

DISS. ETH NO. 21852

# **Machine Learning Approaches for Structure Analysis in Medical Image Data**

A thesis submitted to attain the degree of  
DOCTOR OF SCIENCES of ETH ZURICH  
(Dr. sc. ETH Zurich)

presented by

PETER JOHANNES SCHÜFFLER

M.Sc., Saarland University

born November 9<sup>th</sup>, 1983

citizen of Germany

accepted on the recommendation of

Prof. Dr. Joachim M. Buhmann, examiner

Prof. Dr. Holger Moch, co-examiner

Prof. Dr. Franciscus M. Vos, co-examiner

Prof. Dr. Volker Roth, co-examiner

2014



*For my family*



# ABSTRACT

The principal focus of this thesis lies on the investigation of new computational approaches for the automated processing of medical images to identify normal and diseased structures. Medical images constitute a particularity for computer vision methods as they commonly result from highly complex acquisition techniques. Frequently, they visualize biological objects that are hidden to the naked human eye. Two examples of medical images are investigated in this thesis: *microscopic* immunohistochemically stained cancer tissue images known from pathology and *macroscopic* abdominal magnetic resonance images known from radiology. In both cases, the medical task is to diagnose and if needed quantify abnormal structures such as cancer or bowel diseases. We aim to solve this task in computer-aided or fully automated manner to be able to improve daily clinic diagnostics and scientific research on large patient cohorts. This thesis is structured in following parts:

First, we introduce technical terms relevant for this thesis. *Tissue microarrays* (TMA) are medical instruments for experimental cancer research in large patient cohorts. *Magnetic resonance imaging* (MRI) is a medical imaging method used to visualize the gastrointestinal tract. *Crohn's disease* (CD) and *CD severity* are explained.

*Technical  
Background*

The data basis of the computer-aided staining estimation pipeline is formed by eight TMA images of clear cell renal cell carcinoma patients and six TMA images of prostate cancer patients. Each image is fully labeled by two medical doctors. Additionally, two labeled MRI datasets of 27 and 35 CD patients are introduced serving as development and validation datasets of our automatic CD detection and severity estimation.

*Datasets*

For a medical understanding of a variety of cancers, the morphological evaluation of hundreds of histologic tissue images has emerged as fundamental procedure in clinical cancer research. We define in this chapter a new computational pipeline for automated cell nucleus detection, segmentation, classification and staining estimation on this type of images. We investigate similarity based approaches such as (multiple) kernel learning for nucleus classification. Our best classifier reaches a classification accuracy of 83% which is as good as the manual annotation: the inter-pathologist accuracy for classification of 1633 renal clear cell carcinoma nuclei is 80%. Further, active learning for nucleus classification reduces the number of annotated training samples by the factor of two.

*TMA Staining  
Estimation  
Pipeline*

The automated staining estimation pipeline is implemented in the user-friendly and free Java program TMARKER.

*TMARKER*

*Single Cell  
Segmentation*

Single cell analysis of protein expression profiles constitute an emerging field in recent cancer research. *Highly multiplexed mass cytometry* has been introduced as new imaging method to visualize localized and quantified expression rates of dozens of proteins in single tissue slices. In this chapter, we study the single cell segmentation in this new type of medical images. Our watershed based algorithm exploits the highly registered joint information of multiple membrane and nucleus proteins.

*Automatic  
Crohn's Disease  
Detection*

As an example of macroscopic computer-aided disease recognition, we investigate in this chapter the automatic CD detection on MRI data. An automated system on MRI basis is favorable for research and daily clinic to process the increasing amount of patient data. We define a novel hierarchical CD segmentation system with two steps: First, diseased areas are coarsely localized in the images. Second, a pixel-wise classification system precisely segments CD in the aforementioned areas. Dedicated image features, such as texture anisotropy, spatial context and higher order statistics are developed for this task. We achieve a CD segmentation with a Dice metric of 91.9% compared to manual segmentations.

*Crohn's Disease  
Severity  
Assessment*

CD is usually graded in its severity influencing therapy strategies and indicating surgery. Conventional grading systems such as the endoscopic index of severity (CDEIS) are subjective or invasive. Physicians commonly rely on multiple grading systems for a holistic view on the patient. This chapter investigates the potential of MRI to serve as standard for a new CD severity grading system. 14 MRI features manually assessed by four radiologists serve as data basis for an exhaustive model development pipeline. Our proposed CD severity model shows favorable performance compared to the literature models MaRIA and CDA.

*Automatic  
Feature  
Extraction*

As a novel continuation of the MRI based CD severity assessment, we incorporate two new automatically measured MRI features: *automatic wall thickness* and *dynamic contrast enhancement*. The two dedicated CD descriptors improve the correlation to segmental CDEIS to over 80%.

We conclude the work on computational radiology with a suggestion to combine the automatic CD *detection* with CD *severity estimation* to a holistic medical approach. Automated feature extraction has shown to significantly improve CDEIS regression and should therefore be considered in future applications.

# ZUSAMMENFASSUNG

Der Fokus dieser Arbeit liegt in der Erforschung neuer Methoden für die medizinische Bildanalyse zur automatischen Erkennung gesunder und kranker Strukturen. Solche Strukturen sind mit bloßem Auge oft nicht zu erkennen und müssen daher in komplexen Verfahren sichtbar gemacht werden. Zwei Beispiele medizinischer Bilder werden in dieser Arbeit behandelt: *mikroskopische*, immunhistochemisch gefärbte Krebsgewebsbilder aus der Pathologie und *makroskopische* Kernspintomographiebilder des menschlichen Abdomens aus der Radiologie. In beiden Fällen ist die medizinische Aufgabe, abnormale Strukturen wie Krebs oder Morbus Crohn zu erkennen und gegebenenfalls zu quantifizieren. Wir wollen diese Aufgabe möglichst rechnergestützt und automatisiert lösen um sowohl die klinische Medizin als auch die Forschung an grossen Patientengruppen zu verbessern. Diese Arbeit ist in folgende Teile gegliedert:

Zunächst werden technisch relevante Begriffe erörtert. *Tissue Microarrays* (TMA) sind medizinische Instrumente für experimentelle Krebsforschung an großen Patientenkohorten. MRI ist ein medizinisches Bildgebungsverfahren für die Visualisierung des gastrointestinalen Trakts. Morbus Crohn (CD) und dessen Schweregrad werden erklärt.

*Technischer Hintergrund*

Acht TMA Bilder des klarzelligen Nierenzellkarzinoms und sechs Prostatakrebs TMA Bilder sind die Datenbasis für die automatische Färbeschätzung. Alle Nuklei in den Bildern sind von zwei Ärzten unabhängig voneinander annotiert. Für die Arbeit an CD Erkennung in MRI werden zwei annotierte MRI Datensätze von 27 bzw. 35 Patienten involviert.

*Datensätze*

In der klinischen Krebsforschung ist die morphologische Beurteilung von Hunderten von Gewebeproben für ein medizinisches Verständnis vieler Krebsarten nötig. Wir definieren in diesem Kapitel eine neue Methode für die rechnerbasierte Erkennung, Segmentierung, Klassifizierung und Färbeschätzung von Nuklei. Ähnlichkeitsbasierte Klassifikatoren wie (*Multiple*) *Kernel Learner* erreichen eine Genauigkeit von 83% und erreichen damit das Level der Pathologen: Die Inter-Pathologen Genauigkeit für die Erkennung von 1633 Nierenkrebs Nuklei beträgt 80%. Ausserdem beschäftigen wir uns mit *aktivem Lernen*, wodurch die Zahl der benötigten annotierten Nuklei halbiert werden kann.

*TMA Färbeschätzung*

Die automatisierte TMA Färbeschätzung ist in einem benutzerfreundlichen, freien Java Programm namens TMARKER implementiert.

*TMARKER*

Die *Einzelzellanalyse* wird in der Krebsforschung immer bedeutender. *Highly Multiplexed Mass Cytometry* ist ein neues Bildgebungsverfahren

*Zellsegmentierung*

für mehrlagige, hochkorrelierte Proteinbilder aus einem einzigen Gewebeschnitt. Wir stellen eine neue Zellsegmentierungsmethode vor, die von der gemeinsamen Information mehrerer Membranproteine profitiert.

*Automatische  
Morbus Crohn  
Erkennung in MRI*

Als ein Beispiel für makroskopische Strukturanalyse untersuchen wir in diesem Kapitel die automatische Erkennung von CD in MRI Daten. Ein automatisiertes System dafür wäre für die Forschung und klinische Anwendung vorteilhaft, da es die Verarbeitung der zunehmenden Datenmenge vereinfachen kann. Wir stellen eine neue zweistufige CD Segmentierung vor: Zuerst werden erkrankte Bereiche in den Bildern grob lokalisiert. Diese werden in einem zweiten Schritt pixelgenau segmentiert. Spezifische Bildmerkmale wie Textur-Anisotropie, räumlicher Kontext und Statistiken höherer Ordnung werden für diese Aufgabe hergeleitet. Wir erreichen durch das zweistufige Vorgehen eine CD Segmentierung mit einer Genauigkeit von 91,9% (Dice Metrik).

*Bestimmung des  
Morbus Crohn  
Schweregrads*

CD wird in der Regel nach Schweregrad eingestuft, welcher die Therapiestrategie massgeblich beeinflusst. Herkömmliche Bewertungssysteme für den Schweregrad sind aber sehr subjektiv oder invasiv. Häufig verlassen sich Ärzte daher auf mehrere Graduierungssysteme für einen ganzheitlichen Eindruck vom Patienten. Wir untersuchen daher, inwieweit MRI als Standard für eine neue CD Schweregraduierung dienen kann. Dazu werden 14 durch vier Radiologen manuell annotierte CD Merkmale mit dem endoskopischen Schweregrad (CDEIS) in Zusammenhang gebracht. Aus einer umfangreichen Regressionsanalyse resultiert ein Modell mit signifikant höherer Korrelation zum CDEIS als die der beiden bekannten Modelle MaRIA und CDA.

*Automatische  
Gewinnung von  
CD Merkmalen*

Eine Neuheit in der MRI-basierten CD Schweregradbewertung ist die Involvierung der zwei *automatisch* gewonnenen CD Merkmalen *Wanddicke* und *dynamische Kontrastverstärkung*. Die Korrelation zum CDEIS kann mit ihnen auf über 80% angehoben werden.

Am Schluss dieser Arbeit wird die Kombination der automatischen CD *Erkennung* mit der *Bewertung* des Schweregrads zu einer ganzheitlichen medizinischen Untersuchung erörtert. Die automatisch gewonnenen CD Merkmale können die CDEIS Regression deutlich verbessern und sollten auf jeden Fall in zukünftigen Studien berücksichtigt werden.



## ACKNOWLEDGEMENTS

It is a pleasure to me to thank all the people I met during my PhD, the SIMBAD project and the VIGOR++ project, without whom this work would not have been possible.

My first and highest thanks is due to my supervisor Joachim for the successful years that I could participate in the great research projects under his guidance. He gave me a lot of insights into the scientific work in the various fields of machine learning, also due to the manifold research projects in his group. He was always motivating my research with new ideas and support of my own concepts. Joachim always encouraged my interdisciplinary work with various groups inside and outside the ETH for which I am especially thankful.

*Joachim  
Buhmann*

Gladly, I want to thank my co-examiner and mentor Frans. As the scientific coordinator of the VIGOR++ project, he always had an ear for all the questions and ideas which raised in this multidisciplinary field. I appreciate the lot of discussions we had, together with our partners.

*Frans Vos*

Very special thanks go to Rita, who had always an open door for me for advice and help in a wide variety of concerns. The many questions that you get when you do a PhD at the ETH, personnel advice for the life in Zurich or scientific suggestions and coordination – she always had an answer or knew whom to ask.

*Rita Klute*

I can't express my gratitude to Thomas, my first and direct PhD mentor here at the ETH. Thomas has always motivating words for almost everything. He is very inspiring, also because of his infectious enthusiasm. I am very happy to continue his outstanding work on computational pathology. "There are *pe-ta*-bytes of data!"

*Thomas Fuchs*

I owe many thanks to my office mate Dwarika, who is an excellent colleague. I appreciate the many discussions we had not only about features, classifiers and labels, but also on religion, marriage and food.

*Dwarikanath  
Mahapatra*

Special thanks go to Peter who gave me stellar support for the work on computational pathology. Also, he was always willing to provide different kinds of data sets and manual labels. I owe a lot of experiments to his ideas and effort.

*Peter Wild*

I also appreciate the outstanding cooperation with Niels and Jan, who voluntarily labeled lots of medical microarray images during various weekends such that our experiments could have happened.

*Niels Rupp  
Jan Rüschoff*

- Cheng Soon Ong* Special thanks are due to Cheng who accompanied me throughout my time at the ETH. I stayed motivated in my research also due to our many discussions about directions, projects and collaborations.
- Ludwig Busse* Ludwig has my sincere thanks, as he was the first who reviewed my thesis and gave constructive feedback. Throughout my PhD time, we had many delightful on-topic and off-topic discussions.
- Holger Moch* I want to thank my co-examiner Holger Moch for supporting my collaboration with the USZ and reviewing my thesis. I feel honored that he is interested in my work.
- Volker Roth* Many thanks to my co-examiner Volker, as he was a great support especially in my early stage of my PhD.
- Jesica Makanyanga*  
*Jeroen Tielbeek*  
*Carl Puylaert* My earnest thanks go to Jesica, Jeroen and his fellow Carl, who extensively acquired a large patient cohort for our VIGOR++ project. Not only did they tremendous work on accurate labels on a lot of images but also took they time to discuss and explain the various details of Crohn's disease and image acquisition, at any time, besides their daily work in clinic.
- Robiel Naziroglu*  
*Zhang Li* I appreciate the meetings and discussions with my PhD fellows Robiel and Zhang, in which we together elaborated the manifold aspects of our cooperation. The automatic Crohn's disease severity assessment pipeline gains influence especially due to Robiel's and Zhang's substantial contributions on automatic feature extraction.
- Davide Soldini*  
*Simone Brandt* I want to say thank you to Davide and Simone, two great scientists at the university hospital Zurich who let me contribute to their outstanding work on lymphoma research.
- Monika Bieri*  
*Norbert Wey* I am much obliged to Monika Bieri and Norbert Wey who have been a great help during my studies at the ETH. Patiently they have arranged to transfer all the tons of image data acquired at the university hospital Zurich.
- Ruth Hüttenhain*  
*Silvia Surinova*  
*Igor Cima*  
*Ralph Schiess* I am particularly thankful to participate in so many successful cancer research projects with these wonderful people. I gladly remember the many discussions with Igor or the long Skype meetings with Silvia to evolve the best computational approaches in these large-scale prostate or colorectal cancer projects. Also together with Ruth and Ralph, we had close collaboration which gave me great insight into biological research.
- Aydın Ulaş*  
*Umberto Castellani*  
*Manuele Bicego*  
*Mehmet Gönen* The time we spent in Italy and Zurich together for various SIMBAD meetings was always very enjoyable. I want to thank Aydın, Umberto, Manuele and Mehmet for their outstanding contribution and constructive collaboration. Many projects and papers would not have been possible without their excellent effort and support.

Friederike was the first beta tester for our implemented cell nucleus counting program TMARKER. I appreciate her constructive feedback very much, which contributed to the further developments of the program. *Friederike Böhm*

Marcello had the great idea of a scientific book as a crowned closing of the SIMBAD project. I feel honored that I could contribute to “Similarity-Based Pattern Analysis and Recognition”. *Marcello Pelillo*

I owe particular gratitude to Verena, as she was an excellent colleague in our group. Thanks to her, I live today in a wonderful flat on top of town. *Verena Kaynig-Fittkau*

Endless special thanks go to Reto, my dear friend, who always helped me in any concerns. I gladly remember hundreds of hours of remote computer work with his provided mobile internet access. Also, he showed me fantastic places for excellent food and relax after long working days. *Reto Weber*

My deepest gratitude is due to my family for their exceptional support of my decision to study in Switzerland and for their many delightful and cordial visits in Zurich. I am happy to welcome my new niece Anthea and my new nephew Hendrik in my family. *Family*



# CONTENTS

<b>1</b>	<b>Introduction .....</b>	<b>1</b>
1.1	Thesis Structure .....	2
1.2	Scientific Challenges .....	3
1.2.1	Image Interpretation.....	3
1.2.2	Data Fusion .....	3
1.2.3	Lack of Ground Truth .....	4
1.2.4	Clinical Availability and Implementation.....	4
1.3	Social Benefit .....	4
1.4	Interdisciplinary Endeavor .....	5
1.5	Original Contributions .....	5
1.5.1	Extensions of Computational Pathology .....	5
1.5.2	Similarity Based Classification.....	5
1.5.3	Shape Importance for Nucleus Classification.....	6
1.5.4	TMARKER: Implementation of Staining Estimation Pipeline .....	6
1.5.5	Crohn's Disease Detection in MRI.....	6
1.5.6	General Feature Selection Pipeline.....	7
1.5.7	Automated MRI CD Severity Assessment .....	7
1.5.8	Combination of Multiple Data Dimensions.....	7
1.5.9	Combination of Different Data Domains .....	8
1.6	List of Own Publications .....	8
<b>2</b>	<b>Technical Background.....</b>	<b>13</b>
2.1	Tissue Microarray (TMA).....	13
2.2	Magnetic Resonance Imaging (MRI) .....	14
2.2.1	T <sub>1</sub> -weighted MRI (Spin-Lattice Relaxation).....	14
2.2.2	T <sub>2</sub> -weighted MRI (Spin-Spin Relaxation) .....	15
2.2.3	Sequences .....	15
2.2.4	Alternative Techniques .....	16
2.3	Crohn's Disease .....	17
2.3.1	Causes .....	18
2.3.2	Symptoms.....	18
2.3.3	Therapy.....	19
2.3.4	Diagnosis and Monitoring .....	20
2.4	Crohn's Disease Severity .....	20
2.4.1	CDEIS (Endoscopic Index of Severity).....	21
2.4.2	AIS and eAIS (endoscopic Acute Inflammation Score)....	21
2.4.3	CDAI (Crohn's Disease Activity Index).....	22
2.4.4	HBI (Harvey Bradshaw Index).....	24
2.4.5	MaRIA (Magnetic Resonance Index of Activity).....	24
2.4.6	CDA (Crohn's Disease Activity).....	25

2.4.7 CRP (C-Reactive Protein).....	25
2.4.8 Calprotectin .....	25
2.4.9 Further Indices .....	25
<b>3 Datasets.....</b>	<b>27</b>
3.1 Renal Cell Carcinoma TMA Dataset.....	27
3.2 Prostate Cancer TMA Dataset .....	29
3.3 Retrospective Crohn’s Disease MRI dataset.....	29
3.3.1 MRI Protocol for CD patients.....	30
3.3.2 MRI Feature Assessment by Four Radiologists.....	31
3.3.3 Pseudo-Polyps, Node enhancement, Lymph Nodes, Mural Thickness and Edema.....	34
3.3.4 CDEIS Assessment by a Medical Doctor.....	35
3.3.5 Manual CD Segmentation by One Radiologist .....	35
3.3.6 Retrospective Dataset Statistics .....	36
3.4 Prospective Crohn’s Disease MRI Dataset.....	41
3.4.1 MRI Feature Assessment by Two Radiologists .....	42
3.4.2 CDEIS Assessment by Two Medical Doctors .....	42
3.4.3 Prospective Dataset Statistics .....	43
<b>4 TMA Staining Estimation Pipeline.....</b>	<b>45</b>
4.1 Structure .....	47
4.2 Nucleus Detection .....	47
4.2.1 Color Deconvolution Based Nucleus Detection.....	47
4.2.2 Superpixel Based Nucleus Detection.....	50
4.3 Nucleus Segmentation via Graph-Cut .....	55
4.4 Feature Extraction .....	58
4.4.1 Histogram of Patch Intensity (ALL).....	58
4.4.2 Color Histogram (COL) .....	58
4.4.3 Histogram of Foreground Intensity (FG) .....	58
4.4.4 Histogram of Background Intensity (BG) .....	58
4.4.5 Freeman Chain Code (FCC) .....	58
4.4.6 1D-Signature (SIG).....	58
4.4.7 Pyramid Histogram of Oriented Gradients (PHOG).....	59
4.4.8 Region Properties (PROP) .....	59
4.4.9 Local Binary Patterns (LBP).....	59
4.5 Classification with Support Vector Machines.....	59
4.5.1 Experimental Design .....	62
4.5.2 Results .....	62
4.6 Shape Descriptors Boost the Classification Performance.....	63
4.7 Better Nucleus Classification for Better Staining Estimation....	65
4.7.1 Experimental Design .....	65
4.7.2 Results .....	66
4.8 Multiple Kernel Learning for Nucleus Classification .....	67

---

4.8.1	Introduction .....	67
4.8.2	MKL Framework.....	67
4.8.3	Experiments and Results.....	69
4.8.4	Discussion .....	71
4.8.5	Conclusion .....	72
4.9	Nonlinear Data Combination for Nucleus Classification.....	72
4.9.1	MKL .....	72
4.9.2	Linear MKL .....	73
4.9.3	Nonlinear MKL Framework.....	73
4.9.4	Experimental Design .....	75
4.9.5	Results.....	75
4.9.6	Discussion .....	76
4.9.7	Conclusion .....	76
4.10	Active Learning for Nucleus Classification.....	77
4.10.1	Random Forests as Probabilistic Classifier Ensemble.	77
4.10.2	Experiments and Results.....	78
4.11	Survival Analysis.....	80
4.11.1	Results .....	80
<b>5</b>	<b>TMARKER: A Free Software Toolkit For Staining Estimation ....</b>	<b>83</b>
5.1	Implementation Details .....	84
5.2	Usage Statistics.....	86
5.3	Conclusion.....	86
<b>6</b>	<b>Single Cell Segmentation on Highly Multiplexed Images.....</b>	<b>87</b>
6.1	Introduction.....	87
6.2	Methods .....	89
6.2.1	Image Acquisition .....	89
6.2.2	Single Cell Segmentation .....	89
6.2.3	Segmentation Score.....	92
6.3	Results .....	93
6.3.1	Mutual Information between Membrane Proteins .....	93
6.3.2	Cell Segmentation .....	94
6.3.3	Implementation .....	96
6.4	Discussion.....	96
6.5	Conclusion.....	97
<b>7</b>	<b>Shape Features for Automatic Crohn’s Disease Detection and Segmentation .....</b>	<b>99</b>
7.1	Introduction.....	99
7.2	Voxel Based Classification.....	100
7.2.1	Feature extraction.....	101
7.2.2	Comparison Features.....	103
7.2.3	Dataset .....	105
7.2.4	Classification Scenario.....	105

---

7.2.5 Results – 1 <sup>st</sup> Stage: Intestine vs. Background .....	107
7.2.6 Single Feature Contribution .....	109
7.2.7 Results – 2 <sup>nd</sup> Stage: Diseased vs. Normal Intestine .....	109
7.2.8 Whole Patient Classification.....	111
7.3 Automated Preprocessing and Post-Processing .....	115
7.4 Novel Context Features Refine Automatic CD Detection .....	117
7.5 Supravoxels for VOI Detection .....	118
7.5.1 Effect of Supravoxel Size.....	119
<b>8 A Model Development Pipeline for Crohn’s Disease Severity Assessment.....</b>	<b>121</b>
8.1 CD Analysis Pipeline with Manually Read Features.....	121
8.1.1 Feature Extraction.....	121
8.1.2 Feature Selection and Model Training.....	122
8.1.3 Results .....	123
8.1.4 First Order Statistics for Feature Selection .....	124
8.1.5 Second Order Statistics for Feature Selection .....	127
8.1.6 A Heuristic Approach .....	130
8.1.7 Biological Constraints for Model Selection .....	131
8.1.8 Validation of Models on Two Data Sets .....	132
<b>9 Automated Feature Extraction Methods Improve CD Severity Assessment.....</b>	<b>137</b>
9.1 Low Mutual Information between MRI and CDEIS .....	138
9.2 Automated Bowel Wall Thickness (ABWT).....	140
9.2.1 Retrospective Dataset Expansion .....	140
9.2.2 ABWT Corresponds to Manual Wall Thickness Scoring.....	140
9.2.3 Does ABWT Increase the Correlation to CDEIS? .....	144
9.3 Automated Dynamic Contrast Enhancement (DCE) .....	148
9.3.1 DCE as Feature .....	148
9.3.2 Dataset Expansion .....	150
9.3.3 DCE Contributes to CD Severity Assessment .....	151
9.3.4 DCE and Manually Annotated Diseased Regions .....	154
9.4 Combining ABWT and DCE.....	156
<b>10 Conclusion.....</b>	<b>157</b>
10.1 Computational Pathology .....	157
10.1.1 Nucleus Classification .....	157
10.1.2 Nucleus Detection .....	158
10.2 Computational Radiology .....	158
10.2.1 CD Severity Score.....	158
10.2.2 Multimodal Features for Severity Assessment .....	158
10.2.3 Exhaustive Search vs. Heuristic .....	159
10.2.4 CD Detection with MRI.....	159
10.2.5 Combining CD Detection and Severity – Outlook....	159



10.3 Computer Aided CD Assessment with MRI – Outlook ..... 160

**References ..... 163**

**Abbreviations ..... 179**



# 1 INTRODUCTION

Medical imaging comprises all techniques which allow to visualize hidden parts of the body for clinical purposes or medical research. Several techniques that have been developed over the last decades are known for this task and they substantially improve modern medical decision making as many morbid alterations of the body happen inside, invisible to every naked human eye. One prominent example is abdominal magnet resonance imaging (MRI) in radiology which allows a macroscopic insight into the abdominal part of the body. A second common example are tissue microarrays (TMA) in pathology which allow a microscopic magnification of human cells to visualize cell alterations. Such medical images are widely used for identification, diagnosis and grading of various types of diseases, disorders or cancers. While the *image acquisition* of e.g. MRI is more and more computer aided, the *image interpretation* with a subsequent decision making process depends on visual inspection by the attending physician, commonly of multiple images derived from several scan *sequences*.

In the advent of improved and cheaper imaging techniques, the *computational medical imaging* is more and more in focus of scientific research. Therein, the image analysis is driven by automatic image processing, computer vision and machine learning algorithms supporting and automating the manifold processes involved such as e.g. disease or cancer detection, object segmentation and object classification. Machine learning algorithms further play more and more a role in medical research when it comes to feature selection for modeling complex biological systems.

This thesis is largely motivated by the improvement of existing and the development of new computational image analysis algorithms for abdominal MRI and TMA images. Two multidisciplinary research projects from the European Community's 7<sup>th</sup> Framework Program constitute a strong support for this thesis: the **SIMBAD** project (grant no. 213250) and the **VIGOR++** project (grant no. 270379). Both projects partially focus on the exploration of machine learning approaches for specific problems in recent medical research.

In the context of **SIMBAD** ("Similarity Based Pattern Analysis and Recognition"), we study the advantage of similarity based classification approaches for *renal clear cell nucleus classification*. Similarity based classification allows the design of individual distance measures between

*SIMBAD*  
founded in 2008.  
[www.simbad-fp7.eu](http://www.simbad-fp7.eu)

objects which are tailored to the underlying problem. Several classification systems based on support vector machines dedicated to exploit such individual similarity measures are investigated.

VIGOR++  
started in 2011.  
[www.vigorpp.eu](http://www.vigorpp.eu)

More medically driven, **VIGOR++** (“Virtual Gastrointestinal Tract”) pioneers in exploring the multifaceted field of *automatic Crohn’s disease assessment in abdominal MRI* (Tielbeek *et al.* 2012). The demand of automated evaluation methods for Crohn’s disease (CD) is motivated by the fact that only moderate inter-observer agreement for radiologic severity measures are reported (Vos *et al.* 2012). Two essential parts of this research field are covered by this thesis: the computational *CD detection/segmentation* on MRI, and the automatic endoscopic *CD severity estimation*.

## 1.1 Thesis Structure

After this introduction of the thesis, we will shortly outline the underlying medical and technical background of TMA, MRI and CD in **chapter 2**. Subsequently, the medical datasets comprising TMA of renal clear cell carcinoma patients and prostate cancer patients as well as the MRI scans from CD patients are introduced in **chapter 3**. **Chapter 4** then proposes our new processing pipeline for computational staining estimation on immunohistochemically stained tissue microarrays. The cell nucleus classification within this pipeline is exhaustively studied with similarity based approaches. Also, the importance of shape features is outlined. Further, we demonstrate on a simplistic approach the advantage of active learning for cell nucleus classification for medical research. The number of training labels can be reduced when active learning for labeling is incorporated. An implementation of some of the proposed algorithms as a freely available Java program TMARKER is reported in **chapter 5**. Thereafter in **chapter 6**, we consider whole cell segmentation based on watersheds in highly multiplexed TMA images. This watershed based method differs from the previously studied nucleus segmentation as it can incorporate the multidimensional information of various, highly aligned membrane proteins. **Chapter 7** investigates the use of shape features for Crohn’s disease detection and segmentation in MR images, which highlights the importance of morphological shape features in various problems in computational medical imaging. **Chapter 8** introduces a feature selection pipeline based on exhaustive search for the CD severity assessment in MR images. This pipeline is used for the development of a severity score which incorporates computer-read CD related MRI images for the first time in such a model in **chapter 9**. **Chapter 10** finally critically concludes the work with an outlook to possible future research directions.

## 1.2 Scientific Challenges

In various medical imaging problems, a common ultimate goal can be formulated as **to identify qualitatively and/or quantitatively clinically relevant objects in a medical image with as little user input as possible**. E.g. for VIGOR++, this means on the one hand to detect and segment CD in MR images and on the other hand to quantify the severity of CD as a clinically relevant factor. For computational pathology however, the detection and classification of cell nuclei is a scientific problem. Several scientific challenges come together with the aforementioned formulation, which we will explain in the following sections.

### 1.2.1 Image Interpretation

Medical images differ from natural images in several aspects. They might originate from artificial image source, e.g. magnetic resonance signals, microscopic light sources. MRI 3D scans are computationally reconstructed from a series of 1D signals and consequently differ according to the applied normalization. Imaging techniques have individual difficulties such as examination artifacts (e.g. poor bowel distention, air or fecal remains in the bowel, motion or breathing artefacts) for MRI or experimental sample preparation for TMA. As a consequence, the interpretation of images might largely be driven by the experience of the physician.

The computational image interpretation of CD in MRI is extremely difficult due to the high variability of the non-rigid anatomic structure of the bowel. Hence, to the best of our knowledge, no study exists which addresses the computational CD detection and segmentation in MRI. We examine the image interpretation by human domain experts as well as by computer vision methods. Four radiologists have exhaustively examined the MRI scans of 30 patients and screened for 17 different signs of CD. We show that such images can be sufficient for a consistent CD assessment among independent experts and that the proposed features can be interpreted as endoscopic severity index. On the other hand, we investigate the scientific question whether modern computer vision methods and machine learning algorithms are able to detect the signs of CD in the same images in a standardized manner.

### 1.2.2 Data Fusion

Scientifically very interesting is the advent of multiple data domains in a single research problem. As Crohn's disease is a multifaceted disorder with intra- and extra-intestinal manifestations, its extensive diagnosis and severity assessment commonly includes patient's medical history, physical examinations, clinical biomarkers, endoscopy, MRI examination

and pathological studies. The fusion of these different data domains poses itself a major challenge of the project. Especially the definition of an adequate reference standard for validation is important.

### 1.2.3 *Lack of Ground Truth*

As well for a supervised analysis as for educated model validation, the lack of ground truth is a challenging problem which commonly occurs in a wide range of natural sciences. Computer algorithms have to be designed to account for this problem. In our medical research, we incorporate the *gold standard* by multiple experts to estimate the variance of the underlying dataset and thus the “difficulty” of the problem as well as to train classifiers based on a larger range of data variability.

### 1.2.4 *Clinical Availability and Implementation*

Computer science has to provide the capacity to store and process an increasing amount of data with parallel computing. Databases have to be created which tolerate the sovereignty of the hospitals and research institutes for the medical datasets. Further, the developed scientific algorithms have to be implemented and designed for practical usability. While the VIGOR++ project incorporates a professional partner (Biotronics3D) for this task, we contribute with an own Java implementation of the proposed cancer cell nucleus classification in the SIMBAD project.

## 1.3 *Social Benefit*

This thesis largely contributes to a social benefit in several manners. The improvement of cancer cell nucleus classification will increase the accuracy and confidence in computational analysis methods by medical research. Computational pathology will not only improve the treatment of every single patient by standardized and reproducible medical decisions. It further facilitates the creation of larger patient cohorts and research projects, too, by decreased time and costs of specimen evaluation.

Further, this study explores automatic CD severity assessment and CD localization on MRI for the first time. This drastically facilitates the individualized medicine for this type of disease. Physicians can benefit from computer-driven techniques for the automatic CD assessment in MRI. These algorithms also autonomously propose standardized and patient-specific decision support which only has to be reviewed by the medical doctor instead of completely surveyed from the raw data.

## 1.4 Interdisciplinary Endeavor

This thesis clearly accrues from the fruitful collaboration of scientists from various domains. Medical doctors with a deep and experienced understanding of the underlying medical problems, radiologists with the technical and medical expertise in abdominal MRI analysis and interpretation, pathologists with their medical expertise for cell nucleus classification and description in various types of cancer tissue, computer scientists from image processing, computer vision, machine learning, visualization and bioinformatics with their knowledge of computational research in medical sciences and information scientists from software and business companies with the experience to bring new technologies into clinic have to cooperate for a successful contribution. The excellent collaboration and frequent communication of scientists and experts from various areas is a central characteristic of this PhD work.

## 1.5 Original Contributions

Motivated by the aforementioned scientific problems in the medical imaging domain, this thesis studies following computer science and machine learning approaches with the specified contributions:

### 1.5.1 Extensions of Computational Pathology

We continue the scientific investigation of **computational pathology**, a research field that has been defined by Fuchs and Buhmann (2011b) as “*the investigation of a complete probabilistic treatment of scientific and clinical workflows in general pathology, combining experimental design, statistical pattern recognition and survival analysis in a unified framework to answer scientific and clinical questions in pathology*”. Fuchs *et al.* (2008a; 2011b) systematically developed and validated a computational pathology workflow for the automated cancer cell nucleus detection and staining estimation of renal clear cell carcinoma TMA specimen with subsequent survival analysis. We extend this workflow by partitioning it into the subsequent steps (i) *nucleus detection*, (ii) *nucleus segmentation*, (iii) *nucleus classification* and (iv) *staining estimation*. This workflow allows the detailed investigation of the single steps and their influence on the whole **new TMA analysis pipeline**. The pipeline is validated on renal clear cell carcinoma and prostate carcinoma data.

Chapter 4

### 1.5.2 Similarity Based Classification

Within the new staining estimation pipeline, we treat **similarity based classification** approaches for the distinction of malignant and benign cell

Section 4.5

nuclei by the introduction of new multiple kernel learning algorithms which combine the information of specially designed features in a classification ensemble. This classification ensemble is favorable compared to single support vector machines. 15 kernel and distance functions are included in the study.

### 1.5.3 *Shape Importance for Nucleus Classification*

*Section 4.6* We further introduce dedicated **shape measurements** for cell nucleus classification. Shape is an important descriptor of nucleus characteristics which influences the design of a nucleus detection and classification pipeline. We quantify the influence of shape descriptors in cell nucleus classification on the example of renal clear cell carcinoma. The information of shape is statistically discriminative for classifying nuclei as malignant or benign. Although this hypothesis is already known in pathology, there is no study which quantifies the information gain.

### 1.5.4 *TMARKER: Implementation of Staining Estimation Pipeline*

*Chapter 5* We introduce the new and freely available software package **TMARKER** which is tailored to automatic nuclear staining estimation of immunohistochemically stained TMA images. We implement our developed modern machine learning algorithms in the platform independent programming language Java. The algorithms are validated on renal clear cell carcinoma images and prostate cancer images in terms of accuracy, precision, recall and survival grouping. The program is developed for scientific and clinical use and allows the unspecific cell nucleus detection and counting as well as the comprehensive nucleus classification and staining estimation with subsequent survival analysis. TMARKER is available on [www.comp-path.inf.ethz.ch](http://www.comp-path.inf.ethz.ch).

### 1.5.5 *Crohn's Disease Detection in MRI*

*Chapter 7* To the best of our knowledge, no study exist which tried to solve the difficult task of **fully automated CD detection and segmentation in abdominal MR** images. The difficulty of this task lies in the high variance in image signal inside and outside the bowel, the low resolution compared to endoscopic camera images and the lack of a ground truth which we encounter with an appropriate gold standard. We firstly describe a framework for automated Crohn's disease detection in MR images based on standard image features, textures, shape, and context information.



### 1.5.6 *General Feature Selection Pipeline*

We propose a new pipeline for the exhaustive and elaborate feature and model selection for the quantification of Crohn's disease severity. The pipeline comprises combinatorial feature selection with exhaustive search for linear regression models quantifying endoscopic disease severity and combining different data sources. We show the generality of the pipeline applying it on different problems, e.g. Crohn's disease severity estimation based on MR image features, and cancer diagnosis prediction based on protein feature selection.

Chapter 8

### 1.5.7 *Automated MRI CD Severity Assessment*

One principal idea of this thesis is to automate the CD severity assessment procedure for MRI examinations by computational support with standardized and validated algorithms. Although a lot of research is ongoing concerning the detection and segmentation of rigid organs in MRI (e.g. liver segmentation (Masoumi *et al.* 2012), kidney segmentation (Zollner *et al.* 2012) or heart segmentation (Petitjean and Dacher 2011)), the exploration and development of automatic methods for CD detection, segmentation and severity estimation in MRI is completely new. The flexible anatomic structure of the bowel, the non-localized and highly variant phenotype of CD and the high variability of MRI scan quality complicate the computer-driven interpretation of the images. We show that it is not only possible to automatically extract severity related image features such as wall thickness or dynamic contrast enhancement, but also that these **newly measured quantities** arising from the VIGOR++ project qualify for an **enhanced severity prediction** for Crohn's disease patients.

Chapter 9

### 1.5.8 *Combination of Multiple Data Dimensions*

Many biological classification problems in medical tasks show a highly complex environment which complicates the choice of appropriate features. In fact, the classification of a medical object can depend on numerous different features. Often, the incorporation of different data dimensions requires a prior registration of the data sources (e.g. image registration). We show that the registration of different protein expression images improves the proper single cell segmentation on breast cancer tissue microarray images. A new technique called *Highly Multiplexed Mass Cytometry* allows for the simultaneous quantitative co-localized expression scan of dozens of proteins in tissue microarrays. The highly registered information is used for improved cell segmentation.

Chapter 6

### 1.5.9 Combination of Different Data Domains

Chapters 8-9

Extending the idea of using a variety of data dimensions for a given learning task, one scientific challenge of recent medical research the connection of different data sources to constitute an even higher dimensional, but more holistic view of the patient and to gain information from the mutual interactions of the different data sources. We contribute to this research by illustrating how Crohn's disease severity information from MR images, clinical data, colonoscopy, and computer vision can be fused to form a holistic severity measure superior to any single sensor modality.

## 1.6 List of Own Publications

Following publications are fully or partly covered in this thesis:

1. P.J. Schüffler, D. Schapiro, C. Giesen, H.a.O. Wang, B. Bodenmiller and J.M. Buhmann. **Single Cell Segmentation with Watersheds on Highly Multiplexed Images**. In submission. (2014)
2. P.J. Schüffler, T.J. Fuchs, C.S. Ong, P.J. Wild, N.J. Rupp and J.M. Buhmann. **TMARKER: A Free Software Toolkit for Histopathological Cell Counting and Staining Estimation**. *Journal of Pathology Informatics*, 4(2). (2013) DOI: 10.4103/2153-3539.109804
3. P.J. Schüffler, N.J. Rupp, C.S. Ong, J.M. Buhmann, T.J. Fuchs and P.J. Wild. **TMARKER: A Robust and Free Software Toolkit for Histopathological Cell Counting and Immunohistochemical Staining Estimation**. *Der Pathologe*, 34: p. 30. (2013)
4. P.J. Schüffler, D. Mahapatra, J.a.W. Tielbeek, F.M. Vos, J. Makanyanga, D.A. Pendsé, C.Y. Nio, J. Stoker, S.A. Taylor and J.M. Buhmann. **A Model Development Pipeline for Crohn's Disease Severity Assessment from Magnetic Resonance Images**. *Abdominal Imaging. Computation and Clinical Applications*, H. Yoshida, S. Warfield and M. Vannier, Editors. Springer Berlin Heidelberg. p. 1-10. (2013) DOI: 10.1007/978-3-642-41083-3\_1
5. P.J. Schüffler, T.J. Fuchs, C.S. Ong, V. Roth and J.M. Buhmann. **Automated Analysis of Tissue Micro-Array Images on the Example of Renal Cell Carcinoma**. *Similarity-Based Pattern Analysis and Recognition*, M. Pelillo, Editor. Springer London. p. 219-45. (2013) DOI: 10.1007/978-1-4471-5628-4\_9
6. P. Schüffler, A. Ulaş, U. Castellani and V. Murino. **A Multiple Kernel Learning Algorithm for Cell Nucleus Classification of Renal Cell**

Outstanding  
Paper Award

- Carcinoma. Image Analysis and Processing – ICIAP**, G. Maino and G. Foresti, Editors. Springer Berlin Heidelberg. p. 413-22. (2011) DOI: 10.1007/978-3-642-24085-0\_43
7. P.J. Schüffler, T.J. Fuchs, C.S. Ong, V. Roth and J.M. Buhmann. **Computational TMA Analysis and Cell Nucleus Classification of Renal Cell Carcinoma**. *Proceedings of the 32nd DAGM Conference on Pattern Recognition*. Springer-Verlag Berlin: Darmstadt, Germany. p. 202-11. (2010) DOI: 10.1007/978-3-642-15986-2\_21
  8. C. Giesen, H.a.O. Wang, D. Schapiro, N. Zivanovic, A. Jacobs, B. Hattendorf, P.J. Schüffler, D. Grolimund, J.M. Buhmann, S. Brandt, Z. Varga, P.J. Wild, D. Günther and B. Bodenmiller. **Highly Multiplexed Imaging of Tumor Tissues with Subcellular Resolution by Mass Cytometry**. *Nature Methods*, 11(4): p. 417-22. (2014) DOI: 10.1038/nmeth.2869
  9. D. Mahapatra, P.J. Schüffler, J.a.W. Tielbeek, J.M. Buhmann and F.M. Vos. **A Supervised Learning Approach for Crohn's Disease Detection Using Higher-Order Image Statistics and a Novel Shape Asymmetry Measure**. *Journal of Digital Imaging*, 26(5): p. 920-31. (2013) DOI: 10.1007/s10278-013-9576-9
  10. D. Mahapatra, P.J. Schüffler, J.a.W. Tielbeek, J.C. Makanyanga, J. Stoker, S.A. Taylor, F.M. Vos and J.M. Buhmann. **Automatic Detection and Segmentation of Crohn's Disease Tissues from Abdominal MRI**. *IEEE Transactions on Medical Imaging*, 32(12): p. 2332-47. (2013) DOI: 10.1109/TMI.2013.2282124
  11. D. Mahapatra, P.J. Schüffler, J.a.W. Tielbeek, F.M. Vos and J.M. Buhmann. **Localizing and Segmenting Crohn's Disease Affected Regions in Abdominal MRI Using Novel Context Features**. *Proc. SPIE, Medical Imaging 2013: Image Processing*, 8669. (2013) DOI: 10.1117/12.2006698
  12. D. Mahapatra, P. Schüffler, J.W. Tielbeek, F. Vos and J. Buhmann. **Semi-Supervised and Active Learning for Automatic Segmentation of Crohn's Disease**. *Medical Image Computing and Computer-Assisted Intervention*, K. Mori, I. Sakuma, Y. Sato, C. Barillot and N. Navab, Editors. Springer Berlin Heidelberg. p. 214-21. (2013) DOI: 10.1007/978-3-642-40763-5\_27
  13. D. Mahapatra, P.J. Schüffler, J. Tielbeek, F.M. Vos and J.M. Buhmann. **Crohn's Disease Tissue Segmentation from Abdominal MRI Using Semantic Information and Graph Cuts**. *IEEE 10th International Symposium on Biomedical Imaging*. San Francisco. (2013) DOI: 10.1109/ISBI.2013.6556486

14. D. Mahapatra, A. Vezhnevets, P.J. Schüffler, J.a.W. Tielbeek, F.M. Vos and J.M. Buhmann. **Weakly Supervised Semantic Segmentation of Crohn's Disease Tissues from Abdominal MRI.** *IEEE 10th International Symposium on Biomedical Imaging.* (2013) DOI: 10.1109/ISBI.2013.6556607
15. D. Soldini, C. Montagna, P. Schüffler, V. Martin, A. Georgis, T. Thiesler, A. Curioni-Fontecedro, P. Went, G. Bosshard, S. Dehler, L. Mazzuchelli and M. Tinguely. **A New Diagnostic Algorithm for Burkitt and Diffuse Large B-Cell Lymphomas Based on the Expression of Cse11 and Stat3 and on Myc Rearrangement Predicts Outcome.** *Annals of Oncology*, 24(1): p. 193-201. (2013) DOI: 10.1093/annonc/mds209
16. D. Mahapatra, P.J. Schüffler, J.a.W. Tielbeek, J.M. Buhmann and F.M. Vos. **A Supervised Learning Based Approach to Detect Crohn's Disease in Abdominal MR Volumes.** *Abdominal Imaging. Computational and Clinical Applications*, H. Yoshida, D. Hawkes and M. Vannier, Editors. Springer Berlin Heidelberg. p. 97-106. (2012) DOI: 10.1007/978-3-642-33612-6\_11
17. F.M. Vos, J.a.W. Tielbeek, R.E. Naziroglu, L. Zhang, P.J. Schüffler, D. Mahapatra, A. Wiebel, C. Lavini, J.M. Buhmann, H. Hege, J. Stoker and L.J. Van Vliet. **Computational Modeling for Assessment of IBD: To Be or Not to Be?** *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE.* (2012) DOI: 10.1109/EMBC.2012.6346837
18. I. Cima, R. Schiess, P. Wild, M. Kaelin, P. Schüffler, V. Lange, P. Picotti, R. Ossola, A. Templeton, O. Schubert, T. Fuchs, T. Leippold, S. Wyler, J. Zehetner, W. Jochum, J. Buhmann, T. Cerny, H. Moch, S. Gillessen, R. Aebersold and W. Krek. **Cancer Genetics-Guided Discovery of Serum Biomarker Signatures for Diagnosis and Prognosis of Prostate Cancer.** *Proceedings of the National Academy of Sciences of the USA*, 108(8): p. 3342-7. (2011) DOI: 10.1073/pnas.1013699108
19. M. Gönen, A. Ulaş, P. Schüffler, U. Castellani and V. Murino. **Combining Data Sources Nonlinearly for Cell Nucleus Classification of Renal Cell Carcinoma.** *Similarity-Based Pattern Recognition*, M. Pelillo and E. Hancock, Editors. Springer Berlin Heidelberg. p. 250-60. (2011) DOI: 10.1007/978-3-642-24471-1\_18

**Following publications are not included in this thesis as they do not cover the material here:**

20. S. Brandt, C. Montagna, A. Georgis, P. Schüffler, M. Buhler, B. Seifert, T. Thiesler, A. Curioni-Fontecedro, I. Hegyi, S. Dehler, V. Martin, M. Tinguely and D. Soldini. **The Combined Expression of the Stromal Markers Fibronectin and Sparc Improves the Prediction of Survival in Diffuse Large B-Cell Lymphoma.** *Experimental Hematology & Oncology*, 2(1): p. 27. (2013) DOI: 10.1186/2162-3619-2-27
21. M.R. Ghigna, T. Reineke, P. Rince, P. Schüffler, B. El Mchichi, M. Fabre, E. Jacquemin, A. Durrbach, D. Samuel, I. Joab, C. Guettier, M. Lucioni, M. Paulli, M. Tinguely and M. Raphael. **Epstein-Barr Virus Infection and Altered Control of Apoptotic Pathways in Posttransplant Lymphoproliferative Disorders.** *Pathobiology*, 80(2): p. 53-9. (2012) DOI: 10.1159/000339722
22. J. Veeck, P.J. Wild, T. Fuchs, P.J. Schüffler, A. Hartmann, R. Knuchel and E. Dahl. **Prognostic Relevance of Wnt-Inhibitory Factor-1 (Wif1) and Dickkopf-3 (Dkk3) Promoter Methylation in Human Breast Cancer.** *BMC Cancer*, 9: p. 217. (2009) DOI: 10.1186/1471-2407-9-217
23. M. Bicego, A. Ulaş, P.J. Schüffler, U. Castellani, V. Murino, A. Martins, P. Aguiar and M. Figueiredo. **Renal Cancer Cell Classification Using Generative Embeddings and Information Theoretic Kernels.** *Pattern Recognition in Bioinformatics*, M. Loog, L. Wessels, M.T. Reinders and D. Ridder, Editors. Springer Berlin Heidelberg. p. 75-86. (2011) DOI: 10.1007/978-3-642-24855-9\_7
24. A. Ulaş, P.J. Schüffler, M. Bicego, U. Castellani and V. Murino. **Hybrid Generative-Discriminative Nucleus Classification of Renal Cell Carcinoma.** *Similarity-Based Pattern Recognition*, M. Pelillo and E. Hancock, Editors. Springer Berlin Heidelberg. p. 77-89. (2011) DOI: 10.1007/978-3-642-24471-1\_6



## 2 TECHNICAL BACKGROUND

### 2.1 *Tissue Microarray (TMA)*

Cancer research in pathology frequently concerns the quantitative protein expression rate of nuclear expressed proteins in cancer tissue samples. Tissue microarrays (TMA) constitute a common diagnostic tool for this target (Meyer *et al.* 2010; Meyer *et al.* 2012). The preparation of a TMA is simplified on the following example of renal clear cell carcinoma, and different variations of that protocol are known:

- A tissue biopsy of affected renal tissue is taken from a renal clear cell carcinoma (RCC) patient and formalin-fixed and paraffin embedded for conservation.
- On the biopsy specimen, the cancer site is localized and punched out with an inflation needle of 0.6 mm in diameter. The extracted tissue cylinder is transferred to a separate, empty paraffin block.
- Repeating this procedure with biopsies from several patients, the new paraffin array can carry dozens to hundreds of small tissue samples.
- A thin slice of approximately 2-50  $\mu\text{m}$  is abraded from the array with a microtome and fixed on a glass plate.
- In an immunohistochemical assay, the slice is exposed to MIB-1 monoclonal antibodies which specifically bind to the nuclear proliferation protein Ki-67, unspecific to cancer cells or normal cells. The antibodies are linked to the chromatic enzyme peroxidase. Unbound antibodies are washed out.
- The chromogenic substrate 3,3'-Diaminobenzidine is incubated to the slice and will be processed by bound peroxidase to a brown reaction product.
- A bluish hematoxylin counterstain reveals the morphological structure of the tissue for better visibility.

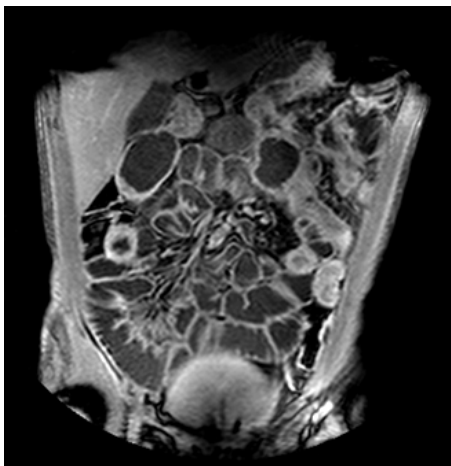
The resulting *immunohistochemically (IHC) stained* TMA is then manually examined under the light microscope, and cancerous cell nuclei which express protein are counted. The crucial advantage of TMA is the simultaneous preparation of samples of a whole patient cohort under equal experimental conditions, such as temperature, incubation times, pH levels, salt concentrations and other factors, which immediately influence the quality of staining.

Our used TMA datasets of RCC tissue and prostate cancer (PCa) tissue are introduced in section 3.1 and 3.2. Both datasets are MIB-1 stained as explained above against Ki-67, a nuclear proliferation protein. The presence of this protein indicates cell proliferation which is an unwanted behavior of tumor cells. The estimation of the fraction of proliferating cancerous cells, called *staining estimation*, is a standard measure for various types of cancer and is commonly related to the cancer prognosis and estimated survival of patients.

## 2.2 Magnetic Resonance Imaging (MRI)

Magnetic resonance Imaging (MRI) is a non-invasive medical imaging method to visualize inner parts of the body. The technique is based on magnetic fields of strengths 1-7 T in which atomic nuclei show their resonant behavior to react on electromagnetic radiation. The MRI scanner detects these resonances as induced electricity and calculates the resulting MR signal image. Especially hydrogen atoms show a distinct resonance behavior. Since the time to reach the thermodynamic equilibrium of the stimulated nuclei differs depending on the chemical and physiological structure and environment (and thus in different tissues, fluids, etc.), the measured induced electromagnetic signals differ which will be visible in the reconstructed signal images as different intensities. Two possible “time-signals” are explained here:  $T_1$ -weighted MRI and  $T_2$ -weighted MRI.

### 2.2.1 $T_1$ -weighted MRI (Spin-Lattice Relaxation)



**Figure 2.1:  $T_1$ -weighted image (coronal THRIVE).**

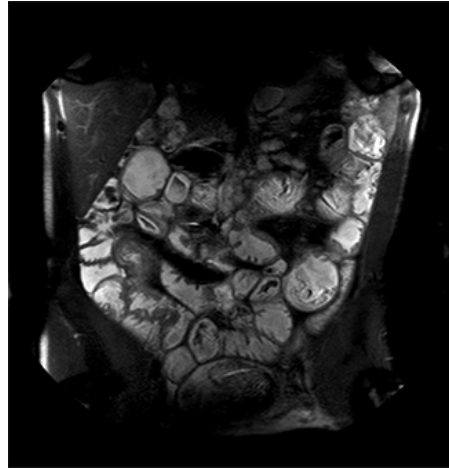
After being stimulated by an external electromagnetic signal, the spin of an atomic nucleus is excited from its aligned position and immediately tries to recover the magnetic equilibrium with its surroundings.  $T_1$ -images refer to the time it takes for the stimulated magnetic moment in an atomic nucleus to recover 63 % ( $1-1/e$ ) of its initial value after stimulation. The initial value refers to the relaxed parallel orientation to the static magnetic field before an external electromagnetic field has been switched on for stimulation.

The  $T_1$ -time for clean water is in the range of a few seconds. Typically,  $T_1$  equals a few seconds for blood and approximates 100ms for body fat. Therefore, fat appears brighter in  $T_1$ -images than water. Figure 2.1 shows a slice of a  $T_1$ -weighted scan of patient 3 of our dataset. Structures like bowel wall are clearly visible.



### 2.2.2 $T_2$ -weighted MRI (Spin-Spin Relaxation)

A stimulated atomic nucleus shows a temporary magnetization orthogonal to the magnetic field. After stimulation, this temporary magnetization exponentially decays.  $T_2$  is the time it takes for the magnetization reach 37% ( $1/e$ ) of its maximum value.  $T_2$  is usually smaller than  $T_1$  depending on the material or tissue. The  $T_2$ -times of aqueous tissues lie in the range of 40-200ms, while fat based tissues lie in the range of 10-100ms. Therefore, water appears brighter than fat.  $T_2$ -images are especially useful for detecting edema and cancerous abnormalities, since they appear brighter due to the higher water content. Figure 2.2 shows a slice of a  $T_2$ -weighted scan of patient 3 of our dataset. Structures like lumen are visible.



**Figure 2.2:  $T_2$ -weighted image (coronal SPAIR).**

### 2.2.3 Sequences

MR images can be generated in different *sequences*. The aim of distinct sequences is either to visualize different characteristics of the object or to reduce the time needed to record the signals. As explained above, the electromagnetic recovery time of the magnetic moments ranges from milliseconds to seconds. Especially motion artefacts (such as breathing or moving) and noise can therefore disturb the image quality. Several techniques have been invented to increase the speed and signal accuracy of MRI scans. Relevant for this thesis are following sequences:

#### **Spin Echo (SE)**

When the magnetic moment of a nucleus is refracted by  $90^\circ$  by the stimulating magnetic field, the orthogonal magnetization is not stable. Due to a slight spatial inhomogeneity of the field, the spins start to “drift” away from the refracted position, a phenomenon called *dephasing*. The stimulating magnetic field is now used to turn the magnetic moment again by  $180^\circ$ , by switching it on twice as long as before. After rotation, the spins “drift” together at the same rate as they had drifted away before. When all spins again align together orthogonally to the static field, the electromagnetic signal shows a clear peak (echo).

### **T1-weighted High Resolution Isotropic Volume Examination (THRIVE)**

THRIVE is an optimized fast T<sub>1</sub>-weighted 3D imaging technique combining sensitivity encoding, large volume coverage and uniform fat suppression. THRIVE improves for example dynamic liver, small bowel, breast, prostate and pancreas MRI, providing isotropic images with high resolution in short breath-hold times.

### **Spectral Selection Attenuated Inversion Recovery (SPAIR)**

The MRI fat suppression technique SPAIR is characterized by a low sensitivity to radio frequency field inhomogeneity. The used adiabatic radio frequency pulses for spectral saturation ensure a high uniformity and lower specific absorption rate (SAR). SPAIR is suitable for offset and difficult to suppress regions such as liver, pelvis and shoulder.

### **Dynamic Contrast Enhanced MRI (DCE-MRI)**

This is a series of T<sub>1</sub>-weighted MRI scans which are consecutively taken after application of a gadolinium based contrast agent to the patient. Gadolinium causes the T<sub>1</sub> relaxation time to decrease and thus will be visible as enhanced bright areas. The contrast agent is distributed in the whole body by blood vessels. Damaged tissue will accumulate the contrast agent over time which will be visible in the MRI scans. In contrast to the higher temporal resolution of the DCE-MRI sequence, these images usually show a lower spatial resolution or have a smaller field of view.

## *2.2.4 Alternative Techniques*

Medical MRI constitutes a 3D visualization of hidden body parts at high resolution. Although alternative techniques for visualization are known in radiology and medical imaging, all of them arise with different benefits and drawbacks, which will be shortly reviewed below.

**X-rays** are used in radiology to visualize hard tissues such as bones. Electromagnetic radiation on the wave length between ultraviolet light and gamma rays from a radiation source are passed through an object (body) and detected on the other side by either a photographic film or a digital detector. Depending on the composition of the irradiated tissue, more or less energy of the X-rays is absorbed which will be visible on the generated image. Since soft tissues and fluids have a comparable low absorption rate for X-rays, they show low contrast on such images making them hardly visible, in opposite to bones, teeth and medical metal implants being clearly visible due to high contrast. X-rays present a high ionizing

radiation dose to the human body, which is why they are only used when absolutely indicated.

**Computer tomography (CT)** is a further possibility to display inner parts of the body. Here, X-rays are again used to permeate the object while the radiation source and the detector move around the object allowing 3D reconstructions from the absorption profile. To enhance the tissue contrast in human bodies, iodine containing contrast agents are injected to the patient. The agent accumulates in soft tissues and shows different absorption profiles depending on its concentration. The high resolution of CT images comes at a price of high radiation dose that limits the usage of CT in daily clinic.

**Ultrasonography** is a radiation-free technique for medical imaging. It reconstructs usually 2D images (but also 3D sonography is possible) from the echo signals of underlying tissue exposed to ultrasound. Although sonography is cheap and safe for the patient, it has limitations in the comparably low resolution of the images and the lower contrast between different tissues. Also, air-filled lumens such as lung and bowel as well as bones inhibit the application of sonography, since large density changes reflect the ultrasound signal making these borders impenetrable. Still, ultrasonography has shown to be useful for initial diagnosis, assessment of disease activity, identification of fistulas, stenosis and abscesses in CD.

MRI as a further widespread imaging technique tries to overcome the drawbacks of the aforementioned methods while still delivering high resolution 3D images with high contrast for soft and hard tissues. Due to its lack of electromagnetic radiation (X-rays), MRI can potentially be used for frequent patient examinations and is also the first choice in research for novel applications. In theory, MRI can achieve much higher resolution images than X-rays or CT, since it is not limited by the resolution of a detector. But high resolution for MRI requires high static magnetic fields (7.0 T or above). The extremely high energy consumption might be problematic in clinical diagnostics. Therefore, 1.5 T and 3.0 T units are used in practice as they show sufficient resolution for diagnosis, medical decisions and guided surgery.

## 2.3 Crohn's Disease

One emerging global disease and healthcare problems are inflammatory bowel diseases (IBD), which divide in to the two groups *Crohn's disease (CD)*, also known as regional enteritis, and *ulcerative colitis (UC)* (M'Koma 2013). CD occurs in North America at the largest annual incidence rate of over 20 persons per 100,000 per year, whereas in Europe, around 13 newly CD patients per 100,000 per year are registered

(Molodecky *et al.* 2012). Europe and USA further show the largest number of prevalence of CD, with approximately 320 patients per 100,000 persons (Molodecky *et al.* 2012). These numbers are estimated to be increasing (Kirsner 1988; Molodecky *et al.* 2012), which is why more and more attention is paid in clinics and research to this disease. Whereas IBD have been known since 1761 (Kirsner 1988), Crohn's disease was firstly described by Antoni Leśniowski (1903) and later by Burril Bernad Crohn (1932).

