.28).

Validation experiments: DS 3-2013 and RDF

Setup III was used to evaluate the performance of DALSA on the basis of the BraTS 2013 challenge data (c.f. dataset DS 3), in contrast to other segmentation approaches that are trained on complete segmentations. For the experiments, the pipeline of Kleesiek, Biller, et al. (2014), who scored third on the on-site BraTS 2014 challenge, was adapted. The same preprocessing, features, and post-processing as in the original work were used. Only the sample selection was varied. Instead of randomly drawing a fixed number of samples for each tissue class (which corresponds to the $LCA_{\%}$ training scheme), LSA or DALSA were used on the basis of the SURs that had been defined.

On the basis of setup III, it is assessed whether it is possible to integrate

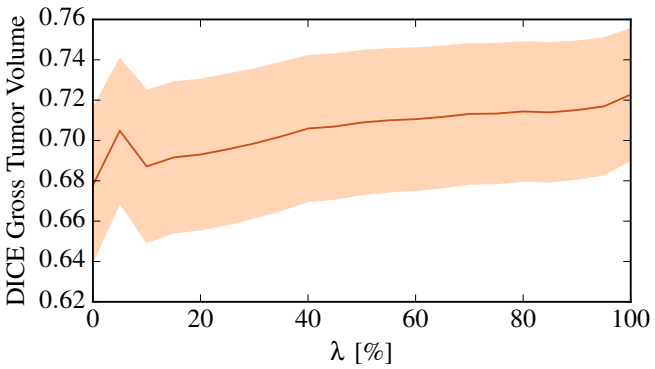


Figure 5.32: Mean and standard error for the leave-one-out experiments acquired with different λ. No parameter beside λ is changed during the experiment, tree depth is set to four in all iterations.

the approach in another existing tumor segmentation pipeline and compared the obtained results for LSA and DALSA with state-of-the-art methods that were trained on complete manual annotations on the basis of the ongoing BraTS 2013 challenge. It is further evaluated the positive effect of DALSA on a second, independent dataset.

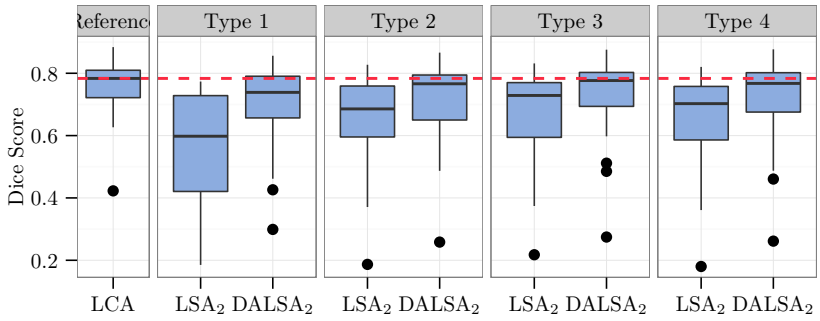
Validation results: DS 3-2013 and RDF

Examples of BraTS 2013 challenge results are shown in Figure 5.34. On the 10 test datasets DALSA yield a visible increase in segmentation accuracy with respect to Dice score (0.84 to 0.86) and Sensitivity (0.74 to 0.78) for GTV[5]. The Positive Predictive Value was reduced from 0.94 to 0.93. The resulting segmentation quality was similar to those achieved in the original approaches of Kleesiek et al. and Peres et al. that were trained on complete segmentations (reported Dice scores for both was 0.86). While sensitivity was clearly lower than in both previous approaches (0.91 and 0.87), this was compensated by the Positive Predictive Value (0.83 and 0.85 in the previous approaches). P-values are not calculated due to the small number of test subjects.

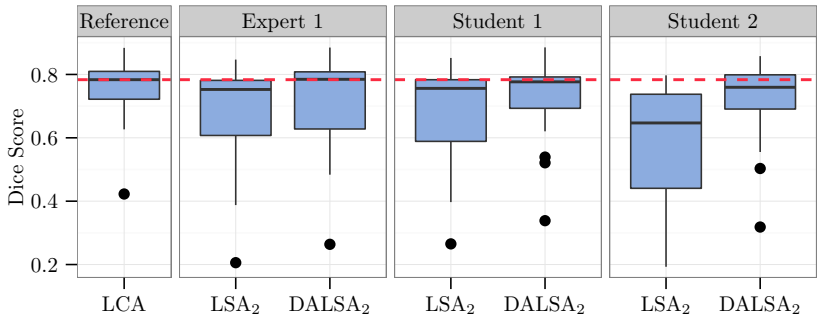
Annotation Time and Performance

The mean times required both for creation of the training data and for training and application of the forests are provided in Table 5.6. The SUR-based training was faster than training with sampled or complete data. Since less data needed to be labelled and labelling was more straightforward, the sparse annotation took less than five minutes per patient (for all annotation strategies), while the full annotation took more than six hours (a reduction of labelling time by a factor of more than 70).

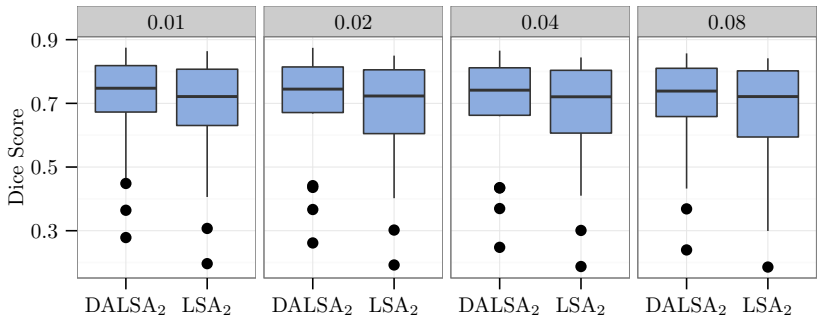
[5]LSA and DALSA-results were obtained with the implementation of the approach of Kleesiek, Biller, et al. (2014). They differ from the original results due to the different training setting.



(b) Different labeling schemes



(b) Different experts



(c) SVM-based classifiers

Figure 5.33: Dice scores obtained by leave-one-patient-out experiments. (a): Evaluation of different labelling schemes. (b) Variability between different rater that drew the SURs independently and blindfolded to the complete tumour segmentation. (c) Comparison of LSA_2 and $DALSA_2$ using SVM instead of RDF. The cost factor determines the noise sensitivity.

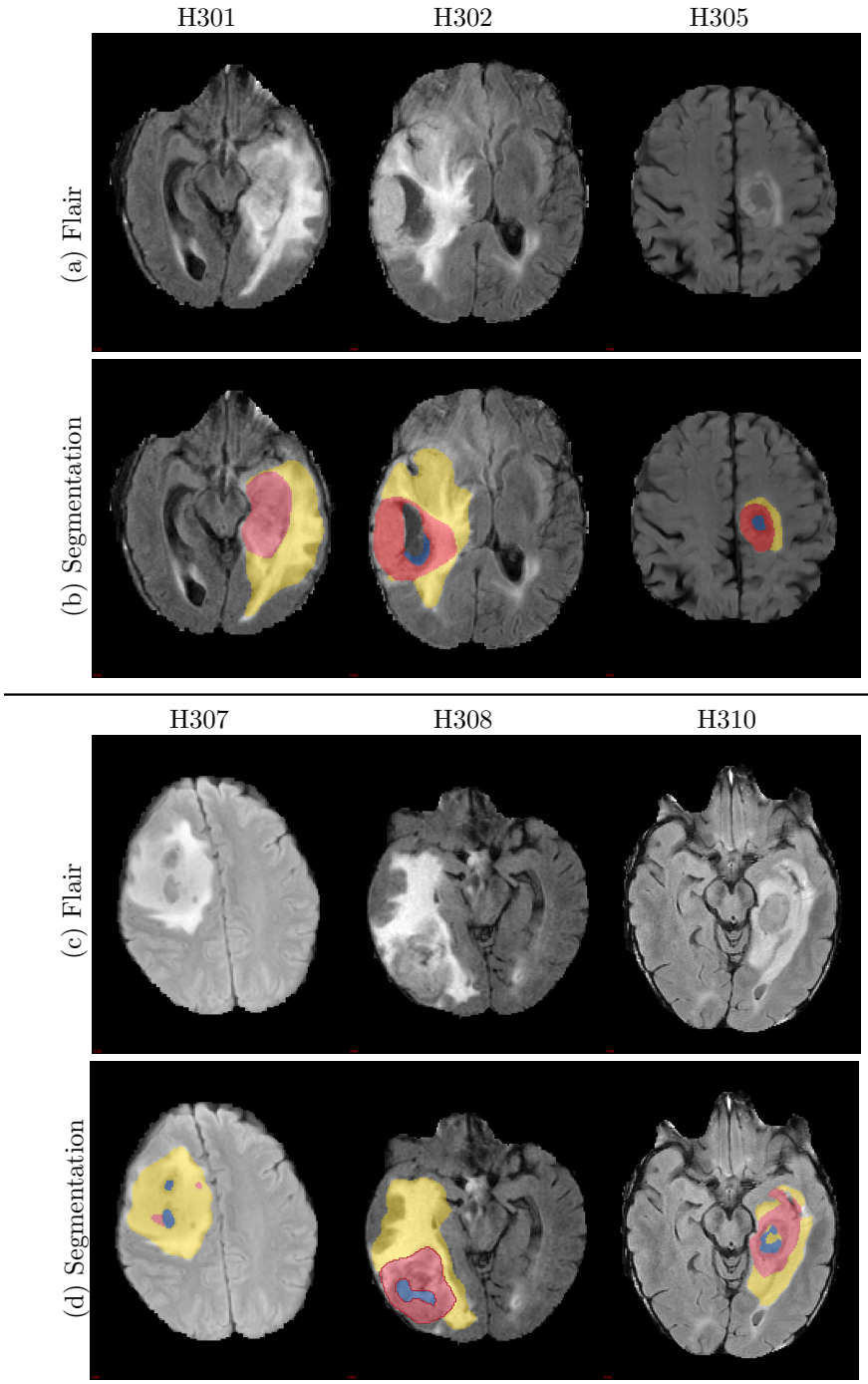


Figure 5.34: Example axial slices from the validation set using DS-3 (BraTS 2013 data set). The colour coding is: 'yellow': edema, 'red': active tumour, 'blue': necrosis. The subject name is identical to the identifier within the original training dataset.

TABLE 5.6
RUNTIMES

Method	Training time	Prediction time	Tree depth
LCA	192.5 ± 1.67 min	226 ± 41.0 sec	12
$LCA_{\%}$	46.9 ± 1.1 sec	149.3 ± 16.3 sec	10
LSA_2	12.4 ± 1.1 sec	45.7 ± 4.3 sec	4
$DALSA_2$	63.8 ± 14.4 sec	74.4 ± 8.3 sec	6

Labelling- and runtime for the different learning schemes. The labelling time is measured using DS-1. The runtime is measured using a standard PC (Intel Core i7 @3,2GHz, 32 GB Ram)

5.4.2 Learning from Only Positive Annotations

The evaluation of the proposed PU-learning approach is done using the in house glioma dataset DS-1. For the small annotations, the previous described SUR annotation is used. The sparse annotations covers $0.25\,\% \pm 0.15\,\%$ of thee complete brain volume and $2.6\,\% \pm 1.5\,\%$ of tumorous tissue. Of the sparse annotated voxels $55.1\% \pm 16.5\%$ belonged to negative samples, i.e. tumorous areas. SURs that cover healthy areas are discharged for the evaluation of the PU-learning.

Estimation of π

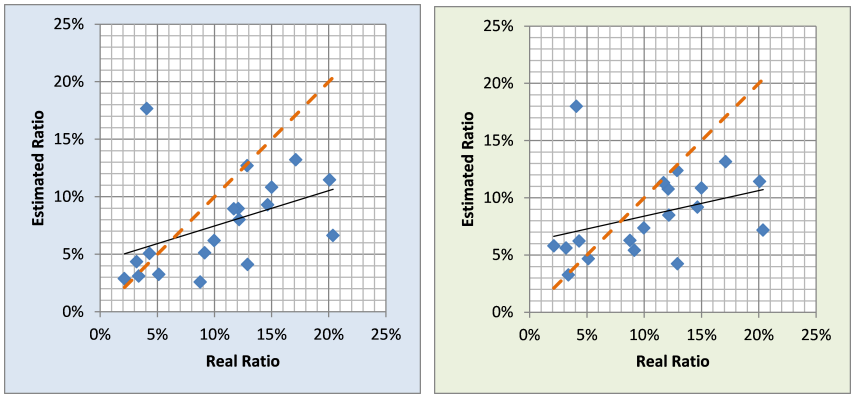
The estimation of the π is a crucial step in the proposed approach. To evaluate this approach, the tumour rate is calculatewd based on the full GTV segmentation. This is done by dividing the number of voxels labelled as 'GTV' by the number of voxels within the whole brain mask. The result of four different π estimation methods are compared to this ratio. Two additonally selected manual approaches and the two described fully automatic approaches. The manual approaches are chosen, as they allow to model the clinical routine. Since these data might be available due to clinical reasons, they could be used without further need of labelling. In comparison, the automatic approaches (PEPE and DA-PEPE) can be used if these manual estimations are not already available.

The quality of the estimation is measured with different approaches. First, the mean tumour ratios are reported, i.e. the average estimated ratio between tumorous and healthy tissue. Beside showing the differences in the estimations, this indicates whether a methods over- or underestimates the tumour volume ratio. Further, the mean absolute error compared to the GTV-based ratio estimation. Finally, the Pearson correlation between the ground truth and the estimations are calculated.

Table 5.7 shows the result of the analysis. The algorithmic estimation, which requires no manual interaction, is less accurate and underestimate the tumour prior. Both algorithmic approaches produced one outlier for the same patient (Figure 5.35). A visual inspection of the images of the corresponding subject reveal a low contrast between tumorous to healthy tissue. Without this outlier the Pearson correlation coefficient between the estimated and reference class prior is 0.761 and 0.690 for PEPE and DA-PEPE respectively. The outlier are kept within the training base for further experiments.

TABLE 5.7
ESTIMATIONS OF CLASS PRIOR π AND THE CORRELATION BETWEEN THE REFERENCE RATIO AND THE ESTIMATION

	Real Ratio	Manual 1	Manual 2	PEPE	DA-PEPE
Mean tumour ratio	10.5%	11.1%	11.7%	7.6%	8.5%
Mean absolute error	0%	16.8%	20.1%	51.2%	55.6%
Pearson Correlation	1	0.95	0.95	0.41	0.34
Labelling Time	4h	1 min	1 min	0	0



PEPE DA-PEPE

Figure 5.35: Distribution of the estimated tumour vs. healthy ratio. The real ratio is based on the GTV segmentation. The red dashed line gives the optimal line, the black line is the regression line.

Segmentation performance

Several leave-one-out experiments with all 19 subjects are used to evaluate the proposed method. The same training and data pipeline is used as for the previous experiments on 'Learning from sparse annotations' (sec. 5.4.1). The feature vector for each voxel is composed of the intensity values of all available contrasts and diffusion-based maps. RDF-based classifier are used for classification – although a slightly different implementation was used for these experiments. For the evaluation, the Dice score is mainly used, following the reasoning of B. H. Menze, Jakab, et al. (2015) and excluding distance measures.

Two reference modes were included in this comparison. First, a traditionally trained classifier, trained on the whole segmentation (LCA). Second, a classifier trained using the described DALSA scheme using only partly annotations, but including both types of tissue. These reference values are compared to the results of 10 different classifiers. For each of the five different tissue ratio estimation methods, two classifier are trained. One using the state of art 'global mode' and one using the proposed 'batch mode'.

Figure 5.36 shows the obtained Dice score for each method and table 5.8 shows the significance differences. Given the reference π the proposed workflow

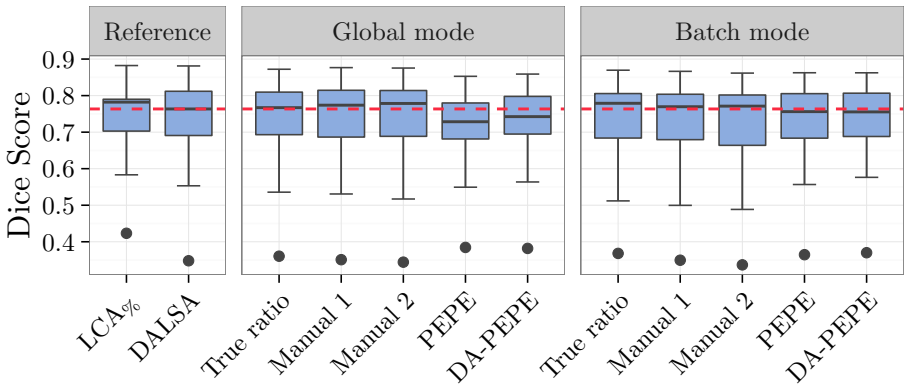


Figure 5.36: Dice scores for different leave-one-out configurations. All experiments except 'complete' used importance weighting. The horizontal dashed line shows the median of the current state-of-the-art.

yielded classifiers that are comparable to learning from positive and negative samples (SPN, i.e. DALSA). Compared to the reference prior based classifier the results improved when the manual estimation was used in batch mode. Generally, batch-mode training gives better results than global mode training. There is a small drop in the quality of results when using automated estimation of π. However, the retrieved Dice scores are still comparable to the state-of-the-art that uses both positive and negative labels.

DA-PEPE gives slightly better results than PEPE even though it has a less favourable correlation coefficient and a higher mean error. I think this is because it overestimated low π and generally performed better for high π. Since the proposed workflow is less sensitive to overestimation the so-made estimation error might lead to improved results.

Influence of π

To analyse the influence of π on the final segmentation result, multiple leave-one-out experiments were run using an artificially falsified class prior both for global and batch mode. For this, the same classifier setup is used as for the previous experiments.

Figure 5.37 shows the results. In general, both modes were stable against wrongly estimated π and were more robust against overestimation than underestimation.

Validation with DS 3-2013

To validate the findings, the experiments are repeated using the BraTS 2013 challenge training set. The same features are used as for the experiments with sparse annotations using the BraTS 2013 dataset. The results are obtained using leave-one-patient-out scheme. For this the SUR-annotation previously described and the true ratio that is calculated based on the GTV segmentation were used.

TABLE 5.8
STATISTIC SIGNIFICANCE BASED ON WILCOXON SIGNED-RANK TEST FOR PU EVALUATION

	DALSA	True Ratio(Global)	Manual 1(Global)	Manual 2 (Global)	PEPE (Global)	DA-PEPE (Global)	True Ratio (Batch)	Manual 1(Batch)	Manual 2 (Batch)	PEPE (Batch)	DA-PEPE (Batch)
True Ratio (Global)	0.71	X									
Manual 1 (Global)	0.87	0.69	X								
Manual 2 (Global)	0.78	0.49	.469	X							
PEPE (Global)	0.12	**.005** ↑ .039	**.044** ↑ .045	.070	X						
DA-PEPE (Global)	0.72	.070	.260	.494	**.000** ← .014	X					
True Ratio (Batch)	0.66	**.027** ← .012	.159	.147	.059	.355	X				
Manual 1 (Batch)	0.24	.159	**.024** ↑ .004	**.020** ↑ .009	.171	.99	.314	X			
Manual 2 (Batch)	0.24	**.005** ← .004	**.022** ↑ .003	**.003** ↑ .008	.295	.840	**.014** ↑ .008	.494	X		
PEPE (Batch)	0.94	.059	.398	.717	**.003** ← .028	.494	.376	.748	.421	X	
DA-PEPE (Batch)	0.97	.059	.546	.809	**.001** ← .027	.227	.494	.778	.398	.084	

Uncorrected p-values of Wilcoxon signed-rank test indicating differences in segmentation results based on the Dice score for the gross tumour volume. Beside DALSA, the legend gives always the ratio that is used for the comparison. $p \leqslant .05$ is shown in bold. The absolute difference of the group median Dice scores is shown below the significant p-values. Arrows point to the group with higher median score. For example, using the True Ratio in a global way performs significantly better than PEPE in a global way.

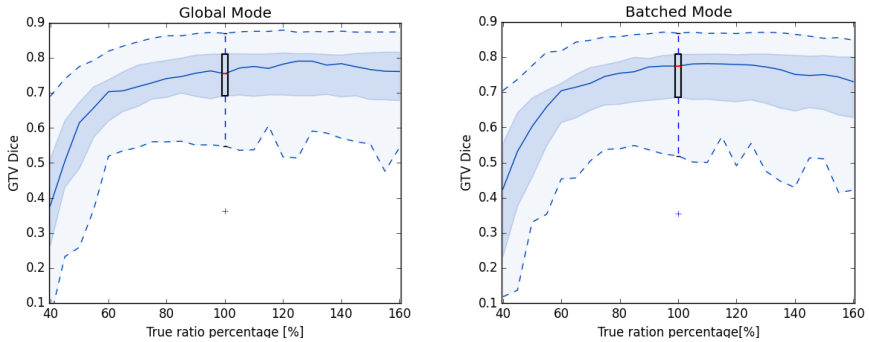


Figure 5.37: Results of leave-one-out experiments with artificially falsified π. All experiments are conducted with the same random forest tree depth (Global 5, Batched 4). The box in the diagram follows standard boxplot definitions and gives the results that are obtained with the true ratio.

Examples of the results are shown in Figure 5.38 and Figure 5.39. The obtained mean Dice score is $78.7\,\% \pm 15.7\,\%$, and the sensitivity and specificity are $78.3\,\% \pm 15.9\,\%$ and $98.8\,\% \pm 1.2\,\%$ respectively.

5.4.3 Learning from Bag-Wise-Annotations

Synthetic experiments

F. Yu et al. (2013) designed an synthetic experiment to validate the function of LLP algorithm. Based this suggestion the experiment was reproduced, naming it experiment Yu-1. The training data consists of two data bags. The samples of each bag are drawn randomly from a two-dimensional normal distribution with a standard deviation of 0.1. The first data does have a class ratio of $60 : 40$ and a mean of $(-0.25 \mid -0.25)$ and $(0.75 \mid 0.75)$ for class A and class B, respectively. The second bag does have a class ratio of $40 : 60$ and a mean of $(-0.75 \mid -0.75)$ and $(0.25 \mid 0.25)$ for class A and class B, respectively. Figure 5.40 visualizes the resulting data points. Remember, that the class label is not known during training.

A classifier is then trained using these data. As this experiment does not check the classification power of the classification algorithm but only if it is possible to create a correct classifier, the parameters can be tuned using the training data. The evaluation is then done by reusing the training data as test data – this time including the known labels.

Beside the experiments introduced by F. Yu et al. (2013), an additional setting is proposed by Patrini et al. (2014). He states that bags are often not randomly sampled but influenced by one or more covariants. He proposed to simulate this by creating an setting of 17 different experiments, each consisting of 16 bags. I will refer to these experiments as Patrini-0 to Patrini-16. Each bag consists of 100 samples, with a random ratio between the two labels. The observations of each label are randomly drawn from two normal distributions with a mean of $(\pm 1 | \pm 0.5 + \delta_b)$ and a standard deviation of 0.63 for the two features. The offset depends on the bag number b and the current experiment

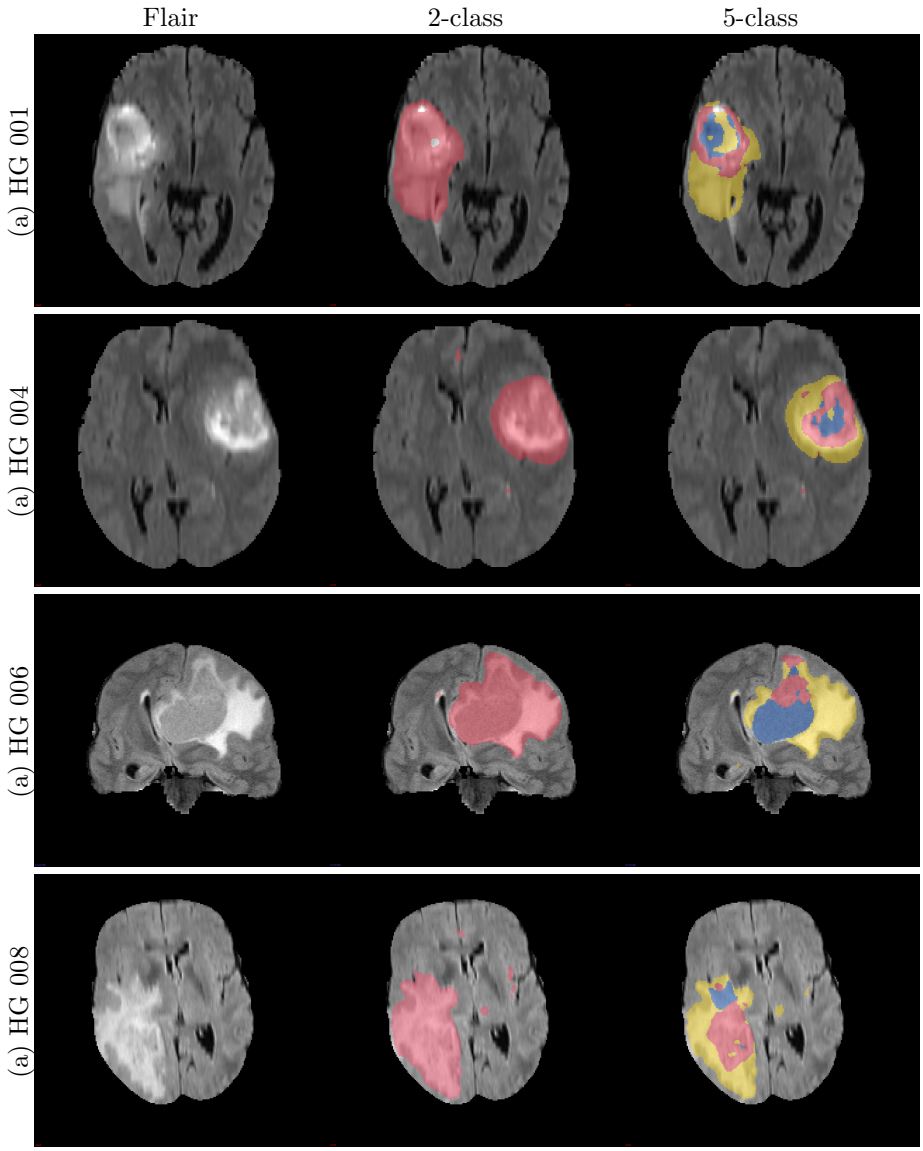


Figure 5.38: Example axial and coronal (HG 006) slices from the validation set using DS-3 (BraTS 2013 data set). The segmentation is obtained by the proposed PU-learning method. The colour coding is: 'yellow': edema, 'red': enhancing tumour, 'green': non-enhancing tumour, 'blue': necrosis. The subject name is identical to the identifier within the original training dataset.

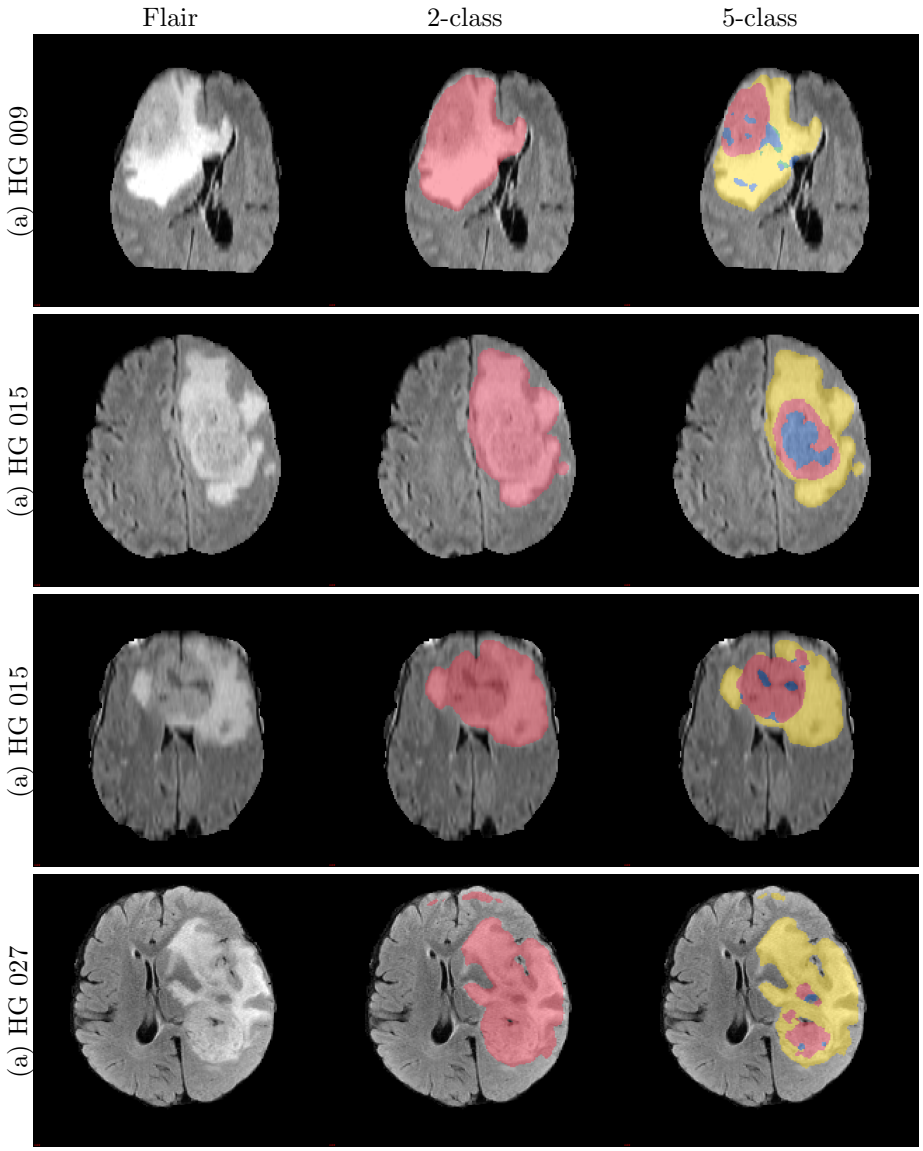


Figure 5.39: Example axial slices from the validation set using DS-3 (BraTS 2013 data set). The segmentation is obtained by the proposed PU-learning method. The colour coding is: 'yellow': edema, 'red': enhancing tumour, 'green': non-enhancing tumour, 'blue': necrosis. The subject name is identical to the identifier within the original training dataset.

]

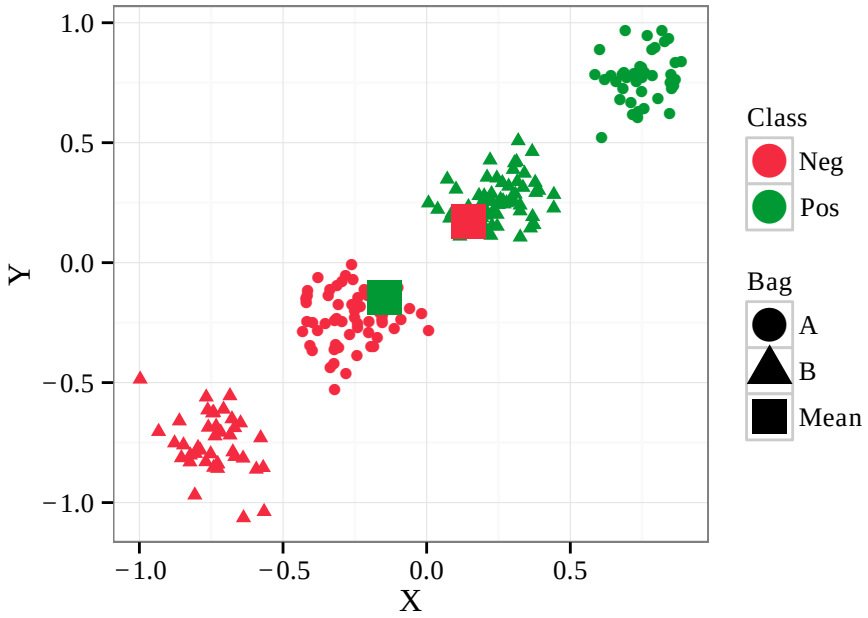
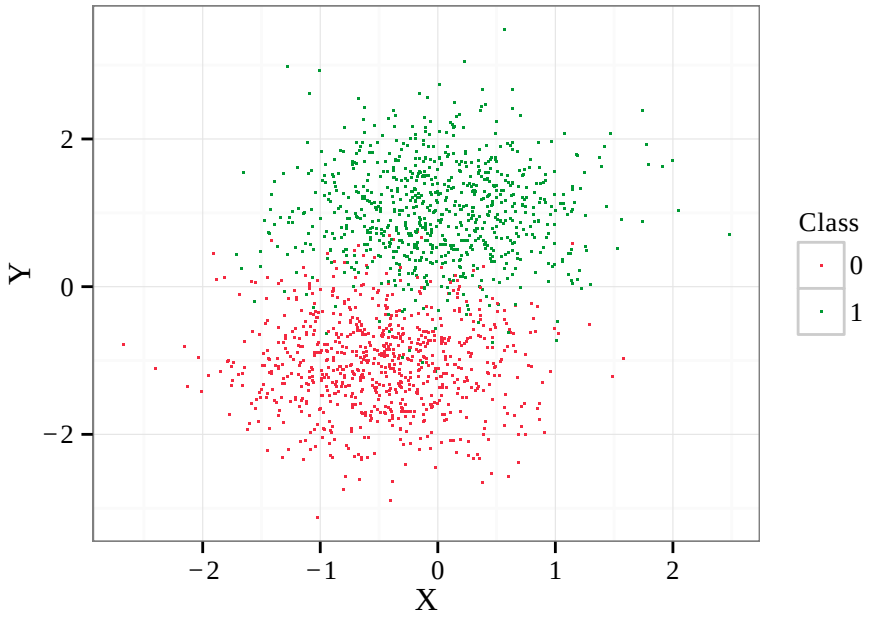


Figure 5.40: Scatter-plot of the ground truth data experiment Yu-1. The mean values of each class are highlighted since some SVM-based approaches depend on these values.