### 2.3.1 Causes

*Crohn's disease arises from genetic predisposition, environment and lifestyle.*

The typical onset of the disease is in the young adulthood between 25 and 35 years of age, and 20-25% of patients have onset of symptoms even earlier during childhood or adolescence (Baumgart and Sandborn 2012). Although the biological causes of the disease are not fully understood, they are assumed to be an adverse mixture of genetic predisposition, lifestyle and environmental factors. A German study showed CD to be concordant in 35% of monozygotic twins with this disorder, but only in 3% of dizygotic pairs (Spehlmann *et al.* 2008). This genetic concordance has been confirmed even for CD phenotype such as location, behavior and age of diagnosis, as well at diagnosis and longitudinally (Ng *et al.* 2012). Ethnic groups with a traditionally low incidence rate such as Hispanics and Asians and immigrants moving from regions from low incidence rates to areas with high rates showed an increasing risk of developing IBD and CD (Joossens *et al.* 2007; Hou *et al.* 2009), stressing the importance of the environmental factors. Further, a range of environmental factors and changes in lifestyle have been associated with increased prevalence of CD, such as e.g. more adverse life events (Lerebours *et al.* 2007), less women breastfeeding (Barclay *et al.* 2009), smaller families with improved hygiene and sanitation (Gent *et al.* 1994), air pollution (Kaplan *et al.* 2010) or increased tobacco usage (Seksik *et al.* 2009; Jones *et al.* 2008).

### 2.3.2 Symptoms

Crohn's disease is basically a systemic and chronic inflammatory disease, and mainly affects the gastrointestinal tract and especially the whole colon (see Figure 2.3). The various manifestations of CD vary from patient to patient complicating the proper diagnosis and therapy of the disease. Acute CD can manifest with abdominal pain, fever, weight loss, diarrhea with passage of blood or mucus, or bowel obstruction (Baumgart *et al.* 2012). CD further can be accompanied by anemia and autoimmune disorders such as arthropathy, osteoporosis and pyoderma gangrenosum,

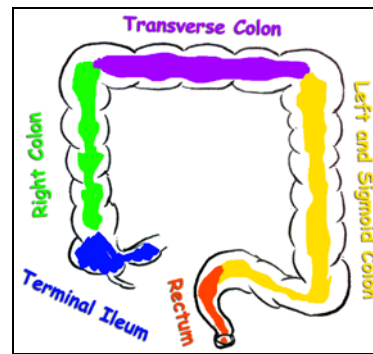
making a monitoring of the immune system necessary (Baumgart *et al.* 2012).

Studies showed a significantly reduced diversity of the mucosal bowel flora and a dysfunction of the mucosal layer in patients with CD, meaning a considerable disturbance of the first lines of defense of the immune system (Baumgart *et al.* 2012).

### 2.3.3 Therapy

There is no general top down treatment model for patients with CD, and the disease shows a high variability among patients, multifaceted in terms of symptoms, severity and secondary disorders. The aim of the individual therapy is therefore to achieve a sustained clinical remission of CD and to minimize the risk of associated complications (Baert *et al.* 2010), and medical treatment is highly varying among patients. Among conventional treatment possibilities are diets, antibiotics, autoimmune-suppressors, steroids and surgery (Baumgart *et al.* 2012). New therapeutic strategies have been invented and engineered over the past years involving monoclonal antibodies, fusion proteins, small molecules, recombinant growth factors and oligonucleotides (Baumgart *et al.* 2012). Stem cell therapies from haematopoietic, mesenchymal stromal or adipose tissue have been suggested as new alternative strategies with promising outcome (Garcia-Olmo *et al.* 2009; Burt *et al.* 2010; Duijvestein *et al.* 2010). A careful differential diagnosis considering intra- and extraintestinal symptoms is needed for an individualized therapy respecting the patient's situation (e.g. family planning, immune system state or drug tolerance).

A long-term treatment of CD patients with combinational therapy has emerged as common strategy to provoke remission. However, there is no unique consent about the optimal therapy duration among the physicians (Talley *et al.* 2011; Louis *et al.* 2012; Van Assche *et al.* 2013). Therapy might become unresponsive over time making a switch of drug classes or agents necessary (Baumgart *et al.* 2012). Surgery with operative removal of affected bowel parts does not cure Crohn's disease and should only be considered when indicated, e.g. for abscess, complex fistulas which are unresponsive to therapy, fibrostenotic strictures, high grade dysplasia or cancer (Larson and Pemberton 2004).



**Figure 2.3: Luminal Crohn's disease commonly affects the terminal ileum, right colon, transverse colon, left and sigmoid colon and the rectum.**

### 2.3.4 Diagnosis and Monitoring

Due to the multiple manifestations of the systemic disorder, various diagnosis techniques and screening methods exist to assess Crohn's disease. The diagnosis often requires a holistic view of the patient, considering medical history and environment, physical examination (heart rate, blood pressure, body mass index or external signs of disease), laboratory studies (blood test, urine strip, C-reactive protein), microbial studies (stool cultures), pathology and histology (biopsies from bowel taken during endoscopy), endoscopy, imaging examinations (CT, MRI, capsule endoscopy), and specialist consultations for possible extraintestinal symptoms (rheumatology, dermatology, urology, surgery) (Baumgart *et al.* 2012). Several infectious or non-infectious diseases can trigger similar symptoms as CD and have to be excluded, e.g. tuberculosis (Lee *et al.* 2003), John's disease (Lee *et al.* 2009a), the irritable bowel syndrome or Behçet's disease (Lee *et al.* 2009b), or enteroviruses (Pawlowski *et al.* 2009). After diagnosis, patients are commonly categorized according to the *Montreal classification* into groups of differential loci and characteristics (CD affecting the terminal ileum, colon, upper gastrointestinal tract or combination of these and with or without strictures and with or without penetration) (Silverberg *et al.* 2005; Satsangi *et al.* 2006). Further, the **disease activity** is monitored according to a range of possible scores, helping to further group patients and to pick the optimal therapeutic strategies. Disease severity scores have been developed as important predictors for course and complications of CD.

## 2.4 Crohn's Disease Severity

The absolute and objective measurement of CD severity is still an unsolved problem. Although several CD activity related indices exist, they all focus on different aspects of the disease, such as patient's quality of life, endoscopy, protein level, histopathology or MRI. Ileo-colonoscopy is considered the gold standard for CD severity assessment (Baumgart *et al.* 2012), but several research projects in the recent years have focused on abdominal MRI as principal CD activity or severity assessment modality (Rimola *et al.* 2009; Rimola *et al.* 2011; Steward *et al.* 2012; Vos *et al.* 2012; Ziech *et al.* 2012a; Ziech *et al.* 2012b). CD severity derived from MRI might be cheaper than colonoscopy, less painful and more compliant to the patients, especially on a regular examination basis. Prominent studies for MRI assessment are these of Rimola *et al.* (2009; 2011) and Steward *et al.* (2012). Rimola *et al.* have formulated a precise protocol for MRI interpretation, in which specific signs of CD (visible bowel wall thickness, relative contrast enhancement (RCE), edema and ulceration) are predictive

*The gold-standard for CD severity assessment is ileo-colonoscopy.*

for endoscopic CD severity and activity. Steward *et al.* have developed an activity index derived from MRI scans which closely relates to CD activity as indicated by histopathology. Both approaches clearly emphasize the high potential of MRI for CD severity assessment. Still, they both rely on the manually interpretation of various types of MRI sequences by medical doctors, which is not always unproblematic due to a potentially subjective perception, different level of medical experience and different types of imaging protocols. Sostegni *et al.* (2003) comprehensively reviewed available CD activity measures, of which the most important ones are explained here.

#### 2.4.1 CDEIS (Endoscopic Index of Severity)

The Crohn's Disease Endoscopic Index of Severity was developed by Mary and Modigliani (1989). Until nowadays, it remains the gold-standard for clinical studies evaluating CD activity and severity. The CDEIS is completely based on endoscopic findings on five bowel segments *terminal ileum, right colon, transverse colon, sigmoid and left colon and rectum* (see Table 2.1) and ranges from 0 to 44. The relation between clinical CD activity and endoscopic CD activity is not clear, yet (Cellier *et al.* 1994). However, novel biologic therapies aim more and more endoscopic remission measured by CDEIS. CDEIS assessment is more time consuming and complex than other scores, shows discomfort for the patient and can be impeded by stenosis and the risk of bowel perforation. Therefore, the use of CDEIS in daily clinic is still significantly limited.

##### **Local CDEIS**

For our Crohn's disease analysis pipeline, we calculate a local severity score which we call "**local CDEIS**" for every bowel segment explored. The local CDEIS is the sum of all four segment wise CDEIS characteristics (*deep ulcerations, superficial ulcerations, surface involved by disease, surface involved by ulcerations*). Table 2.1 illustrates the calculation of CDEIS together with the local CDEIS.

#### 2.4.2 AIS and eAIS (endoscopic Acute Inflammation Score)

These scores refer to Crohn's disease activity based on histopathology with methods from Borley *et al.* (2000). Steward *et al.* (2012) used this histopathological grading of CD to form the trans-mural histopathological scoring of acute inflammation (AIS), ranging from 0 to 13. Here, mucosal ulceration, edema, neutrophils and depth of neutrophil penetration are evaluated in surgical resection specimens. The authors further developed an endoscopic biopsy acute inflammation score (eAIS, range 0 to 6), which is also applicable on specimens without surgical resection.

**Table 2.1: CDEIS calculation scheme. Four endoscopic findings are evaluated for five bowel segments and summed up to "TOTAL A". The mean of the explored segments is calculated in "TOTAL B". Three points add to B if ulcerated or non-ulcerated stenosis is present anywhere, respectively, resulting in the CDEIS. Highlighted are the new *local CDEIS* scores for each segments as the sum of the four endoscopic findings per segment. Table from Daperno *et al.* (2004).**

CDEIS						
	Ileum	Right colon	Transverse	Sigmoid and left colon	Rectum	
Deep ulcerations (0 if none; 12 if present)						Total 1+
Superficial ulcerations (0 if none; 6 if present)						Total 2+
Surface involved by disease (cm)						Total 3+
Surface involved by ulcerations (cm)						Total 4=
	<b>Local CDEIS Ileum</b>	<b>Local CDEIS Right colon</b>	<b>Local CDEIS Transverse</b>	<b>Local CDEIS Sigmoid and left colon</b>	<b>Local CDEIS Rectum</b>	<b>TOTAL A</b>
Number of segments explored (1-5)						<b>n</b>
Total A / n						<b>TOTAL B</b>
If ulcerated stenosis is present anywhere add 3						<b>C</b>
If non-ulcerated stenosis is present anywhere add 3						<b>D</b>
<b>Total B + C + D = CDEIS</b>						

The authors showed on 16 patients that the eAIS correlates to the MRI findings *mural thickness* and *T2 signal* (Kendall's tau = 0.4, p=0.02). AIS and eAIS have been developed for study purposes. Clinical use in daily practice is limited due to the need of surgery or endoscopic biopsies.

### 2.4.3 CDAI (Crohn's Disease Activity Index)

Crohn's Disease Activity Index (CDAI) is one of the oldest activity indices, developed by Best *et al.* (1976). It ranges from 0 to around 600 and is calculated by a questionnaire with 8 questions answered by the patient (see Table 2.2). This self-assessment – especially of abdominal pain and



general well-being – makes the score highly subjective and less comparable among different patients. A further limitation of the CDAI as marker in everyday clinical practice is the time consumption: the CDAI is elaborated over 7 days. Further, the index is not applicable for patients with fistula or stenosis since it does not reflect these complaints accurately.

**Table 2.2: Questionnaire for CDAI (Best *et al.* 1976).**

<b>Parameter</b>	<b>Description</b>	<b>Weight</b>
<b>Number liquid stools</b>	<i>Sum of 7 days</i>	x2
<b>Abdominal pain</b>	<i>Sum of 7 days ratings</i> 0 = none 1 = mild 2 = moderate 3 = severe	x5
<b>General well-being</b>	<i>Sum of 7 days ratings</i> 0 = generally well 1 = slightly under par 2 = poor 3 = very poor 4 = terrible	x7
<b>Extra-intestinal complications</b>	<i>Number of listed complications</i> Arthritis / arthralgia, iritis / uveitis, erythema nodosum, pyoderma gangrenosum, aphtous stomatitis, anal fissure / fistula / abscess, fever > 37.8 °C	x20
<b>Anti-diarrhoeal drugs</b>	<i>Use in the previous 7 days</i> 0 = no 1 = yes	x30
<b>Abdominal mass</b>	0 = no 2 = questionable 5 = definite	x10
<b>Hematocrit</b>	<i>Expected - observed Hct</i> Males: 47 - observed Females: 42 - observed	x6
<b>Body weight</b>	<i>Ideal / observed ratio</i> [1-(ideal/observed)]x100	x1 (not <-10)

#### 2.4.4 HBI (Harvey Bradshaw Index)

Harvey and Bradshaw (1980) developed a simplified version of the CDAI. It is composed of only clinical parameters and uses a different scale than the CDAI (Table 2.3).

**Table 2.3: Clinical parameters for HBI (Harvey *et al.* 1980). These are a closely related to the CDAI.**

Parameter	Score
<b>Number liquid stools (previous day)</b>	
<b>Abdominal pain (previous day)</b>	0 = none 1 = mild 2 = moderate 3 = severe
<b>General well-being (previous day)</b>	0 = very well 1 = slightly below par 2 = poor 3 = very poor 4 = terrible
<b>Complications (each score 1)</b>	arthralgia, uveitis, erythema nodosum, aphthous ulcers, pyoderma gangrenosum, anal fissure, new fistula, abscess
<b>Abdominal mass</b>	0 = none                      1 = dubious 2 = definite      3 = definite and tender

#### 2.4.5 MaRIA (Magnetic Resonance Index of Activity)

An elaborated MRI-based CD activity score is MaRIA (Magnetic Resonance Index of Activity). Rimola *et al.* (2009) reported a significant correlation ( $r = 0.81$ ,  $p < 0.001$ ) between the CDEIS and a MR index built by a linear tobit regression model of the MRI findings wall thickness, RCE, edema and ulceration in 50 CD patients. This index has been validated and confirmed in a second study with 48 CD patients by the same authors (correlation to CDEIS  $r = 0.80$ ,  $p < 0.001$ ) (Rimola *et al.* 2011). The patient's severity is calculated as the sum of the segmental scores. The reported MaRIA score serves as a baseline for this thesis.

$$\begin{aligned}
 \text{MaRIA (segment)} &= 1.5 * \text{wall thickness (mm)} + 0.02 * \text{RCE} + 5 * \text{edema} \\
 &+ 10 * \text{ulceration}
 \end{aligned}$$

#### 2.4.6 CDA (Crohn's Disease Activity)

Since MRI is a non-invasive technology enabling an insight into the patient, it is also coming into the focus of CD activity scoring research. Steward *et al.* (2012) have shown MRI findings such as bowel wall (mural) thickness and T2 signal to be predictive for acute inflammation indicated by biopsy histology (correlation to eAIS, Kendall's  $\tau = 0.4$ ,  $p=0.02$ ). Their definition of the Crohn's disease activity is:

$$CDA = 1.79 + 1.34 * \text{mural thickness} + 0.94 * \text{mural T2 score}$$

#### 2.4.7 CRP (C-Reactive Protein)

C-Reactive Protein is an inflammation related protein. It can serve as a biochemical CD severity indicator, since it is commonly increased in patients with active CD. Although it is fast and easy to measure, it is not solely specific to CD and thus can only support the indication of CD activity.

#### 2.4.8 Calprotectin

This fecal marker protein is measured in stools. A calprotectin level over 50 mg/L during remission has been shown to predict CD relapse within 1 year with a sensitivity and specificity of 90% and 83%, respectively (Tibble *et al.* 2000).

#### 2.4.9 Further Indices

Further indices for Crohn's disease activity are the *Dutch Index* (van Hees *et al.* 1980), the *Organization Mondiale de Gastroenterologie (OMGE) Index* (Myren *et al.* 1984), the *Cape Town Index* (Wright *et al.* 1985), the *Inflammatory Bowel Disease Questionnaire (IBDQ)* (Guyatt *et al.* 1989), the *Rutgeerts' score for postsurgical recurrence* (Rutgeerts *et al.* 1990), the *Perianal Disease Activity Index (PDAI)* (Irvine 1995), *intestinal permeability* (Suenart *et al.* 2002), and the *Simple Endoscopic Score for Crohn's Disease (SES-CD)* (Daperno *et al.* 2004). They are reviewed by Sostegni *et al.* (2003).



## 3 DATASETS

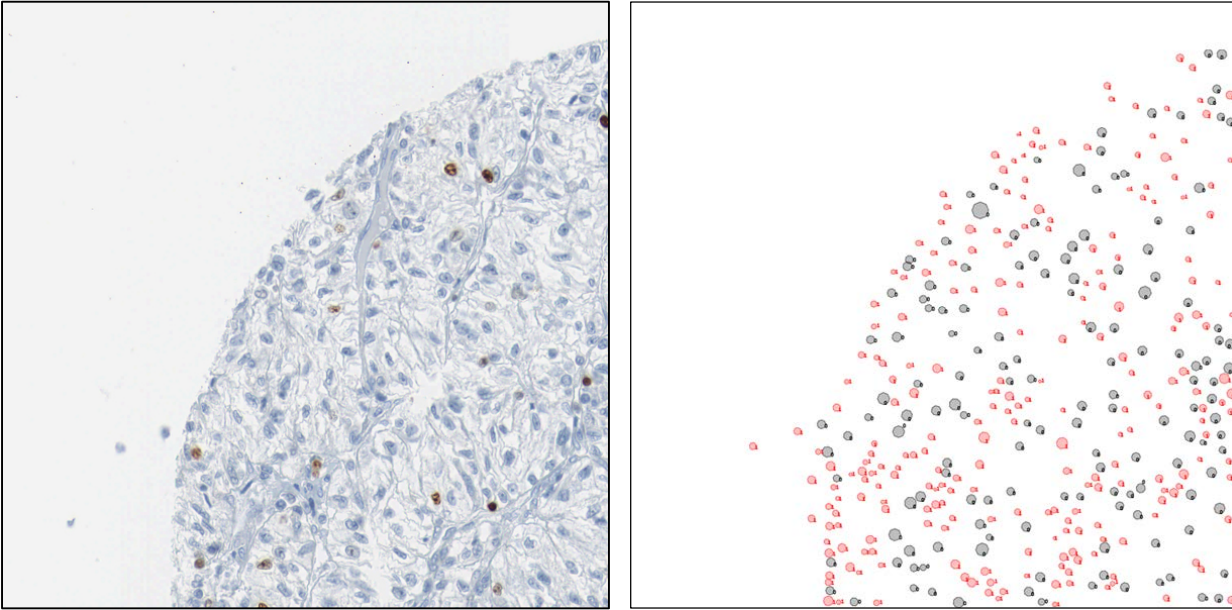
In this chapter, the datasets involved in the development and validation of our computational imaging methods are introduced. The labeled TMA dataset of clear cell renal cell carcinoma (ccRCC) images and the image database for retrospective Crohn’s disease patients are adopted from previous work (Fuchs *et al.* 2008a; Ziech *et al.* 2012a). All other datasets have been newly created during the projects discussed in this thesis and during the SIMBAD and VIGOR++ projects (Schüffler *et al.* 2013c; Schüffler *et al.* 2013d).

### 3.1 Renal Cell Carcinoma TMA Dataset

Renal Cell Carcinoma (RCC) belongs to the 10 most common cancers in western societies’ mortality (Grignon *et al.* 2004). Clear cell renal cell carcinoma (ccRCC) is a subtype of RCC occurring on cells with clear cytoplasm. Since this cancer develops metastases in a very early stage, commonly already before the diagnosis, the prognosis for RCC patients is usually poor (Tannapfel *et al.* 1996). One research field is therefore the discovery of early stage biomarkers for diagnosis and prognosis. Tissue microarrays (see section 2.1) are an important tool for molecular biomarker discovery, since they allow the screening of dozens or even hundreds of specimen simultaneously. Subgroups of patients with differential protein expression patterns can be identified under unique experimental settings.

In a retrospective patient cohort, 133 ccRCC specimens with clinical survival data were collected on a TMA at the University Hospital Zurich, Switzerland. The specimens have been immunohistochemically (IHC) stained against the monoclonal Ki-67 antibody Mib-1. Ki-67 is a human nuclear proliferation protein (Scholzen and Gerdes 2000). Positively stained nuclei on the images indicate cell proliferation and thus tumor growth. The proliferation rate of ccRCC is related to clinical prognosis, which makes Mib-1 staining estimation an indicator for survival. The 133 images were acquired with a Nanozoomer C9600 virtual slide light microscope scanner (Hamamatsu Photonics K.K.), with a 40x magnification (3000x3000px for one TMA spot). The per-pixel resolution is 0.23 $\mu$ m.

To train our TMA staining estimation pipeline, two trained pathologists with experience over 10 years have independently classified all cell nuclei on the top left quarters of 8 of these ccRCC TMA images into benign



**Figure 3.1: One example image of the eight fully labeled ccRCC TMA image quarters. LEFT: The top left quarter of one original MIB-1 stained TMA spot. The original size of the quarter is 1500x1500px. An unspecific hematoxylin staining makes cell nuclei appear as bluish roundish objects in the image. Due to the IHC staining, Ki-67 expressing nuclei appear brown in the image. The task is to estimate the percentage of brown nuclei among cancer nuclei. Computationally difficult is the nucleus detection and classification. RIGHT: One of the two manual labels of the pathologists who detected all nuclei on the image and classified them into cancerous (red) or benign (gray) class. Note that the classification of cell nuclei is independent from their color, proliferation takes place in cancer and normal cells.**

and malignant nuclei. Figure 3.1 shows one of the eight labeled TMA images. Every image shows 100-300 cell nuclei. In total, pathologist 1 discovered 2091 nuclei and pathologist 2 1908 nuclei. 1781 nuclei have been found by both pathologists commonly with a radius of 10 pixels. 1379 of these have consistent label: 978 (55%) are recognized as benign nuclei and 401 (23%) as cancerous nuclei. The 402 remaining nuclei (22%) are variably classified by the two pathologists (see Table 3.1).

**Table 3.1: Nuclei identified by two pathologists independently. 310 (15%) and 127 (7%) nuclei have been recognized by either pathologist 1 or 2, respectively.**

	Malignant	Benign	Unknown / Discrepancy	Total
<b>Pathologist 1</b>	649 (31%)	1442 (69%)	-	2091
<b>Pathologist 2</b>	581 (30%)	1327 (70%)	-	1908
<b>Consensus</b>	401 (23%)	978 (55%)	402 (22%)	1781

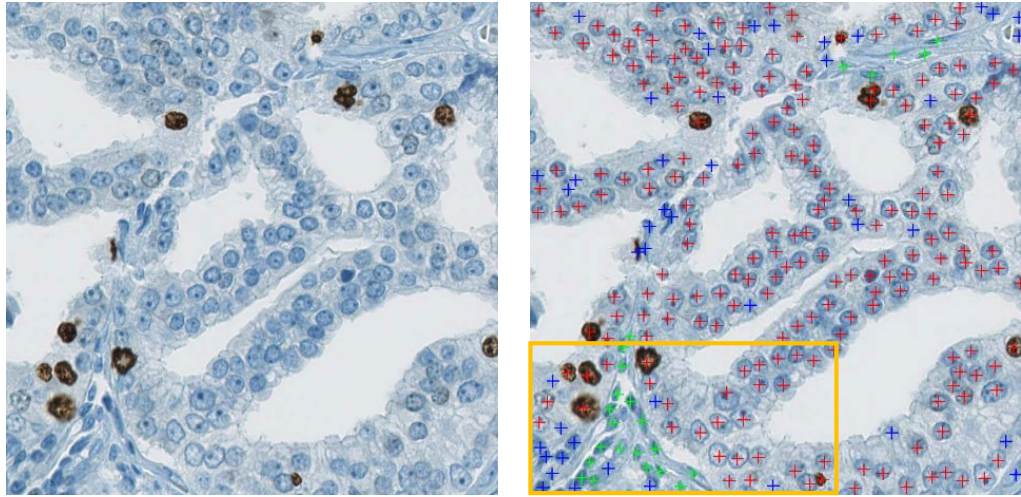
### 3.2 Prostate Cancer TMA Dataset

Prostate cancer (PCa) is one of the most common cancer types in western male society. It is the second most frequently diagnosed cancer for human males worldwide, and the sixth leading cause of cancer related death (Jemal *et al.* 2011). Although up to 80 % of the over 70 year old men have developed PCa highlighting the high incidence rate of this type of cancer, the mortality of PCa is relatively small (Breslow *et al.* 1977). However, research is ongoing for the development of specific biomarkers for the early diagnosis and the deeper understanding of PCa (Cima *et al.* 2011; Kalin *et al.* 2011).

To validate our TMA analysis pipeline on a wider range of cancers and to show its generality, we included a dataset of 227 TMA images of a PCa patient cohort. 2 pathologists exhaustively labeled six image patches of these. Pathologist 1 identified 1297 cell nuclei (407 benign, 892 malignant) in total and pathologist 2 labeled 1416 cell nuclei (314 benign, 784 malignant and 318 unclassified). Given a radius of 10 pixels, the two pathologists agreed on the location of 1195 nuclei and on the label of 985 nuclei (250 benign, 735 malignant). Figure 3.2 and Figure 3.3 show two examples of the labeled PCa dataset. Figure 3.3 illustrates the high concordance for nucleus detection for this type of images, while the nucleus classification is more difficult.

### 3.3 Retrospective Crohn's Disease MRI dataset

For our research on computational CD interpretation on MRI, two patient cohorts with clinical data and CDEIS were acquired. A *retrospective* dataset (explained in this section) of 27 CD patients comprising raw MRI scans (section 3.3.1), manual MRI scorings of 14 features by four radiologists (section 3.3.2), clinical data including CDEIS (section 3.3.4), CRP level and CDAI and manual CD segmentations on MRI (section 3.3.5) has been included for model generation. The initial retrospective dataset comprised 33 patients. Three of these patients were not scored by the radiologists and three did not sign the written consent of publish, such that 27 patients were used in this thesis. A second *prospective* dataset of 35 patients has been acquired comprising raw MRI scans and manual MRI scorings of 16 features by two radiologists for model validation and is detailed in section 3.4.



**Figure 3.2:** Example of the nucleus labeling of a pathologist for one of six PCa images. The pathologist indicated location and class of cancerous nuclei (red), normal nuclei (green) and unknown nuclei (blue). Original image patch dimensions are 800x800px. The orange rectangle is shown in original size in Figure 3.3.



**Figure 3.3:** The bottom left part of Figure 3.2 is shown illustrating the original magnification (40x) of the image. Further, the independent labels of two pathologists are overlaid (red, malignant nucleus; green, benign nucleus; blue, unknown nucleus).

### 3.3.1 MRI Protocol for CD patients

MRI scans of 27 CD patients with written consent of data usage have been acquired at the Academic Medical Center (AMC), Amsterdam, The Netherlands with a 3 Tesla MRI scanner (Intera, Philips Healthcare, Best, The Netherlands) according to following protocol (Ziech *et al.* 2012a):



1. Patients fasted for 4 h before examination.
2. To distend the bowel for better visibility in MRI, the patients drank 1.6 l of mannitol (2.5%, Baxter, Utrecht, The Netherlands), 60 minutes before the scans.
3. Axial and coronal T2-weighted single shot fast spin echo sequences (SSFSE) with and without fat saturation were acquired.
4. A coronal 3D T1-weighted spoiled gradient echo sequence (SPGE) with fat saturation was recorded.
5. The antispasmodic butylscopolaminebromide (20mg, Buscopan, Boehringer, Ingelheim, Germany) was injected to stop bowel motility.
6. A dynamic contrast enhanced (DCE-MRI) sequence with contrast agent gadobutrol (0.1 ml/kg, Gadovist 1.0 mmol/ml, Bayer Schering Pharma, Berlin, Germany) was performed: A coronal DCE-MRI sequence with 450 scans over 6 min (temporal resolution: 0.82s, spatial resolution: 2.78x2.78x2.5mm at 227x227x14 px).
7. Butylscopolaminebromide (20mg) was injected a second time.
8. Post contrast axial and coronal 3D T1-weighted SPGE sequences with fat saturation were acquired with a resolution of 1.02x1.02x2mm (400x400x100 vx).

For medical inspection and scoring by four radiologists, all sequences have been used. Automatic CD segmentation was performed only on post contrast images.

### 3.3.2 MRI Feature Assessment by Four Radiologists

Four expert radiologists independently scored 14 segmental and 3 global CD specific MRI features in all 27 patients. For the segmental scores, the visible bowel was virtually partitioned into five segments (*terminal ileum*, *ascend (right) colon*, *transverse colon*, *descend (left) and sigmoid colon* and *rectum*) and every bowel segment was individually scored. The bowel segments were identified by the medical experts coherently by visible landmarks such as e.g. the ileocecal valve, splenic flexure and hepatic flexure.

Table 3.2 lists the segmental features and Table 3.3 lists the global CD-related features, evaluated for every patient. Among these are MRI features reported in literature and commonly used by abdominal radiologists (Ziech *et al.* 2012b), and MRI features used in the two available CD MRI scoring systems *MaRIA* and *CDA*. However, it is not clear, which of these features represent CD severity and activity best, and radiologists disagree in the weighting of the features for CD assessment (Ziech *et al.* 2012b).

**Table 3.2: Scoring sheet of 14 segmental MRI features which have been manually scored by four radiologists independently. Features with a star (\*) are not used for this study (see section 3.3.3).**

Feature	Description	Values
<i>abscess</i>	Indicator whether or not abscesses are found in the bowel segment.	0: absent 1: present
<i>comb_sign</i>	Indicator whether or not the comb sign can be seen in the bowel segment. The comb sign refers to visible intestinal arcades due to increased flow, fibro fatty proliferation and perivascular inflammatory infiltration.	0: absent 1: present
<i>edema</i>	Indicator whether or not edema are present in the bowel segment.	0: absent 1: present
<i>enhancement_T1</i>	The enhancement of T <sub>1</sub> signal in the bowel segment.	0: normal 1: minor 2: moderate 3: marked
<i>fistula</i>	Indicator whether or not fistula are present in the bowel segment.	0: absent 1: present
<i>length</i>	The length of affected bowel wall in the segment.	0: 0 cm 1: 0-5 cm 2: 5-15 cm 3: > 15 cm
<i>muralT2</i>	The mural T <sub>2</sub> signal of the bowel segment.	0: normal 1: minor increase 2: moderate increase 3: marked increase
<i>mural thickness</i>	The largest mural thickness in the bowel segment.	0: 1-3 mm 1: 3-5 mm 2: 5-7 mm 3: >7 mm

<i>pattern</i>	The mural enhancement pattern in the bowel segment.	0: not allocable 1: homogeneous 2: mucosal 3: layered
<i>peri-mural_T2</i>	The perimural T <sub>2</sub> signal in the bowel segment.	0: normal 1: increased 2: small fluid rim 3: large fluid rim
<i>rce</i>	<p>The relative contrast enhancement between pre-contrast MRI and post-contrast MRI in the bowel segment. Three regions of interest (ROI) with largest bowel wall thickness are identified. The wall signal intensity (WSI) is obtained before and after application of the contrast agent gadolinium as the mean intensity of the three ROI. The <i>rce</i> is then defined as:</p> $RCE = 100 * \frac{WSI_{postgadolinium} - WSI_{pregadolinium}}{WSI_{pregadolinium}} * \frac{SD_{noise\ pregadolinium}}{SD_{noise\ postgadolinium}}$ <p>where the standard deviation (<i>SD</i>) noise pre- and post-gadolinium is measured as average of three <i>SD</i> of intensities outside the body before and after gadolinium intake, respectively (<a href="#">Semelka et al. 1991</a>).</p>	$\in \mathbb{R}$
<i>ulcers</i>	Indicator whether or not ulcers are present in the bowel segment.	0: absent 1: present
<i>wall-thickness</i>	The thickness of affected bowel wall in mm.	$\in \mathbb{R}^+$
<i>pseudo-polyps*</i>	Indicator whether or not pseudo-polyps are detected in the bowel segment.	0: absent 1: present

**Table 3.3: Scoring sheet of 3 global MRI features which have been manually scored by four radiologists independently. “\_pP”, per Patient. Features with a star (\*) are not used for this study (see section 3.3.3).**

Name	Description	Scores
<i>Enlarged_lymphnodes_pP</i>	Indicator whether or not enlarged lymph nodes are present in the patient.	0: absent 1: present
<i>lymph_nodes_pP*</i>	Reflects the number of lymph nodes found in all bowel segments.	0: no lymph nodes 1: lymph node cluster 2: 1 lymph node >1 cm 3: 3 lymph nodes >1 cm
<i>node_enhancement_pP*</i>	Describes the visual enhancement of the lymph nodes per patient.	0: less than vascular structure 1: equivalent to vascular structure

### 3.3.3 *Pseudo-Polyps, Node enhancement, Lymph Nodes, Mural Thickness and Edema*

In our medical problem, 17 potentially CD related MRI features are scored by four experienced radiologists. Four features showed to be redundant, unspecific or extremely difficult to score.

*Pseudo-polyps* are clinically important biological markers for past inflammation. However, while they are preferably useful in cancer research, they are not considered to be relevant for CD activity. Further, pseudo-polyps are often smaller than five mm and hard to identify on MRI scans. They are usually detected by high-resolution endoscopy.

Similarly, present *lymph nodes* and *lymph node enhancement* are considered as important general, non-specific markers of current active inflammation. Their meaning for CD activity is clinically not clear as well as their scoring is not well-defined as they appear outside the bowel.

Finally, *mural\_thickness* and *edema* are redundant features already expressed in *wall\_thickness* and *muralT2*, respectively. While *wall\_thickness* is measured in millimeters, *mural\_thickness* is scored in four categories (1-

3 mm, 3-5 mm, 5-7 mm and >7 mm). Nevertheless, both scorings are recorded separately, meaning *mural\_thickness* is not a post-processed stratification of *wall\_thickness*. Therefore, *mural\_thickness* and *wall\_thickness* are not perfectly correlated. The Spearman rank correlation between the two features throughout all radiologists is  $r=0.81$  ( $p<2.2e-16$ ). Analogously, *edema* is the binary formulation of the categorical *muralT2* ("Normal", "Minor", "Moderate", "Marked"). Their correlation is  $r=0.97$  ( $p<2.2e-16$ ).

We decided to exclude *pseudopolyps*, *lymph\_nodes\_pP*, *node\_enhancement\_pP* as they are not reasonably detectable or not CD severity related. On the other hand, we keep *edema* and *mural\_thickness* since they are included in the comparison literature scores. The resulting 14 manual MRI features are subjected to further analysis.

The size of the final dataset is estimated as 27 patients \* 5 bowel segments \* 4 radiologists = 540 samples. Since 7 segments could not be assessed by the radiologists (*ascend colon* in patients 12, 14 and *rectum* in patients 13, 15, 18, 26, 27; 2 due to resection, 5 due to poor bowel distension), the MRI dataset comprises  $128 * 4 = 512$  samples. This number will further reduce when the samples are matched to the label CDEIS in the next section.

### 3.3.4 CDEIS Assessment by a Medical Doctor

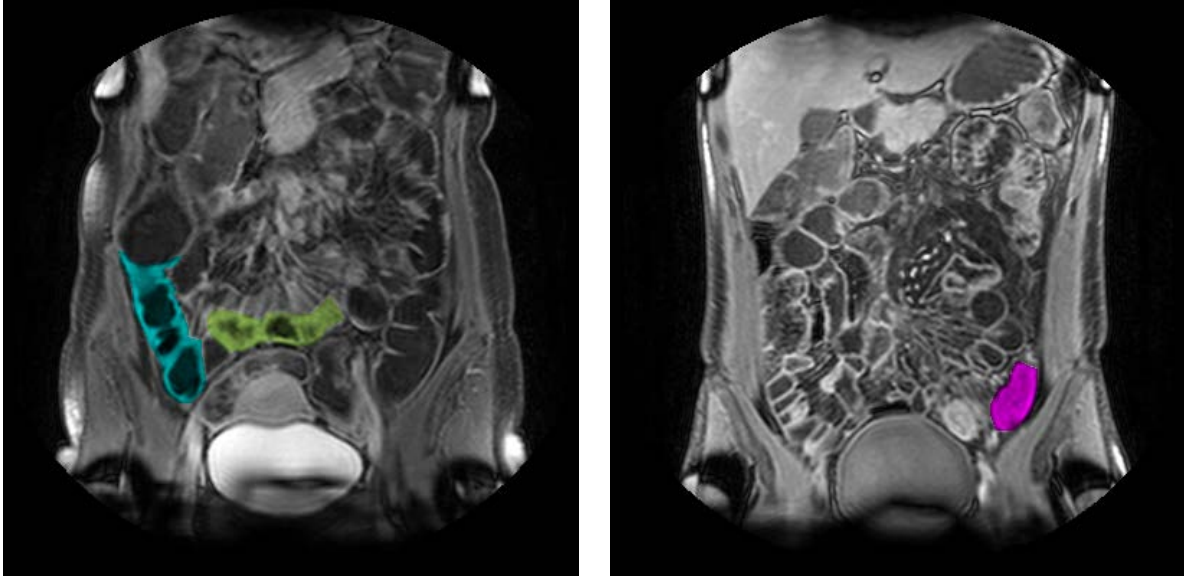
33 CD patients from Academic Medical Center (AMC), Amsterdam, The Netherlands, underwent ileo-colonoscopy prior to MRI examination. The *Crohn's Disease Endoscopic Index of Severity (CDEIS)* is determined according to the scheme in section 2.4.1. Thus, every bowel segment obtains a local severity score which we call "**local CDEIS**".

The CDEIS serves as gold-standard for disease severity. Three of all 33 patients (no 10, 16 and 30) do not have a MRI report. The *terminal ileum* was not accessible for colonoscopy in six patients (no 11, 20, 21, 24, 28, 33; e.g. due to stenosis). Combining the MRI dataset with the corresponding local CDEIS results in  $122 * 4 = 488$  samples with 14 MRI features and one CDEIS label, each, which forms the final retrospective dataset.

*Retrospective dataset with 122 samples and 14 features.*

### 3.3.5 Manual CD Segmentation by One Radiologist

To develop a framework for automatic CD detection and segmentation in MRI, we collected the manual segmentation of a trained domain expert as gold-standard. In 26 retrospective patients, 28 3D-regions of enhanced bowel wall signal were identified and manually segmented in the post-contrast VIBE sequences. For every diseased region, a normal region in the same segment was depicted as counter example. Figure 3.4 illustrates two example images of the manual CD segmentation. The bowel segment is encoded by the color in the drawing.



**Figure 3.4:** Two example images of THRIVE MRI scans of two patients with manually segmented bowel wall regions with enhanced signal. LEFT: Patient 4, 3D-slice No. 33. RIGHT: Patient 18, slice No. 26 (100 slices in total for each patient). The color encodes the bowel segment: green, terminal ileum; blue, ascend colon; violet, descend and sigmoid colon. Resolution is 1.02x1.02x2mm (400x400 px).

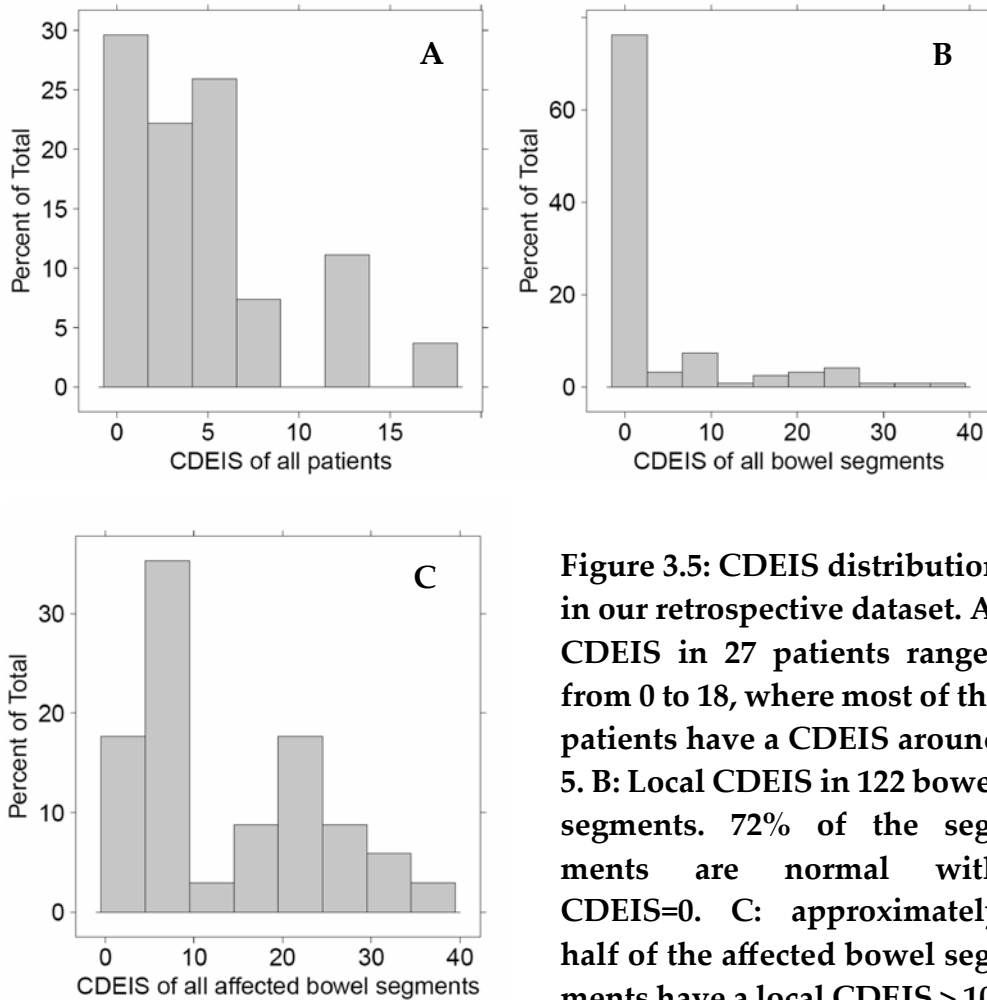
### 3.3.6 Retrospective Dataset Statistics

#### CDEIS Distribution

Four patients (no 3, 7, 15, 26) showed no endoscopic evidence of disease (CDEIS = 0). The CDEIS in all patients ranges from 0 to 18 (mean = 4.67, median = 3.6). The local CDEIS ranges from 0 to 38 (mean = 3.85, median = 0). From the 122 bowel segments in our dataset, 88 (72%) are normal (CDEIS = 0) and 34 (28%) have a CDEIS > 0. From the 34 affected bowel segments, 16 (47%) have a local CDEIS larger than 10 (see Figure 3.5).

#### MRI Feature Data Types

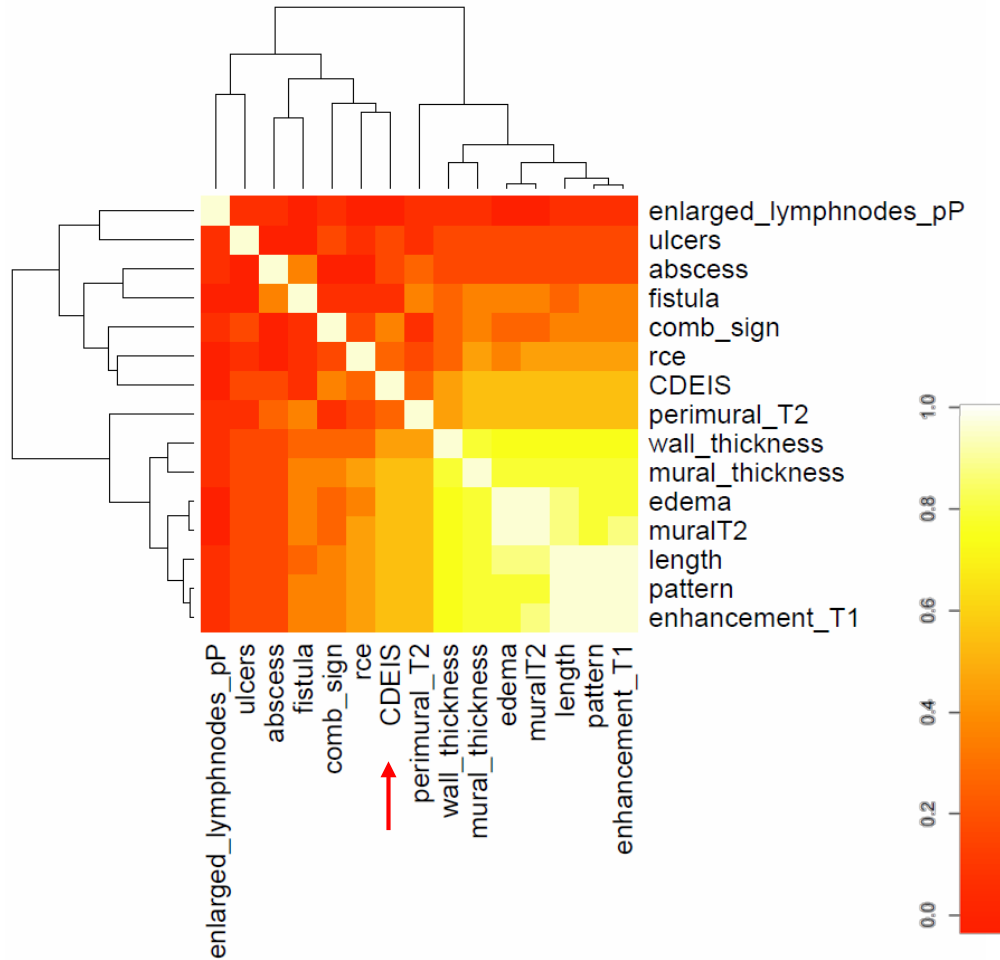
Two MRI features are numerical: *rce* (min = -48.5, max = 234.6, mean = 52.8, median = 43.6) and *wall\_thickness* (min = 1.8, max = 15.1, mean = 3.7, median = 2.8). An *rce* value of 0 means there is no contrast enhancement after application of contrast agent. A wall thickness < 3 mm is considered normal. 15 MRI features are categorical with two or four categories, with most segments being normal. E.g. abnormal *abscess*, *pseudopolyps*, *fistula* or *ulcers* were identified in only 5, 7 or 6 of 548 samples, respectively, making these features extremely sparse (~1%).



**Figure 3.5: CDEIS distribution in our retrospective dataset. A: CDEIS in 27 patients ranges from 0 to 18, where most of the patients have a CDEIS around 5. B: Local CDEIS in 122 bowel segments. 72% of the segments are normal with CDEIS=0. C: approximately half of the affected bowel segments have a local CDEIS > 10.**

### Correlation of Features

A univariate correlation analysis reveals the cross-correlation of the 14 MRI features and their correlation to the local CDEIS (see Figure 3.6). A significant Spearman rank correlation to the local CDEIS larger than  $r=0.5$  throughout all radiologists can be found for *length* ( $r=0.54$ ,  $p<0.001$ ), *muralT2* ( $r=0.53$ ,  $p<0.001$ ), *enhancement\_T1* ( $r=0.52$ ,  $p<0.001$ ) and *mural\_thickness* ( $r=0.5$ ,  $p<0.001$ ). Further, *wall\_thickness*, *mural\_thickness*, *edema*, *muralT2*, *length*, *pattern* and *enhancement\_T1* show a significant and strong inter-correlation ( $r>0.7$ ,  $p<0.001$ ) and therefore a yellow to white entry in Figure 3.6. Note, that only three variables (*CDEIS*, *wall\_thickness* and *rce*) are quantitative measurements. For Spearman rank correlation calculation, all categorical variables are considered ordinal (e.g. *comb\_sign* “absent”, “present” is represented by 0 and 1).



**Figure 3.6: Symmetric cross-correlation (Spearman’s  $r=0.0-1.0$ ) of 14 MRI features and local CDEIS (middle, red arrow). A dendrogram clusters the features according to their means.**

### Inter-Observer Agreement of Features

To evaluate the reproducibility of the features among the four expert radiologists, the overall agreement (OA),  $\kappa$ -statistics (Cohen 1960; Fleiss 1971; Light 1971) and agreement coefficient (AC1) (Gwet 2008) was calculated for categorical variables and an intra-class-correlation (ICC) (Fleiss and Cohen 1973) was calculated for numeric variables. These statistics are shortly explained here.

#### Overall Agreement OA

Given two raters A and B classifying  $N$  objects into  $k$  categories, the overall agreement  $p_o$  is defined as the proportion of objects in which the raters agree with the category (see contingency Table 3.4):

$$p_o = \frac{1}{N} \sum_{i=1}^k f_{i,i}$$

where  $f_{i,i}$  is the frequency of objects classified by both raters into category  $i$ .



**Table 3.4: Contingency table of two raters classifying N objects into k categories. The overall agreement and  $\kappa$  coefficient is calculated with the frequencies  $f_{i,\cdot}$ .**

		Rater B			Frequencies $f_{i,\cdot}$
		Cat 1	...	Cat k	
Rater A	Cat 1	$f_{1,1}$	...	$f_{1,k}$	$f_{1,\cdot} = \sum_i^k f_{1,i}$
	...	...	...	...	...
	Cat k	$f_{k,1}$	...	$f_{k,k}$	$f_{k,\cdot} = \sum_i^k f_{k,i}$
Frequencies $f_{\cdot,i}$		$f_{\cdot,1} = \sum_i^k f_{i,1}$	...	$f_{\cdot,k} = \sum_i^k f_{i,k}$	$\sum \sum = N$

$\kappa$ -statistics

Cohen (1960) defined the  $\kappa$ -coefficient as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where  $p_e$  is the proportion of objects with expected agreement by chance:

$$p_e = \frac{1}{N^2} \sum_{i=1}^k f_{i,\cdot} * f_{\cdot,i}$$

Fleiss (1971) and Light (1971) have expanded the  $\kappa$ -coefficient to multiple raters. While Fleiss introduced  $\kappa_F$  by extending the contingency table to more dimensions and adjusting  $p_o$  and  $p_e$  accordingly, Light proposed  $\kappa_L$  as average over all possible pairwise  $\kappa$  coefficients. We use both Fleiss' and Light's methods as they both form standard agreement measures in literature and are comparable to other studies (Tielbeek *et al.* 2013).

Further, Cohen (1968) introduced a weighted  $\kappa_W$  coefficient for ordered categories. He suggested weights for the distance of the categories of the two raters from the contingency's diagonal. The values of the features in our dataset can be ordered in the case of multiple categories, therefore we use a linear weight for the disagreement between the raters for  $\kappa_L$ .

*Agreement Coefficient AC1*

Gwet (2008) proposed a new agreement coefficient AC1 for categorical data which is supposed to be more robust on data where the overall agreement is very high. In these cases, the  $\kappa$ -value can be very small, indicating a poor agreement, which is not intuitive. This observation is known as a paradox of the  $\kappa$ -value and explained later on the example of

*pseudopolyps* and *ulcers* in our dataset. The agreement coefficient AC1 is defined as:

$$AC_1 = \frac{p_o - p_{ey}}{1 - p_{ey}}$$

where

$$p_{ey} = \frac{1}{k-1} \sum_{i=1}^k p_i * (1 - p_i) \text{ and } p_i = \frac{f_{i,+} + f_{+,i}}{2N}.$$

#### *Intra-Class-Coefficient ICC*

For numerical data and multiple raters, the intra-class-correlation (ICC) is the standard measure for agreement validation. This measure is based on variance analysis between the objects and raters. Fleiss and Cohen (1973) have shown that the ICC is the numeric equivalent to the weighted  $\kappa$ -coefficient. Table 3.5 lists the agreement of the features in our dataset of 27 CD patients by four radiologists. According to the nomenclature of  $\kappa$ -value interpretation by Landis and Koch (1977) (<.00: poor; .00-.20: slight; .21-.40: fair; .41-.60: moderate; .61-.80: substantial; .81-1.00: almost perfect), *length*, *edema*, *wall\_thickness*, *CDA* and *MaRIA* show a substantial agreement among the four observers. *Pattern*, *fistula*, *muralT2*, *enhancement\_T1*, *mural\_thickness*, *node\_enhancement\_pP* and *rce* still show a moderate agreement, whereas the remaining features have a slight or fair agreement among the four radiologists.

The example of *pseudopolyps* illustrates the delicate interpretation of the  $\kappa$ -value: of 128 bowel segments in the dataset, four radiologists did not see any *pseudopolyps* in 127 segments. In one segment, one expert detected a *pseudopolyp*. Intuitively, the inter-observer agreement is expected to be high, though the  $\kappa_L$  value is 0.0. In this example,  $p_o$  as the overall agreement is 1.00. But  $p_e$ , the expected agreement *by chance* is also close to one. Since the observed agreement is not different than the agreement by chance, the  $\kappa_F$  value is comparable small.

The feature *ulcers* further illustrates the paradox of a negative  $\kappa$ -value. Consider the pairwise Cohen  $\kappa$  of two observers: of 128 bowel segments, both observers agree on absence of *ulcers* in 123 segments. 4 segments show *ulcers* only for the first radiologist, one segment shows *ulcers* only for the second radiologist. No segment shows *ulcers* for both radiologists. Therefore, the observed agreement is:  $p_o = 123/128 \approx 0.9609$ . The expected agreement with this contingency is:  $p_e = \frac{127*124+4*1}{128^2} \approx 0.9614$ . Since  $p_e > p_o$ , the resulting  $\kappa$  will be negative.

Table 3.5: Inter-observer agreement of 4 radiologists within all MRI findings and two MRI severity scores in 27 patients. Values are ordered by Gwet's (2008) robust and intuitive agreement coefficient (AC1). All coefficients range from 0 (poor agreement) to 1 (perfect agreement). K can be negative if the expected agreement between raters is larger than the overall agreement. Overall agreement (OA), Fleiss'  $\kappa_F$ , Light's  $\kappa_L$  and AC1 for categorical data, intra-class-correlation (ICC) for numerical data.

Feature	OA	$\kappa_F$	$\kappa_L$	AC1	ICC
<i>pseudopolyps</i>	0.99	0.00	0.00	1.00	
<i>ulcers</i>	0.95	-0.01	-0.01	0.98	
<i>abscess</i>	0.95	0.24	0.38	0.98	
<i>fistula</i>	0.94	0.48	0.50	0.97	
<i>comb_sign</i>	0.86	0.26	0.20	0.92	
<i>edema</i>	0.77	0.64	0.64	0.80	
<i>enlarged_lymphnodes_pP</i>	0.74	0.35	0.34	0.80	
<i>perimural_T2</i>	0.68	0.25	0.33	0.79	
<i>pattern</i>	0.64	0.55	0.60	0.75	
<i>muralT2</i>	0.66	0.42	0.51	0.74	
<i>length</i>	0.59	0.48	0.61	0.71	
<i>enhancement_T1</i>	0.56	0.39	0.53	0.67	
<i>wall_thickness</i>					0.66
<i>mural_thickness</i>	0.52	0.39	0.57	0.64	
<i>node_enhancement_pP</i>	0.59	0.50	0.50	0.56	
<i>lymph_nodes_pP</i>	0.37	0.30	0.38	0.48	
<i>rce</i>					0.44
<b>Score</b>					
<i>CDA</i>					0.77
<i>MaRIA</i>					0.72

### 3.4 Prospective Crohn's Disease MRI Dataset

During this study, a prospective dataset is being prepared by the medical VIGOR++ partners at the Academic Medical Center (AMC), Amsterdam, The Netherlands and the University College London (UCL), London, United Kingdom. A shared MRE protocol has been developed to generate comparable MRI scans from new CD patients. The protocol is almost identical to that for the retrospective data. The main change in the prospective MRI sequence is the use of a 3 Tesla scanner for UCL, instead of a 1.5 Tesla machine. The stronger magnetic field results in a better image quality with higher signal to noise ratio.

Initially, 35 new patients have been scanned at UCL, with written consent of data usage. The basic specification for inclusion in this dataset is an age over 18, confirmed Crohn's disease, assessed CDEIS and MRE with the same MRI features as for the retrospective study.

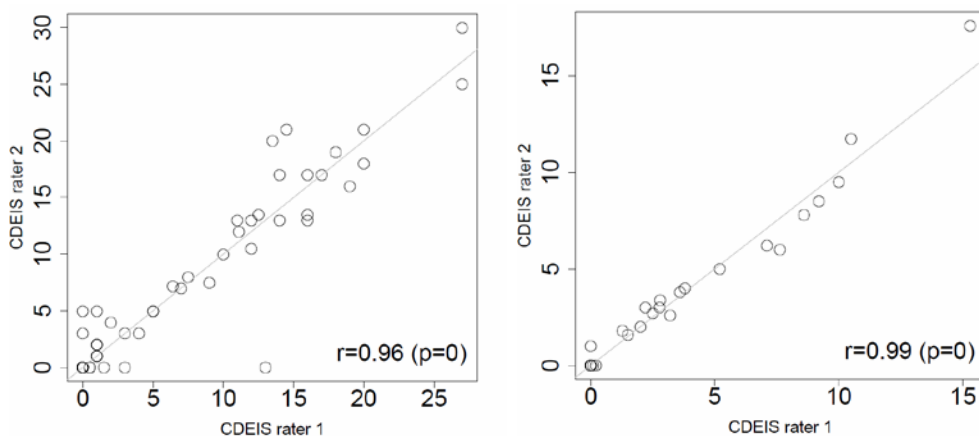
### 3.4.1 MRI Feature Assessment by Two Radiologists

Similar to the retrospective data, two radiologists (one from each institute) have independently scored the MRI features listed in Table 3.2 and Table 3.3, except of the feature *enlarged\_lymphnodes\_pP*. Of the 35 patients, one radiologist scored 34 patients and the second radiologist scored 24 patients. 23 patients have been scored by both radiologists. In total, 172 bowel segments were scored by the radiologists, 114 segments in intersection.

### 3.4.2 CDEIS Assessment by Two Medical Doctors

As a peculiarity in the prospective dataset, the CDEIS is diagnosed by *two* medical doctors independently. The second CDEIS is derived through a video examination by the second gastroenterologist without communication to the performing physician.

For two of the 35 patients (patients 5 and 6) a second CDEIS is not available, as well as for seven individual segments in patients 15, 24, 30, 44 and 50. The correlation of the local CDEIS for 155 segments is  $r=0.96$  ( $p=2.2e^{-16}$ ) (see Figure 3.7, left). The root mean square deviation is  $\text{RMSD}=1.51$ . The large number of non-affected bowel segments does not influence these values: on only 44 segments with signs of CD for at least one labeler, the CDEIS correlation is  $r=0.93$  ( $p=2.2e^{-16}$ ) and the  $\text{RMSD}=2.53$ . The high segmental CDEIS agreement propagates to the whole-patient assessment: the correlation of the global CDEIS in 33 patients is  $r=0.99$  ( $p=2.2e^{-16}$ ) and the  $\text{RMSD}=0.60$  (see Figure 3.7, right). This high correlation nicely indicates how suitable the endoscopic CDEIS is as gold standard reference.



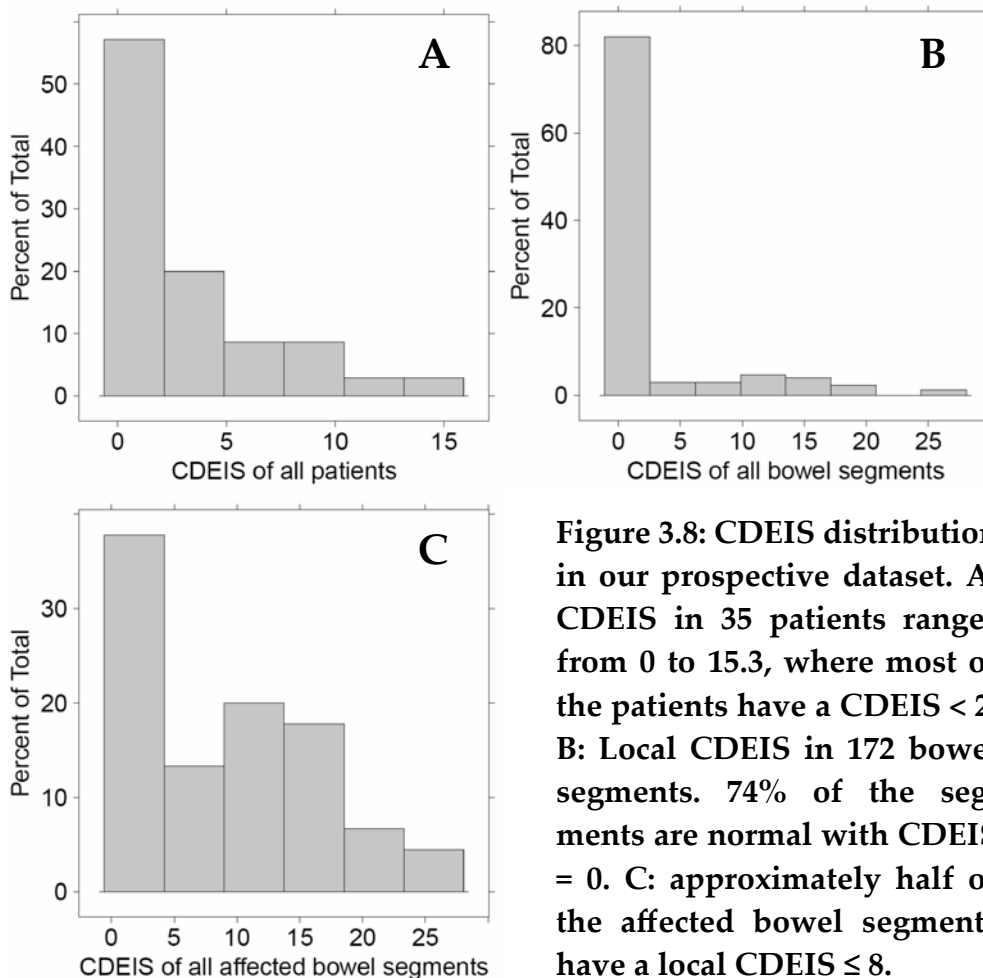
**Figure 3.7: Pearson correlation of independently labeled CDEIS of two independent medical doctors. LEFT: Per segment CDEIS of 155 segments. RIGHT: Per patient CDEIS of 33 patients. The high concordance of this severity justifies its use as reference standard.**

### 3.4.3 Prospective Dataset Statistics

35 CD patients have been acquired for this study. 13 patients (37 %) show no endoscopic evidence of disease with CDEIS = 0. 16 patients (46 %) had  $\text{CDEIS} \leq 1$  and 20 patients (57 %) had a  $\text{CDEIS} \leq 2$ . The CDEIS of all patients ranges from 0 to 15.3 (mean=2.9, median=1.5). The local CDEIS of all 172 bowel segments ranges from 0 to 27 (mean = 2.4, median = 0). 127 (74%) segments are normal with  $\text{CDEIS} = 0$  and 45 (26%) have a  $\text{CDEIS} > 0$ . From the 45 affected bowel segments, 22 (49%) have a local CDEIS larger than 8 (see Figure 3.8).

#### Inter-Observer Agreement of Features

23 patients have been scored by two radiologists for 17 MRI features. The inter-observer agreement is similar and slightly higher than in the retrospective dataset (see Table 3.6). Note that this might mainly arise from the fact that only two observers instead of four have labeled weigh fewer patients (23 patients instead of 27). The chance of agreement is therefore higher than on the retrospective data set. Note that all perfect agreements of 1 (100% agreement) are exclusively a result of all samples being normal, with no positive sample according to both radiologists.



**Figure 3.8: CDEIS distribution in our prospective dataset. A: CDEIS in 35 patients ranges from 0 to 15.3, where most of the patients have a  $\text{CDEIS} < 2$ . B: Local CDEIS in 172 bowel segments. 74% of the segments are normal with  $\text{CDEIS} = 0$ . C: approximately half of the affected bowel segments have a local  $\text{CDEIS} \leq 8$ .**

Table 3.6: Inter-observer agreement of two radiologists within all MRI findings and two MRI scores for CD severity in 23 prospective patients. Values are ordered by Gwet's (2008) robust and intuitive agreement coefficient (AC1). All coefficients range from 0 (poor agreement) to 1 (perfect agreement). Overall agreement (OA), Fleiss'  $\kappa_F$ , Light's  $\kappa_L$  and AC1 for categorical data, intra-class-correlation (ICC) for numerical data.

Feature	OA	$\kappa_F$	$\kappa_L$	AC1	ICC
<i>lymph_nodes_pP</i>	1.00			1.00	
<i>node_enhancement_pP</i>	1.00			1.00	
<i>abscess</i>	1.00			1.00	
<i>fistula</i>	1.00			1.00	
<i>pseudopolyps</i>	1.00			1.00	
<i>perimural_T2</i>	0.95	0.38	0.18	0.94	
<i>comb_sign</i>	0.95	0.64	0.64	0.94	
<i>enlarged_lymphnodes_pP</i>	0.91	0.05	0.00	0.90	
<i>ulcers</i>	0.85	0.25	0.29	0.82	
<i>length</i>	0.84	0.54	0.63	0.81	
<i>pattern</i>	0.82	0.49	0.63	0.79	
<i>mural_thickness</i>	0.80	0.34	0.53	0.77	
<i>edema</i>	0.84	0.47	0.48	0.76	
<i>enhancement_T1</i>	0.76	0.36	0.49	0.73	
<i>muralT2</i>	0.71	0.19	0.36	0.67	
<i>wall_thickness</i>					0.61
<i>rce</i>					0.14
Score					
<i>CDA</i>					0.75
<i>MaRIA</i>					0.63

## 4 TMA STAINING ESTIMATION PIPELINE

The percentage staining estimation of immunohistochemically stained tissue microarrays (TMA) is essential in a variety of medical studies. E.g. in the context of cancer research, we refer to the pathologic *staining estimation* as the *quantitative evaluation of stained cancerous cell nuclei in a given TMA spot*. TMA specimens can thus be grouped to non-expressing, low expressing, moderately expressing or highly expressing samples. Staining estimation is widely employed in medical research and life sciences for the development of diagnostic or therapeutic biomarkers.

Pathologists typically estimate the staining percentage of TMA samples visually with a light microscope. Cell nuclei are counted and classified within a well-defined area, similar to a haemocytometer. The stained fraction of malignant cells for the whole TMA spot is then estimated.

*Manual staining estimation*

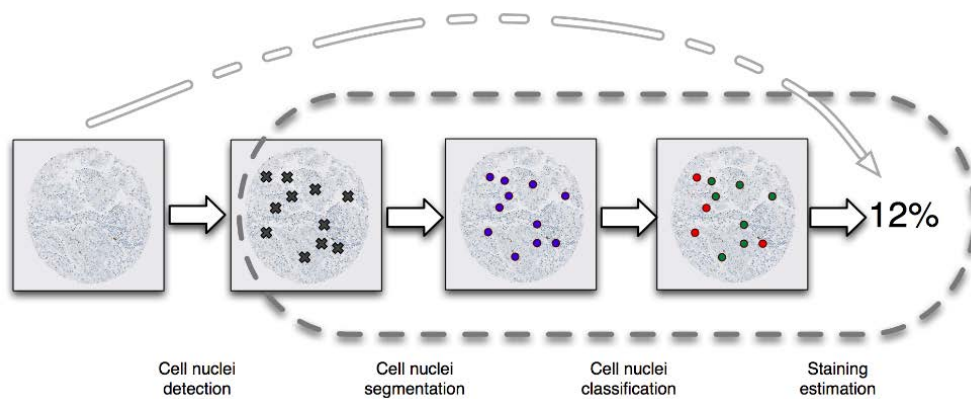
This visual procedure is of course highly time consuming, partly subjective and potentially prone to error. Although the experienced pathologist can easily judge the quality of the TMA spot and select a highly representative subarea for the cell counts, still the manual staining estimation depends on the selected count window, the homogeneity of the TMA spot and the correct identification of malignant cells. Fuchs *et al.* (2011b) have shown that the human perception of malignant renal clear cell carcinoma cell nuclei varies among five trained pathologists: of 180 ccRCC nuclei, all radiologists agreed on only 105 (58%) on the label (tumor or non-tumor) whereas they disagreed on 75 (42%) nuclei, indicating a high inter-expert variability. The authors further illustrate the intra-expert variability when the five pathologists had to label the nuclei twice: The second time, they had an intra-expert classification disagreement of 21.2% on average. The inter- and intra-expert variance in nucleus classification propagates to TMA staining estimation: also here, a high standard deviation of staining percentage can be found, especially on TMA spots with a mean staining percentage over 10% (Fuchs *et al.* 2011b).

Therefore, a fast, standardized and reproducible staining estimation procedure is desirable, especially for large patient cohorts. For this reason, Fuchs *et al.* (2008a) have formalized a computer-aided TMA analysis pipeline on renal clear cell carcinoma samples. In principle, the pipeline works with following steps: (i) cancer-nucleus detection as pixel-wise classification problem, (ii) staining estimation among the detected nuclei as color classification and (iii) survival prediction on the whole patient

*Related Work: Computational Staining Estimation by Fuchs et al. (2008a).*

cohort as validation of the procedure. This automatic process comprises a classifier specifically designed to directly detect malignant nuclei on the image. A random forest classifier ensemble is used to solve this step. Local binary patterns and color features are extracted as descriptors (Fuchs *et al.* 2008a).

To study the difficult part of nucleus classification, we consider a new TMA analysis pipeline (Figure 4.1) which separates the nucleus detection and classification as two consecutive steps (Schüffler *et al.* 2013a). As an additional modulation, we introduce the cell nucleus segmentation as an essential part of the pipeline (Schüffler *et al.* 2010).



**Figure 4.1: Overview of the new computational TMA analysis pipeline. The dashed arrow indicates the manual staining estimation process by pathologists: From a TMA spot, cell nuclei are identified, classified and counted by human eye in nearly one step. We mimic this highly complex process in a computer vision pipeline and partition it into the four steps (i) nucleus detection, (ii) nucleus segmentation (iii) nucleus classification and (iv) staining estimation on malignant nuclei. The dashed circle encloses the nucleus classification which we want to improve.**

Our proposed TMA analysis pipeline therefore consists of the following steps: (i) nucleus detection, (ii) nucleus segmentation, (iii) nucleus classification, and (iv) staining estimation.

The new design of the computational staining estimation pipeline enables the isolated study of nucleus detection, nucleus segmentation and nucleus classification. We will show in section 4.7 that the isolated improvement of single steps in the pipeline also enhances the overall process of staining estimation. In this thesis, we focus on the nucleus segmentation and classification with new shape measurements and new classifier ensembles. Morphology and shape play an important role in manual and visual nucleus identification.



## 4.1 Structure

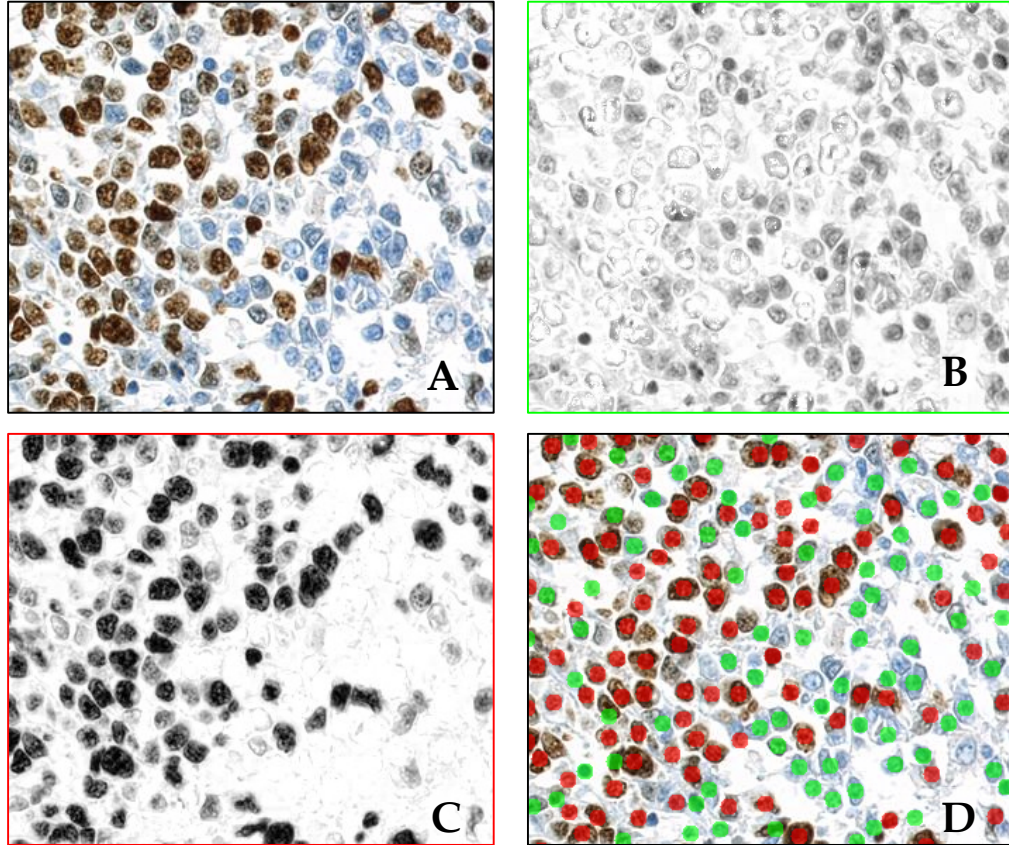
First, we will present two methods for nucleus detection in section 4.2: Color deconvolution has been developed by Ruifrok and Johnston (2001). This method is completely unsupervised and does not support cell nucleus classification. As a new alternative, we study the potential of superpixels for nucleus detection with inherent segmentation and classification. Cell nucleus segmentation is further detailed using graph-cuts as an isolated step in section 4.3. In the following section, we consider the nucleus classification step of the TMA analysis pipeline as a similarity based classification problem on the example of 1272 ccRCC nuclei, labeled by two pathologists with a unique label cancerous or benign. Support vector machines (SVM) provide a possible object classification with custom measures for similarity or distance. First, we propose a simple classification scenario with dedicated features described in section 4.4 and SVM in section 4.5. Shape measures are beneficial for this task, as we will explain in section 4.6. We extend this approach by the use of multiple kernel learning in section 4.8 as classifier ensemble. The kernels are then combined in a nonlinear manner in section 4.9. These chapters are an important example of the joint contribution within the SIMBAD project.

## 4.2 Nucleus Detection

### 4.2.1 Color Deconvolution Based Nucleus Detection

When the underlying biological problem aims for staining estimation without prior classification (i.e. the cells are homogeneous on the image), color deconvolution by Ruifrok *et al.* (2001) provides a fast and unsupervised way for identifying and counting stained nuclei. The image is deconvolved into distinct color channels (e.g. hematoxylin channel and DAB channel) which are then smoothed with a Gaussian blur filter. Subsequently, the images are screened for local intensity maxima to localize the detected nuclei. Figure 4.2 shows an example image section of a ccRCC TMA on which cell nuclei are detected with color deconvolution. Note that the different classes of nuclei (green and red nuclei) refer to unstained or stained as they arise from the hematoxylin or DAB channel. Few parameters are needed for local maxima detection: The radius  $r$  of cell nuclei describing the size of the local environment, and the intensity threshold  $t$  per channel, above which a local maximum is accepted. These parameters vary between tissues and experimental staining protocols of the experiments. The nucleus detection and staining estimation with color deconvolution provides facilitated parameter settings without

classical machine learning influence. If the pathology goal does not depend on the nucleus type (e.g. overall counting tasks), or the nucleus types are known throughout a given image set (e.g. all are cancer nuclei), this method is a fast alternative to the more comprehensive classification



**Figure 4.2: Color Deconvolution of an example ccRCC image patch: A: The original image. B: The hematoxylin channel image. C: The DAB channel image. D: Found nuclei based on the intensities on the two channels B and C.**

### Validation of Nucleus Detection via Color Deconvolution

To quantify the performance of the presented algorithms in this thesis, we calculate the match statistics between the gold standard labels of two trained pathologists and the computationally detected cell nuclei. Two nuclei with distance  $d$  are called *matching* each other, if  $d \leq 2r$ , where  $r$  is the nucleus radius. If more than two nuclei in a radius are found, the closest ones are matched to each other. Based on this radius, *precision*, *recall* and *Fscore* are calculated. After classification of the detected cell nuclei into malignant and benign, *sensitivity*, *specificity* and overall classification *accuracy* are evaluated. Consider following confusion table for nucleus detection and classification (Table 4.1). Note that *precision* is also known as *positive predictive value*.

Table 4.1: Confusion table for nucleus **detection** (brown) and **classification** (blue). Note for the **detection**, “true negative” is not defined, thus there is no “detection accuracy”.

		Gold standard		
		Present / Positive	Absent / Negative	
Machine	Present / Positive	<b>Found</b> <i>True Positive</i>	<b>Hallucinated</b> <i>False Positive</i>	<i>Precision</i> $\frac{TP}{TP + FP}$
	Absent / Negative	<b>Missed</b> <i>False Negative</i>	- <i>True Negative</i>	
		<i>Recall</i> = <i>Sensitivity</i> $\frac{TP}{TP + FN}$	<i>Specificity</i> $\frac{TN}{FP + TN}$	

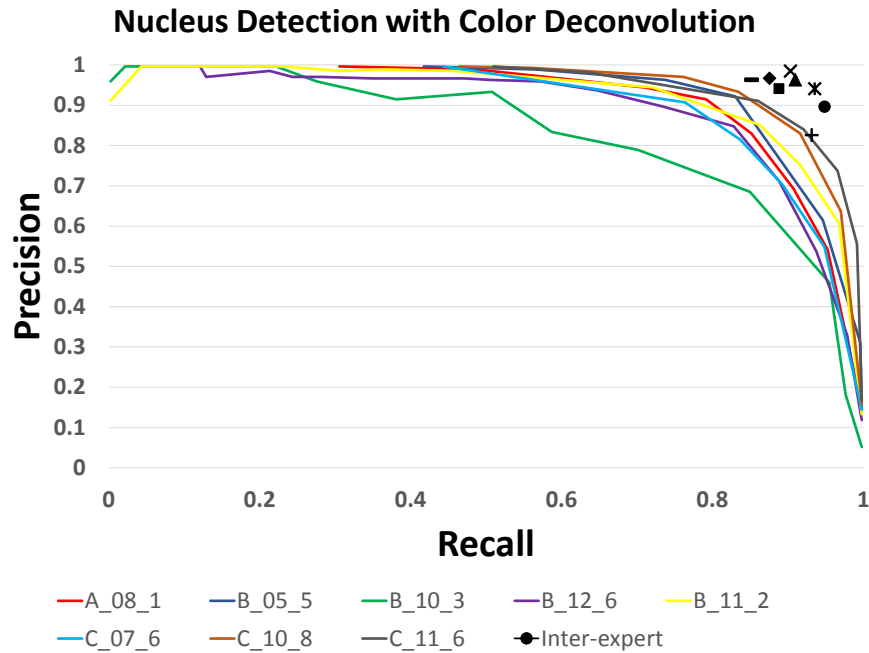
Then, the *Fscore* is defined as:

$$Fscore = 2 * \frac{Precision * Recall}{Precision + Recall} ,$$

and the overall classification accuracy is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} .$$

We tested the algorithm on 8 fully labeled images of ccRCC TMA. Two trained pathologists exhaustively identified all visible nuclei on the images. The detected nuclei of our algorithm are matched against one of the pathologists. As shown in Figure 4.3, color deconvolution achieves a reproducible and high precision and recall in nucleus detection. Each patient is represented by a colored curve. The shifting parameter of the method in this figure is the radius  $r$ : a high radius entails a high precision but a low recall, since less nuclei are found, whereas a small radius involves a high recall but a low precision. The performance hereby is still comparable to two individual pathologists with their inter-expert precision and inter-expert recall for the same patients, represented as black landmarks in the same plot (Schüffler *et al.* 2013b).



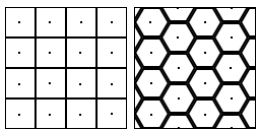
**Figure 4.3: Precision/Recall curve for the nucleus detection via color deconvolution with varying radius  $r$ .** Each curve represents one of eight TMA images. Higher radius  $r$  reveals less nuclei and therefore influences the precision and recall. The automatically detected nuclei were validated against the labels of a trained pathologist. The inter-expert values are denoted for all eight images (top right) as the precision and recall of one pathologist to “match” the other.

#### 4.2.2 Superpixel Based Nucleus Detection

As pathological staining estimation commonly requires prior nucleus classification to identify the subset of *malignant* nuclei, we incorporate a proper segmentation and classification system using superpixels. The idea is to envelope roundish nuclei with superpixels which can then be classified as (cancerous) nuclei or background (Schüffler *et al.* 2013b).

##### Superpixels

For a fast nucleus segmentation, we make use of the superpixel over-segmentation algorithm “Simple Linear Iterative Clustering” (SLIC) (Achanta *et al.* 2012). Superpixels are connected image pixel clusters which show a unique image characteristics such as intensity or morphology. SLIC fully partitions the underlying image into roughly equally sized segments, the superpixels. To this end, the seeds or cluster centers of the superpixels are distributed over the image in a regular grid. Then, pixels in the local environment are iteratively assigned to a cluster center according to intensity and spatial constraints. We implemented an adapted version of the SLIC algorithm, where a comb-shaped prior structure of the superpixels supports the roundish shape of the nuclei (see



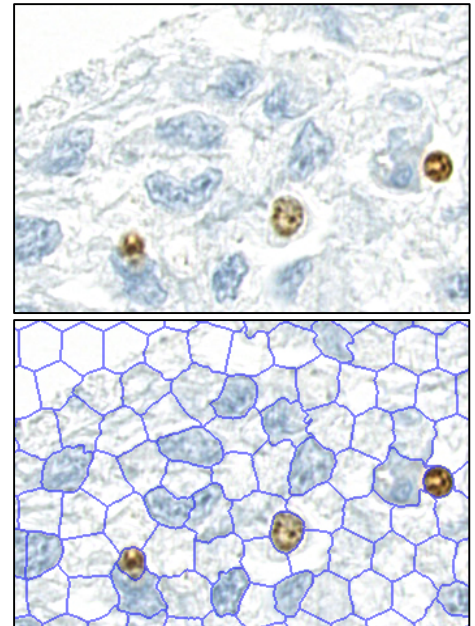
**Figure 4.4: LEFT: Original SLIC implementation. The uniform seeds result in quadratic superpixels. RIGHT: Our implementation favors comb-shaped roundish superpixels.**

Figure 4.4). The size of a superpixel should roughly cover the typical size of a nucleus. The number of superpixels  $n$  for an image with width  $w$  and height  $h$  can therefore be estimated as  $n = (w * h)/(4 * r^2)$ , where  $r$  is the typical nucleus radius. Superpixels have several advantages for computational staining estimation:

- They provide a fast and unsupervised segmentation which is favorable in a clinically used software program.
- The over-segmentation of the image can be used for the whole staining estimation pipeline: in the first stage, nucleus detection is based on nucleus superpixels and background superpixels. In the second stage, the same superpixels can be classified into malignant and benign.
- They enable a fast processing of the image compared to pixel-wise classification since the number of samples is determined by the number of superpixels and not by the size of the image.

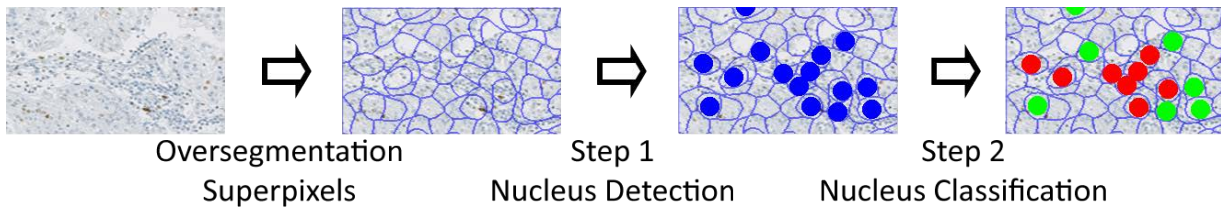
See Figure 4.5 for an example of superpixels. However, their main disadvantages are:

- SLIC superpixels are not scaling invariant. Thus, images with nuclei highly varying in size might not profit from the superpixel approach. Nuclei might be partitioned into two or more superpixels.
- The smoothness of the segmentation can be regulated by a parameter which weights spatial constraints (smooth) against intensity constraints (precise) for pixel assignments to the superpixel centers. Still, irregular shape of abnormal cell nuclei especially on images with poor differentiation of background and foreground might be a problem.



**Figure 4.5: Example of Superpixels on a ccRCC image.**

The processing of TMA images with superpixels enables the implementation of the nucleus detection, segmentation and classification as a two-step classification scenario. In the first step, nucleus detection is solved as classification of *foreground* and *background* superpixels. The labels for training the classifier are given by the user. Alternatively, the background labels can be generated via Voronoi-sampling over a fully annotated image (see below). In the second step, all foreground superpixels are classified as malignant or benign, according to the training labels given by the user. The TMA staining estimation pipeline therefore is reordered with the (over-)segmentation in the beginning (see Figure 4.6).



**Figure 4.6: Scheme of two-stage nucleus classification with superpixels.** The SLIC over-segmentation segments cell nuclei and other structures. In step 1, cell nuclei are detected with a binary classifier for foreground (nuclei) and background. In step 2, the foreground superpixels are classified as malignant or benign.

### Voronoi-Sampling for Background Labels

Voronoi sampling for TMA has been described by Fuchs *et al.* (2009). To train a classifier in step 1, examples of foreground and background are needed. Since in our case, two pathologists exhaustively fully annotated eight TMA images identifying all nuclei, we can assume that not annotated pixels belong to background. A Voronoi diagram around all annotated nuclei reveals loci with largest distance to the surrounding nuclei (Figure 4.7). These loci form the background samples. To reduce the number of samples with similar image information (Figure 4.7 D, multiple background samples on one locus), a post-processing of the Voronoi sampling filters overlapping background labels (Figure 4.7 E). In total, 1584 background loci have been sampled.

### Step 1: Superpixel Based Nucleus Detection as Classification Problem

The tissue image is first partitioned into superpixels which are used as samples to train a binary foreground/background classifier. From each superpixel, a feature vector is calculated considering three features which proved valuable for histology in the past: color histograms (3\*16 bins), local binary patterns (LBP) (Ahonen *et al.* 2004) (size 256), and pyramid histograms of oriented gradients (PHOG) (Bosch *et al.* 2007) (size 338). These features are detailed in section 4.4. The resulting concatenated feature vector has a length of 642. We employed a random forest (Breiman 2001) as default classifier. Based on the labels provided by a pathologist and by Voronoi-sampling, the classifier learns to discriminate between superpixels which represent a nucleus (foreground) and superpixels belonging to the background. The foreground superpixels are subjected to the subsequent nucleus classification.

### Step 2: Superpixel Based Nucleus Malignancy Classification

After the detection of the superpixels with inherent cell nuclei, the goal is to classify them into malignant and benign. We are using the same feature vector as before, but the classifier is now trained only on cell nuclei

labeled by the pathologist, ignoring the background superpixels and labels. Figure 4.8 illustrates the superpixel segmentation and superpixel classification of an example image. For illustration purposes, we trained the classifier to discriminate between stained and unstained cell nuclei (rather than between malignant and benign cell nuclei).

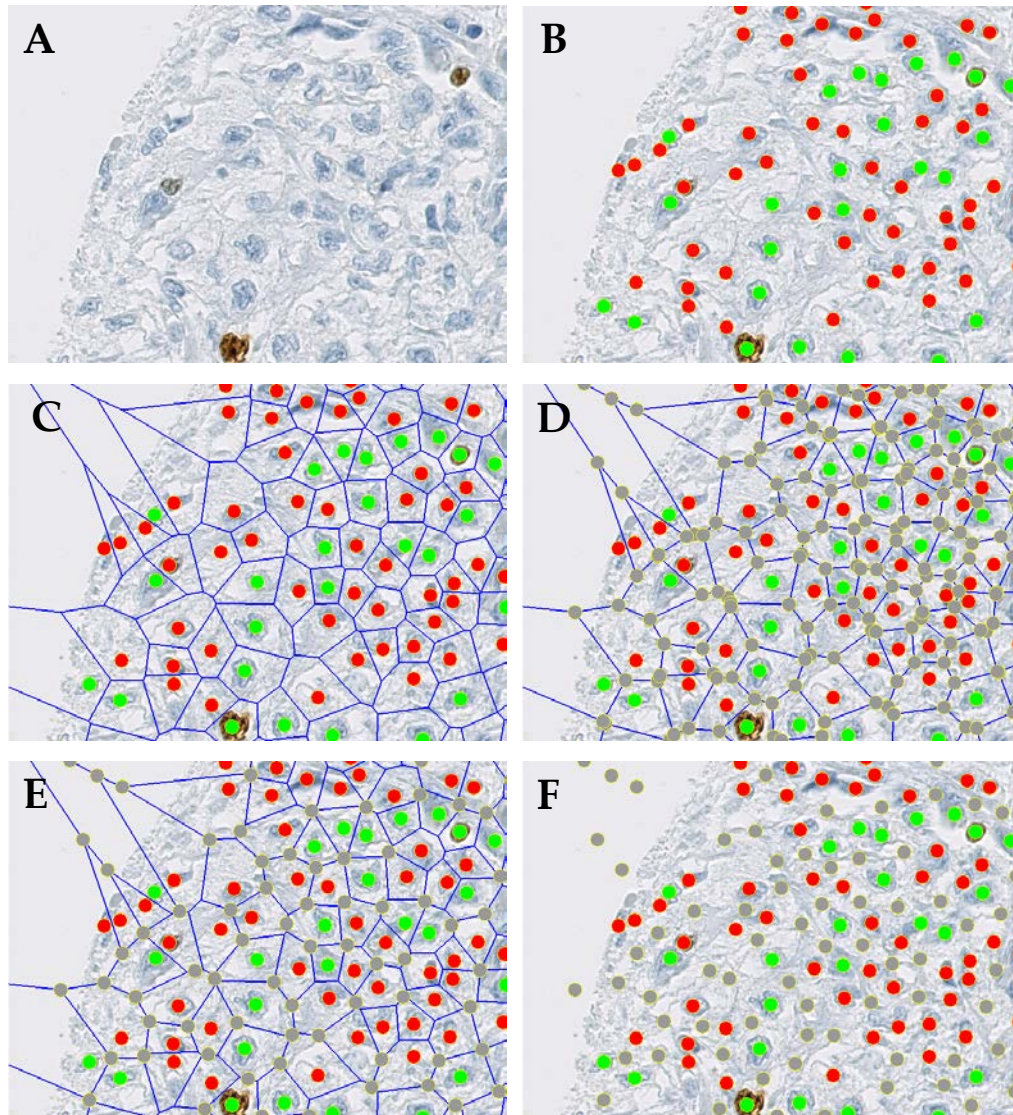
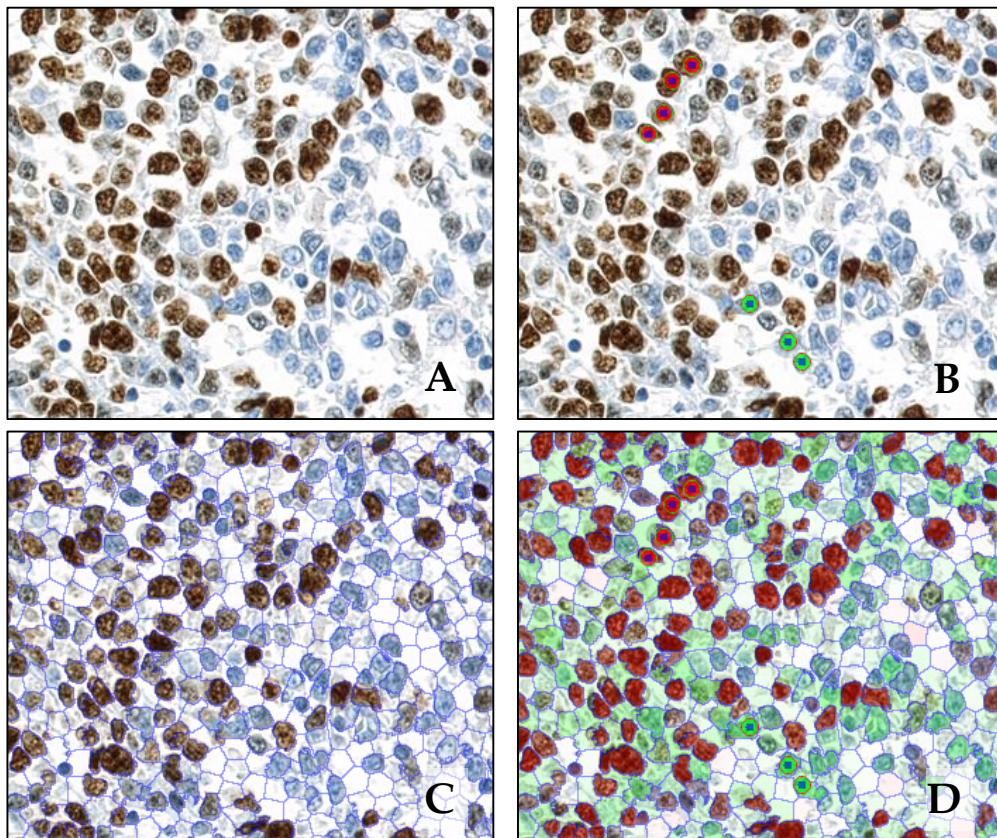


Figure 4.7: Illustration of Voronoi-sampling for TMA images to get background labels. A: The original image patch. B: A pathologist labeled all malignant (red) and benign (green) nuclei. C: A Voronoi diagram tessellates the nuclei in the image. D: The nodes of the Voronoi graph have largest distance to the surrounding nuclei and form the background samples. E: Overlapping background points are filtered in a post-processing step: For each locus, all background loci within a radius  $r$  are merged to one locus. F: The resulting labeled image without Voronoi diagram.

### Validation of Nucleus Detection and Classification via Superpixels

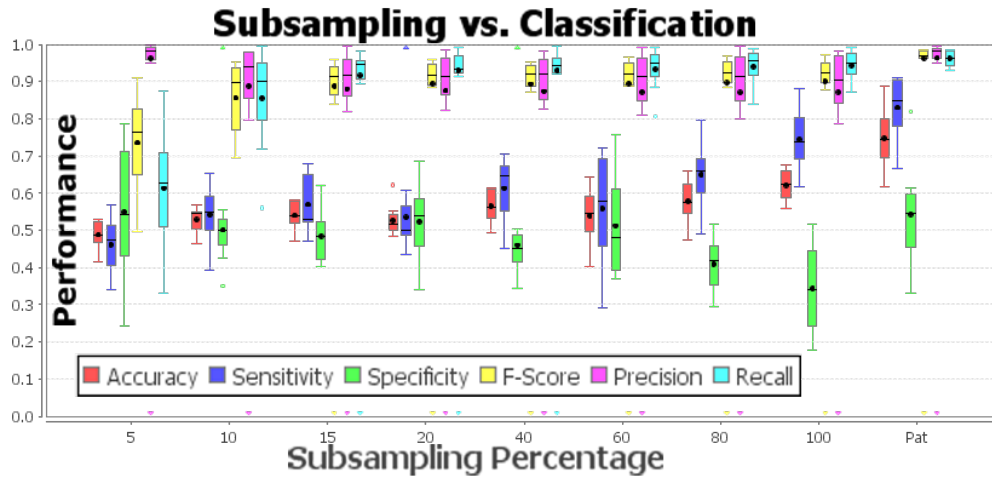
Eight ccRCC IHC images have been labeled by two pathologists who independently identified and classified all visible cell nuclei. A leave-one-patient-out cross-validation has been performed, in which 7 TMA images serve as training set for the classifier which is then tested on the remaining TMA image. The quantitative detection accuracy (F-Score) of 92% and classification accuracy of 64% touches the accuracy range achieved by trained pathologists with an inter-expert accuracy of 97% and 74% on the same dataset, respectively. This holds also true for the sensitivity and specificity for nucleus classification (see Figure 4.9).



**Figure 4.8: Superpixels for cell nucleus segmentation, detection and classification. A: Original image. B: A user labelled positive and negative cell nuclei (red and green). C: SLIC superpixels segment nuclei. D: Superpixels are classified into red and green superpixels according to the user labels. The color intensity reflects the class probability.**

We also tested the performance when only a fraction of the available training set is incorporated. The high level **detection** accuracy is already reached, when only 15% of all available nuclei are used for training (Figure 4.9, yellow bars). However, the cell nucleus **classification** profits from more samples, and the accuracy steadily improves until 100% of available training samples are used (see Figure 4.9, red bars). *Detection works faster than classification.*






**Figure 4.9:** Performance of nucleus detection and classification via superpixels in 8 TMA spots. Depicted are precision, recall and F-score for the nucleus detection, as well as sensitivity, specificity and accuracy for the nucleus classification. Experiments were conducted with training set sizes from 5% to 100% (x-axis). Each box represents a leave-one-patient-out cross-validation. The performance stabilizes with 15% of training samples. The inter-expert performances of two pathologists is plotted last (“Pat”): for each of the 8 images, pathologist A is taken as reference for pathologist's B guesses.

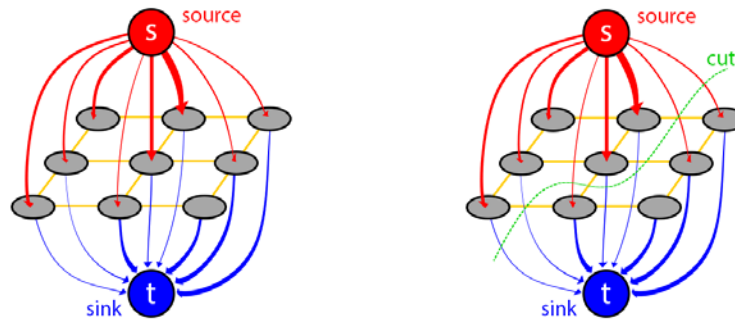
### 4.3 Nucleus Segmentation via Graph-Cut

It is known from pathology that cancer nuclei tend to alter their size and shape compared to normal nuclei. This fact is extremely helpful for the detection and classification of cell nuclei (Schüffler *et al.* 2010). Table 4.2 lists the most important morphological differences between normal and cancerous nuclei in ccRCC tissue. E.g. benign and healthy nuclei appear roundish and regular whereas malignant nuclei are commonly larger and more irregular in shape.

To exploit the shape information of cancerous nuclei, we suggest a prior binary segmentation of the nucleus boundary via graph-cuts (Boykov *et al.* 2001; Boykov and Funka-Lea 2006). Graph-cut represents the image pixels as a meshed weighted graph with two additional nodes  $s$  (source, foreground) and  $t$  (sink, background) to which every pixel is connected. The graph is then cut with a minimum  $s$ - $t$  cut to separate the foreground from the background (see Figure 4.10). The segmentation information is thus encoded in the weights of the graph.

**Table 4.2: Typical guidelines for pathologists to distinguish normal and malignant ccRCC nuclei. The image patches show example nuclei from our ccRCC dataset with both pathologists agreeing on the label.**

	Normal ccRCC nucleus	Cancerous ccRCC nucleus
<b>Shape</b>	Roundish	Irregular
<b>Membrane</b>	Regular	Thick/thin irregular
<b>Size</b>	Smaller	Larger
<b>Nucleolus</b>	None	Dark spot in the nucleus
<b>Texture</b>	Smooth	Irregular
<b>Example</b>		



**Figure 4.10: Concept of graph-cut segmentation for images. The gray nodes represent the image pixels. Source and sink represent foreground and background. LEFT: The weights for the graph (red, yellow and blue edges) are properly set. RIGHT: A minimum cut (green) partitions the graph into foreground and background pixels.**

Figure 4.11 illustrates the graph-cut process for our nucleus segmentation. To favor the naturally roundish shape of nuclei, we incorporate a roundish shape prior in the weights of  $s$  and  $t$ . The nucleus image patch centering the nucleus is gray-scaled and smoothed with a Gaussian smooth filter of radius 5, which worked best for segmentation. The edges from the source to the image pixels are weighted by the normalized squared distance of the pixels to the image center. The edges from the image pixels to the sink are weighted by 1 minus the normalized squared distance of the pixels to the image center. The meshed graph weights are initialized with the intensity differences between neighbor-pixels. After cutting, the connected component in the middle of the patch represents the shape of the nucleus. Figure 4.12 depicts five typical examples. The contours of the nuclei were subjected to shape related feature extraction.

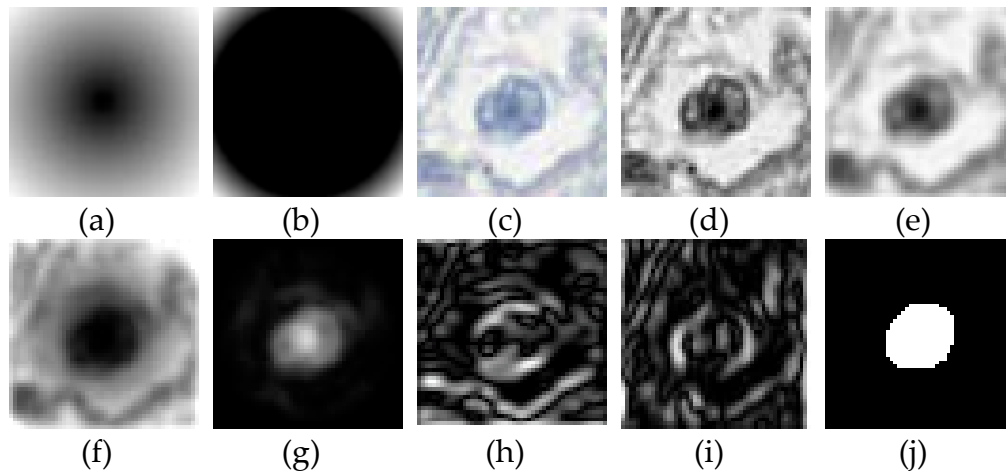


Figure 4.11: Illustration of segmentation via graph-cut. (a), (b): roundish shape priors of source and sink. (c), (d) original image patch with centered nucleus and gray scaled version. (e): smoothed gray-scaled nucleus patch. (f), (g): the roundish priors from a) and b) applied on the image patch define the source and the sink weights for the graph. (h), (i): the North-South and East-West difference image of e) form the weights for the connected graph. (j): a max-flow algorithm cuts the graph into the final segmentation.

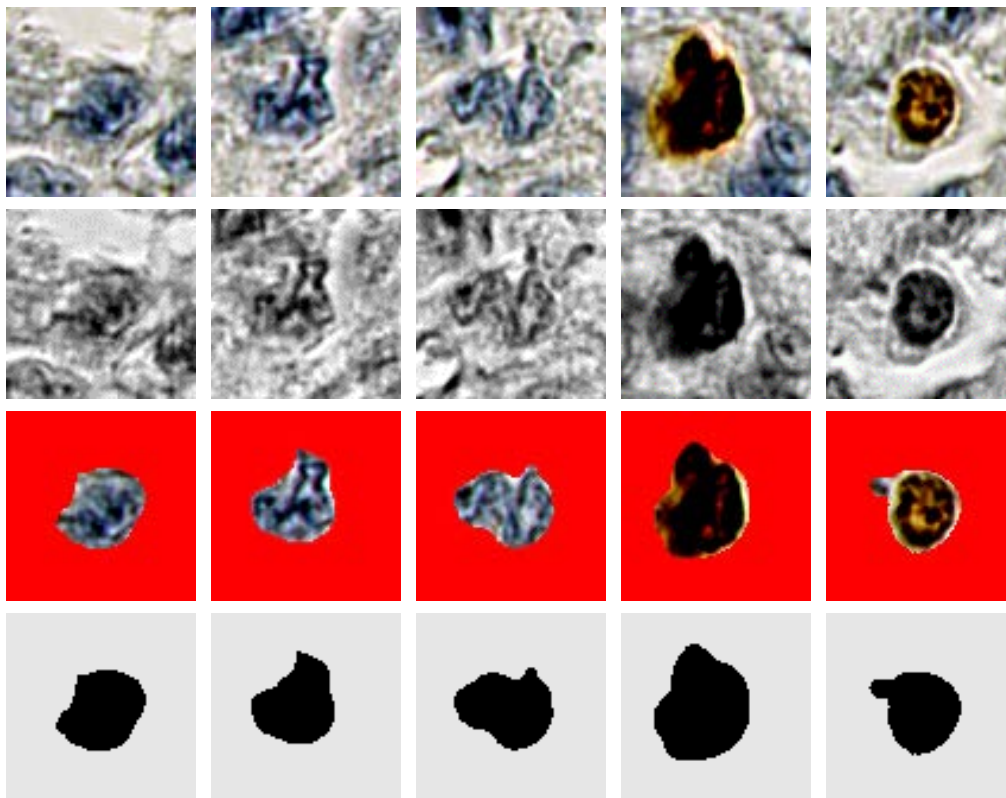


Figure 4.12: Five examples for nucleus segmentation via graph-cut. TOP: Original 80x80px image patches with nucleus centered in the image. 2<sup>nd</sup> ROW: The patches are gray-scaled. 3<sup>rd</sup> ROW: The graph-cut algorithm with a roundish shape preference cuts the centered object. BOTTOM: The resulting nucleus shape.

## 4.4 Feature Extraction

To get a comprehensive description of the nuclei, we use following set of image and shape related features. All nuclei are represented by 80x80px image patches with the nucleus centered in the patch. All histogram-like features are normalized to unique sum.

### 4.4.1 Histogram of Patch Intensity (ALL)

A 32-bin intensity histogram is calculated over the whole gray scaled nucleus patch, comprising the nucleus in the center and the immediate environment. The histograms are normalized to unique sum.

### 4.4.2 Color Histogram (COL)

The patch is rescaled to size 5x5px. The 25 intensity values of the three color channels red, green and blue are concatenated to a feature vector of length 75.

### 4.4.3 Histogram of Foreground Intensity (FG)

A 32-bin intensity histogram is calculated over the segmented nuclear area. A cancer cell's nucleolus is represented as dark spot in a nucleus and is to be represented in a nuclear intensity histogram.

### 4.4.4 Histogram of Background Intensity (BG)

Similar to the foreground histogram, we calculate a 32-bin intensity histogram of the surrounding area of a nucleus (i.e. of the background area in the patch). This can capture differences of the microenvironment of a nucleus.

### 4.4.5 Freeman Chain Code (FCC)

The FCC is a string representation of a shape's boundary (Freeman 1961). The boundary is pixel-wise locally described by a number of 1 to 8, according to the direction in which the boundary points at that pixel. To be rotation invariant and scale invariant, the FCC is not directly used (it would change with the starting point and with the size of the boundary). Instead, the 8-bin histogram of the first derivative of the FCC forms the final FCC feature.

### 4.4.6 1D-Signature (SIG)

This shape descriptor is especially suitable for closed and roundish objects. Starting from the center of the nucleus, the radius of the object is

measured in all circular 360 directions (Gonzalez *et al.* 2004). To respect rotational invariance, the nucleus shape was first rotated to the direction with maximum radius (note that this rotation might be sensitive to outliers). The signature is normalized by the maximum value to be scale invariant. A 16-bin histogram of the signature creates the final signature feature.

#### 4.4.7 *Pyramid Histogram of Oriented Gradients (PHOG)*

PHOG has been presented by Bosch *et al.* (2007). First, an 8-bin intensity histogram over the whole image is calculated. Then, the image is quartered and four intensity histograms for the quarters are calculated. Thereafter, each quarter is again quartered into smaller sub-images and histograms are calculated. The resulting histograms are concatenated to a 168-bin feature vector.

#### 4.4.8 *Region Properties (PROP)*

Additionally to these elaborated features, we collected standard region descriptors to the feature vector. Area size, bounding box size, major axis length, minor axis length, eccentricity, convex area, equivalent diameter, solidity, extent, perimeter, mean intensity, min intensity and max intensity of the nucleus region were measured. To count for their relative impact rather than absolute, the PROP features were normalized to sum up to one.

#### 4.4.9 *Local Binary Patterns (LBP)*

Local Binary Patterns are illumination invariant and showed advantageous behavior in medical image processing (Fuchs *et al.* 2008a). For each pixel of the gray-scaled patch, an 8-bit string is generated where each bit corresponds to one neighbor of the pixel. A bit is set to 1, if the neighbor's intensity value is smaller than the intensity of the original pixel, and otherwise to 0. The resulting binary number is converted to a decimal number and a 256-bin histogram is calculated to capture the distribution of local binary patterns.

## 4.5 *Classification with Support Vector Machines*

Support vector machines (Schölkopf and Smola 2002) are in widespread use and highly successful for bioinformatics tasks (Ben-Hur *et al.* 2008). SVMs exhibit very competitive classification performance with similarity based classification, and they can conveniently be adapted to the specific problem at hand. This adaptation is achieved by designing individual

kernel functions. Kernel functions can be seen as problem specific similarity functions between examples. A kernel function implicitly maps examples from their input space  $X$  to a Hilbert space  $\mathcal{H}$  of real-valued features (e.g.  $\mathcal{H} = \mathbb{R}^n$ ,  $n \in \mathbb{N} \cup \{\infty\}$ ) via an associated function  $\Phi: X \rightarrow \mathcal{H}$ . For two samples  $u, v \in X$ , the kernel function  $k$  provides an efficient method for computing dot products in the feature space  $\mathcal{H}$  via:

$$k(u, v) = \langle \Phi(u), \Phi(v) \rangle$$

The resulting optimization problem is convex and the global optimum can be found efficiently, for which many freely available software packages can be used. The essential question is the choice of features or kernel functions  $k(u, v)$  to be used for a particular problem.

To use a dedicated distance function for nucleus classification as a kernel function, we calculated a  $n * n$  squared distance matrix between all  $n$  samples. For explicit kernel functions, the resulting kernel matrix  $K$  is symmetric and positive semi-definite – a requirement for support vector machines. Other distance functions may result in non-metric distance matrices  $D$  (e.g. non-symmetric or not positive semi-definite). To use such a distance matrix  $D$  as kernel matrix  $K$ , it has to be centered to zero mean:

$$D_{centered} = -0.5 * Q * D * Q$$

where

$$Q = \begin{bmatrix} 1 - \frac{1}{n} & & -\frac{1}{n} \\ & \ddots & \\ -\frac{1}{n} & & 1 - \frac{1}{n} \end{bmatrix}$$

Then,  $D_{centered}$  is checked for being positive semi-definite. Negative Eigenvalues are mirrored to calculate the corresponding positive semi-definite kernel matrix  $K$ :

$$K = V * |\Lambda| * V'$$

where  $V$  is the Eigenvector matrix and  $\Lambda$  is the Eigenvalue diagonal matrix of  $D_{centered}$ . The resulting kernel matrix  $K$  can be used for SVM classification (Schüffler *et al.* 2010). For a comparison of different similarity measures, we incorporated 10 kernel functions and 8 distance functions as listed in Table 4.3.

Table 4.3: Kernel and distance functions for nucleus classification.  $u$  and  $v$  are scalar feature vectors of length  $p$ . For the histogram-like feature vectors, all functions were used. For the PROP feature, only the linear, polynomial and Gaussian kernel functions were applied.

### Kernel Functions

Linear	$u' * v$
--------	----------

Polynomial (degree $d \in \{3, 5, 7, 10\}$ )	$\left(\frac{u' * v}{p}\right)^d$
--	-----------------------------------

Gaussian	$e^{-\frac{1}{p}\sum_i(u_i * v_i)^2}$
----------	---------------------------------------

Hellinger (1909)	$\sum_i \sqrt{u_i * v_i}$
------------------	---------------------------

Jensen Shannon	$-\frac{1}{\log 2} \sum_i \left[ u_i \log \frac{u_i}{u_i + v_i} + v_i \log \frac{v_i}{u_i + v_i} \right]$
----------------	---

Total Variation	$\sum_i \min(u_i, v_i)$
-----------------	-------------------------

$\chi^2$	$\sum_i \frac{u_i * v_i}{u_i + v_i}$
----------	--------------------------------------

### Distance Functions

Euclidean	$\sqrt{\sum_i (u_i - v_i)^2}$
-----------	-------------------------------

Intersection	$\min\left(\sum_i u_i, \sum_i v_i\right) * \left(1 - \frac{\sum_i \min(u_i, v_i)}{\min(\sum_i u_i, \sum_i v_i)}\right)$
--------------	---

Bhattacharyya (1943)	$-\log \sum_i \sqrt{u_i * v_i}$
----------------------	---------------------------------

$\chi^2$	$\sum_i \frac{(u_i - v_i)^2}{u_i + v_i}$
----------	--

Kullback Leibler (1951)	$\sum_i u_i \log \frac{u_i}{v_i} + \sum_i v_i \log \frac{v_i}{u_i}$
-------------------------	---

Earth Mover (Rubner <i>et al.</i> 2000)	$\sum_{i=1}^p \left  \sum_{j=1}^i u_j - v_j \right $
---	--

$$\sum_{l=0}^L |d_l(x)|$$

**Diffusion (Haibin and Okada 2006)** with

$$d_0(x) = u - v$$

$$d_l(x) = [d_{l-1}(x) * \phi(x, \sigma)] \downarrow_2$$

where  $\phi$  is a Gaussian filter with standard deviation  $\sigma$ .

$$\ell_1 \quad \sum_i |u_i - v_i|$$

#### 4.5.1 Experimental Design

The experiments for nucleus classification using SVM were conducted with Matlab (2010) and the libSVM package (Chang and Lin 2011). 1273 ccRCC nucleus patches with unique label (cancerous or benign) from two pathologists were subjected to nucleus classification. Each nucleus is represented as 80x80px patch centering the nucleus. Graph-cut segmentation further segmented the nuclei as explained before. Six feature vectors were extracted per nucleus (FG, BG, FCC, SIG, PHOG, PROP) and 18 kernel matrices were calculated for each feature with the 18 kernel and distance functions listed in Table 4.3 (polynomial kernel with degrees 3, 5, 7 and 10). Kernel matrices for different features were normalized by their trace and combined (summed up) to a SVM classification matrix. The normalization is important to cope with matrices and features on different scales. The addition of kernel matrices enabled the combination of different features as well as the combination of different distance measures.

Support vector machines were trained and validated in a 10-fold cross-validation. The capacity parameter  $C$  of the SVM was set to 0.1, 1, 10, 100 and 1000, to search for the optimal parameter. All cross-validated models were ranked and features and distance measures of the top performing parameters were investigated (Schüffler *et al.* 2010).

#### 4.5.2 Results

The classification task can efficiently be solved with models using all features FG, BG, FCC, SIG, PHOG and PROP. Due to the patch-wise approach, nuclei on the image borders were excluded, and 1273 nuclei of all 1379 with consistent label have been used for this study. The median classification accuracy of the best model is 83%. Figure 4.13 plots the misclassification error of the 15 best models and 16 typical median models. These models all lie in the range or are slightly better than the guess of

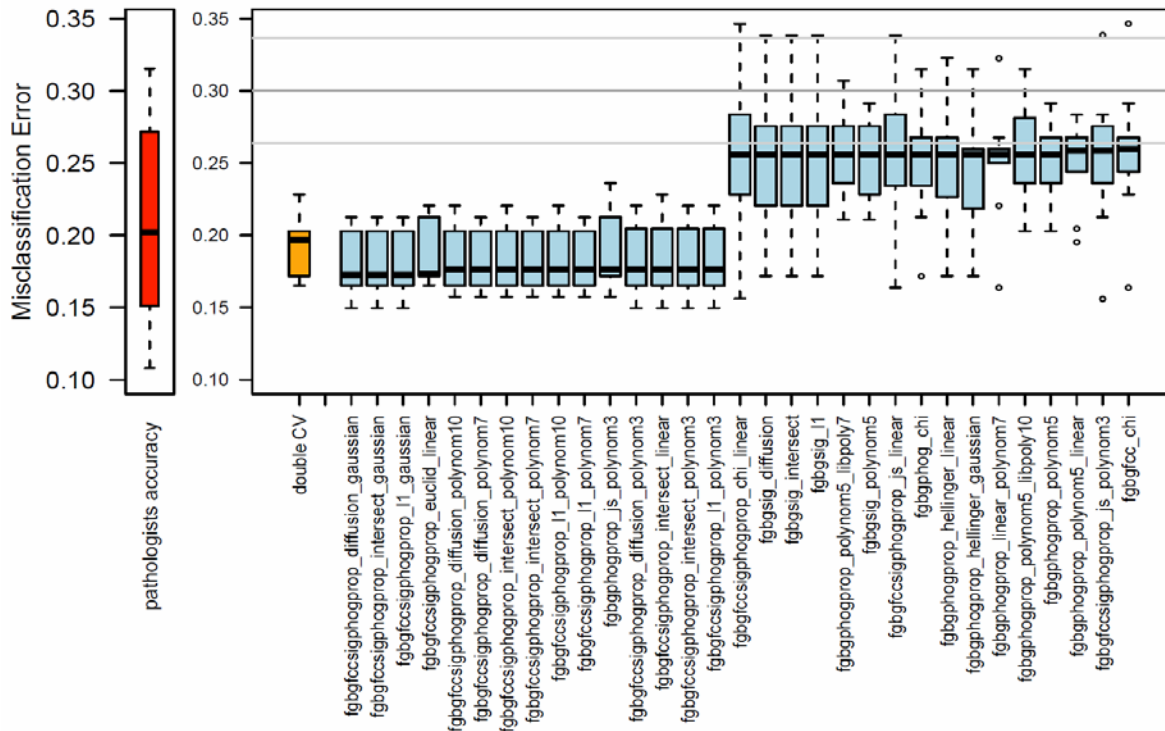


two pathologists: their median “misclassification disagreement” is 20%, meaning they disagree on every 5<sup>th</sup> nucleus. To evaluate a random level of classification performance, we cross-validated the best performing model (using all features and diffusion distance) on the dataset with randomly permuted label. 100 repetitions for hundred permutations resulted in a mean misclassification error of  $30 \pm 3.5\%$ , meaning that the information in the data can be exploited with our model.

Further, we tested our approach for overfitting. A double layer 10-fold cross-validation was performed on the whole dataset of 1273 nuclei. The outer layer divided the dataset into 90% training data and 10% test data. On the training data, our approach as described above is run to find the best model in this fold, forming the inner layer. The top model is then tested on the 10% “external” test data which have never been seen during model training and ranking. The misclassification error on the test data is recorded and the scenario is repeated as the next fold in the outer cross-validation. The 10 test errors have a median value of 20%, and lie in the same range as the cross-validation errors itself (see Figure 4.13, orange model “double CV”). This indicates that our cross-validated models are not overfitted, since they can hold their performance even on new, unseen data which have not been used for training. Interestingly, in six of the 10 folds, the best ranked model used all features with the diffusion distance as divergence measure for the histogram-like features, indicating this model’s stability.

## 4.6 Shape Descriptors Boost the Classification Performance

To answer the question, how far shape descriptors influence the performance of the ccRCC nucleus classification (malignant vs. benign), we grouped the features into intensity features (FG and BG), shape features (FCC, SIG and PROP) and PHOG as a combination of both (Schüffler *et al.* 2010). With these groups, a double layer cross-validation was performed as described before. On our ccRCC dataset, the intensity histograms solely show a median misclassification error of 23%. The error is considerably lowered when shape features are incorporated: the misclassification error drops to 21% (with shape features) or to 19% (with PHOG feature), respectively (see Figure 4.14). This consistent classification improvement motivates the nucleus segmentation with graph-cuts. Interestingly, PHOG seems to inherit shape information: The combination of all three feature types does not improve the nucleus classification accuracy compared to intensity and PHOG alone.



**Figure 4.13: Cross-validated nucleus classification.** Shown are the 15 top models and 16 medium models (blue boxes). Each model is named by the features involved and the distance measure used to calculate the kernel matrices. For the PROP feature, only linear, polynomial or Gaussian kernels are used since PROP is not a typical histogram-like feature. The best model uses all features FG, BG, FCC, SIG, PHOG and PROP with the diffusion distance (respectively Gaussian kernel for PROP). The median misclassification rate of this model is with 17% lower and less variant than the inter-rater disagreement of the two pathologists (red box, median 20%). Mean and standard deviation of 100 permutation tests are drawn as gray lines (random level, 30%). An additional 10-fold cross-validation level was employed to test for overfitting. The best found model of each fold was tested on the unseen samples, yielding one misclassification error of the orange box (median 20%). This external validation procedure reveals that the best model of a run holds its performance also on new data and lowers the risk of overfitting.

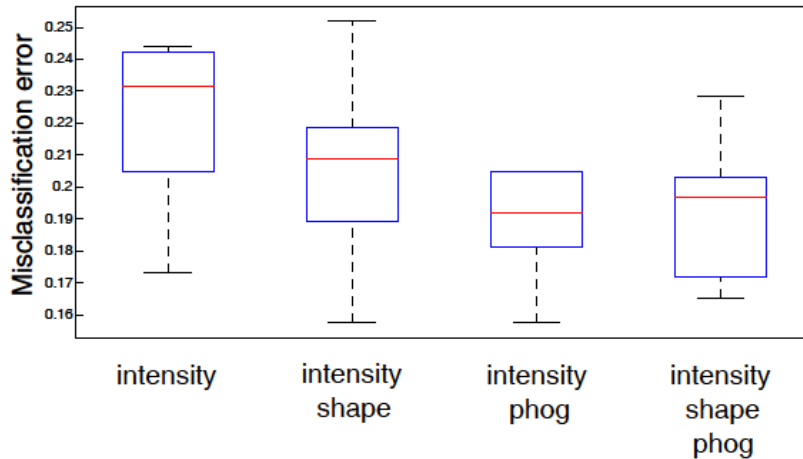


Figure 4.14: Classification accuracy of 1273 ccRCC benign or malignant nuclei with three feature groups *intensity* (FG, BG), *shape* (FCC, SIG, PROP) and *PHOG*. The misclassification rate lowers when shape features are included in the classifier.