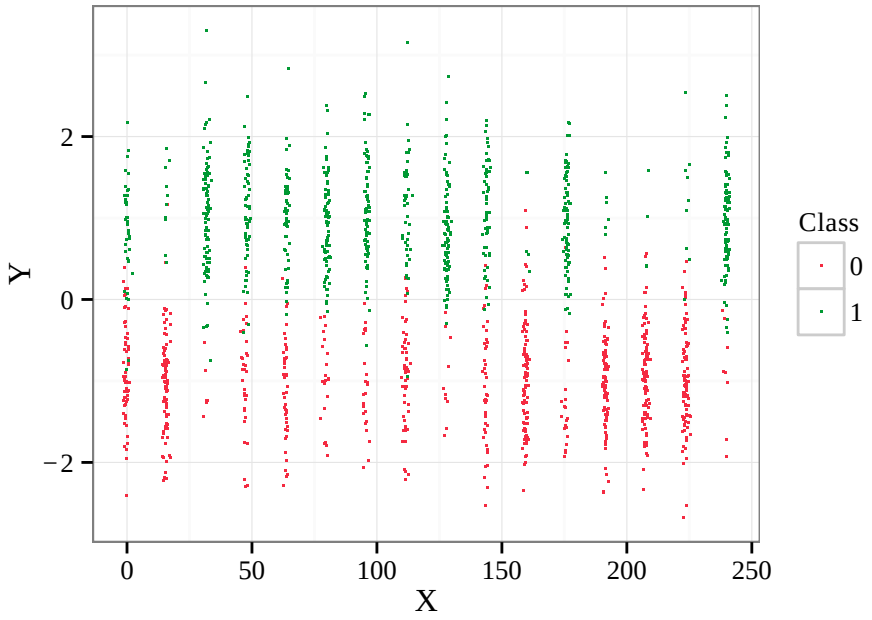
e with $\delta = b \cdot e$. Bags in experiments with higher numbers are therefore more affected by a covariate shift along the first axis (Figure 5.41). For each of these experiments, a classifier is trained with all 16 bags and the parameters being tuned using five-fold cross validation. The prediction accuracy on the same set is then obtained and reported – a separate test data set is not necessary for these synthetic data.

Two additional synthetic experiments are designed in addition to previously published experiments. The first experiment is designed to test whether the classifier can be successfully trained if the observations of one class are surrounded by the observation of the other class. For this, 10 bags with each 100 observation are created, with varying class ratios. The samples for first class are randomly drawn from a normal distribution with mean $(0|0)$, while the data of the second are drawn from two distributions with mean values of $(\pm 0.5|\pm 0.5)$ with varying ratios. The standard deviation of all distributions is set to 0.1. A scatter-plot of the experiment data is given in Figure 5.43a.

The second new experiment is designed to test the multi-label capabilities of the proposed algorithm. 16 bags with 100 observations each and a random ratio between three or four classes are generated by randomly drawing from either three or four distributions. The mean values of the distributions are $(0.5|0.5)$, $(-0.5|0.5)$, $(0.5|-0.5)$, and $(-0.5|-0.5)$. The experiments are conducted twice, once with a standard deviation of 0.1 and 0.77, respectively. The experiments are named Multiclass-a-b, while a is indicating the standard deviation and b

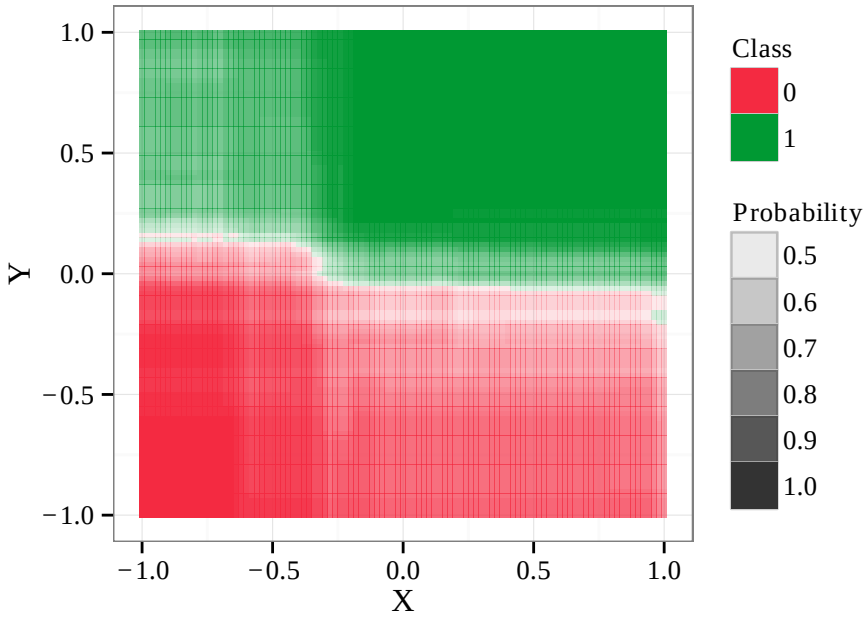


(a) Ground Truth for Patrini-0

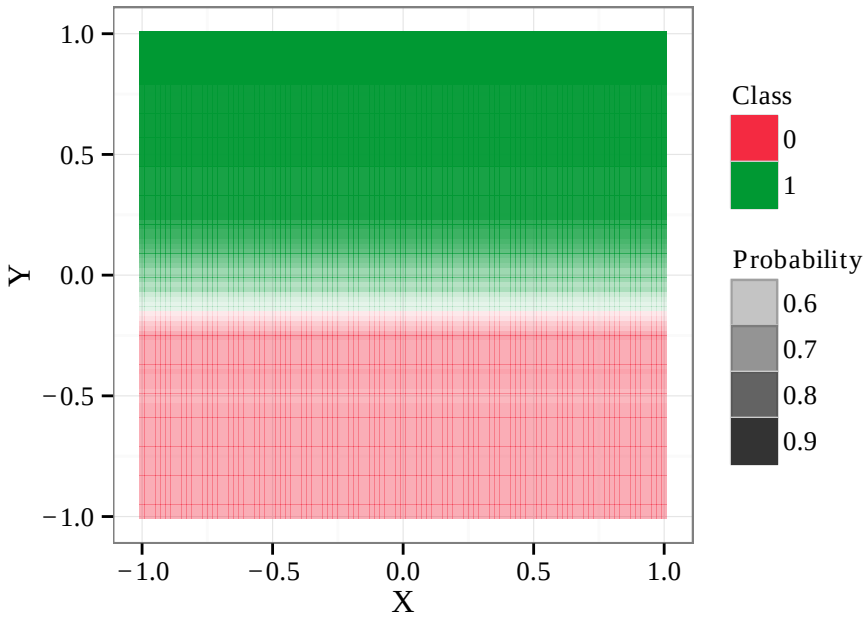


(b) Ground Truth for Patrini-16

Figure 5.41: Scatter-plot of the ground truth data for experiment Patrini-0 and Patrini-16. The shift introduced by δ is clearly visible for experimental Patrini-16.



(a) Decision Boarder for Patrini-0



(b) Decision Boarder for Patrini-16

Figure 5.42: Decision boards obtained by the two most-extreme Patrini-experiments. The colour coding indicates the class at each location, while the intensity indicates the degree of certainty. Both images show only parts of the used feature-space.

TABLE 5.9
ACCURACY OBTAINED FOR DIFFERENT SYNTHETIC EXPERIMENTS.

Experiment	Accuracy
Yu-1	100.0 %
Patrini-0	95.1 %
Patrini-1-15	92.6 % - 95.3 %
Patrini-16	95.1 %
Middle-1	99.4 %
Multiclass-1-3	100 %
Multiclass-1-4	99.0 %
Multiclass-2-3	94.4 %
Multiclass-2-4	89.9 %

the number of classes used. Figure 5.43b gives shows the training data for Multiclass-2-4.

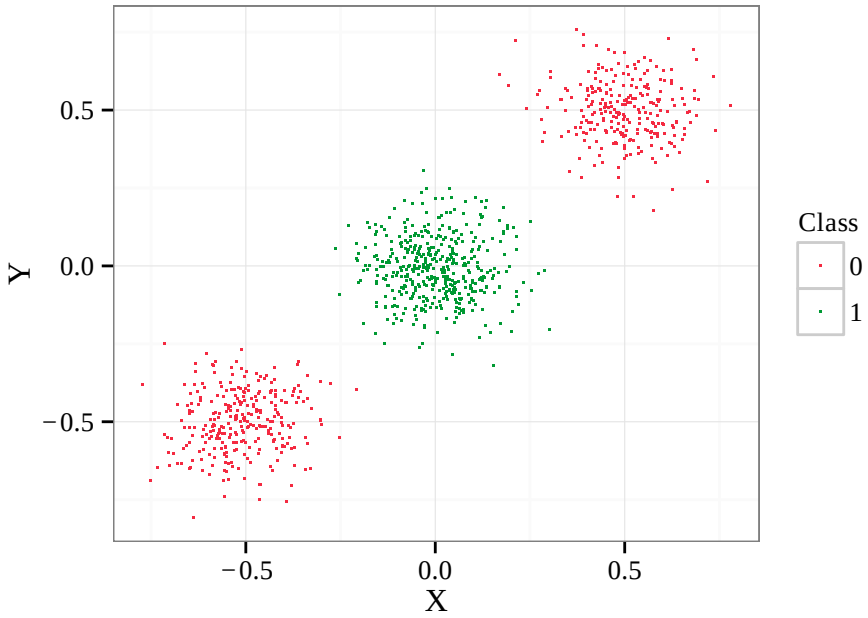
All tested synthetic experiments are successfully classified (Table 5.9). The obtained classification accuracy is 100 % if this is possible with the given training data set. For other experiments, the results is high enough to show that the training is finished successfully. This is even true for experiments that are designed for the use with SVM like Yu-1. The proposed classification algorithm is able to find the decision boarder even though it is not aligned with a single feature.

The results of the experiments Patrini-0 to Patrini-16 indicate show that the proposed algorithm is not affected by the shift of a covariate. The obtained scores are similar for all experiments, there are no outliers, the results are all in the same range and seem to differ rather by chance. Both, the worst and the best results are not the extreme experiments Patrini-0 or Patrini-16 but some intermediate experiments. Further, there are no patterns in the accuracy that indicate an optimum. This is also validated by plotting the decision boarders (Figure 5.42). While the decision boarder is only affected by the non-shifting feature for Patrini-16, the decision boarder of Patrini-0 seems to be more adapted to the small offset within the training data.

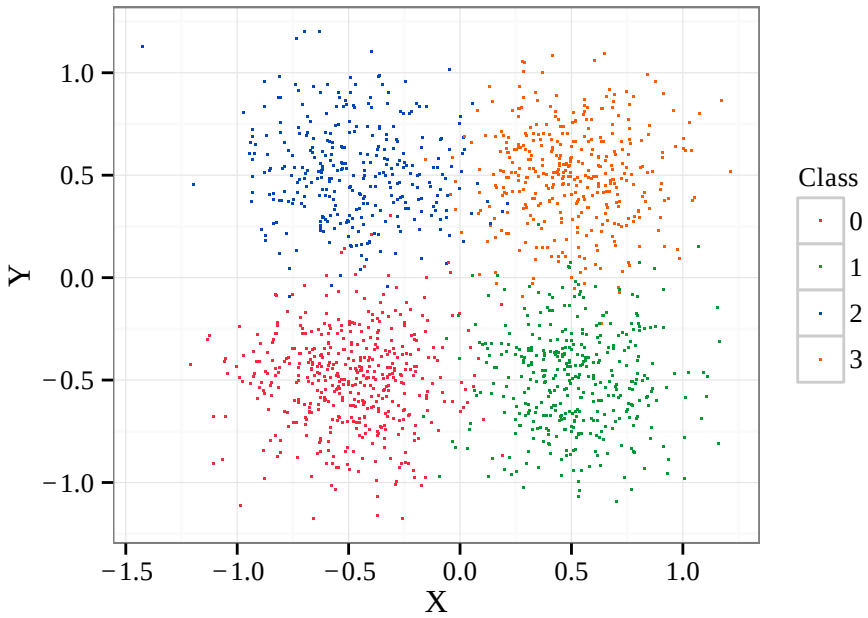
The unclear decision boarders obtained from the other synthetic experiments (Figure 5.44) doe show that the decision boarder is affected by the provided training data. While the decision boarder is in principal similar to the general decision boarder, there are some small adaptations to the random distributed training data.

Parameter evaluation

Parameter influence is investigated using well-known machine learing datasets (DS-5). The three datasets that are citet most often are selected. Dataset ‘a1a’ and ‘satimage’ are not used since as they have uncommonly large feature spaces or binary features. For the other datasets the best configuration obtained by the previous experiments are taken. To evaluate the influence of a single parameter, this parameter is changed over a meaningful range, while all other parameters are kept fix. The classification accuracy is obtained by conducting a 5-fold cross validation for each parameter configuration. This is repeated for all seven

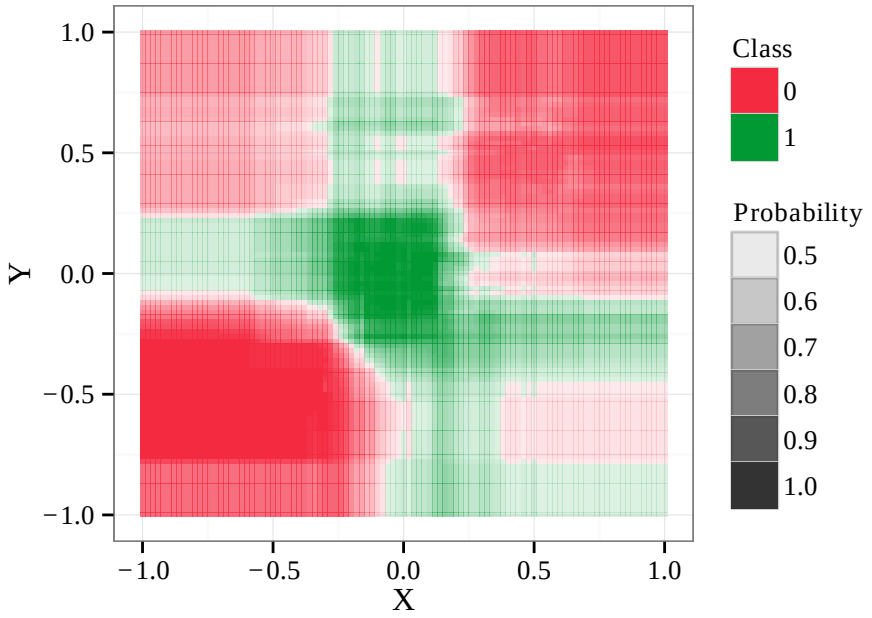


(a) Ground Truth for Middle-1

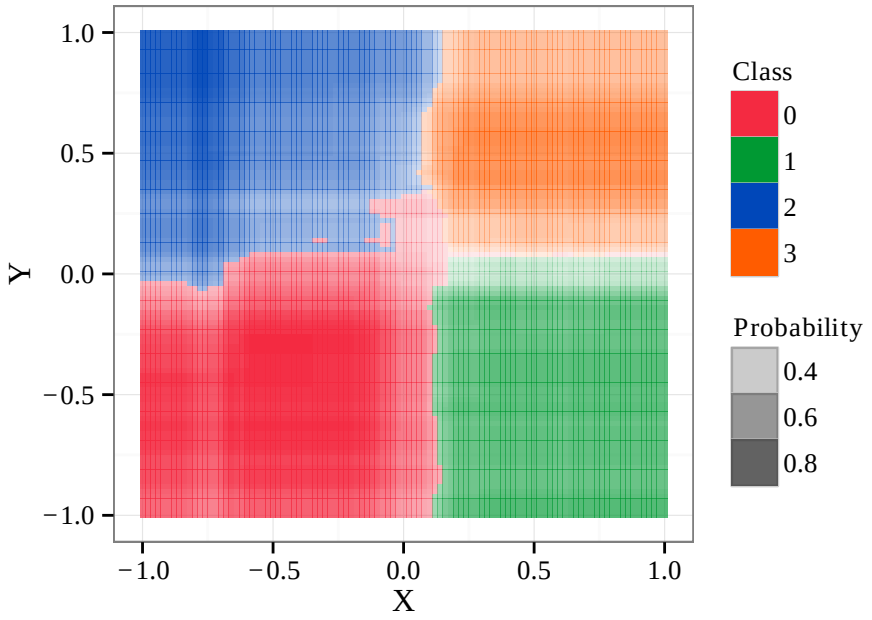


(b) Ground Truth for Multiclass-2-4

Figure 5.43: Scatter-plots of the ground truth data for experiment Middle-1 and Multiclass-2-4. Multiclass-2-4 is more challenging that Multiclass-1-4 as the data are not clearly separable. The multiclass experiments are similar to their three class counterparts, which are only missing one class.



(a) Decision Boarder for Middle-1



(b) Decision Boarder forMulticlass-2-4

Figure 5.44: Decision boundary obtained by the experiment Middle-1 and Multiclass-2-4. The color coding indicates the class at each location, while the intensity indicates the degree of certainty.

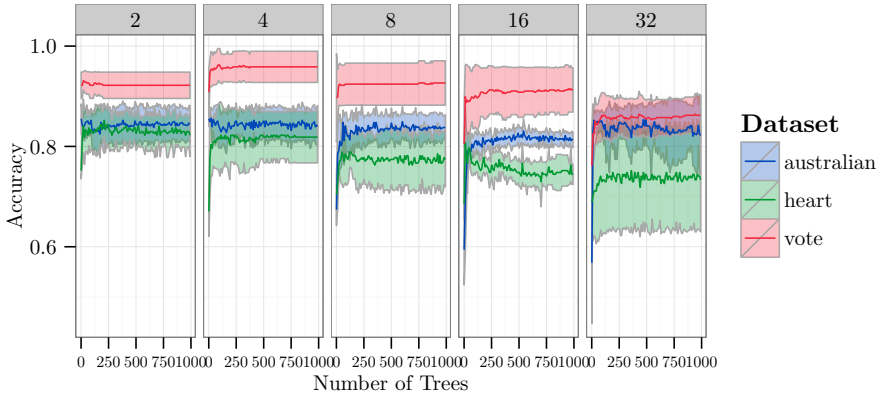


Figure 5.45: Results of parameter sweep controlling number of trees.

different parameters.

Figure 5.45, 5.46, and 5.47, gives the results of these runs. Splitted by different sub-datasets and different parameters the influence of each parameter can be seen in these experiments.

Comparison with State of the Art

The experiments of the training data are based on the experiment description of Rueping (2010). Following this scheme allows to compare the findings to the results reported in other papers as this scheme is usually taken for the experiments.

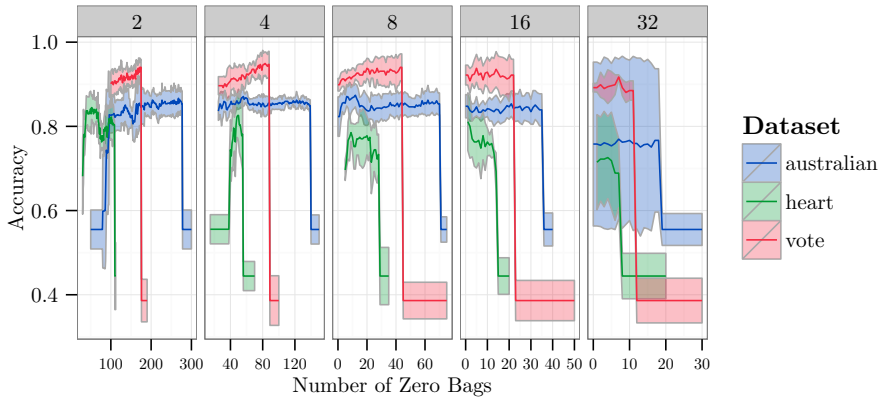
As already described, the available samples are randomly slitted into bags of various size. These data are used for a five-fold cross evaluation. The evaluation is performed based on the mean accuracy obtained. This process is repeated five times to account for the random elements in the data.

The parameter tuning is based on the previous finding and done using an inner cross validation on the training data. As the true labels are unknown at this stage. Following the suggestion of F. Yu et al. (2013) the bag-level error will be used as error measure:

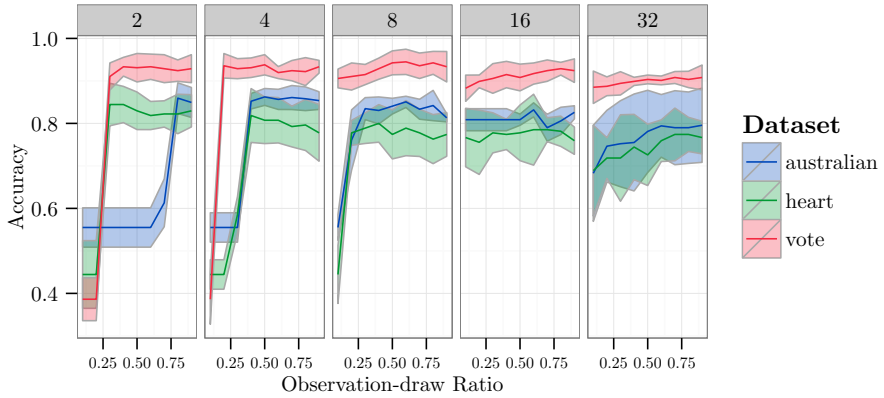
$$Err = \sum_{k} \mid \tilde{p}_k - p_k \mid \tag{5.1}$$

with $\tilde{p}_k$ being the estimated class proportion for bag k and p_k being the given class proportion.

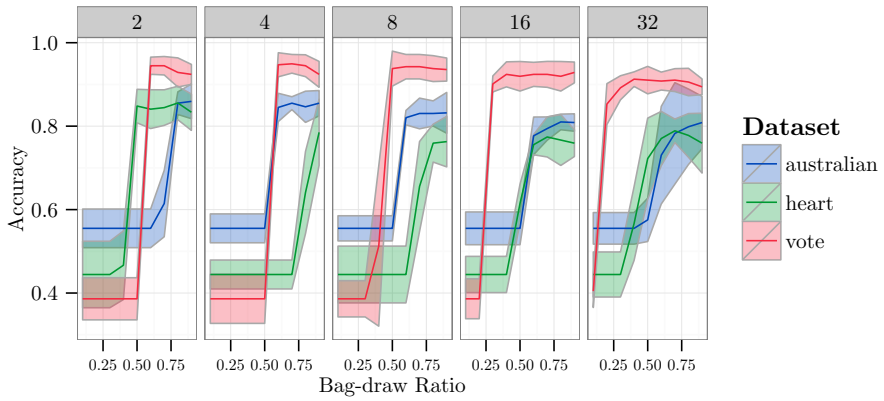
The results of the experiments are given in Table 5.10. The table contains also results for other tree ensemble learning based approaches that are taken from the literature as baseline. As these results are obtained within different settings – for example by using additional test sets or 10-fold cross validation instead of 5-fold – the results cannot be completely compared.



(a) Parameter sweep over allowed number of zero bags

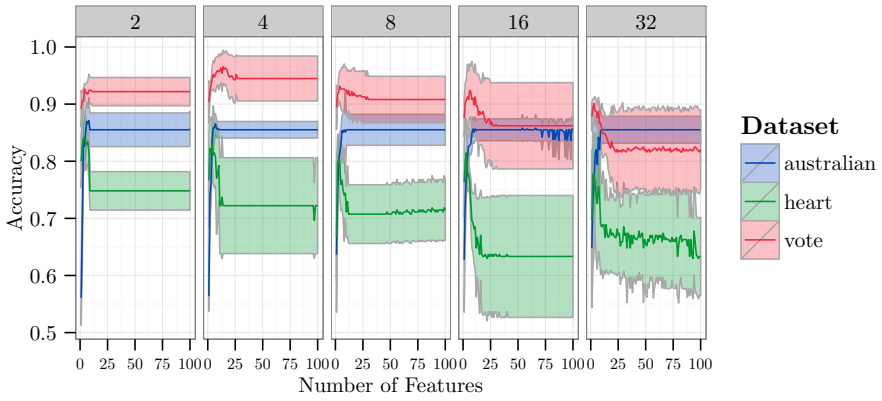


(b) Parameter sweep over observation bagging

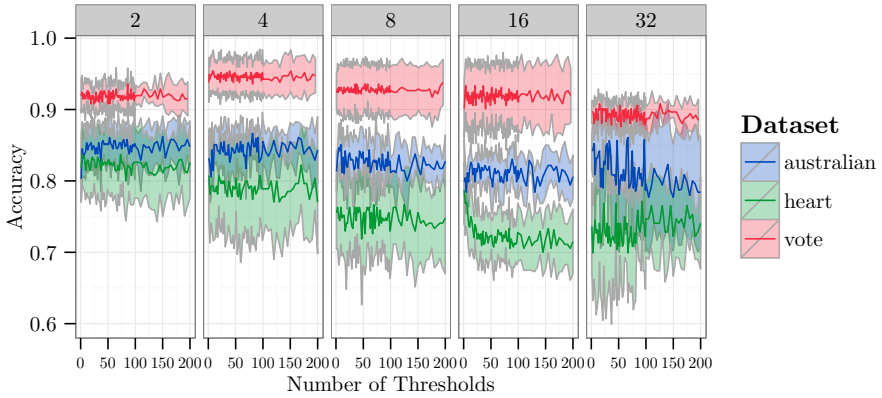


(c) Parameter sweep over bag-wise bagging

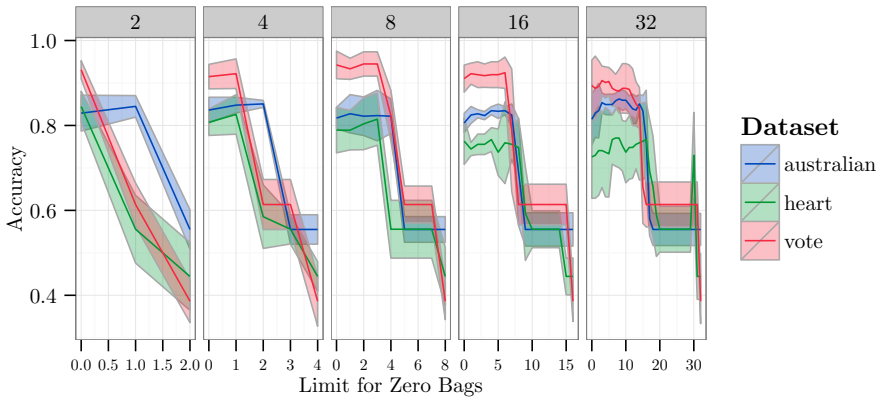
Figure 5.46: Results of parameter sweep controlling the bagging and the number of empty bags.



(a) Parameter sweep over allowed number of zero bags



(b) Parameter sweep over observation bagging



(c) Parameter sweep over bag-wise bagging

Figure 5.47: Results of parameter sweep controlling the randomness in feature and threshold selecting as well as the definition of empty bags.

TABLE 5.10
LP-Forest classification accuracy

Dataset	Method	2	4	8	16	32
a1a	LP-Forest	78.82	78.82	76.20	76.38	70.59
	Rand.-Forest			94.9		
australian	LP-Forest	84.93	85.21	82.90	86.67	83.62
	Rot. Feature			86.57		
satimage	LP-Forest	87.37	86.79	86.72	84.58	65.91
	CO2-Forest			91.1		
vote	LP-Forest	94.71	94.48	93.10	92.64	89.43
	Rot. Forest			96.26		
Statlog (heart)	LP-Forest	80.37	78.89	75.19	80.00	78.89
	Rot. Forest			82.25		

Average accuracy obtained on different public datasets. The results for the other methods are taken from literature, and might be obtained by using additional data. (Rotation Forests: Rodriguez, Kuncheva, and Alonso (2006) , Rotated Features: L. Zhang, Ren, and Suganthan (2013), CO2-Forest: Norouzi, M. D. Collins, et al. (2015), Random Forest: Fernández-Delgado et al. (2014))

Real life experiments

To validate the use of the new proposed algorithm within the field of tissue characterization a semantic segmentation algorithm is trained using dataset DS-1. The features and preprocessing is kept the same as for the previous described experiments with these datasets. The only thing that changed is the annotation. There are no manual annotations used during the training. Instead the class ratio is given by estimating it from the ground truth segmentation.

Using a leave-one-out scheme, a classifier is trained for every subject. Using this classifier, the subject is segmented and the Dice score is calculated. Parameter selection and tuning is again based on previous findings and is done individually for each patient. The achieved classification accuracy is $90.4\,\% \pm 5.4\,\%$ and $91.3\,\%$ for mean and median respectively. The Dice score for non-brainmatter-tissue is $79.3\,\% \pm 9.9\,\%$ and $82.1\,\%$ for mean and median respectively.

5.5 Assessment of image acquisition

In this section the experiments are described that have been carried out to evaluate the two proposed methods for assessing the information content of image modalities.

5.5.1 Multi-rater region of interest comparison

Synthetic experiments

The effect of the proposed algorithm for the combination of different ROIs is firstly evaluated using synthetic data. To simulate the voxels covered by the

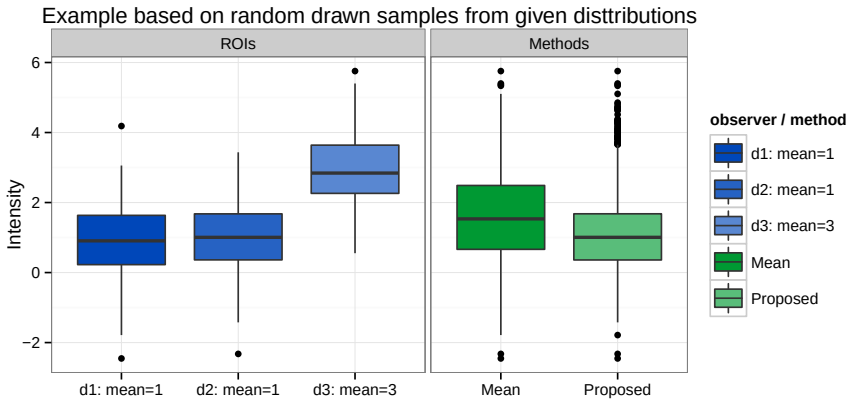


Figure 5.48: Toy example illustrating the effect of different fusion methods. Three artificial ROIs are created by randomly sampling from different normal distributions. For each ROI 200 samples are drawn with a standard deviation of 1. The results of the standard (mean) and proposed fusion are shown.

ROIs artificial voxels are sampled from Gaussian distributions. Three different ROIs are samples. The standard deviation of each Gaussian distribution is set to 1. Two of the ROIs are sampled from distributions with the mean value set to 1, and one distribution is sampled with a mean value of 3. The last distribution simulates a wrong placed ROI.

The so created three different ROIs are combined using two approaches: The standard approach, where each observation is weighted with the same value to create a new mean distribution. As second approach, the three simulated ROIs are combined using the proposed approach.

Figure 5.48 shows boxplot of the simulated ROIs and the resulting combination of them. The resulting distribution is more affected by the outlier ROI if the standard approach is used. This is not the case with the proposed algorithm. The final weight that is assigned to the 'outlier' ROI is close to zero, the actual value depends on the randomly drawn samples of each of the simulated ROI.

Real life data experiments

In addition to the evaluation based on synthetic data, the performance of the algorithm is evaluated if real clinical data are used. For this, it was decided to use dataset DS-1. Beside the advantage, that this dataset contains diffusion weighted images, it does contain annotations similar to ROIs – namely the SURs-based annotation – made by different observer. As already mentioned, there are seven sets of ROIs available for every image made by three different observer.

To make the evaluation easier to read, the experiments were limited to a single contrast – namely mean diffusivity (MD) – and a single type of tissue – namely edema. MD is well suited for the detection of edema, as it can be seen in Figure 5.1. Further, the SURs are created by using mainly FLAIR and

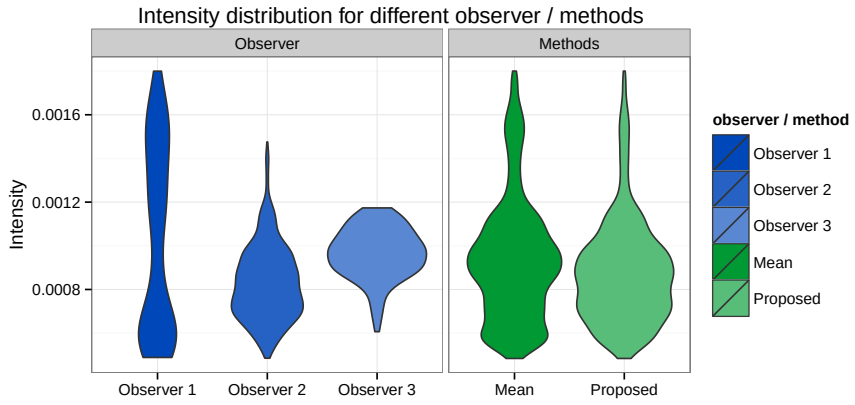


Figure 5.49: Fusion result on a single patient. Three ROIs are drawn by different observer. The huge differences in the appearance of each ROI can be clearly seen. The two different fusion approaches are compared to the individual ROIs.

$T1_w$c. I think that this allows a realistic simulation of ROIs, As the placement is usually done using a different contrast than those which is actually analysed.

Figure 5.49 shows the distribution of intensities covered by the ROIs of the three observer within a single image. It also shows the combined distributions of all ROIs for a simple mean combination and the proposed combination. The average run-time of the proposed algorithm for a single patient is $0.05\,\mathrm{s} \pm 0.003\,\mathrm{s}$ using a standard PC[6]. The whole calculations for all 18 patients are finished in less then half a second. On average, 11.7 ± 4.9 iterations were run until convergence (threshold is set to 0.001) is reached.