## 4.7 Better Nucleus Classification for Better Staining Estimation

To estimate the effect of nucleus classification on the TMA staining estimation pipeline, we pulled the diverse model performances and the staining estimation together. As mentioned before, the staining estimation only refers to cancerous stained nuclei. Therefore, it can be expected that a more accurate classification of nuclei would result in a more accurate staining estimation (Schüffler *et al.* 2010).

The staining color of the nucleus patches can be separated with a single threshold approach. A color histogram in the 30x30px center of each patch encodes the nucleus' predominant color. Since blue and red are clearly separable color channels, the staining information of a patch can be expressed as a fraction  $f$  of the red channel intensity mean ( $r$ ) and the blue channel intensity mean ( $b$ ):

$$f = \frac{r}{b}$$

If  $f > 1$ , the nucleus is considered stained, and non-stained otherwise.

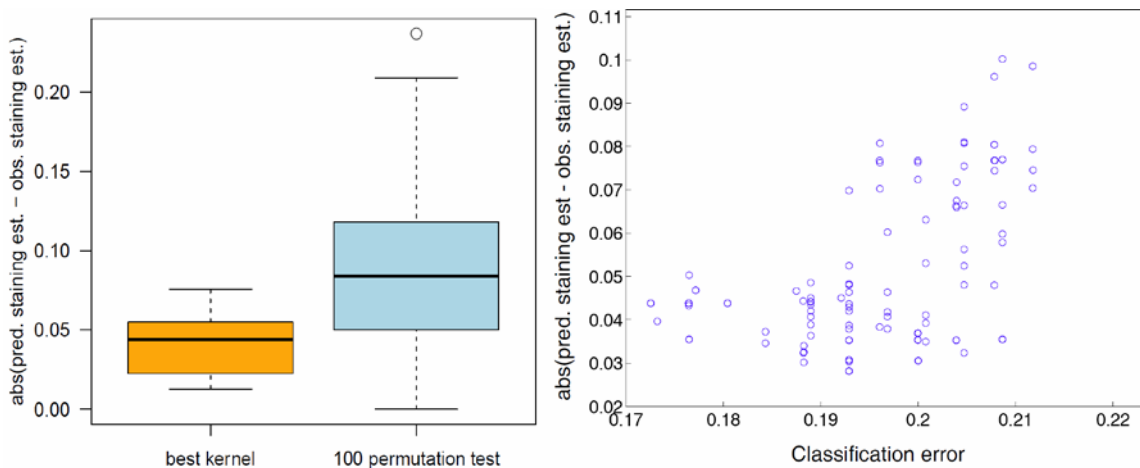
### 4.7.1 Experimental Design

The best ranked SVM model from 4.5.2 was used for this experiment. The kernel was trained in a 10-fold cross-validation scenario, similar as before. But instead of evaluating the misclassification rate on the nuclei in the test set, we calculated the staining estimation on the predicted cancerous nuclei in the test set. The predicted staining estimation is then

subtracted from the staining estimation of a pathologist. The absolute difference of the predicted and the annotated staining estimation should be as small as possible, if the task is to mimic the pathologists staining procedure.

#### 4.7.2 Results

The computationally calculated staining estimation has a deviance of only 5% from the pathologist (Figure 4.15 left). A permutation test shows that guessing the nucleus label without any knowledge can yield a staining estimation difference of up to 20%. Figure 4.15, right shows a plot of the classification error of the best kernels with a misclassification error below 22% versus the staining estimation error clearly shows that the correct classification of cell nuclei as part of the pipeline has a positive effect on the predicted staining estimation: a better nucleus classification propagates to a more accurate staining estimation. This is a nice motivation to optimize single steps in the pipeline for a holistic improvement of the global medical task.



**Figure 4.15: A more accurate cell nucleus classification improves the predicted staining estimation. LEFT: Our best nucleus classification model (orange) predicts staining estimation in a 10-fold CV with 5% deviance to a pathologist. A random model has a twice as high error rate and reaches up to 20% deviance to a pathologist. RIGHT: The more accurate the nucleus classification is performed as a single step in the TMA analysis pipeline, the more accurate is the subsequent staining estimation compared to a pathologist. Each point represents one top classifier with classification error < 22%. Shown are the absolute differences of predicted and observed staining estimation.**

## 4.8 Multiple Kernel Learning for Nucleus Classification

We consider in this section a Multiple Kernel Learning (MKL) framework for nuclei classification of renal cell carcinoma (Schüffler *et al.* 2011). The features extracted from the nuclei are identical to these introduced in section 4.4. MKL is then applied for classification. We compare our results with an incremental version of MKL, SVM with single kernel and voting. We demonstrate that MKL inherently combines information from different input spaces and creates statistically significantly more accurate classifiers than single kernel SVMs and voting for renal cell carcinoma classification on nuclear level.

### 4.8.1 Introduction

For various classification tasks, SVM classifiers use one data set and maximize the margin between different classes. This poses a restriction on some problems, where different data representations are used. Combining the contribution of different characteristics and properties is especially important in discriminating between cancerous and healthy cells. MKL is a recent and promising paradigm, where the decisions of multiple kernels are combined to achieve better accuracies (Bach *et al.* 2004). MKL allows the beneficial utilization of data from multiple sources. We compare MKL using global combination of multiple kernels with the conventional combination of outputs of multiple classifiers.

We used 1273 nucleus image patches with consistent label from 8 renal clear cell carcinoma TMA images. The images have been immunohistochemically stained against the proliferation protein MIB-1 (Ki-67 antigen), as described in section 3.1. The nuclei on these images have been labeled by two trained pathologists as cancerous or benign. Patches have the size of 80x80 pixels and center a cell nucleus, each. 891 (70%) of the nuclei are benign, 382 (30%) are malignant. Nuclei were segmented via graph-cut as described in section 4.3. ALL, BG, COL, FCC, FG, LBP, PHOG, SIG and PROP features for each nucleus have been extracted as explained in section 4.4.

### 4.8.2 MKL Framework

The main idea behind support vector machines is to transform the input feature space to another space with a possibly greater dimension, where the classes are linearly separable. After training, the discriminant function of SVM becomes  $f(x) = \langle w, \Phi(x) \rangle + b$ , where  $w$  are the weights,  $b$  is the threshold and  $\Phi(x)$  is the mapping function. Using dual formulation

and the kernels one does not have to define this  $\Phi(x)$  explicitly and the discriminant results as:

$$f(x) = \sum_{i=1}^N \alpha_i y_i k(x, x_i) + b$$

where  $k(x, x_i)$  is the *kernel*,  $N$  is the number of samples with label  $y$ , each, and  $\alpha$  is the weight of each sample. A single kernel SVM will be restricted to the use of one feature set (or a concatenation of all feature sets) and thus complicates the possibility to exploit the manifold information coming from different sources. As known in classifier combination (Kuncheva 2004), the combination of multiple kernels using different feature sets can come up with more accurate classifiers (Lee *et al.* 2007). In a simple way, this can be achieved by using an unweighted sum of kernel functions (Moguerza *et al.* 2004). Lanckriet *et al.* (2004) formulated this semi definite programming problem which allows finding the combination weights and support vector coefficients simultaneously. Bach *et al.* (2004) reformulated the problem and proposed an efficient algorithm using sequential minimal optimization (SMO). Using Bach's formulation with  $P$  kernels, the discriminant function results as:

$$f(x) = \sum_{m=1}^P \eta_m \sum_{i=1}^N \alpha_i y_i k_m(x, x_i) + b$$

where  $\eta_m$  is the weight of the  $m^{\text{th}}$  kernel.

This method allows us to combine different kernels of different feature spaces. In this study, the kernels are combined globally, i.e. the kernels are assigned the same weights for the whole input space. It has been shown by various studies that using a subset of given classification algorithms increases accuracy rather than using all the classifiers (Ruta and Gabrys 2005; Ulař *et al.* 2009). Against this background, we apply the same idea to incrementally adding kernels to the MKL framework and compare the results. The incremental algorithm works as follows: Starting with the most accurate kernel (classifier) on the validation folds (leave-the-other-fold-out), kernels (classifiers) are added to the combination one by one. This procedure continues until all kernels (classifiers) are used or the average validation accuracy does not increase (Ulař *et al.* 2009). The algorithm starts with  $E^0 \leftarrow \emptyset$ , then at each step  $t$ , all the kernels (classifiers)  $k_j \notin E^{(t-1)}$  are combined with  $E^{(t-1)}$  to form  $S_j^t = E^{(t-1)} \cup M_j$ . We select  $S_{j^*}^t$  which is the ensemble with the highest accuracy. If accuracy of  $S_{j^*}^t$  is higher than of  $E^{(t-1)}$ , we set  $E^t \leftarrow S_{j^*}^t$  and continue, else the algorithm stops and returns  $E^{(t-1)}$ .

### 4.8.3 Experiments and Results

The data of 1273 nucleus samples is divided into ten folds with stratification. Support vector machines (*svl*, *sv2*, *svg*, see below) and MKL are trained and cross-validated (CV) using these folds. We also combine the support vector machines using voting and report average accuracies using 10-fold CV. For the Gaussian kernel,  $\sigma$  is chosen using a rule of thumb:  $\sigma = \sqrt{D}$  where  $D$  is the number of features for data representation. We compare our results using a 10-fold CV t-test at  $p=0.05$ . In the incremental learning part, we apply leave-the-other-fold-out cross-validation (used for validation) to estimate which kernel and classifier should be added.

Nine nucleus representations are collected in total (ALL, BG, COL, FCC, FG, LBP, PHOG, SIG and PROP), as well as three different kernels (linear kernel: *svl*, polynomial kernel with degree 2: *sv2*, and Gaussian kernel: *svg*), and two combination algorithms (MKL, VOTE). The SVM accuracies with each individual kernel are reported in Table 4.4. The best accuracy using a single SVM is 76.9 %. For most representations (except PHOG and COL), the accuracies of different kernels are comparable.

Next, we use the same kernel and combine all the feature sets we extracted. As shown in Table 4.5 (top), we can reach an accuracy of 81.3% using the linear kernel, combining all feature representations. This fact stresses that the combination of information from multiple sources might be important and, by using MKL, the accuracy can be increased around 5 % compared to single kernel SVM. Further, using all kernels with *sv2*, the accuracy decreases compared to the single best support vector machine (72.0% vs. 76.9%). This is analogous to combining all classifiers in classifier combination. The combination of inaccurate classifiers may decrease the accuracy.

**Table 4.4: Single support vector accuracies ( $\pm$  std) in %.**

	<i>SVL</i>	<i>SV2</i>	<i>SVG</i>
<b>ALL</b>	70.0 $\pm$ 0.2	71.6 $\pm$ 2.9	72.0 $\pm$ 3.2
<b>BG</b>	70.0 $\pm$ 0.2	71.2 $\pm$ 2.6	68.9 $\pm$ 2.3
<b>COL</b>	70.1 $\pm$ 0.2	63.6 $\pm$ 3.5	66.2 $\pm$ 2.3
<b>FCC</b>	70.0 $\pm$ 0.2	70.0 $\pm$ 0.2	67.4 $\pm$ 1.6
<b>FG</b>	70.0 $\pm$ 0.2	70.0 $\pm$ 3.2	70.5 $\pm$ 3.5
<b>LBP</b>	70.0 $\pm$ 0.2	66.9 $\pm$ 3.0	68.7 $\pm$ 4.4
<b>PHOG</b>	76.5 $\pm$ 3.7	72.0 $\pm$ 3.3	<b>76.9 <math>\pm</math> 3.6</b>
<b>SIG</b>	70.0 $\pm$ 0.2	68.6 $\pm$ 2.5	66.6 $\pm$ 2.6
<b>PROP</b>	75.7 $\pm$ 2.3	75.6 $\pm$ 2.6	74.1 $\pm$ 1.8

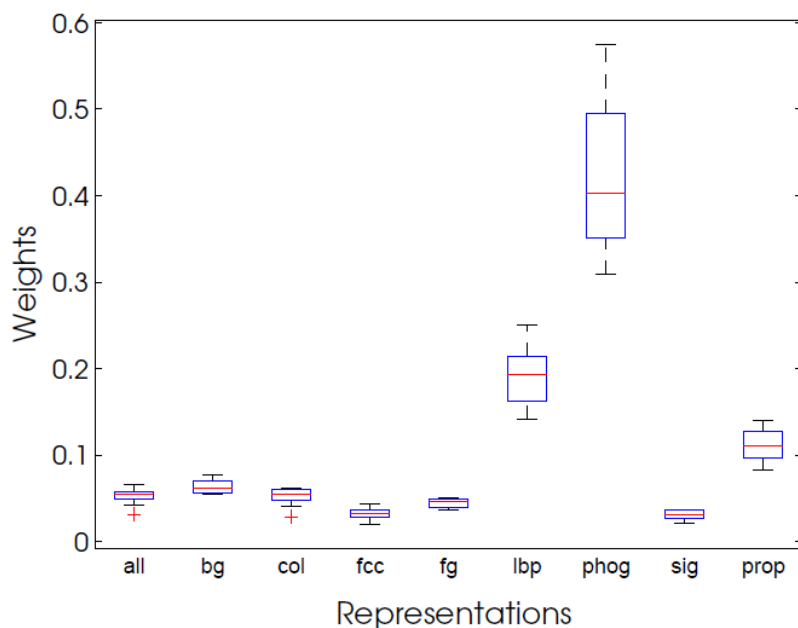
Instead, it might be better to choose a subset of classifiers. A medical interpretation of this result indicates that all the different information of the nuclei is complementary and should be used to achieve better accuracy. In Figure 4.16, the weights of MKL using the linear kernel are plotted. As expected, the two best representations PHOG and PROP have high weights. But the representation LBP, having very low accuracy when considered as a single classifier, increases the accuracy when considered in combination with others. This illustrates that a combination of even solely more inaccurate classifiers might gain in performance, indicating hidden interaction of the different features. From this, we also deduce that these three features are useful in discriminating between malignant and benign cell nuclei and we may focus our attention on these properties.

**Table 4.5: MKL accuracies (in %). TOP: Accuracy ( $\pm$  std) of combining all kernels. BOTTOM: Accuracies calculated using the incremental algorithm, the number  $P$  of kernels / classifiers selected.**

	<i>SVL</i>	<i>SV2</i>	<i>SVG</i>
<b>MKL</b>	<b>81.3 <math>\pm</math> 3.6</b>	72.0 $\pm$ 3.3	76.9 $\pm$ 3.6
<b>VOTE</b>	70.0 $\pm$ 0.2	71.3 $\pm$ 1.7	72.4 $\pm$ 1.2
<b>MKL</b>		76.9 $\pm$ 3.6, $P = 1$	
<b>VOTE</b>		78.9 $\pm$ 2.5, $P = 4$	

Table 4.5 (bottom) depicts the results using the incremental algorithm. There is no increase in accuracy compared to the best single support vector machine. In fact, the incremental algorithm cannot find a second complementary kernel increasing the accuracy when added to the single best. Generally, we expect the incremental algorithm to reach higher accuracies than combining all classifiers. This behavior can be observed for *sv2*, but for *svl*, combining all kernels seems to be better than the subset selection strategy. This might partially result from the optimization formulation of MKL. In the incremental search, we discard kernels which do not improve the overall accuracy. On the other hand, in MKL, every kernel is weighted and all kernels contribute to the solution of the problem. Therefore, we state that the framework of MKL is superior to combining outputs of support vector machines using voting. Table 4.5 supports this statement: using voting, combining all classifiers always results in lower accuracy than the single best kernel and MKL, since the optimization procedure does not “see” the data, but only combines outputs of all classifiers. On the other hand, applying the incremental paradigm is superior to MKL due to complementary classifiers that increase the accuracy.





**Figure 4.16: Combination weights in MKL using the linear kernel *svl*. The weights are determined in a leave-one-out cross-validation.**

#### 4.8.4 Discussion

We have seen that MKL performs better than VOTE and SVMs with single kernel, when all kernels are combined. This phenomenon is explained by the fact that the optimization procedure takes into account all data and gives weights to all kernels, so it can use all representations. On the other hand, when we apply the incremental algorithm, classifier combination achieves better accuracies than combining all classifiers. MKL combines the underlying feature sets to make a better combination. In this work, we used three different kernels and two combination schemes to study how the change of each parameter effects the classification accuracy. All kernels have comparable stand-alone-accuracies. The importance of each kernel function increases when their combination is considered, and combining outputs is less effective than combining the kernels themselves using optimization. We experienced an accuracy gain of 5% when using the multiple kernel learning algorithm instead of single kernels. Combining all kernels here comes with a drawback. All kernels have to be used and all features have to be extracted in order to use this model, but the increase in accuracy might be worth the costs. When the incremental algorithm is applied, no kernel is added, which is equivalent to be in a local minimum. When the classifiers are combined on the other hand, the incremental algorithm achieves more accurate results. Nevertheless, the best reached result so far is obtained when we use all representations using *svl*.

#### 4.8.5 Conclusion

We propose the use of the multiple kernel learning paradigm for the classification of nuclei in TMA images of renal clear cell carcinoma. We studied support vector machines extensively through different feature sets in our previous work. This study extends those works by using several feature sets in a multiple kernel learning paradigm and compares the results with single support vector machines and combining outputs of support vector machines using voting. MKL performs better than SVMs or VOTE in most of the experiments. MKL exploits the underlying individual contribution of each feature set, and by using multiple kernels, achieves better results in terms of classification accuracy than single kernels or voting of classifiers.

In this work, we used image based feature sets for creating multiple features. In a further application of this scenario, the use of other modalities or features (e.g. SIFT), as well as the incorporation of complementary information of different modalities is possible in order to achieve better classification accuracy.

### 4.9 Nonlinear Data Combination for Nucleus Classification

In the previous section, we studied the linear combination of multiple kernels and reported a beneficial behavior of the combination of different data sources. As a logical consequence, we want to explore now how far a nonlinear combination of kernels can further improve nucleus classification (Gönen *et al.* 2011). First, we formulate a nonlinear MKL variant and then we apply it for nuclei classification in tissue microarray images of renal cell carcinoma (RCC). The proposed variant is tested on several feature representations extracted from the automatically segmented nuclei. We compare our results with single-kernel support vector machines trained on each feature representation separately and three linear MKL algorithms from the literature. We use the dataset of 1273 renal clear cell carcinoma nucleus patches as already explained in section 0. The nonlinear MKL approach is compared with single-kernel SVMs and linear MKL algorithms. Our experiments clearly indicate that although it is more costly to use the proposed nonlinear MKL approach, the increase in accuracy is worth its computational complexity.

#### 4.9.1 MKL

MKL algorithms found in the literature frequently combine kernels linearly (e.g., linear sum, convex sum, and conic sum) (Bach *et al.* 2004;

*Acknowledgement*  
I want to thank  
Mehmet Gönen for  
the support and  
collaboration during  
the SIMBAD project.  
He prepared the man-  
uscript for the work  
on nonlinear kernel  
combinations.

Lanckriet *et al.* 2004; Rakotomamonjy *et al.* 2008). Similar to nonlinear classifier combination rules, we can also formalize nonlinear kernel combinations obtain high performing classifiers (Lewis *et al.* 2006; Cortes *et al.* 2009; Gönen and Alpaydin 2013). Our nonlinear MKL variant is based on polynomial kernel combination (Cortes *et al.* 2009). We start with a formalization of multiple kernel learning, in which a combination  $k_\eta$  of multiple kernels is learned:

$$k_\eta(x_i, x_j; \eta) = f_\eta(\{k_m(x_i^m, x_j^m)_{m=1}^P\}; \eta)$$

where the combination function  $f_\eta$  forms a single kernel from  $P$  base kernels using the parameters  $\eta$ .

#### 4.9.2 Linear MKL

Considerable research is ongoing on the theory and application of MKL and most of the published frameworks use linear combination functions (e.g. convex sum or conic sum). Fixed rules use the combination function  $f_\eta$  as a fixed function of the kernels, without any training. The combined kernel  $k_\eta$  is then treated as single kernel SVM. One example for this scenario is the non-weighted mean of the base kernels.

Instead of using a fixed combination function, a parameterized function can be employed. A learning procedure would then optimize these parameters as well. A simple case is to parameterize the sum rule as a weighted sum:

$$k_\eta(x_i, x_j; \eta) = \sum_{m=1}^P \eta_m k_m(x_i^m, x_j^m)$$

with  $\eta_m \in \mathbb{R}$ . Different versions of this approach put different restrictions on the kernel weights. For example, arbitrary weights (i.e., linear combination), nonnegative kernel weights (i.e., conic combination), or weights on a simplex (i.e., convex combination) are possible.

#### 4.9.3 Nonlinear MKL Framework

Cortes *et al.* (2009) developed a nonlinear kernel combination method based on kernel ridge regression (KRR) and polynomial combination of kernels. The nonlinear combination of  $P$  kernels can be formulated as:

$$k_\eta(x_i, x_j) = \sum_{q \in \mathcal{Q}} \eta_{q_1 q_2 \dots q_P} k_1(x_i^1, x_j^1)^{q_1} k_2(x_i^2, x_j^2)^{q_2} \dots k_P(x_i^P, x_j^P)^{q_P}$$

where  $\mathcal{Q} = \{q: q \in \mathbb{Z}_+^P, \sum_{m=1}^P q_m \leq d\}$  and  $\eta_{q_1 q_2 \dots q_P} \geq 0$ . The number of parameters to be learned is too large and the combined kernel is simplified in order to reduce the learning complexity:

$$k_\eta(x_i, x_j) = \sum_{q \in \mathcal{R}} \eta_1^{q_1} \eta_2^{q_2} \dots \eta_P^{q_P} k_1(x_i^1, x_j^1)^{q_1} k_2(x_i^2, x_j^2)^{q_2} \dots k_P(x_i^P, x_j^P)^{q_P}$$

where  $\mathcal{R} = \{q: q \in \mathbb{Z}_+^P, \sum_{m=1}^P q_m = d\}$  and  $\eta \in \mathbb{R}^P$ . For example, when  $d = 2$ , the combined kernel function becomes

$$k_\eta(x_i, x_j) = \sum_{m=1}^P \sum_{h=1}^P \eta_m \eta_h k_m(x_i^m, x_j^m) k_h(x_i^h, x_j^h)$$

The combination weights are optimized by solving the following min-max optimization problem:

$$\min_{\eta \in \mathcal{M}} \max_{\alpha \in \mathbb{R}^N} y^T \alpha - \frac{1}{2} \alpha^T (K_\eta + \lambda I) \alpha$$

where  $\mathcal{M}$  is a positive, bounded, and convex set. Two possible choices for  $\mathcal{M}$  are the  $l_1$ -norm-bounded and the  $l_2$ -norm-bounded sets  $\mathcal{M}_1$  and  $\mathcal{M}_2$  defined as:

$$\mathcal{M}_1 = \{\eta: \eta \in \mathbb{R}_+^P, \quad \|\eta - \eta_0\|_1 \leq \Lambda\}$$

$$\mathcal{M}_2 = \{\eta: \eta \in \mathbb{R}_+^P, \quad \|\eta - \eta_0\|_2 \leq \Lambda\}$$

where  $\eta_0$  and  $\Lambda$  are two model parameters. A projection-based gradient-descent algorithm can be utilized to solve this min-max optimization problem. At each iteration,  $\alpha$  is obtained by solving a KRR problem with the current kernel matrix and  $\eta$  is updated with the gradients calculated using  $\alpha$  while considering the bound constraints on  $\eta$  due to  $\mathcal{M}_1$  or  $\mathcal{M}_2$ . We formulate a variant of this optimization problem by replacing KRR with SVM as the base learner:

$$\min_{\eta \in \mathcal{M}} \max_{\alpha \in \mathcal{A}} J_\eta = 1^T \alpha - \frac{1}{2} \alpha^T ((yy^T) \odot K_\eta) \alpha$$

where  $\mathcal{A}$  is defined as:

$$\mathcal{A} = \{\alpha: \alpha \in \mathbb{R}_+^P, \quad y^T \alpha = 0, \quad \alpha \leq C\}$$

We solve this optimization problem again using a projection-based gradient-descent algorithm. When updating the kernel parameters at each iteration, the gradients of  $J_\eta$  with respect to  $\eta$  are used. These gradients can be written as

$$\frac{\partial J_\eta}{\partial \eta_m} = -\frac{1}{2} \sum_{h=1}^P \eta_h \alpha^T ((yy^T) \odot K_h \odot K_m) \alpha \quad .$$

#### 4.9.4 Experimental Design

The same dataset of 1273 RCC nucleus image patches as in the global MKL approach in the previous section was used, as well as the same experimental design (10-fold cross-validation). Analogously, three different single-kernel functions were used: *svl* (linear kernel), *sv2* (polynomial kernel of degree 2) and *svg* (Gaussian kernel).

One single kernel SVM and four MKL algorithms have been implemented. SVM denotes the single-kernel SVMs trained on each feature representation separately. RBMKL stands for the rule-based MKL algorithm that trains an SVM with the mean of the combined kernels. SimpleMKL denotes the iterative algorithm of (Rakotomamonjy *et al.* 2008) using projected gradient updates and training single-kernel SVMs at each iteration. GLMKL abbreviates the group Lasso-based MKL algorithms proposed in (Xu *et al.* 2010; Kloft *et al.* 2011). Our implementation involves the  $l_1$ -norm on the kernel weights learning a convex combination of the kernels. NLMKL denotes the nonlinear MKL variant derived from (Cortes *et al.* 2009), which uses the quadratic kernel ( $d = 2$ ) and selects the kernel weights from the set  $\mathcal{M}_1$ . As a simple starting position, we choose  $\eta_0 = 0$  and  $\Lambda = 1$ .

In total, eight representations (ALL, FG, BG, LBP, COL, FCC, SIG, and PHOG), three kernels (*svl*, *sv2*, and *svg*), and five algorithms (SVM, RBMKL, SimpleMKL, GLMKL, and NLMKL) are incorporated.

#### 4.9.5 Results

The single-kernel SMV accuracies for all feature representations and kernel functions are listed in Table 4.6. The best performance is obtained by *svg* with feature PHOG with 76.9 % classification accuracy. In general, feature representations BG and PHOG gave consistently higher accuracies than other representations.

**Table 4.6: Classification accuracies of single-kernel classifiers. Values are mean and standard deviation of 10-fold cross-validation.**

	<i>SVL</i>	<i>SV2</i>	<i>SVG</i>
<b>ALL</b>	70.0 ± 0.2	71.9 ± 2.9	68.7 ± 2.9
<b>FG</b>	70.0 ± 0.2	71.2 ± 3.7	65.9 ± 4.3
<b>BG</b>	70.2 ± 0.6	72.7 ± 3.8	69.6 ± 3.1
<b>LBP</b>	70.0 ± 0.2	63.6 ± 2.7	68.4 ± 6.3
<b>COL</b>	70.2 ± 3.0	62.9 ± 3.5	67.2 ± 3.4
<b>FCC</b>	70.0 ± 0.2	69.8 ± 0.7	62.9 ± 5.5
<b>SIG</b>	70.0 ± 0.2	69.6 ± 3.4	66.0 ± 3.0
<b>PHOG</b>	76.0 ± 3.4	70.5 ± 3.3	<b>76.9 ± 2.7</b>

Next, using four different MKL algorithms, we combined eight kernels calculated on the feature representations with the same kernel function. Table 4.7 lists the results of best single-kernel SVMs and four MKL algorithms trained. The highest accuracy of 83.3% is achieved by combining eight *svg* kernels via NLMKL. The accuracy is considerably higher than for all other MKL settings or single-kernel SVMs. Table 4.7 lists in the last column the classification accuracies when combining all pairs of feature representations and kernel functions (i.e., 24 kernels) *in a single classifier*. Still, NLMKL shows the highest accuracy of 83.1%, strengthening its position as best MKL variant.

**Table 4.7: Classification accuracies of MKL classifiers. Values are mean and standard deviation of 10-fold cross-validation.**

	<i>SVL</i>	<i>SV2</i>	<i>SVG</i>	<i>SVL+SV2+SVG</i>
<b>SVM</b>	76.0 ± 3.4	72.7 ± 3.8	76.9 ± 2.7	-
<b>RBMKL</b>	77.3 ± 4.0	77.2 ± 2.4	82.7 ± 3.6	81.8 ± 3.8
<b>SIMPLEMKL</b>	77.1 ± 3.3	77.3 ± 2.3	81.8 ± 3.8	81.6 ± 3.9
<b>GLMKL</b>	77.1 ± 3.5	76.5 ± 3.2	81.8 ± 4.3	81.8 ± 3.8
<b>NLMKL</b>	77.9 ± 3.9	79.2 ± 3.8	<b>83.3 ± 3.6</b>	83.1 ± 3.5

#### 4.9.6 Discussion

This study clearly supports the idea of multiple kernel learning. We have formulated a nonlinear MKL algorithm derived from polynomial kernel combination (Cortes *et al.* 2009) with which we could achieve a higher classification performance than single-kernel SVMs and three linear MKL algorithms. When combining linear kernels on the feature representations, we observed linear MKL algorithms to outperform single-kernel SVMs, whereas the nonlinear MKL algorithm improved the average accuracy at most – supposedly due to the beneficial nonlinear kernel combination. Even for the combination of nonlinear base kernel (*sv2* and *svg*), the nonlinear MKL algorithm led to higher accuracies than single-kernel SVMs and linear MKL algorithms. The gain in classification accuracy when using nonlinear MKL can be quantified as 6.4% compared to single-kernel SVMs.

#### 4.9.7 Conclusion

This study extends our previous work on MKL (see section 0) by the use of a nonlinear MKL setting and clearly indicates that the nonlinear combination of kernels can further improve cell nucleus classification. The proposed nonlinear MKL variant learns a better similarity metric than linear MKL algorithms by combining the input kernels nonlinearly.

## 4.10 Active Learning for Nucleus Classification

Supervised cell nucleus classification based on IHC images requires a set of manually labeled cell nuclei serving as gold standard training labels. In our example, two trained pathologists have detected and classified all cell nuclei into malignant or benign on eight tissue microarray image quarters, in total a number of 1633 cell nuclei identified by both pathologists. This process of label acquisition commonly poses an expensive part in medical research. We therefore investigated the possibility of reducing the number of needed training labels (Schüffler *et al.* 2013b), by active learning (AL). AL is a scenario where the classifier itself decides which labels from an external labeler are most informative for an improved classification result (Cohn *et al.* 1996). Given an initial set of labeled test data, the probabilistic classifier calculates an uncertainty score for each unlabeled sample and queries the label for the most uncertain sample from the labeler. The uncertainty score  $U$  for a sample  $s$  is calculated by the entropy of classification probability:

$$U(s) = - \sum_{\hat{y} \in \{MAL, BEN\}} p(\hat{y}|s) * \log(p(\hat{y}|s))$$

where  $p(\hat{y}|s)$  denotes the probability of sample  $s$  to be malignant or benign, respectively. The samples with largest uncertainty is selected for supervision.

### 4.10.1 Random Forests as Probabilistic Classifier Ensemble

Random forests (RF) (Breiman 2001) are being used increasingly by many medical applications like cancer classification or tissue segmentation (Geremia *et al.* 2010; Fuchs *et al.* 2011b). They are computationally efficient for large training data, can solve multiclass classification problems, and can be interpreted in a probabilistic manner. An RF is an ensemble of binary decision trees, where each tree is typically trained with a different subset of the training set (“bagging”), thereby improving the generalization ability of the classifier (Mahapatra *et al.* 2013c). Samples are processed along a path from the root to a leaf in each tree by performing a binary test at each internal node along this path. A test compares a certain feature with a threshold. Training a forest amounts to identifying the set of tests that best separate the data into the different training classes. At each node, the feature space is searched for a test that maximizes the reduction of class impurity, typically measured by class entropy (Mahapatra *et al.* 2013c).

Rather than inspecting the full space of features at each node of a tree, a random subset is sampled, and the best one is selected. Even if this choice

renders the individual trees weaker, it decreases the correlation between their outputs, increasing the performance of the forest as a whole. Each training sample is sent to the corresponding child depending on the result of the test. Comparison of a feature subset with a threshold continues iteratively until convergence. The convergence criteria for stopping the recursive comparison of feature values to a threshold are:

1. The number of samples in a node falls below a threshold;
2. A predefined maximum tree depth is reached;
3. All samples belong to the same class. In that case, the node becomes a leaf, and the most frequent class of the training data at the node is stored.

During testing, a new sample is processed by applying respective tests according to the path from the root node to the leaf it traverses. When a leaf node is reached, the tree casts a vote corresponding to the class assigned to this node in the training stage. The final decision for a test sample is given by the class with the majority of votes. Moreover, the probability that a test sample belongs to a class can be estimated as the fraction of votes for that class cast by all trees (Mahapatra *et al.* 2013c).

#### 4.10.2 Experiments and Results

As we solely want to explore the potential of active learning in this field, we implemented a very basic active learning scenario (Algorithm 4.1). 20 randomly chosen malignant and benign nuclei (ten from each class) serve as initial supervised training set for a random forest classifier ensemble with 50 trees. The classifier tries to predict all other nuclei on the test images. The classification accuracy is shown in Figure 4.17 on the left side, as a baseline for a weak classification performance (77.5% accuracy). The training set then grows by additional 20 nuclei, and the labels of the new nuclei are provided by the pathologist. In one experiment, the additional twenty nuclei are chosen randomly from all available nuclei. In a second experiment, the twenty new nuclei are systematically chosen by their classification confidence: those nuclei with the highest classification uncertainty are added to the training set. The new classification accuracy is determined and the procedure is successively repeated until all nuclei from the image are queried. As shown in Figure 4.17, the classification accuracy performance increases when more nuclei are provided for training. But interestingly, the performance increases much faster when the new nuclei are queried according to their classification confidence. The classifier mostly profits from additional information of cases with high uncertainty. The plateau of maximum information is at around 96% classification accuracy and is reached twice as fast with AL as in the random sampling scenario (160 samples vs. 320 samples) (Schüffler *et al.* 2013b).



Based on these results, we emphasize the potential of active learning algorithms especially in medical imaging, when the datasets can be very large and labeling very expensive.

**Algorithm 4.1: AL process of nucleus classification.** A pathologist labels the image step by step, thus correcting and improving the resulting classifier  $C$ . The underscored line is used for the AL approach and is omitted in the random approach.

**Data:** Set of unlabeled cell nuclei  $U = \{u_1, \dots, u_n\}$ ,

**Input:** Trained pathologist  $P$ . **Output:** Probabilistic classifier  $C$ .

---

```

1  let  $P$  label ten malignant and ten benign nuclei with labels  $l_1, \dots, l_{20}$ ;
2  train classifier  $C$ ;
3  while ( $\exists u_i \in U | u_i$  is not labeled by pathologist)
4      predict labels for  $u_i$ ;
5      sort  $u_i$  according to classification uncertainty  $U(u_i)$ ;
6      let  $P$  label twenty malignant and benign unlabeled nuclei;
7      retrain classifier  $C$ ;
8  end
9  return  $C$ ;

```

---

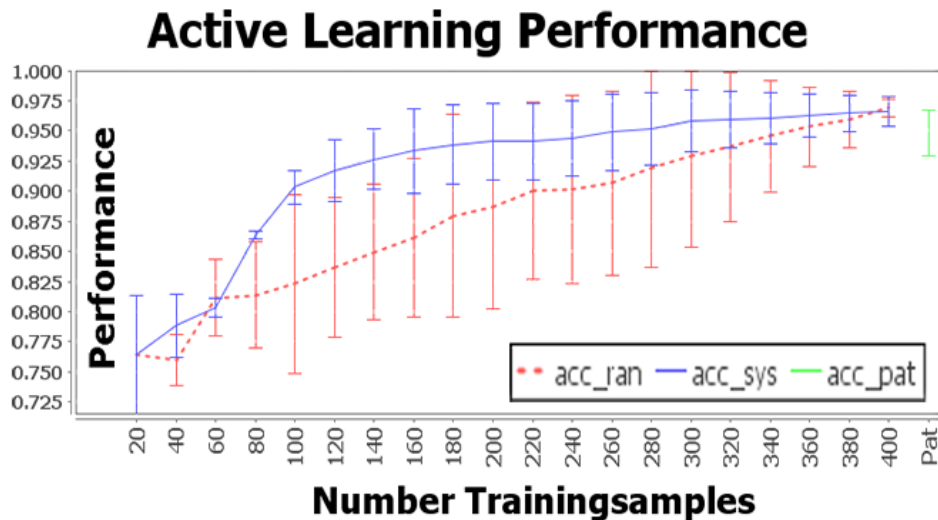


Figure 4.17: Proof of concept for the active learning approach for nucleus classification. For three given TMA images, initially 10 malignant and 10 benign nuclei were selected to train a random forest classifier with 50 trees. The classification result on all nuclei is shown as accuracy on the y-axis. Consecutively, 20 additional nuclei were added repeatedly to the training (x-axis) thus improving the classification performance. The additional nuclei were chosen at random ( $acc\_ran$ ) or systematically according to the highest classification uncertainty ( $acc\_sys$ ). The systematic approach saturates much faster. The classification accuracy reaches the level of the two pathologists shown as green bar on the right ( $acc\_pat$ ).

## 4.11 Survival Analysis

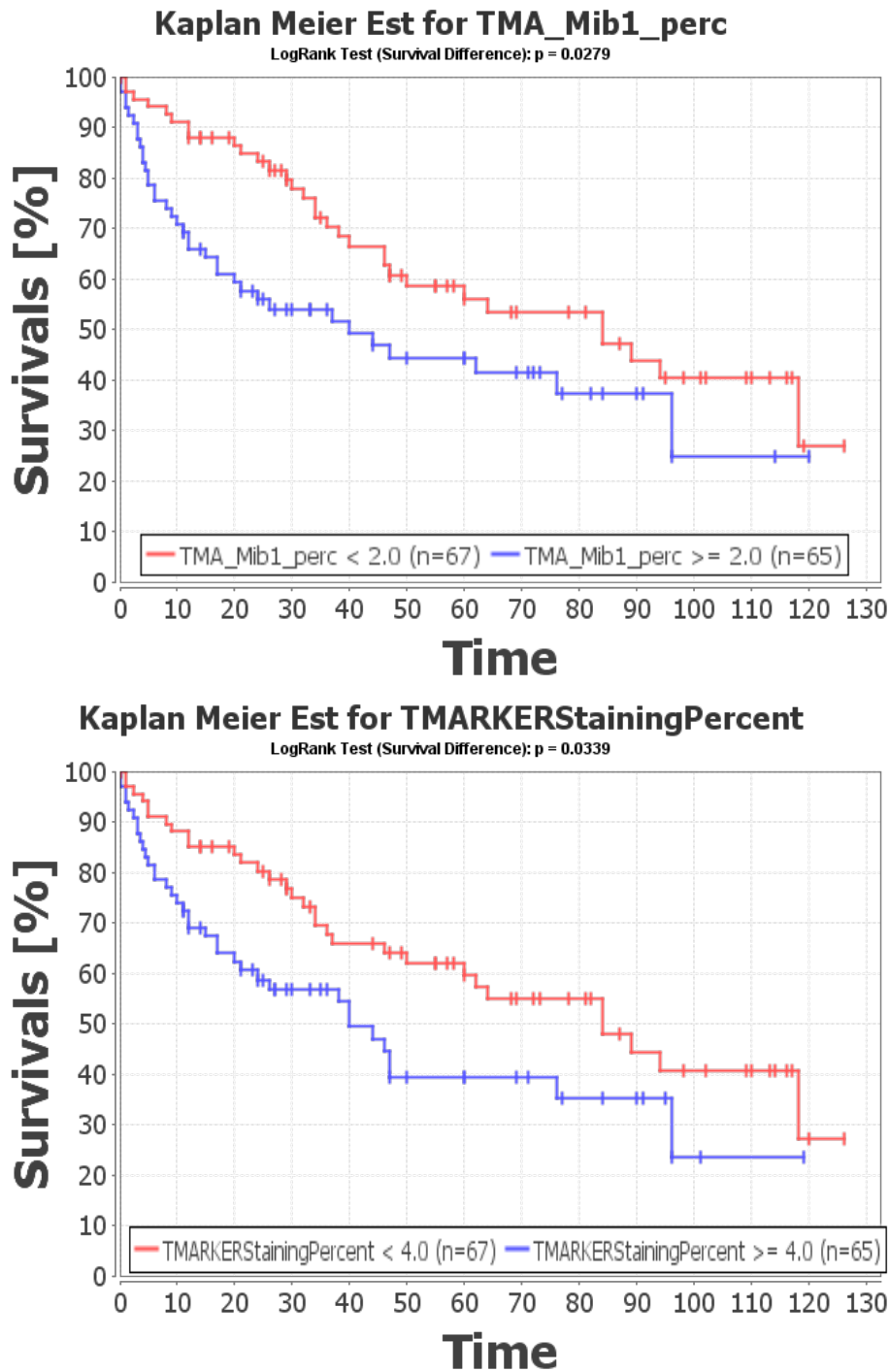
For a holistic analysis of the staining estimation pipeline, a set of 132 TMA images of ccRCC patients was analyzed in a fully automatic manner. Eight fully labeled TMA images served as training set showing 1633 cell nuclei. 891 of these are benign and 382 are malignant. 360 are undetermined or unsure. A Voronoi sampling revealed 784 background loci to train the two-step classification pipeline. For training, images have been segmented using the SLIC superpixel algorithm (Achanta *et al.* 2012). Standard image features as described in 4.4 (Intensity, LBP, PHOG, SIG, FCC). Random Forests (Breiman 2001) with 50 trees were used to train a nucleus detector, classifying foreground and background superpixels. Using the same features, a second classifier was learned exclusively to discriminate superpixels on benign and malignant nuclei.

During testing, the 132 TMA images were processed in a similar way: After over-segmentation with SLIC and subsequent feature extraction, superpixels were classified into foreground or background using the first classifier. Superpixels classified as foreground were processed by the second classifier discriminating malignant from benign superpixels. Since MIB-1 (Ki-67 antigen) positive nuclei in ccRCC TMA are dark brown stained in contrast to bluish negative nuclei, the fraction  $f$  of the mean red intensity ( $r$ ) and the mean blue intensity ( $b$ ),  $f = r/b$ , defined the staining state of a superpixel. If  $f > 1$ , the nucleus was considered stained, and non-stained otherwise.

Ki-67 protein is involved in cell proliferation. In ccRCC, a high percentage of proliferating cancer cells is associated with poor patient prognosis. Therefore, patients can be classified as “MIB-1 positive” and “MIB-1 negative” and a Kaplan-Meier estimate indicates the difference in survival between these two groups. As a reference, the MIB-1 staining was estimated by a trained pathologists on the same TMA image data. To compare the automatic method with the manual reference, the survival groups were kept at equal size: around 66 patient in each group. A log-Rank test quantifies the statistical difference of the two survival curves.

### 4.11.1 Results

The reference groups are divided at the MIB-1 percentage level of 2% (67 patients in group “negative”). The difference of survival curves is significant ( $p=0.028$ ), supporting the hypothesis that Ki-67 is a prognostic marker for ccRCC (Figure 4.18, top). A similar assumption can be done with the fully automated staining estimation pipeline (Figure 4.18, bottom): with a threshold at 4% (67 patients in group “negative”), the two survival groups are discriminated at similar significance level ( $p=0.034$ ).



**Figure 4.18: Survival time (months after diagnosis) for ccRCC patients with high or low percentage of MIB-1 positive cancer cells. TOP: A pathologist estimated the staining percentage of 132 TMA images (TMA\_Mib1\_perc). Patients of the “MIB-1 negative” group ( $\geq 2\%$ , red, upper curve) have a better prognosis than those of the positive group. The difference of survival is significant (log-Rank test  $p=0.027$ ). BOTTOM: Our staining estimation pipeline (TMARKERStainingPercent) stratified the patients in the same dataset similarly. While the threshold is at 4%, the survival difference still is significant ( $p=0.034$ ).**

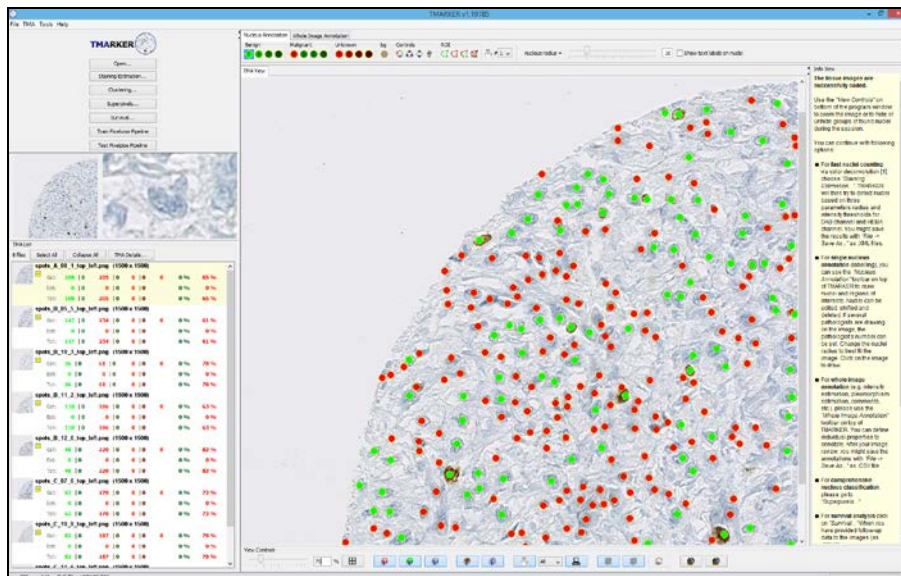


# 5 TMARKER: A FREE SOFTWARE TOOLKIT FOR STAINING ESTIMATION

We provide an open-source and freely available software package which implements the aforementioned staining estimation pipeline (Schüffler *et al.* 2013b; Schüffler *et al.* 2013d). TMARKER is a user-friendly Java GUI, platform independent and can be used by any personal computer with internet connection and Java runtime environment 1.7 or higher. The software aims to fit following needs based on the pipeline presented in chapter 4:

- Semiautomatic, reproducible, accurate and fast cell nuclei detection and counting for a given set of IHC stained images.
- Automatic cell nucleus classification into malignant and benign.
- Platform independence, open source, user-friendly Java webstart user interface for an easy distribution.

The program TMARKER is implemented in Java v1.7 and is publicly available at <http://www.comp-path.inf.ethz.ch>.



**Figure 5.1:** Screenshot of TMARKER with a TMA image of MIB-1 (ki-67) stained ccRCC. Detected cancerous and benign nuclei are marked with red and green points, respectively.

TMARKER implements the complete staining estimation pipeline from chapter 4. The program supports nucleus detection via color deconvolution (section 4.2.1), nucleus segmentation via superpixels (section 4.2.2)

or graph-cuts (section 4.3). All image and shape features explained in section 4.4 are implemented. Nucleus classification can then be performed with support vector machines (section 4.5) or random forests (section 4.10.1). Further, labeling histopathological images is supported with the active learning approach from section 4.10. Implementation details of the individual modules are explained in the next section.

## 5.1 Implementation Details

TMARKER has been built with NetBeans IDE 7.4 under Java SDK 7. If not stated separately, Java modules for individual functions have been adopted from the Fiji software package (Schindelin *et al.* 2012), JFeatureLib (Graf 2012) or LIRE (Lucene Image Retrieval, (Lux and Chatzichristofis 2008)). Multiple histopathological tissue images can be loaded for computational analysis. Nucleus detection on various staining channels (e.g. DAB staining, HE staining, methyl green staining and others) is performed with color deconvolution by Ruifrok *et al.* (2001).

For patch-wise nucleus segmentation via graph-cuts, a Java implementation of the graph-cut MAXFLOW algorithm by Boykov and Kolmogorov (2004) has been employed. For superpixels, the *Simple Linear Iterative Clustering (SLIC)* algorithm (Achanta *et al.* 2012) has been translated to native Java code and adjusted for comb-shaped superpixels.

We implemented a set of image features which have been described in section 4.4. Local binary patterns (Ahonen *et al.* 2004) and pyramid histograms of oriented gradients (Bosch *et al.* 2007) have been adopted from JFeatureLib (Graf 2012). Random forest classifiers from WEKA package (Hall *et al.* 2009) and support vector machines from libSVM (Chang and Lin 2001) were incorporated for nucleus classification. TMARKER visualizes the probabilistic classification results for nucleus detection and classification as an overlay over the histological image such that the user can immediately supervise and retrain the classifier.

Background samples with Voronoi tessellation is calculated with Delaunay triangulation (Chew 2014). For visualization and plotting, jFreeChart (Gilbert 2000) has been incorporated. The library iTextPDF (Lowagie 2010) enables TMARKER analyses to be exported as PDF file.

Java is an object oriented programming language and supports modular implementation. TMA spots and nuclei are represented as individual objects. Figure 5.2 illustrates the architecture of TMARKER with a simplified unified modeling language (UML) class diagram. Classes for visualization, feature extraction, parallel threading, and additional computing are omitted. In total, TMARKER comprises 67 classes.

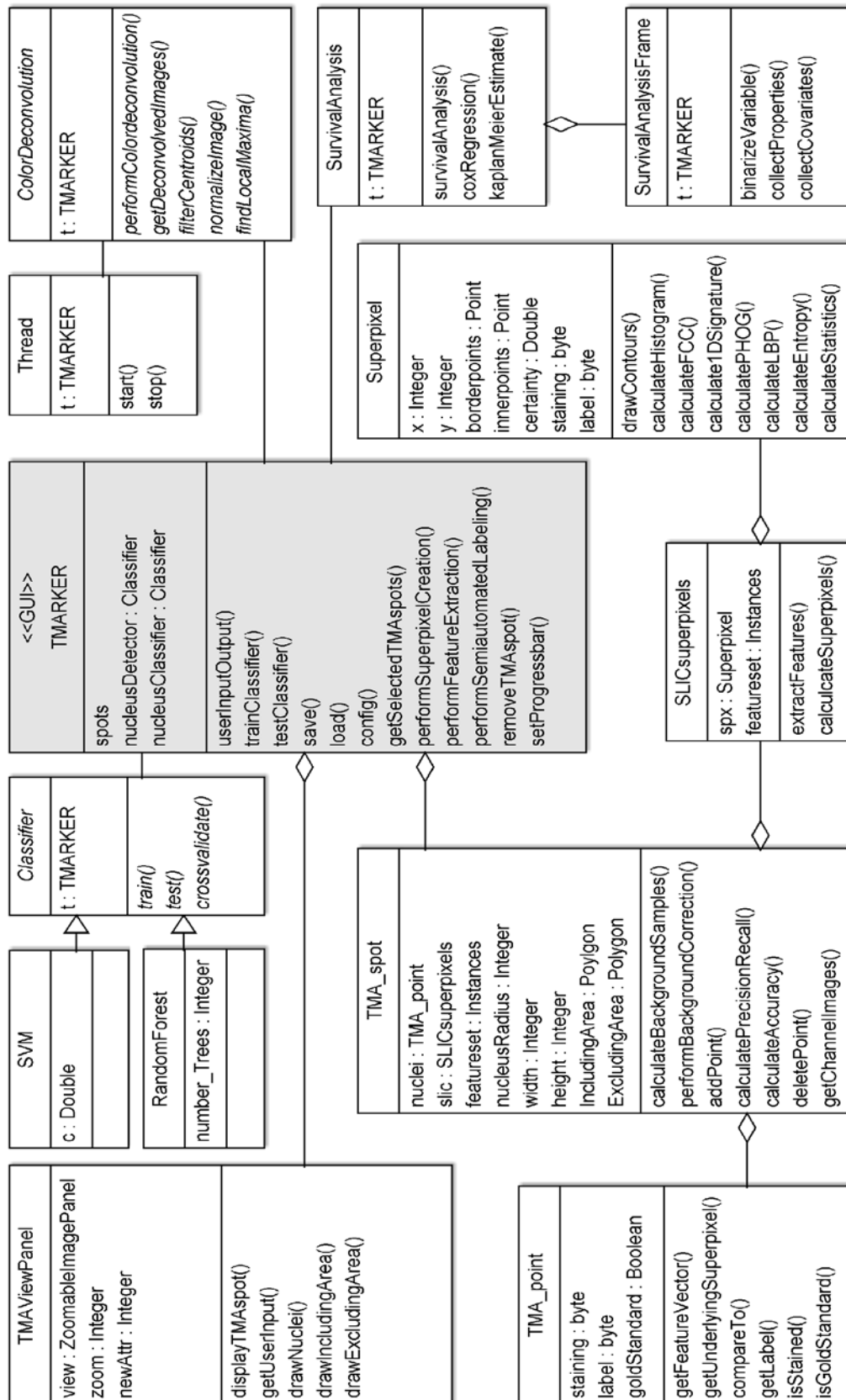
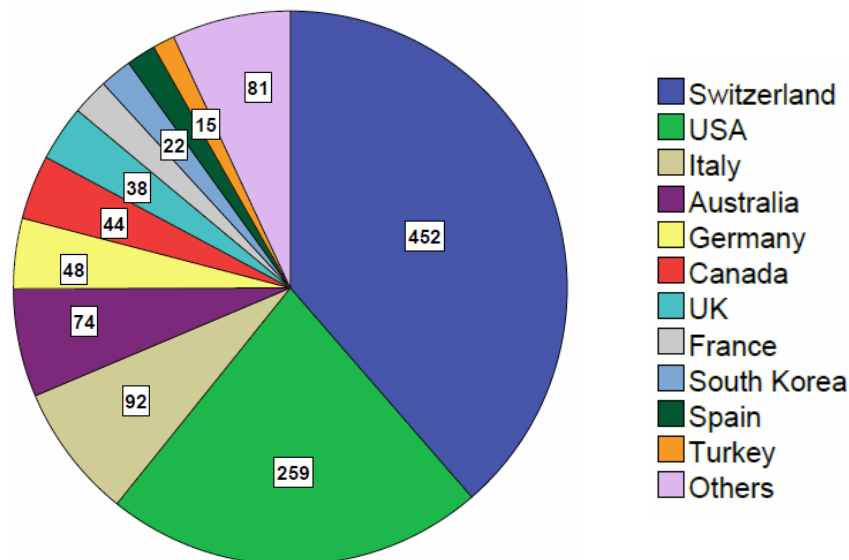


Figure 5.2: Simplified UML class diagram of TMARKER illustrating dependencies of most important Java classes, attributes and functions. The main class is TMARKER (GUI, highlighted).

## 5.2 Usage Statistics

TMARKER is programmed for stand-alone computer systems in hospitals and other institutes. Figure 5.3 documents the usage of TMARKER webstart from launch to today. The program was launched over 1100 times in over 29 different countries worldwide. The cities with the most usages are Zurich, Napoli, Melbourne, London, Los Angeles, Mountain View, Nashville and New York City. TMARKER is used daily.



**Figure 5.3: Usage numbers of TMARKER webstart by country from launch until today (Oct 11 2011 – Jan 22 2014). Data are collected by IP address, excluding those of ETHZ. Numbers of downloads of the desktop version are not included.**

## 5.3 Conclusion

TMARKER is a free software toolkit with large potential in cell counting and staining estimation of histopathological IHC stained tissue images. A superior benefit of TMARKER is the computational reproducibility of competitive cell counts. A fast cell counting and staining estimation method is provided by the integrated color deconvolution method. When only relevant cells are to be considered for staining estimation, e.g. with distinction between malignant and benign cells, TMARKER offers modern machine learning algorithms for nucleus detection and classification. The potential of TMARKER has been shown on renal clear cell carcinoma images, prostate cancer images and mamma carcinoma datasets and the program is already in frequent use for diverse applications.



## 6 SINGLE CELL SEGMENTATION ON HIGHLY MULTIPLEXED IMAGES

In the previous sections, we investigated the nucleus segmentation and classification on tissue microarrays (TMAs). These TMAs provide a qualitative and partially quantitative expression pattern of typically one target protein (singleplex). Modern cancer research commonly incorporates *single cell analysis* on affected tissues with multiple target proteins. Single cell analysis comprises whole cell segmentation with subsequent *protein expression analysis* in individual cells. Therefore, multiple singleplex protein images have to be aligned to each other for a holistic view of the cells segmented either on the individual images or on the overlay image. A new immunohistochemical mass cytometry approach, that has recently been introduced (Giesen *et al.* 2014), simultaneously and quantitatively measures multiple protein expression and modification profiles in a single tissue sample making the alignment unnecessary. We present a new cell segmentation approach based on watersheds which exploits the highly registered multidimensional cell morphology information of the multiplexed images. The joint information of cell membrane proteins B-catenin, Her2 and Cytokeratin as well as the nucleus protein H3 improve the segmentation. We define a score for segmentation validation without manually segmented gold-standard.

### 6.1 Introduction

Computational pathology typically addresses the automated analysis of digitalized immunohistochemically (IHC) stained cancer tissue images (Fuchs and Buhmann 2011a). The visualized expression pattern of specific marker proteins such as e.g. the proliferation factor ki-67 can be indicative for diagnosis, prognosis or therapy decision (Fuchs *et al.* 2008b). Machine learning algorithms have been invented which automate cell nucleus detection, segmentation and classification (Schuffler *et al.* 2010).

Classical pathology usually examines the expression profiles of individual proteins separately in multiple IHC experiments on different sample slices (singleplex), as the simultaneous staining of multiple proteins inherits technical problems, such as lack of a variety of different dyes and non-separable color mixture on the prepared image. Immuno-fluorescence (IF), however, can achieve multiplexed staining protocols for the use of up to seven proteins in one sample by using three filter sets at different wavelengths (Tsurui *et al.* 2000), or more proteins by repeated

singleplex experiments with self-inactivating dyes (Gerdes *et al.* 2013). Still, such methods rely on post-processed registering of image stacks, are time consuming and prone to changes during experiment (Tsurui *et al.* 2000; Gerdes *et al.* 2013).

Single-cell analysis in cancer research tries to gain information from spatial protein expression patterns in individual cells and their interacting neighbors. IHC is a medical tool commonly used for the spatial and quantitative analysis of individual target proteins. To get a multidimensional view of the sample with multiple proteins, several IHC experiments have to be aligned to each other such that the spatial information of the samples corresponds to each. The alignment of the various IHC stained tissue samples is not unproblematic due to changing experimental settings, anisotrop tissue slices and varying image quality. Single cells in one image can disappear on the subsequent image. Illumination changes, sharpness variances, scaling variances (linear or nonlinear), dust and tissue scratches can complicate or even impede a sufficient image registration. A technical system with multidimensional highly registered spatial protein expression patterns would therefore be favorable.

Giesen *et al.* (2014) developed a mass cytometry technique for the multiplexed spatial and quantitative measurement of dozens of protein expression patterns at sub cellular resolution in a single tissue specimen. The sample is scanned pixel-wise quantifying all targeted proteins separately in each pixel. Thus, the resulting scan comprises multiple highly registered protein channel images. A registration process of these images is not required anymore facilitating single cell analysis drastically.

For subsequent single cell analysis, a cell segmentation algorithm is needed which is able to detect and segment individual cells on this new type of images. The advantage of such a segmentation algorithm is the separate use of cell border information of multiple dimensions (proteins). In conventional microscopic IHC images, cell borders are usually not highlighted and their staining is mixed with the IHC information of the target protein, which complicates whole cell segmentation.

We present a new weighted watershed based cell segmentation which uses separate information of the epithelial cell-cell junction proteins Beta-catenin (B-catenin), Her2 and Cytokeratin 8/18 (Keratin) together with the nucleosomal DNA package protein histone H3 for single cell segmentation in human breast cancer images.

The **validation of a given segmentation** is a crucial to compare different segmentation algorithms. In many medical segmentation problems, trained medical doctors therefore provide manual segmentations of the objects to which the algorithms are compared. In our case, no such gold

standard exists. We therefore define a segmentation score considering several aspects of a valuable segmentation, which (i) should present appropriately sized cell segments, (ii) should encapsulate maximal one nucleus per tissue cell, (iii) should largely overlap with membrane marker proteins and (iv) should not overlap with nucleus marker proteins.