For each patient, two fusion ROIs are calculated using the standard and the proposed approach. Figure 5.50 shows the distribution of the mean values of all nine ROIs, i.e. the seven observer placed ROIs and the two fusion ROIs and Figure 5.51 shows the resulting weights distribution. The mean distance between the mean intensities of all observer-created ROIs and the simulated ROIs is $2.04 \cdot 10^{-5}$ for the standard approach and $-9.23 \cdot 10^{-6}$ for the proposed approach. The median distance between the mean intensities is $4.32 \cdot 10^{-5}$ and $8.30 \cdot 10^{-6}$, respectively.

Further the mean value of an observer-created ROI is compared directly to the corresponding mean intensity of the artificial ROI. In 84 cases, the difference is smaller for the proposed approach, and in 42 cases the difference is smaller for the standard approach.

5.5.2 Classifier-based information assessment

DS-2 is chosen for this experiment as it contains two different modalities which are usually not taken together. As additional advantage, the relation between both modalities is well-known (Stiller et al., 2015). The used patient collective is divided into two groups for the experiments. The first group consists of 15 randomly selected patients and is used for the parameter tuning. I refer to this

[6]Intel Core i7 @ 3,2 GHz, 32 GB Ram

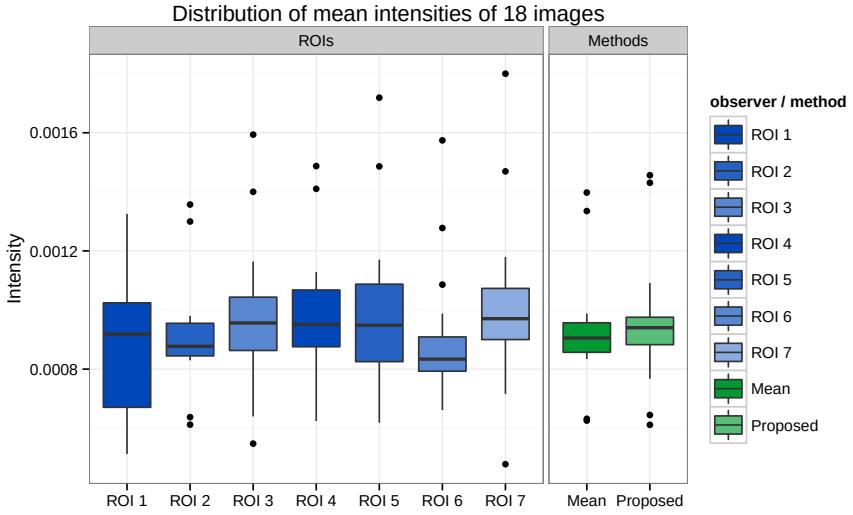


Figure 5.50: Distribution of mean values from 18 patients. For each patient two fused ROI are calculated, one using the standard mean method and one using the proposed approach.

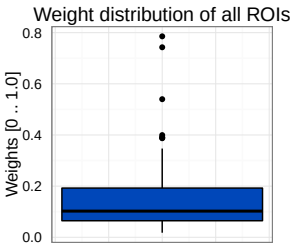


Figure 5.51: Weight distribution obtained by applying the proposed method to the real-life data.

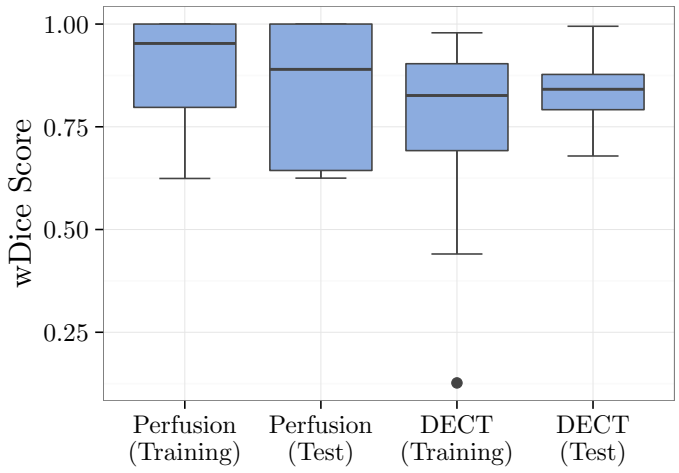


Figure 5.52: Classification results for tumorous tissue. For every modality (Perfusion CT or DECT) two results are shown. The parameters on the training set were obtained by cross-validation, while those of test were estimated on the training group.

group as training group. The remaining 5 patients are used as testing group. Using this group, it is possible to get unbiased results for the algorithm and the parameters found by using the training group. The classifier setup is done as described in the method section (sec. 4.4.2).

For parameter tuning, the training group is randomly divided into four groups and performed cross-validation. The parameter sets that gave the best mean weighted Dice (wDice) score for healthy and tumorous tissue is then chosen. This was done twice, once for DECT-based classifier and once for Perfusion-based classifier. Using these parameters, the remaining patients of the test groups are then classified. Qualitative results of the segmentation algorithms are shown in Figure 1 (Perfusion CT) and Figure 2 (DECT). For both methods a standard CT, an exemplary measurement map, the obtained tumour segmentation, and the tumour risk map are shown for the same patients.

Quantitative results are shown in Figure 5.52 and 5.53. It shows the wDice-scores for tumorous and healthy tissue. For every method the results obtained on the training group and the test groups are separately shown. Within the test group, one patient had a significant lower perfusion / iodine values. This led to completely false classifications, regardless of the used modality. Beside this patient the results between test and training groups are usually comparable.

The quality of the segmentations and risk maps is further evaluated by the rating of a medical expert. To avoid a bias toward one modality, the used modality is hidden by showing the rater only the labels ‘a’ and ‘b’. The assignment of the modalities to these labels is done randomly per subjects. The task of the expert was to rate each segmentation and risk map individually by using the grades 1 (very good) to 6 (unusable). The segmentations and risk maps were shown as overlay to an contrast enhanced CT image slice. The quality is evaluated both with respect to the anatomical correctness as well as the quality compared to the underlying image. The scores obtained for the four different

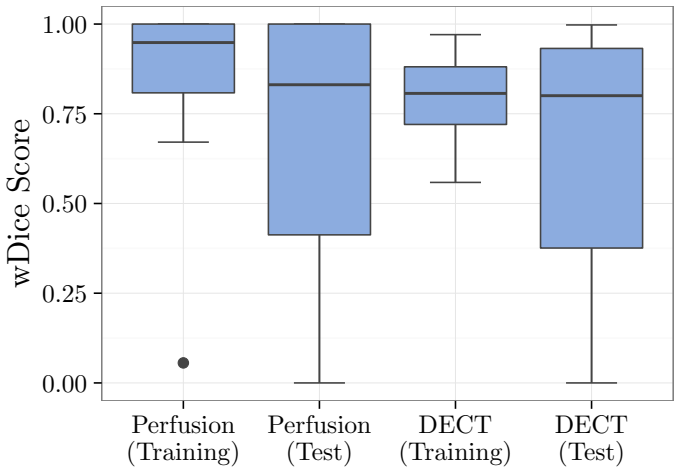


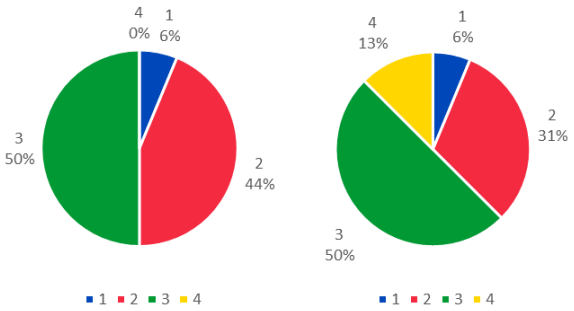
Figure 5.53: Classification results for healthy tissue. For every modality (Perfusion CT or DECT) two results are shown. The parameters on the training set were obtained by cross-validation, while those of test were estimated on the training group.

methods are visualized in Figure 5.54. In addition, the rater compared the quality obtained by the two different modalities, i.e. Perfusion CT and DECT. The quality difference is assessed using a five-step scale, which indicates whether one modality gives much better, better, or equal results. This scale is visualized as a -2 to 2 scale in Figure 5.55. One subject was excluded from this experiment as the corresponding CT images slice contains not enough information to make a valid rating. The used questionnaire is given in Appendix C.

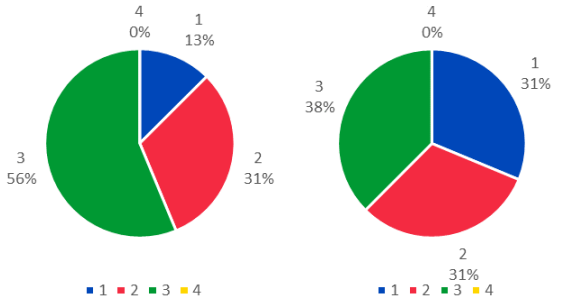
5.6 Summary of Experiments

The experiments reported in the first part of this chapter gives an insight into the influence of different pipeline steps. A clear result is the fact that normalization is necessary. The classification performance is clearly reduced without the normalization, and the choice of normalization algorithm does have a clear impact on the final result. This is evaluated on a training and a test dataset. Similar, the classification algorithm is important, which is also evaluated on a training and test dataset. The influence of both parameter, i.e. normalization and classification strength, is measured by reporting the obtained classification accuracy for different pipelines.

The two proposed methods to deal with data variation are also evaluated. First, the variation in a single dataset is qualitative reported and it is then evaluated if it is possible to estimate the similarity between two images. Based on the previous experiments, the IDAL approach is then evaluated and a performance boost is found, if the number of training images is reduced. Further, the proposed semiautomatic classifier adaptation is evaluated. For this, the number of manual interactions is evaluated against the obtained accuracy and the best obtained accuracy is reported which are both in favour of the proposed

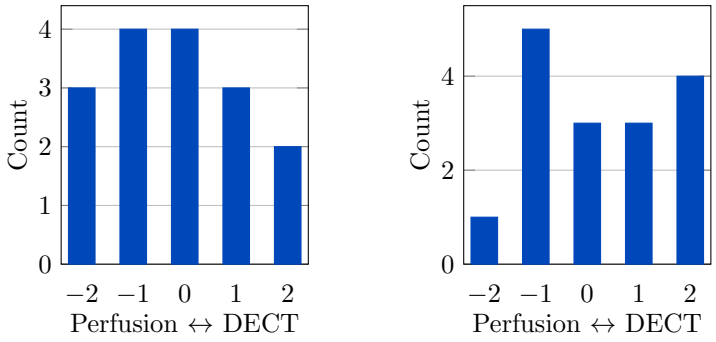


(a) Segmentation (Perfusion) (b) Segmentation (DECT)



(c) Risk map (Perfusion) (d) Risk map (DECT)

Figure 5.54: Result of the grading of the segmentations and risk maps obtained on the two modalities. The grading was from 1 (really good) to 6 (unusable). No segmentation or risk map is graded with 5 or 6.



(a) Comparison of Segmentations (b) Comparison of Risk Maps

Figure 5.55: Comparison of (a) segmentations and (b) risk maps obtained on the two modalities. The sign indicates which result is better. A negative value indicates that the rater voted for the perfusion-based results. The value indicates whether it is slightly (1) or much better (2).

approach.

The three proposed methods to learn from reduced annotations are evaluated. The annotation time is clearly reduced, from four hours to less than five minutes. The hypothesis that a sampling bias is introduced is evaluated and confirmed together with the proposed correction method. It is shown that the results of classifiers which are trained with DALSA or PU-Learning are not significant different from results obtained from traditional trained classifiers. Further, the new proposed LP-Forest algorithm is evaluated using synthetic datasets from the literature and s artificially bagged datasets.

Two approaches for assessing information in different image modalities are evaluated. Using synthetic datasets the proposed method for ROI fusion is analysed. The findings are then again validated on real ROI placements. The use of the DALSA annotation scheme for the brain is evaluated by comparing the information in Perfusion CT and DECT images. The results are evaluated quantitative (comparison of Dice Scores) and qualitative (rater study).

6

Discussion

Within this chapter the findings about each of the proposed methods and hypothesis are put into context and discussed. This is done by referring to the experiments and results reported in the previous chapter as well as reporting results and findings that are published in the literature. This chapter has the same order as the previous chapter as each finding is discussed on its own.

6.1 Classification pipeline

The two evaluated parts of the whole segmentation algorithm does have a high impact on the final performance of the whole algorithm. Both, the normalization algorithm for MRI images and the classification algorithm need to be carefully selected and the right choice can increase the performance of the whole system. The influence of both parts is now discussed in more detail.

6.1.1 Evaluation of MR Normalization

The experiments shows the huge impact of MRI-normalization for the classification results. The preprocessing of the data does have a significant influence on the classification accuracy – especially if the selected features are sensitive to normalization methods. In all the experiments done for this thesis, the results obtained from the classifiers without any normalization were consistently worse than those achieved with some normalization (See Figure 5.7a and Figure 5.9). The use neighbourhood information through an additional image-smoothing does have a lower effect than the normalization and does not reduce effects introduced by the normalization. Contrary, all other methods do benefit from the additional information provided by incorporating neighbourhood information except those without any normalization. I think that the low performance with unnormalized data is caused by the differences between the training and test scans.

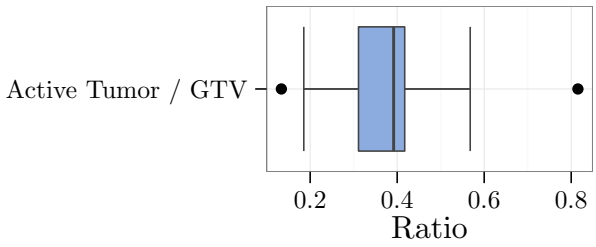


Figure 6.1: Distribution of the ratio of active tumour vs. Gross Tumour Volume (GTV) within the training set of the 2012 BraTS challenge. The ratio is calculated using the provided ground truth.

Rather surprisingly, the simpler normalization methods give better results in the experiments (Figure 5.7a). Using either **Statistic** or **Peak** gives always the best results. While **Polynomial** is still comparable, a normalization using an interval-based methods (like **Percentile**) gives clearly worse results. I think that this is due to the more strict assumptions made by the more complex methods. A major assumptions of the **Percentile**-based normalization is that the distribution of tissue is similar for all patients. While this holds true for healthy patients, it is easily violated if tumorous tissue is present. The size of tumorous tissue can vary significantly between different patients. In the used data-set (DS-3, training) the ratio of tumorous to brain tissue is between 13.3 % and 81.5 % (Figure 6.1). If the used normalization method does enforce a similar histogram shape this means that different tissue types will be assigned the same intensity value. This effect can be seen more clearly for the **Percentile** than the **Polynomial** because the strictness of the later is reduced by using a polynomial function with lower order.

The method that is used to create the necessary reference histogram does not reduce this effect. Both methods that make use of an reference histogram are rather stable with regard to the way in which these are created. Although there are some small changes (Figure 5.8), the obtained results are still very similar and no trend is visible within the data. One methods gives slightly better results for one score while the other performs slightly better for the other. I therefore concluded that the reference methods does only have a small effect on the final results. This seems legit, especially since the reference histograms defines only the shape of the normalized histograms. Therefore, there should be no big differences as long as the reference is reasonable similar to most histograms.

Even though the performance of both simple methods is similar, I prefer **Peak** normalization over **Statistic** normalization. A more detailed analysis of the data shows that the latter is more likely to produce outlier – for the same reason that is causing more complex methods to fail. The mean intensity, that is used in **Statistic**, is more affected by the actual tissue distribution. For example, the mean intensity of FLAIR images will be higher if more edema is present[1]. On the other side, the histogram maximum usually corresponds to an image value that is most frequent in healthy tissue. The method is therefore not affected by the varying ratio of the tumorous tissue and gives a stable reference

[1] Edema is hyper-intense in FLAIR images.

point. A further simplification of the **Peak**-algorithm is possible if the mode is calculated using the histogram and used as peak reference. Although not completely investigated, I found that this gives even more stable results, which is in agreement with the recent findings of Shinohara et al. (2014).

In theory, MRI normalization requires a non-linear transformation to be able to create a complete intensity matching. Nevertheless, the linear transformations seem to be superior, mainly due to the more stable reference points, which seems to have a higher impact. To combine the advantages of both – linear and non-linear transformation – a non-linear transformation using more stable markers might lead to an additional boost in accuracy. Possible markers could be segmented tissue classes like showed by Kleesiek, Biller, et al. (2014) or the incorporation of markers like mode, or minimum.

Bias field correction seems to improve the results regardless of the used normalization algorithm (Figure 5.9). Only if no normalization is used at all, a bias field correction leads to decreased classifier performance for the detection of edema. In all other cases, the results that are obtained using corrected data are similar or better than those obtained without. But since the effect is rather small, and was not found in other datasets as well, I recommend to use these results with caution only. From previous experience with other datasets, it is important to check the influence independently for each dataset as it might decrease the detection accuracy in some cases.

Although the findings presented in this section are validated during later experiments on other datasets as well, the used dataset is a major limitation. Only the BraTS 2012 dataset was available when these experiments were conducted. So I decided to use this dataset to enable a reproduction of the experiments. But the used dataset is rather small, consisting of only 20 training and five test subjects. This limits the use of tests for significance, although it would be possible. An even bigger obstacle is the quality of the training data. The provided ground truth does contain a significant amount of false annotations (Figure 6.2). It is possible that an improved classification result actually decreases the score if an incorrectly labelled area is affected. This made it especially important to verify the findings on the test dataset of the challenge (Figure 5.10).

6.1.2 Evaluation of Classification Algorithms

Using ExtraTrees instead of canonical RDF seems to give an additional boost in classification accuracy, which matches the findings of Geurts, Ernst, and Wehenkel (2006). The segmentation quality for all tissue classes is increased using ExtraTrees (Figure 5.11).

Based on the final segmentation (Figure 5.12) the results obtained by ExtraTrees classifier are less sensitive – the fewer areas are labelled as tumorous. This leads to a significantly lower number of false positives. At the same time, the most tumorous areas are still detected correctly.

These results show that existing algorithm can be easily improved by using slightly modified learning algorithm. As the properties of ExtraTrees and canonical RDF – like integrated feature selection – are the same, a change of learning algorithm does not require additional changes. I was able to change the learning algorithm without any other changes in the pipeline. No other step beside the training algorithm was adapted. Although I havae not investigated it yet, I believe that an even larger improvement is possible, if newer training

(a) (b) (c)

Figure 6.2: Example slices from from subject 15 from the 2012 BraTS challenge. The ground truth annotation of edema is given in yellow, and for active tumour in red. Areas with wrong labels are marked with green circles. **(a)** and **(b)** use $T1_w$ with contrast, **(c)** shows a FLAIR image.

algorithms are used, for example Rotation forests (Rodriguez, Kuncheva, and Alonso, 2006).

The normalization of the MRI images was done using a linear transformation with a fixed mode. This is similar to the **Peak** normalization. It is therefore possible to compare the obtained scores with the previous results obtained during the experiments analysing the normalization algorithms. Compared to the results obtained using a kNN classifier, the ExtraTrees scores are significantly better (Figure 5.10, 5.11). Although obtained on two different datasets, the results from RDF are clearly better. Especially as the modalities and tasks in both datasets are identical. A major reason for this difference might be the better integrated feature selection of RDF algorithm compared to kNN algorithm. Using more features actually reduced the segmentation quality if a kNN classifier was used during the experiments. During this experiments, this effect could not be observed for RDF based algorithms. This supports the current trend to use RDF based algorithms for multispectral tissue characterization.

6.2 Pre- and post-training data selection

Based on the hypothesis that medical images are highly variable two different methods to control the variability of the training data were proposed and evaluated. For both methods it is shown that the reduction, either automatic before the training or afterwards, can increase the quality of the obtained results.

6.2.1 Input Data Adaptive Learning (IDAL)

The main idea behind the IDAL algorithm is to limit the training data by findings well-suited training data for each image individually. This is based on the idea that some images are more similar to each other than others.

Not all training data are well-suited

Calculating the similarity[2] between all images in the training data set reveals that the initial assumption that there are huge differences between medical images, holds true in the given case (Figure 5.13). No image is suited as training image for all images. There are always some test images on which the resulting classifier does not perform well. This shows that it is important to have enough training data and gives an idea about the diversity that must be included in a large classifier if only a single classifier is used. This is also illustrated by Figure 5.14.

It is also interesting to notice that there are only negative outliers. While no image is suitable as training base for all test images, there are some images that are not suitable for training a good classifier for any of the other images. The three worst-performing images (image / subject 8, 16, and 27) are not only not suited as training source but also hardly segmented by the proposed approach at all. A closer inspection of these images revealed that more tissue affected by partial volume was included in the reference segmentation as compared to other images. I suspect therefore that these differences lead to wrong trained classifier.

Classifier similarity can be estimated

The qualitative evaluation of the similarity shows that an estimation of the similarity is possible (cf. Figure 5.15 and Figure 5.16). The selected top three training subjects are usually the ones with higher similarity to the corresponding test subject – although there are some exceptions. These exceptions occur mostly for test subjects that are generally hard to classify, such as subject 16, or subject 22. I assume that these images are either rather special and the corresponding similarity therefore not trained using the previous data or the features used to train the similarity classifiers do not capture the important elements for these images. Nevertheless, the best training subjects are still ranked rather high for these patients, as indicated by the similarity of the ranking matrix (Figure 5.16) and the similarity matrix (Figure 5.13).

The quantitative evaluation confirms the qualitative results. The average rank that is selected per test image is mostly below the expectancy value if the training subjects are randomly selected[3] (Figure 5.17). While there are a few cases, where the estimated mean rank is higher, the reason for this is either the close distance between two similarity values or the affected images are generally difficult to label. As expected, the mean rank becomes more and more similar to the expectancy value if more training images are used as fewer images are left out.

The average similarity of the selected training images usually decreases for more training subjects (Figure 5.18). This indicates that images with higher similarity are usually ranked higher than those with lower similarity. This is still true, even if there is a slight increase in the median, like it is the case between six and seven training patients. As the median is the value of a single patient, it is likely that this increase is caused by only one training subject.

[2] I.e. the ability to train a classifier on image A that can be used to segment image B.
[3] Expectancy value for the rank of random selected training images is 14.

Although my experiments proof that the similarity classifier is working, there is a clearly visible difference between the estimated similarity and the true similarity. This is visible both for the average rank (Figure 5.18) as well as for the average similarity (Figure 5.18) of the selected training patients. I think that is mainly because very basic features were used to describe the whole images. These features might not be sufficient for the description. Since the provided training base is rather small I refrain from evaluating more complex features to avoid over-fitting of the presented results. But I think that it is possible to further improve on the quality of the similarity estimation by more problem-tailored image description features.

Training on similar images improves accuracy

Using only the most similar images to train a specific classifier for each image improves the quality of the obtained segmentation. The quality of the segmentation in all experiments that use a limited number of training subjects is better than the quality that is achieved if the traditional approach, i.e. using all training subjects, is used. This can not only be observed in the training data but also in the validation set.

The achieved improvement depends mainly on the selected number of training subjects. Using too few training subjects increases the chance that important information might not be trained and therefore leads to a drop in the obtained score. Using too many training subjects includes unimportant or wrong training information and also leads to a drop in accuracy. Beside the actual task[4], the quality of the similarity estimation seems to be important. With a better similarity estimation, less training data are needed. Only four subjects are needed for the best result, if the best training subjects are defined using the known training similarity. If the similarity is estimated, this number increases to eight. These additional patients are needed to account for the lesser quality of the training data that are selected. Also, including more patients increases the chance of including the best training patient.

Eight training subjects turned out to be a good number for the training patients in the used training dataset. Even if the used training subjects are determined by other methods than a fixed number of training subjects, the best results are always obtained if around eight patients are used in the mean case. Beside this, there seems to be no advantage in choosing a dynamic selection of the number of training patient. Neither selection all training subject above an given similarity threshold nor limiting the sum of the votes of the training patients led to better results. In contrary, the latter one is even less stable – making the selection of a final parameter more difficult.

Using the true similarity matrix does not give the best result using only a single patient. This is mainly because even the true similarity matrix is still an approximation. To create this similarity matrix, a classifier is trained using the DALSA training scheme and a wDice score calculated using the SUR annotation of the test patient. I did this to indicate the possibility to combine both approaches. But the predictions made on SURs differ from the final prediction, mainly because per definition a SUR is usually placed in areas that are easy to annotate. A further evaluation of the differences to a similarity approach that

[4]Including available modalities, problem formulation, training data, features etc..

is estimated using full annotations for test subjects is given in the appendix, B.

6.2.2 Pre-trained semi-automatic tissue characterization

A new semi-automatic method to improve the results of an automatic segmentation method is introduced. It is shown that the conversion rate improves, if previously trained classifiers are improved instead of training a new classifier from scratch.

Pre-trained classifier can improve conversion rate

A faster convergence to the limit of the classifier is reached if a pre-trained classifier is improved instead of training a completely new classifier. Even after annotating 30 different training points, the new trained classifier does not reach the final classification performance. At the same time, the improvements of an adapted classifier is reached with less than ten manual annotations. The weighting approach allows to correct more segmentations in less time as fewer interactions are needed. In addition, the results are less sensitive to the used training data, it is thus easier to correct images using this method.

It seems that the final classification accuracy does not differ significantly if enough interactions are made. Both methods, new learning and weighting, seem to produce classifier of similar power. This indicates that the limitation of the segmentation accuracy is due to the underlying combination of classification algorithm and extracted features and not depended of the chosen semi-automatic approach. I therefore conclude that both methods could be improved by using different features.

Being able to improve an existing classifier to the same level as a manually trained classifier indicates that the necessary information is already in the training data. The main problem seems to be to select the correct training samples. This correlates with the findings about pre-selected training data (IDAL, section 6.2.1). Implementing the suggested approach might lead to a further improvement of the results and can reduce the need for manual interactions.

Manual data reduction leads to additional improvement

Requiring that all areas segmented as 'lesion' are connected to a point manual labelled as 'lesion' helps to improve the classifier results. The improvement does not depend on the way the final classifier is trained but occurs for both, new training of classifier and weighting of existing classifier.

A major reason for this is the introduction of additional information. There are some tissue types within the brain that do have similar appearance as lesions. While a human is able to identify such areas based on global anatomical knowledge, this is difficult for voxel-based classification systems.

6.3 Methods for Reduced Annotation Effort

A main obstacle for automatic tissue segmentation with learning based algorithms is the amount of necessary training data. Therefore three different methods have been proposed that allow a reduction of the manual annotations that are necessary for the training, following the main idea that the training data

needs only to be partially annotated. The three methods differ mainly in the type of necessary annotation.

6.3.1 Learning from Sparse Annotations

I presented a new approach (DALSA, that allows training of classifiers in automatic tumour segmentation using easy-to-annotate SURs instead of complete segmentations without sacrificing segmentation accuracy. The proposed domain adaptation technique correctly compensates for sampling selection errors and yields results that are comparable to state-of-the-art methods that require tedious full annotations. This alleviates a major obstacle of learning-based methods with regard to clinical applicability and will facilitate the transfer of methods to different clinical domains and settings.

Learning on SUR is Time-efficient

Using SURs saves time during manual creation of the training data. Fewer voxels need to be labelled and the labelling of these voxels is more straightforward, since areas of uncertainty can be avoided. For the experiments these effects add up to an overall reduction of the labelling time by a factor of more than 70.

The use of SURs also reduces the mean training time by a factor greater than 180. The main reason for this – beside the reduction of training data points (voxels) – is the more coherent structure of the data. Since a full segmentation is prone to incorrectly labelled voxels and inconsistent border definitions, the separation of the classes is more difficult. This also explains why lower tree depths perform better if SURs are used.

The usual approach of reducing the training time is to learn only from a randomly drawn subset of all training data. While this approach does not reduce the required labelling time, it does reduce the training time significantly, resulting in times comparable with LSA and DALSA. The times that are report for DALSA include the calculation time of the correction weights. Since this is an independent step, it can be performed separately from the training. This will reduce the overall training time if multiple training runs are performed – for example during parameter optimization or cross validation.

Learning on sampled training data also reduces the prediction time. Although the effect is not as pronounced as it is on the training time, it still takes twice as long to predict an unseen patient using conventional classifiers compared to using DALSA classifiers. This is mainly due to the more coherent training data, which allow the use of trees with a lower tree depth. The decreased prediction time will be especially important for interactive applications.

Learning on SUR Introduces a Selection Bias

Learning on reduced training data results in a drop in the quality of the prediction results. Neither randomly sampled nor sparsely annotated training data yield classifiers of similar quality as the ones trained on complete annotations (Figure 5.27). If the reduction of training data is not done randomly, e.g. when annotating with SURs, a selection bias is introduced. This leads to classifiers which are not optimal for the given problem, as shown in Figure 5.31. A simple correction of this effect by an adapted decision threshold is not possible

for several reasons: 1. The threshold depends on the unknown $P_{\text{Test}}(x)$ and $P_{\text{Train}}(x)$ and the corresponding $P_{\text{Test}}(y)$ and $P_{\text{Train}}(y)$ and therefore on the rater and his selection of SUR (c.f. Figure 5.31). 2. The decision threshold is multi-dimensional for classifiers with more than two classes. 3. Optimal decision thresholds can only be determined for a known gold standard (i.e. fully annotated training data), while the proposed classifier training is bases on SURs only. Figure 5.31 could only be plotted because I had complete segmentations available for validation of the proposed approach.

Domain Adaptation Compensates Selection Bias

The proposed domain adaptation successfully compensates this disproportion of label representations in the training data (Figure 5.31). All the experiments show that the use of domain adaptation increases the Dice score and results in a segmentation quality similar to random sampling at comparable ratios.The experiments with combined SUR sets of different raters show that the level of quality reached by learning from all voxels in the complete annotations can be reached by investing more time into labelling multiple SURs per subject.The DALSA Dice score improves with increasing values of λ, i.e. with a higher influence of the corrective weights (Figure 5.32), further demonstrating the positive effect of DALSA. My experiment with SVMs and the BraTS challenge data (DS-3) supported these finding.

Training on SURs can increase the classifier's sensitivity for tumorous tissue and at the same time result in an increasing amount of false positive decisions (Figure 5.28). This could be caused by the increased tumour-to-tissue ratio in the annotations, as suggested by the finding that domain adaptation lowers the effect when correcting for this ratio. In addition, the classifier's sensitivity is influenced by the training data quality. Due to ambiguities in the data, complete annotations potentially include a higher number of falsely labelled voxels, resulting in less distinctive label classes. Further between-class ambiguities could be caused by healthy tissue voxels that contain inflammation, above-average blood-volume, or chronic stroke and thus have similar appearances to tumorous tissue. The labelling of these voxels would likely be avoided when annotating SURs. The increased amount of false positive decisions that are not connected to the main tumour could likely be reduced by simple post-processing of the results. However, these results could also help identify as yet undiscovered signs of tumorous tissue. The reversed findings in setup III (lower sensitivity and higher positive predictive value for DALSA) are not contradictory, since Kleesiek *et al.* originally sampled the data non-i.i.d., thereby introducing artificial class weights.

The effect of domain adaptation on $\text{LCA}_{\%}$ is only marginal and the absolute improvement is unlikely to be relevant for real applications. The low p-value of the Wilcoxon signed-rank test can be explained by the fact that random sampling is never perfectly i.i.d.. Thus the proposed correction does have a minimal effect on each data point, which then adds up to a high sum of ranks in the test.

DALSA Applicable Under Different Conditions

The experiments demonstrate that DALSA can be easily integrated into existing classification pipelines, such as the one of Kleesiek, Biller, et al. (2014). In cases where a classification algorithm does not offer native support for observation-based weighting, solutions based on classifier ensembles could help and make the proposed approach applicable in combination with virtually any classification algorithm (Zadrozny, Langford, and Abe, 2003). This is important, since previous studies of W. Fan et al. (2005) and the experiments with SVM demonstrate the impact of the sampling bias also on other classifiers. The experiments also show a positive effect of DALSA under varying data and feature sets. Using the pipeline of Kleesiek, Biller, et al. (2014), state-of-the-art performance was achieved on the basis of sparse training data.

The experiments also show that the performance of DALSA does depend on the way SURs had been selected (c.f. Figure 5.33a). An arbitrarily varying placement of SURs produced the lowest quality end result while at the same time, mostly profiting from the proposed domain adaptation. The advantage of this sampling scheme was its time efficiency: On average it took only 50 seconds to annotate a single patient and add the annotation to the training data. The other annotation schemes (Type $2-4$) all yielded similar results in terms of Dice scores while requiring an annotation time between two and three minutes per patient. Type 3 has the additional advantage of not requiring tissue annotations close to tissue borders. Irrespective of the labelling scheme, all SUR-based classifiers could be improved by the use of domain adaptation.

6.3.2 Learning from Only Positive Annotations

I showed that it is possible to train a voxel-based classifier for tissue characterization using only positive annotations by using tissue rates. Using this technique, new use cases are possible that were not possible before.

Tissue annotations can be learned from positive annotations only

Using only positive annotated data in combination with the class ratios of unlabelled data allows to train complete classifier. Using these sparsely annotated data does not lead to a significant decrease in segmentation quality. None of the results based on PU-learning differs significantly from the results that are obtained if DALSA is used to train the classifier (Table 5.8). In contrary, some of the trained classifiers are no longer significantly different from the results obtained by a standard training approach.

On average, more than 40 % of the annotated tissue belongs to the healthy class. Removing the need to annotate this class leads to a considerable reduction of annotation load. This is especially true since most of the brain is considered healthy and a huge amount of variation is found within this class. Picking up the right areas is therefore not always straight forward and might need annotations in different slices of the brain. This advantage becomes even more prominent as even a manual estimation of the tumour volume can be done in less than a minute, which is clearly below the time that is needed for the annotation of the corresponding areas.