## 6.2 Methods

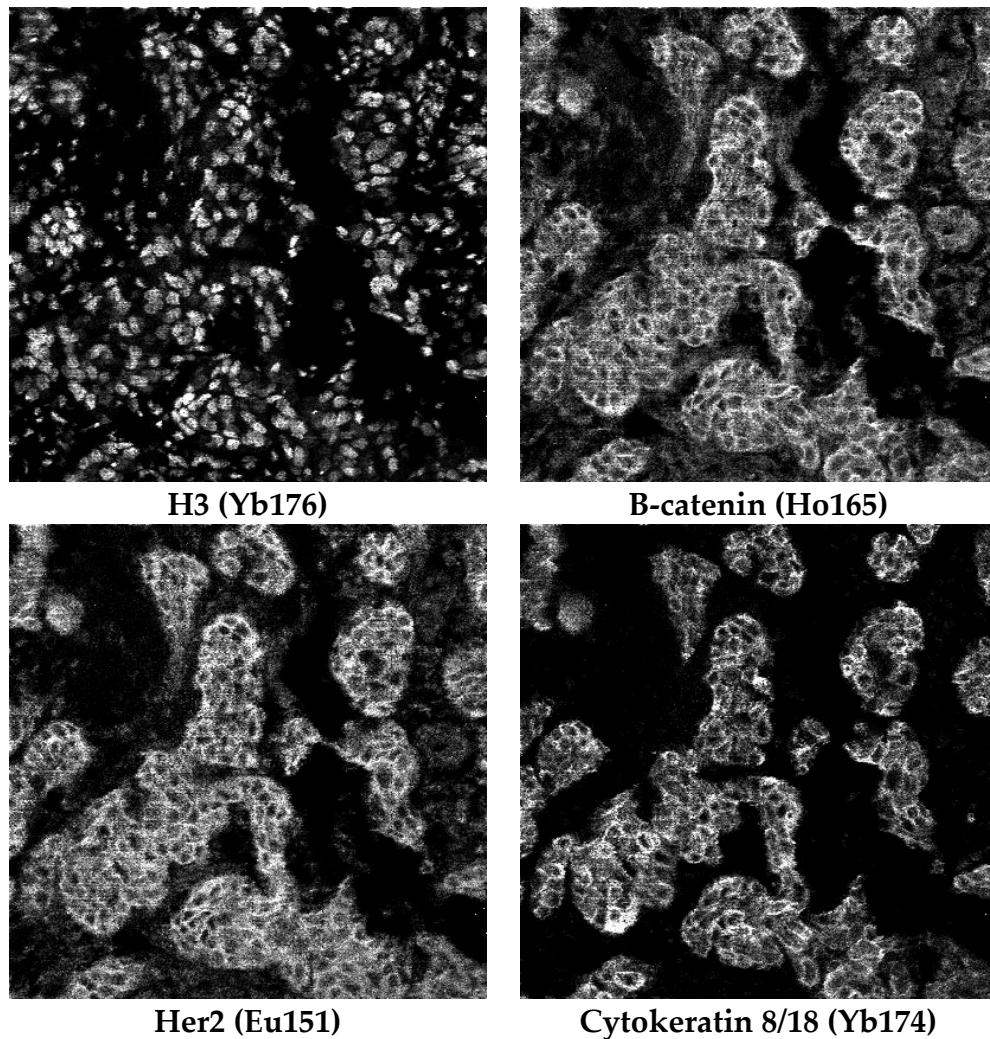
### 6.2.1 Image Acquisition

The breast cancer dataset has been described by Theurillat *et al.* (2007). The multiplexed IHC staining protocol described by Giesen *et al.* (2014) in principle follows a classical IHC staining protocol. 32 breast cancer relevant antibodies have been selected for staining. Before staining, the antibodies are labeled with a unique rare earth metal isotope with defined atomic mass. The stained sample is processed in a laser ablation chamber for high resolution, high-throughput and high sensitive analysis (Wang *et al.* 2013). A 193 nm argon fluoride laser beam (Gunther *et al.* 1997) at  $3.5 \text{ J cm}^{-2}$  fluence ablates the sample at a frequency of 20 Hz in a regular grid of  $1 \text{ }\mu\text{m}$  resolution. The ablated substance is immediately subject to a time-of-flight inductively coupled plasma mass spectrometer (ICP-MS, CyTOF™) where the metal isotopes are identified and counted. The raw data counts are normalized to gray-scaled intensity images between 0 (black) and 1 (white). For contrast enhancement and to remove signal outliers, a histogram adjustment mapped the intensity values to new values such that 1% of data is saturated at low and high intensities. 32 rare earth metal isotopes have been used to label antibodies to 32 different breast cancer target proteins. Each protein is visible in a separate channel image corresponding to the tagged metal isotope. We use the channels of four isotopes.

### 6.2.2 Single Cell Segmentation

For downstream single cell analysis, a prior single cell segmentation is necessary. The cell membrane proteins B-catenin, Her2 and Keratin shown in Figure 6.1 visualize the morphological structure of cell membranes best. Further, the nuclear protein H3 indicates the loci of all cell nuclei as complementary information.

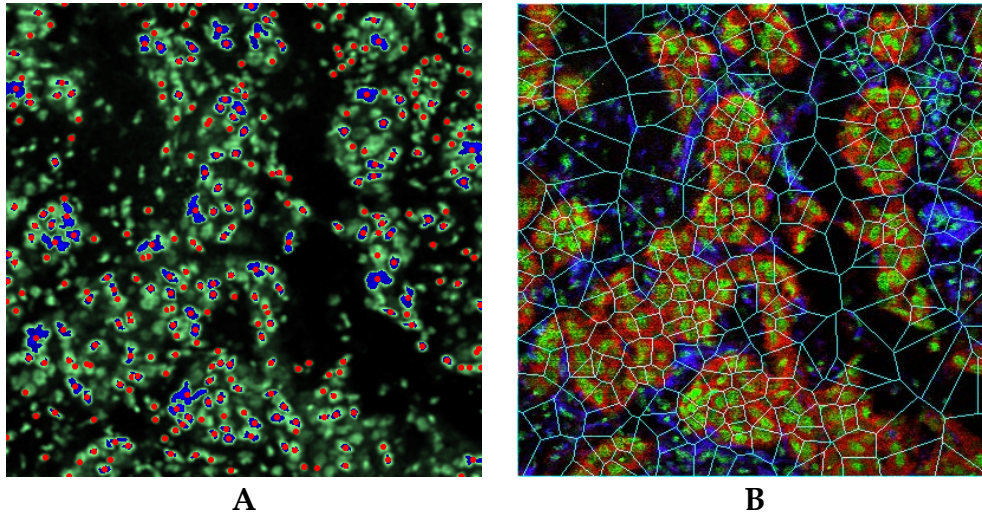
For improved cell segmentation, we want to exploit the information of all four channels, treating the inverse of H3, namely  $1\text{-H3}$ , as a pseudo-membrane protein. All methods have been implemented using MATLAB (R2013b), standard imaging toolbox.



**Figure 6.1:** The channel images of H3 (nucleus protein), B-catenin, Her2 and Keratin (epithelial cell-cell junction proteins mainly expressed in membrane). In brackets are the applied metal isotope for antibody labeling. For contrast enhancement, the images are normalized and adjusted for their histogram: the intensity values are mapped to new values such that 1% of data is saturated at low and high intensities.

### **Voronoi Segmentation with Dirichlet Clustering**

As a reference cell segmentation, a Voronoi diagram (Aurenhammer 1991; Okabe *et al.* 2009) around the cell nuclei is created (see Figure 6.2). Therefore, the H3 nucleus image (Figure 6.1 top left) is smoothed with a Gaussian blur filter of radius 2 to reduce single pixel noise. To find the centroids of the cell nuclei, the image is dichotomized at an intensity threshold 0.5 and the centroids are determined as the centers of mass of single connected components (Figure 6.2 A). For each centroid, the Voronoi polygon is defined as a boundary that encloses all intermediate pixels lying close to the centroid than to other centroids. The Voronoi diagram is then defined as the set of all Voronoi polygons in the image. The



**Figure 6.2: Process of Voronoi tessellation for single cell segmentation without cell membrane information. A: The original image H3 shown in Figure 6.1 top left is smoothed with a Gaussian blur filter and dichotomized. Centroids (red) of the resulting clusters (blue) determine the centers of nuclei. B: Voronoi tessellation of found centroids. The colored image is an overlay of Her2 (red, membranes), H3 (green, nuclei) and Vimentin (blue, cytoplasm).**

segmentation is illustrated as overlay image (Figure 6.2 B) visualizing cell membranes (Her2, red), cell nuclei (H3, green) and cytoplasm (Vimentin, blue). Voronoi segmentation does not include explicit information of cell boundaries and only considers the cell nuclei as centers of cells.

### Watershed Segmentation

As cell membrane proteins naturally indicate boundaries of the cells, we propose to use their information for single cell segmentation via watersheds (Beucher and Lantuejoul 1979; Meyer 1994). Watersheds interpret a gray-scaled image as a topographic relief with basins of low intensity and walls of high intensity. The relief is then uniformly flooded starting at the lowest point. When two neighboring catchment basins would merge due to the flood, a damn is erected to prevent the mixture. The collection of all damns over the image then describes the watershed segmentation. Since the tissue cells in membrane channel images can be interpreted as basins, the use of watersheds is intuitive.

To capture the information of multiple membrane proteins (B-catenin, Her2 and Keratin), the images first are weighted and averaged to a single membrane image. Further, since a cell nucleus is considered to be exclusive to cell membrane, the inverse of the nucleus image H3, namely  $1-H3$ , is incorporated as pseudo-membrane channel. Two standard averages have been considered, the arithmetic mean and the geometric mean.

Given  $N$  gray-scale images  $I_i$ , each weighted with  $w_i \in \mathbb{R}$ , with  $\sum w_i = 1$ , the arithmetic mean image  $I_{Arith}$  is defined as:

$$I_{Arith} = \sum_{i=1}^N w_i * I_i$$

Similarly, the geometric mean image  $I_{Geom}$  is defined as:

$$I_{Geom} = \prod_{i=1}^N I_i^{w_i} = e^{(\sum_i w_i \log(I_i))}$$

Two additional averages are defined as *disjunction* or *conjunction*:

$$I_{OR} = \max(I_i) \quad I_{AND} = \min(I_i)$$

Note that all operations are performed element-wise, i.e.  $I_{Geom}$  uses the pixel-wise product of exponentiated intensities and  $I_{OR}$  and  $I_{AND}$  use the pixel-wise max and min intensity of all images, respectively.

### 6.2.3 Segmentation Score

Let  $I$  be the original multiplexed tissue image of size  $m \times n \times 32$  and  $M \in \{0,1\}^{m \times n}$  a cell segmentation mask (0, no border; 1, border). Four terms define our segmentation score:

- 1) The score should approximate the expected number of cells  $N_{exp} = \frac{m*n}{4r^2}$ , where  $r$  is the typical nucleus radius. We address this by sampling the observed number of cells  $N$  from a normalized Gaussian distribution with  $\mu = N_{exp}$  and  $\sigma = 2 * N_{exp}$  as a probability  $p_{size} \in [0; 1]$ . To avoid non-sense segmentation with no cells or too small cells,  $N$  must not be larger than  $4 * N_{exp}$  or smaller than  $N_{exp}/4$ .
- 2) Multiple nuclei in one cell should be penalized. The total number of nuclei  $N_{Nucl}$  is determined as explained for the Voronoi method. The number of nuclei lying in cells with one or more nuclei  $N_{nucl+}$  is determined. The sub-score of the nucleus constraint is given by  $p_{n\_nuclei} = 1 - N_{Nucl+}/N_{Nucl}$ .
- 3) To measure the wanted mask overlap with the membranes, we define a membrane image  $I_{MEM} = \sqrt[3]{I_{Bcatenin} * I_{Her2} * I_{Keratin}}$ . The mean overlap is then quantified as  $p_{membrane} = \frac{1}{\sum M} \sum M * I_{MEM}$ .
- 4) The inverse mean overlap of the mask  $M$  with the nuclei is quantified as  $p_{nuclei} = 1 - \frac{1}{\sum M} \sum M * I_{H3}$ .

The final segmentation score for an image  $I$  and a mask  $M$  is then defined as the geometric mean of the sub-scores:

$$\text{score}(I, M) = \sqrt[4]{p_{\text{size}} * p_{n\_nuclei} * p_{\text{membrane}} * p_{\text{nuclei}}}$$

This score ranges from zero to one and enables to compare different segmentation as a higher score implies a better segmentation.

## 6.3 Results

### 6.3.1 Mutual Information between Membrane Proteins

Since the membrane protein channels B-catenin, Her2 and Keratin are perfectly registered to each other, their mutual information pixel-wise can be calculated in a pixel-wise manner. The images are scaled down to half edge length in order to smooth scanning artifacts. The images are dichotomized according to a threshold  $t$  and the normalized mutual information (MI) between two binary images  $I, J$  is:

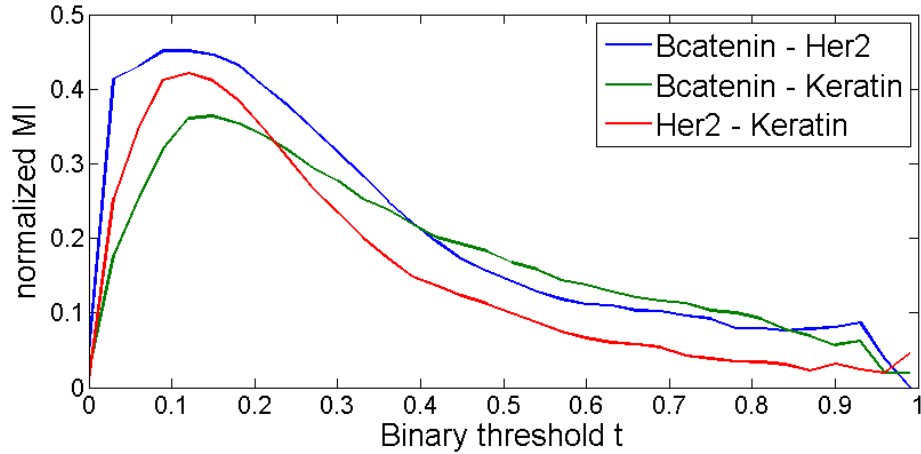
$$MI_{\text{normalized}}(I, J) = \frac{H(I) + H(J) - H(I, J)}{\max(H(I), H(J))}$$

where  $H(I)$  is the entropy of  $I$  and  $H(I, J)$  is the joint entropy of  $I$  and  $J$ .

$$H(I) = - \sum_{X \in \{I=0, I=1\}} p(X) * \log_2 p(X)$$

$$H(I, J) = - \sum_{X \in \{I=0, I=1\}} \sum_{Y \in \{J=0, J=1\}} p(X, Y) * \log_2 p(X, Y)$$

MI is normalized according to the image with largest entropy. The dependence of MI to  $t$  is illustrated in Figure 6.3. The MI shows a chi-squared behavior with a maximum of 0.35 to 0.45. Although the three proteins are known membrane proteins, they apparently are not completely identical in terms of location and expression rate, which emphasizes to combine their information for better membrane description.



**Figure 6.3: Pairwise normalized mutual information (MI) of the membrane proteins Her2, Bcatenin and Keratin after dichotomization of the images at a shifting threshold  $t \in [0; 1]$ .**

### 6.3.2 Cell Segmentation

Single cell segmentation in multiplexed images is different to the known nucleus segmentation in conventional IHC stained images, where the cell membranes are not specifically stained (Schüffler *et al.* 2013b). Superpixels, for example, would not identify whole cells as uniform image compartments with homogeneous content. Also graph-cut segmentation used for cell nucleus segmentation (Schuffler *et al.* 2010) will not be able to segment the multifaceted structure of a whole cell as a single unit. Watersheds on the other hand are best suited for cell segmentation when the cell is stained with membrane marker proteins. We tested our approach on human breast cancer images (Theurillat *et al.* 2007).

Due to the lack of a gold-standard, different segmentation masks were compared according to the presented segmentation score. An exhaustive search through all parameters revealed the best segmentations possible with the proposed method. The Dirichlet tessellation of cell nuclei achieved a maximum segmentation score of 0.65 which mainly results from the high over-segmentation of unpopulated areas in the image. High dimensional watershed segmentations yield a higher segmentation score of 0.70 (arithmetic and geometric combination, Figure 6.4). The segmentation algorithm further profits from the independent contribution of individual registered channels. The best segmentation with a score of 0.70 can be found with weights 0.63, 0.25 and 0.12 for  $\beta$ -catenin, Her2 and Keratin, respectively. This clearly demonstrates the advantage of multiplexed cytometry with multidimensional registered information over singleplexed methods accessing only one dimension.



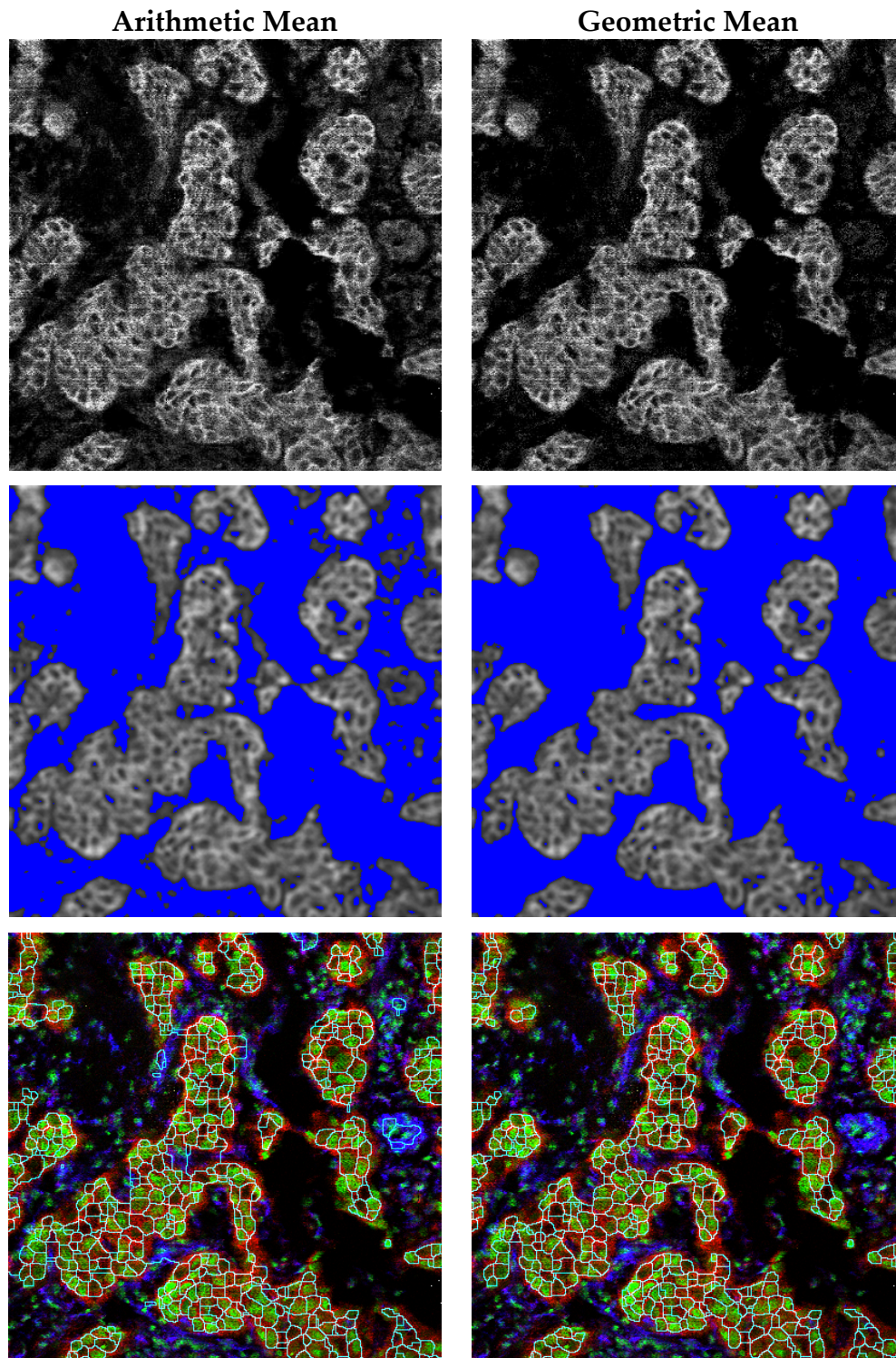


Figure 6.4: Process of watershed based single cell segmentation with multiplexed images  $\beta$ -catenin, Her2, Keratin and 1-H3 (see Figure 6.1). TOP: The averaged image as arithmetic mean (left) or geometric mean (right). MIDDLE: The combined image is smoothed with a Gaussian blur filter of radius 3. A threshold  $t=0.18$  excludes all image areas where no tissue cells are visible (blue). BOTTOM: A watershed segmentation of foreground area is overlaid to Her2 (red, membranes), H3 (green, nuclei) and Vimentin (blue, cytoplasm).

### 6.3.3 Implementation

The algorithms used in this paper are implemented in MATLAB (R2013b) and compiled as a GUI for Windows called *MultiplexedCellSegmentation* (see Figure 6.5). The program is freely available for non-commercial use at [www.comp-path.inf.ethz.ch](http://www.comp-path.inf.ethz.ch).

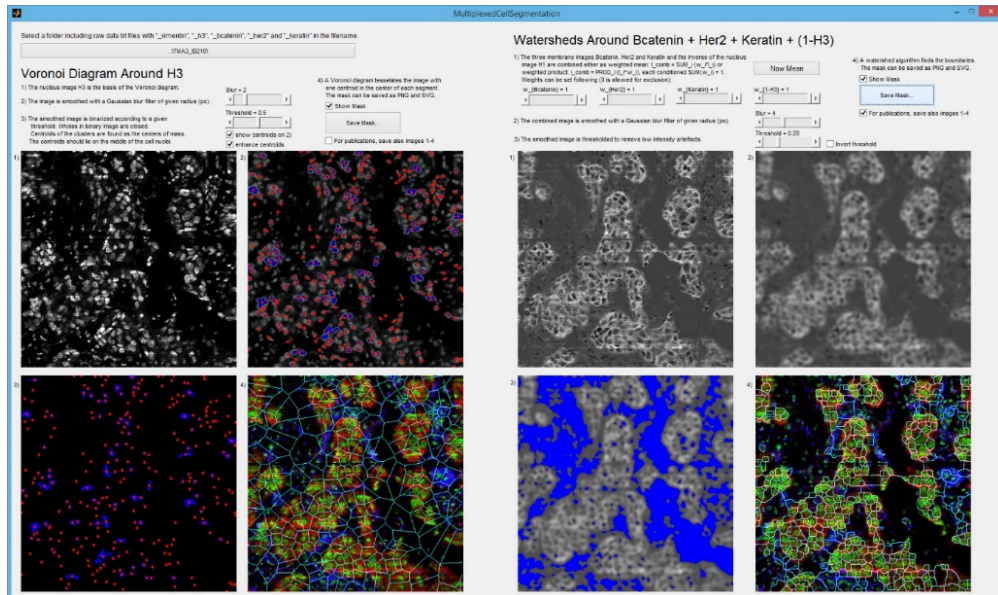


Figure 6.5: Screenshot of *MultiplexedCellSegmentation*, a program which implements the Voronoi tessellation (left) and the watershed segmentation (right) of tissue cells in multiplexed images.

## 6.4 Discussion

Cell segmentation in medical images is mandatory for single cell cancer research. IHC mass cytometry is a new promising imaging technique for multidimensional and perfectly registered quantitative and spatial protein expression profiles. The idea of this paper is to exploit the registered information of several protein channels for an improved segmentation compared to single-image segmentations. Since the value of a given segmentation dependent on the underlying medical task and since a gold-standard segmentation is missing for the new type of images at hand, we defined a new segmentation score to compare different segmentation masks. The simplistic score accounts for favorable passages of the segmentation overlapping with membrane signal and for unfavorable segmentations through a nucleus, which is a good start for comparison of different segmentation algorithms. Though, more sophisticated scores are possible, e.g. considering the number of non-segmented nuclei.

## 6.5 Conclusion

In contrast to IHC or IF, highly multiplexed mass cytometry provide a perfect registration of the target protein images. IHC and IF are commonly used to detect the expression of one or two proteins simultaneously. Other proteins are then processed on a subsequent tissue slice. For single cell analysis, a computational registration process is therefore necessary to align the different slices to each other. This process can be complicated or impeded when individual cells disappear on the subsequent tissue slice. Highly multiplexed mass cytometry circumvents this issue by simultaneously measuring dozens of proteins on *one tissue slice*.

Highly registered protein channels can be exploit to improve single cell segmentation which is mandatory for downstream single cell analysis. We have shown that the combination of the information of three membrane proteins Bcatenin, Her2 and Keratin as well as the nucleus protein H3 improves the single cell segmentation in human breast cancer images, compared to a segmentation which is based on only H3. A single cell segmentation in conventional IHC or IF images is typically based on singular protein markers for general morphological structure (e.g. hematoxylin staining), but not specific for cell membrane. The multidimensional view on single cells with highly registered dimensions enables more accurate cell segmentation.



# 7 SHAPE FEATURES FOR AUTOMATIC CROHN'S DISEASE DETECTION AND SEGMENTATION

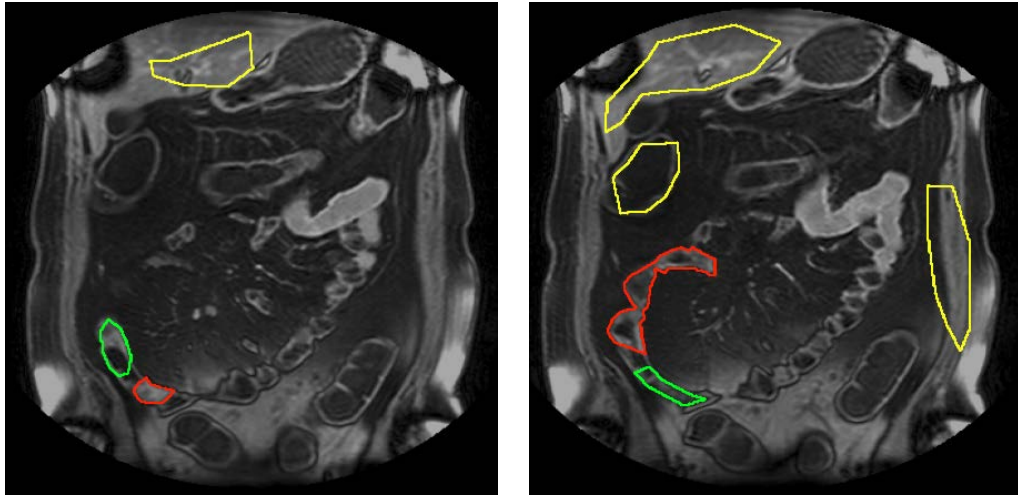
In the previous chapters, a shape measurement for cancerous renal clear cell nuclei was derived exploiting their description for improved cancer cell classification. The shape information was basically hidden in the dedicated feature design. We will now study this shape impact for medical image segmentation and demonstrate the importance of shape information for the detection and segmentation of Crohn's disease (CD) in MRI images.

## 7.1 Introduction

T1-weighted high-resolution isotropic volume examinations (THRIVE) after contrast agent supply are sharp and high-contrast magnetic resonance imaging (MRI) sequences for the visualization of CD. On 3D signal images, organic structures of small bowel, terminal ileum and colon are usually clearly visible. Signal enhancement in bowel wall as a result of CD activity can be recognized. THRIVE was therefore chosen to acquire manual segmentations by trained radiologists for training an automatic CD detection system. Two medical experts with more than 7 years of experience in abdominal MRI analysis identified and segmented diseased areas and healthy counter examples in bowel segments as well as background areas in MRI volumes of 26 patients (Figure 7.1). The scientific task is then to learn a computer algorithm to autonomously detect and segment similar regions in unseen images.

To achieve this goal, we start with a voxel-based method. Every image voxel is individually classified and the final segmentation results as a connected component of equally classified voxels. This approach is packed in a two-stage classification scenario: First, the pixels are classified into intestine or background. Second, the intestine is further classified into enhanced (diseased) or normal regions. Thereafter, we will refine the method in several ways. New features (e.g. spatial context features) are introduced. Further, the automatic detection of a region of interest as starting point is studied. Additionally, we investigate superpixels as unsupervised object segmentation in the context of CD. Finally, we evaluate weakly supervised learning methods as well as active learning scenarios for this label intensive task of CD detection in MRI.

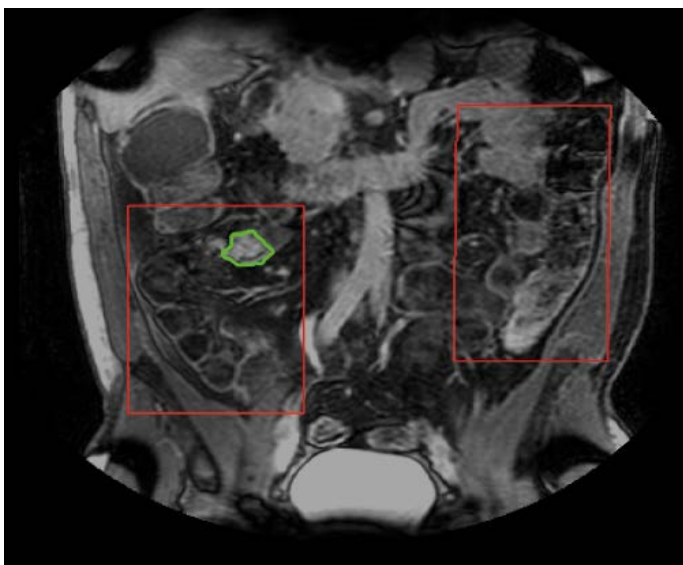
*Acknowledgement  
I want to thank  
Dwarikanath  
Mahapatra for the  
numerous discus-  
sions about Crohn's  
disease detection and  
segmentation. He  
prepared the manu-  
script for this section.*



**Figure 7.1:** Typical manual annotations of CD in MRI sequences on two different slices from one patient. Red: CD affected bowel segments. Green: normal bowel segments. Yellow: non-intestine regions. 26 CD patients are manually annotated analogously.

## 7.2 Voxel Based Classification

A voxel-wise MRI segmentation has been presented (Mahapatra *et al.* 2012b, 2013b) for the detection of CD. Since voxel (vx) based classification can be computationally infeasible for large images (in our case MRI volumes of  $400 \times 400 \times 100$ vx), we restrict this classification approach to a specific region of interest (ROI) within the whole volume. This localization is reasonable since the bowel appears in the middle slices, and even there, the bowel covers the area in the center of the images. Large peripheral parts of the scans show other organs or even background. Processing of the complete image would result in a large false positive rate. Figure 7.2 illustrates an example MRI slice with two ROIs (red) with a manually segmented diseased area (green).



**Figure 7.2:** Example MRI scan of human abdomen. Two rectangular regions of interest (red) contain intestine, background and manually segmented diseased bowel (green).

### 7.2.1 Feature extraction

For voxel-wise classification, texture and shape asymmetry features for each voxel in three different scales, each within a local neighborhood around the voxel were extracted. Trained radiologists usually rely on multiple data sources to identify diseased regions in a patient (e.g., different MRI sequences), whereas the automatic approach does not incorporate multimodal information without comprehensive inter-modality registration. In order to reveal hidden structures in the images which make the classification still possible on only one MRI sequence, we targeted image features that are not discernible by the human eye (Mahapatra *et al.* 2013b). It has been shown in psychological experiments, that the human visual perception is especially sensitive to first order and second order image features (mean and variance of intensity) (Julesz *et al.* 1973). **Higher order statistics**, such as *skewness* and *kurtosis*, are hidden to the observer. Still, they frequently vary among different image objects in MR images (Petrou *et al.* 2006), which is to be exploited for voxel classification. Therefore, we calculate for every image voxel in a given neighborhood the mean, variance, skewness and kurtosis for intensity and for texture. To respect scaling issues in the images, the neighborhoods are chosen at the scales 25x25, 30x30 and 35x35 pixels. We call a pixel with its given neighborhood a *patch*.

#### Skewness

The *skewness* of an image patch is defined by its third order moment: Let  $S_i$  be an image patch with  $N$  voxels where  $S_i(j)$  is the intensity value of its  $j$ th voxel,  $\bar{S}_i$  the mean intensity and  $\sigma_i^2$  the variance. Then the skewness  $Sk_i$  is defined by:

$$Sk_i = \left[ \frac{1}{N} \sum_{j=1}^N (S_i(j) - \bar{S}_i)^3 \right] * \frac{1}{\sigma_i^3}$$

Skewness can be understood as a measure of the symmetry of a distribution:  $Sk_i < 0$  indicates a shift of most values to the right side of the mean, and  $Sk_i > 0$  implicates a shift to the left side of the mean. A skewness is close to zero when the values are relatively equally distributed, e.g. for a Gaussian distribution.

#### Kurtosis

Similarly, the kurtosis  $Ku_i$  is defined by the fourth order moment:

$$Ku_i = \left[ \frac{1}{N} \sum_{j=1}^N (S_i(j) - \bar{S}_i)^4 \right] * \frac{1}{\sigma_i^4}$$

Kurtosis describes the style of the peaks a distribution: A high kurtosis means sharper peaks with larger, wider tails, whereas a low kurtosis describes more roundish peaks with smaller, thinner tails.

### Texture

2D Gabor filter banks are rich descriptors of texture in images and inherit various desired properties. Multi-scale and multi-orientation Gabor filter banks can capture visual characteristics such as spatial localization, orientation and spatial frequency. They have previously been used as texture descriptors in medical imaging, e.g. for simple cortical cell representation (Devalois *et al.* 1982; Manjunath and Ma 1996; Liu and Wechsler 2002). A Gabor filter bank is defined as:

$$g_{\gamma,\omega}(x,y) = a^\gamma g(a^\gamma(x * \cos(\omega\psi) + y * \sin(\omega\psi)))a^\gamma(y * \cos(\omega\psi) - x * \sin(\omega\psi))$$

where  $\gamma = 0, \dots, \Gamma - 1$  and  $\omega = 0, \dots, \Omega - 1$ .  $\psi = \pi/\Omega$  is the rotation factor and  $a = (\frac{U_h}{U_l})^{\frac{1}{\Gamma-1}}$  the scaling factor.  $U_h$  and  $U_l$  determine the frequency range of the filter bank and  $W$  is a shifting parameter in the frequency domain. Our Gaussian function  $g$  is defined by:

$$g(x,y) = \left( \frac{1}{2\pi\sigma_x\sigma_y} \right) \exp \left[ -\frac{1}{2} \left( \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} \right) + 2\pi j W x \right]$$

where  $\Gamma = 6$  is the total number of orientations and  $\Omega = 2$  the total number of scales. We obtain 12 texture maps by oriented Gabor filters in six orientations ( $0^\circ, 30^\circ, 60^\circ, 90^\circ, 120^\circ$  and  $150^\circ$ ) and at two scales (1 and 0.5) (Mahapatra *et al.* 2013b). A texture map is a result of the convolution of the intensity image patch with the corresponding Gabor filter.

### Shape Asymmetry

Shape asymmetry for skin lesion identification by Liu *et al.* (2011) is calculated as *bin* differences over the principal axis of a feature histogram. We show a novel shape asymmetry measure to quantify local orientation distributions by entropy. A given intensity patch is radially partitioned into 18 sectors by a circle centered in the patch. In each sector, the single pixel orientations are calculated and the entropy of the angle distribution is determined. The entropy is high for uniform distributions with unstructured orientation in the sector. Low entropy indicates a peaked distribution or a structured sector image. Also, a peaked distribution leads to a lower shape asymmetry. For a given sector  $r$  with its angle distribution  $p_\theta^r$ , the shape asymmetry is defined by the entropy:

$$Sh_{Asymmetry}^r = - \sum_{\theta} p_\theta^r * \log(p_\theta^r)$$



Figure 7.3 illustrates the calculation of the shape asymmetry for a diseased example image patch and a healthy example. Interestingly, the shape asymmetry feature has a more linear profile for the healthy patch than for the diseased patch. This difference might result from the fact that healthy regions in the image show a smoother, regular shape, while the diseased areas separate by structured shape due to lesions, ulcerations or other abnormalities.

According to the above description, the pixel-wise feature vector comprises 70 dimensions (intensity: 4, texture:  $4 \times 6 \times 2 = 48$  (mean, variance, skewness and kurtosis of texture maps) and shape asymmetry: 18). To respect image characteristics over multiple scales, we calculate the feature vector over three different neighborhood sizes:  $25 \times 25$ ,  $30 \times 30$  and  $35 \times 35$ , expanding the final feature vector length to  $3 \times 70 = 210$ .

### 7.2.2 Comparison Features

We compare the performance of our new features with two state-of-the-art methods: Dual tree complex wavelet transform (*DTCWT*) (Berks *et al.* 2011) and a shape-asymmetry based method (*Asy*) (Liu *et al.* 2011), both explained in the following sections.

#### *DTCWT*

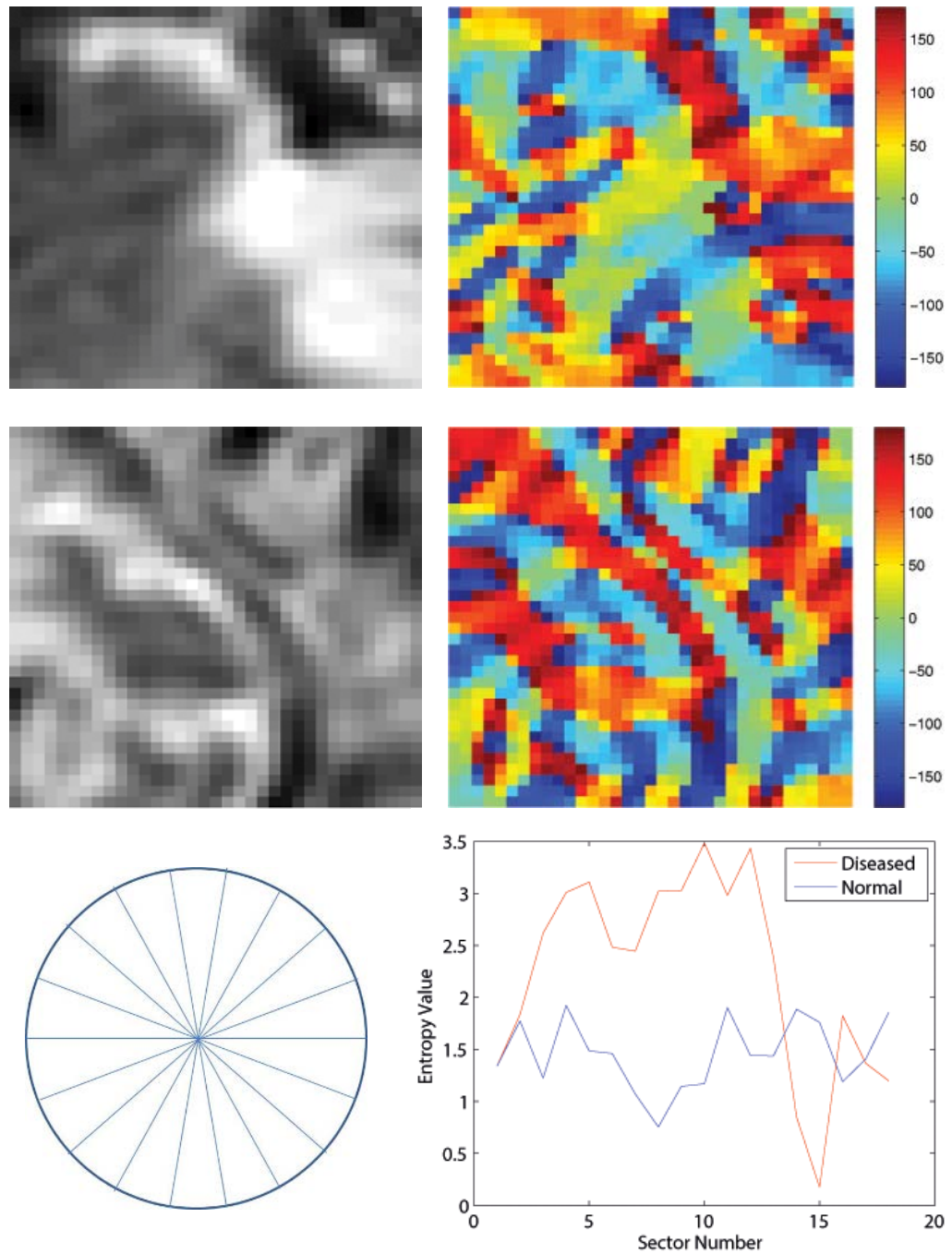
Rich descriptors of local structure are *dual tree complex wavelet transforms* (*DTCWT*), which have extensively been used in various kinds of image processing. *DTCWT* combines two discrete transforms, both shifted in their phase by  $90^\circ$ . For 2D images, *DTCWT* calculates 6 directional sub-bands oriented at  $\pm 15^\circ$ ,  $\pm 45^\circ$  and  $\pm 75^\circ$ . The *DTCWT* feature is calculated from the center pixel of a certain neighborhood. Applying the three neighborhoods from above, we obtain a  $3 \times 6 = 18$  dimensional feature vector per pixel. See Berks *et al.* (2011) for a detailed description of *DTCWT*.

#### *Asy*

Based on the *reflectional* asymmetry measure by Liu *et al.* (2011) for a pigmentation model of skin lesions, we develop a similar shape measure. Instead of global point signatures, we use orientation angle information. For every image patch, a 40-bin histogram of orientation angles is calculated with the magnitude of angles lying between  $-180^\circ$  and  $180^\circ$  preserving an equal number of bins for positive and negative magnitudes. *Asy* is then defined as:

$$Asy = \sum_{i=1}^{20} (h_i - h_{-i})$$

where  $i$  is the number of bins of the orientation histogram  $h$  of 40 bins.



**Figure 7.3: Illustration of the new shape asymmetry feature used for CD detection. For each pixel in the shown patches (left), the orientation angle is calculated ( $-180^{\circ}$ - $180^{\circ}$ ) and displayed as heatmap (right). TOP ROW: Diseased image patch. MIDDLE ROW: Normal image patch. BOTTOMLEFT: The patches are partitioned into 18 radial segments and the entropy of the orientations per segment is calculated. BOTTOMRIGHT: The entropy per segment is shown for the two examples (red line, diseased patch; blue line, healthy patch). Diseased and healthy patches differ in their entropy distribution.**

From every histogram count with positive orientation magnitude, the corresponding histogram count with negative orientation magnitude is subtracted (without absolute value). The 20 differences are then summed up to  $Asy$ . A negative  $Asy$  means that the orientation angles have a net negative leaning and a positive  $Asy$  indicates a net positive leaning, yet yielding a measure for asymmetry.

### 7.2.3 Dataset

We used the CD patient MRI data of 26 patients (19-72 years, mean 36 years, 17 females and 9 males), acquired at the AMC, Amsterdam, The Netherlands. The corresponding MRE protocol has been introduced in section 3.3. The resulting images have a spatial resolution of  $1.02 \times 1.02 \times 2$ mm at a dimension of  $400 \times 400 \times 100$  voxels.

Two radiologists have annotated regions that show strong wall enhancement as diseased areas as well as normal wall regions. Further, normal background (non-intestine) regions have been labeled to provide a supervised gold standard. In total, 6827 pixels from diseased regions, 5156 pixels from normal regions and 3725 pixels from background regions could be extracted from the annotations in all 26 patients. Every pixel forms a sample with its calculated feature vector in the three described neighborhoods.

### 7.2.4 Classification Scenario

We compare three different feature sets ( $DCTWT$ ,  $Asy$  and  $Our Features$ ). Additionally, we compare three different classifiers to show their ability to separate these types of MRI data. Therefore, we incorporated a Random Forest (RF), a Support Vector Machine (SVM) and a Bayesian Classifier (BC).

#### **Random Forest**

Random Forests are classifier ensembles of decision trees (Breiman 2001). They have been successfully used in various kinds of machine learning domains, such as computer vision, medical imaging and bioinformatics (Cima *et al.* 2011; Fuchs *et al.* 2011b; Soldini *et al.* 2013). Each decision tree is trained on a different subset of the training data, to generalize the ensemble's prediction accuracy. For each tree, a different subset of features and feature parameters is significant for prediction. A majority vote among all trees then forms the probabilistic outcome of the forest. We used a RF with 100 trees in our experiments.

### Support Vector Machine

Support Vector Machines construct hyper-planes in the feature space which separate the data points. The hyper-plane is constructed such that the distance to the nearest data points of any class (support vectors) is maximum, since this assures best separation also for unseen data. SVM are very flexible for classification, since they can be designed with any metric distance measure between data (called kernel function), at any dimension. This makes them highly favorable for a variety of tasks such as brain tumor segmentation (Bauer *et al.* 2011), chest pathologies (Avni *et al.* 2011) or cell nucleus classification (Schüffler *et al.* 2010; Schüffler *et al.* 2011). We use the publicly available LIBSVM package (Chang *et al.* 2011) with a radial basis kernel function (RBF).

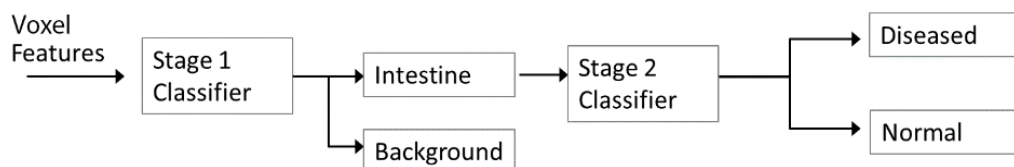
### Bayesian Classifier

As a third comparison, we chose a Bayesian Classifier as it stresses the non-linear nature of our data. We use the default naïve Bayesian Classifier in MATLAB.

All classifiers and feature sets were tested with 10-fold cross validation. In this scenario, the total data are partitioned into 10 equally sized subsets. In each fold, nine of the ten subsets are used to train the classifiers and the remaining data part is used to test them.

To reduce the feature space hierarchically and to obtain better classification results, we designed a two-stage classification scenario illustrated in Figure 7.4: Stage 1 classifies pixels into intestine and background (trained on the full dataset fold). Stage 2 then classifies the intestine samples into diseased and normal (trained only on the intestine samples of the fold). The classification results are reported in the following sections.

Since we perform a voxel-wise classification, we record accuracy, sensitivity, specificity and precision of a classifier's test with the definitions via a contingency table given in Table 7.1. Later in section 7.2.8, we will also refer to Dice metric and Hausdorff distance as performance measure.



**Figure 7.4: Hierarchical two-stage classification of voxels for CD segmentation. In stage 1, the samples are classified as intestine and background to reduce the variability of the following second stage classification. Stage 2 only considers diseased and normal intestine.**

**Table 7.1: Definition of accuracy, sensitivity, specificity and precision (positive predictive value) used in our experiments via contingency table. Indicated are true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).**

		Gold standard		
		Positive	Negative	
Machine	Positive	TP	FP	$Precision = \frac{TP}{TP + FP}$
	Negative	FN	TN	
		$= \text{Sensitivity}$ $= \frac{TP}{TP + FN}$	$\text{Specificity}$ $= \frac{TN}{FP + TN}$	$\text{Accuracy}$ $= \frac{TP + TN}{TP + FP + FN + TN}$

### 7.2.5 Results – 1<sup>st</sup> Stage: Intestine vs. Background

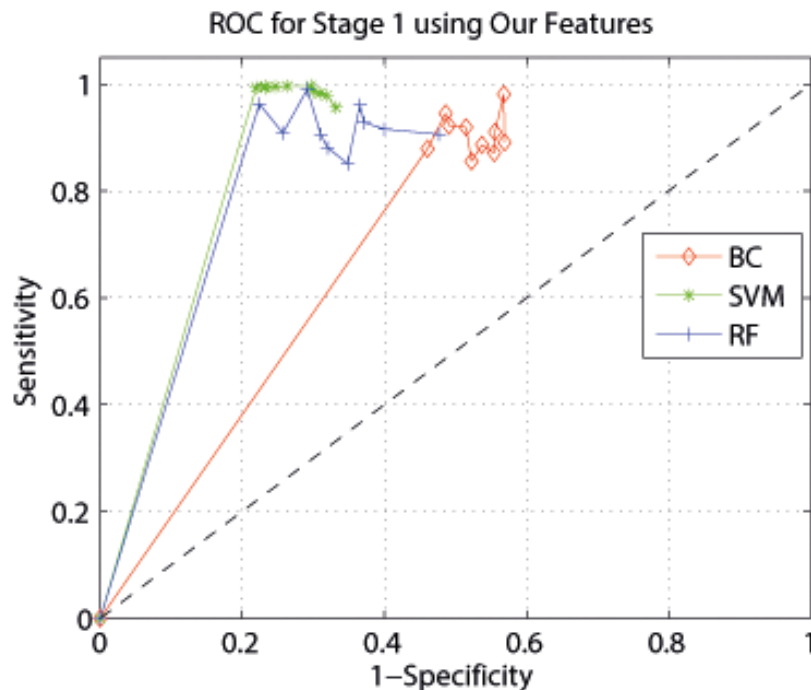
Each pixel is classified into intestine or background in a 10-fold cross-validation approach. *Our features* obtain the highest classification result with a support vector machine (RBF kernel). Table 7.2 compares the classification results. Note that we want to achieve a high sensitivity or true positive rate at this stage, even at the cost of a poorer overall accuracy. Samples that are classified as intestine are subjected to the subsequent 2<sup>nd</sup> stage classification. We observe background samples wrongly classified as intestine are invariably identified as normal. Further, background samples which are classified as diseased could easily be discarded by a clinician. On the other hand, intestine samples which are classified as background in this 1<sup>st</sup> stage are sorted out for disease classification which is highly undesirable.

The receiver operating characteristic (ROC) curves of the SVM classifiers using our features show overall a high sensitivity over 90 % in each cross-validation run (see Figure 7.5). In contrast, the specificity is considerably lower indicating the large number of false positives – background samples classified to intestine. This type I error is more desired than type II error, since the background samples are classified as normal in the 2<sup>nd</sup> classification stage. The naïve Bayesian classifier shows the lowest specificity (<50 %) and overall accuracy (<75 %), which might result from the non-linearity of the data. Naive Bayesian classifiers find optimal decision

boundaries on linearly separable data or non-overlapping data, while RF and SVM with RBF kernel can handle non-linearities to certain extent. Our results suggest that the data show these non-linear characteristics. Still, all classifiers perform better than random.

**Table 7.2: Quantitative classification results of Stage 1 classification. Listed are mean (and standard deviation) of 10-fold cross-validation.**

	<i>Asy</i>			<i>DTCWT</i>			<i>Our Features</i>		
	RF	SVM	BC	RF	SVM	BC	RF	SVM	BC
<b>Accuracy (%)</b>	79.9 (2.2)	80.4 (2.6)	72.0 (2.3)	80.1 (2.5)	82.2 (2.4)	71.3 (2.9)	83.3 (4.1)	<b>86.4 (1.5)</b>	73.2 (4.4)
<b>Specificity (%)</b>	68.0 (1.7)	67.9 (1.8)	41.5 (1.8)	67.6 (1.8)	68.1 (1.6)	42.7 (1.7)	70.6 (2.2)	<b>71.1 (1.8)</b>	49.1 (2.1)
<b>Sensitivity (%)</b>	84.6 (1.8)	86.2 (1.9)	81.5 (1.4)	85.7 (1.9)	93.9 (2.7)	88.3 (2.1)	92.1 (4.1)	<b>96.7 (1.2)</b>	90.1 (6.5)
<b>Precision (%)</b>	89.6 (1.3)	90.1 (1.1)	77.9 (1.8)	89.5 (1.1)	92.1 (2.1)	78.8 (1.7)	90.9 (2.9)	<b>96.3 (1.8)</b>	80.4 (4.2)



**Figure 7.5: ROC curves of the classifiers using our features. Each curve consists of 10 data points, each from one cross-validation run. While the SVM and RF show both high specificity and sensitivity, the BC classifies weigh more false positives. This illustrates non-linear data structures which can be handled by SVM and RF.**

### 7.2.6 Single Feature Contribution

We tested the contribution of the single feature components (intensity, texture and shape asymmetry). Table 7.3 lists the accuracy and sensitivity of the RF classifier using only a subset of features. Interestingly, intensity alone shows lowest accuracy as it does not incorporate only coarsely structured information. Incorporating texture or shape features significantly improves the classification result. A t-test on the cross-validation for *Shape+Tex* and *all features* from Table 7.2 with  $p < 0.032$  suggests that the combination of all three feature vectors clearly improves the classification statistically significantly.

**Table 7.3: Quantitative measures for the contribution of individual feature groups for the RF. Values indicate mean (and std. dev.) of 10-fold cross validation.**

	Int	Tex	Shape	Tex + Int	Shape + Int	Shape + Tex
<b>Accuracy</b>	77.1	81.6	79.1	79.2	79.5	<b>82.3</b>
<b>(%)</b>	(2.3)	(2.1)	(2.7)	(1.3)	(2.4)	<b>(1.3)</b>
<b>Sensitivity</b>	79.3	<b>86.9</b>	82.3	83.1	83.8	86.6
<b>(%)</b>	(3.2)	<b>(2.1)</b>	(1.9)	(3.1)	(2.3)	(2.8)

### 7.2.7 Results – 2<sup>nd</sup> Stage: Diseased vs. Normal Intestine

Samples which have been classified as intestine in the 1<sup>st</sup> stage are now subjected to classification into diseased or normal. Due to the hierarchical classification workflow, it is necessary to detail the definition of performance measures for this stage.

Let  $N$  be the number of intestine samples at the beginning of stage 1.  $N_d$  from these are diseased, and  $N_n$  are normal, i.e.  $N = N_d + N_n$ . After stage 1,  $N_2$  is the number of correctly classified intestine samples, out of which  $N_{d2}$  are diseased and  $N_{n2}$  are normal. In stage 1, the sensitivity is calculated using  $N_2$  and  $N$ . In stage 2,  $N_2$  is considered: supposed the number of correctly classified diseased samples is  $N_{d3}$  and the number of correctly classified normal samples is  $N_{n3}$ . The performance measures of stage 2 are based on the original number of samples at the start of stage 1, i.e.  $N_d$  and  $N_n$ . Table 7.4 defines our performance measures (Mahapatra *et al.* 2013b):

**Table 7.4: Definition of performance measures for stage 2 classification classifying exclusively intestine samples. The performance is based on the original number of samples  $N$ .**

<b>Accuracy:</b> number of diseased and normal samples correctly classified.	$\frac{Nd3 + Nn3}{N}$
<b>True positives (TP):</b> number of correctly classified diseased samples.	$Nd3$
<b>True negatives (TN):</b> number of correctly classified normal samples.	$Nn3$
<b>False positives (FP):</b> number of normal samples classified as diseased.	$Nn - Nn3$
<b>False negatives (FN):</b> number of diseased samples classified as normal.	$Nd - Nd3$
<b>True positive rate (TPR):</b> same as <i>sensitivity</i> or <i>recall</i> .	$\frac{TP}{TP + FN} = \frac{Nd3}{Nd}$
<b>True negative rate (TNR):</b> same as <i>specificity</i> .	$\frac{TN}{TN + FP} = \frac{Nn3}{Nn}$
<b>Precision: the fraction of retrieved diseased instances.</b>	$\frac{TP}{TP + FP} = \frac{Nd3}{Nd3 + Nn - Nn3}$

The 2<sup>nd</sup> stage classification performance of all features and classifiers is listed in Table 7.5. Again, we observe the SVM to separate the data best. All measures reach approximately 90%. On the other hand, we observe the BC to drastically loose accuracy compared to stage 1 classification, a drop of over 14% in accuracy which is at the same time not observable for the other classifiers. Again this indicates that the data are not linearly separable, and the data points overlap even more complex when only considering diseased and normal samples. The differences between background and intestine seems to be larger (and more capable for BC) than between diseased and normal intestine. Due to the larger differences in the first stage, the BC could comparably better classify background and intestine. However, SVN and RF still are able to classify the smaller differences of diseased and normal bowel enhancement samples at very high accuracy around 90%.



**Table 7.5: Quantitative classification result of Stage 2 classification. Listed are mean (and standard deviation) of 10-fold cross-validation.**

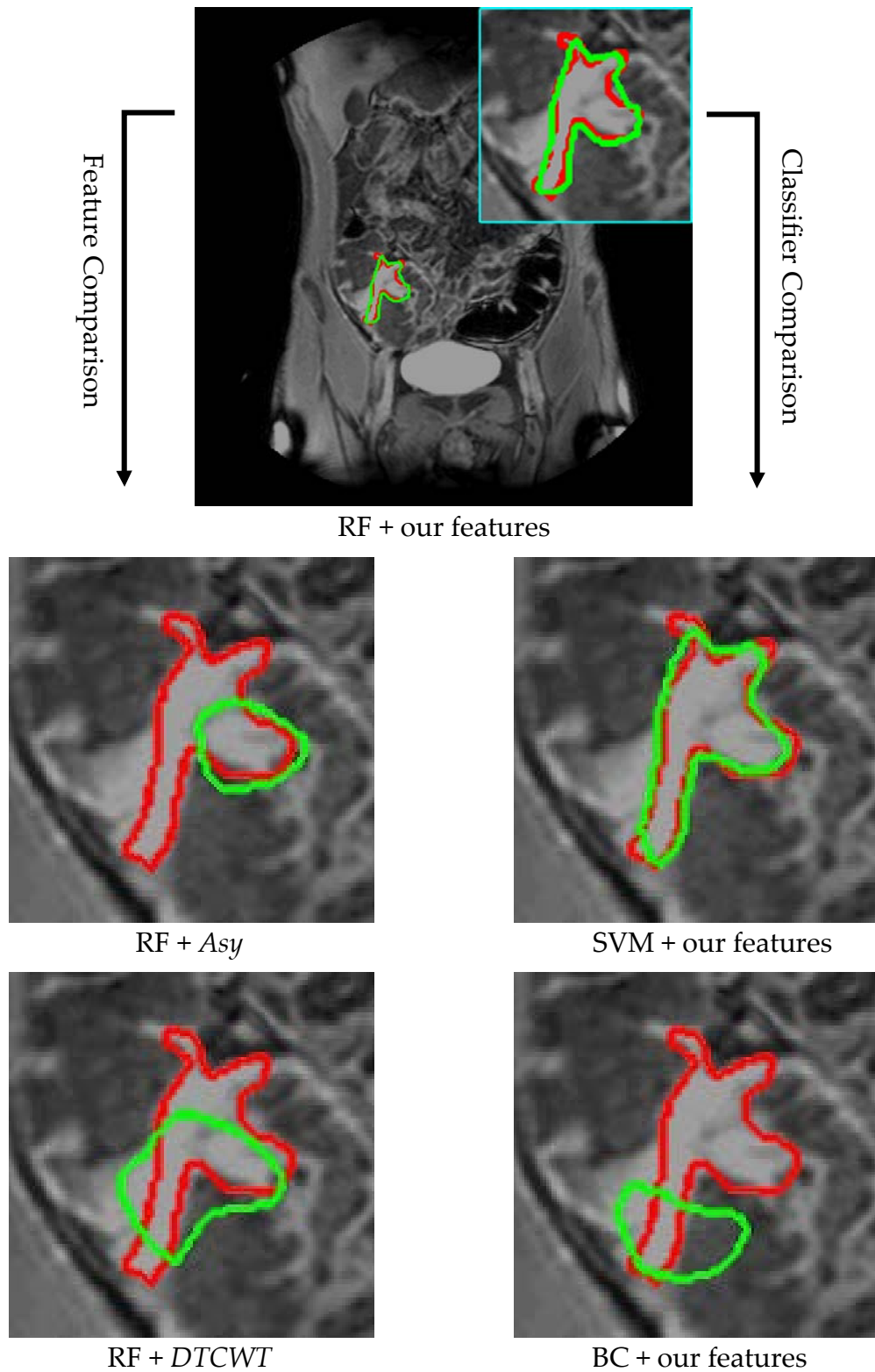
	<i>Asy</i>			<i>DTCWT</i>			<i>Our Features</i>		
	RF	SV M	BC	RF	SV M	BC	RF	SV M	BC
<b>Accuracy (%)</b>	81.7 (1.2)	81.5 (1.3)	59.1 (0.9)	81.9 (1.2)	82.2 (1.4)	58.4 (6.1)	88.9 (1.5)	<b>89.5 (2.6)</b>	62.8 (5.4)
<b>Specificity (%)</b>	83.4 (2.4)	80.8 (3.1)	35.1 (4.8)	84.2 (1.9)	82.8 (1.4)	37.7 (2.7)	90.1 (1.6)	<b>90.2 (1.7)</b>	39.3 (4.1)
<b>Sensitivity (%)</b>	84.9 (1.8)	84.5 (1.9)	60.5 (1.2)	86.1 (1.9)	86.9 (1.7)	61.3 (8.2)	90.4 (1.2)	<b>91.9 (2.6)</b>	64.8 (9.7)
<b>Precision (%)</b>	82.7 (1.5)	82.9 (1.4)	59.7 (1.9)	84.9 (1.4)	85.3 (1.4)	59.7 (5.4)	88.9 (1.3)	<b>90.2 (2.0)</b>	63.3 (4.3)

### 7.2.8 Whole Patient Classification

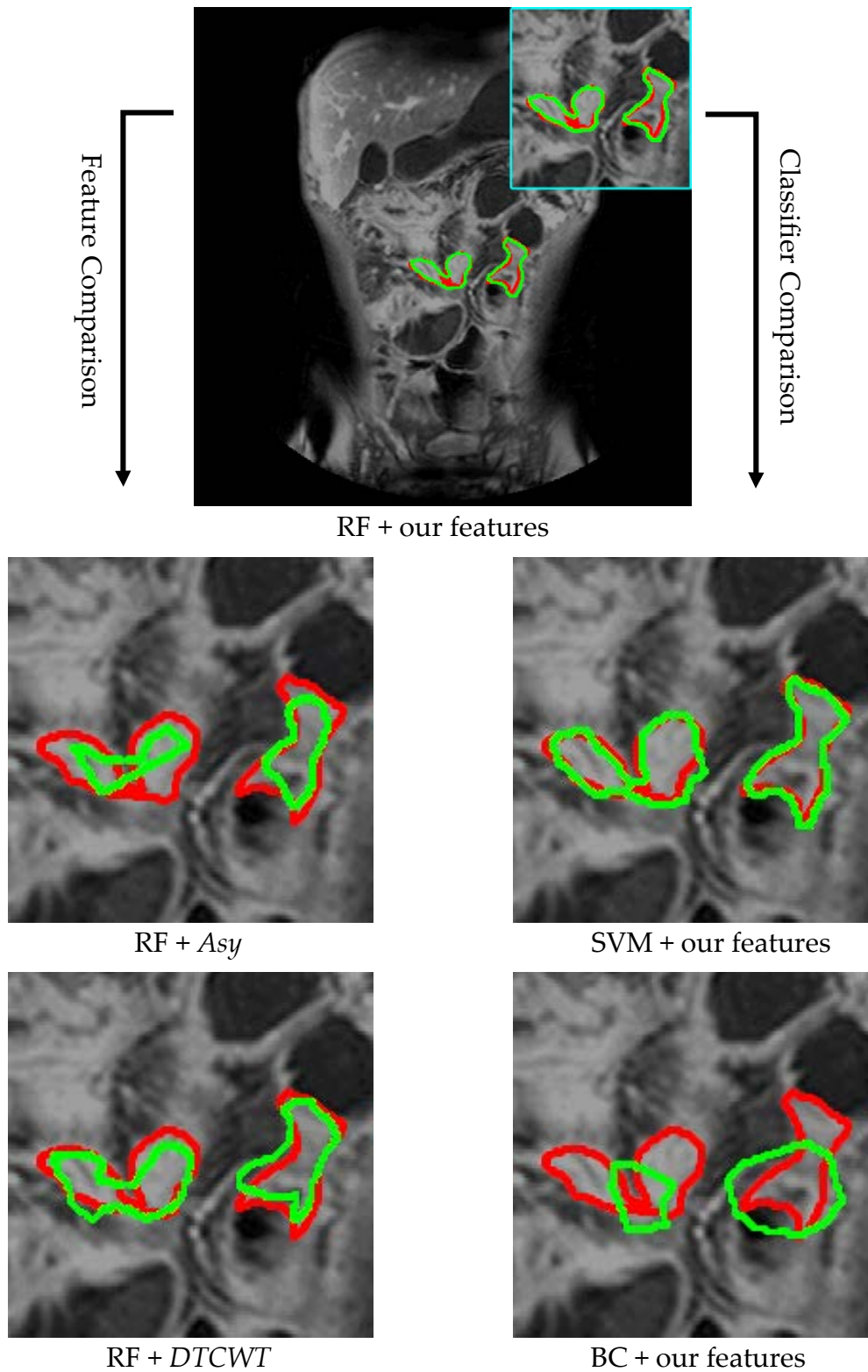
In this section, we demonstrate the visual validation of our approach in a leave-one-patient-out cross-validation (LOPO-CV) with whole patient classification. The classifiers are trained on the annotations of 25 patients as described above. Thereafter, all ROIs of the remaining patient are classified with the two-stage classification voxel per voxel. Those voxels classified as diseased constitute then the predicted CD affected region in the patient, which is compared to the doctor’s manual annotations.