Being able to train classifiers even in use-cases where the annotation of a second class is difficult is one of the additional benefits beside the reduced

annotation time. The separation between two tissue classes can not always easily made using solely medical images. Creating a training ground truth that includes labelled samples of all available tissue classes is difficult in such cases. Using PU-learning can help to provide decision support systems for such cases. The tissue risk maps and segmentation created by a classifier can be used to aid the diagnosis and lead to more stable and safe diagnosis.

Tumour ratio estimation is straight forward

Not surprisingly, the best results are achieved if the true class ratio is used. But a full annotation is necessary to be able to calculate these data from the medical images, making the use of PU-learning useless. But for some cases, the actual volume of a tumour might be already known, for example from a post-surgery histological measurement. In this cases, this ratio should be used, assuming that the measurement is accurate.

If the true tumour size is not known it needs to be estimated. All four methods that are evaluated in this work give reasonable results. Since the proposed method is stable against false tissue rates (Figure 5.37). Reasonable results are obtained even with an estimation error of $\pm 20\,\%$. This allows to use fast, but slightly more inaccurate methods in order to reduce manual annotation time without sacrificing too much accuracy.

Using manual estimations of the tumour volume is known to over-estimate the actual tumour volume Galanis et al. (2006). This is also shown in the experiments, where manual volume estimations lead to an overestimation of about $0.6\,\%$ or even $1.2\,\%$ in the complete tumour ratio. But these measurements can be made fast, and are commonly applied during clinical routine Jaffe (2006). Therefore, these measurements might be already available, allowing to use the proposed approach without an additional labelling effort.

Creating the manual tumour volume estimations takes still some time and might not be possible in every use-case. For tissue types that do not appear with circular shapes or at several different spots, a manual estimation might become too time consuming. Especially if multiple images are annotated for the training. Using an algorithmic estimation is a suitable alternative in such cases.

The prediction accuracy of the automatic algorithms is clearly below the manual estimation. But they do not require any additional manual interaction and the results of the obtained classifiers are not significantly different from those of the manual estimation. It is worth noticing that the results reported in Table 5.7 contain a clear outlier, as it can be seen in Figure 5.35. Removing this outlier, that is caused by low contrast within the image, leads to more comparable results.

Batched mode improves classifier quality

Using an individual class prior per image, i.e. the batched mode, seems to improve the estimation quality. While this is obvious if the true ratio is used, the obtained results are better for the global mode if the data are manually annotated. This is mainly due to the fact that the manual annotation overestimates the tumour proportion. The results for global mode further improves (Figure 5.37) if the ratio is overestimated. The reason for this effect is not known. I

suspect that it is mainly due to random effects of the used data. Without this effect, there would be no difference between the global and batched mode for the manual ratio estimation.

The findings correspond well with the findings of Hernández-González, Inza, and Jose A. Lozano (2015) and Hernández-González, Inza, and Jose A Lozano (2016), who showed that using different class prior and positive labelled data only can improve the result on the training data. I therefore suggest to use the presented batched mode instead of the global mode.

6.3.3 Learning from Bag-Wise-Annotations

The proposed training algorithm, Label Proportion Forest (LP-Forest), allows for the first time the training of RDF based on bag-wise annotations. These RDF-based classifiers can be used in the same way as tree-based classifiers that are trained on data that are trained on data that are annotated on observation level.

LP-Forest avoids SVM pitfalls

Using tree-based classifiers prevents some of the typical problems that arise if SVM-based learning algorithms are used in a LLP setting. Due to that, the proposed classification challenges, namely Yu-1, and Patrini-0 to Patrini-16, are solved without problem. The most successful SVM-based algorithms assume a specific mean distribution, like for example $\propto$-SVM (F. Yu et al., 2013). A violation of the underlying assumption, as constructed in Yu-1 can therefore lead to a failure of the algorithm. The proposed learning algorithm is not affected by this assumption, and is passing the corresponding test successfully.

RDF-based classifiers incorporate an internal feature selection process (Ishwaran et al., 2011). This makes the training insensitive to covariates assigned to each bag. There is no clear difference between the different test data sets of Patrini (Patrini-0 to Patrini-16). The trained classifier are all similar in respect to the achieved accuracy. The small differences in the detection accuracy are due to the random generation process of the different test sets. It can also be seen that the decision border depends only on the y-feature if it is not a covariate of the bagging process (Figure 5.42). The absence of the mean assumption allows to train classifier even if the means of both classes are identical, as it is the case with the toy example Middle-1.

LP-Forest inherits RDF properties

Since the resulting classifier cannot be distinguished from a traditionally trained RDF classifier, most of the properties are shared with these classifiers. The learning algorithm inherited the multi-class capability from other RDF based algorithm. It is no problem to train a classifier that distinguishes between different classes as each leaf can be assigned an individual class. I validated this behaviour using the newly introduced Multiclass tests. Here it is shown, that not only the results are mostly correctly classified, but also that the corresponding decision rules align nicely with the underlying class distributions (Figure 5.44b).

Using a learning algorithm that is based on the design principles of other RDF-based classification algorithms, like Random Forests, allows to use well-

established feature selection methods. This can be done either by using the feature importance proposed by Breiman (2001) or other similar measures or even more complex methods like those proposed by Dramiński et al. (2008) which also allow to detect the correlation between features. This allows to gain further insight into the relevance of different features even if it is impossible to label single observations, as long as it is possible to obtain bag-wise class distributions.

LP-Forest is comparable to the State-of-the-art

The discriminative power of the proposed LP-Forests is comparable to other state-of-the-art classifiers (See Table 5.10). While the reference value is better in most cases the differences are small. It must be further considered, that for each dataset the best result based on a RDF algorithm is taken from the literature. Considering the large number of different reported approaches, it is very unlikely to find the best results for all datasets with the given setting. Also a rather conservative approach was chosen to to estimate the classification accuracy – the reported results are therefore in favour of the state-of-the-art. This was done following the arguments of Fernández-Delgado et al. (2014) about multiple testing of different algorithms. The experiments using the brain tumor images further showed that this technique can be used to train a classifier for brain tumour segmentation without pre-labeling data.

LP-Forest benefits from more bags

An observation made during the experiments is that the performance of LP-Forests depends on the number of available bags. Reducing the number of bags available for training leads to a reduced classification accuracy. This can also be seen in Table 5.10 which shows that the obtained accuracy decreases with the bag size. As the number of overall observations is kept constant, this means that the number of bags used during training is reduced. This is clearly visible with the largest bag size of 32, especially for small datasets, like 'satimage'.

This property is not surprising, as a main source of information is the coherence between different bags. If too few bags are available, there are too many options to create a classifier that explains the given data without being correct. This can be best seen if only a single bag is used. In this case, it is impossible to determine the true distributions of two classes without additional information. There are simply too many possible ways to split the data. To resolve this ambiguity, additional information – like more bags – is needed. The shown behaviour is therefore not specific to the LP-Forest algorithm but is shared with all LLP algorithms.

It is difficult to name a rule for the minimum or maximum number of necessary bags. This is highly depended on the data that is used – including the difficulty to separate the classes and the number of features. A possibility to find an upper limit of bags is to increase the number of bags used for training. If the accuracy does not further improve, it can be assumed that enough bags are used.

6.4 Assessment of image acquisition

The selection of the right imaging modality is an important question during medical imaging. Two different approaches have been evaluated for this task. Both do have a different focus, with the first allowing the combination of multiple rater and the second giving a more detailed information about the information content provided from each modality.

6.4.1 Multi-rater region of interest comparison

I proposed a new method for the combination of multiple Regions of Interest (ROIs). An individual weight is estimated for each ROI by estimating the similarity of this ROI to an artificial best ROI. My experiments show that the proposed method allows the estimation of a common ROI from different rater or different images.

Mean fusion is affected by outliers

The common method to fuse different Regions of Interest, i.e. weighting each ROI with the same weights, is sensitive to outliers. A single misplaced ROI can have a severe effect on the final result. An example for this is shown in Figure 5.48. While two out of three ROIs are similar, the final result is heavily affected by the third ROI. It is easy to imaging that the final result would be even more influenced if the third ROI would be even more different from the other two, showing the sensitivity of this method to a single outlier.

The same effect can be observed in in-vito data. Figure 5.49 shows that the resulting distribution of intensity values is affected by single results. Only one result might lead to a significantly different distribution of values. The same is true for Figure 5.50. The distribution of mean values is heavily influenced by 'ROI 1', giving the impression of a relatively low mean intensity value.

Removing outlying results is an option to reduce the effect of such, probably misplaced, areas. But as long as there is no obvious reason to remove data during the evaluation, such an approach can easily lead to biased results. Even if this is not intended, there is always the danger of introducing an artificial bias. Also, removing outliers should be avoided to circumvent the impression of doctored findings.

Virtual ROI fusion is less outlier sensitive

Using the proposed fusion strategy gives a artificial ROI that is less affected by outliers in the training data. This is shown most prominently using synthetic data (Figure 5.48). There, the result of the proposed method is clearly not affected by the single outlier in the data, giving a result that corresponds more closely to the two most-similar ROIs. The mean value and the standard deviation are similar to those of these two. The same observation can be made with the real-life data (Figure 5.49 and Figure 5.50). For both results, the experiments are closer to other approaches. This is also validated by evaluating the mean distances between the ROIs and the created artificial ROIs. For all combinations, more results are similar to the artificial ROI that is created using the proposed method compared to the mean fusion.

The distribution of weights indicates that the proposed method tends to incorporate all provided ROIs. Even though there are a few cases with high weights about 0.4 or more, most final weights are between 0 and 0.3. Considering that the mean weight would be around 0.15 if seven ROIs were contributing the same, it seems that in most cases nearly all provided ROIs contribute to the final results. The proposed method does therefore behave like a 'median' function for non-distributional data in terms of outlier sensitivity but still incorporates all data.

Small overhead

Estimating a virtual ROI that is similar to all others can be done quite fast. The introduced overhead by the proposed method is quite low. As the average data size is relatively small compared to a whole image, the calculations can be done in an efficient manner. Additionally, the whole process converges fast. In my experiments, it needed less then 12 iterations per run until convergence. For this, the whole process can be run in less than a second. This allows to include the proposed method in an existing evaluation pipeline.

Can be combined with other methods

The proposed method can be combined with multiple different approaches. Either using traditional ROI-based comparison of different contrasts or using more advanced methods. Because the final result of this approach is a virtual ROI, there are only few limitations. Most of the approaches that are working on ROIs or similar data structures might use this method.

This is not only true for tissue comparison but might also be used if sparse annotations – like SURs are used as input. The proposed method then allows the combination of different methods. In the use-case of DALSA it is therefore now possible to have multi-rater annotations. This was not possible before, because previous methods for multi-rater fusion, like STAPLE (Warfield, Zou, and W. M. Wells, 2004) do require that the same areas are marked within the original image. While this requirement does not hinder for complete segmentation it is in fact a strict limitation if the rater is free to select the area he wants to annotate.

6.4.2 Classifier-based information assessment

I proposed a new method to estimate the information content of a new contrast and compare the results with other contrasts. The proposed approach is validated by reproducing previous findings from literature.

The proposed method can reproduce previous findings

My experiments showed that Perfusion CT is well-suited for the detection of pancreatic carcinoma. I was able to achieve median wDice for the test and the training group of over 90% for healthy and tumorous tissue. This result is similar to other automatic tumour segmentation approaches. Training classifiers on DECT-images results in classifiers which perform worse on unseen patients. The median wDice is slightly lower than it is for Perfusion CT, both for healthy and tumorous tissue. Nevertheless, the results are clearly above 80% for the both

test groups. These findings align well with previous findings, which indicates that the difference between tumorous and healthy tissue is lower for DECT than it is for Perfusion CT, but still sufficient to discriminate both tissue classes.

These findings correlate with previous findings from literature and the estimation of information content from medical doctors. This indicates that the proposed method allows the estimation of the image content and a clear indication on the information content.

The proposed method allows Computer Supported Diagnosis

An advantage of this method is the fact that a tumour risk map is produced as side-product of the analysis. Since a classifier is trained for every modality, this can then be used to further improve the diagnostic process by fusing multiple contrasts and offering a 'second opinion'. I expect that this reduces the required time for diagnosis and improves the diagnosis quality. It will be especially helpful for training physicians, who can benefit from the more experienced colleagues without requiring their time.

The result of the rating experiments indicates that the resulting segmentation and risk maps are accurate enough to be used for diagnosis. This is especially remarkable as the training data is not annotated for the training of the classifiers but rather for the comparison of different classifiers. These findings corresponds nicely with the findings about DALSA that are reported earlier – there is no need for a complete annotation of the training data to train an automatic tissue characterization scheme.

Some of the image combinations that were used in the experiments are not common within the clinical routine. For example, it is uncommon to look at the mixed images if images from two other energies are available since they contain only little additional information. So the time for looking at these images separately is not reasonable for clinical daily routine. My approach allows to incorporate the information of these images, and presenting them might improve the classifier accuracy since the contained information is easier to access for the machine. Further experiments should evaluate the impact of these images on the obtained classifier accuracy.

The proposed method can detect different textures

The proposed method offers the advantage of comparing two modalities not only with respect to the intensity structure but also to structural appearance. While I used very basic features in the experiments, it is possible to improve more complex texture features. The reasoning for my choice was to allow a better comparison between my findings and those of other state-of-the-art methods – there are no technical reasons to limit the used features.

Incorporating more complex features, for example 'Local Binary Pattern' (Ojala, Pietikainen, and Maenpaa, 2002), will not only render the comparison more meaningful but will also improve the final results of the indicated classifier. This can also be achieved by improving the used classifier, for example by incorporating the proposed IDAL training scheme. But at the same time, it will be more challenging to validate the developed method by comparing it to previous findings.

7

Conclusion

Learning-based tissue characterization can improve the clinical decision making process by making the important information quickly accessible. But creating the necessary training dataset is challenging – due to the time-consuming and error prone process. This hinders the application of such algorithms in a clinical setting. High variability of imaging modalities and across different diseases are further adding to this challenge. Other approaches for tissue characterization are maily focusing on other challenges and make use of small or publicly available datasets during the development, which reduces the adaptability to other use-cases or data variations.

For this thesis, two different approaches to deal with data variability were evaluated. First, the variation within the training dataset and the capability of the classification algorithm were evaluated. I showed that the right selection of preprocessing steps can increase the classification accuracy by making the complete pipeline more robust and reducing the variation within a dataset. Choosing the right normalization can make a significant difference and is a crucial step in each application. As a second approach, two algorithms were proposed that allow an adaptation of the classifier to the current image that is classified. Both, pre- and post-training adaptation increase the classification accuracy clearly, with an improvement up to 33 %.

Furthermore, different methods were developed to reduce the annotation effort by proposing learning methods that can make use of partially or weakly annotated data. Using only sparse and unambiguously annotated regions (SURs), the annotation effort for a single patient can be reduced from more than four hours to less than five minutes. This reduction by a factor of more than 70 can be achieved without sacrificing accuracy. Even further reductions of the annotation effort can be achieved by using patient wise-class ratios that are commonly available or might be estimated from the data. I showed that, by using these ratios, it is possible to either avoid the annotation of a single class or even avoid all voxel annotations at all. Since these data might be already available in the clinical setting, such systems might be trained without any additional labelling effort.

All proposed approaches are flexible – they are not limited to a specific classification algorithm or fixed set of features. This makes the proposed methods potentially applicable to a wide range of different applications and allows incorporating the proposed methods in existing and new workflows. This is also true for the training using only bag-wise annotations. Here, the classification algorithm is an important part of the method. I therefore proposed the first decision tree based approach, bringing this popular classifier class with all advantages associated to it within this problem setting. While this prevents changing the classification algorithm, it still allows using the proposed algorithm with different features and different learning tasks.

This thesis describes the concept, design, and evaluation of these methods. It shows that the proposed solutions are well-suited to solve the challenge of data variability and annotation.

7.1 Summary of contributions

The main focus of this thesis is the reduction of variability within training data and the reduction of annotation time. Most contributions are therefore within the field of machine learning and preprocessing of radiologic images. All proposed methods are applied to the field of tissue characterization from medical images, solving the described challenges in this area. A short list of the main contributions is given below:

- Evaluation of different algorithms in respect to variability reduction and development of algorithms for the adaption to variable appearances:
  - Evaluation of MRI normalization algorithms with respect of the influence on classification accuracy for brain tumour segmentation.
  - Evaluation of different classification algorithms for brain tumour segmentation on a publicly available challenge dataset.
  - Development of an algorithm for patient-specific classification training to reduce the training variability.
  - Development of an algorithm for semi-automatic improvement of segmentations based on classifier correction.
- Development and evaluation of methods for training from weakly supervised training data to reduce the annotation effort:
  - The introduction of a domain-adaptation based learning algorithm for sparse and unambiguous annotations that reduces the annotation time.
  - A learning scheme for learning from class ratios and sparse, positive annotations that allows for the first time to train a standard classifier for tissue characterization without annotating both classes.
  - Development of a new, general classification algorithm for learning from bag-wise annotations. By proposing the first RDF-based algorithm for this problem, the properties of these classifiers are introduced to these problems. Allowing now to train decision trees using only the ratio of classes in different patients.

- Application of the proposed techniques to extract global characterizations:
  - Development and evaluation of a new method to fuse different ROIs, which allows a more reliable comparison of different characteristics.
  - Using sparse annotations and classification algorithms to enable ROI-based comparisons that incorporate not only contrast but also texture information.

7.2 Future work

The proposed methods can be used to reduce the annotation effort and uncertainties. While they have been evaluated and developed for the task of tissue characterization, they can also be used in other domains. It should be evaluated if the proposed methods are also beneficial within different fields. This might further increase the impact of the proposed methods.

Deep convolutional networks have become more and more important for computer vision and also medical image analysis. With the huge amount of training data that are needed for those methods, the initial challenge becomes even more important. Therefore, it will be interesting to combine the proposed methods and findings with Deep Convolutional Networks, allowing to benefit from their increased classification performance while still being able to reduce the necessary training time.

The proposed methods allows now to train and apply fully automatic segmentation algorithms for medical images from weakly annotations. This makes it feasible to annotate large datasets within short time. This allows to make even more advanced predictions and extract higher level information – for example using the now popular Radiomics approach. Without doubt, this is an important field of research that should be evaluated further.

Bibliography

Harold P Adams et al. "Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST". In: *Stroke* 24.1 (1993), pp. 35–41.

Hugo JWL Aerts et al. "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach". In: *Nature communications* 5 (2014).

Tariq Javid Ali, Pervez Akhtar, M. Iqbal Bhatti, and M. Abdul Muqeet. "Significance of Region of Interest Applied on MRI and CT Images in Teleradiology-Telemedicine". In: *Medical Imaging and Informatics: 2nd International Conference, MIMI 2007, Beijing, China.* Ed. by Xiaohong Gao et al. Springer, 2008, pp. 151–159. DOI: 10.1007/978-3-540-79490-5_20.

Mahdi Alizadeh et al. "Intensity inhomogeneity correction in clinical pediatric spinal cord MRI images". In: *2015 41st Annual Northeast Biomedical Engineering Conference (NEBEC).* IEEE. 2015, pp. 1–2.

Yali Amit and Donald Geman. "Shape quantization and recognition with randomized trees". In: *Neural Computation* 9.7 (Oct. 1997), pp. 1545–1588. DOI: 10.1162/neco.1997.9.7.1545.

Dimitris Ampeliotis et al. "A computer-aided system for the detection of prostate cancer based on magnetic resonance image analysis". In: *Communications, Control and Signal Processing, 2008. ISCCSP 2008. 3rd International Symposium on.* IEEE, 2008, pp. 1372–1377.

Roberto. Andreani, Jose M. Martinez, and Maria L. Schuverdt. "On the Relation between Constant Positive Linear Dependence Condition and Quasinormality Constraint Qualification". In: *Journal of Optimization Theory and Applications* 125.2 (2005), pp. 473–483. DOI: 10.1007/s10957-004-1861-9.

Elsa D. Angelini et al. “Glioma dynamics and computational models: a review of segmentation, registration, and in silico growth algorithms and their clinical applications”. In: *Current Medical Imaging Reviews* 3.4 (2007), pp. 262–276.

Mohammad Ali Balafar. “Review of intensity inhomogeneity correction methods for brain MRI Images”. In: *International Journal on Technical and Physical Problems of Engineering* 4.4 (2012), pp. 60–66.

Peter J Basser and Carlo Pierpaoli. “Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI”. In: *Journal of magnetic resonance* 213.2 (2011), pp. 560–570.

Stefan Bauer, Thomas Fejes, and Mauricio Reyes. “A skull-stripping filter for ITK”. In: *Insight Journal* (2012).

Stefan Bauer, Roland Wiest, Lutz-P Nolte, and Mauricio Reyes. “A survey of MRI-based medical image analysis for brain tumor studies”. In: *Physics in medicine and biology* 58.13 (2013).

Christian Beaulieu. “The biological basis of diffusion anisotropy”. In: *Diffusion MRI: From quantitative measurement to in vivo neuroanatomy* (2009), pp. 105–126.

Matt A Bernstein, Kevin F King, and Xiaohong Joe Zhou. *Handbook of MRI pulse sequences*. Elsevier, 2004.

BfS. *Bundesamt f”ur Strahlenschutz*. `https://www.bfs.de/DE/themen/ion/anwendung-medizin/diagnostik/roentgen/haeufigkeit-exposition.html`. Accessed; 15.08.2016. 2016.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. “Discriminative learning for differing training and test distributions”. In: *Proceedings of the 24th international conference on Machine learning*. ACM. 2007, pp. 81–88.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. “A training algorithm for optimal margin classifiers”. In: *Proceedings of the fifth annual workshop on Computational learning theory*. ACM. 1992, pp. 144–152.

Adam C Braithwaite, Brian M Dale, Daniel T Boll, and Elmar M Merkle. “Short-and Midterm Reproducibility of Apparent Diffusion Coefficient Measurements at 3.0-T Diffusion-weighted Imaging of the Abdomen 1”. In: *Radiology* 250.2 (2009), pp. 459–465.

BraTS. *Multimodal Brain Tumor Segmentation Challenge*. `http://braintumorsegmentation.org`. Accessed; 15.06.2016. 2016.

Leo Breiman. “Bagging predictors”. In: *Machine learning* 24.2 (1996), pp. 123–140.

Leo Breiman. “Random Forests”. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984. ISBN: 9780412048418.

Andrew J. Buckler et al. “A Collaborative Enterprise for Multi-Stakeholder Participation in the Advancement of Quantitative Imaging”. In: *Radiology* 258.3 (2011), pp. 906–914. DOI: `10.1148/radiol.10100799`.

Andrew J. Buckler et al. "Quantitative Imaging Test Approval and Biomarker Qualification: Interrelated but Distinct Activities". In: *Radiology* 259.3 (2011), pp. 875–884. DOI: 10.1148/radiol.10100800.

Luigi Cerulo, Charles Elkan, and Michele Ceccarelli. "Learning gene regulatory networks from only positive and unlabeled data". In: *Bmc Bioinformatics* 11.1 (2010), p. 1.

Ian Chan et al. "Detection of prostate cancer by integration of line-scan diffusion, T2-mapping and T2-weighted magnetic resonance imaging; a multichannel statistical classifier". en. In: *Medical Physics* 30.9 (2003), p. 2390. DOI: 10.1118/1.1593633.

Bee-Chung Chen, Lei Chen, Raghu Ramakrishnan, and David R Musicant. "Learning from aggregate views". In: *22nd International Conference on Data Engineering (ICDE'06)*. IEEE. 2006, pp. 3–3.

Shuo Chen, Bin Liu, Mingjie Qian, and Changshui Zhang. "Kernel k-Means based framework for aggregate outputs classification". In: *2009 IEEE International Conference on Data Mining Workshops*. IEEE. 2009, pp. 356–361.

Matthew C Clark et al. "Automatic tumor segmentation using knowledge-based techniques". In: *IEEE transactions on medical imaging* 17.2 (1998), pp. 187–201.

LP Clarke et al. "MRI segmentation: methods and applications". In: *Magnetic resonance imaging* 13.3 (1995), pp. 343–368.

Fergus V Coakley et al. "Pancreatic imaging mimics: part 1, imaging mimics of pancreatic adenocarcinoma". In: *American Journal of Roentgenology* 199.2 (2012), pp. 301–308.

Guylaine Collewet, Michal Strzelecki, and Francois Mariette. "Influence of MRI acquisition protocols and image intensity normalization methods on texture classification". In: *Magnetic Resonance Imaging* 22.1 (2004), pp. 81–91. DOI: 10.1016/j.mri.2003.09.001.

Jason J. Corso et al. "Efficient multilevel brain tumor segmentation with integrated bayesian model classification". In: *Medical Imaging, IEEE Transactions on* 27.5 (2008), pp. 629–640.

Corinna Cortes, Larry D Jackel, and Wan-Ping Chiang. "Limits on Learning Machine Accuracy Imposed by Data Quality". In: *Advances in Neural Information Processing Systems*. 1995, pp. 239–246.

Corinna Cortes and Vladimir N Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

Antonio Criminisi, Jamie Shotton, and Ender Konukoglu. "Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning". In: *Microsoft Research Cambridge, Tech. Rep. MSRTR-2011-114* 5.6 (2011), p. 12.

Annarita D'Addabbo, Nicola Ancona, Palma N. Blonda, and Roberto A. De Blasi. "Detection of multiple sclerosis lesions in MRIs with an SVM classifier". In: vol. 5032. 2003, pp. 1367–1374. DOI: 10.1117/12.480407.

Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. IEEE. 2005, pp. 886–893.

Matthew S Davenport et al. "Inter-and intra-rater reproducibility of quantitative dynamic contrast enhanced MRI using TWIST perfusion data in a uterine fibroid model". In: *Journal of Magnetic Resonance Imaging* 38.2 (2013), pp. 329–335.

Francesco De Comité, François Denis, Rémi Gilleron, and Fabien Letouzey. "Positive and unlabeled examples help learning". In: *International Conference on Algorithmic Learning Theory*. Springer. 1999, pp. 219–230.

Matthew A Deeley et al. "Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study". In: *Physics in medicine and biology* 56.14 (2011), p. 4557.

Wankai Deng, Wei Xiao, He Deng, and Jianguo Liu. "MRI brain tumor segmentation with region growing method based on the gradients and variances along and inside of the boundary curve". In: *2010 3rd International Conference on Biomedical Engineering and Informatics*. Vol. 1. 2010, pp. 393–396. DOI: 10.1109/BMEI.2010.5639536.

François Denis. "PAC learning from positive statistical queries". In: *International Conference on Algorithmic Learning Theory*. Springer. 1998, pp. 112–126.

François Denis, Remi Gilleron, and Marc Tommasi. "Text classification from positive and unlabeled examples". In: *Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU'02*. 2002, pp. 1927–1934.

Lee R Dice. "Measures of the amount of ecologic association between species". In: *Ecology* 26.3 (1945), pp. 297–302.

Annette J Dobson and Adrian Barnett. *An introduction to generalized linear models*. CRC press, 2002. ISBN: 978-1584889502.

Olaf Dössel. "Bildgebende Verfahren in der Medizin". In: *Von der Technik zur medizinischen Anwendung: Springer Verlag* (2000).

Michał Dramiński et al. "Monte Carlo feature selection for supervised classification". In: *Bioinformatics* 24.1 (2008), pp. 110–117.

Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 1999. ISBN: 978-0-471-05669-0.

Cynthia Dwork and Aaron Roth. *The algorithmic foundations of differential privacy*. Vol. 9. 3-4. 2014. ISBN: 978-1601988188.

Cynthia Dwrok. *Invited Talk: Privacy in the Land of Plenty*. NIPS 2014. 2014.

Lucas Elijovich, Pratik V Patel, and J Claude Hemphill. "Intracerebral hemorrhage". In: *Seminars in neurology*. Vol. 28. 05. Thieme Medical Publishers. 2008, pp. 657–667.

Charles Elkan. "The foundations of cost-sensitive learning". In: *International joint conference on artificial intelligence* (2001).

Charles Elkan and Keith Noto. "Learning classifiers from only positive and unlabeled data". In: *14th ACM SIGKDD int. conf. on Knowledge discovery and data mining* (2008).

Kai Fan et al. "Learning a generative classifier from label proportions". In: *Neurocomputing* 139 (2014), pp. 47–55.

Wei Fan, Ian Davidson, Bianca Zadrozny, and Philip S Yu. "An improved categorization of classifier's sensitivity on sample selection bias". In: *Fifth IEEE International Conference on Data Mining (ICDM'05)*. IEEE. 2005, 4–pp.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. "Do we need hundreds of classifiers to solve real world classification problems". In: *J. Mach. Learn. Res* 15.1 (2014), pp. 3133–3181.

Martijn Froeling, Pim Pullens, and Alexander Leemans. "DTI Analysis Methods: Region of Interest Analysis". In: *Diffusion Tensor Imaging: A Practical Handbook*. Ed. by Wim Van Hecke, Louise Emsell, and Stefan Sunaert. Springer New York, 2016, pp. 175–182. ISBN: 978-1-4939-3118-7. DOI: 10.1007/978-1-4939-3118-7_9.

Evanthia Galanis et al. "Validation of neuroradiologic response assessment in gliomas: Measurement by RECIST, two-dimensional, computer-assisted tumor area, and computer-assisted tumor volume methods". In: *Neuro-Oncology* 8.2 (2006), pp. 156–165. DOI: 10.1215/15228517-2005-005.

Juan M. García-Gómez et al. "Multiproject–multicenter evaluation of automatic brain tumor classification by magnetic resonance spectroscopy". In: *Magnetic Resonance Materials in Physics, Biology and Medicine* 22.1 (2008), pp. 5–18. DOI: 10.1007/s10334-008-0146-y.

Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. "Variable selection using random forests". In: *Pattern Recognition Letters* 31.14 (2010), pp. 2225–2236.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees". In: *Machine learning* 63.1 (2006), pp. 3–42.

Vicky Goh et al. "Quantitative Assessment of Colorectal Cancer Tumor Vascular Parameters by Using Perfusion CT: Influence of Tumor Region of Interest 1". In: *Radiology* 247.3 (2008), pp. 726–732.

Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. "Domain adaptation for object recognition: An unsupervised approach". In: *2011 international conference on computer vision*. IEEE. 2011, pp. 999–1006.

Dirk J Gouma et al. "Rates of complications and death after pancreaticoduodenectomy: risk factors and the impact of hospital volume". In: *Annals of surgery* 232.6 (2000), pp. 786–795.

Kevin A Hallgren. "Computing inter-rater reliability for observational data: an overview and tutorial". In: *Tutorials in quantitative methods for psychology* 8.1 (2012), p. 23.

Robert M Haralick, Karthikeyan Shanmugam, et al. "Textural features for image classification". In: *IEEE Transactions on systems, man, and cybernetics* 6 (1973), pp. 610–621.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Vol. 10. Springer series in statistics Springer, Berlin, 2013. ISBN: 978-0-387-84858-7.

Mohammad Havaei, Axel Davy, et al. "Brain tumor segmentation with deep neural networks". In: *Medical Image Analysis* (2016).

Mohammad Havaei, Pierre-Marc Jodoin, and Hugo Larochelle. "Efficient Interactive Brain Tumor Segmentation as Within-Brain kNN Classification." In: *International Conference on Pattern Recognition, ICPR*. 2014, pp. 556–561.

James Hays and Alexei A Efros. "IM2GPS: estimating geographic information from a single image". In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE. 2008, pp. 1–8.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015).

James J Heckman. "Sample selection bias as a specification error (with an application to the estimation of labor supply functions)". In: *Technical Report* (1977).

James J. Heckman. "Sample selection bias as a specification error". In: *Econometrica: Journal of the econometric society* (1979), pp. 153–161.

Tobias Heimann, Peter Mountney, Matthias John, and Razvan Ionasec. "Real-time ultrasound transducer localization in fluoroscopy images by transfer learning from synthetic training data". In: *Medical image analysis* 18.8 (2014), pp. 1320–1328.

Pierre Hellier. "Consistent intensity correction of MR images". In: *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*. Vol. 1. IEEE. 2003, pp. I–1109.

Jerónimo Hernández and Iñaki Inza. "Learning naive Bayes models for multiple-instance learning with label proportions". In: *Conference of the Spanish Association for Artificial Intelligence*. Springer. 2011, pp. 134–144.

Maria del C Valdés Hernández et al. "On the computational assessment of white matter hyperintensity progression: difficulties in method selection and bias field correction performance on images with significant white matter pathology". In: *Neuroradiology* (2016), pp. 1–11.

Jerónimo Hernández-González, Iñaki Inza, and Jose A Lozano. "Learning Bayesian network classifiers from label proportions". In: *Pattern Recognition* 46.12 (2013), pp. 3425–3440.

Jerónimo Hernández-González, Iñaki Inza, and Jose A. Lozano. "A Novel Weakly Supervised Problem: Learning from Positive-Unlabeled Proportions". In: *Advances in Artificial Intelligence: 16th Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2015 Albacete, Spain, November 9–12, 2015 Proceedings*. Springer International Publishing, 2015, pp. 3–13. DOI: 10.1007/978-3-319-24598-0_1.

Jerónimo Hernández-González, Iñaki Inza, and Jose A Lozano. "Learning from Proportions of Positive and Unlabeled Examples". In: *International Journal of Intelligent Systems* (2016).

Tobias Heye et al. "Reproducibility of dynamic contrast-enhanced MR imaging. Part II. Comparison of intra-and interobserver variability with manual region of interest placement versus semiautomatic lesion segmentation and histogram analysis". In: *Radiology* 266.3 (2013), pp. 812–821.

Tin Kam Ho. "Random decision forests". In: *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*. Vol. 1. IEEE. 1995, pp. 278–282.

Tin Kam Ho. "The random subspace method for constructing decision forests". In: *IEEE transactions on pattern analysis and machine intelligence* 20.8 (1998), pp. 832–844.

Jiang Hsieh. *Computed tomography: principles, design, artifacts, and recent advances*. 2009. ISBN: 9780819475336.