For whole patient classification, we apply two post processing steps after prediction, since we frequently obtain several small detected diseased regions and one or two larger clusters. We firstly remove all detected diseased regions smaller than 10 pixels (smaller than one cm). Second, we transform the larger clusters to connected continuous diseased region using contour fitting. We show in Figure 7.6 and Figure 7.7 two examples of patient 23 with one annotated diseased region (red) and patient 16 with two annotated diseased regions (red). On the left column, the three different features *Asy*, *DTCWT* and *our features* are visually compared for whole patient annotation. The suggested prediction is depicted in green. Our features produce the best overlap with the annotated red region. On the right column of the image, the three classifiers are compared to our features. Here, we observe the similar performance of RF and SVM while the BC shows its difficulties to especially detect the diseased peripheral pixels of the annotation.

As illustrated in Figure 7.6 and Figure 7.7, we can compute the region-matching quantifications Dice metric (DM) or Hausdorff distance (HD), as explained below.



**Figure 7.6:** Visual detection and segmentation results of diseased CD regions of patient 23. The gold standard manual annotation of CD activity is indicated in red. The computer's segmentation is delineated in green. The left column compares different features with RF. The right column compares different classifiers with our features.



**Figure 7.7:** Visual detection and segmentation results of diseased CD regions of patient 16. The gold standard manual annotation of CD activity is indicated in red. The computer's segmentation is delineated in green. The left column compares different features with RF. The right column compares different classifiers with our features.

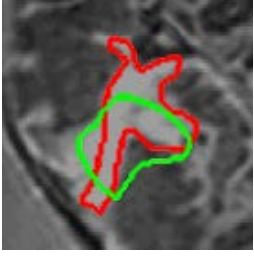


Figure 7.8: Example of manual segmentation  $M$  (red) and algorithmic segmentation  $A$  (green).

### Dice Metric

A measure for the relative amount of overlap of two datasets was proposed by Dice (1945). We use this measure for the evaluation of two overlapping areas  $A$  (algorithm) and  $M$  (manual annotation) (c.f. Figure 7.8). The Dice metric  $DM$  is defined as:

$$DM = \frac{2 * |A \cap M|}{|A| + |M|}$$

DM ranges from 0 (no overlap) to 1 (perfect overlap) and might be interpreted as  $F$ -score for detection.

### Hausdorff Distance

While DM considers the relative amount of overlap of two segmentation, the distance of the boundaries is not reflected. The Hausdorff distance  $HD$  (Hausdorff 1957) aims to quantify the distance between two corresponding contours. Let  $A = \{a_1, a_2, \dots\}$  be the set of points of an algorithmic contour and  $M = \{m_1, m_2, \dots\}$  be the set of points of a manually annotated contour. For each point  $a_i \in A$ , the distance to the closest point (DCP) on  $M$  is calculated, and *vice versa*. We define then  $HD \in \mathbb{R}_0^+$  as:

$$HD(A, M) = \max \left( \max_i (DCP(a_i, M)), \max_j (DCP(m_j, A)) \right)$$

A  $HD = 0$  indicates perfect overlap of two contours. The larger HD is, the more distinct are two contours.

The manual segmentation serves as gold-standard to which the computed segmentation is to be matched. Table 7.6 summarizes the average DM and HD for all regions in our dataset. RF and SVM lead to CD segmentation in MRI scans with highest DM (>90% match) and lowest HD (approximately 2 pixels difference at the maximum deviance). BC again indicates the difficulty to separate the samples in a linear space. Experiments with a SVM with linear kernel have confirmed the assumption of linear non-separable data (data not shown).

Table 7.6: Average Dice metric (DM) and Hausdorff distance (HD) for CD detection with different features and classifiers. Values are mean (and standard deviation) of a LOPO-CV.

	Asy			DTCWT			Our Features		
	RF	SVM	BC	RF	SVM	BC	RF	SVM	BC
<b>DM</b>	84.7	84.1	79.3	85.6	85.3	80.1	<b>90.9</b>	90.3	81.8
<b>(%)</b>	(1.5)	(1.1)	(2.4)	(1.4)	(2.1)	(2.3)	<b>(1.2)</b>	(2.1)	(2.1)
<b>HD</b>	3.5	3.4	6.3	3.2	3.1	6.7	<b>2.0</b>	2.1	4.8
<b>(px)</b>	(1.8)	(2.1)	(2.8)	(2.1)	(1.2)	(2.2)	<b>(1.1)</b>	(1.1)	(1.6)

### 7.3 Automated Preprocessing and Post-Processing

Manual interventions in the aforementioned segmentation approach appear mainly in two important time points. Before image processing, a ROI respectively VOI has to be identified to reduce the computational effort and also to reduce the variable search space to regions which show intestine. Our experiments have shown that human anatomic structures and neighboring organs such as kidneys, liver, aorta and others often show similar voxel signals and structures as the intestine thus complicating the whole-scan analysis drastically. This effect can significantly be scaled down when concentrating on specific ROIs which scheme the rough area of the bowel beforehand and therefore exclude a large but non-relevant image part.

Second, after voxel-wise segmentation, the classified voxels are post-processed to exclude classification noise and diseased regions smaller than 10 voxels. Single voxels can be classified as diseased even if they do not form spatial clusters with neighboring voxels as our approach does not consider smoothness or spatial regularization.

We extend the method to automate these processes as far as possible (Mahapatra *et al.* 2013a). A method has been developed to detect slices and ROIs containing bowel information. For this, the image slices are partitioned into 30x30px image patches. A separate RF classifies these patches into “bowel” or “non-bowel”, based on similar features as stated above (mean, variance, skewness and kurtosis of intensity, texture and curvature values (Mahapatra *et al.* 2013a)). A set of classified “bowel” patches defines the ROI and a set of ROIs in adjacent slices defines the VOI (see Figure 7.9). For training the RF, the image patches with manual segmentation have been used as “bowel” examples. “Non-bowel” patches have been sampled by visual inspection (Mahapatra *et al.* 2013a).



Figure 7.9: Example of localized ROIs in three MRI slices of one patient. Green are the manual CD annotations of the radiologists. The red rectangle indicates the detected ROI around the annotations.

To incorporate spatial smoothness in the voxel-based classification and to reduce the number of singular diseased voxels, a second order Markov random field (MRF) energy function  $E$  has been introduced which is solved by graph-cut (Mahapatra *et al.* 2013a):

$$E(L) = \sum_{s \in P} D(L_s) + \lambda \sum_{s \in P} \sum_{t \in N_s} V(L_s, L_t)$$

where  $P$  is the set of pixels  $s$  of the patch,  $L_s$  is the label of pixel  $s$  and  $N_s$  is the neighborhood of  $s$ . The cost function is optimized using graph-cuts (Boykov *et al.* 2001).  $\lambda$  regularizes the relative contribution of the penalty cost  $D$  and the smoothness cost  $V$ .

The penalty  $D$  is defined by:

$$D(L_s) = -\log(\text{Pr}(L_s) + \varepsilon)$$

where  $Pr$  refers to the probability heatmaps of the pixel  $s$  to belong to class *background*, *normal* or *diseased*, obtained by the RF classification.  $\varepsilon = 0.00001$  ensures real logarithm values.

The smoothness cost function  $V$  incorporates semantic information from the feature rankings of the RF. The three different feature types *intensity*, *texture* and *curvature* are ranked by the RF according to their importance providing their semantic information. Let  $w_I$  (intensity),  $w_T$  (texture) and  $w_C$  (curvature) be the RF weights of the features, then  $V$  is defined by:

$$V(L_s, L_t) = \begin{cases} w_I V_I + w_T V_T + w_C V_C & L_s \neq L_t \\ 0 & L_s = L_t \end{cases}$$

where  $V_I$ ,  $V_T$ ,  $V_C$  are the contributions by *intensity*, *texture* and *curvature* to the smoothness:

$$V_I(L_s, L_t) = \frac{1}{\|s - t\|} * e^{-\frac{(I_s - I_t)^2}{2\sigma^2}}$$

where  $I$  denotes the intensity of pixels  $s$  and  $t$  and  $\sigma$  is the intensity variance over  $N_s$  (8 neighbors of  $s$ ). Analogously are defined  $V_T$  and  $V_C$  with *texture* and *curvature* instead of *intensity* (Mahapatra *et al.* 2013a). After training on a subset of 10 patients, following parameters were determined:  $w_I = 0.23$ ,  $w_T = 0.33$ ,  $w_C = 0.33$  and  $\lambda = 0.02$ .

A further addition in this work is the exploitation of spatial context information of the voxels (Mahapatra *et al.* 2013a). Since the human anatomy is relatively constant throughout the images (neglecting missing organs), we expect to gain information of the voxels by their relative position in the image. Context information has already been shown to improve oriented organ segmentation (Tu and Bai 2010; Mahapatra and Buhmann 2012a; Mahapatra and Sun 2012c). We therefore extend a

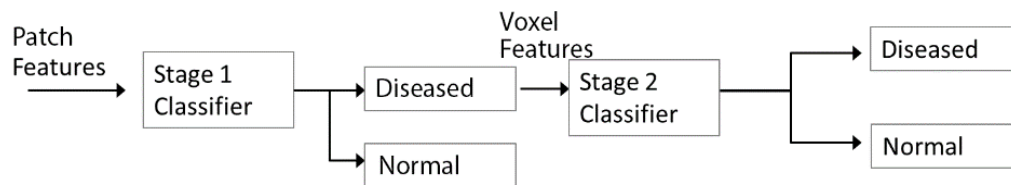
voxel's feature vector by image information of remote image areas at distances of 3, 8, 15 and 22 pixels in eight cardinal directions. On each remote sample point, the mean intensity, texture and curvature values of a  $3 \times 3 \times 3 \times v_x$  region was extracted and concatenated to the original feature vector.

These modifications of the voxel-based method lead to a much more automated classification design, in which (a) the preprocessing for ROI definition is completely automated and (b) the post-processing for false positive filtering of small non-connected voxel clusters as minimized due to smooth spatial constraints. Our experiments show that we still yield an accurate segmentation compared with the original method. In a leave-one-patient-out cross-validation, the mean DM among 26 patients is 85.5 % for the more automated method and the mean HD is 3.2 pixels (Mahapatra *et al.* 2013a).

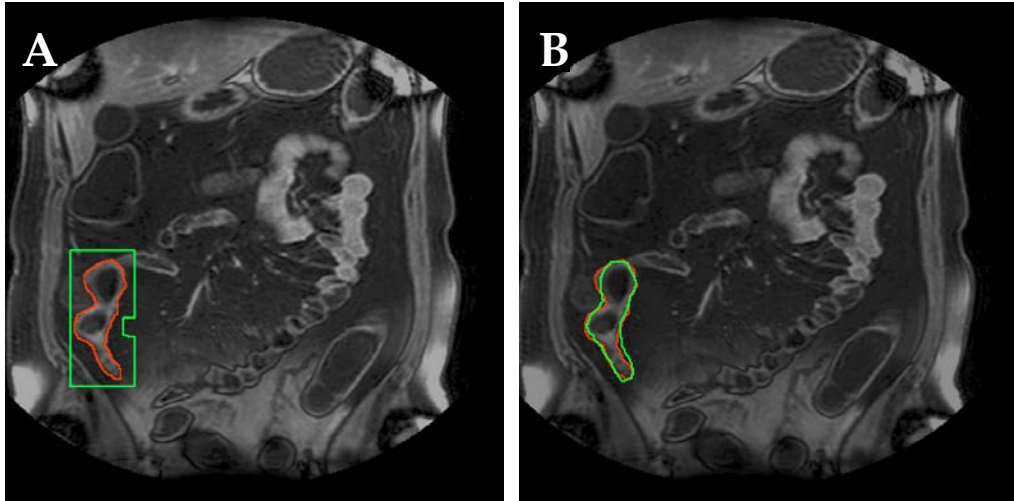
## 7.4 Novel Context Features Refine Automatic CD Detection

We study a slightly different workflow for CD detection and segmentation in MRI images as described in Mahapatra *et al.* (2013d). We could successfully improve the CD segmentation by a redesign of the two stage classification. Figure 7.10 illustrates the new classification pipeline: in the first stage, the whole image is partitioned into  $11 \times 11 \times 7$  3D patches from which intensity and texture features are extracted as explained below. Additionally, spatial context features are incorporated to exploit spatial constraints and to smooth the classification. The classification of the image patches replaces the former definition of a ROI. Those patches classified as diseased are subjected to the second stage, where a voxel-wise classification is performed to further segment the proper location of the diseased region within the patch (Mahapatra *et al.* 2013d).

*Acknowledgement*  
My sincere thanks to my colleague Dwari-kanath who prepared the manuscript and refined the CD detection method.



**Figure 7.10: Scheme of the modified hierarchical two-stage classification framework. In the first stage, larger image patches ( $11 \times 11 \times 7 \times v_x$ ) are classified to contain disease information or not. The patches form the samples for the first stage classifier. In the second stage, a voxel-wise classification sharpens the resolution of disease segmentation. Here, one voxel is one sample for the sage 2 classifier.**



**Figure 7.11: Visual example of the modified classification pipeline. A:** Image patches of size  $11 \times 11 \times 7$  voxels within the green rectangle are classified as “diseased”. The red area outlines the manual CD affected annotation. **B:** Within the green VOI of A, a per-voxel classification further refines the predicted diseased area (green) which aligns to the manual annotation (red).

With this modification, we achieve with standard RF classifiers a DM of 91.9% ( $\pm 1.9\%$ ) and a HD of 5.9px ( $\pm 2.3$ px) (mean and SD of LOPO-CV) (Mahapatra *et al.* 2013d). Figure 7.11 shows an example patient processed with the two new stages. On the left, the diseased image patches are detected as “VOI”. On the right, the voxels inside the VOI are classified as normal or diseased and a contour is fitted around the diseased voxels.

## 7.5 Supervoxels for VOI Detection

To scale down the computational load of voxel based classification, the definition of a prior volume of interest (VOI) is an essential need. In the previous sections, we showed that the use of a VOI as basis for voxel classification reduces the number of false positives in the outlying area. We proposed a sliding window approach of an  $11 \times 11 \times 7$  patch to classify rectangular coarse diseased areas in the image. A connected set of identified diseased areas then forms a VOI.

This method is now expanded for the use of supervoxels (Mahapatra *et al.* 2013c). We implement the algorithm of SLIC supervoxels (Achanta *et al.* 2012), already used for cell nucleus classification in section 4.2.2, to over-segment the whole MRI volumes into areas with homogeneous texture values. Since supervoxels inherently segment homogeneous regions with similar texture or intensity, we expect from the use of supervoxels a more accurate VOI detection which is not constrained on rectangular shape but rather respect the irregular image structures as boundaries.



For training, the supervoxels are assigned to the class of the majority of comprised manually annotated voxels (*diseased*, *normal* or *background*). Although label ambiguities are rare, the best suitable size of superpixels has to be determined empirically (see section 7.5.1). Intensity statistics, texture anisotropy and curvature anisotropy (analogous to texture) are used as features for the supervoxels (Mahapatra *et al.* 2013c). After supervoxel classification, we observe following effects:

First, despite the fact that manually annotated diseased regions commonly are only few relatively large clusters in one slice comprising more than one supervoxels, single supervoxels are occasionally wrongly detected as diseased. This is not a large drawback since these false positives will be classified as normal in the 2<sup>nd</sup> stage classification. Still, the computational load should be decreased if possible. Therefore, we filter single diseased supervoxels only keeping clusters of more than one supervoxels. In the case exclusively singular supervoxels are detected in an image, we keep only the largest supervoxel (Mahapatra *et al.* 2013c).

Second, especially when the number of labeled diseased voxels is low, the supervoxel tends to be classified as non-diseased, escaping the subsequent voxel-wise segmentation. This is explainable due to the fact that the evidence of disease might be small in this supervoxel as well as the contribution of the diseased part to the feature vector. Hence, the false negative rate will considerably increase after segmentation which is unwanted in this medical problem. To overcome this shortcoming, the final detected VOI is defined as the detected diseased supervoxels plus their immediate neighbor supervoxels. This enlargement of the VOI maximally reduces the occurrence of false negatives voxels (Mahapatra *et al.* 2013c). Figure 7.12 demonstrates particularly well the advantages of the proposed VOI detection and processing on an example slice.

### 7.5.1 *Effect of Supervoxel Size*

As small supervoxels tend to be more homogenous, their features are more representative for a single class. However they may not always provide sufficient number of voxels to estimate stable features. Large supervoxels often contain enough voxels to calculate statistically stable features but they may contain voxels from more than one class generating label ambiguities. Consequently, the extracted features may not be representative of one class. Thus, training with features from very large supervoxels leads to low classification accuracy. Table 7.7 summarizes the classification accuracy for different supervoxel sizes. Our experiments clearly demonstrate that an empirically optimal tradeoff between accuracy and homogenous samples is achieved at a number of 1800 – 2200 voxels per supervoxel.

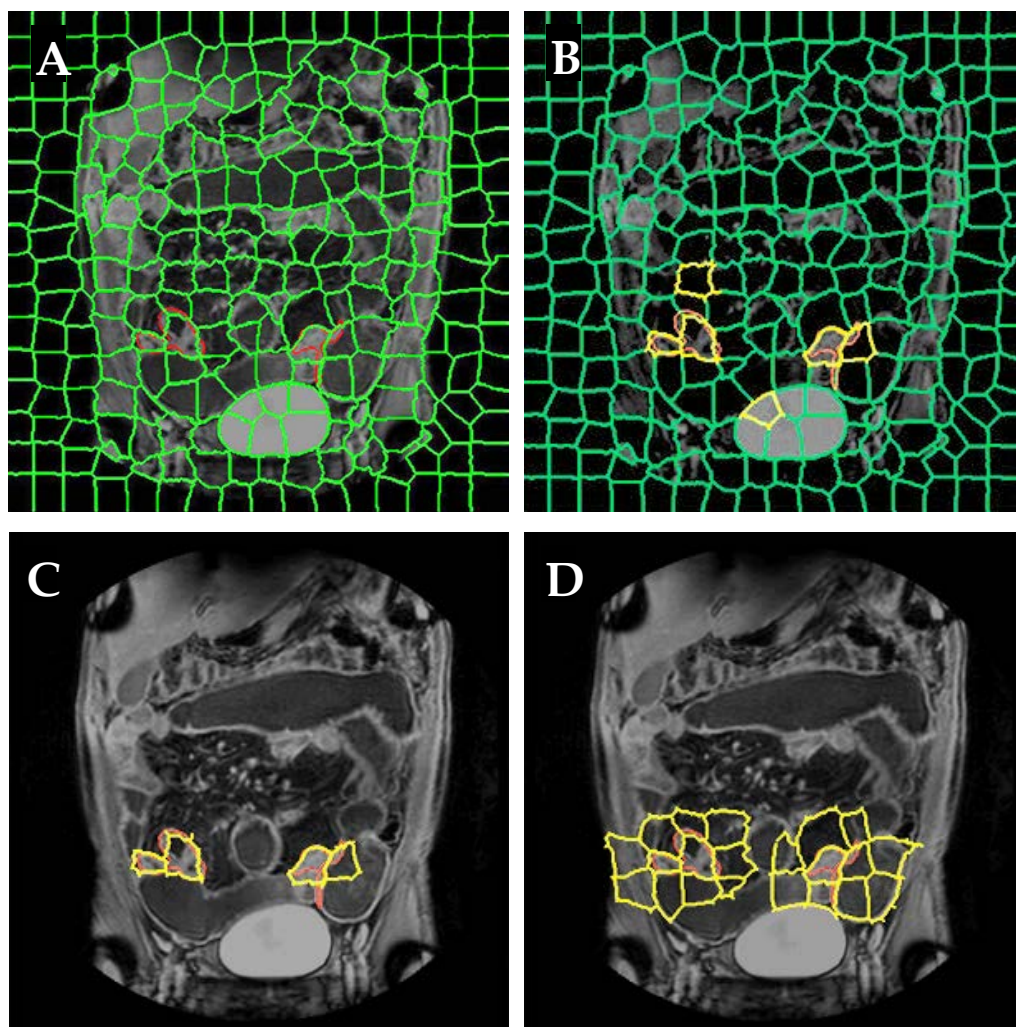


Figure 7.12: VOI detection with SLIC supervoxels on one example slice. A: The MRI image is tessellated with supervoxels. The manual annotated diseased regions are outlined in red. B: Single supervoxels are classified as to be “diseased” with the stage 1 classifier based on intensity, texture and curvature features. Detected diseased supervoxels are drawn yellow. C: Detected supervoxels are filtered: the largest cluster or clusters with more than one connected supervoxels are selected for further processing. D: To ensure that the whole diseased region is covered, all neighboring supervoxels are added to the detected ones to form the final VOI.

Table 7.7: Classification accuracy for supervoxels of varying size ( $N$ , number of voxels per supervoxel).

$N$	500- 1000	1000- 1500	1500- 1800	1800- 2200	2200- 2500	2500- 3000
Acc	77.9 $\pm$ 1.6	81.1 $\pm$ 2.4	83.5 $\pm$ 2.7	90.3 $\pm$ 2.9	83.3 $\pm$ 1.4	79.4 $\pm$ 3.3

## 8 A MODEL DEVELOPMENT PIPELINE FOR CROHN'S DISEASE SEVERITY ASSESSMENT

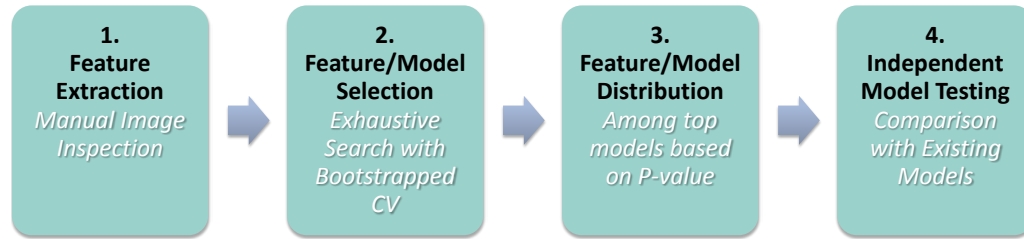
In the previous chapter, we studied a two-stage machine learning pipeline for the automated CD detection and segmentation in MRI scans. A further scientific problem in medical imaging is the development of a *CD severity* representation in MRI. CD occurs as a chronic disease and its severity influences the therapy strategy. The state-of-the-art CD severity measurement is the *CDEIS* derived by a full ileo-colonoscopy (Baumgart *et al.* 2012). The aim of this study is therefore *to find a model consisting of available MRI features which significantly correlates to the CDEIS*. To tackle this problem, we will present in this chapter an exhaustive search pipeline proposal for elaborate feature selection (Schüffler *et al.* 2013c). This pipeline is unspecific, and we showed its generic application to various kinds of feature selection problems, such as the discovery of a serum biomarker for the diagnosis and prognosis of prostate cancer (Cima *et al.* 2011) or the selection of diagnostic markers for lymphoma (Brandt *et al.* 2013; Soldini *et al.* 2013). Thereafter in chapter 9, we introduce computational non-standard image features tailored to CD severity measurement which have been developed in the scope of the VIGOR++ project. These significantly improve the CDEIS correlation.

### 8.1 *CD Analysis Pipeline with Manually Read Features*

For the discovery of highly predictive MRI models with high correlation to the endoscopic CD severity CDEIS we set up a model development pipeline of four consecutive steps: (1) Feature Extraction, (2) Feature selection, (3) Feature Distribution and (4) Model testing (Figure 8.1).

#### 8.1.1 *Feature Extraction*

Feature extraction as the first step has been performed during dataset generation of the retrospective dataset, as described in chapter 3.3: Four radiologists independently scored 14 MRI-based CD features in 27 CD patients, each patient with 5 bowel segments, yielding a dataset of 488 bowel segments à 14 features.



**Figure 8.1: Model development pipeline (Schüffler *et al.* 2013c) for the discovery of predictive features and models for CD MRI activity correlating CDEIS. Feature extraction (1) has been done manually at this stage as described in section 3.3. The steps (2), (3) and (4) are described in the text.**

### Dataset Division into Training Set and Test Set

Regarding the independent model testing in step (4) in the pipeline, the dataset is randomly divided into training and test set. The training set comprises 332 samples from 18 patients and the independent test set includes 156 samples of 9 randomly selected patients. Note that the separation is performed on patient basis and not on segment basis to account for independency of the sets. The following sections referring to model development consider exclusively the training set, while we will test the models on the test set in section 8.1.8. A further external validation procedure on the prospective dataset is described in section 8.1.8.

#### 8.1.2 Feature Selection and Model Training

Since the task is to find selected features correlating to CDEIS, comparable to the existing MaRIA score, we decide to train linear regression models which are easy to interpret and highly comparable to other studies (Rimola *et al.* 2009; Rimola *et al.* 2011; Steward *et al.* 2012). In an exhaustive search, all feature combinations as potential CDEIS predicting linear regression models are built. Since 14 features are available, in total  $2^{14}-1 = 16383$  models are validated in a 50-fold bootstrapped cross-validation. Algorithm 8.1 outlines the bootstrapped cross-validation procedure.

Thus, every generated model is evaluated with 50 correlation values, each developed from a different subset as training set and test set. The bootstrapped cross-validation with random sampling method provides a realistic mimicry of the heterogeneity in the population, especially for biological data, which are often highly variable by nature.

**Algorithm 8.1: Bootstrapped cross-validation on patient basis.**


---

**Dataset:** Set of patients  $P = \{p_1, \dots, p_n\}$  with features and CDEIS

**Input:** Number of folds  $K = 50$ ;

Features  $F = \{f_1, \dots, f_m\}$ ,

**Output:** Validation of model  $M$  consisting of features  $F$ .

---

```

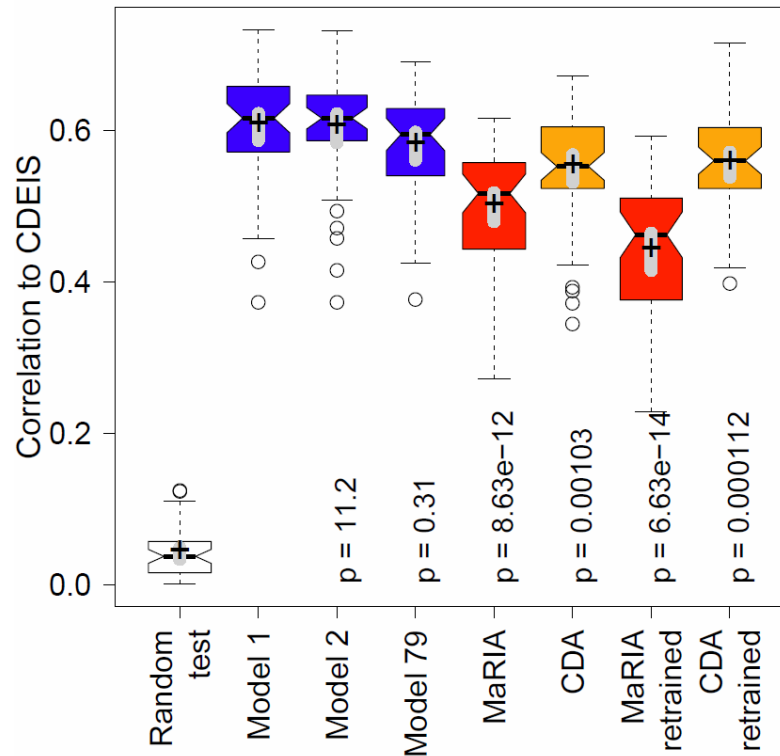
1  bootstrapped_validation_score = [ ];
2  for  $k = 1..K$ 
3    Draw 18 patients randomly with replacement from the 18 patients;
4    # Approximately 12 (67%) patients are drawn during this process;
5    The data of the 6 out-of-bag patients form the fold's test set;
6    Train a linear regression model  $M^*$  with the features  $F$  on the 18
7      patients;
8    Predict the CDEIS of the test set with  $M^*$ ;
9    Append bootstrapped_validation_score by the Person
10     correlation of the CDEIS and the predicted CDEIS;
11  end
12  return bootstrapped_validation_score;

```

---

## 8.1.3 Results

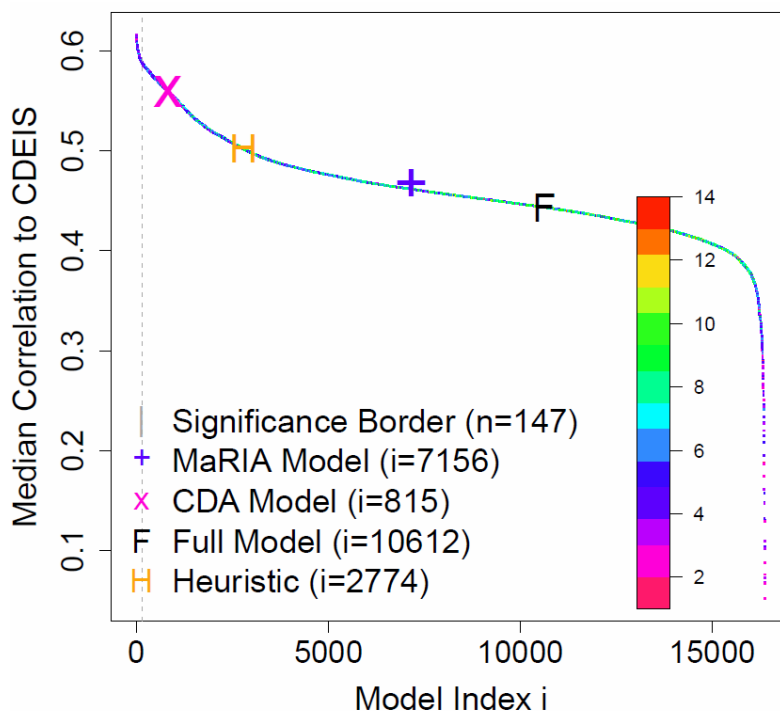
We compare three generated models by their cross-validated correlation to local CDEIS in Figure 8.2. The top three models (blue) show clearly higher CDEIS correlation than the reference models MaRIA (red) and CDA (orange). The best correlating model, *Model 1 (comb\_sign, muralT2)*, has a considerably high median Spearman rank correlation of  $r=.62$ . A random test reveals that shuffling the local CDEIS before cross-validation will destroy the information in the labels and end up in a correlation value of  $r=.04$  (no correlation, white box on the left in Figure 8.2). The second and third blue model represent the model with a particularly small number of features (three features *comb\_sign, muralT2, abscess*), and the model with lowest variance (*abscess, comb\_sign, fistula, muralT2, ulcers*). The reference models MaRIA and CDA are fully parameterized and can be applied on our dataset as they are. Within the same folds as in the cross-validation procedure, the MaRIA score has a median correlation of  $r=.46$  to the local CDEIS and the CDA a correlation of  $r=.56$ . We further consider retraining of their weights on our dataset (“MaRIA retrained” and “CDA retrained”) to test the performance change when the two models would have been newly developed. As expected, the median correlation does not change significantly, and is  $r=.46$  and  $r=.56$  for MaRIA and CDA, respectively. As every cross-validation runs on the same folds, pairwise Student’s t-test is used to test for statistically significant differences of the reference models to our new model 1. The tests have been corrected for multiple testing with the Bonferroni method (Abdi 2007).



**Figure 8.2: Cross-validated correlation to CDEIS of different models.** The blue models 1 and 2 (the best ranked model and the model with only three features *comb\_sign*, *muralT2* and *abscess*) show a superior median correlation to the CDEIS as the MaRIA and CDA. The third blue model (rank 79) uses additionally *fistula* and *ulcers* and has lowest variance of all top models. The models MaRIA (red) and CDA (orange) were either applied as fully parameterized models as reported in literature or retrained on our dataset. The P-values below the boxes indicate the difference to the first blue box (pairwise t-test, Bonferroni corrected for multiple testing).

#### 8.1.4 First Order Statistics for Feature Selection

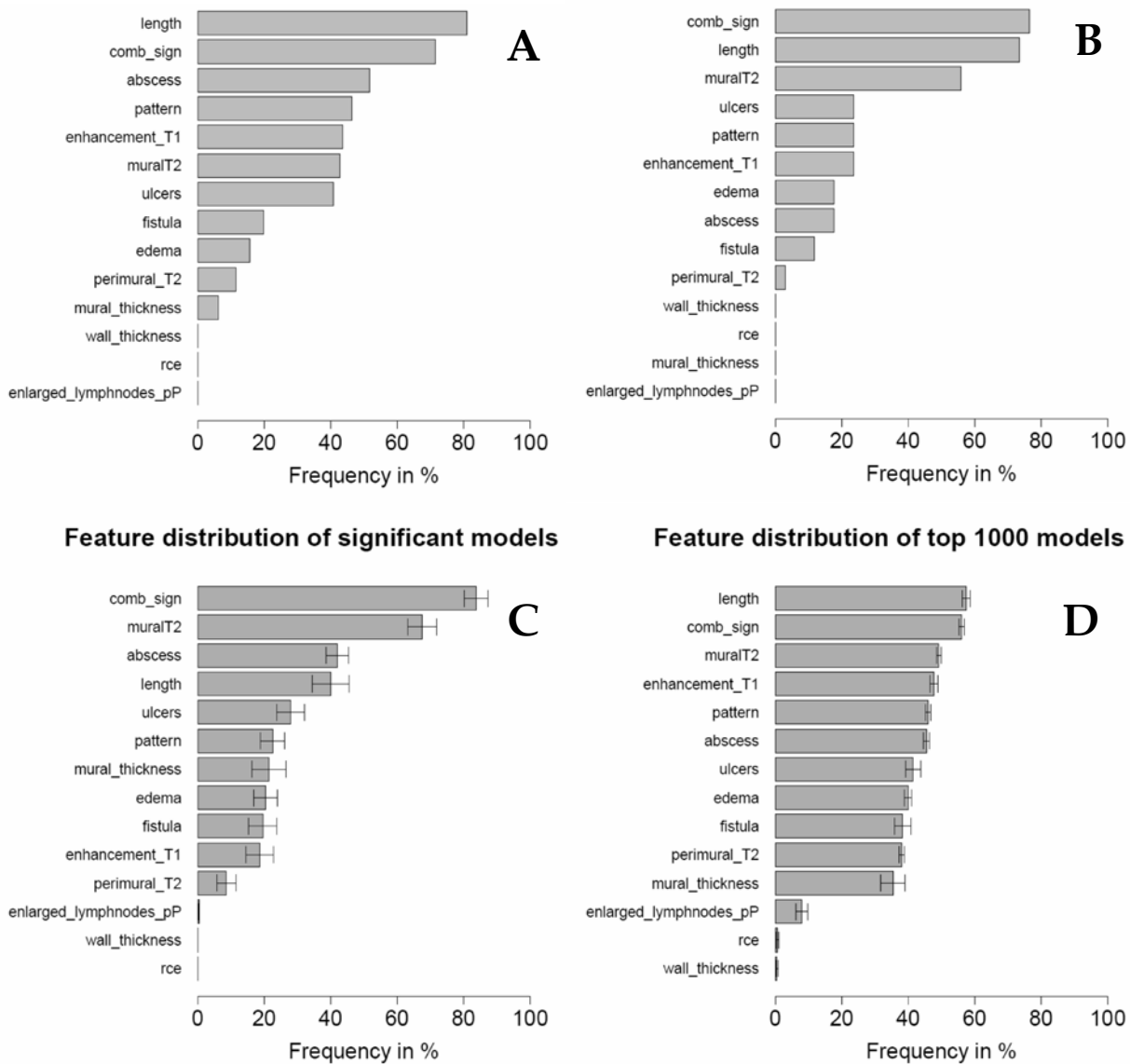
To determine the most relevant features for CDEIS regression, we rank them by their relative presence in the best classifiers. First, all 16383 classifiers are ordered by their median cross-validation performance (Figure 8.3). This ordering follows a sigmoidal shape with best classifier at a correlation level of  $r=0.6$  and poorest models at  $r=0.1$ . Interestingly, the best classifiers are not those with most features. E.g., the full model  $F$  with all 14 features appears on rank 10612. This might be due to the fact that some features are not indicative for CDEIS or have a larger noise levels, thus disturbing the CDEIS regression. Figure 8.3 further depicts the retrained MaRIA (rank 7156) and retrained CDA (rank 815), both distributing in the middle range of all models.



**Figure 8.3:** Ranking of all models according to their median correlation to the CDEIS. The color code indicates the number of features used in the model (1-14). Univariate models appear usually in the tail of the curve. The model with all 14 features (“Full model” F) ranks on the last third. The models with the features of the MaRIA and CDA are on rank 7156 and 815, respectively. A heuristic approach would suggest model H ranking on place 2774 (see section 8.1.6). A gray significance border line indicates a class of 147 models which are not distinguishable by their cross-validation correlation (pairwise t-test P-value  $\geq 0.05$ ).

### Significance Border

Pairwise Student’s t-tests for differences of the top models to *model 1*, Bonferroni-corrected for multiple testing, reveal that the first 147 models have a P-value of equal or larger than 0.05. This indicates that these top models are statistically not significantly different in their performance. On the other hand, they differ in the feature sets they use. In this sense, different classifiers result in the same classification performance and the question about the best classifier to be reported cannot be answered uniquely. To solve this ambiguity, we incorporate a feature ranking among the top models under the assumption that relevant features with high predictive power will result in better classifiers, which we call *first order statistics*. Figure 8.4 (top left) shows the high frequency of the features *length* and *comb\_sign* among the top 147 models, appearing in nearly all of these. Note that every single feature can appear only once in a model.



**Figure 8.4: Feature distribution among top models.** **A:** Almost all 147 models use the features *length* and *comb\_sign*, indicating their importance for CDEIS regression. **B:** A repetition of the model development pipeline. A different feature ranking results from a different random seed, which assigns 18 other patients to the training set for exhaustive search. After ranking, only 34 models are statistically not significantly different. **C:** The average rankings over 10 runs with different random seeds are shown. *comb\_sign* and *muralT2* hold their high ranks. **D:** The first order statistics over the top 1000 models instead of top significant models. Features reaching the 50% level are *length*, *comb\_sign*, *muralT2* and *enhancement\_T1*.

### Stability of Feature Ranking

To judge the stability of our pipeline, we repeat the model selection procedure several times with changed random seed. The seed influences the selection of patients for exhaustive search, thus changing the cross-validated performances. We thereby discover the stability of the model selection pipeline when the data set is altered. Two repetitions are shown



together with a consensus result of 10 repetitions in Figure 8.4. Although slight changes in the rankings are observable, there seems to be a stable high ranking of *comb\_sign* and *muralT2* and a consistent low ranking of *enlarged\_lymphnodes\_pP*, *rce* and *wall\_thickness*. Note that the ranking might especially fluctuate for sparse features such as *fistula* or *perimural\_T2*, since their occurrence in the training data is highly influenced by the subset of training patients.

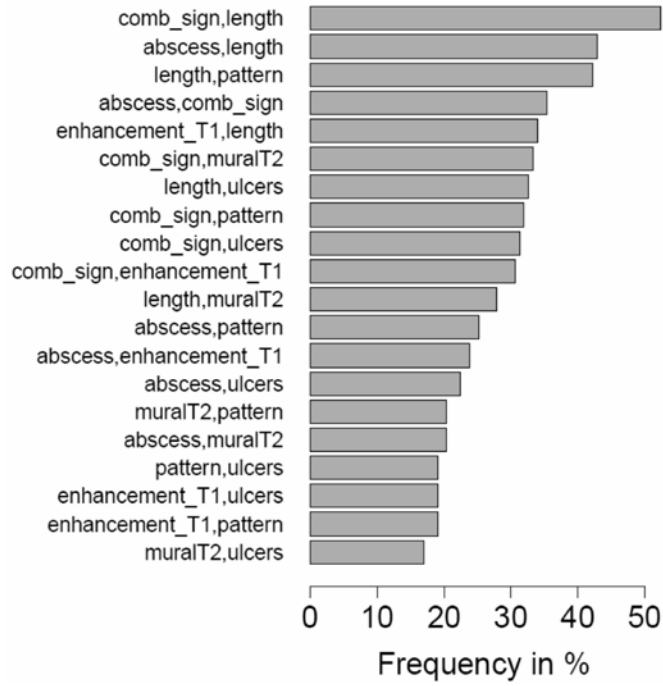
In biomedical feature selection problems similar to the problem at hand, it is often wished to result in a specific feature set which solves the medical task. Although the feature ranking provides a valuable basis for the decision which features should be validated in follow-up experiments, the decision of a final model is still not unique. Note that we cannot use external validation data to select the best performing model there, since the validation data must not be used for model selection to avoid overfitting. To decide for a specific model, two ideas are considered:

- A randomly selected model out of the class of best models can be reported to be the solution of the medical problem, e.g., *model 1* which is top ranked here. Also most heuristic approaches would end up with such a solution, as we will see in section 8.1.6.
- Further constraints given by the biological problem might be considered, as e.g. a specific feature should be excluded or included, the number of features should be small, or the variance of the model should be small. We discuss this idea in section 8.1.7.

### 8.1.5 Second Order Statistics for Feature Selection

While the first order statistics refers to the relative frequency of individual features, the *second order statistics* analyses the distribution of pairwise occurrences. This analysis can reveal dependences between features such as protagonists and antagonists. Analog to the individual distribution of the features in Figure 8.4, the distribution of pairs of co-occurring features in the class of best classifiers is visualized in Figure 8.5.

One interesting question in biomedical feature selection problems is the interaction of features. As known from several biological problems, biological features commonly depend on each other. For example, expression patterns of different proteins can be correlated to each other when proteins lie in the same biological pathway. Referring to the present problem, different Crohn's disease features might occur together or chronologically staggered since they belong to the same disease. A measure for statistical dependency of feature pairs  $F_i$  and  $F_j$  is the probability of co-occurrence  $p(F_i, F_j)$  in a classifier. Stochastically, two features are independent if  $p(F_i, F_j) = p(F_i) * p(F_j)$ .



**Figure 8.5: Frequency of feature pairs in the top 1000 models. Since the features *length*, *comb\_sign* and *pseudopolyps* occur in nearly 100% of the models, also their pairwise combination is prominent in the top class. For visibility, only the top 20 pairs are shown.**

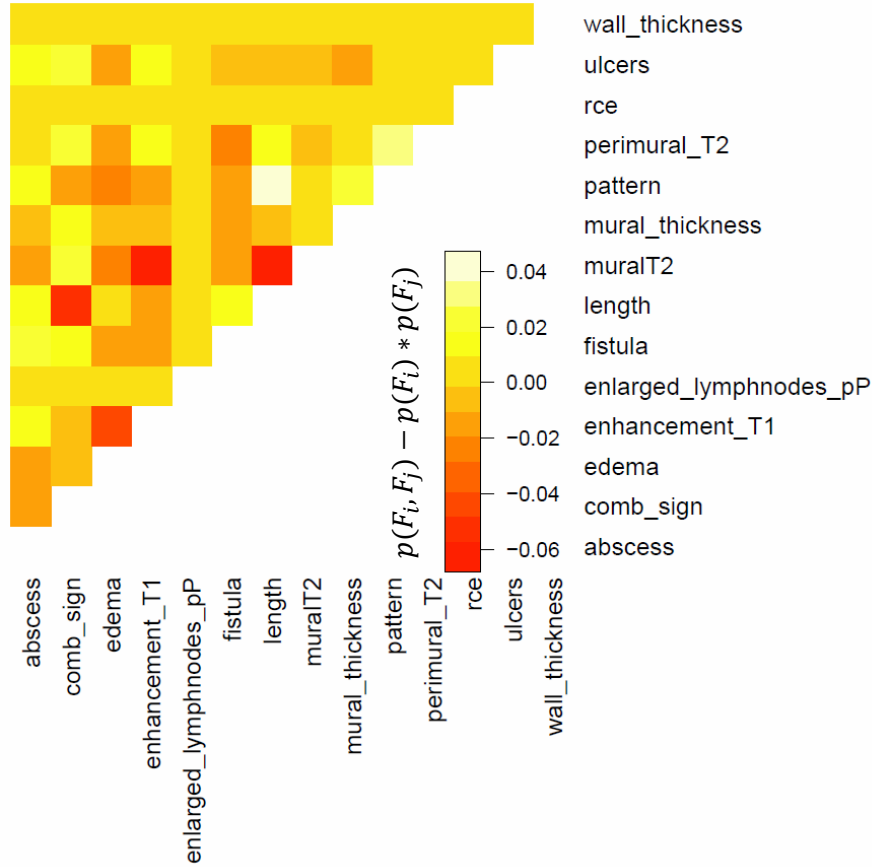
Figure 8.6 displays a matrix with the pairwise statistical dependences of the features  $F_i$ . The color of a feature pair  $(F_i, F_j)$  reflects the value  $v = p(F_i, F_j) - p(F_i) * p(F_j)$ . For independent features,  $v$  is zero, whereas for co-occurring or exclusionary features,  $v$  larger than zero (yellow) or less than zero (red), respectively. This observation might indicate synergistic or antagonistic effects of the corresponding features, respectively. As observed in our dataset, the maximum deviation of observed and expected frequency is  $v = 0.06$  for *muralT2* and *enhancement\_T1*. This means they occur 6% less frequent together as expected. As this is a comparable small number, we do not consider specific synergistic or antagonistic dependencies in our dataset.

A second measure for the pairwise interaction between features is given by the mutual information. Here, features are considered as random variables with two states 0 and 1. Given a classifier, a feature  $F_i$  has the state 0, if it is not used in the classifier (*i.e.* its weight equals 0). Otherwise, its state is 1. With this notation, the mutual information  $MI$  between two features  $F_i$  and  $F_j$  is defined as:

$$MI_{i,j} = \sum_{F_i \in \{0,1\}} \sum_{F_j \in \{0,1\}} p(F_i, F_j) \log \left( \frac{p(F_i F_j)}{p(F_i) p(F_j)} \right) ,$$

where

$p(F_i = 1, F_j = 1)$  is the frequency of models with features  $i$  and  $j$ ,



**Figure 8.6: Statistical independence of all feature pairs in the top 1000 models. The color of a feature pair  $(F_i, F_j)$  encodes  $p(F_i, F_j) - p(F_i) * p(F_j)$  (yellow, positive; red, negative). E.g., the features *pattern* and *length* occur more often together in the classifiers as expected from their individual frequency. On the other hand, the features *muralT1* and *enhancement\_T1* are under-represented according to their individual frequency.**

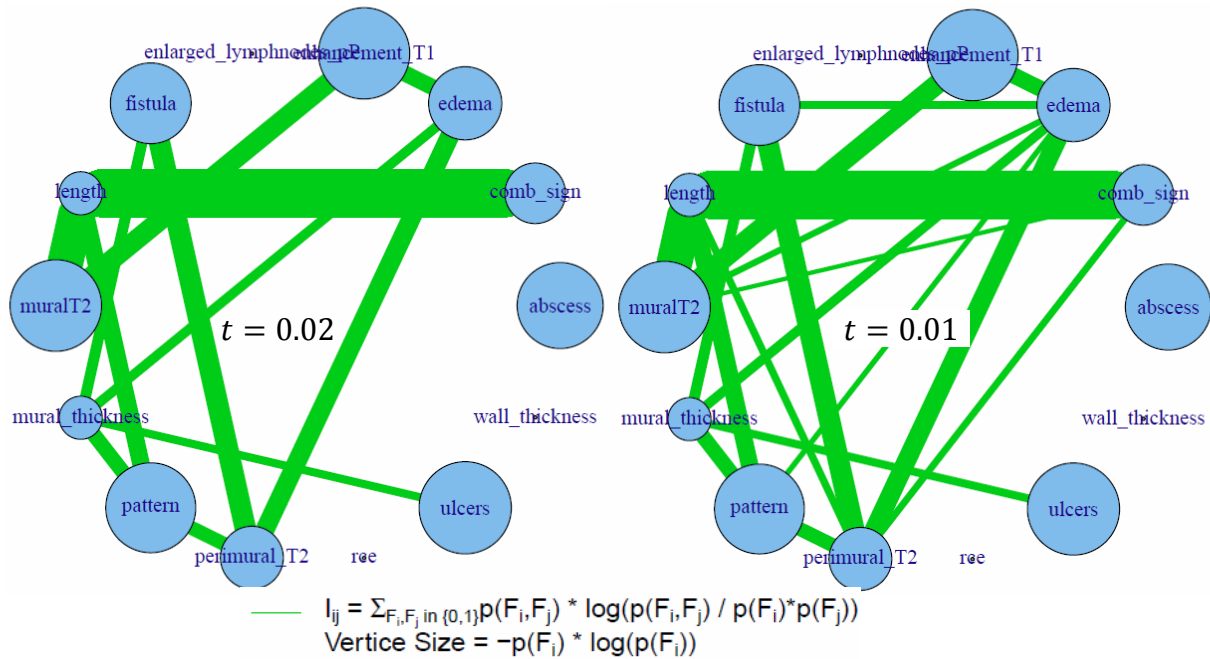
$p(F_i = 1, F_j = 0)$  and  $p(F_i = 0, F_j = 1)$  are the frequencies of models with either feature  $i$  or feature  $j$ , respectively, and

$p(F_i = 0, F_j = 0)$  is the frequency of models without features  $i$  and  $j$ .

$MI_{i,j}$  is normalized by the maximum Shannon entropy  $H_{max}$  of the two corresponding features  $H_{max} = \max(H_i, H_j)$ , where

$$H_i = - \sum_{F_i \in \{0,1\}} \frac{p(F_i)}{\log(p(F_i))}$$

Figure 8.7 shows the mutual information graph of our dataset with two different thresholds  $t=0.02$  and  $t=0.01$  for graph density regularization. In our dataset, the highest mutual information can be found between features *node\_enhancement\_pP* and *pattern* with  $MI = 0.08$ .



**Figure 8.7: Mutual information (MI) between all features in the best 147 classifiers with two thresholds  $t=0.02$  (LEFT) and  $t=0.01$  (RIGHT). Edges with weight  $\geq t$  are shown. The node size is equivalent to  $-p(F_i) * \log(p(F_i))$ . Therefore, nodes of features which are present in all or none of the 147 classifiers are very small, meaning that they cannot carry high mutual information (e.g. conditioned on the knowledge about *rce*, which is never present, we cannot say anything about other features). The largest normalized MI is found between *comb\_sign* and *length* with  $MI=0.12$ . Note that this is not a typical graphical model, as it does not consider *conditional* mutual information.**

### 8.1.6 A Heuristic Approach

In many biological problems, an exhaustive search through all possible models might not be feasible due to the computational overload. A combinatorial search among a large number of features (e.g. gene expression profiles) or a large number of samples (e.g. in micro array based experiments) can be problematic in terms of computational time and memory. Since our problem at hand deals with 14 MRI features for 27 Crohn's disease patients, an exhaustive search is feasible, especially on a high performance computing cluster for parallel computing.

Heuristic approaches which successively approximate an optimal solution of the problem are computationally cheaper as exhaustive search, since they do not calculate all possible solutions. On the other hand, they can result in locally optimal solutions of the problem and do not allow an exhaustive analysis of the solutions as not all solutions are computed. We explore a stepwise backward selection method as a heuristic resulting in one of the optimal solutions calculated by the exhaustive search pipeline.

The *fast backward variable selection* method by Lawless and Singhal (1978) is outlined in Algorithm 8.2. Starting from the full linear regression model with all predictors, predictors are successively excluded whenever the sub-model fits the data better as the larger model. This procedure is repeated until no improvement is achieved. We use the implementation in R (v3.0.1) package “rms” (Team 2008). Stepwise ordinary linear regression selects the “heuristic model”  $H$  ranked on position 2774 of the exhaustive search ranking (see Figure 8.3). The model consists of *muralT2*, *comb\_sign*, *length*, *rce* and *pattern*, of which the first three are among the top 4 features in the first order statistics in section 8.1.4.

**Algorithm 8.2: Principal algorithm of stepwise backward selection.**

---

**Input:** Set of patients  $P = \{p_1, \dots, p_n\}$  with features  $F = \{f_1, \dots, f_m\}$  and CDEIS

**Output:** Parameterized model  $M$  consisting of features  $F^* \subseteq F$ .

---

```
1  Train model  $M$  on all features  $F$ ;  
2   $model\_improved = true$ ;  
3  while  $model\_improved$   
4    for  $F_i \in M$   
5      Train  $M_i^*$  as  $M \setminus \{F_i\}$ ;  
6    end  
7    Select  $M_{max}^*$  which improves  $M$  best;  
8    if  $M_{max}^* \neq \emptyset$   
9       $M = M_{max}^*$ ;  
10    $model\_improved = true$ ;  
11  else  
12    $model\_improved = false$ ;  
13  end  
14 end  
15 return  $M$ ;
```

---

### 8.1.7 Biological Constraints for Model Selection

In section 8.1.4, we reported the top 147 models forming a cluster with statistically not distinguishable cross-validated correlations. This set of models can be further reduced by additional medical specifications:

- The final model should preferably consist of a medium number of features (4-9 features). If the number of features is too small, the model might not be robust for new data. If the number of features is too large, the feature extraction might be too costly for daily clinic usage. Note that features such as presence of *pseudopolyps*, *length* of affected disease or *mural thickness* have to be extracted manually by the physician.

- The resulting model should preferably not rely on *rce* (relative contrast enhancement). The reading of this feature is highly time consuming and subjective (see section 3.3.2). Substituting or neglecting *rce* in a model would therefore already be practically useful for daily clinic.

Considering these specifications, one model carries only one feature (*length*, on rank 35), 9 models carry two features (combinations of *length*, *comb\_sign*, *abscess*, *muralT2*, *enhancement\_T1*, *pattern*, *ulcers* or *edema*), 30 models have three features (again combined with additional features), 43 models have four features, 38 models have five features, 21 models have six features, four models have seven features and one model has eight features. All of them belong to the top models and do not rely on *rce* making them clinically interesting for further study.

### 8.1.8 Validation of Models on Two Data Sets

The retrospective dataset has been used for model generation and feature selection. A split of this dataset into 18 training patients and 9 test patients resulted in a specific feature ranking which showed to be unstable considering the selection of training patients. Therefore, this split has been repeated 10 times, each time with 18 randomly chosen training patients. The average feature selection in these runs is stable and shown in Figure 8.4 (bottom). We decided to validate the features with averaged high ranking among the top 1000 models over 10 runs, as it has a minimum number of four features, and employs T1 and T2 MRI sequences. The parameterization of this model trained on our full retrospective dataset is:

$$Model_{avg} = 7.6 * comb\_sign + 2.5 * length + 1.9 * muralT2 + 0.9 * enhancement\_T1$$

We cross-validate the model on the full retrospective dataset, i.e. all 27 patients (leave-one-patient-out cross-validation). For this, the model is re-trained on the data of 26 patients and four observers, and the last patient is predicted. This is repeated for all 27 patients to be predicted. The overall Spearman correlations of the MRI score with CDEIS are  $r=.56$  on a segmental basis and  $r=.36$  on a global per patient score. The global score per patient is the mean of all four or five observed bowel segments, respectively. All correlations are significant ( $p=0$ ) (see Figure 8.8). As comparison, Figure 8.9 shows the correlation of the CDEIS to the MaRIA score. MaRIA is slightly lower on segmental basis than our MRI score. On the other hand, the global MaRIA still outperforms our global MRI score. We will show later in chapter 9 that automatic features can significantly improve the CDEIS correlation.

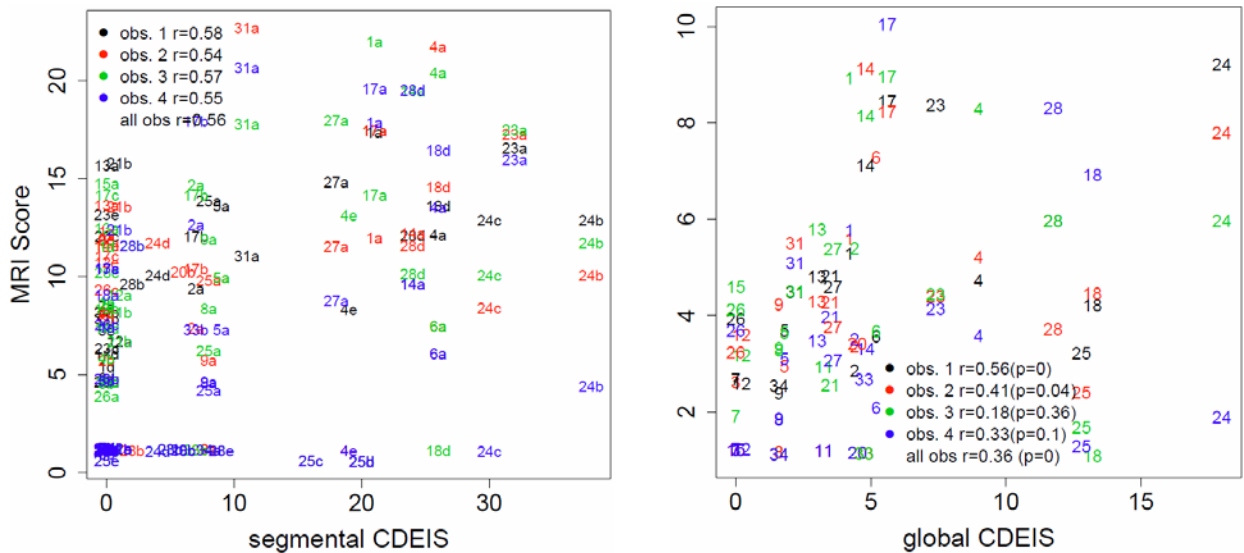


Figure 8.8: Leave-one-patient-out cross-validation for CDEIS regression on MRI in 27 retrospective patients. The average model is trained on 26 patients and the left-out patient is predicted. Each patient is denoted by a number and each observer by a color. LEFT: The segmental predictions are plotted for each of the four observers. Each bowel segment is denoted by a letter (a, TI; b, CA; c, CT; d, CD; e, RE). All correlations are significant ( $p=0$ ). RIGHT: The segmental scores per patient are averaged to the global CDEIS per patient. The global correlation of the CDEIS with the MRI score ranges from  $r=0.18$  (observer 3) to  $r=0.56$  (observer 1).