Jiayuan Huang et al. "Correcting sample selection bias by unlabeled data". In: *Advances in neural information processing systems* 19 (2007), p. 601.

Hemant Ishwaran, Udaya B Kogalur, Xi Chen, and Andy J Minn. "Random survival forests for high-dimensional data". In: *Statistical analysis and data mining* 4.1 (2011), pp. 115–132.

Isles. *Ischemic Stroke Lesion Segmentation*. `http://www.isles-challenge.org`. Accessed; 15.06.2016. 2016.

Michael A Jacobs et al. "Benign and malignant Breast lesions: diagnosis with multiparametric MR imaging 1". In: *Radiology* 229.1 (2003), pp. 225–232.

C Carl Jaffe. "Measures of response: RECIST, WHO, and new alternatives". In: *Journal of Clinical Oncology* 24.20 (2006), pp. 3245–3251.

Thorsten RC Johnson. "Dual-energy CT: general principles". In: *American Journal of Roentgenology* 199.5 (2012), S3–S8.

Thorsten Johnson, Christian Fink, Stefan O Schönberg, and Maximilian F Reiser. *Dual energy CT in clinical practice*. Springer Science & Business Media, 2011.

Michael R Kaus et al. "Automated segmentation of MR images of brain tumors 1". In: *Radiology* 218.2 (2001), pp. 586–591.

Brad Keller et al. "Adaptive multi-cluster fuzzy C-means segmentation of breast parenchymal tissue in digital mammography". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2011, pp. 562–569.

Miriam Klauss, Andreas Lemke, et al. "Intravoxel incoherent motion MRI for the differentiation between mass forming chronic pancreatitis and pancreatic carcinoma". In: *Investigative radiology* 46.1 (2011), pp. 57–63.

Miriam Klauss, Wolfram Stiller, et al. "Computed tomography perfusion analysis of pancreatic carcinoma". In: *Journal of computer assisted tomography* 36.2 (2012), pp. 237–242.

Jens Kleesiek, Armin Biller, et al. “Ilastik for multi-modal brain tumor segmentation”. In: *Proceedings MICCAI BraTS (Brain Tumor Segmentation Challenge)* (2014), pp. 12–17.

Jens Kleesiek, Gregor Urban, et al. “Deep MRI brain extraction: a 3D convolutional neural network for skull stripping”. In: *NeuroImage* 129 (2016), pp. 460–469.

Ender Konukoglu. *Personal website of Ender Konukoglu*. `http://www.nmr.mgh.harvard.edu/~enderk/software.html`. Accessed; 15.08.2016. 2016.

Ender Konukoglu, Ben Glocker, Darko Zikic, and Antonio Criminisi. “Neighbourhood approximation forests”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2012, pp. 75–82.

Ender Konukoglu, Ben Glocker, Darko Zikic, and Antonio Criminisi. “Neighbourhood approximation using randomized forests”. In: *Medical image analysis* 17.7 (2013), pp. 790–804.

Nikolaus Kriegeskorte, W Kyle Simmons, Patrick SF Bellgowan, and Chris I Baker. “Circular analysis in systems neuroscience: the dangers of double dipping”. In: *Nature neuroscience* 12.5 (2009), pp. 535–540.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

Hendrik Kuck and Nando de Freitas. “Learning about individuals from group statistics”. In: *Proceedings of the Twenty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*. Arlington, Virginia: AUAI Press, 2005, pp. 332–339.

Brenda F. Kurland et al. “Promise and pitfalls of quantitative imaging in oncology clinical trials”. In: *Magnetic Resonance Imaging* 30.9 (2012). Quantitative Imaging in Cancer, pp. 1301–1312. DOI: `10.1016/j.mri.2012.06.009`.

Dongjin Kwon et al. “PORTR: Pre-operative and post-recurrence brain tumor registration”. In: *IEEE transactions on medical imaging* 33.3 (2014), pp. 651–667.

Frederic Lachmann and Christian Barillot. “Brain tissue classification from MRI data by means of texture analysis”. In: *Medical Imaging VI*. International Society for Optics and Photonics. 1992, pp. 72–83.

David H Laidlaw, Kurt W Fleischer, and Alan H Barr. “Partial-volume Bayesian classification of material mixtures in MR volume data using voxel histograms”. In: *IEEE transactions on medical imaging* 17.1 (1998), pp. 74–86.

Doenja MJ Lambregts et al. “Tumour ADC measurements in rectal cancer: effect of ROI methods on ADC values and interobserver variability”. In: *European radiology* 21.12 (2011), pp. 2567–2574.

Frederik B Laun, Klaus H Fritzsche, Tristian A Kuder, and Bram Stieltjes. “Introduction to the basic principles and techniques of diffusion-weighted imaging”. In: *Der Radiologe* 51.3 (2011), pp. 170–179.

Wee Sun Lee and Bing Liu. "Learning with positive and unlabeled examples using weighted logistic regression". In: *ICML*. Vol. 3. 2003, pp. 448–455.

Fabien Letouzey, François Denis, and Rémi Gilleron. "Learning from positive and unlabeled examples". In: *International Conference on Algorithmic Learning Theory*. Springer. 2000, pp. 71–85.

Rongjian Li et al. "Deep learning based imaging data completion for improved brain disease diagnosis". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2014, pp. 305–312.

Wenkai Li, Qinghua Guo, and Charles Elkan. "A positive and unlabeled learning algorithm for one-class classification of remote-sensing data". In: *IEEE Transactions on Geoscience and Remote Sensing* 49.2 (2011), pp. 717–725.

Xiaoli Li and Bing Liu. "Learning to classify texts using positive and unlabeled data". In: *IJCAI*. Vol. 3. 2003, pp. 587–592.

Chunquan Liang, Yang Zhang, Peng Shi, and Zhengguo Hu. "Learning very fast decision tree from uncertain data streams with positive and unlabeled samples". In: *Information Sciences* 213 (2012), pp. 50–67.

Laura Liberman et al. "The breast imaging reporting and data system: positive predictive value of mammographic features and final assessment categories." In: *AJR. American journal of roentgenology* 171.1 (1998), pp. 35–40.

Issel Anne L. Lim et al. "Human brain atlas for automated region of interest selection in quantitative susceptibility mapping: Application to determine iron content in deep gray matter structures". In: *NeuroImage* 82 (2013), pp. 449–469.

Bing Liu, Wee Sun Lee, Philip S Yu, and Xiaoli Li. "Partially supervised classification of text documents". In: *ICML*. Vol. 2. 2002, pp. 387–394.

Ce Liu, Jenny Yuen, and Antonio Torralba. "Nonparametric scene parsing via label transfer". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.12 (2011), pp. 2368–2382.

Ce Liu, Jenny Yuen, and Antonio Torralba. "Sift flow: Dense correspondence across scenes and its applications". In: *IEEE transactions on pattern analysis and machine intelligence* 33.5 (2011), pp. 978–994.

Xin Liu and Imam Samil Yetik. "Automated prostate cancer localization without the need for peripheral zone extraction using multiparametric MRI". In: *Medical physics* 38.6 (2011), pp. 2986–2994.

Xavier Lladó et al. "Segmentation of multiple sclerosis lesions in brain MRI: a review of automated approaches". In: *Information Sciences* 186.1 (2012), pp. 164–185.

Melissa A LoPresti et al. "Hematoma volume as the major determinant of outcomes after intracerebral hemorrhage". In: *Journal of the neurological sciences* 345.1 (2014), pp. 3–7.

David N Louis et al. "The 2007 WHO classification of tumours of the central nervous system". In: *Acta neuropathologica* 114.2 (2007), pp. 97–109.

David G Lowe. "Object recognition from local scale-invariant features". In: *Computer vision, 1999. The proceedings of the seventh IEEE international conference on.* Vol. 2. Ieee. 1999, pp. 1150–1157.

Oskar Maier et al. "Extra Tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences". In: *Journal of neuroscience methods* 240 (2015), pp. 89–100.

Lena Maier-Hein, Sven Mersmann, Daniel Kondermann, et al. "Can Masses of Non-Experts Train Highly Accurate Image Classifiers?" In: *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2014, pp. 438–445.

Larry M Manevitz and Malik Yousef. "One-class SVMs for document classification". In: *Journal of Machine Learning Research* 2.Dec (2001), pp. 139–154.

Anna Margolis. "A literature review of domain adaptation with unlabeled data". In: *Tec. Report* (2011), pp. 1–42.

Katharina Marten et al. "Inadequacy of manual measurements compared to automated CT volumetry in assessment of treatment response of pulmonary metastases using RECIST criteria". In: *European radiology* 16.4 (2006), pp. 781–790.

John E Mayhew and John P Frisby. "Texture discrimination and Fourier analysis in human vision." In: *Nature* 275 (1978), pp. 438–439.

Gloria P Mazzara et al. "Brain tumor target volume determination for radiation treatment planning through automated MRI segmentation". In: *International Journal of Radiation Oncology* Biology* Physics* 59.1 (2004), pp. 300–312.

Medline Trend. `http://dan.corlan.net/medline-trend.html`. Accessed; 15.06.2016, Search Term: ("tissues"[MeSH Terms] OR "tissues"[All Fields] OR "tissue"[All Fields] OR ("tumour"[All Fields] OR "neoplasms"[MeSH Terms] OR "neoplasms"[All Fields] OR "tumor"[All Fields])) AND (characterization[All Fields] OR "classification"[Subheading] OR "classification"[All Fields] OR "classification"[MeSH Terms] OR Segmentation[All Fields]) AND (multimodal[All Fields] OR multispectral[All Fields] OR multicontrast[All Fields]). 2016.

Raphael Meier et al. "A hybrid model for multimodal brain tumor segmentation". In: *Proceedings of NCI-MICCAI BRATS* 1 (2013), pp. 31–37.

Raphael Meier et al. "Patient-specific semi-supervised learning for postoperative brain tumor segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer. 2014, pp. 714–721.

Bjoern H Menze, Andras Jakab, et al. "The multimodal brain tumor image segmentation benchmark (BRATS)". In: *IEEE Transactions on Medical Imaging* 34.10 (2015), pp. 1993–2024.

Bjoern H Menze, Koen Van Leemput, et al. "A generative model for brain tumor segmentation in multi-modal images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer. 2010, pp. 151–159.

Bjoern Menze, Andras Jakab, et al. *Proceedings of the MICCAI Challenge on Multimodal Brain Tumor Image Segmentation (BRATS) 2012*. 2012.

Bjoern Menze, Mauricio Reyes, Keyvan Farahani, and Jayashree Kalpathy-Cramer. *Proceedings of the MICCAI Challenge on Multimodal Brain Tumor Image Segmentation (BRATS) 2014*. 2014.

Bjoern Menze, Mauricio Reyes, Keyvan Farahani, Jayashree Kalpathy-Cramer, and Dongjin Kwon. *Proceedings of the MICCAI Challenge on Multimodal Brain Tumor Image Segmentation (BRATS) 2015*. 2015.

Bjoern Menze, Mauricio Reyes, Andras Jakab, et al. *Proceedings of the MICCAI Challenge on Multimodal Brain Tumor Image Segmentation (BRATS) 2013*. 2013.

Kenneth A Miles and MR Griffiths. "Perfusion CT: a worthwhile enhancement?" In: *The British journal of radiology* 76 (2003), pp. 220–231.

Anthony B. Miller, Barth Hoogstraten, Maurice Staquet, and Ashley E. Winkler. "Reporting results of cancer treatment". In: *Cancer* 47.1 (1981), pp. 207–214. DOI: 10.1002/1097-0142(19810101)47.

UCL MLR. *UCL Machine Learning Repository*. `https://archive.ics.uci.edu/ml/datasets.html`. Accessed; 15.06.2016. 2016.

David R Musicant, Janara M Christensen, and Jamie F Olson. "Supervised learning by training on aggregate outputs". In: *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. IEEE. 2007, pp. 252–261.

Neat. *Ischemic Stroke Lesion Segmentation*. `http://neatbrains15.isi.uu.nl/`. Accessed; 15.06.2016. 2016.

Aljaž Noe and James C Gee. "Partial volume segmentation of cerebral MRI scans with mixture model clustering". In: *Biennial International Conference on Information Processing in Medical Imaging*. Springer. 2001, pp. 423–430.

Marco Nolden et al. "The medical imaging interaction toolkit: challenges and advances". In: *International journal of computer assisted radiology and surgery* 8.4 (2013), pp. 607–620.

Mohammad Norouzi, Maxwell D Collins, David J Fleet, and Pushmeet Kohli. "Co2 forest: Improved random forest by continuous optimization of oblique splits". In: *arXiv preprint arXiv:1506.06155* (2015).

Mohammad Norouzi, Maxwell Collins, et al. "Efficient non-greedy optimization of decision trees". In: *Advances in Neural Information Processing Systems*. 2015, pp. 1720–1728.

László G Nyúl, Jayaram K Udupa, and Xuan Zhang. "New variants of a method of MRI scale standardization". In: *Medical Imaging, IEEE Transactions on* 19.2 (2000), pp. 143–150.

Timo Ojala, Matti Pietikainen, and Topi Maenpaa. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns". In: *IEEE Transactions on pattern analysis and machine intelligence* 24.7 (2002), pp. 971–987.

Annegreet van Opbroek, M Arfan Ikram, Meike W Vernooij, and Marleen De Bruijne. "Transfer learning improves supervised image segmentation across imaging protocols". In: *IEEE transactions on medical imaging* 34.5 (2015), pp. 1018–1030.

Annegreet van Opbroek, Meike W Vernooij, M Arfan Ikram, and Marleen de Bruijne. "Weighting training images by maximizing distribution similarity for supervised segmentation across scanners". In: *Medical image analysis* 24.1 (2015), pp. 245–254.

Andrés Ortiz et al. "Automatic ROI Selection in Structural Brain MRI Using SOM 3D Projection". In: *PLoS ONE* 9.4 (Apr. 2014), pp. 1–12. DOI: 10.1371/journal.pone.0093851.

Sedat Ozer et al. "Supervised and unsupervised methods for prostate cancer segmentation with multispectral MRI". en. In: *Medical Physics* 37.4 (2010), p. 1873. DOI: 10.1118/1.3359459.

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. "Domain adaptation via transfer component analysis". In: *IEEE Transactions on Neural Networks* 22.2 (2011), pp. 199–210.

Sinno Jialin Pan and Qiang Yang. "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10 (2010), pp. 1345–1359.

Sarah Parisot et al. "Concurrent tumor segmentation and registration with uncertainty-based sparse non-uniform graphs". In: *Medical image analysis* 18.4 (2014), pp. 647–659.

Ofer Pasternak et al. "Free water elimination and mapping from diffusion MRI". In: *Magnetic Resonance in Medicine* 62.3 (2009), pp. 717–730.

Clifford S Patlak, Ronald G Blasberg, and Joseph D Fenstermacher. "Graphical evaluation of blood-to-brain transfer constants from multiple-time uptake data". In: *Journal of Cerebral Blood Flow & Metabolism* 3.1 (1983), pp. 1–7.

Giorgio Patrini, Richard Nock, Tiberio Caetano, and Paul Rivera. "(Almost) No Label No Cry". In: *Advances in Neural Information Processing Systems*. 2014, pp. 190–198.

Valentina Pedoia et al. "Manual labeling strategy for ground truth estimation in MRI glial tumor segmentation". In: *Proceedings of the 1st International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications*. ACM. 2012, p. 8.

Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (Nov. 2011), pp. 2825–2830.

Tao Peng, Wanli Zuo, and Fengling He. "SVM based adaptive learning method for text classification from positive and unlabeled documents". In: *Knowledge and Information Systems* 16.3 (2008), pp. 281–301.

Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. "Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images". In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1240–1251.

Marthinus C du Plessis, Gang Niu, and Masashi Sugiyama. "Analysis of Learning from Positive and Unlabeled Data". In: *Advances in Neural Information Processing Systems 27* (2014).

Marthinus Christoffel du Plessis and Masashi Sugiyama. "Class Prior Estimation from Positive and Unlabeled Data". In: *IEICE TRANSACTIONS on Information and Systems* (2014).

Leonid Portnoy, Eleazar Eskin, and Sal Stolfo. "Intrusion detection with unlabeled data using clustering". In: *In Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001*. Citeseer. 2001.

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

Marcel Prastawa, Elizabeth Bullitt, and Guido Gerig. "Simulation of brain tumors in MR images for evaluation of segmentation efficacy". In: *Medical image analysis* 13.2 (2009), pp. 297–311.

Marcel Prastawa, Elizabeth Bullitt, Sean Ho, and Guido Gerig. "A brain tumor segmentation framework based on outlier detection". In: *Medical image analysis* 8.3 (2004), pp. 275–283.

Yanjun Qi. "Random forest for bioinformatics". In: *Ensemble machine learning*. Springer, 2012, pp. 307–323.

Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. "Estimating labels from label proportions". In: *Journal of Machine Learning Research* 10.Oct (2009), pp. 2349–2374.

J. Ross Quinlan. "Induction of decision trees". In: *Machine learning* 1.1 (1986), pp. 81–106.

J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014. ISBN: 978-1558602380.

Johann Radon. "Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten". In: *Akad. Wiss.* 69 (1917), pp. 262–277.

Islem Rekik, Stéphanie Allassonnière, Trevor K Carpenter, and Joanna M Wardlaw. "Medical image analysis methods in MR/CT-imaged acute-subacute ischemic stroke lesion: Segmentation, prediction and insights into dynamic evolution simulation models. A critical appraisal". In: *NeuroImage: Clinical* 1.1 (2012), pp. 164–178.

International Agency for Research on Cancer. *World Cancer Report 2014 (International Agency for Research on Cancer)*. World Health Organization, 2014. ISBN: 9283204298.

Juan José Rodriguez, Ludmila I Kuncheva, and Carlos J Alonso. "Rotation forest: A new classifier ensemble method". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 28.10 (2006), pp. 1619–1630.

Su Ruan, Stéphane Lebonvallet, Abderrahim Merabet, and Jean-Marc Constans. "Tumor segmentation from a multispectral MRI images by using support vector machine classification". In: *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on*. IEEE. 2007, pp. 1236–1239.

Stefan Rueping. "SVM classifier estimation from group probabilities". In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 2010, pp. 911–918.

Bryan Russell et al. "Object recognition by scene alignment". In: *Advances in Neural Information Processing Systems*. 2007, pp. 1241–1248.

DV Sahani. *Perfusion CT: an overview of technique and clinical applications*. 2012. URL: http://cds.ismrm.org/protected/10MProceedings/files/Tues%20E09_02%20Sahani.pdf.

Mark Schmidt et al. "Segmenting brain tumors using alignment-based features". In: *Fourth International Conference on Machine Learning and Applications (ICMLA'05)*. IEEE. 2005, 6–pp.

Mohak Shah et al. "Evaluating intensity normalization on MRIs of human brain with multiple sclerosis". In: *Medical Image Analysis* 15.2 (2011), pp. 267–282. DOI: 10.1016/j.media.2010.12.003.

Katie Sharkey and Richard Beare. *Histogram-based Thresholding*. https://blog.kitware.com/histogram-based-thresholding. Accessed; 15.06.2016. 2016.

Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. "Get another label? improving data quality and data mining using multiple, noisy labelers". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2008, pp. 614–622.

Hidetoshi Shimodaira. "Improving predictive inference under covariate shift by weighting the log-likelihood function". In: *Journal of statistical planning and inference* 90.2 (Oct. 2000), pp. 227–244. DOI: 10.1016/S0378-3758(00)00115-4.

Russell T. Shinohara et al. "Statistical normalization techniques for magnetic resonance imaging". In: *NeuroImage: Clinical* 6 (2014), pp. 9–19. DOI: 10.1016/j.nicl.2014.08.008.

Dirk Simon et al. "Diffusion-weighted imaging-based probabilistic segmentation of high-and low-proliferative areas in high-grade gliomas". In: *Cancer Imaging* 12.1 (2012), p. 89.

Phil Simon. *Too Big to Ignore: The Business Case for Big Data*. Vol. 72. John Wiley & Sons, 2013. ISBN: 978-1118638170.

Stephan Skornitzke et al. "Qualitative and quantitative evaluation of rigid and deformable motion correction algorithms using dual-energy CT images in view of application to CT perfusion measurements in abdominal organs affected by breathing motion". In: *The British journal of radiology* 88.1046 (2015), p. 20140683.

John G Sled, Alex P Zijdenbos, and Alan C Evans. "A nonparametric method for automatic correction of intensity nonuniformity in MRI data". In: *Medical Imaging, IEEE Transactions on* 17.1 (1998), pp. 87–97.

Stephen M Smith et al. "Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data". In: *Neuroimage* 31.4 (2006), pp. 1487–1505.

Jeffrey Solomon, John A Butman, and Arun Sood. "Segmentation of brain tumors in 4D MR images using the hidden Markov model". In: *Computer methods and programs in biomedicine* 84.2 (2006), pp. 76–85.

Christoph Sommer, Christoph Straehle, Ullrich Köthe, and Fred A. Hamprecht. "Ilastik: Interactive learning and segmentation toolkit". In: *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. Mar. 2011, pp. 230–233. DOI: 10.1109/ISBI.2011.5872394.

Bram Stieltjes et al. "Diffusion tensor imaging in primary brain tumors: Reproducible quantitative analysis of corpus callosum infiltration and contralateral involvement using a probabilistic mixture model". In: *NeuroImage* 31.2 (2006), pp. 531–542.

Wolfram Stiller et al. "Correlation of Quantitative Dual-Energy Computed Tomography Iodine Maps and Abdominal Computed Tomography Perfusion Measurements: Are Single-Acquisition Dual-Energy Computed Tomography Iodine Maps More Than a Reduced-Dose Surrogate of Conventional Computed Tomography Perfusion?" In: *Investigative radiology* 50.10 (2015), pp. 703–708.

Marco Stolpe and Katharina Morik. "Learning from label proportions by optimizing cluster model selection". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2011, pp. 349–364.

Roger Stupp et al. "Changing paradigms—an update on the multidisciplinary management of malignant glioma". In: *The Oncologist* 11.2 (2006), pp. 165–180.

Nagesh K Subbanna, Doina Precup, D Louis Collins, and Tal Arbel. "Hierarchical probabilistic Gabor and MRF segmentation of brain tumours in MRI volumes". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2013, 751–758.

Nagesh Subbanna, Doina Precup, and Tal Arbel. "Iterative multilevel MRF leveraging context and voxel information for brain tumour segmentation in MRI". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 400–405.

Masashi Sugiyama and Motoaki Kawanabe. "Machine learning in non-stationary environments: introduction to covariate shift adaptation". In: MIT Press, 2012. ISBN: 9780262017091.

Xiaofei Sun et al. "Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions". In: *Biomedical engineering online* 14.1 (2015), p. 1.

Chikako Suzuki et al. "Radiologic Measurements of Tumor Response to Treatment: Practical Approaches and Limitations". In: *RadioGraphics* 28.2 (2008), pp. 329–344. DOI: 10.1148/rg.282075068.

GuangJian Tian, Yong Xia, Yanning Zhang, and Dagan Feng. "Hybrid genetic and variational expectation-maximization algorithm for Gaussian-mixture-model-based brain MR image segmentation". In: *IEEE transactions on information technology in biomedicine* 15.3 (2011), pp. 373–380.

Joseph Tighe and Svetlana Lazebnik. "Superparsing". In: *International Journal of Computer Vision* 101.2 (2013), pp. 329–349.

Nicholas J Tustison et al. "N4ITK: improved N3 bias correction". In: *Medical Imaging, IEEE Transactions on* 29.6 (2010), pp. 1310–1320.

Nicholas J. Tustison et al. "Optimal Symmetric Multimodal Templates and Concatenated Random Forests for Supervised Brain Tumor Segmentation (Simplified) with ANTsR". In: *Neuroinformatics* 13.2 (2015), pp. 209–225. DOI: 10.1007/s12021-014-9245-2.

Michael W Vannier et al. "Multispectral analysis of magnetic resonance images." In: *Radiology* 154.1 (1985), pp. 221–224.

Vladimir N. Vapnik. *Statistical learning theory*. Vol. 2. Wiley New York, 1998. ISBN: 978-0-471-03003-4.

Vladimir N Vapnik and Alexey Chervonenkis. "A note on one class of perceptrons". In: *Automation and remote control* 25.1 (1964).

Various. *Wikipedia, Human Brain*. https://en.wikipedia.org/wiki/Human_brain. Accessed; 15.08.2016. 2016.

Various. *Wikipedia, Pancreatic Cancer*. https://en.wikipedia.org/wiki/Pancreatic_cancer. Accessed; 15.08.2016. 2016.

Ragini Verma et al. "Multiparametric tissue characterization of brain neoplasms and their recurrence using pattern classification of MR images". In: *Academic radiology* 15.8 (2008), pp. 966–977.

Uro Vovk, Franjo Pernus, and Botjan Likar. "A Review of Methods for Correction of Intensity Inhomogeneity in MRI". In: *IEEE Transactions on Medical Imaging* 26.3 (2007), pp. 405–421. DOI: 10.1109/TMI.2006.891486.

Liqun Wang et al. "Correction for variations in MRI scanner sensitivity in brain studies with histogram matching". In: *Magnetic Resonance in Medicine* 39.2 (1998), pp. 322–327.

Shijun Wang and Ronald M Summers. "Machine learning and radiology". In: *Medical image analysis* 16.5 (2012), pp. 933–951.

Zhijun Wang et al. "Multimodality 3D Tumor Segmentation in HCC Patients Treated with TACE". en. In: *Academic Radiology* 22.7 (July 2015), pp. 840–845. DOI: 10.1016/j.acra.2015.03.001.

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

Christian Weber et al. "Tumor Progression Mapping: An Intuitive Visualization of Glioblastoma Progression in MR Follow-ups". In: *Proc. Annual Meeting ISMRM, Mailand* (2014).

Website. *Grand Challenges in Biomedical Image Analysis*. http://grand-challenge.org/. Accessed; 15.06.2016. 2016.

Jeffrey C Weinreb et al. "PI-RADS prostate imaging–reporting and data system: 2015, version 2". In: *European urology* 69.1 (2016), pp. 16–40.

William Wells et al. "Multi-modal volume registration by maximization of mutual information". In: *Medical image analysis* 1.1 (1996), pp. 35–51.

Caroline Weltens et al. "Interobserver variations in gross tumor volume delineation of brain tumors on computed tomography and impact of magnetic resonance imaging". In: *Radiotherapy and Oncology* 60.1 (2001), pp. 49–59.

Ivo Wolf et al. "The medical imaging interaction toolkit". In: *Medical image analysis* 9.6 (2005), pp. 594–604.

Min-Jie Yang et al. "Common and unusual CT and MRI manifestations of pancreatic adenocarcinoma: a pictorial review". In: *Quantitative imaging in medicine and surgery* 3.2 (2013), p. 113.

Xulei Yang, Qing Song, and Yue Wang. "A weighted support vector machine for data classification". In: *International Journal of Pattern Recognition and Artificial Intelligence* 21.05 (Aug. 2007), pp. 961–976. DOI: 10.1142/S0218001407005703.

Tang-Kai Yin, Bi-Fang Lee, Yen Kuang Yang, and Nan-Tsing Chiu. "Differences of Various Region-of-Interest Methods for Measuring Dopamine Transporter Availability Using-TRODAT-1 SPECT". In: *The Scientific World Journal* 2014 (2014).

Felix Yu et al. "$\propto$SVM for Learning with Label Proportions". In: *Proceedings of The 30th International Conference on Machine Learning*. 2013, pp. 504–512.

Hailong Yu, Wanli Zuo, and Tao Peng. "A new PU learning algorithm for text classification". In: *Mexican International Conference on Artificial Intelligence*. Springer. 2005, pp. 824–832.

Hwanjo Yu, Jiawei Han, and Kevin Chen-Chuan Chang. "PEBL: positive example based learning for web page classification using SVM". In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2002, pp. 239–248.

Bianca Zadrozny, John Langford, and Naoki Abe. "Cost-sensitive learning by cost-proportionate example weighting". In: *Third IEEE International Conference on Data Mining, 2003. ICDM 2003.* IEEE, 2003, pp. 435–442.

Le Zhang, Ye Ren, and Ponnuthurai N Suganthan. "Instance based random forest with rotated feature space". In: *Computational Intelligence and Ensemble Learning (CIEL), 2013 IEEE Symposium on*. IEEE. 2013, pp. 31–35.

Liang Zhao, Wei Wu, and Jason J. Corso. "Semi-automatic Brain Tumor Segmentation by Constrained MRFs Using Structural Trajectories". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Ed. by Kensaku Mori et al. Springer Berlin Heidelberg, 2013, pp. 567–575. DOI: 10.1007/978-3-642-40760.

Darko Zikic, Ben Glocker, and Antonio Criminisi. "Atlas encoding by randomized forests for efficient label propagation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2013, pp. 66–73.

Darko Zikic, Ben Glocker, Ender Konukoglu, et al. "Decision Forests for Tissue-Specific Segmentation of High-Grade Gliomas in Multi-channel MR". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2012, pp. 369–376. DOI: 10.1007/978-3-642-33454.

List of Figures

List of Tables

Publication List

Part of the content of this thesis, like some ideas, descriptions, figures, and tables are previously published in the following publications:

Peer reviewed Journals

1. Sebastian Bickelhaupt, et al., **Michael Götz**, and et al. "Prediction of malignancy by a radiomic signature from contrast agent-free diffusion MRI in suspicious breast lesions found on screening mammography." In: *Journal of Magnetic Resonance Imaging* ((accepted)).
2. Franciszek Binczyk, et al., **Michael Götz**, and et al. "MiMSeg - an algorithm for automated detection of tumor tissue on NMR apparent diffusion coefficient maps." In: *Information Sciences* (2016). DOI: 10.1016/j.ins.2016.07.052.
3. **Michael Götz**, Christian Weber, Franciszek Binczyk, et al. "DALSA: Domain Adaptation for Supervised Learning from Sparsely Annotated MR Images". In: *IEEE Transactions on Medical Imaging* 35 (Jan. 2016), pp. 184–196. DOI: 10.1109/TMI.2015.2463078.
4. Philipp Kickingereder, et al., **Michael Götz**, and et al. "Radiomic Profiling of Glioblastoma: Identifying an Imaging Predictor of Patient Survival with Improved Performance over Established Clinical and Radiologic Risk Models". In: *Radiology* (), p. 160845. DOI: 10.1148/radiol.2016160845.
5. Philipp Kickingereder, **Michael Götz**, et al. "Large-scale radiomic profiling of recurrent glioblastoma identifies an imaging predictor for stratifying anti-angiogenic treatment response". In: *Clinical Cancer Research* (submitted).
6. Oskar Maier, et al., **Michael Götz**, and et al. "ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI". In: *Medical Image Analysis* (2016). DOI: 10.1016/j.media.2016.07.009.

7. Patrick Philipp, et al., **Michael Götz**, and et al. "Toward cognitive pipelines of medical assistance algorithms". In: *International Journal of Computer Assisted Radiology and Surgery* 1 (2015), pp. 1–11. DOI: 10.1007/s11548-015-1322-y.

8. Moritz Scherer, et al., Michael Götz, and et al. "Development and Validation of an Automatic Segmentation Algorithm for Quantification of Intracerebral Hemorrhage". In: *Stroke* (accepted).

Peer reviewed Book Chapter

9. **Michael Götz**, Christian Weber, Christoph Kolb, and Klaus Maier-Hein. "Input Data Adaptive Learning (IDAL) for Sub-acute Ischemic Stroke Lesion Segmentation". In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: First International Workshop, Brainles 2015, Held in Conjunction with MICCAI 2015, Munich, Germany, October 5, 2015, Revised Selected Papers*. Ed. by Alessandro Crimi et al. Springer International Publishing, 2016, pp. 284–295. ISBN: 978-3-319-30858-6. DOI: 10.1007/978-3-319-30858-6_25.

Peer reviewed Conferences and Workshops

10. David Bonekamp, et al., **Michael Götz**, and et al. "Prostata-MRT vor radikaler Prostatektomie: Radiomics-Parameter zur Differenzierung von primärem Gleason Pattern 3 von hochgradigem Prostatakarzinom". In: *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*. Vol. 188. S 01. 2016, SP303_2.

11. Philipp Gemmeke, et al., **Michael Götz**, and et al. "Using Linked Data and Web APIs for Automating the Pre-processing of Medical Images". In: *Fifth International Workshop on Consuming Linked Data (COLD2014)*. Riva del Garda, Italy, 2014.

12. **Michael Götz**, Eric Heim, et al. "A learning-based, fully automatic liver tumor segmentation pipeline based on sparsely annotated training data". In: *SPIE Medical Imaging*. International Society for Optics and Photonics. 2016, pp. 97841I–97841I.

13. **Michael Götz**, Christoph Kolb, et al. "Fallspezifisches Lernen zur automatischen Läsionssegmentierung in multimodalen MR-Bildern". In: *Bildverarbeitung für die Medizin 2016*. Springer, 2016, pp. 62–67.

14. **Michael Götz**, Stephan Skornitzke, et al. "Machine-learning based comparison of CT-perfusion maps and dual energy CT for pancreatic tumor detection". In: *SPIE Medical Imaging*. International Society for Optics and Photonics. 2016, 97851R–97851R.

15. **Michael Götz**, Christian Weber, F. Binczyck, et al. "Automatische Tumorsegmentierung mit spärlich annotierter Lernbasis". In: *Bildverarbeitung für die Medizin 2015*. Informatik aktuell. Springer Berlin Heidelberg, 2015, pp. 486–491.

16. **Michael Götz**, Christian Weber, Josiah Blöcher, et al. "Extremely randomized trees based brain tumor segmentation". In: *Proceedings of MICCAI 2014 Brain Tumor Segmentation Challenge*. Boston, USA, 2014.

17. **Michael Götz**, Christian Weber, and Klaus Maier-Hein. "Input Data Adaptive Learning (IDAL) for sub-acute Ischemic Stroke Lesion Segmentation". In: *Proceedings of MICCAI Workshop on Ischemic Stroke Lesion Segmentation 2015*. 2015, pp. 39–42. URL: `http://www.isles-challenge.org/pdf/20150930_ISLES2015_Proceedings.pdf`.