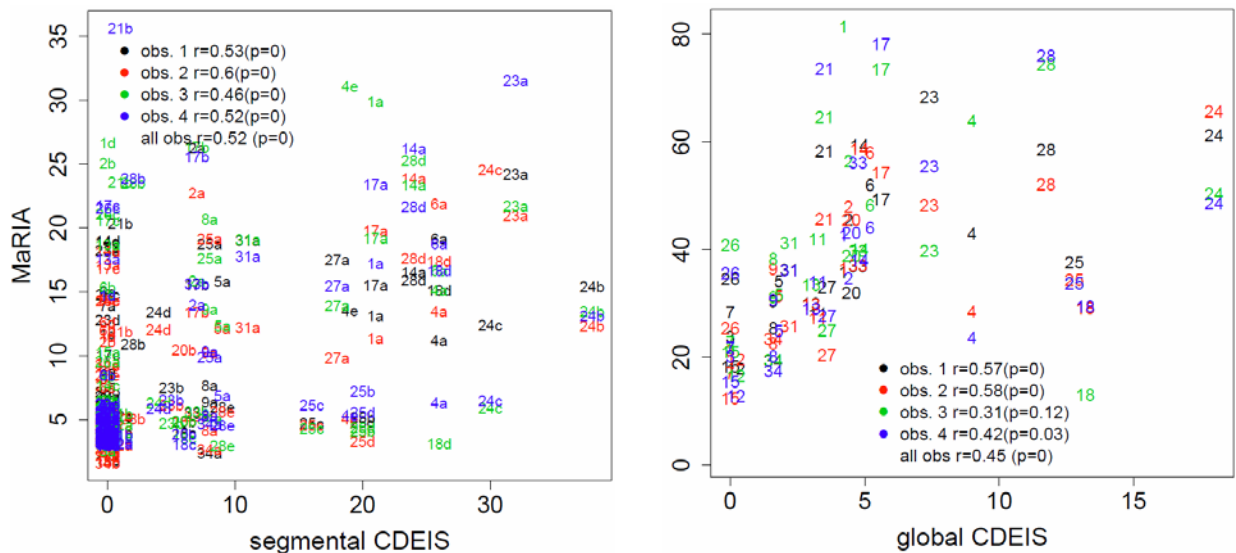
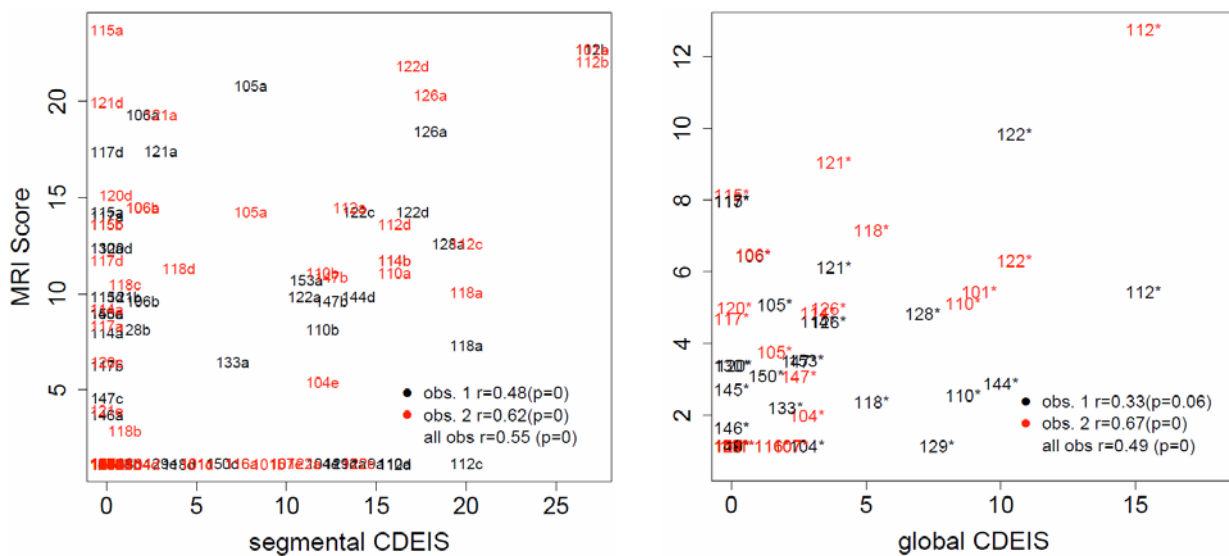


Figure 8.9: Correlation of MaRIA to segmental (LEFT) and global (RIGHT) CDEIS in 27 retrospective patients. The global MaRIA is the sum of the segmental scores. Each observer is denoted by a color, each patient by a number and each bowel segment by a letter (a, TI; b, CA; c, CT; d, CD; e, RE). While the segmental correlation is slightly lower than for our MRI score ( $r=0.52$ ), the global correlation of MaRIA is still larger ( $r=0.45$ ) than the correlation of our MRI score. Nevertheless, our MRI score will be significantly improved by the new automatic features as shown in chapter 9.

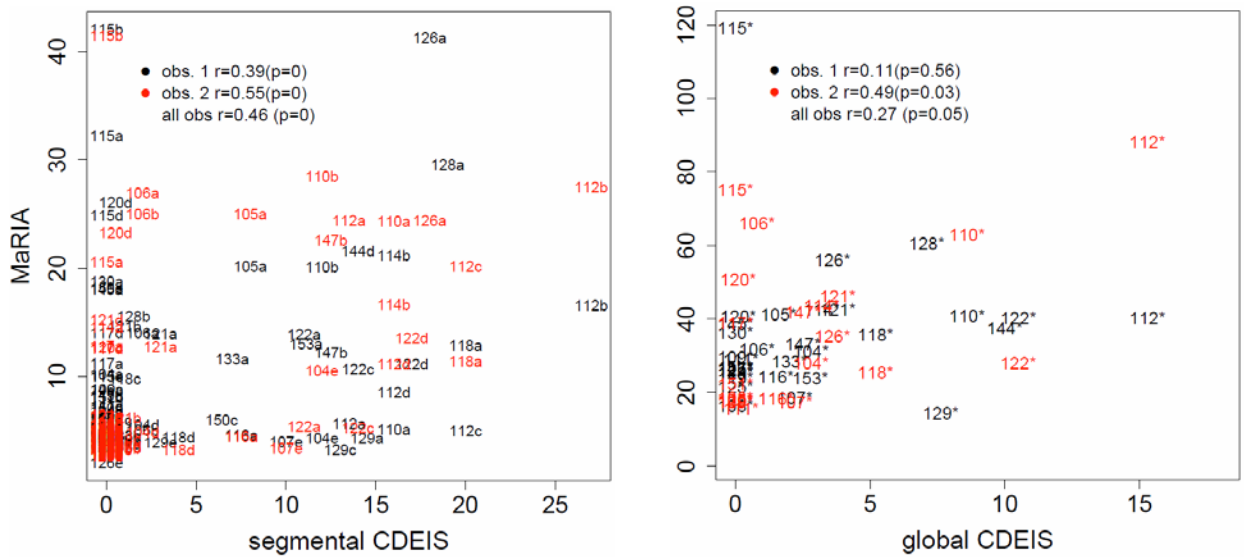
As we also want to evaluate the performance of our algorithms on data from different patient cohorts, a second validation is employed as the prospective dataset. This dataset is completely independent from the retrospective set, as it is collected on a different hospital with different patients and MRI scanners.

We report the tested correlations to CDEIS in Figure 8.10 for our MRI model and in 11 for the MaRIA score. Our model clearly outperforms MaRIA as well on segmental basis ( $r=.55$  vs.  $r=.46$ ) as on global basis ( $r=.49$  vs.  $r=.27$ ). MaRIA makes use of *rce*, a difficult MRI feature which is prone to error or outliers. E.g. patient 115 (11 RIGHT, top left corner) has a very high MaRIA score due a high *rce* value. Our score does not incorporate *rce*. Further, both scores show a high inter-observer variance which might arise from the difference of time spending the two observers annotating.



**Figure 8.10: Spearman correlation of MRI score to CDEIS on 35 prospective test patients. The patients are numbered starting from 101. The two observers are colored in red and black. LEFT: Segmental CDEIS correlation is  $r=.48$  or  $r=.62$  for the two observers. Each segment is denoted by a letter (a, TI; b, CA; c, CT; d, CD; e, RE). RIGHT: The global CDEIS and MRI score are shown. The stars (\*) indicate global CDEIS. The correlations are  $r=.33$  or  $r=.67$  for both observers.**





**Figure 8.11: Spearman correlation of MaRIA to CDEIS on 35 prospective test patients. The patients are numbered starting from 101. The two observers are colored in red and black. LEFT: Segmental CDEIS correlation is  $r=0.39$  or  $r=0.55$  for the two observers. Each segment is denoted by a letter (a, TI; b, CA; c, CT; d, CD; e, RE). RIGHT: The global CDEIS and MaRIA are shown. The stars (\*) indicate global CDEIS. The correlations are  $r=0.11$  or  $r=0.49$  for both observers.**



## 9 AUTOMATED FEATURE EXTRACTION METHODS IMPROVE CD SEVERITY ASSESSMENT

To automate CD severity assessment based on MRI, the VIGOR++ project aims to develop CD related features that can be computationally assessed in a preferably automated manner. Once such features were established, the whole CD MRI assessment pipeline would benefit threefold: (1) the readings of MRI scans might be considerable more objective across different clinics and medical doctors, (2) the readouts of MRI scans would be reproducible and documented and (3) given a fully automated approach, MRI scans could be processed overnight much faster and in a high throughput manner. Two feature types for automated readout are imaginable:

1. *Unspecific* or standard image features such as signal intensity, texture, morphological characteristics or similar features.
2. *CD specific* image features that mimic the readouts of trained radiologists, such as wall thickness, relative contrast enhancement or similar features.

Note that the group of specific features might be deducted as to be a special case of the first group. The difference of the groups lies more the interpretation of the features and their medical relation. For the proposed CD severity assessment model in chapter 8, we exclusively used manual CD specific features for the following reasons:

- There is a lack of whole bowel segmentation in MRI. Since the bowel is not as rigid as e.g. a kidney, the localization and segmentation of the bowel and its diseased sites is considerably more difficult.
- There is no colon partition it is known for the CDEIS (five segments *terminal ileum, ascend colon, transverse colon, descend and sigmoid colon* and *rectum*). This partitioning would be needed for segment wise CDEIS prediction. Implicitly, the medical experts did this segmentation in the manual read-outs by eye.
- Highly diseased areas in the MRI scans annotated by medical domain experts have not shown to be mutually related to CDEIS (see 9.1). We can learn from this experiment that sparse labeling it is not sufficient for severity assessment.

Nevertheless, research is ongoing to use standard image features for localization of highly diseased areas in MRI. We showed in chapter 7 on 26 patients that intensity, texture, curvature and context information might be sufficient to detect highly diseased areas in MRI qualitatively: 85.8% of the diseased area could be detected.

*Differences to  
chapter 8*

Note that in contrast to chapter 8, we use 26 patients in this study (those with manual disease segmentation), as well as Pearson correlation as accuracy measure (instead of Spearman). Further, we do not use *edema* and *mural\_thickness* for model development. Therefore, the correlations, rankings and top models might slightly change compared to chapter 8.

## 9.1 Low Mutual Information between MRI and CDEIS

The relation between unspecific standard image features of MRI and endoscopic CDEIS are studied in this section. 26 retrospective 3D-MRI scans CD patients (post contrast sequence) of a size of approximately  $400 \times 400 \times 100$  vx per scan have manually been segmented by an expert radiologist. Visible areas of enhanced bowel signal intensity related to CD have been outlined extensively in the five bowel segments *terminal ileum*, *ascend colon*, *transverse colon*, *descend and sigmoid colon* and *rectum*. Figure 9.1 shows four typical example images with or without found evidence of disease in two patients.

The same patients have undergone endoscopic colonoscopy to assess the segment wise CDEIS by an independent medical doctor. The match between the MRI dataset and the CDEIS dataset covers 117 samples. Two patients do not have a visible ascend colon (patients 12 and 14) and five patients do not have a visible rectum (patients 13, 15, 18, 26 and 27) (e.g. due to resection). Further, the terminal ileum of 6 patients is not accessible for colonoscopy (patients 11, 20, 21, 24, 28 and 33) (e.g. due to stenosis). 16 samples have both a positive CDEIS and a MRI annotation, 17 samples have a positive CDEIS but no sign of disease in MRI, 10 samples have no endoscopic sign of disease but enhancement in MRI and 74 samples show neither endoscopic nor radiologic evidence of disease (see Table 9.1). The mutual information  $MI$  between CDEIS and MRI defined as

$$MI_{CDEIS,MRI} = \sum_{CDEIS \in \{>0,=0\}} \sum_{MRI \in \{+,-\}} p(CDEIS, MRI) \log \left( \frac{p(CDEIS, MRI)}{p(CDEIS)p(MRI)} \right)$$

equals 10% non-normalized and 12% normalized by the maximum entropy. This indicates we can learn 12% of the MRI enhancement information given CDEIS or vice versa, which is less than expected. The CDEIS should therefore not be replaced by MRI enhancement solely.

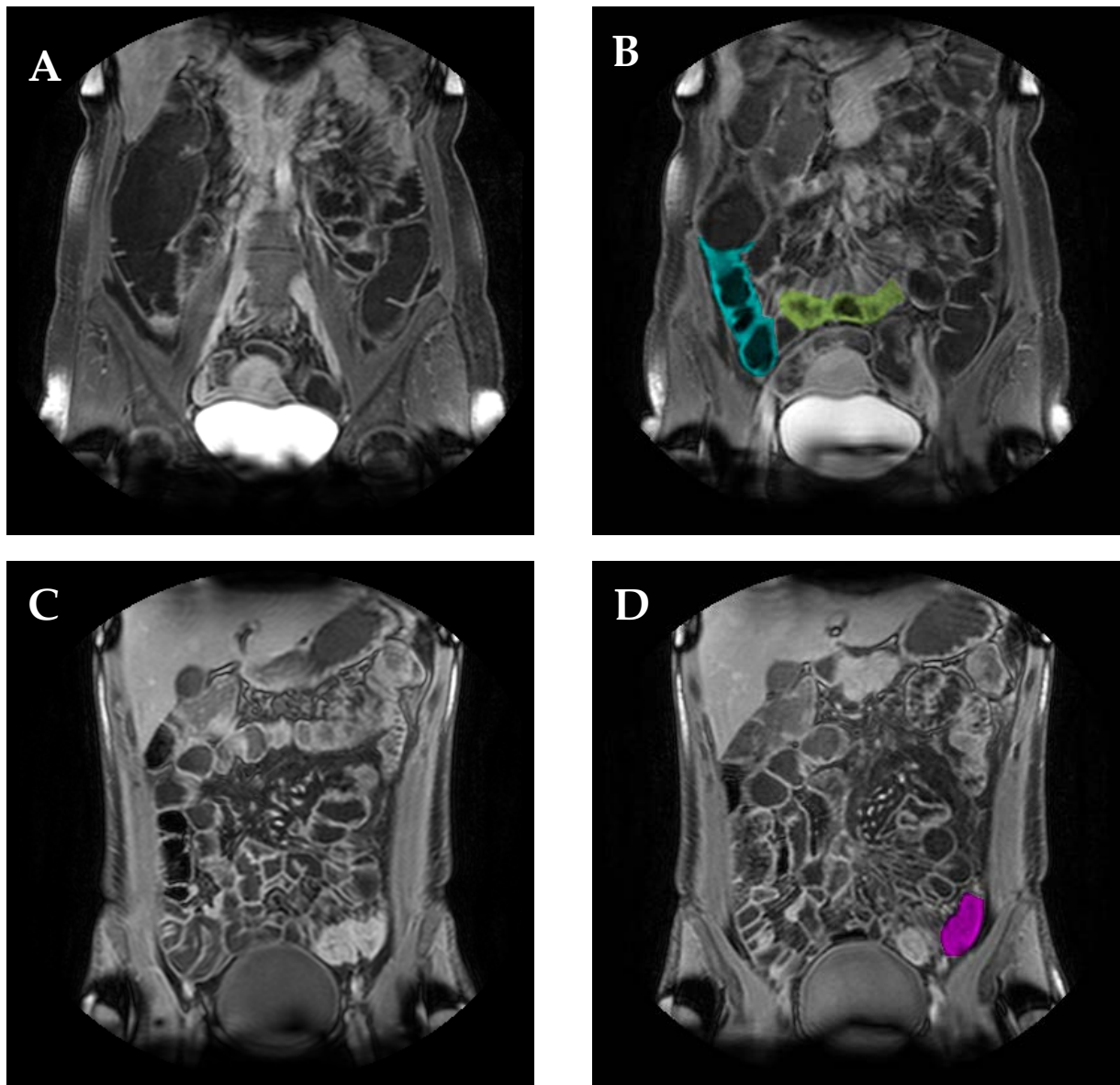


Figure 9.1: Four example images of MRI scans of two patients with and without radiologic signs of disease. **A:** Patient 4, 3D-slice no. 44. **B:** Slice no. 33. **C:** Patient 18, slice no. 21. **D:** Slice no. 26. 100 such slices exist in total for each patient. The local CDEIS for these patients are: Patient 4 (top): 26/0/0/0/19 (terminal ileum (TI) / ascend colon (AC) / transverse colon / descend and sigmoid colon (DC) / rectum); Patient 18 (bottom): 0/0/6/26/34. Drawn are manual CD annotations: green, TI; blue, AC; violet, DC.

Table 9.1: Contingency table for endoscopic disease severity (CDEIS) and MRI signal enhancement. 117 bowel segments have independently been labeled for both by a medical doctor and radiologist, respectively. 24% of the samples show either endoscopic or radiologic signs for disease, while 76% agree on the two assessment methods.

117 samples	MRI +	MRI -	
CDEIS > 0	16 (14%)	17 (15%)	33 (29%)
CDEIS = 0	10 (9%)	74 (62%)	84 (71%)
	26 (23%)	91 (77%)	

## 9.2 Automated Bowel Wall Thickness (ABWT)

*Acknowledgement*  
I want to thank  
Robiel Naziroglu for  
our pleasant  
VIGOR++ collabora-  
tion. He patiently de-  
livered all ABWT  
measurements for  
this work.

Bowel wall thickness is a CD related feature visible in MRI. It is represented categorical (*mural\_thickness*, 0: 1-3 mm, 1: 3-5 mm, 2: 5-7 mm, 3: >7 mm) and numerical (*wall\_thickness*, in mm) in our manually annotated MRI datasets (see sections 3.3 and 3.4). A separate part of the VIGOR++ project is the development of automatic methods for bowel wall thickness (ABWT) measurement in MRI for reproducible, fast and accurate feature extraction. The development of such features goes beyond the scope of this thesis and is not explained here. Rather, we study in this section the quantitative improvement of CD severity assessment for computationally measured bowel wall thickness.

One method to quantify the bowel wall thickness in MRI is to take the difference between outer bowel wall segmentation and inner bowel wall segmentation. This method is based on regions of interest (ROI), indicating the centerline of a bowel section. Within the ROI and starting from the inner lumen of the bowel, the inner wall and the outer wall are segmented with intensity gradient based methods. The wall thickness of the ROI is then represented by either the minimal thickness in the ROI (*ABWT.min*), the maximum thickness found in the ROI (*ABWT.max*) or the averaged thickness of the ROI (*ABWT.mean*).

### 9.2.1 Retrospective Dataset Expansion

For CDEIS regression with automated wall thickness measurement, min, max and mean ABWT have been measured on all bowel segments with endoscopic signs of disease on 26 retrospective patients. For bowel segments without endoscopic signs of disease, a standard value of 2 mm has been used to account for normal wall thickness to complete the dataset. In total, 16 of 117 bowel segments have been explicitly measured, and 101 bowel segments were considered as normal.

### 9.2.2 ABWT Corresponds to Manual Wall Thickness Scoring

*Mural\_thickness* and *wall\_thickness* both describe the thickness of the recognized bowel wall. Still, they are scored independently and by human eye. On the other side, *ABWT.min*, *ABWT.max* and *ABWT.mean* are *computational* measurements per region of interest.

Figure 9.2 shows the segmental correlation of the two manual and the three automated wall thickness scores. The two manual scores show an almost perfect correlation of  $r=0.92$  ( $p=2.2e-16$ ). On the other hand, the highest achievable correlation of ABWT and manual bowel wall thickness measuring is  $r=0.63$  (*ABWT.max*).

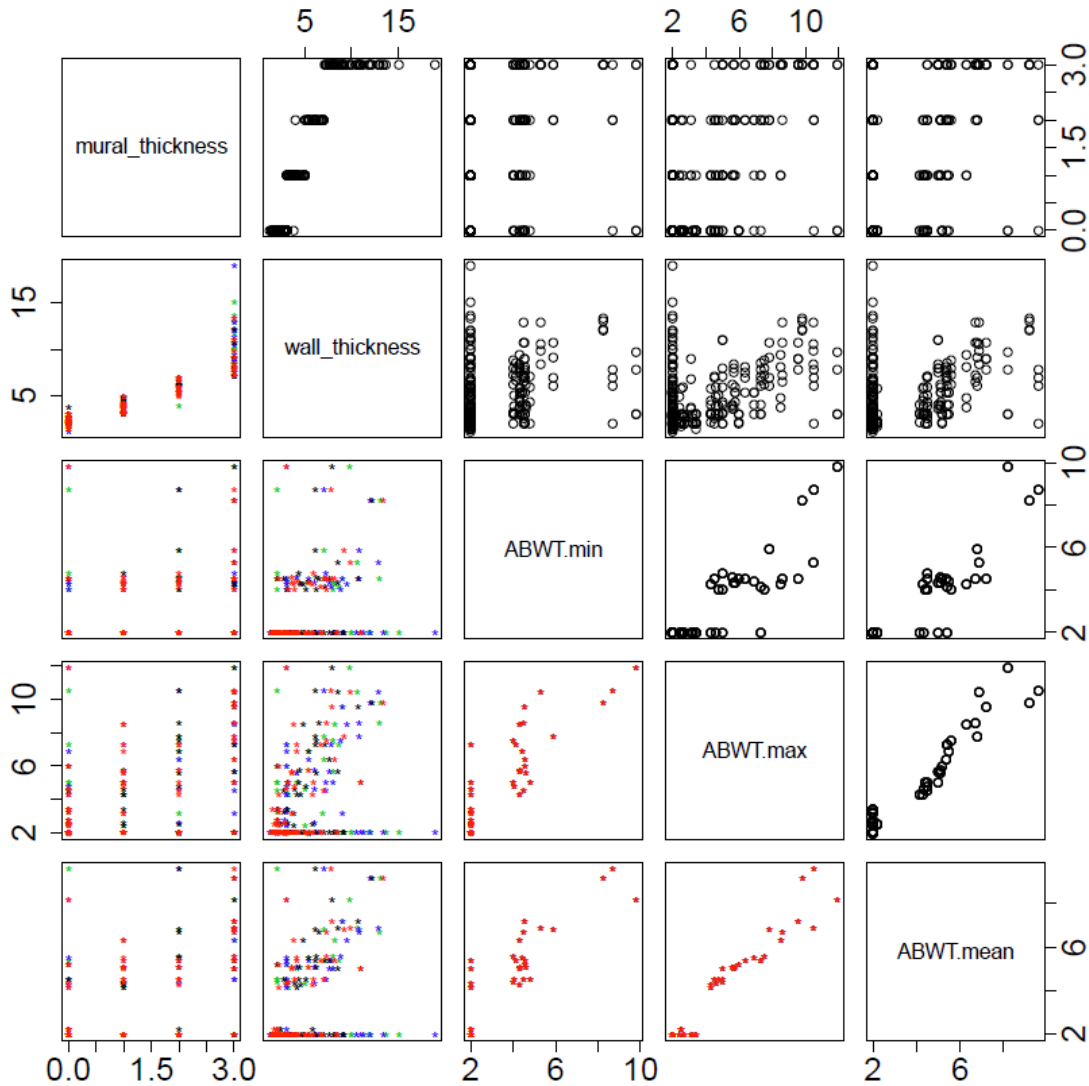
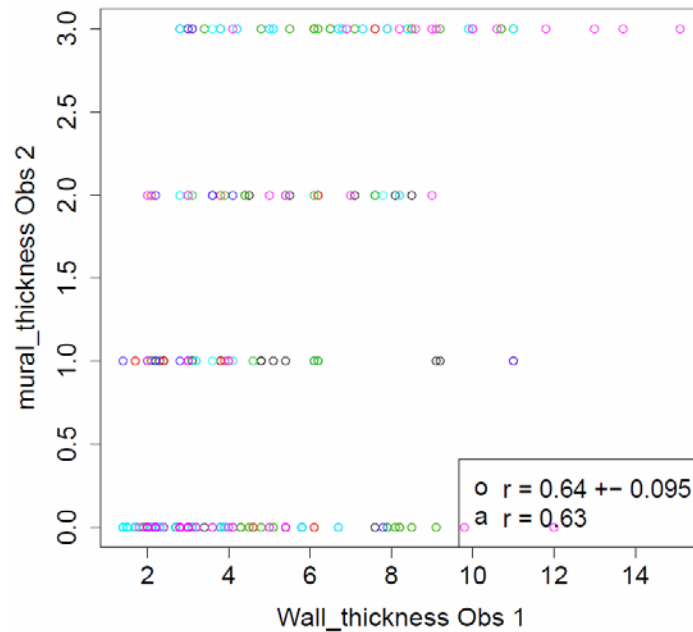


Figure 9.2: Cross-correlation of all bowel wall thickness related MRI features. The two manually scored thickness measures *mural\_thickness* and *wall\_thickness* are highly correlated: their Pearson correlation coefficient is  $r=0.92$  ( $p<2.2e-16$ ). Note that these features are not independent from each other: Shown are only samples which are measured by the same expert (inter-feature, but not inter-expert). The automatically read features range for the correlation to *wall\_thickness* from  $r=0.37$  (*ABWT.min*,  $p=0.02$ ) to  $r=0.63$  (*ABWT.max*,  $p=2.0e-08$ ). Note that 484 ( $121*4$ ) samples have constant ABWT values (2mm). Solely 64 ( $16*4$  observers) bowel segments have a measured ABWT value larger than 2mm. The correlations are calculated for the latter, more interesting part. The bottom left triangle stratifies the manual scores features by the four observers (four colors).

#### Correlation of *wall\_thickness* and *mural\_thickness*

Figure 9.2 implies that the inner-expert correlation of *wall\_thickness* and *mural\_thickness* is almost perfect, compared to a considerably lower relation of the automatic measures to mural thickness. But it has to be noted

that only samples measured by the same human expert are shown: a radiologist discovering a high *wall\_thickness* is also measuring a high *mural\_thickness*. On the other hand, ABWT acts as a new, independent “expert” which is plotted against the human experts. We therefore also compare in Figure 9.3 the *inter-expert* correlations of *wall\_thickness* to *mural\_thickness*, considering the human experts as independent labelers. Then, the correlation of *wall\_thickness* and *mural\_thickness* drops to  $r=0.64$ .



**Figure 9.3: Pairwise correlation of *mural\_thickness* and *wall\_thickness*.** Each pair of human labelers is indicated by a separate color. Six pairs of all four labelers are potted in total. The mean correlation is  $r=.64$ , and the total correlation is  $r=.63$ , similar to the correlation of *ABWT.max* to *mural\_thickness* ( $r=.59$ , data not shown).

#### Inter-Expert and Machine-Expert Variance of *wall\_thickness*

An inter-expert stability comparison of ABWT and *wall\_thickness* is shown in Figure 9.4. For this, the segmental ABWT values are co-related to the manual scorings stratified by observer. Each observer is assigned to a specific color in the plot and each segment is denoted by a circle. Most segments have a normal, unmeasured ABWT of 2 mm and form a vertical cluster in these plots. The interesting part of measured segments shows a considerable high mean correlation to all four observers, with  $r=0.6$  and a particular low standard deviation of 0.06 (*ABWT.max*). This reaches the level of inter-observer discrepancies. In Figure 9.4 top left, the pairwise wall thickness measurements from all observer pairs are plotted. Each pair is assigned to one color (six pairs in total). Interestingly, trained radiologists occasionally disagree in the observed thickness by the factor four. Still, the averaged correlation over all observers is  $r=0.62 \pm 0.1$  on all cases or  $r=0.65 \pm 0.1$  on cases with measured ABWT.



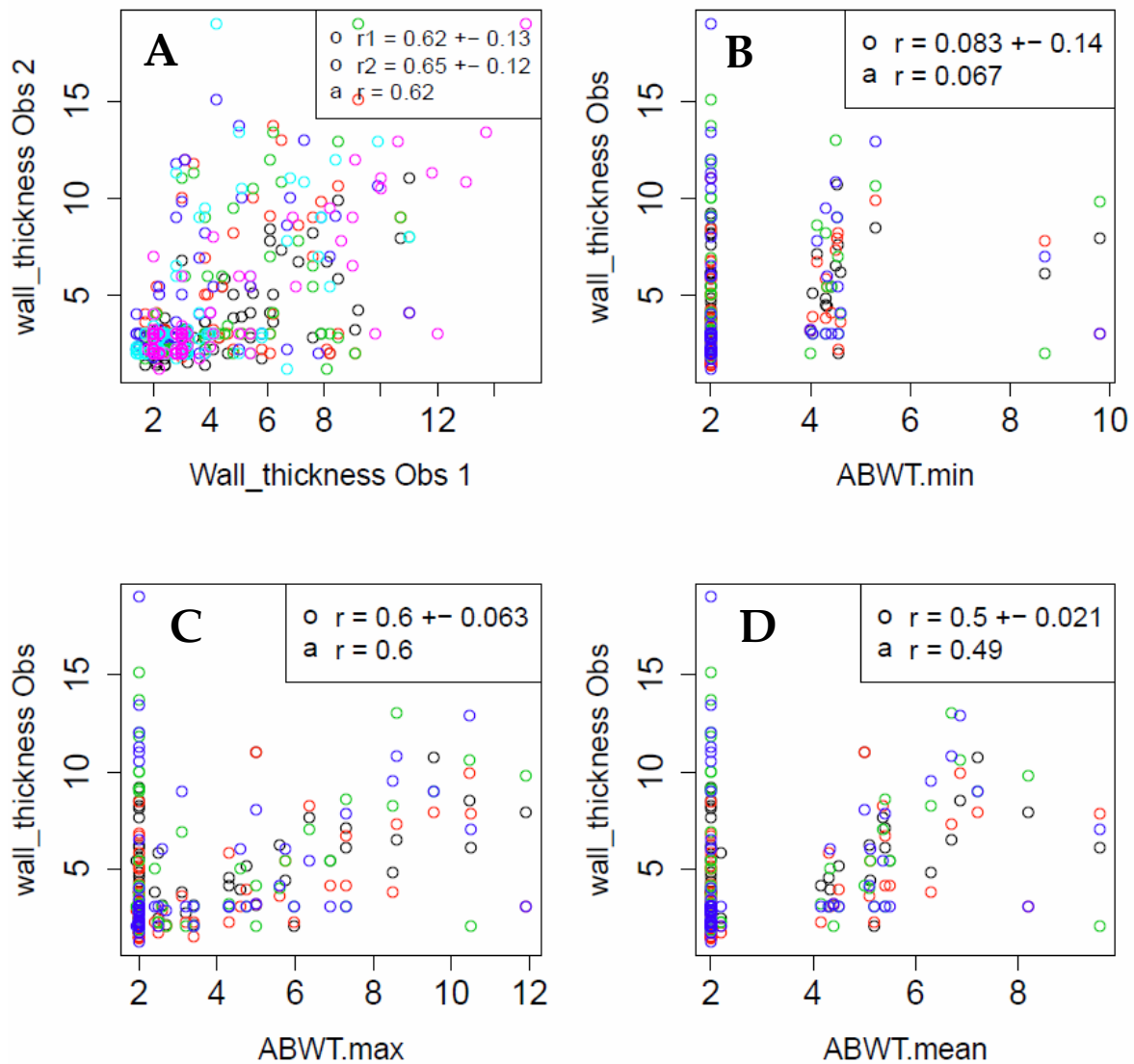
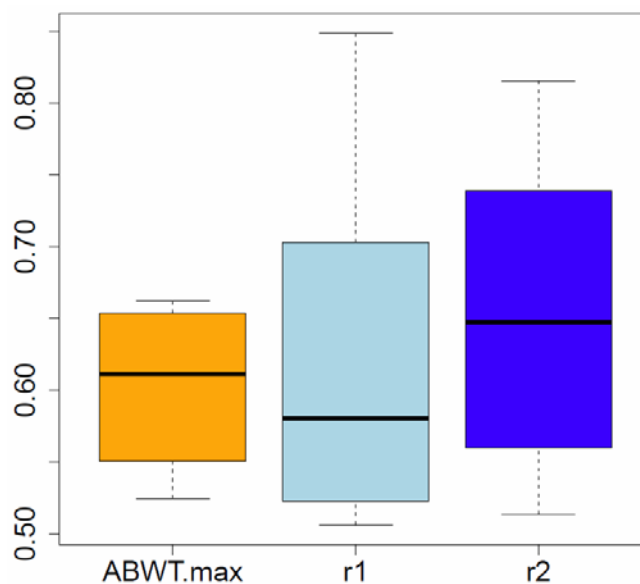


Figure 9.4: Detailed correlation of ABWT and manual readings. **A:** The inter-expert correlation of pairwise *wall\_thickness* correlation is  $r_1=0.62\pm0.1$  (mean and standard deviation of six possible expert pairs), or  $r_2=0.65\pm0.1$  on the subset of samples with measured ABWT or  $r=0.62$  all samples together. Each expert-pair is assigned to one color. **B, C and D:** Detailed view on Figure 9.2, 2<sup>nd</sup> row (black). The automatic features *ABWT.min*, *ABWT.max* and *ABWT.mean* are plotted against the *wall\_thickness* measurement of four observers (= four colors). The correlation is calculated on cases with measured ABWT ( $ABWT>2$ ). As highest value, *ABWT.max* shows a correlation of  $r=0.6\pm0.06$  and reaches range of human experts' discrepancy. o, stratified by observer; a, all observers together.



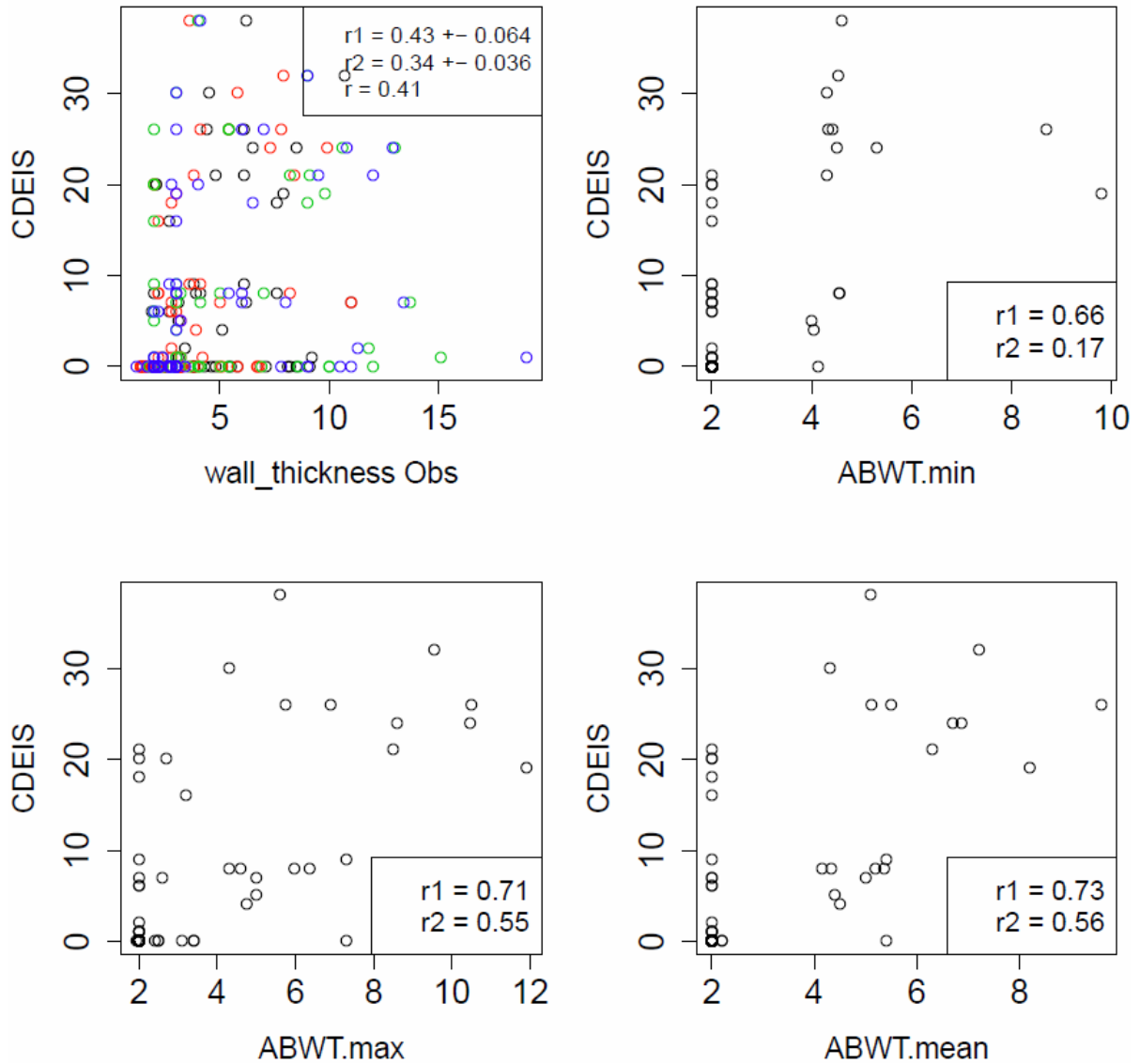
**Figure 9.5: Machine-observer correlation is in the range of inter-observer correlation for wall thickness. LEFT: Median correlation of *ABWT.max* to four observers (Figure 9.4 C). MIDDLE: Median pairwise inter-observer correlation of *wall\_thickness* (from Figure 9.4 A). RIGHT: Median pairwise inter-observer correlation of *wall\_thickness* on same subset as for *ABWT.max* (see Figure 9.4 A).**

### 9.2.3 Does ABWT Increase the Correlation to CDEIS?

To answer the question how far ABWT would influence the CDEIS regression, we first describe the univariate correlation to CDEIS. The maximum achievable correlation is  $r=0.73$  ( $p=2.2e^{-16}$ ) for *ABWT.mean* (see Figure 9.6). As expected, this is considerably higher than the correlation of the manually assessed *wall\_thickness* by four observers ( $r=0.43\pm 0.06$ , Figure 9.6 top left), since ABWT is only assessed on cases with positive CDEIS. *ABWT.mean* still shows superior correlation to CDEIS on cases with raised CDEIS ( $r=0.56$ ). While radiologists assign thickened *wall\_thickness* larger than 3 mm to 53 of 336 normal segments (16%), the machine discovers only 7 thickened segments among 84 normal segments (8%) (Table 9.2). On the other hand, the automatic measurement shows a considerable amount of normal bowel wall on segments with raised CDEIS (12 of 33 segments (36%) with  $CDEIS > 0$  have  $ABWT.max \leq 2$ ). This large “false negative” rate arises mainly from the fact that these segments could not explicitly be measured with the automatic method.

**Table 9.2: Summary of severe and normal segments with (+) or without (-) manual (*wt*) or automatic (ABWT) signs of thickened bowel wall.**

	<i>wt</i> -	<i>wt</i> +	ABWT -	ABWT +
CDEIS -	283 (60%)	53 (11%)	77 (66%)	7 (6%)
CDEIS +	52 (11%)	80 (17%)	12 (10%)	21 (18%)

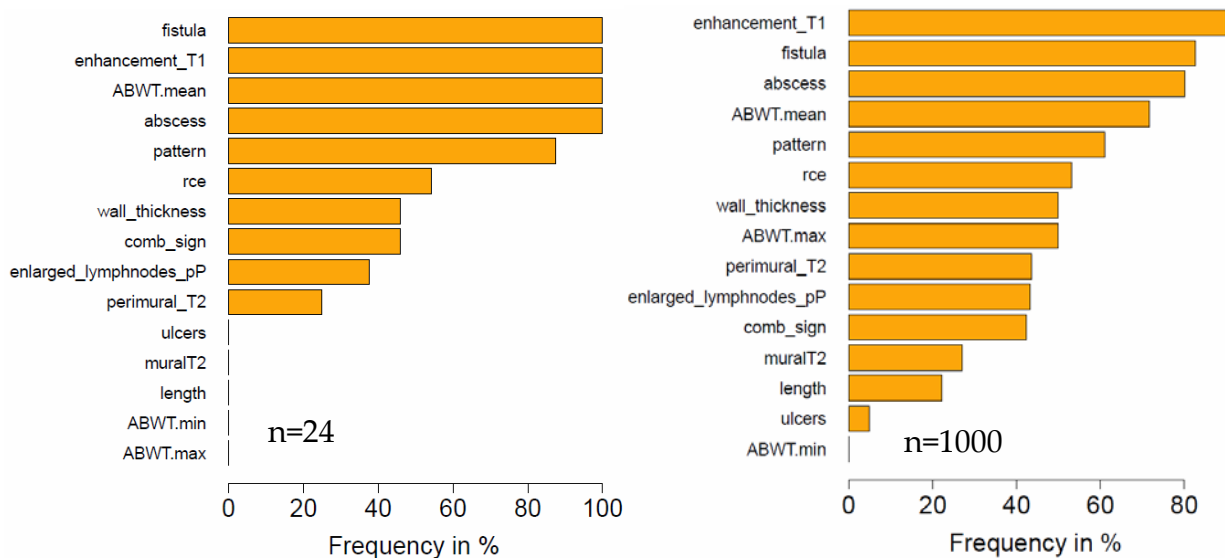


**Figure 9.6: Correlation of ABWT features to CDEIS compared to the observer's scored *wall\_thickness* correlation to CDEIS. The correlations  $r_1$  and  $r_2$  correspond to the Pearson correlation on all 117 samples ( $r_1$ ) or on the subset of samples with measured ABWT ( $r_2$ ), respectively.**

We repeated the CD severity assessment pipeline with and without the additional features *ABWT.min*, *ABWT.max* and *ABWT.mean*. The class of top models with statistically undistinguishable cross-validated correlation to CDEIS comprises 24 models according to our t-test ( $p < 0.05$ ). The predictor distribution among these models clearly suggests high impact of *ABWT.mean* for CDEIS regression (Figure 9.7 left). To support this thesis at a larger scale, the predictor distribution of top 1000 models is shown Figure 9.7 right. The top seven features including *ABWT.mean* are stable in the ranking.

Figure 9.8 shows the significant performance gain in CDEIS correlation when including *ABWT.mean* to a conventional classifier. The basis is the best classifier without *ABWT*, consisting of *comb\_sign*, *length*, *muralT2* and *rce*. When adding *ABWT.mean*, the median correlation to CDEIS significantly increases from  $r=0.65$  ( $p=2.8e^{-14}$ ) to  $r=0.84$  ( $p=0$ ) (Figure 9.8, middle box and right box).

Apparently, *ABWT.mean* is a favorable predictor on the retrospective dataset. The univariate correlation to CDEIS is  $r=0.73$  ( $p=2.2e^{-16}$ ). With the manual features solely, we can achieve a correlation of  $r=0.65$  to CDEIS, which is already lower than the univariate correlation of *ABWT.mean*. Therefore, all models using the new feature will most likely drastically gain in correlation performance. The plot in Figure 9.9 ranks all models of the exhaustive search by their median correlation will therefore show a phase transition when models start to use the new features. A heuristic approach will select a model including an automatic feature.



**Figure 9.7: Feature distribution among the top 24 models (LEFT) and top 1000 models (RIGHT) with ranking of the automated features *ABWT.min*, *ABWT.mean* and *ABWT.mean* in the CDEIS regression pipeline. *ABWT.mean* is used in all 24 top models and 70% of top 1000 models indicating its relative importance.**

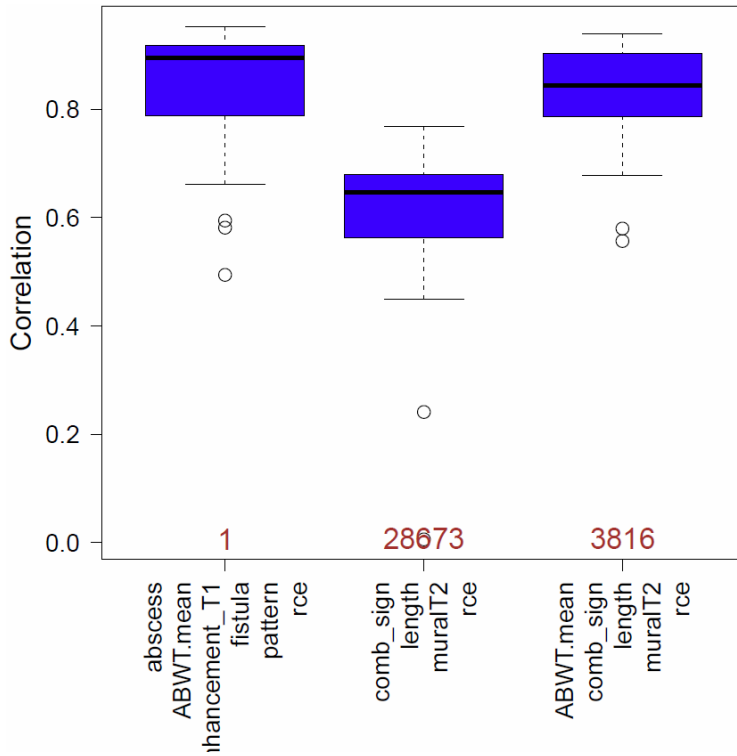


Figure 9.8: Comparison of automated and manual features for CDEIS prediction. The top ranked model 1 (LEFT) includes *ABWT.mean*, and its median Pearson correlation to CDEIS outperforms the best model without automated features (MIDDLE) completely (model 1:  $r=0.89$ ,

$p=0$ , model 28673:  $r=0.65$ ,  $p=2.8e^{-14}$ ). Adding *ABWT.mean* to the manual model (RIGHT) already increases the rank to 3816 ( $r=0.84$ ,  $p=0$ ).

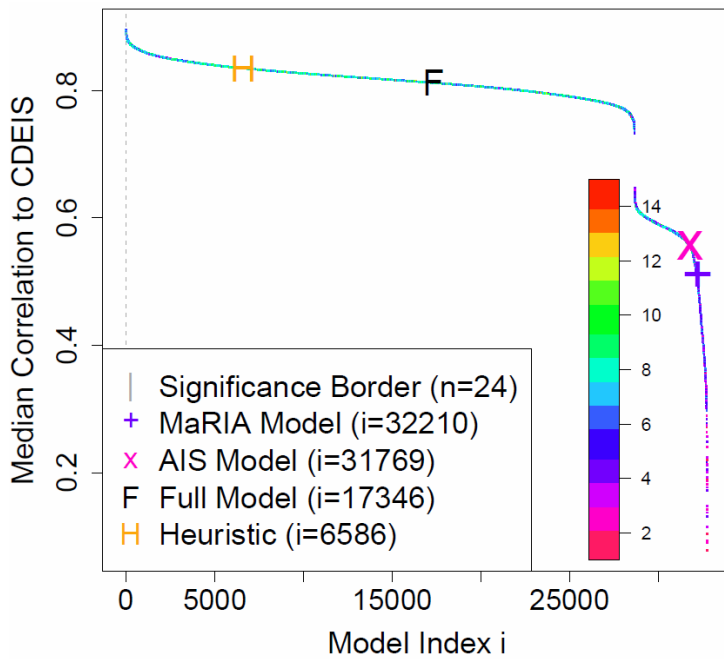


Figure 9.9: Ranking of all models with and without ABWT features as CDEIS predictors. The first model without ABWT is on rank 28673 with a median Pearson correlation to CDEIS of  $r=0.65$ , using *comb\_sign*, *length*, *muralT2* and *rce*. On the neighboring rank 28672, *ABWT.min* joins the manual

features and pushes the correlation to  $r=0.73$  (“phase transition”). Also shown are ranks of the MaRIA model, AIS model, the full model with 15 features and the model found with stepwise selection as heuristic approach. The color codes the number of features used per model.

### 9.3 Automated Dynamic Contrast Enhancement (DCE)

*Acknowledgement*  
Zhang Li developed the DCE features used for CDEIS regression. I want to express my special thanks to him for the provided code and images.

The MRI protocol for radiologic CD examination comprises the uptake of the contrast agent gadobutrol for the dynamic contrast enhancement MRI sequence (DCE-MRI). DCE-MRI runs over six minutes scanning 450 individual recordings while the applied contrast agent distributes in the blood vessels. The temporal resolution of DCE-MRI is 0.82s per scan and the spatial resolution is  $2.78 \times 2.78 \times 2.5$  mm ( $227 \times 227 \times 14$  px) per slice. Gadolinium is commonly used for blood vessel visualization (Lentschig *et al.* 1998). Due to micro lesions in blood vessels and capillaries, the contrast agent accumulates in inflamed areas and metabolizes at different rates. These effects result in differentiable visible body areas over time which are disease-affected and which might clearly be silhouetted against the healthy surrounding tissue. DCE-MRI can therefore be used to identify inflamed and damaged bowel wall. Figure 9.10 illustrates DCE-MRI on one patient. The top left image shows the T<sub>1</sub>-weighted high-resolution isotropic volume examination (THRIVE) before contrast agent application and the top right image the same after contrast agent application. The bottom row shows two DCE-MRI scans before and after contrast agent uptake. Note the different resolution of the two sequences.

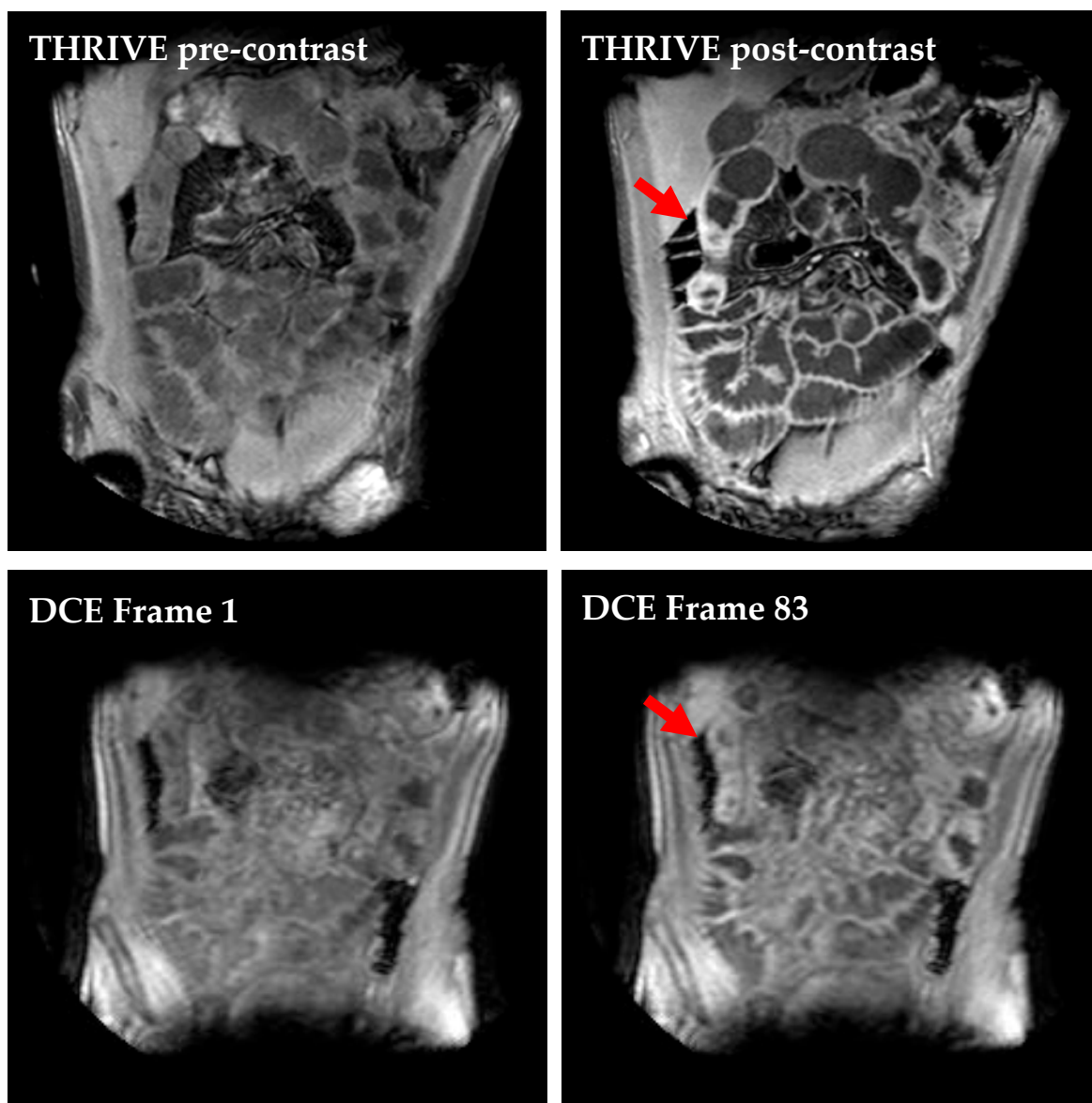
A particularity of abdominal DCE-MRI is the movement of the body caused by respiratory motion. This movement can blur important structures in the image and complicate the registration of subsequent images which is mandatory for automated feature extraction. Breath-hold techniques, e.g. the continuous *volumetric interpolated breath hold examination* (VIBE), and other respiratory motion compensation techniques try to minimize such effects and are a field of current research (Schaffter *et al.* 1999; Lin *et al.* 2008; Yankeelov and Gore 2009).

#### 9.3.1 DCE as Feature

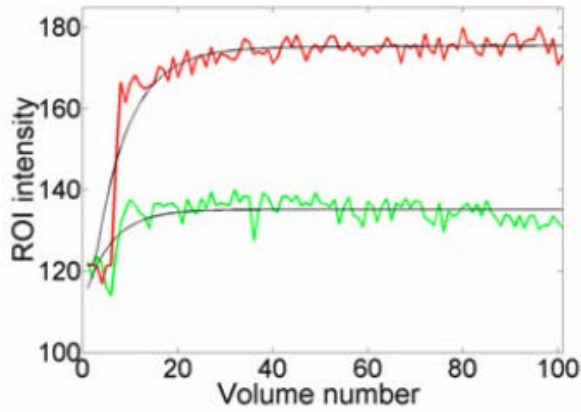
Li *et al.* (2013) developed a new DCE feature extraction method for MRI. First, DCE-MRI scans are registered to post-contrast MRI. The change of intensity over time in a given region of interest (ROI) describes then the *time intensity curve* (TIC). A bi-exponential model  $S(t)$  is fitted to the TIC:

$$S(t) = A_1 e^{-\lambda_1 t} - A_2 e^{-\lambda_2 t}$$

$A_1$  in this model is related to the steepness of the TIC during agent uptake and the steepest ascend reflects the final DCE feature. Figure 9.11 shows two typical TICs for a diseased (red) and a normal (green) ROI, each with raw values and fitted model. The TIC is usually much steeper for diseased ROIs than for normal ROIs resulting in a larger  $A_1$  (Li *et al.* 2013).



**Figure 9.10: Example of DCE-MRI on patient no 3. TOP LEFT: Single 2D slice no 35 of the pre-contrast MRI scan (THRIVE) in our dataset. TOP RIGHT: Corresponding post-contrast THRIVE slice no 33 after DCE-MRI. CD affected regions show a slightly enhanced signal, as indicated by the arrow. BOTTOM: Two registered 2D slices of the DCE sequence of the same patient (time frames 1 and 83). 450 time frames are shot during six minutes. The spatial resolution of DCE-MRI is four times smaller ( $227 \times 227 \times 14$  px) than of THRIVE. During DCE-MRI, a contrast agent is applied to the patient.**



**Figure 9.11:** Illustration of TIC for a healthy bowel segment (green, bottom) and a CD affected bowel segment (red, top). The mean MR signal intensity is measured at a given ROI in all 100 registered DCE frames. At time frame 1, the contrast agent is applied to the patient and the diseased part has a significantly enhanced drug uptake. The black curves represent the fitted bi-exponential models, whose coefficients  $A_1$  serve as DCE features in our dataset.

### 9.3.2 Dataset Expansion

For automatic DCE extraction in our dataset, the manually drawn contours of enhanced bowel wall signals in the post-contrast sequence serve as initial ROIs. The complete data comprise 26 patients or 117 bowel segments. A trained radiologist outlined and segmented all visible major wall enhancements in the post contrast THRIVE sequences of the patients. In 26 bowel segments (22%), strong evidence of wall enhancement is found, while the other segments do not show enhancement of T1 signal. 17 contours are drawn in terminal ileum, 7 contours in the right colon and two contours have been outlined in the left and sigmoid colon. Six patients do not show enhancement in any segment and do not have any contours drawn in the MRI. 11 bowel segments (9%) could successfully be subjected to automatic *DCE* measurement (see Table 9.3). Problematic for 15 segments were poor DCE-MRI registration or mismatch of the field of view of DCE-MRI and the drawn contours. A constant value of zero (no contrast enhancement) has been assigned to normal or unmeasured segments. Similar to the ABWT measures, we estimate the contribution of *DCE* by adding them to the pipeline for model development and feature selection.

**Table 9.3:** 117 bowel segments had 26 ROIs, 11 of which could be used for *DCE* extraction.

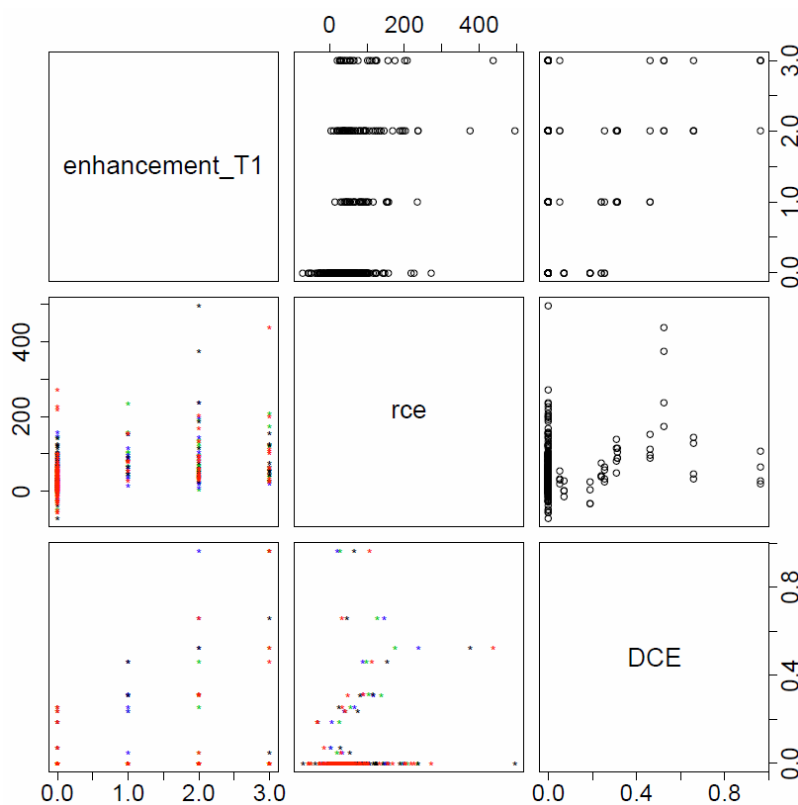
	ROI -	ROI ✓
<i>DCE</i> -	91	15
<i>DCE</i> ✓	0	11



## 9.3.3 DCE Contributes to CD Severity Assessment

**Inter-Feature Variance**

Manually measured enhancement scores are *enhancement\_T1* and *rce*. While *enhancement\_T1* categorizes the observed T1-signal into *normal*, *minor increase*, *moderate increase* and *marked increase* per segment, the *rce* (relative contrast enhancement) quantifies the signal enhancement before and after contrast agent application (see section 3.3.2). In contrast to these features, the automatic *DCE* incorporates the DCE-MRI sequence to extract the enhancement values. Since all three enhancement measures arise from different MRI sequences and measurement protocols, the inter-feature correlation should not be expected to be too high. Figure 9.12 shows that there is indeed a statistical correlation between *DCE* and *rce* (Pearson  $r=0.76$ ,  $p=7.8e-8$ ) and *DCE* and *enhancement\_T1* ( $r=0.63$ ,  $p=4.831e-06$ ). On the other hand, the three cases with largest *DCE* and therefore marked contrast enhancement are scored with highly variant *rce* or consistent lower *rce* by all four radiologists. This nicely illustrates the problematic of an accurate ROI for *rce* and *DCE*.