18. **Michael Götz**, Christian Weber, and Klaus H. Maier-Hein. "Similarity based fusion of multiple Regions of Interests for MR sequence evaluation". In: *Proceedings of International Society of Magnetic Resonance in Medicine*. 2016.

19. **Michael Götz**, Christian Weber, Bram Stieltjes, et al. "Learning from small amounts of labeled data in a brain tumor classification task". In: *Second Workshop on Transfer and Multi-Task Learning: Theory meets Practice, Montreal, Canada*. 2014.

20. Philipp Kickingereder, **Michael Götz**, and et al. "Large-scale radiomic profiling of glioblastoma identifies an imaging signature for predicting and stratifying antiangiogenic treatment response". In: *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*. Vol. 188. S 01. 2016, WISS301_1.

21. Philipp Kickingereder, **Michael Götz**, John Muschelli, et al. "Large-scale radiomic profiling of glioblastoma identifies an imaging signature for predicting and stratifying antiangiogenic treatment response". In: *Proceedings of International Society of Magnetic Resonance in Medicine*. 2016.

22. Lena Maier-Hein, et al., **Michael Götz**, and et al. "Crowd-Algorithm Collaboration for Large-Scale Endoscopic Image Annotation with Confidence". In: *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2016, accepted.

23. Peter F. Neher, **Michael Götz**, and et al. "A machine learning based approach to fiber tractography". In: *Proceedings of International Society of Magnetic Resonance in Medicine*. 2015.

24. Peter F. Neher, **Michael Götz**, and et al. "A machine learning based approach to fiber tractography using classifier voting". In: *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2015.

25. Patrick Philipp, et al., **Michael Götz**, and et al. "Automatisierung von Vorverabeitungsschritten für medizinische Bilddaten mit semantischen Technologien". In: *Bildverarbeitung für die Medizin 2015*. Informatik aktuell. Springer Berlin Heidelberg, 2015.

26. Moritz Scherer, et al., **Michael Götz**, and et al. "Introduction of a fully automatic segmentation algorithm in brain computed tomography . Prerequisite for analysis of intracranial volume proportions in the context of spontaneous intracerebral hemorrhage (ICH)". In: *66th Annual Meeting of the German Society of Neurosurgery (DGNC)*. 2015. DOI: `10.3205/15dgnc533`.

27. Christian Weber, **Michael Götz**, and et al. "Brain Tumor Progression Modeling - A Data Driven Approach". In: *Proceedings of Image-Guided Adaptive Radiation Therapy MICCAI Workshop 2014*. 2014, pp. 71–78.

28. Christian Weber, **Michael Götz**, and et al. "Überwachtes Lernen zur Prädiktion von Tumorwachstum". In: *Bildverarbeitung für die Medizin 2015*. Informatik aktuell. Springer Berlin Heidelberg, 2015, pp. 473–478.

29. Frank M Weber, et al., **Michael Götz**, and et al. "Analysis of mitral valve motion in 4d transesophageal echocardiography for transcatheter aortic valve implantation". In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer. 2014, pp. 168–176.

Other (Talks, Poster etc..)

30. Franciszek Binczyk, et al., **Michael Götz**, and et al. *A filtration of cerebrospinal fluid signal based on Gaussian mixture model decomposition of magnetic resonance diffusion weighted imaging data*. Poster presented at 8th Symposium of the Polish Bioinformatics Society, 17-19 September, Lublin, Poland. 2015.

31. Franciszek Binczyk, et al., **Michael Götz**, and et al. *Determination of high-grade brain tumours internal structure based on magnetic resonance diffusion imaging and signal decomposition to Gaussian mixture model*. Poster presented at 19th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2015) April, Warsaw, Poland. 2015.

32. Josiah Blöcher. *Multimodale Segmentierung von Gliomen mit Gaussian-Mixture-Models und k-Nearest-Neighbor*. Master Thesis, supervised by **Michael Götz**. Feb. 2014.

33. Jonas Cordes. *Segmentation of intracerebral hemorrhages in computer tomography images using random forest*. Master Thesis, supervised by **Michael Götz**. Sept. 2015.

34. **Michael Götz**. *Cognition-Guided Radiology*. Invited Talk at the CURAC Annual Conference. Sept. 2015.

35. **Michael Götz**. *Data-driven oncologic image analysis*. Invited Talk at the Gliwice Scientific Meeting 2015. Nov. 2015.

36. Alexander Tschlatscher. *Integration von LIBSVM in MITK und Test anhand einer automatischen Lebertumorsegmentierung*. Bachelor Thesis, supervised by **Michael Götz**. Feb. 2016.

Acknowledgement

Diese Dissertation entstand während meiner Zeit am Deutschen Krebsforschungszentrum (DKFZ) in der Abteilung Medizinische und Biologische Informatik, bzw. der Juniorgruppe Medical Image Computing. Finanziert wurde sie durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen des Projekts I04 – Wissensbasierte Gewebecharakterisierung des Sonderforschungsbereiches SFB 125 – Cognition-Guided Surgery.

An dieser Stelle möchte ich mich herzlich bei meinem Gruppenleiter, Klaus Maier-Hein, für die Förderung und das entgegengebracht Vertrauen bedanken. Meine Arbeit hat sehr von seiner Unterstützung und dem bereitgestellten, optimalen Umfeld profitiert. Auch bei Hans-Peter Meinzer möchte ich mich für seine Unterstützung und seinen Einsatz für uns Doktoranden in seiner Zeit als Abteilungsleiter bedanken. Herzlichen Dank auch an Rüdiger Dillmann für die Betreuung meiner Doktorarbeit am Karlsruher Institut für Technologie (KIT).

Danken möchte ich auch meinen Kooperationspartnern für die gute und fruchtbare Zusammenarbeit. Im Einzelnen: Bram Stieltjes, David Bonekamp, Philipp Kickingereder, Moritz Scherer und Miriam Klauss für das Bereitstellen das medizinischen Know-Hows und die Einblicke in die klinische Praxis, Franciszek Binczyk, Joanna Polańska und Raphael Tarnawski für die enge Kooperation, meinen Partnern aus dem SFB, u.a. Philipp Mayer, Franziska Fritz, Stephan Skornitzke, Patrick Philipp und allen andere, die viele spannende Projekte erst ermöglicht haben. Danke auch an Josiah, Jonas und Jonas, Alex, Sebastian, Jasmin und Timothy, die Arbeit mit euch hat wirklich Spaß gemacht..

Besondere Dank geht an die gesamte MBI-Abteilung für das konstruktive und tolle Arbeitsklima. Besonders hervorheben möchte ich hier Christian für das gemeinsame Erleben aller Höhen und Tiefen einer Promotion aber auch die restlichen Mitglieder der Gruppen H838 und S2123: Es waren tolle gemeinsame Jahre und ich habe sehr vom Teamgeist und der guten Atmosphäre profitiert.

Ein ganz besonderer Dank gilt meiner Familie für die Unterstützung in den letzten Jahren. Und last, but not least geht eine Danke an meine Frau, Doro, für den großen Beitrag zum Gelingen dieser Arbeit in den letzten Jahren. Ich freue mich auf unser neues Abenteuer.

Appendices

A

Proofs

A.1 Proof: Sum over weights equals c times n

According to Equation 4.8, the sum over all estimated weights for the SURs of an image can be written as

$$\sum_{n_{\text{Train}}} \hat{w}(x) = \sum_{n_{\text{Train}}} c \cdot \frac{1 - \hat{p}\,(z = 1 \mid x)}{\hat{p}\,(z = 1 \mid x)} \tag{A.1a}$$

$$= c \cdot n_{\text{Train}} \cdot \mathbb{E}\left[\frac{1 - \hat{p}\,(z = 1 \mid x)}{\hat{p}\,(z = 1 \mid x)}\right] \tag{A.1b}$$

$$= c \cdot n_{\text{Train}} \cdot \frac{1 - \mathbb{E}\left[\hat{p}\,(z = 1 \mid x)\right]}{\mathbb{E}\left[\hat{p}\,(z = 1 \mid x)\right]}. \tag{A.1c}$$

$$= c \cdot n_{\text{Train}} \cdot \frac{1 - \frac{n_{\text{Train}}}{n_{\text{Train}} + n_{\text{Test}}}}{\frac{n_{\text{Train}}}{n_{\text{Train}} + n_{\text{Test}}}} \tag{A.1d}$$

$$= c \cdot n_{\text{Train}} \cdot \frac{n_{\text{Train}} + n_{\text{Test}} - n_{\text{Train}}}{n_{\text{Train}}} \tag{A.1e}$$

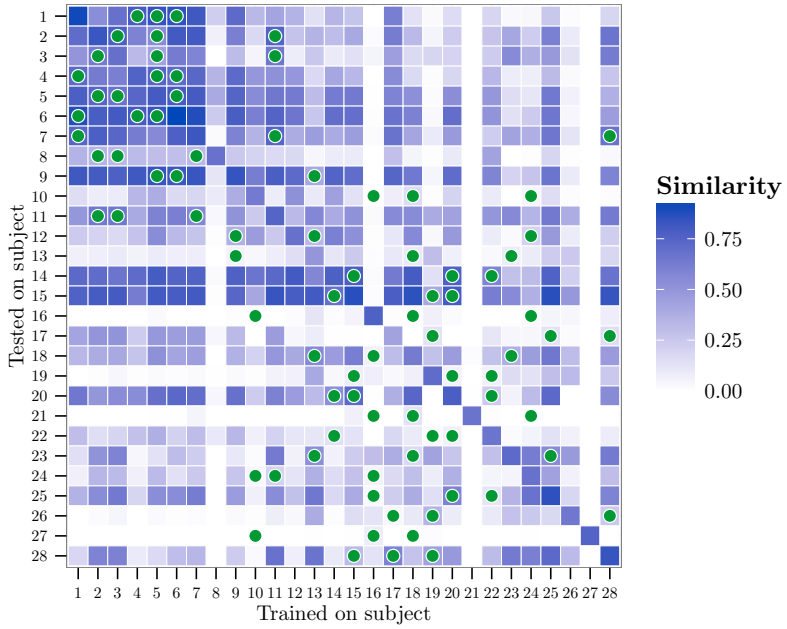
$$= c \cdot n_{\text{Test}} \tag{A.1f}$$

B

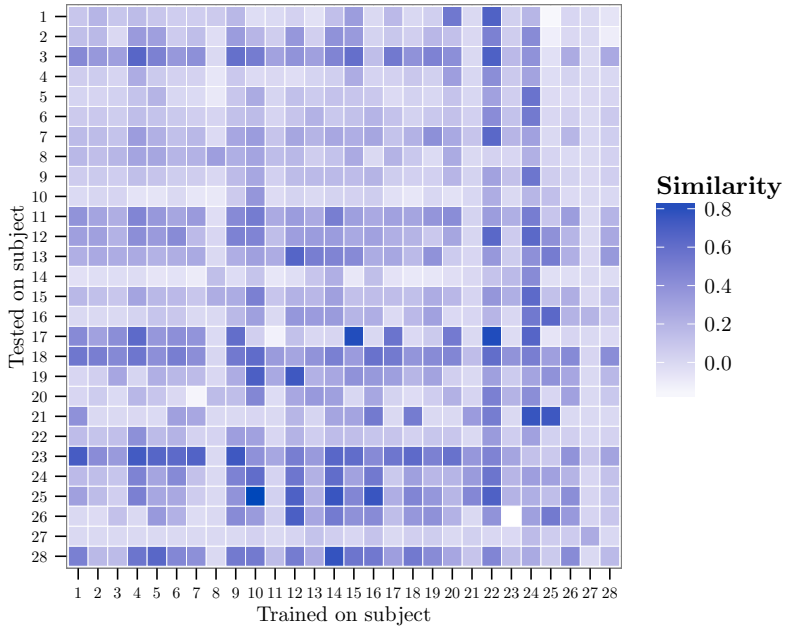
IDAL classifier estimation error

To estimate the similarity matrix for the experiments on Input Data Adaptive Learning (IDAL), a similarity matrix is used that is calculated using only small annotations. This was done to show that the proposed technique can easily be used in conjunction with other proposed methods, in this case DALSA. But evaluation the classifier quality only on a small annotation leads to a bias in the results. Therefore Figure B.1 shows the similarity matrix that is obtained using the complete segmentation and a matrix indicating the difference between both.

As it can be seen, the obtained results are smaller as it would be expected. SURs are per definition placed in areas that clearly belong to a specific type of tissue and by implication are easier to classify. But the differences seem to level out, the error made by each test patient seems to be mostly constant for all training patients.



(a) True similarity matrix estimated on full image



(b) Difference between similarity matrix from full images and SUR

Figure B.1: **(a):** Similarity matrix obtained if the full image is used to estimate the similarity rather than using wDice and SURs. The green points indicate the estimated best training patients. **(b):** The difference between the true similarity estimated on the whole image and the similarity estimated using SURs.

C

Rating Study Questionnaire

The following page contains the questionnaire that is used to evaluate the quality of the segmentations and risk maps from DECT and Perfusion CT.

Rating Study: DECT vs. Perfusion CT

_____ _____

(ID) (Name)

Evaluation of different Segmentation

Comparison

	A				B
Case 01					
Case 03					
Case 04					
Case 05					
Case 06					
Case 07					
Case 09					
Case 10					
Case 11					
Case 12					
Case 13					
Case 15					
Case 16					
Case 18					
Case 20					
Case 21					
Case 22					

Grading A

	1	2	3	4	5	6
Case 01						
Case 03						
Case 04						
Case 05						
Case 06						
Case 07						
Case 09						
Case 10						
Case 11						
Case 12						
Case 13						
Case 15						
Case 16						
Case 18						
Case 20						
Case 21						
Case 22						

Grading B

	1	2	3	4	5	6
Case 01						
Case 03						
Case 04						
Case 05						
Case 06						
Case 07						
Case 09						
Case 10						
Case 11						
Case 12						
Case 13						
Case 15						
Case 16						
Case 18						
Case 20						
Case 21						
Case 22						

Evaluation of different Risk Maps

Comparison

	A				B
Case 01					
Case 03					
Case 04					
Case 05					
Case 06					
Case 07					
Case 09					
Case 10					
Case 11					
Case 12					
Case 13					
Case 15					
Case 16					
Case 18					
Case 20					
Case 21					
Case 22					

Grading A

	1	2	3	4	5	6
Case 01						
Case 03						
Case 04						
Case 05						
Case 06						
Case 07						
Case 09						
Case 10						
Case 11						
Case 12						
Case 13						
Case 15						
Case 16						
Case 18						
Case 20						
Case 21						
Case 22						

Grading B

	1	2	3	4	5	6
Case 01						
Case 03						
Case 04						
Case 05						
Case 06						
Case 07						
Case 09						
Case 10						
Case 11						
Case 12						
Case 13						
Case 15						
Case 16						
Case 18						
Case 20						
Case 21						
Case 22						

D

Additional LLP results

D.1 Results for Yu Experiment

Figure D.1 and D.2 are showing the results obtained for the Yu results. The classification map is showing that the classification areas similar to the expected ones.

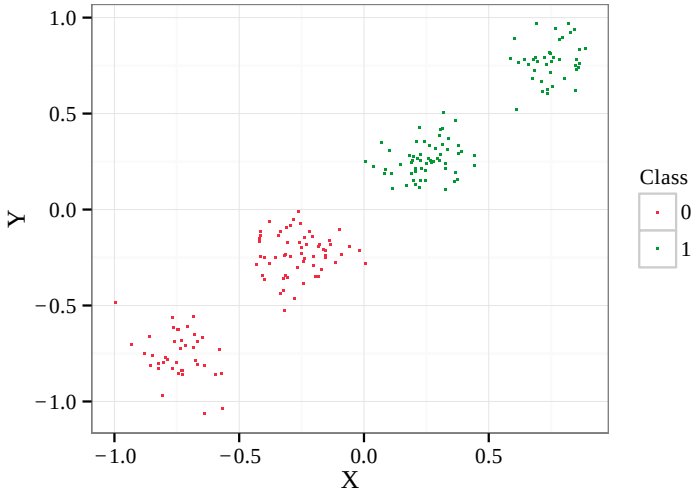
D.2 Results for Patrini Experiments

This sections contains the detailed results for the Patrini-0 to Patrini-16 results. For each experiment, (a) the ground truth, respective training samples, (b) the predicted label for each point in the training group, (c) a heatmap for the area of $(-1|-1)$ to $(1|1)$ colour-coding the label for each position and the classifier accuracy coded by intensity, and (d) a heatmap colour-coding the classifier intensity.

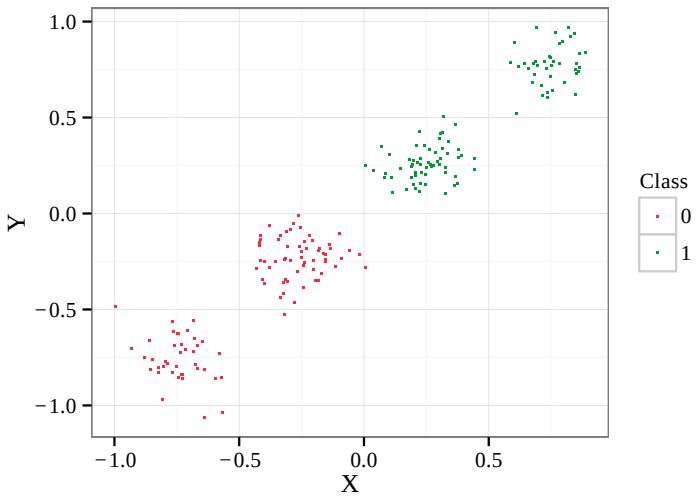
The parameter for each experiment are obtained by running five times a 5-fold cross validation. The range of parameter is given in Table D.1.

TABLE D.1
Parameter range for Patrini-0 to Patrini-16 experiment

Parameter	minimum	maximum	step-size
Number of Folds	5		
Number of repeating runs	5		
Allowed zero bags	0	7	1
Bagging (Observation)	0.5	1.0	0.1
Bagging (Bags)	0.6	1.0	0.1
Thresholds per Node	1	20	1
Feature per node	1	3	1
Trees	50		

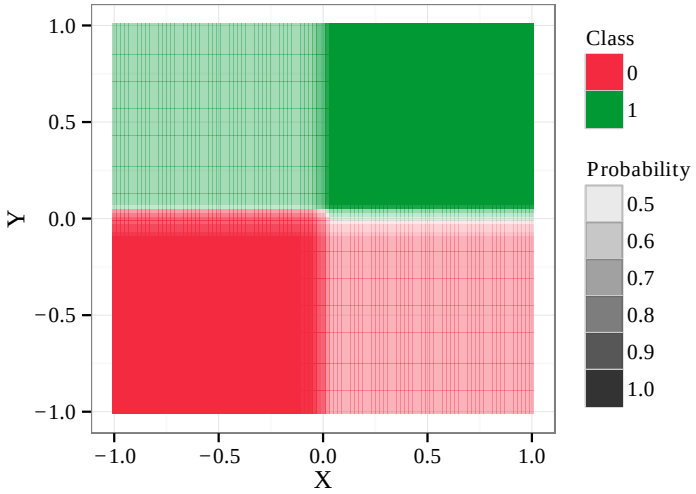


(a) Ground Truth for Patrini-1

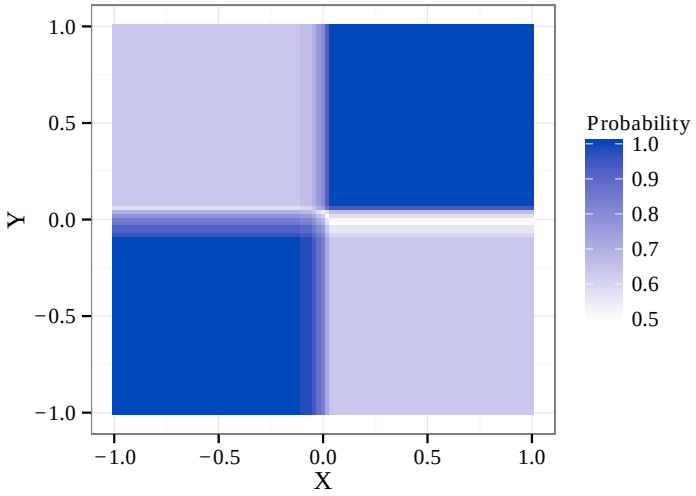


(b) Classification result for Patrini-1

Figure D.1: Scatter-plot of the used training data and the resulting classification for Yu-1 experiment.

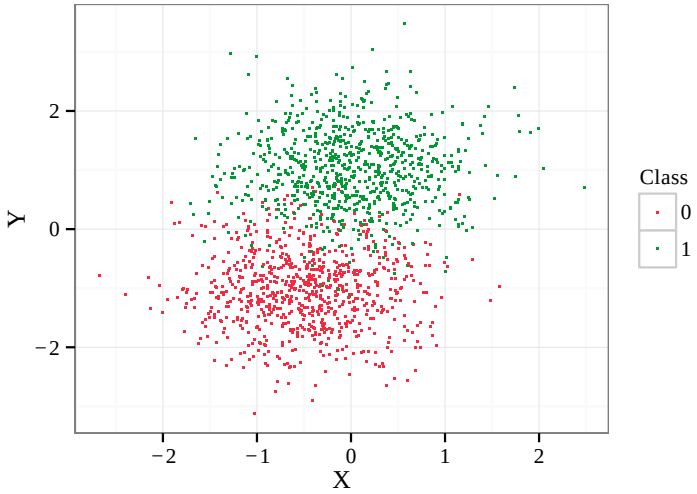


(a) Classification class probability

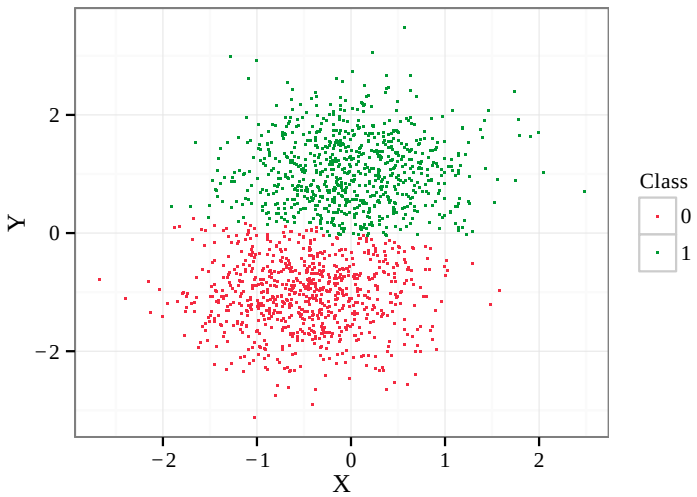


(b) Classifier certainty

Figure D.2: Colour coded classification map and certainty map of the classifier obtained on Yu-1 experiment.

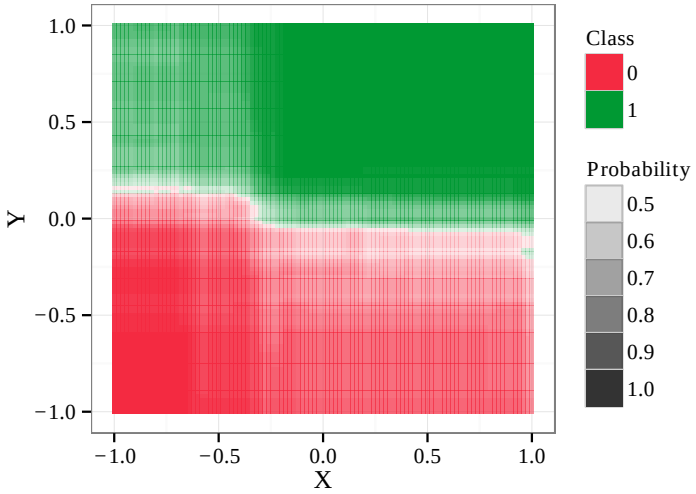


(a) Ground Truth for Patrini-0

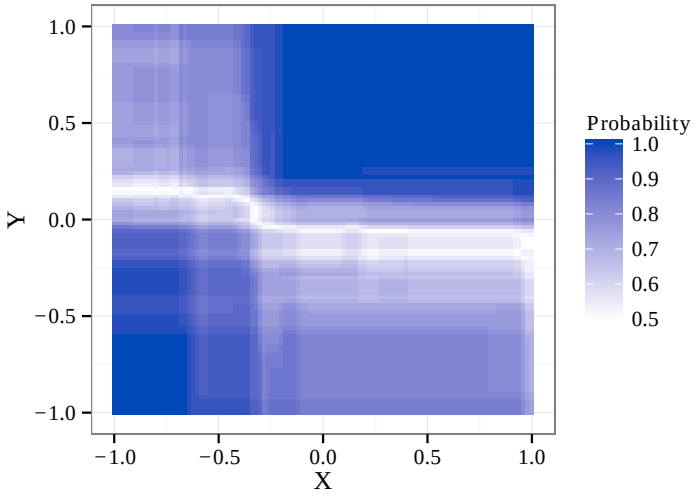


(b) Classification result for Patrini-0

Figure D.3: Scatter-plot of the used training data and the resulting classification for the Patrini-0 experiment.

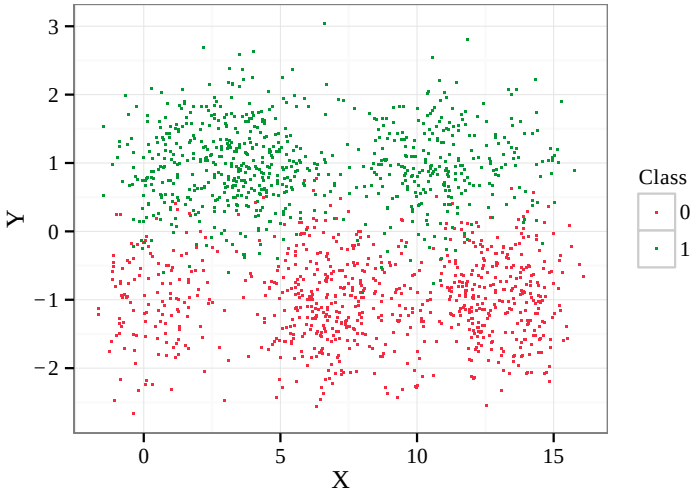


(a) Classification class probability

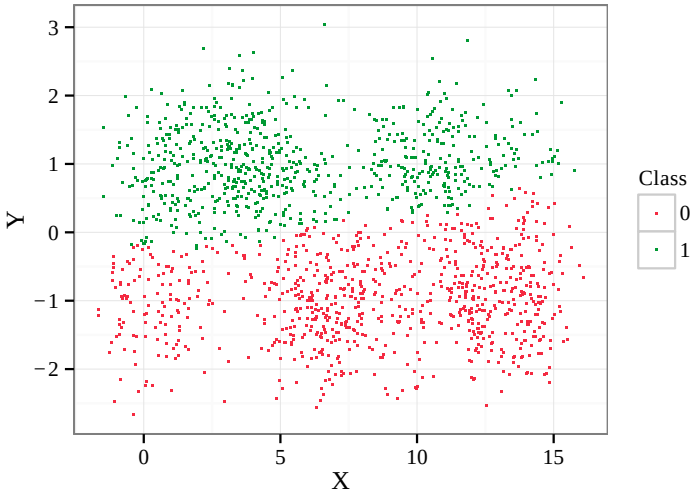


(b) Classifier certainty

Figure D.4: Colour coded classification map and certainty map of the classifier obtained on the Patrini-0 experiment.

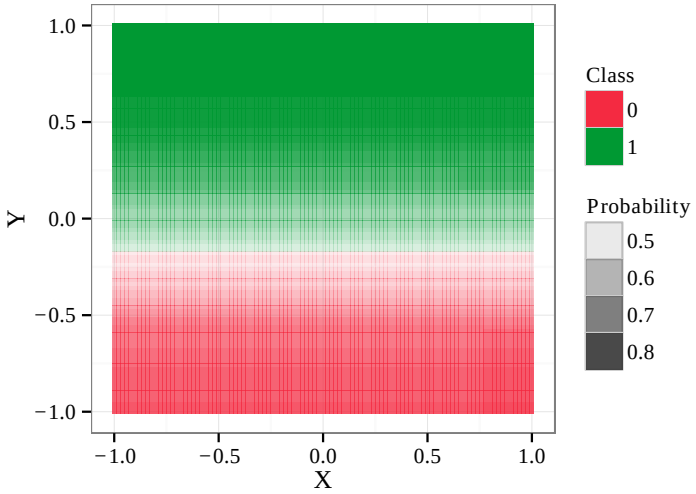


(a) Ground Truth for Patrini-1

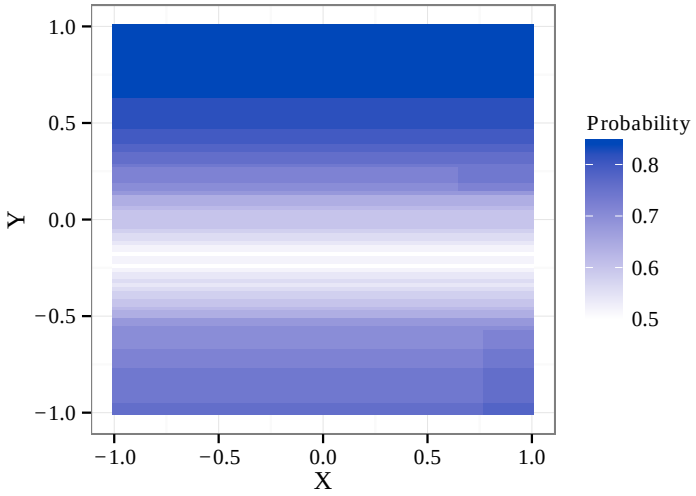


(b) Classification result for Patrini-1

Figure D.5: Scatter-plot of the used training data and the resulting classification for the Patrini-1 experiment.

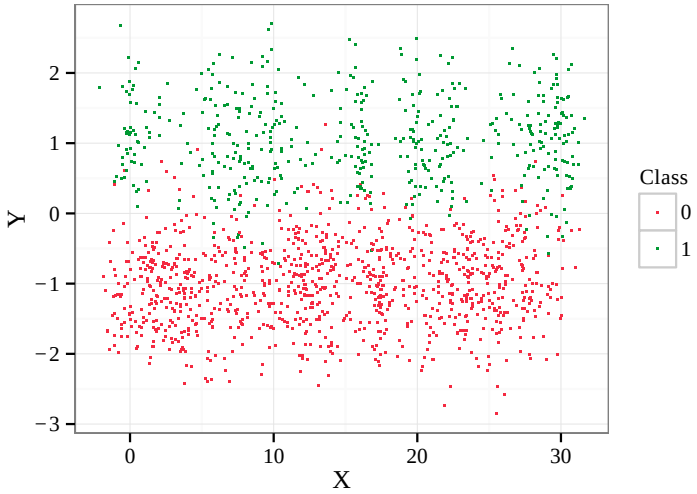


(a) Classification class probability

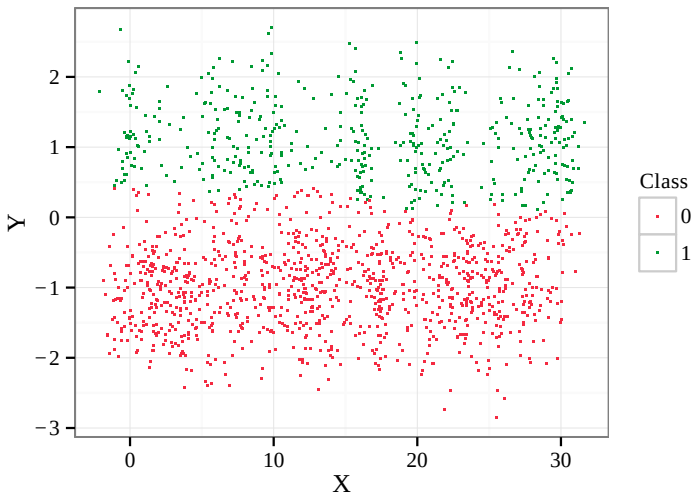


(b) Classifier certainty

Figure D.6: Colour coded classification map and certainty map of the classifier obtained on the Patrini-1 experiment.

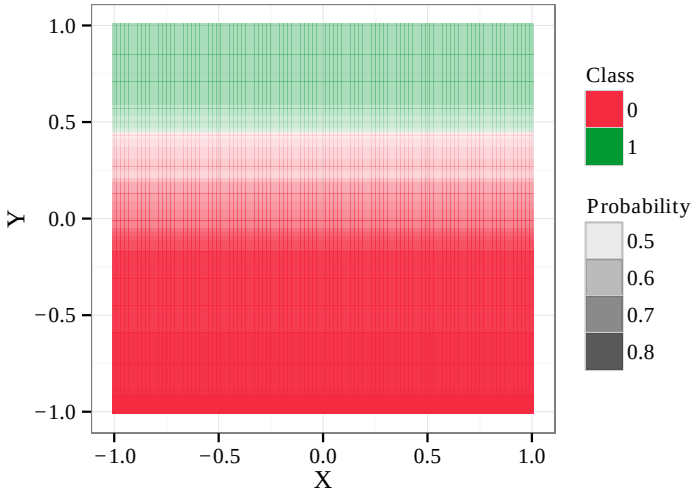


(a) Ground Truth for Patrini-2

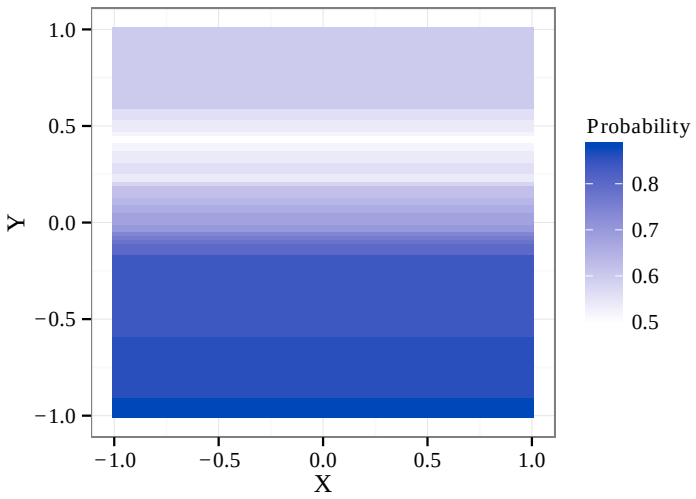


(b) Classification result for Patrini-2

Figure D.7: Scatter-plot of the used training data and the resulting classification for the Patrini-2 experiment.

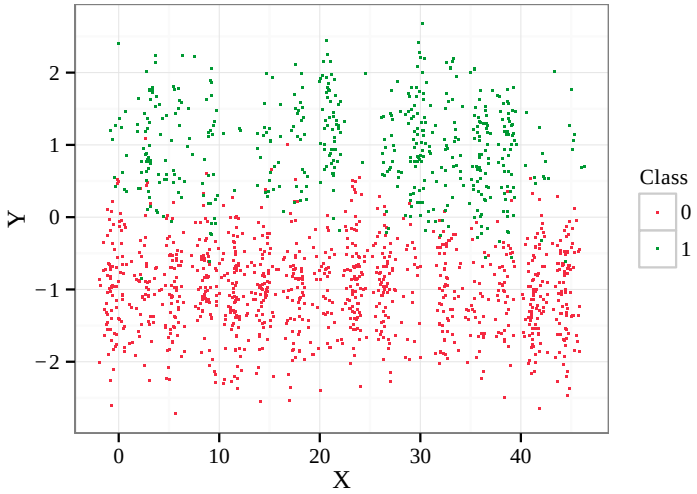


(a) Classification class probability

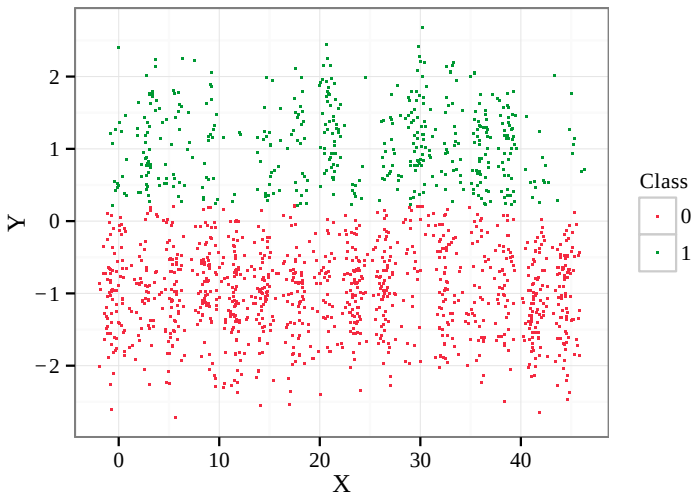


(b) Classifier certainty

Figure D.8: Colour coded classification map and certainty map of the classifier obtained on the Patrini-2 experiment.

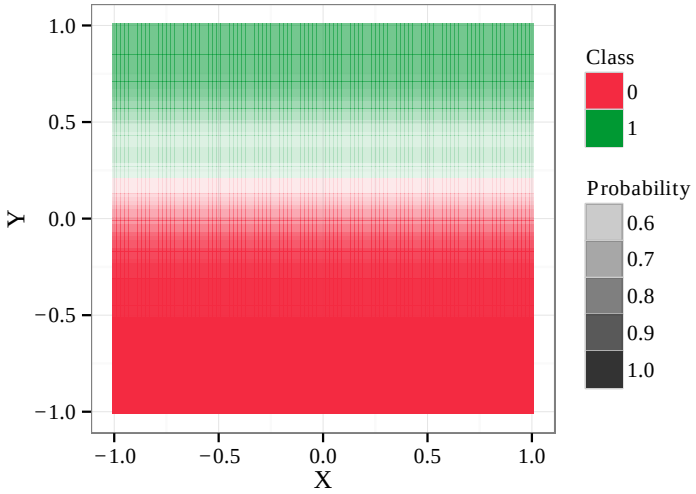


(a) Ground Truth for Patrini-3

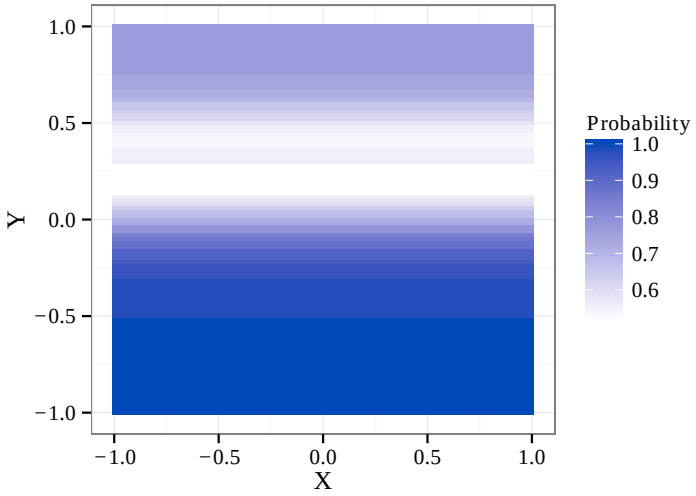


(b) Classification result for Patrini-3

Figure D.9: Scatter-plot of the used training data and the resulting classification for the Patrini-3 experiment.

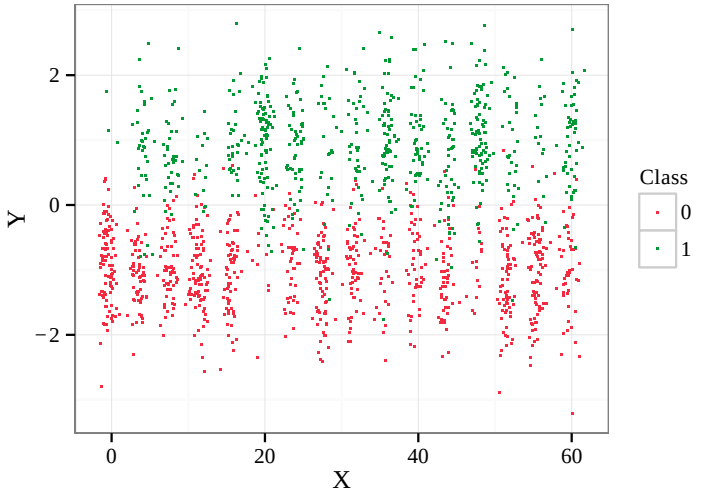


(a) Classification class probability

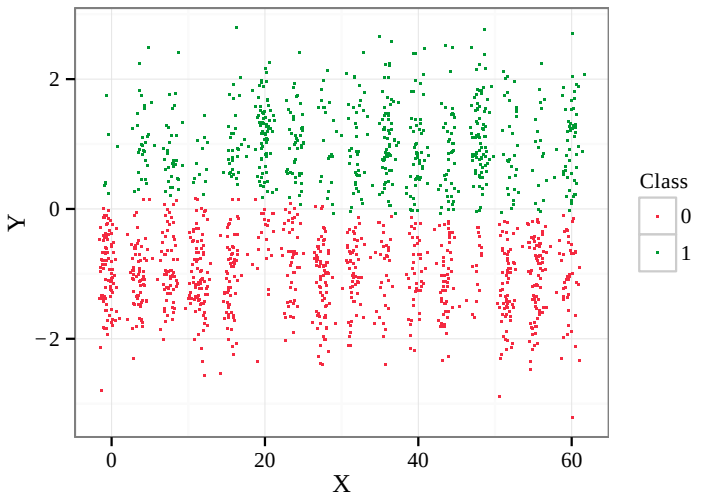


(b) Classifier certainty

Figure D.10: Colour coded classification map and certainty map of the classifier obtained on the Patrini-3 experiment.

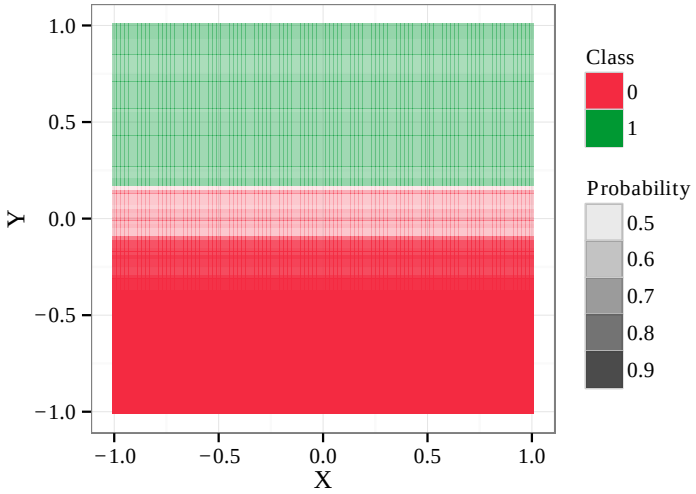


(a) Ground Truth for Patrini-4

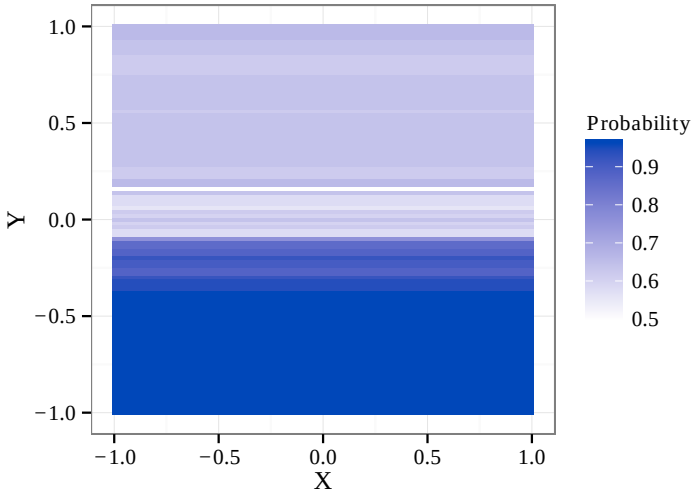


(b) Classification result for Patrini-4

Figure D.11: Scatter-plot of the used training data and the resulting classification for the Patrini-4 experiment.

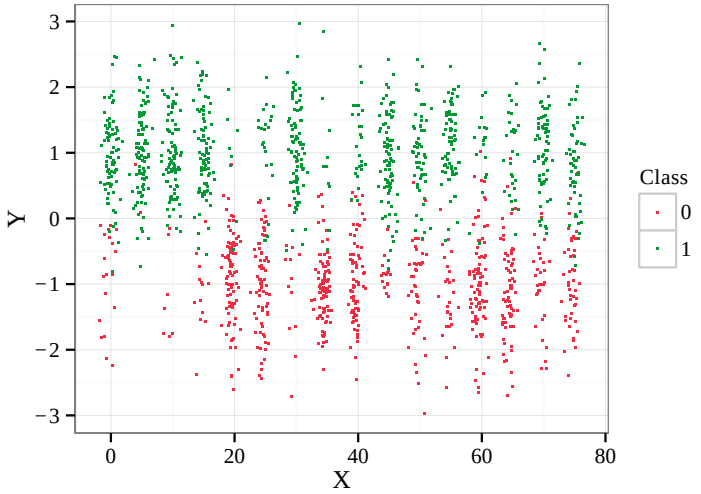


(a) Classification class probability

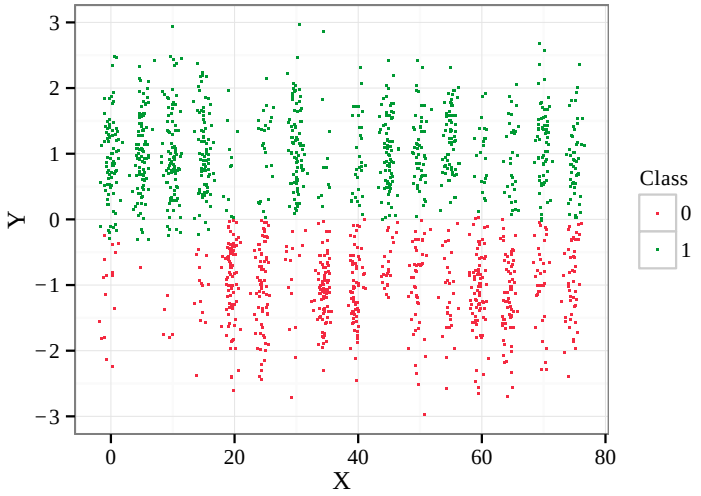


(b) Classifier certainty

Figure D.12: Colour coded classification map and certainty map of the classifier obtained on the Patrini-4 experiment.

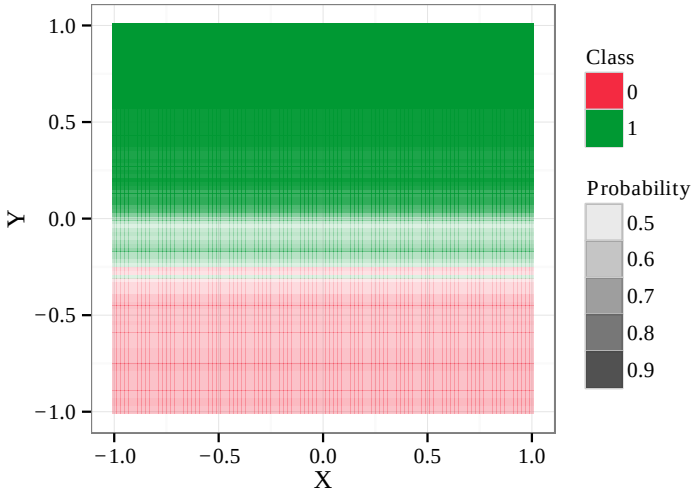


(a) Ground Truth for Patrini-5

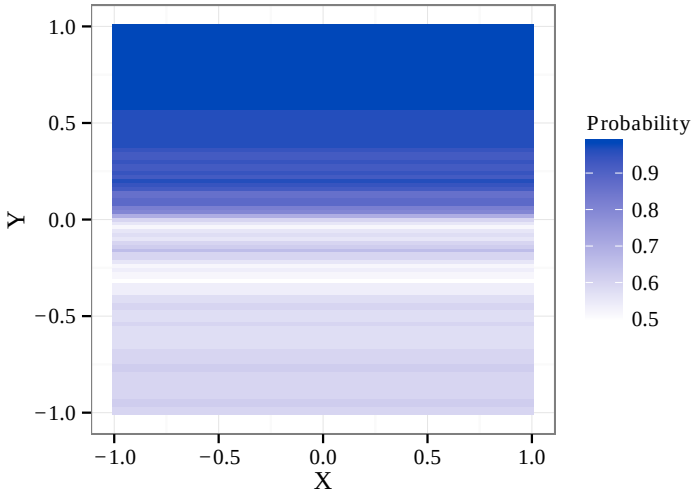


(b) Classification result for Patrini-5

Figure D.13: Scatter-plot of the used training data and the resulting classification for the Patrini-5 experiment.

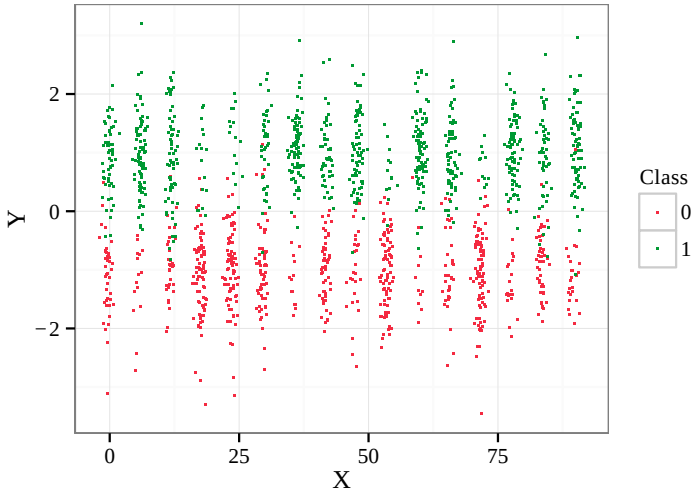


(a) Classification class probability

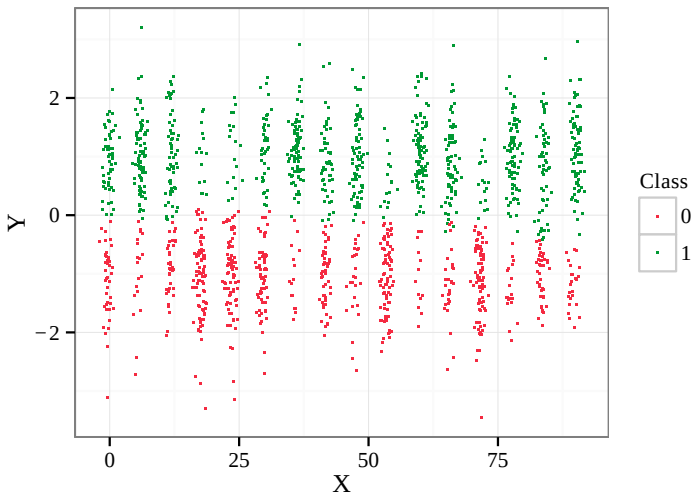


(b) Classifier certainty

Figure D.14: Colour coded classification map and certainty map of the classifier obtained on the Patrini-5 experiment.

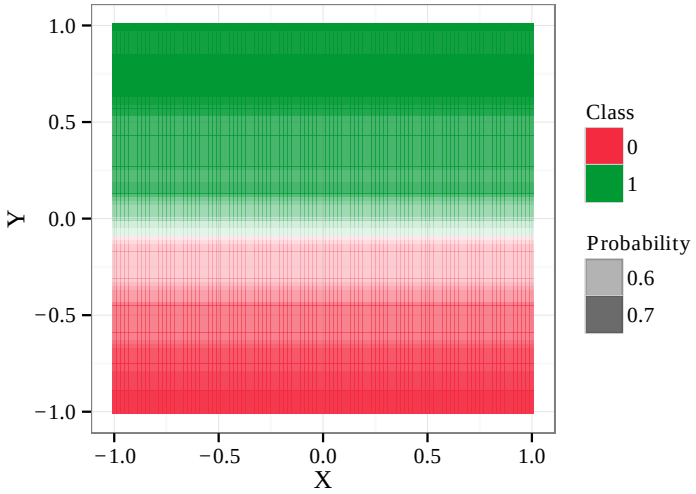


(a) Ground Truth for Patrini-6

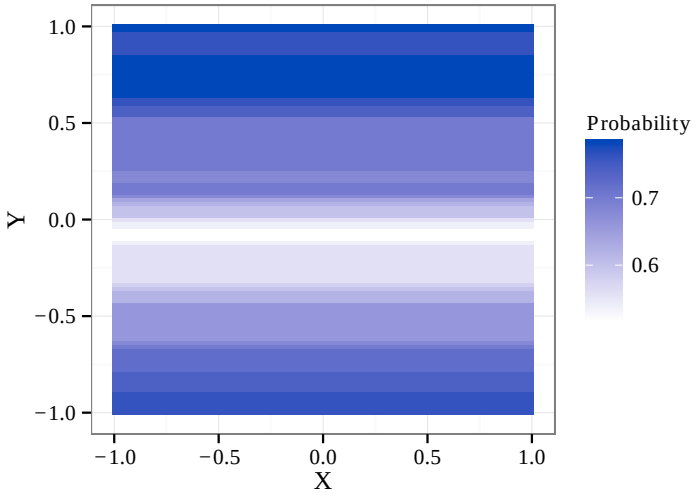


(b) Classification result for Patrini-6

Figure D.15: Scatter-plot of the used training data and the resulting classification for the Patrini-6 experiment.

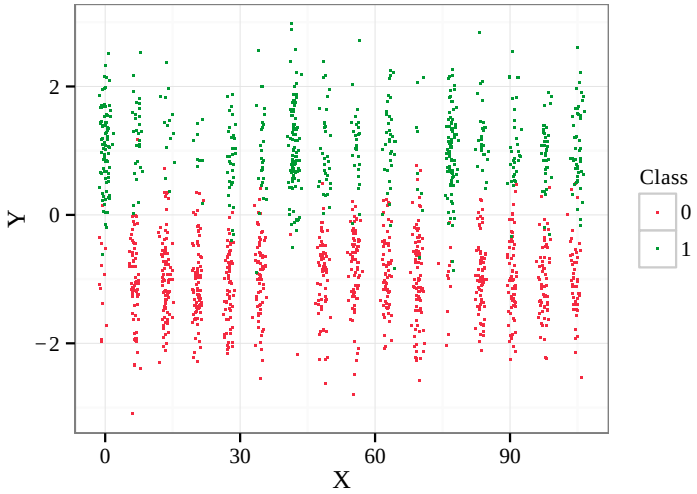


(a) Classification class probability

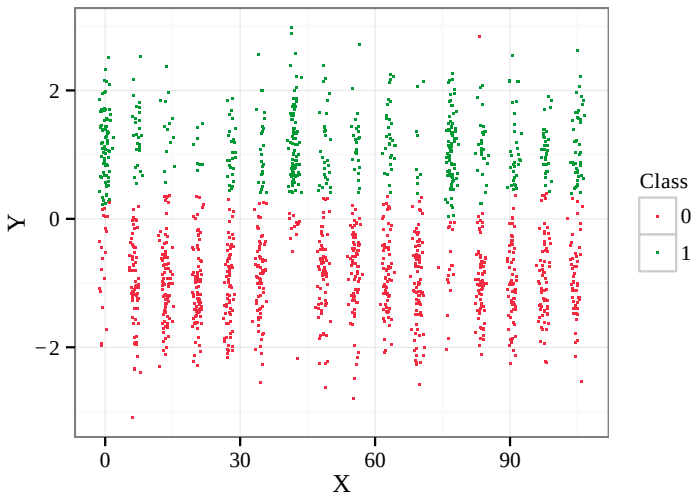


(b) Classifier certainty

Figure D.16: Colour coded classification map and certainty map of the classifier obtained on the Patrini-6 experiment.

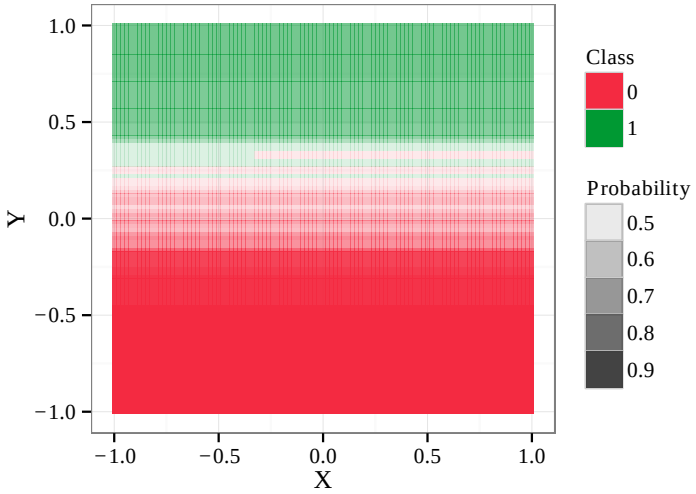


(a) Ground Truth for Patrini-7

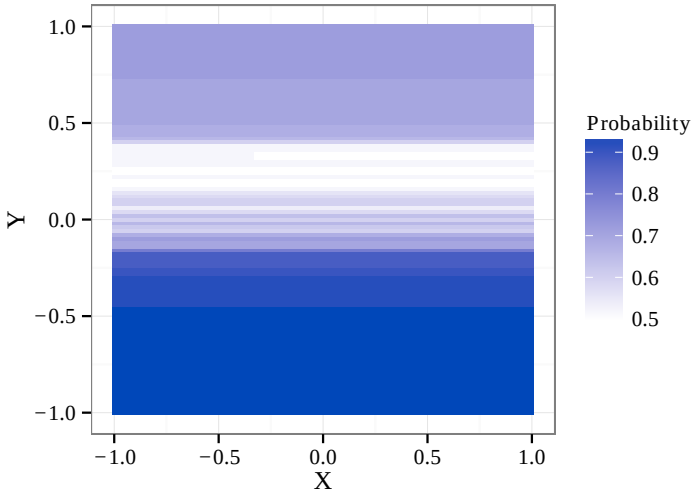


(b) Classification result for Patrini-7

Figure D.17: Scatter-plot of the used training data and the resulting classification for the Patrini-7 experiment.

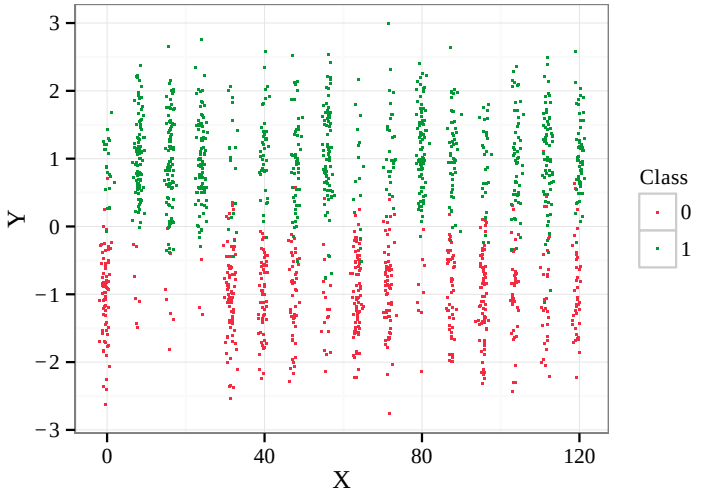


(a) Classification class probability

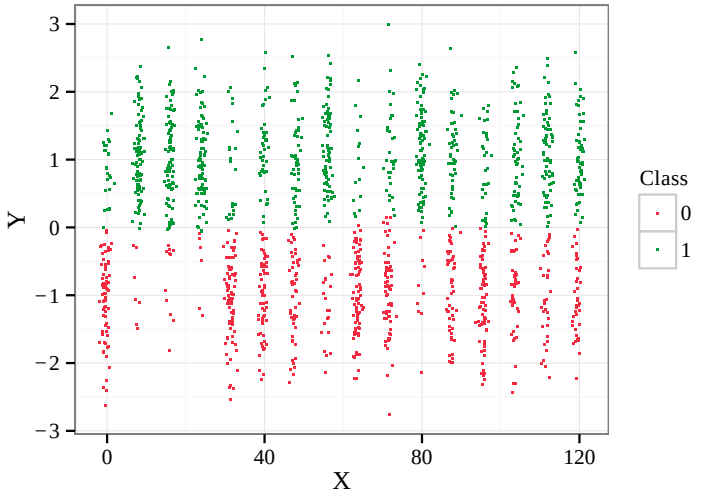


(b) Classifier certainty

Figure D.18: Colour coded classification map and certainty map of the classifier obtained on the Patrini-7 experiment.

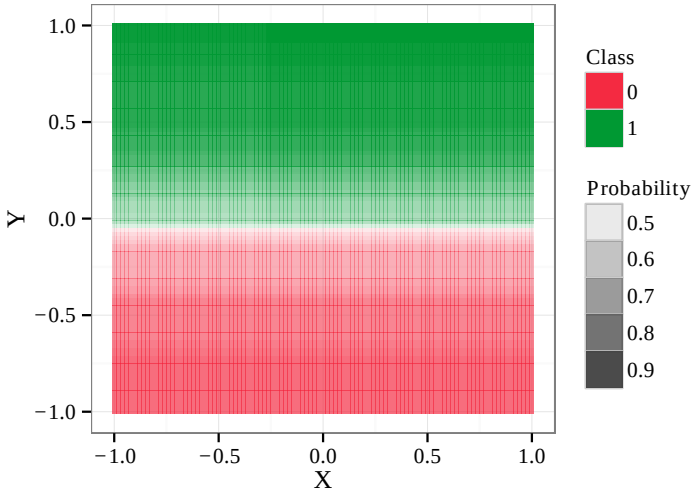


(a) Ground Truth for Patrini-8

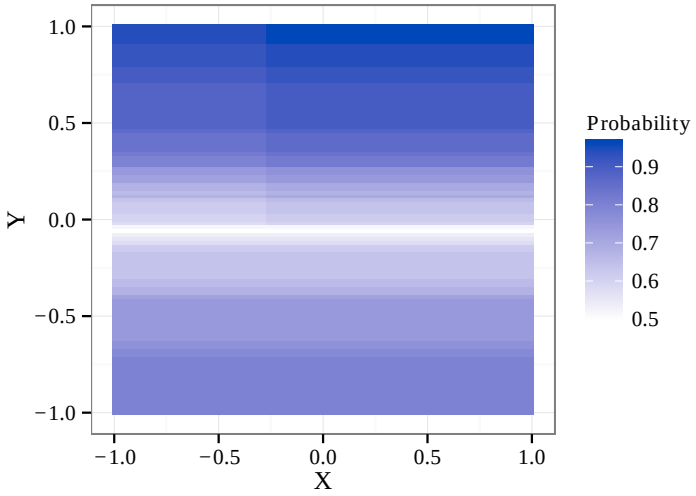


(b) Classification result for Patrini-8

Figure D.19: Scatter-plot of the used training data and the resulting classification for the Patrini-8 experiment.

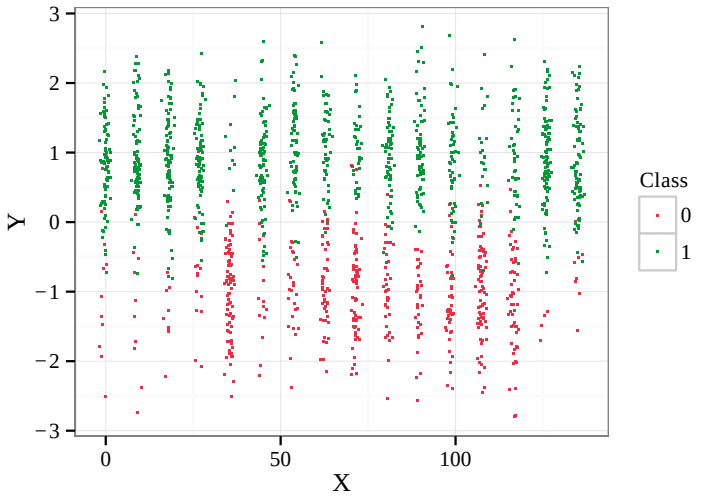


(a) Classification class probability

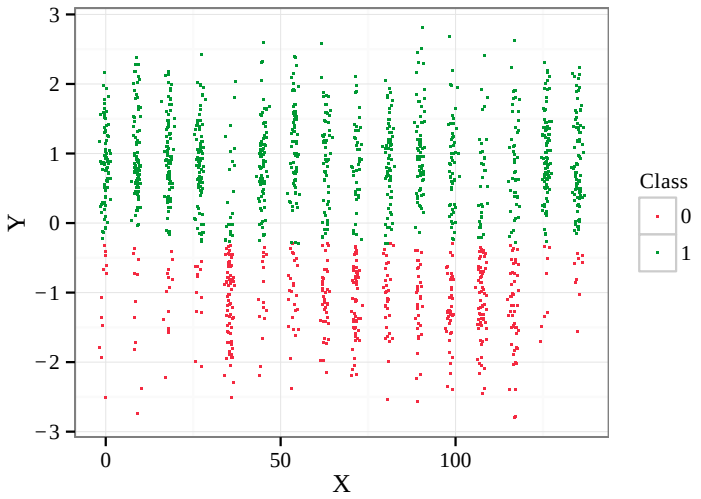


(b) Classifier certainty

Figure D.20: Colour coded classification map and certainty map of the classifier obtained on the Patrini-8 experiment.

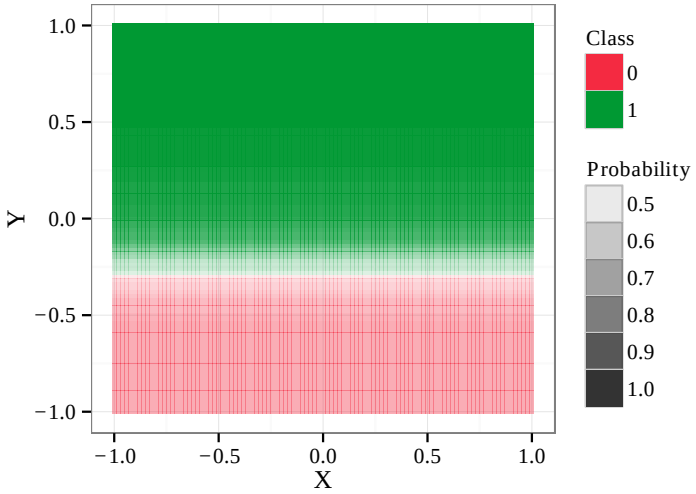


(a) Ground Truth for Patrini-9

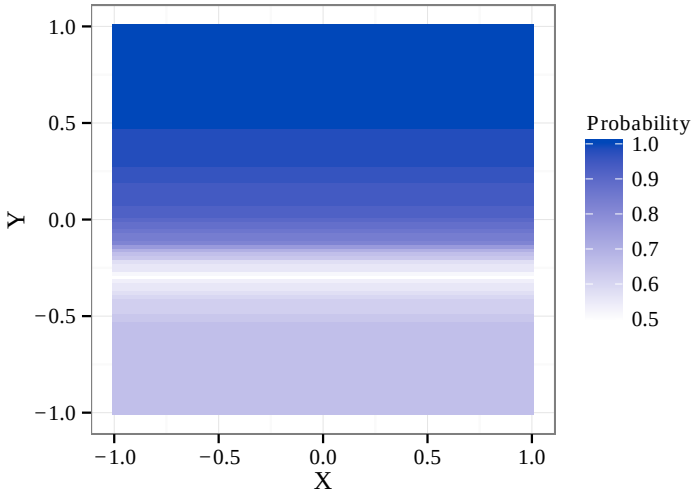


(b) Classification result for Patrini-9

Figure D.21: Scatter-plot of the used training data and the resulting classification for the Patrini-9 experiment.

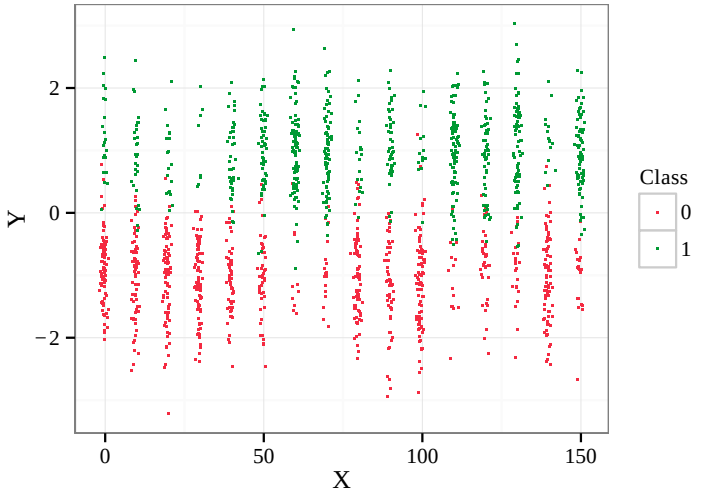


(a) Classification class probability

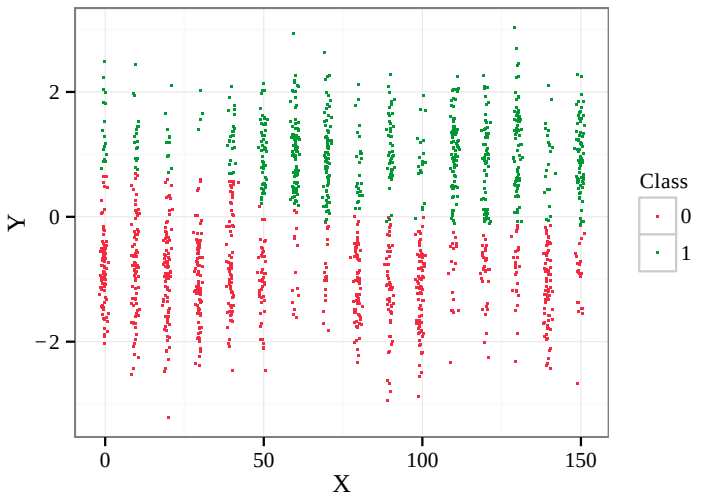


(b) Classifier certainty

Figure D.22: Colour coded classification map and certainty map of the classifier obtained on the Patrini-9 experiment.

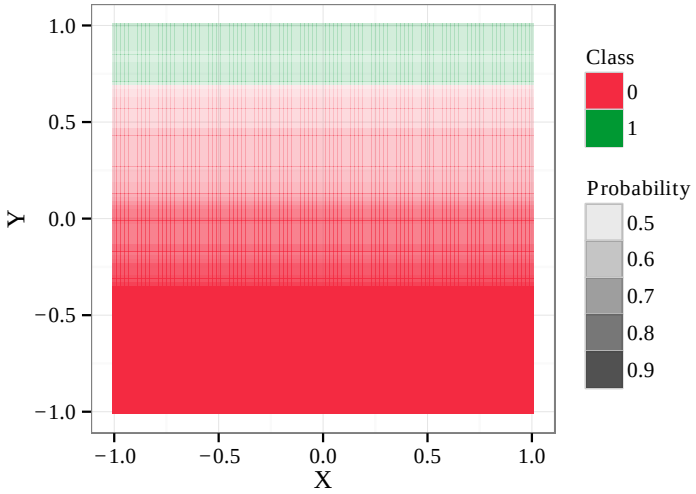


(a) Ground Truth for Patrini-10

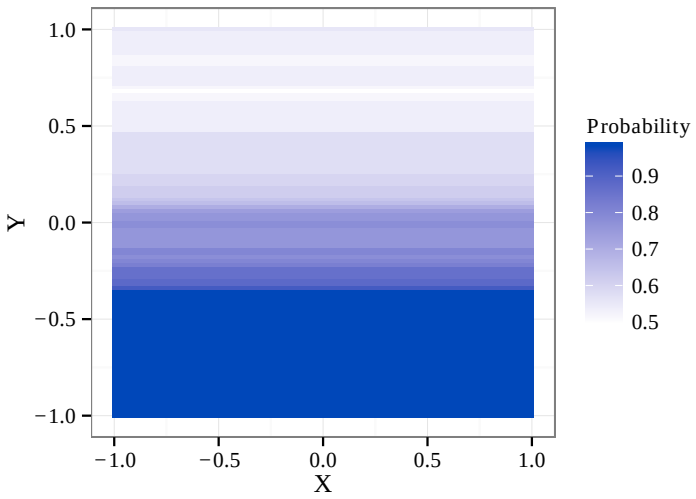


(b) Classification result for Patrini-10

Figure D.23: Scatter-plot of the used training data and the resulting classification for the Patrini-10 experiment.

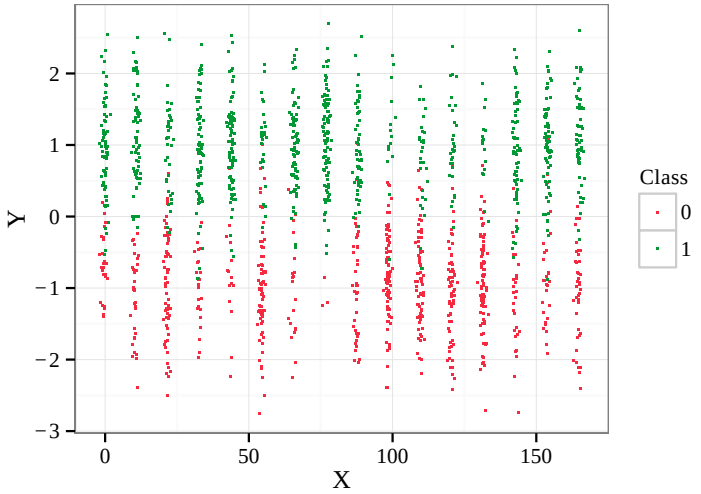


(a) Classification class probability

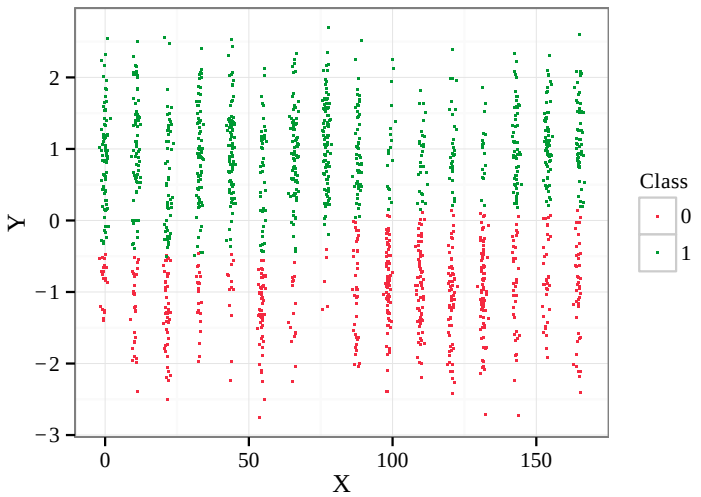


(b) Classifier certainty

Figure D.24: Colour coded classification map and certainty map of the classifier obtained on the Patrini-10 experiment.

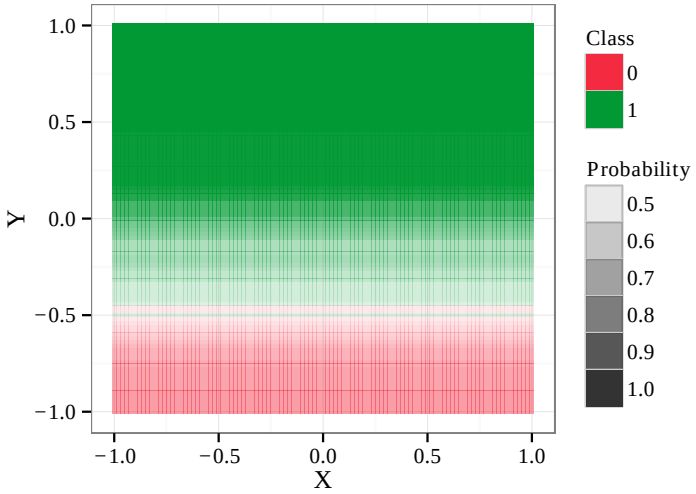


(a) Ground Truth for Patrini-11

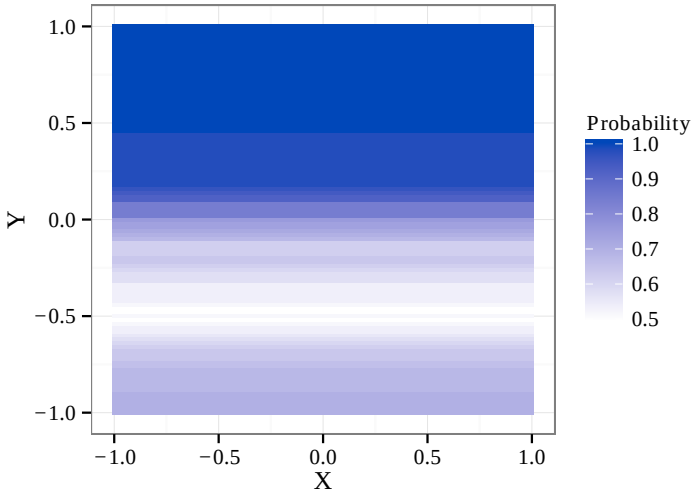


(b) Classification result for Patrini-11

Figure D.25: Scatter-plot of the used training data and the resulting classification for the Patrini-11 experiment.

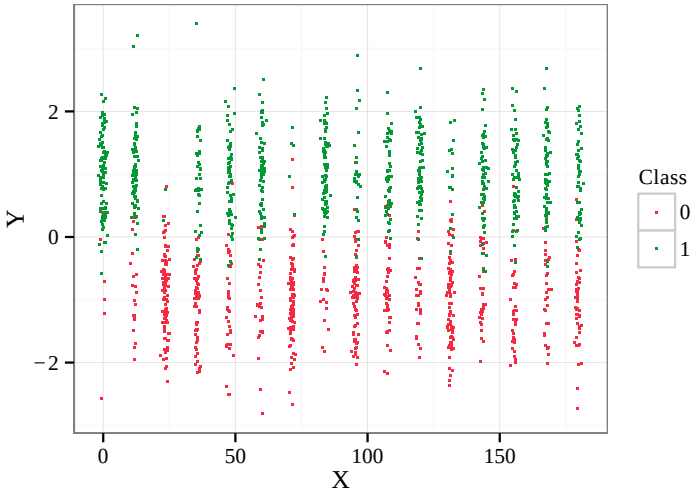


(a) Classification class probability

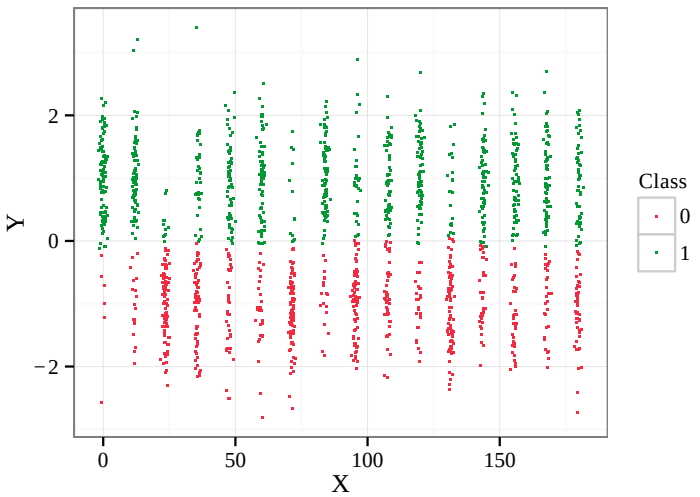


(b) Classifier certainty

Figure D.26: Colour coded classification map and certainty map of the classifier obtained on the Patrini-11 experiment.

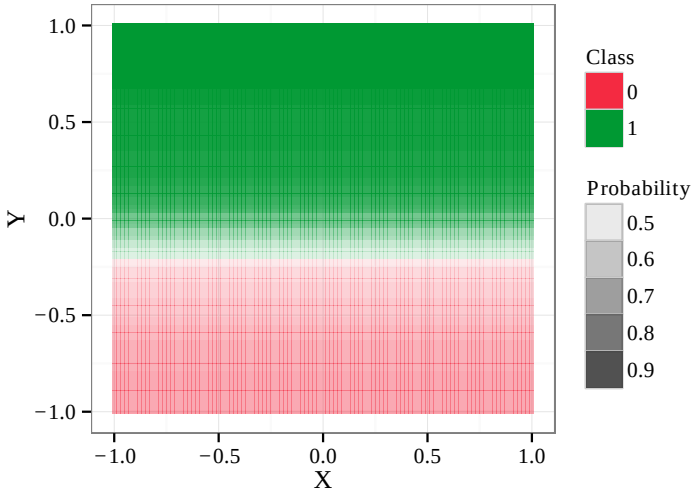


(a) Ground Truth for Patrini-12

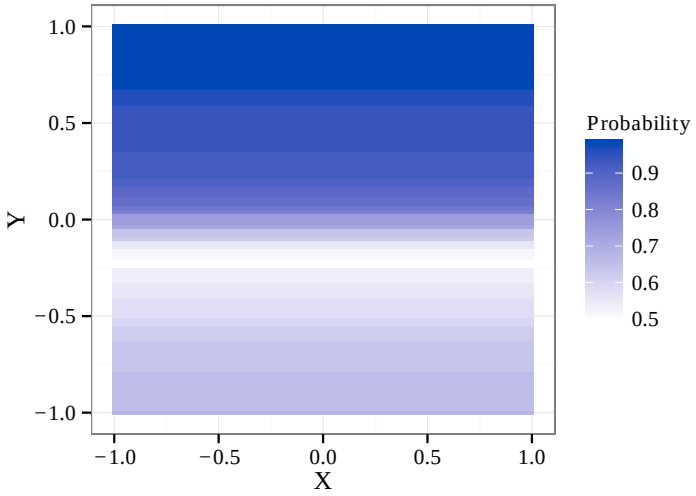


(b) Classification result for Patrini-12

Figure D.27: Scatter-plot of the used training data and the resulting classification for the Patrini-12 experiment.

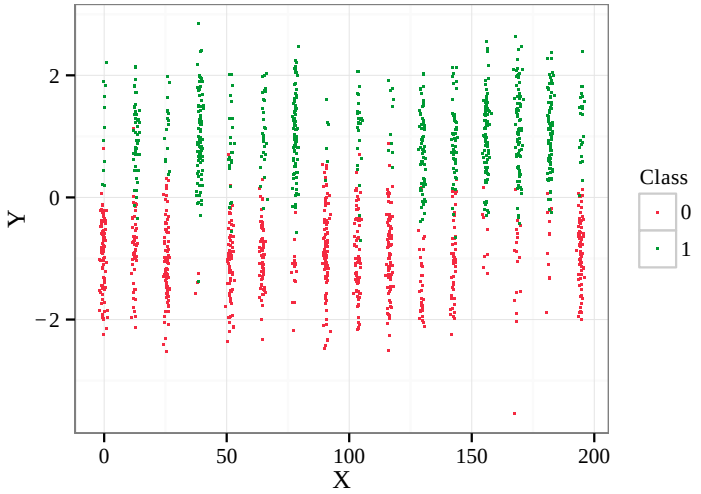


(a) Classification class probability

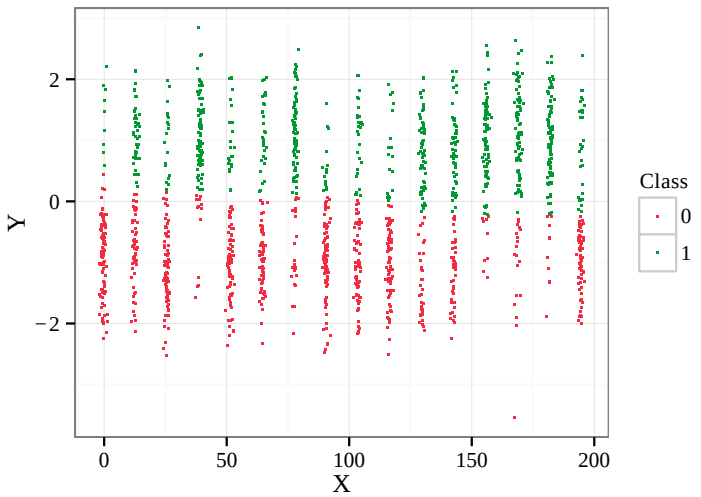


(b) Classifier certainty

Figure D.28: Colour coded classification map and certainty map of the classifier obtained on the Patrini-12 experiment.

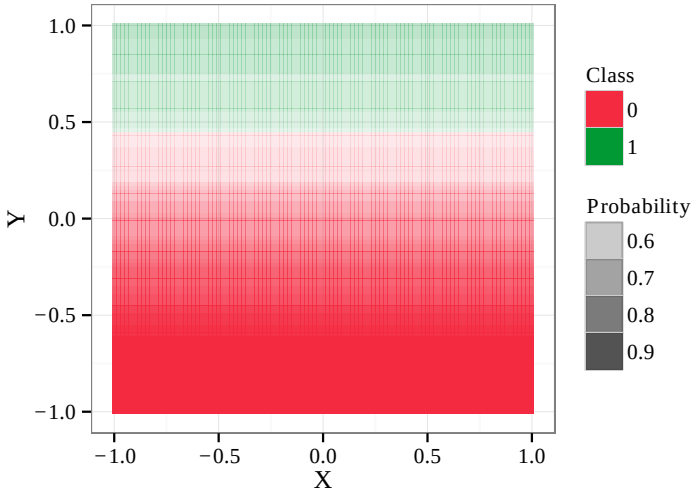


(a) Ground Truth for Patrini-13

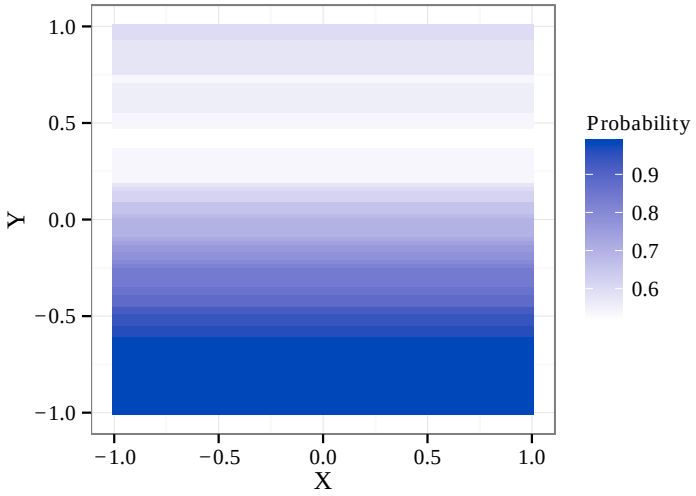


(b) Classification result for Patrini-13

Figure D.29: Scatter-plot of the used training data and the resulting classification for the Patrini-13 experiment.

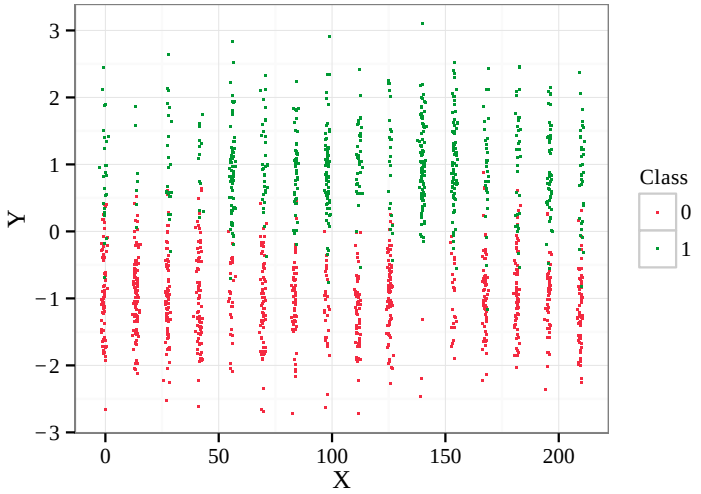


(a) Classification class probability

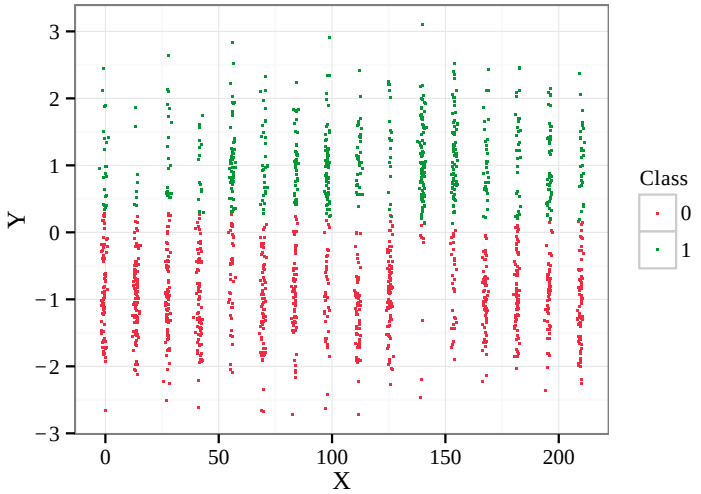


(b) Classifier certainty

Figure D.30: Colour coded classification map and certainty map of the classifier obtained on the Patrini-13 experiment.

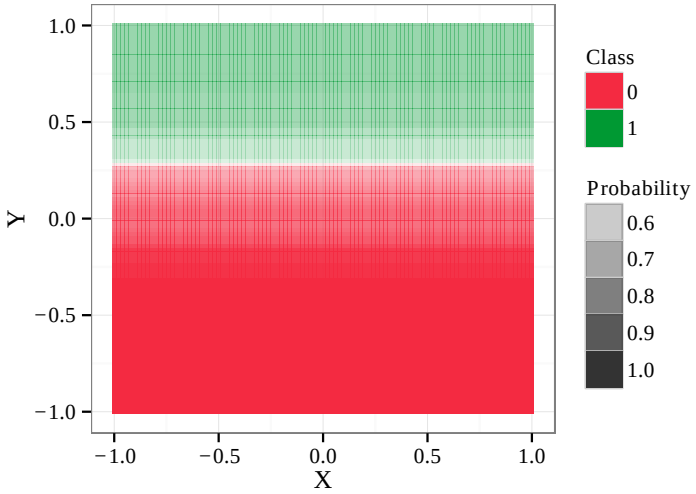


(a) Ground Truth for Patrini-14

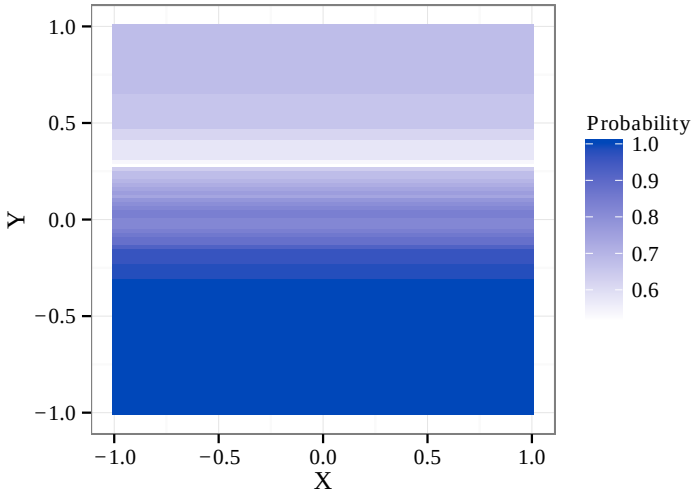


(b) Classification result for Patrini-14

Figure D.31: Scatter-plot of the used training data and the resulting classification for the Patrini-14 experiment.

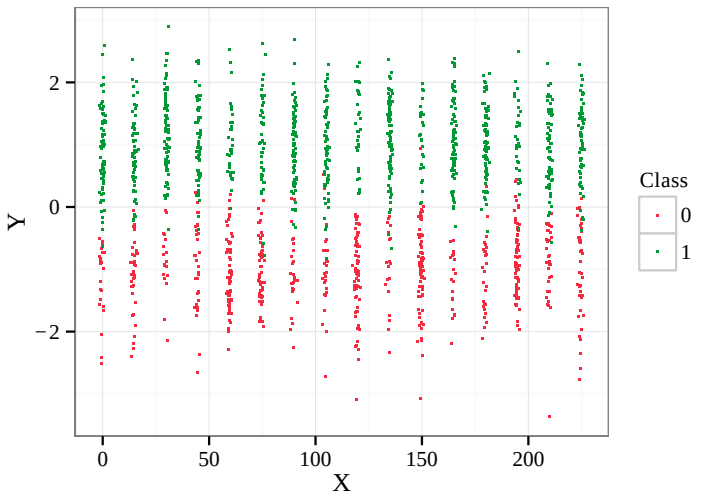


(a) Classification class probability

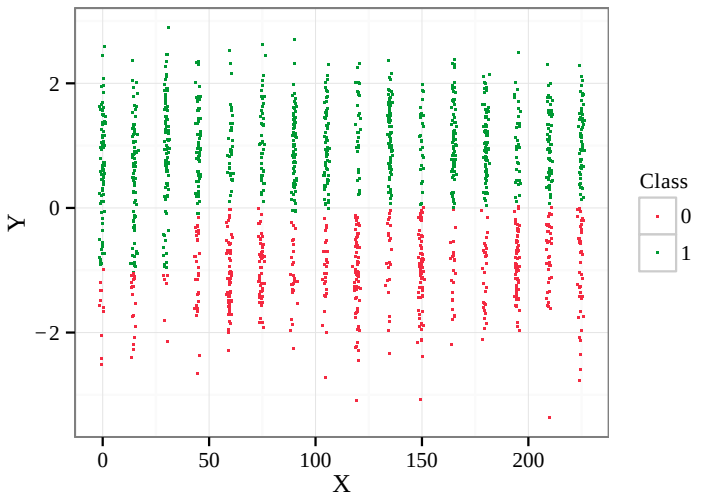


(b) Classifier certainty

Figure D.32: Colour coded classification map and certainty map of the classifier obtained on the Patrini-14 experiment.

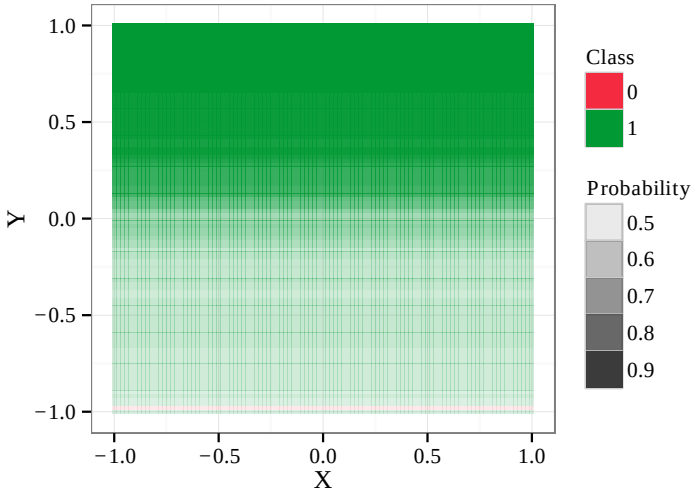


(a) Ground Truth for Patrini-15

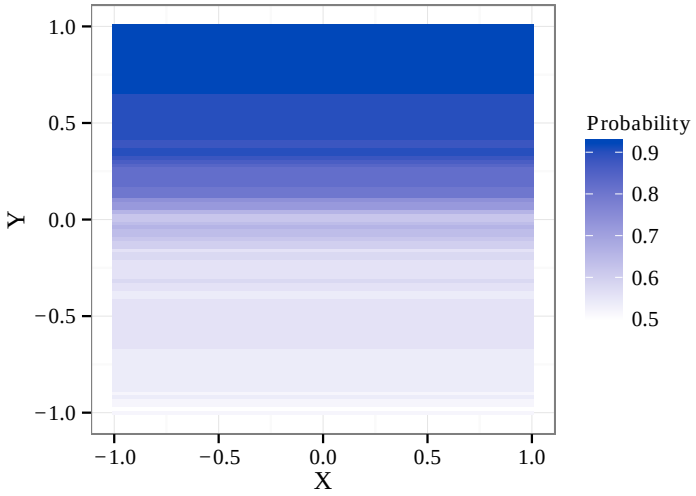


(b) Classification result for Patrini-15

Figure D.33: Scatter-plot of the used training data and the resulting classification for the Patrini-15 experiment.

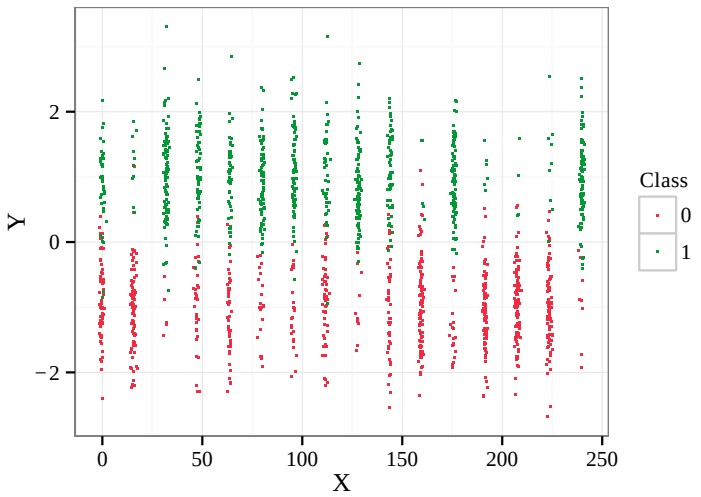


(a) Classification class probability

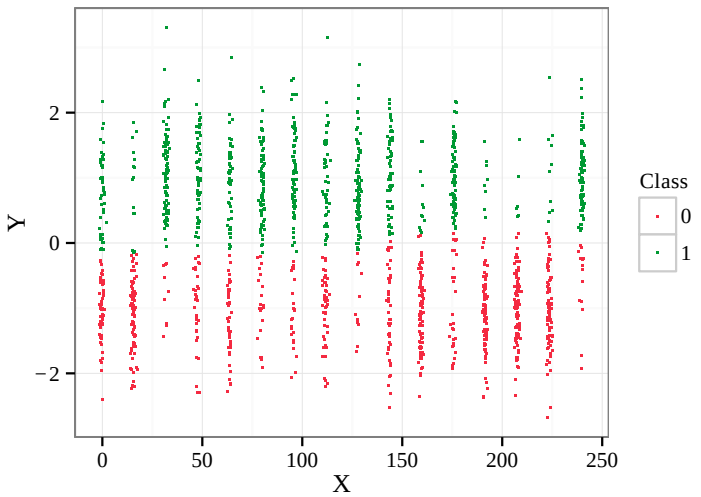


(b) Classifier certainty

Figure D.34: Colour coded classification map and certainty map of the classifier obtained on the Patrini-15 experiment.

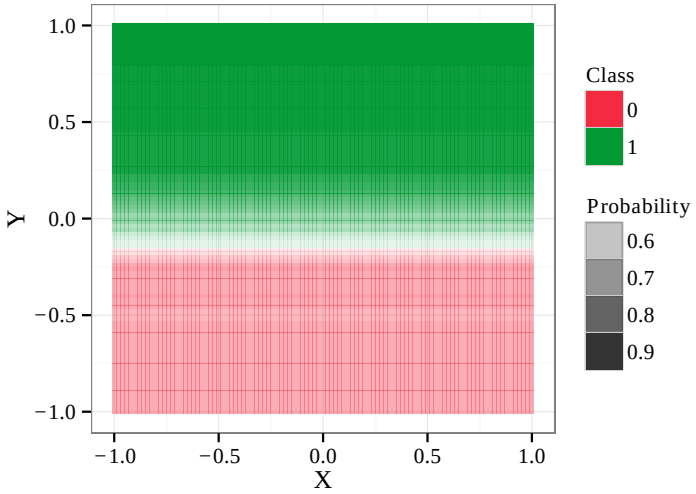


(a) Ground Truth for Patrini-16

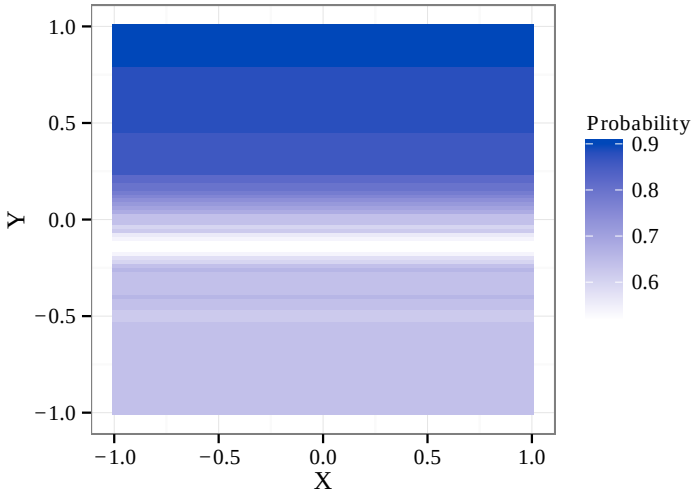


(b) Classification result for Patrini-16

Figure D.35: Scatter-plot of the used training data and the resulting classification for the Patrini-16 experiment.



(a) Classification class probability



(b) Classifier certainty

Figure D.36: Colour coded classification map and certainty map of the classifier obtained on the Patrini-16 experiment.