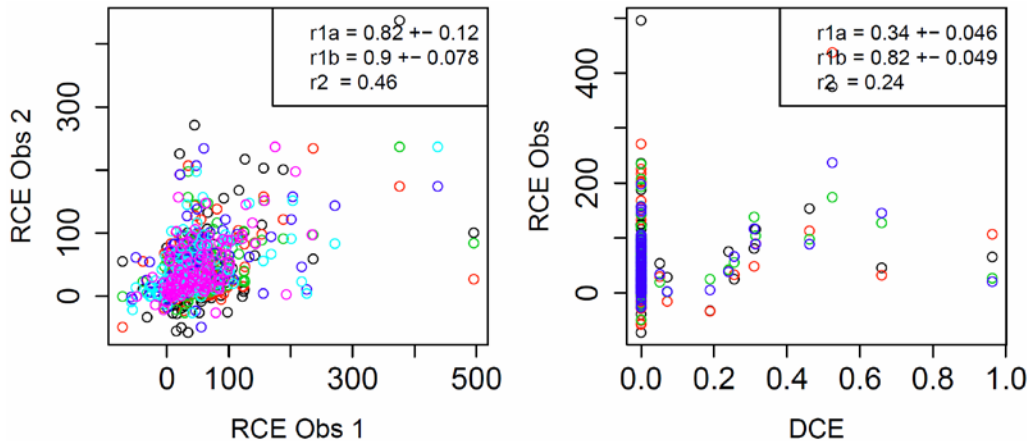


**Figure 9.12: Inter-feature relation of the two manual enhancement measures *enhancement\_T1* and *rce* and the automatic method *DCE*. All three quantify the MRI signal enhancement after gadobutrol application with different methods. Diseased bowel**

wall enhances more than healthy parts. *DCE* is statistically correlated to *enhancement\_T1* ( $r=.63$  for cases with measured *DCE*). The correlation to *rce* is  $r=.33$  ( $r=.76$  on cases with  $DCE < .6$ ). This is slightly higher than *rce* to *enhancement\_T1*:  $r=.47$  (on cases with measured *DCE*) or  $r=.41$  on all cases. All correlations are significant. The four observers are plotted with different colors.

### Inter-Observer Variance

In the best case, *DCE* has a correlation to *rce* stratified by observers of  $r=0.82 \pm 0.05$  on the subset of samples with measured *DCE* and  $DCE < 0.6$  (see Figure 9.13). On the same set, the inter-observer correlation of *rce* is  $r=0.9 \pm 0.08$ . Figure 9.13 visualizes the inter-observer correlation of *rce* and the correlation of *DCE* to *rce* on different subsets of samples (all samples, stratified by observer, subset of measured *DCE* and subset of measured *DCE* without outliers). In all cases, the *DCE* correlated slightly lower to *rce* than the observers correlate to each other.



**Figure 9.13: Inter-observer correlation of *rce* and correlation of *DCE* to *rce*, stratified by observer. Each circle identifies a sample. LEFT: Each color indicates an observer pair. RIGHT: Each color indicates an observer (*DCE* is unique for all observers).  $R_{1a}$ : Pearson correlation on cases with measured *DCE*, stratified by observer (-pair).  $R_{1b}$ : same as  $r_{1a}$  on cases with  $DCE < 0.6$ .  $R_2$ : correlation on all samples. Correlation values are comparable to each other.**

### Univariate CDEIS Correlation

The univariate correlation of *DCE* to CDEIS is  $r_1=0.43$  ( $p=1.3e-6$ ) on all 117 samples and  $r_2=0.64$  ( $p=0.03$ ) on the subset of 11 samples with measured *DCE* (Figure 9.14 right). This is slightly higher than the correlation of *rce* to CDEIS ( $r_1=0.34 \pm 0.07$ ,  $p < 0.001$  on all samples and  $r_2=0.57 \pm 0.07$ ,  $p=0.08$  on cases with measured *DCE*, stratified by observer) (Figure 9.14 left).

### Multivariate CDEIS Correlation

The new automatic feature *DCE* can be evaluated as top player for the CDEIS estimation when inserted to our feature selection pipeline (see Figure 9.15 left). The number of models with statistically not significantly different cross-validated performances is 360 and all top models rely on *DCE* as indicated in Figure 9.15 (left). The complete ranking of all models is displayed in Figure 9.15 (right). The first model without *DCE* is on rank 1049: *comb\_sign + length + muralT2 + rce*. Its cross-validated correlation to

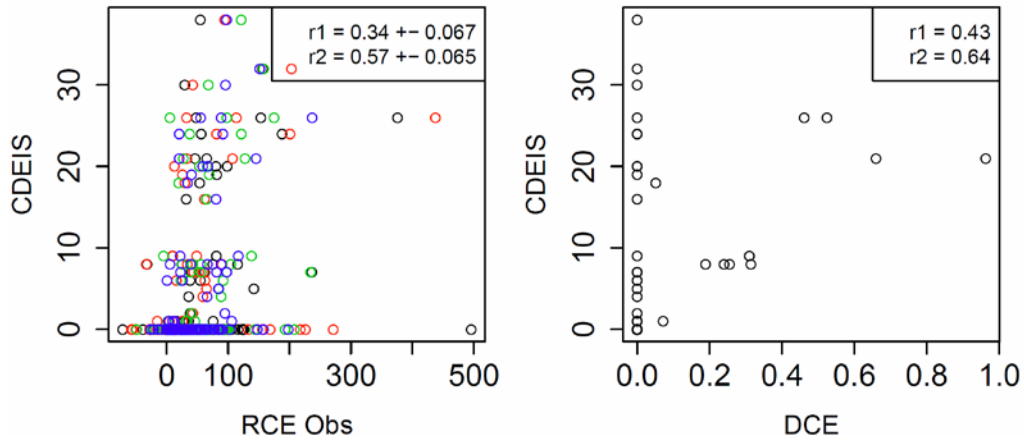


Figure 9.14: Correlation of manual *rce* (LEFT) and automatic *DCE* (RIGHT) to CDEIS.  $r_1$  and  $r_2$  correspond to the Pearson correlation on all samples ( $r_1$ ) or on the subset of 11 samples with *DCE* ( $r_2$ ).

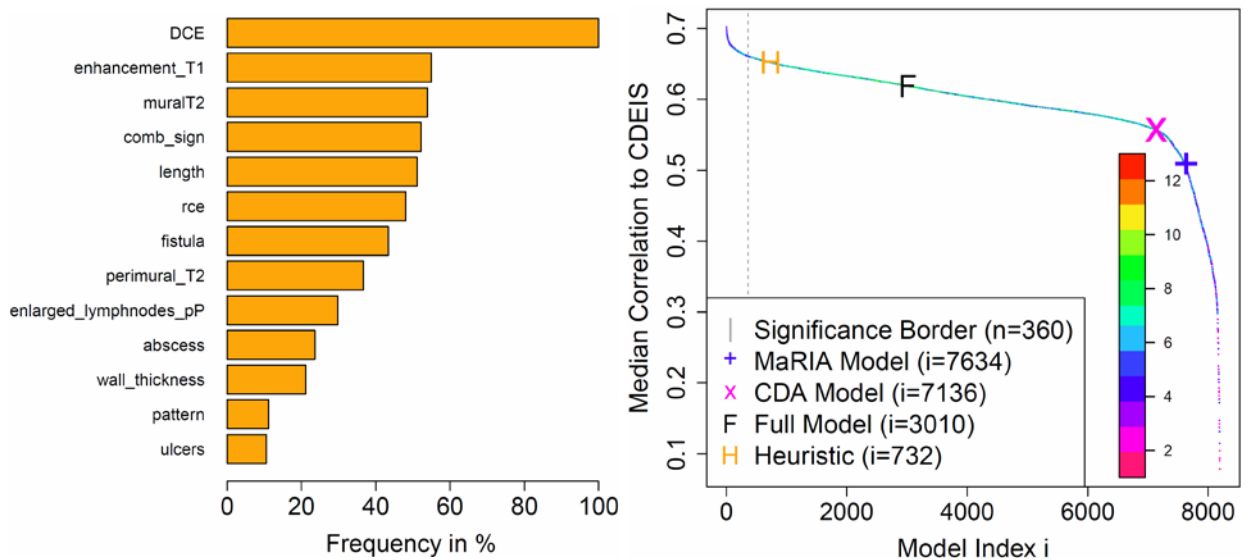
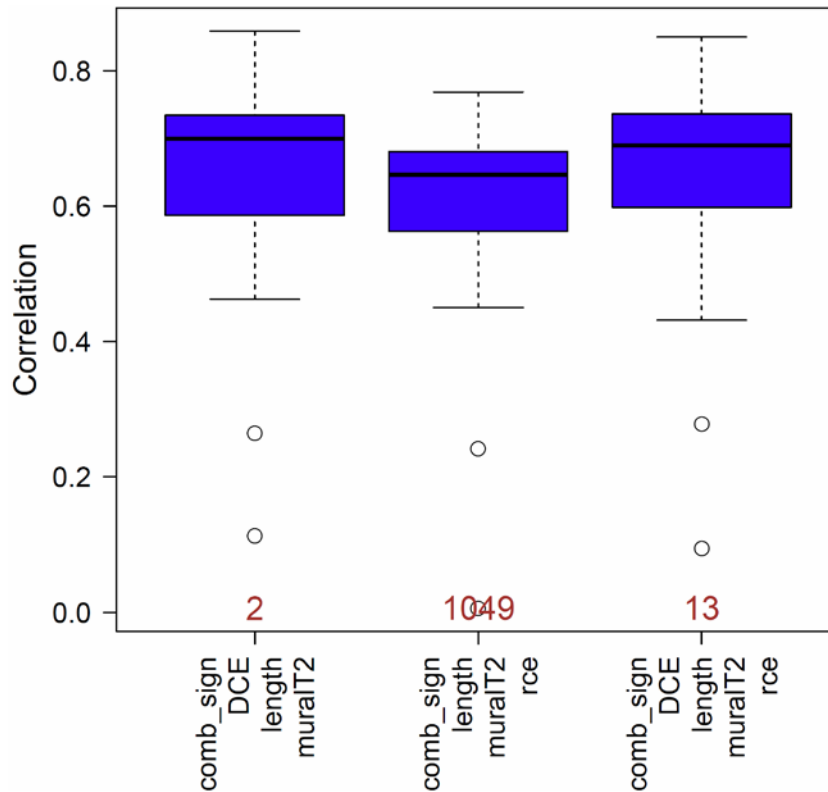


Figure 9.15: LEFT: Feature distribution of the top 360 models. The automatically measured *DCE* feature is present in all top models. RIGHT: Ranking of all models with and without *DCE* as CDEIS predictors. The first model without *DCE* (*comb\_sign*, *length*, *muralT2* and *rce*) is on rank 1049 with a median correlation to CDEIS of  $r=0.65$ . Also shown are ranks of the MaRIA model, AIS model, the full model with 13 features and the model found with stepwise selection as heuristic approach. The color codes the number of features used per model.

CDEIS (median  $r=0.65$ ,  $p=2.8e-14$ ) is shown in Figure 9.16 (middle). When *DCE* is added to this model, it reaches rank 13 (Figure 9.16 right). Interestingly, when *rce* is replaced by *DCE*, the model's rank further increases to rank two with  $r=0.70$  ( $p=0$ ) (Figure 9.16 left), which is a significant improvement of approximately 8% to the original model 1049. This strongly supports the idea to replace the relatively complex *rce* measurement by the automatic *DCE* measure, gaining accuracy to CD severity assessment.



**Figure 9.16: DCE as replacement for *rce*. The best manual model 1049 uses the same features as model 2: *comb\_sign*, *length* and *muralT2*. Additionally, the manual model incorporates *rce*, while model 2 uses *DCE* instead. The resulting improvement in CDEIS correlation from  $r=0.65$  to  $r=0.70$  is significant. Further, including *rce* and *DCE* together does not improve the correlation further (model 13).**

### 9.3.4 DCE and Manually Annotated Diseased Regions


In our dataset, all ROIs are provided by a medical expert who identified regions with visible bowel wall enhancement. The ROIs show a high correlation to the related feature *enhancement\_T1*, as illustrated in Table 9.4: the ROI density is larger in segments with high *enhancement\_T1*. On the other hand, there exist few cases where four radiologists scored high *enhancement\_T1*, but no ROI could be found or vice versa.

Since there are 26 ROI available in the dataset of which only 11 could be processed by *DCE* (due to bad registration result or mismatch of ROI with field of view of DCE-MRI), we investigated whether a larger amount of available data would increase the CDEIS prediction. Therefore, we used two simple ROI features as new computer-derived measurements: *polygon*, which is the size of the ROI in voxels, and *polymean*, a simple normalized signal intensity mean from the TRHIVE post contrast ROIs. Interestingly, *polygon* does not contribute at all to CDEIS prediction and does not occur in any of the top models. *Polymean* as a feature with

signal intensity information occurs in 28 % of the top models. Putting *polymean* together with *DCE* will not change the rankings: *polymean* still is a poor predictor, even if it has more data points available than *DCE*. Two conclusions can be drawn from this fact: First, a larger number of hand-drawn polygons which are converted to intensity-change features result in more CDEIS prediction. Second, *DCE* comes with a superior information of intensity change and therefore drastically improves the severity estimation, even if it has fewer data points. *A rare good feature is better than a frequent bad feature.*

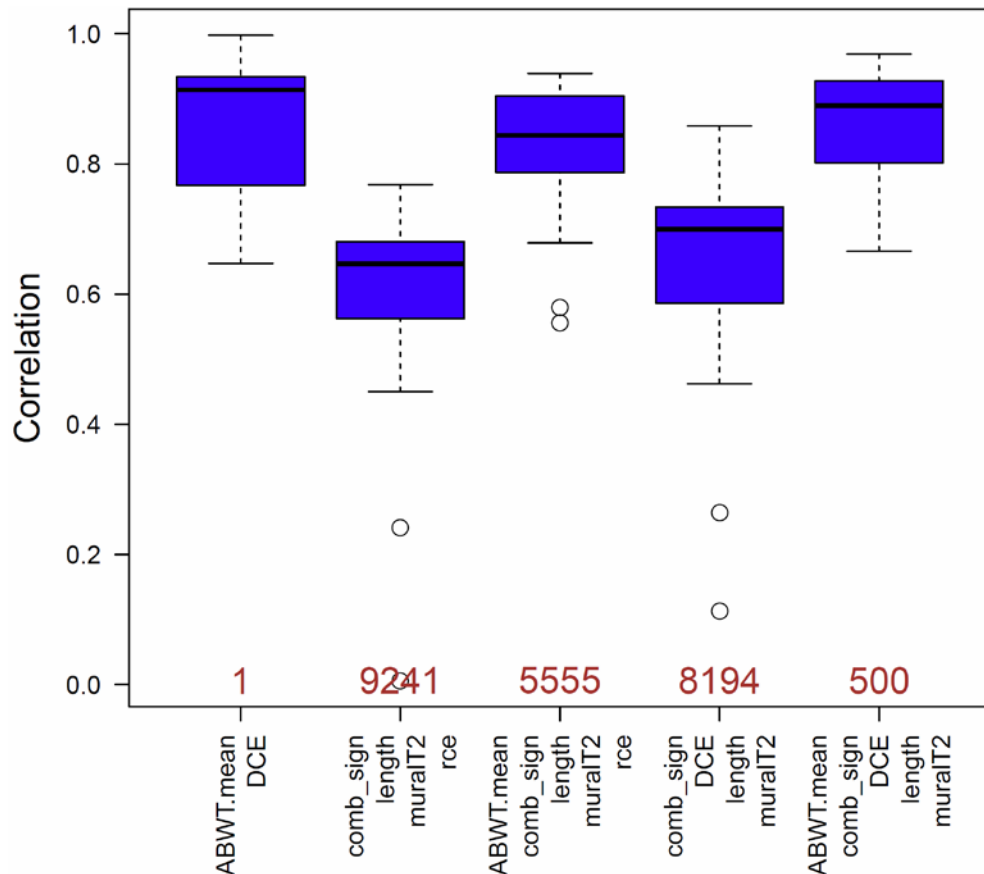
**Table 9.4: Summary of ROI with relation to all four raters' scorings of *enhancement\_T1*. Four radiologists scored *enhancement\_T1* (0, 1, 2 or 3 for normal, minor, moderate and marked) for every bowel segment. A fifth radiologist independently drew ROIs in the bowel segments where he identified signal enhancement. E.g. all radiologists agreed on 74 bowel segments to see no or only minor sign of wall enhancement (group 0). Still, the fifth medical expert outlined 8 ROIs (11%) based on his personal perception. This ratio (and also the number of ROI) correlates well with the strength of the signal enhancement.**

Group		Sum of 4 raters' <i>enhancement_T1</i>	<i>n</i> segments	<i>n</i> ROI	Proportion of present ROI
0	<i>Most raters scored '0'</i>	0	68	4	11 %
		1	6	4	
1	<i>Mild enhancement</i>	2	12	3	26 %
		3	8	1	
		4	2	1	
		5	1	1	
2	<i>Moderate enhancement</i>	6	5	4	60%
		7	5	2	
		8	2	1	
		9	3	2	
3	<i>Most raters scored '3'</i>	10	3	3	80 %
		11	2	1	
		12	0	0	



## 9.4 Combining ABWT and DCE

Putting *ABWT.mean* and *DCE* together reveals that especially automatic bowel wall thickness measurement improves CDEIS regression. Figure 9.17 shows the classifier with highest CDEIS correlation using *ABWT.mean* and *DCE* solely with  $r=0.91$  ( $p=0$ ) which is an improvement of 40% compared to the model with manual features ( $r=0.65$ ).



**Figure 9.17: Comparison of cross-validated classifiers with and without automatic features. Model 1 is the very best model and uses both automatic features solely. Model 9241 is the first model which uses only manual features. Models 5555, 8194 and 500 add either *ABWT.mean*, or replace *rce* by *DCE* or both.**

## 10 CONCLUSION

Computational interpretation of medical images is a wide research field with manifold research questions and applications. Object detection and classification is a central process in this field and requires modern machine learning algorithms. This thesis examines directions in computational pathology and computational radiology on the example of cancer cell classification and Crohn's disease detection, both with regard to the role of the shape of objects for their proper detection and classification.

### 10.1 *Computational Pathology*

In computational pathology we investigated a new algorithmic workflow for histopathological staining estimation on tissue microarrays (TMA). In contrast to already existing alternatives which aim to directly detect cancerous cell nuclei on the image, our approach separates nucleus detection and classification into two consecutive steps. Thus, each individual step can be optimized for accurate predictions. We have shown at the example of nucleus classification that improved modules of the TMA processing pipeline can also improve the complete staining estimation.

#### 10.1.1 *Nucleus Classification*

Considering nucleus classification on TMA, we found nucleus shape to be a crucial descriptor for the computer-aided decision of its diseased state. Shape descriptors such as 1D-signature, Freeman chain code or pyramid histogram of oriented gradients delivered favorable classification outcome, regardless of the underlying classifier. Further, the similarity based classification of nuclei turned out to be privileged for classification accuracy especially when highly nonlinear similarities are considered and classifier ensembles such as random forests or multiple kernels are used. All these facts together with the observed inter-observer agreement of 80% document the non-trivial nature of cell nucleus classification in medical imaging. Nevertheless, we are currently able to design complex automated classification systems with coequal classification performance for renal clear cell carcinoma nuclei as trained human pathologists.

### 10.1.2 Nucleus Detection

We pursued the role of shape descriptors to the task of nucleus detection. Nucleus segmentation has been performed by the use of graph-cuts or superpixels. While graph-cuts presume a prior image patching with centered nuclei in each patch, both, graph-cuts and superpixels require an assumption on the approximate nucleus radius in the images. Once the nuclei are segmented, a binary classifier can immediately differentiate nuclei from background. While the detection of nuclei in MIB-1 stained TMA seems to be more difficult than the subsequent classification, we prove also here with an F-score of 0.85-0.93 to reach the performance range of trained human pathologists with an F-score of 0.9-0.97.

## 10.2 Computational Radiology

### 10.2.1 CD Severity Score

Concerning computational radiology, the aim of this thesis was twofold. On the one hand, a Crohn's disease *MRI severity score* has been developed which clearly correlates to the endoscopic CDEIS. The correlation ranges at 70% on a bowel segment basis, when solely manual read-outs are considered. Further, the score respects the variances among different experts by incorporating the data of four radiologists from two hospitals. The performance of the proposed models is stable for unseen patients, on a segment basis as well as on a per-patient basis, as illustrated on several validation experiments. This correlation is already higher than that of the literature scores MaRIA and CDA. The appropriate development of the new score was possible due to a comprehensive data acquisition with 17 manually assessed predictor candidates (more than ever before), due to an extensive multiple labeling work by four radiologists and due to an exhaustive search algorithm through all potential model candidates.

### 10.2.2 Multimodal Features for Severity Assessment

The correlation to endoscopic CD severity can significantly be improved by the use of additional patient data such as computer-read features. Especially the automatically measured bowel wall thickness and the automatically measured contrast enhancement have shown their potential to increase the correlation of an MRI model to CDEIS up to 90%, which is almost perfect for biomedical research problems. Still, these new automatic features require user interaction to coarsely localize regions of interest for which the features are to be automatically computed.



### 10.2.3 *Exhaustive Search vs. Heuristic*

The proposed pipeline in this thesis for model development of a CD severity assessment model in MRI uses exhaustive search among all possible feature combinations. This concept is of course only feasible when the number of features allows the computational effort, and this is the case in our problem. We have shown in parallel how a heuristic approach would solve the same task: stepwise selection as feature selection method which successively includes features that improve the CDEIS regression. In almost all cases, the MRI model selected by the heuristic approach belongs either to the class of top models or ranks close by this group. This nicely documents the suitability of these types of algorithms. However, it also illustrates that the heuristic would not always find the global optimum. Also, a heuristic might not make aware of the fact that multiple solutions are possible to a given problem. To explore the whole model space, we therefore encourage to use exhaustive search algorithms whenever possible.

Our proposed exhaustive search data analysis pipeline reveals several favorable characteristics: Due to its generality it can be applied on various data types, as we have shown on proteomic and histological datasets in a variety of cancer research projects. Additionally, the pipeline allows a holistic view on the search space, enabling model and feature rankings for easy interpretation of the influence of different features.

### 10.2.4 *CD Detection with MRI*

On the other hand and as the second aim of this study, we elaborated a completely automatic hierarchical classification algorithm for CD detection and segmentation in MRI. The two main steps in this hierarchy are first the coarse localization of diseased regions of interest via supervoxel classification, and second the voxel wise classification for precise CD segmentation. The algorithm uses standard image features such as intensity, texture and curvature, as well as customized and newly designed features tailored to our problem such as higher order statistics, context based features and constraints on spatial distribution and smoothness. This classification scheme has shown high accuracy in terms of Dice metric (approximately 90%) and Hausdorff distance (2 mm deviance at most) to manually segmented areas of CD. The complete algorithm has been validated on a leave-one-patient-out cross-validation basis.

### 10.2.5 *Combining CD Detection and Severity – Outlook*

On the way to automated CD assessment in clinical MRI data, a next step would be to combine the qualitative automated CD localization with the

quantitative automated CD severity estimation. For this, feature extraction algorithms such as wall thickness or dynamic contrast enhancement have to be aligned with CD segmentation procedures. This is the final step in VIGOR++, and several considerations and limitations have to be kept in mind: First, the suggested CD detection and segmentation algorithm is trained to recognize enhanced bowel wall MRI signals. The detection of ulcers, stenosis, fistula, edema, different patterns, comb sign, and other CD related abnormalities in MRI scans was not intended. The manual spatial labeling of such features in MRI is too extensive and are therefore not part of the training set.

Second, the automated wall thickness measurement requires the indication of the centerline of the bowel on a given segment of interest. Also the automated dynamic contrast enhancement measurement requires the definition of an appropriate region of interest. These user interactions are still needed at this stage, even for automatic processing.

Finally, the computation of dynamic contrast enhancement might have limitations due to the potentially difficult registration of the DCE images to the post contrast THRIVE sequence and by the missing DCE image data on the specified loci (note that the DCE images have a smaller field of view and a lower resolution than post contrast THRIVE sequences in order to enable high temporal resolution). However, this problem might be solved in near future, when even higher performing MRI scanners are available which allow DCE sequences at higher spatial resolution.

### ***10.3 Computer Aided CD Assessment with MRI – Outlook***

This thesis explores the field of computational approaches in abdominal MRI at the example of Crohn's disease assessment. Both the CD severity estimation based on manual MRI features as well as the automatic CD detection, segmentation and feature extraction have been investigated. For such a difficult problem, where the exact course of the disease is medically not fully understood, and where even trained human experts show a considerable discrepancy in judging the patients' data, this research describes a first step in the direction of standardized and more objective individualized CD treatments being more and more computer-assisted. The present study clearly contributes to the computational understanding of abdominal MRI.

In the scope of the VIGOR++ project, the further validation of proposed algorithms and approaches is suggested on prospective datasets from our medical partners. Still, for a computer-aided technique to be adopted

by the practical medical community, long term validation on various new datasets is necessary.

The fusion of several data domains such as manual MRI features, endoscopic findings and computer-read features has been a major challenge in this work. Also in future medical problems, the advent of *multiple data sources* will designate the design of appropriate algorithms for diagnosis and grading disease activity. This is the way to account for the manifold facets of the underlying biological problems. Further, this particularly reflects the medical decision making process of medical doctors, who almost always rely on multiple data sources and multiple aspects to build a holistic view of the disease status of the patient.



## REFERENCES

- Abdi, H. 2007. "Bonferroni and Sidak corrections for multiple comparisons." In *Encyclopedia of Measurement and Statistics*, edited by Neil J. Salkind, 103 -7. Sage Pubns.
- Achanta, R., A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. 2012. "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods." Review of. *Ieee Transactions on Pattern Analysis and Machine Intelligence* 34 (11):2274-81.
- Ahonen, T., A. Hadid, and M. Pietikainen. 2004. "Face recognition with local binary patterns." Review of. *Computer Vision - Eccv 2004, Pt 1* 3021:469-81.
- Aurenhammer, F. 1991. "Voronoi Diagrams - a Survey of a Fundamental Geometric Data Structure." Review of. *Computing Surveys* 23 (3):345-405.
- Avni, U., H. Greenspan, and J. Goldberger. 2011. "X-ray Categorization and Spatial Localization of Chest Pathologies." Review of. *Medical Image Computing and Computer-Assisted Intervention, Miccai 2011, Pt Iii* 6893:199-206.
- Bach, F.R., G.R.G. Lanckriet, and M.I. Jordan. 2004. "Multiple kernel learning, conic duality, and the SMO algorithm." In *Proceedings of the twenty-first international conference on Machine learning*, 6. Banff, Alberta, Canada: ACM.
- Baert, F., L. Moortgat, G. Van Assche, P. Caenepeel, P. Vergauwe, M. De Vos, P. Stokkers, et al. 2010. "Mucosal Healing Predicts Sustained Clinical Remission in Patients With Early-Stage Crohn's Disease." Review of. *Gastroenterology* 138 (2):463-8.
- Barclay, A.R., R.K. Russell, M.L. Wilson, W.H. Gilmour, J. Satsangi, and D.C. Wilson. 2009. "Systematic Review: The Role of Breastfeeding in the Development of Pediatric Inflammatory Bowel Disease." Review of. *Journal of Pediatrics* 155 (3):421-6.
- Bauer, S., L.P. Nolte, and M. Reyes. 2011. "Fully Automatic Segmentation of Brain Tumor Images Using Support Vector Machine Classification in Combination with Hierarchical Conditional Random Field Regularization." Review of. *Medical Image Computing and Computer-Assisted Intervention, Miccai 2011, Pt Iii* 6893:354-61.
- Baumgart, D.C., and W.J. Sandborn. 2012. "Crohn's disease." Review of. *Lancet* 380 (9853):1590-605.
- Ben-Hur, A., C.S. Ong, S. Sonnenburg, B. Scholkopf, and G. Ratsch. 2008. "Support Vector Machines and Kernels for Computational Biology." Review of. *Plos Computational Biology* 4 (10).

- Berks, M., Z.Z. Chen, S. Astley, and C. Taylor. 2011. "Detecting and Classifying Linear Structures in Mammograms Using Random Forests." Review of. *Information Processing in Medical Imaging* 6801:510-24.
- Best, W.R., J.M. Bechtel, J.W. Singleton, and F. Kern, Jr. 1976. "Development of a Crohn's disease activity index. National Cooperative Crohn's Disease Study." Review of. *Gastroenterology* 70 (3):439-44.
- Beucher, S., and C. Lantuejoul. 1979. Use of Watersheds in Contour Detection. Paper presented at the International Workshop on Image Processing: Real-time Edge and Motion Detection/Estimation, Rennes, France.
- Bhattacharyya, A. 1943. "On a measure of divergence between two statistical populations defined by their probability distributions." Review of. *Bulletin of the Calcutta Mathematical Society* 35:99-109.
- Biotronics3D. 2013. "3Dnet Medical - Medical Imaging Cloud Service." Biotronics3D, Accessed 12.11.2013. <http://www.3dnetmedical.com>.
- Borley, N.R., N.J. Mortensen, D.P. Jewell, and B.F. Warren. 2000. "The relationship between inflammatory and serosal connective tissue changes in ileal Crohn's disease: evidence for a possible causative link." Review of. *J Pathol* 190 (2):196-202.
- Bosch, A., A. Zisserman, and X. Munoz. 2007. "Representing shape with a spatial pyramid kernel." In *Proceedings of the 6th ACM international conference on Image and video retrieval*, 401-8. Amsterdam, The Netherlands: ACM.
- Boykov, Y., and G. Funka-Lea. 2006. "Graph cuts and efficient N-D image segmentation." Review of. *International Journal of Computer Vision* 70 (2):109-31.
- Boykov, Y., and V. Kolmogorov. 2004. "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision." Review of. *Ieee Transactions on Pattern Analysis and Machine Intelligence* 26 (9):1124-37.
- Boykov, Y., O. Veksler, and R. Zabih. 2001. "Fast approximate energy minimization via graph cuts." Review of. *Ieee Transactions on Pattern Analysis and Machine Intelligence* 23 (11):1222-39.
- Brandt, S., C. Montagna, A. Georgis, P. Schüffler, M. Buhler, B. Seifert, T. Thiesler, et al. 2013. "The combined expression of the stromal markers fibronectin and SPARC improves the prediction of survival in diffuse large B-cell lymphoma." Review of. *Experimental Hematology & Oncology* 2 (1):27.
- Breiman, L. 2001. "Random forests." Review of. *Machine Learning* 45 (1):5-32.

- 
- Breslow, N., C.W. Chan, G. Dhom, R.A.B. Drury, L.M. Franks, B. Gellei, Y.S. Lee, et al. 1977. "Latent carcinoma of prostate at autopsy in seven areas. Collaborative study organized by the International Agency for Research on Cancer, Lyons, France." Review of. *International Journal of Cancer* 20 (5):680-8.
- Burt, R.K., R.M. Craig, F. Milanetti, K. Quigley, P. Gozdzia, J. Bucha, A. Testori, et al. 2010. "Autologous nonmyeloablative hematopoietic stem cell transplantation in patients with severe anti-TNF refractory Crohn disease: long-term follow-up." Review of. *Blood* 116 (26):6123-32.
- Cellier, C., T. Sahmoud, E. Froguel, A. Adenis, J. Belaiche, J.F. Bretagne, C. Florent, et al. 1994. "Correlations between clinical activity, endoscopic severity, and biological parameters in colonic or ileocolonic Crohn's disease. A prospective multicentre study of 121 cases. The Groupe d'Etudes Therapeutiques des Affections Inflammatoires Digestives." Review of. *Gut* 35 (2):231-5.
- Chang, C.C., and C.J. Lin. 2001. "LIBSVM: A Library for Support Vector Machines." Review of. *Acm Transactions on Intelligent Systems and Technology* 2 (3).
- Chang, C.C., and C.J. Lin. 2011. "LIBSVM: A Library for Support Vector Machines." Review of. *Acm Transactions on Intelligent Systems and Technology* 2 (3).
- Chew, P. "Voronoi Diagram / Delaunay Triangulation." Accessed 16.01.2014. <http://www.cs.cornell.edu/home/chew/Delaunay.html>.
- Cima, I., R. Schiess, P. Wild, M. Kaelin, P. Schüffler, V. Lange, P. Picotti, et al. 2011. "Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer." Review of. *Proceedings of the National Academy of Sciences of the USA* 108 (8):3342-7.
- Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales." Review of. *Educational and Psychological Measurement* 20 (1):37-46.
- Cohen, J. 1968. "Weighted Kappa - Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit." Review of. *Psychological Bulletin* 70 (4):213-&.
- Cohn, D.A., Z. Ghahramani, and M.I. Jordan. 1996. "Active learning with statistical models." Review of. *Journal of Artificial Intelligence Research* 4:129-45.
- Cortes, C., M. Mehryar, and A. Rostamizadeh. 2009. Learning non-linear combinations of kernels. Paper presented at the Neural Information Processing Systems.
- Crohn, B.B., L. Ginzburg, and G.D. Oppenheimer. 1932. "Regional ileitis: A pathologic and clinical entity." Review of. *Journal of the American Medical Association* 99 (16):1323-9.
-

- Daperno, M., G. D'Haens, G. Van Assche, F. Baert, P. Bulois, V. Maunoury, R. Sostegni, et al. 2004. "Development and validation of a new, simplified endoscopic activity score for Crohn's disease: the SES-CD." Review of. *Gastrointest Endosc* 60 (4):505-12.
- Devalois, R.L., D.G. Albrecht, and L.G. Thorell. 1982. "Spatial-Frequency Selectivity of Cells in Macaque Visual-Cortex." Review of. *Vision Research* 22 (5):545-59.
- Dice, L.R. 1945. "Measures of the Amount of Ecologic Association between Species." Review of. *Ecology* 26 (3):297-302.
- Duijvestein, M., A.C.W. Vos, H. Roelofs, M.E. Wildenberg, B.B. Wendrich, H.W. Verspaget, E.M.C. Kooy-Winkelaar, et al. 2010. "Autologous bone marrow-derived mesenchymal stromal cell treatment for refractory luminal Crohn's disease: results of a phase I study." Review of. *Gut* 59 (12):1662-9.
- Fleiss, J.L. 1971. "Measuring Nominal Scale Agreement among Many Raters." Review of. *Psychological Bulletin* 76 (5):378-82.
- Fleiss, J.L., and J. Cohen. 1973. "Equivalence of Weighted Kappa and Intraclass Correlation Coefficient as Measures of Reliability." Review of. *Educational and Psychological Measurement* 33 (3):613-9.
- Freeman, H. 1961. "On the Encoding of Arbitrary Geometric Configurations." Review of. *Electronic Computers, IRE Transactions on EC-10* (2):260-8.
- Fuchs, T.J., and J.M. Buhmann. 2011a. "Computational pathology: Challenges and promises for tissue analysis." Review of. *Computerized Medical Imaging and Graphics* 35 (7-8):515-30.
- Fuchs, T.J., and J.M. Buhmann. 2011b. "Computational pathology: challenges and promises for tissue analysis." Review of. *Comput Med Imaging Graph* 35 (7-8):515-30.
- Fuchs, T.J., J. Haybaeck, P.J. Wild, M. Heikenwalder, H. Moch, A. Aguzzi, and J.M. Buhmann. 2009. "Randomized Tree Ensembles for Object Detection in Computational Pathology." Review of. *Advances in Visual Computing, Pt 1, Proceedings* 5875:367-78.
- Fuchs, T.J., P.J. Wild, H. Moch, and J.M. Buhmann. 2008a. "Computational pathology analysis of tissue microarrays predicts survival of renal clear cell carcinoma patients." Review of. *Med Image Comput Comput Assist Interv* 11 (Pt 2):1-8.
- Fuchs, T.J., P.J. Wild, H. Moch, and J.M. Buhmann. 2008b. "Computational Pathology Analysis of Tissue Microarrays Predicts Survival of Renal Clear Cell Carcinoma Patients." Review of. *Medical Image Computing and Computer-Assisted Intervention - Miccai 2008, Pt Ii, Proceedings* 5242:1-8.
- Garcia-Olmo, D., D. Herreros, I. Pascual, J.A. Pascual, E. Del-Valle, J. Zorrilla, P. De-La-Quintana, et al. 2009. "Expanded Adipose-



- Derived Stem Cells for the Treatment of Complex Perianal Fistula: a Phase II Clinical Trial." Review of. *Diseases of the Colon & Rectum* 52 (1):79-86.
- Gent, A.E., M.D. Hellier, R.H. Grace, E.T. Swarbrick, and D. Coggon. 1994. "Inflammatory Bowel-Disease and Domestic Hygiene in Infancy." Review of. *Lancet* 343 (8900):766-7.
- Gerdes, M.J., C.J. Sevinsky, A. Sood, S. Adak, M.O. Bello, A. Bordwell, A. Can, et al. 2013. "Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue." Review of. *Proc Natl Acad Sci U S A* 110 (29):11982-7.
- Geremia, E., B.H. Menze, O. Clatz, E. Konukoglu, A. Criminisi, and N. Ayache. 2010. "Spatial Decision Forests for MS Lesion Segmentation in Multi-Channel MR Images." Review of. *Medical Image Computing and Computer-Assisted Intervention - Miccai 2010, Pt I* 6361:111-8.
- Giesen, C., H.A. Wang, D. Schapiro, N. Zivanovic, A. Jacobs, B. Hattendorf, P.J. Schüffler, et al. 2014. "Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry." Review of. *Nat Methods* 11 (4):417-22.
- Gilbert, D. "JFreeChart." Accessed 16.01.2014. <http://www.jfree.org/jfreechart/>.
- Gönen, M., and E. Alpaydin. 2013. "Localized algorithms for multiple kernel learning." Review of. *Pattern Recognition* 46 (3):795-807.
- Gönen, M., A. Ulaş, P. Schüffler, U. Castellani, and V. Murino. 2011. "Combining Data Sources Nonlinearly for Cell Nucleus Classification of Renal Cell Carcinoma." In *Similarity-Based Pattern Recognition*, edited by Marcello Pelillo and Edwin R Hancock, 250-60. Springer Berlin Heidelberg.
- Gonzalez, R.C., R.E. Woods, and S.L. Eddins. 2004. *Digital Image processing using MATLAB*. Upper Saddle River, NJ: Pearson/Prentice Hall.
- Graf, F. 2012. "JFeatureLib v1.6.0." In.
- Grignon, D., J. Eble, S. Bonsib, and H. Moch. 2004. "Clear cell renal cell carcinoma." In *Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs*, edited by John N. Eble, Guido Sauter, Jonathan I. Epstein and Isabell A. Sesterhenn, 23-5. Lyon: IARC Press.
- Gunther, D., R. Frischknecht, C.A. Heinrich, and H.J. Kahlert. 1997. "Capabilities of an Argon Fluoride 193 nm excimer laser for laser ablation inductively coupled plasma mass spectrometry microanalysis of geological materials." Review of. *Journal of Analytical Atomic Spectrometry* 12 (9):939-44.
- Guyatt, G., A. Mitchell, E.J. Irvine, J. Singer, N. Williams, R. Goodacre, and C. Tompkins. 1989. "A new measure of health status for clinical

- trials in inflammatory bowel disease." Review of. *Gastroenterology* 96 (3):804-10.
- Gwet, K.L. 2008. "Computing inter-rater reliability and its variance in the presence of high agreement." Review of. *Br J Math Stat Psychol* 61 (Pt 1):29-48.
- Haibin, L., and K. Okada. 2006. Diffusion Distance for Histogram Comparison. Paper presented at the Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, 17-22 June 2006.
- Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. 2009. "The WEKA data mining software: an update." Review of. *SIGKDD Explor. Newsl.* 11 (1):10-8.
- Harvey, R.F., and J.M. Bradshaw. 1980. "A simple index of Crohn's-disease activity." Review of. *Lancet* 1 (8167):514.
- Hausdorff, F. 1957. *Set Theory*: American Mathematical Society.
- Hellinger, E. 1909. *Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen*. Berlin: Reimer.
- Hou, J.K., H. El-Serag, and S. Thirumurthi. 2009. "Distribution and Manifestations of Inflammatory Bowel Disease in Asians, Hispanics, and African Americans: A Systematic Review." Review of. *American Journal of Gastroenterology* 104 (8):2100-9.
- Irvine, E.J. 1995. "Usual therapy improves perianal Crohn's disease as measured by a new disease activity index. McMaster IBD Study Group." Review of. *J Clin Gastroenterol* 20 (1):27-32.
- Jemal, A., F. Bray, M.M. Center, J. Ferlay, E. Ward, and D. Forman. 2011. "Global Cancer Statistics." Review of. *Ca-a Cancer Journal for Clinicians* 61 (2):69-90.
- Jones, D.T., M.T. Osterman, M. Bewtra, and J.D. Lewis. 2008. "Passive Smoking and Inflammatory Bowel Disease: A Meta-Analysis." Review of. *American Journal of Gastroenterology* 103 (9):2382-93.
- Joossens, M., M. Simoons, S. Vermeire, X. Bossuyt, K. Geboes, and P. Rutgeerts. 2007. "Contribution of genetic and environmental factors in the pathogenesis of Crohn's disease in a large family with multiple cases." Review of. *Inflamm Bowel Dis* 13 (5):580-4.
- Julesz, B., E.N. Gilbert, L.A. Shepp, and H.L. Frisch. 1973. "Inability of humans to discriminate between visual textures that agree in second-order statistics-revisited." Review of. *Perception* 2 (4):391-405.
- Kalin, M., I. Cima, R. Schiess, N. Fankhauser, T. Powles, P. Wild, A. Templeton, et al. 2011. "Novel Prognostic Markers in the Serum of Patients With Castration-Resistant Prostate Cancer Derived From Quantitative Analysis of the Pten Conditional Knockout Mouse Proteome." Review of. *Eur Urol* 60 (6):1235-43.

- Kaplan, G.G., J. Hubbard, J. Korzenik, B.E. Sands, R. Panaccione, S. Ghosh, A.J. Wheeler, et al. 2010. "The Inflammatory Bowel Diseases and Ambient Air Pollution: A Novel Association." Review of *American Journal of Gastroenterology* 105 (11):2412-9.
- Kirsner, J.B. 1988. "Historical aspects of inflammatory bowel disease." Review of *J Clin Gastroenterol* 10 (3):286-97.
- Kloft, M., U. Brefeld, S. Sonnenburg, and A. Zien. 2011. "l(p)-Norm Multiple Kernel Learning." Review of *Journal of Machine Learning Research* 12:953-97.
- Kullback, S., and R.A. Leibler. 1951. "On Information and Sufficiency." Review of *Annals of Mathematical Statistics* 22 (1):79-86.
- Kuncheva, L.I. 2004. *Combining Pattern Classifiers: Methods and Algorithms*: Wiley-Interscience.
- Lanckriet, G.R.G., N. Cristianini, P. Bartlett, L. El Ghaoui, and M.I. Jordan. 2004. "Learning the kernel matrix with semidefinite programming." Review of *Journal of Machine Learning Research* 5:27-72.
- Landis, J.R., and G.G. Koch. 1977. "The measurement of observer agreement for categorical data." Review of *Biometrics* 33 (1):159-74.
- Larson, D.W., and J.H. Pemberton. 2004. "Current concepts and controversies in surgery for IBD." Review of *Gastroenterology* 126 (6):1611-9.
- Lawless, J.F., and K. Singhal. 1978. "Efficient Screening of Nonnormal Regression-Models." Review of *Biometrics* 34 (2):318-27.
- Lee, J.S., S.J. Shin, M.T. Collins, I.D. Jung, Y.I. Jeong, C.M. Lee, Y.K. Shin, et al. 2009a. "Mycobacterium avium subsp paratuberculosis Fibronectin Attachment Protein Activates Dendritic Cells and Induces a Th1 Polarization." Review of *Infection and Immunity* 77 (7):2979-88.
- Lee, S.K., B.K. Kim, T.I. Kim, and W.H. Kim. 2009b. "Differential diagnosis of intestinal Behcet's disease and Crohn's disease by colonoscopic findings." Review of *Endoscopy* 41 (1):9-16.
- Lee, W.J., S. Verzakov, and R.P.W. Duin. 2007. "Kernel combination versus classifier combination." Review of *Multiple Classifier Systems, Proceedings* 4472:22-31.
- Lee, Y.J., S.K. Yang, T.H. Kim, H.K. Song, S.J. Myung, H.Y. Jung, G.H. Lee, et al. 2003. "Differential diagnosis of intestinal tuberculosis and Crohn's disease by colonoscopic findings." Review of *Gastrointest Endosc* 57 (5):Ab217-Ab.
- Lentschig, M.G., P. Reimer, U.L. Rausch-Lentschig, T. Allkemper, M. Oelerich, and G. Laub. 1998. "Breath-hold gadolinium-enhanced MR angiography of the major vessels at 1.0 T: Dose-response

- findings and angiographic correlation." Review of. *Radiology* 208 (2):353-7.
- Lerebours, E., C. Gower-Rousseau, V. Merle, F. Brazier, S. Debeugny, R. Marti, J.L. Salomez, et al. 2007. "Stressful life events as a risk factor for inflammatory bowel disease onset: A population-based case-control study." Review of. *American Journal of Gastroenterology* 102 (1):122-31.
- Leśniowski, A. 1903. "Przyczynę do chirurgii kiszek." Review of. *Medycyna (Warszawa)* 31:460-4, 83-89, 514-8.
- Lewis, D.P., T. Jebara, and W.S. Noble. 2006. "Nonstationary kernel combination." In *Proceedings of the 23rd international conference on Machine learning*, 553-60. Pittsburgh, Pennsylvania: ACM.
- Li, Z., J.A.W. Tielbeek, M.W.A. Caan, M.L.W. Ziech, C.Y. Nio, J. Stoker, L.J. van Vliet, et al. 2013. "Expiration Phase Template-based Motion Correction of Free-Breathing Abdominal Dynamic Contrast Enhanced MRI." Review of. *in submission*.
- Light, R.J. 1971. "Measures of Response Agreement for Qualitative Data - Some Generalizations and Alternatives." Review of. *Psychological Bulletin* 76 (5):365-77.
- Lin, W., J. Guo, M.A. Rosen, and H.K. Song. 2008. "Respiratory motion-compensated radial dynamic contrast-enhanced (DCE)-MRI of chest and abdominal lesions." Review of. *Magn Reson Med* 60 (5):1135-46.
- Liu, C.J., and H. Wechsler. 2002. "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition." Review of. *Ieee Transactions on Image Processing* 11 (4):467-76.
- Liu, Z., L. Smith, J.A. Sun, M. Smith, and R. Warr. 2011. "Biological Indexes Based Reflectional Asymmetry for Classifying Cutaneous Lesions." Review of. *Medical Image Computing and Computer-Assisted Intervention, Miccai 2011, Pt Iii* 6893:124-32.
- Louis, E., J.Y. Mary, G. Vernier-Massouille, J.C. Grimaud, Y. Bouhnik, D. Laharie, J.L. Dupas, et al. 2012. "Maintenance of Remission Among Patients With Crohn's Disease on Antimetabolite Therapy After Infliximab Therapy Is Stopped." Review of. *Gastroenterology* 142 (1):63-U201.
- Lowagie, B. 2010. *IText in action*. 2nd ed. Greenwich, Conn.: Manning.
- Lux, M., and S.A. Chatzichristofis. 2008. "Lire: lucene image retrieval: an extensible java CBIR library." In *Proceedings of the 16th ACM international conference on Multimedia*, 1085-8. Vancouver, British Columbia, Canada: ACM.
- M'Koma, A.E. 2013. "Inflammatory Bowel Disease: An Expanding Global Health Problem." Review of. *Clinical Medicine Insights: Gastroenterology* 6 (3829):33-47.

- 
- Mahapatra, D., and J.M. Buhmann. 2012a. "Cardiac LV and RV Segmentation Using Mutual Context Information." In *Machine Learning in Medical Imaging*, edited by Fei Wang, Dinggang Shen, Pingkun Yan and Kenji Suzuki, 201-9. Springer Berlin Heidelberg.
- Mahapatra, D., P.J. Schüffler, J. Tielbeek, F.M. Vos, and J.M. Buhmann. 2013a. Crohn's Disease Tissue Segmentation from Abdominal MRI Using Semantic Information and Graph Cuts. Paper presented at the IEEE 10th International Symposium on Biomedical Imaging, San Francisco.
- Mahapatra, D., P.J. Schüffler, J.A.W. Tielbeek, J.M. Buhmann, and F.M. Vos. 2012b. "A Supervised Learning Based Approach to Detect Crohn's Disease in Abdominal MR Volumes." In *Abdominal Imaging. Computational and Clinical Applications*, edited by Hiroyuki Yoshida, David Hawkes and MichaelW Vannier, 97-106. Springer Berlin Heidelberg.
- Mahapatra, D., P.J. Schüffler, J.A.W. Tielbeek, J.M. Buhmann, and F.M. Vos. 2013b. "A Supervised Learning Approach for Crohn's Disease Detection Using Higher-Order Image Statistics and a Novel Shape Asymmetry Measure." Review of. *Journal of Digital Imaging* 26 (5):920-31.
- Mahapatra, D., P.J. Schüffler, J.A.W. Tielbeek, J.C. Makanyanga, J. Stoker, S.A. Taylor, F.M. Vos, et al. 2013c. "Automatic Detection and Segmentation of Crohn's Disease Tissues from Abdominal MRI." Review of. *IEEE Trans Med Imaging* 32 (12):2332-47.
- Mahapatra, D., P.J. Schüffler, J.A.W. Tielbeek, F.M. Vos, and J.M. Buhmann. 2013d. "Localizing and segmenting Crohn's disease affected regions in abdominal MRI using novel context features." Review of. *Proc. SPIE, Medical Imaging 2013: Image Processing* 8669.
- Mahapatra, D., and Y. Sun. 2012c. "Integrating Segmentation Information for Improved MRF-Based Elastic Image Registration." Review of. *Ieee Transactions on Image Processing* 21 (1):170-83.
- Manjunath, B.S., and W.Y. Ma. 1996. "Texture features for browsing and retrieval of image data." Review of. *Ieee Transactions on Pattern Analysis and Machine Intelligence* 18 (8):837-42.
- Mary, J.Y., and R. Modigliani. 1989. "Development and validation of an endoscopic index of the severity for Crohn's disease: a prospective multicentre study. Groupe d'Etudes Therapeutiques des Affections Inflammatoires du Tube Digestif (GETAID)." Review of. *Gut* 30 (7):983-9.
- Masoumi, H., A. Behrad, M.A. Pourmina, and A. Roosta. 2012. "Automatic liver segmentation in MRI images using an iterative watershed algorithm and artificial neural network." Review of. *Biomedical Signal Processing and Control* 7 (5):429-37.
-

- Matlab. 2010. *version 7.10.0 (R2010a)*: The MathWorks Inc.
- Meyer, F. 1994. "Topographic Distance and Watershed Lines." Review of. *Signal Processing* 38 (1):113-25.
- Meyer, S., T.J. Fuchs, A.K. Bosserhoff, F. Hofstadter, A. Pauer, V. Roth, J.M. Buhmann, et al. 2012. "A Seven-Marker Signature and Clinical Outcome in Malignant Melanoma: A Large-Scale Tissue-Microarray Study with Two Independent Patient Cohorts." Review of. *PLoS One* 7 (6).
- Meyer, S., P.J. Wild, H. Moch, V. Roth, T. Fuchs, M. Landthaler, and T. Vogt. 2010. "Patient survival, tumor stage and depth of malignant melanoma are associated with distinct expression profiles: A high-throughput-tissue microarray-based study." Review of. *Experimental Dermatology* 19 (2):210-.
- Moguerza, J.M., A. Munoz, and I.M. de Diego. 2004. "Improving support vector classification via the combination of multiple sources of information." Review of. *Structural, Syntactic, and Statistical Pattern Recognition, Proceedings* 3138:592-600.
- Molodecky, N.A., I.S. Soon, D.M. Rabi, W.A. Ghali, M. Ferris, G. Chernoff, E.I. Benchimol, et al. 2012. "Increasing Incidence and Prevalence of the Inflammatory Bowel Diseases With Time, Based on Systematic Review." Review of. *Gastroenterology* 142 (1):46-54.
- Myren, J., I.A. Bouchier, G. Watkinson, A. Softley, S.E. Clamp, and F.T. de Dombal. 1984. "The O.M.G.E. Multinational Inflammatory Bowel Disease Survey 1976-1982. A further report on 2,657 cases." Review of. *Scand J Gastroenterol Suppl* 95:1-27.
- Ng, S.C., S. Woodrow, N. Patel, J. Subhani, and M. Harbord. 2012. "Role of genetic and environmental factors in British twins with inflammatory bowel disease." Review of. *Inflamm Bowel Dis* 18 (4):725-36.
- Okabe, A., B. Boots, K. Sugihara, and S.N. Chiu. 2009. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*: Wiley.
- Pawlowski, S.W., C.A. Warren, and R. Guerrant. 2009. "Diagnosis and Treatment of Acute or Persistent Diarrhea." Review of. *Gastroenterology* 136 (6):1874-86.
- Petitjean, C., and J.N. Dacher. 2011. "A review of segmentation methods in short axis cardiac MR images." Review of. *Medical Image Analysis* 15 (2):169-84.
- Petrou, M., V.A. Kovalev, and J.R. Reichenbach. 2006. "Three-dimensional nonlinear invisible boundary detection." Review of. *IEEE Trans Image Process* 15 (10):3020-32.
- Rakotomamonjy, A., F.R. Bach, S. Canu, and Y. Grandvalet. 2008. "SimpleMKL." Review of. *Journal of Machine Learning Research* 9:2491-521.

- 
- Rimola, J., I. Ordas, S. Rodriguez, O. Garcia-Bosch, M. Aceituno, J. Llach, C. Ayuso, et al. 2011. "Magnetic resonance imaging for evaluation of Crohn's disease: validation of parameters of severity and quantitative index of activity." Review of. *Inflamm Bowel Dis* 17 (8):1759-68.
- Rimola, J., S. Rodriguez, O. Garcia-Bosch, I. Ordas, E. Ayala, M. Aceituno, M. Pellise, et al. 2009. "Magnetic resonance for assessment of disease activity and severity in ileocolonic Crohn's disease." Review of. *Gut* 58 (8):1113-20.
- Rubner, Y., C. Tomasi, and L.J. Guibas. 2000. "The Earth Mover's Distance as a metric for image retrieval." Review of. *International Journal of Computer Vision* 40 (2):99-121.
- Ruifrok, A.C., and D.A. Johnston. 2001. "Quantification of histochemical staining by color deconvolution." Review of. *Analytical and Quantitative Cytology and Histology* 23 (4):291-9.
- Ruta, D., and B. Gabrys. 2005. "Classifier selection for majority voting." Review of. *Information Fusion* 6 (1):63-81.
- Rutgeerts, P., K. Geboes, G. Vantrappen, J. Beyls, R. Kerremans, and M. Hiele. 1990. "Predictability of the postoperative course of Crohn's disease." Review of. *Gastroenterology* 99 (4):956-63.
- Satsangi, J., M.S. Silverberg, S. Vermeire, and J.F. Colombel. 2006. "The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications." Review of. *Gut* 55 (6):749-53.
- Schaffter, T., V. Rasche, and I.C. Carlsen. 1999. "Motion compensated projection reconstruction." Review of. *Magn Reson Med* 41 (5):954-63.
- Schindelin, J., I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, et al. 2012. "Fiji: an open-source platform for biological-image analysis." Review of. *Nat Methods* 9 (7):676-82.
- Schölkopf, B., and A.J. Smola. 2002. "A short introduction to learning with kernels." Review of. *Advanced Lectures on Machine Learning* 2600:41-64.
- Scholzen, T., and J. Gerdes. 2000. "The Ki-67 protein: from the known and the unknown." Review of. *J Cell Physiol* 182 (3):311-22.
- Schüffler, P., A. Ulaş, U. Castellani, and V. Murino. 2011. "A Multiple Kernel Learning Algorithm for Cell Nucleus Classification of Renal Cell Carcinoma." In *Image Analysis and Processing – ICIAP*, edited by Giuseppe Maino and GianLuca Foresti, 413-22. Springer Berlin Heidelberg.
- Schuffler, P.J., T.J. Fuchs, C.S. Ong, V. Roth, and J.M. Buhmann. 2010. "Computational TMA Analysis and Cell Nucleus Classification of Renal Cell Carcinoma." Review of. *Pattern Recognition* 6376:202-11.
-

- Schüffler, P.J., T.J. Fuchs, C.S. Ong, V. Roth, and J.M. Buhmann. 2010. "Computational TMA analysis and cell nucleus classification of renal cell carcinoma." In *Proceedings of the 32nd DAGM conference on Pattern recognition*, 202-11. Darmstadt, Germany: Springer-Verlag Berlin.
- Schüffler, P.J., T.J. Fuchs, C.S. Ong, V. Roth, and J.M. Buhmann. 2013a. "Automated Analysis of Tissue Micro-Array Images on the Example of Renal Cell Carcinoma." In *Similarity-Based Pattern Analysis and Recognition*, edited by Marcello Pelillo, 219-45. Springer London.
- Schüffler, P.J., T.J. Fuchs, C.S. Ong, P.J. Wild, N.J. Rupp, and J.M. Buhmann. 2013b. "TMARKER: A free software toolkit for histopathological cell counting and staining estimation." Review of. *J Pathol Inform* 4 (2).
- Schüffler, P.J., D. Mahapatra, J.A.W. Tielbeek, F.M. Vos, J. Makanyanga, D.A. Pendsé, C.Y. Nio, et al. 2013c. "A Model Development Pipeline for Crohn's Disease Severity Assessment from Magnetic Resonance Images." In *Abdominal Imaging. Computation and Clinical Applications*, edited by Hiroyuki Yoshida, Simon Warfield and Michael Vannier, 1-10. Springer Berlin Heidelberg.
- Schüffler, P.J., N.J. Rupp, C.S. Ong, J.M. Buhmann, T.J. Fuchs, and P.J. Wild. 2013d. "TMARKER: a robust and free software toolkit for histopathological cell counting and immunohistochemical staining estimation." Review of. *Pathologie* 34:30.
- Seksik, P., I. Nion-Larmurier, H. Sokol, L. Beaugerie, and J. Cosnes. 2009. "Effects of Light Smoking Consumption on the Clinical Course of Crohn's Disease." Review of. *Inflamm Bowel Dis* 15 (5):734-41.
- Semelka, R.C., J.P. Shoenuit, R. Silverman, M.A. Kroeker, C.S. Yaffe, and A.B. Micflikier. 1991. "Bowel disease: prospective comparison of CT and 1.5-T pre- and postcontrast MR imaging with T1-weighted fat-suppressed and breath-hold FLASH sequences." Review of. *J Magn Reson Imaging* 1 (6):625-32.
- Silverberg, M.S., J. Satsangi, T. Ahmad, I.D.R. Arnott, C.N. Bernstein, S.R. Brant, R. Caprilli, et al. 2005. "Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: Report of a Working Party of the 2005 Montreal World Congress of Gastroenterology." Review of. *Canadian Journal of Gastroenterology* 19:5A-36A.
- Soldini, D., C. Montagna, P. Schüffler, V. Martin, A. Georgis, T. Thiesler, A. Curioni-Fontecedro, et al. 2013. "A new diagnostic algorithm for Burkitt and diffuse large B-cell lymphomas based on the expression of CSE1L and STAT3 and on MYC rearrangement predicts outcome." Review of. *Annals of Oncology* 24 (1):193-201.



- Sostegni, R., M. Daperno, N. Scaglione, A. Lavagna, R. Rocca, and A. Pera. 2003. "Review article: Crohn's disease: monitoring disease activity." Review of. *Aliment Pharmacol Ther* 17 Suppl 2:11-7.
- Spehlmann, M.E., A.Z. Begun, J. Burghardt, P. Lepage, A. Raedler, and S. Schreiber. 2008. "Epidemiology of inflammatory bowel disease in a German twin cohort: Results of a nationwide study." Review of. *Inflamm Bowel Dis* 14 (7):968-76.
- Steward, M.J., S. Punwani, I. Proctor, Y. Adjei-Gyamfi, F. Chatterjee, S. Bloom, M. Novelli, et al. 2012. "Non-perforating small bowel Crohn's disease assessed by MRI enterography: derivation and histopathological validation of an MR-based activity index." Review of. *Eur J Radiol* 81 (9):2080-8.
- Suenaert, P., V. Bulteel, L. Lemmens, M. Noman, B. Geypens, G. Van Assche, K. Geboes, et al. 2002. "Anti-tumor necrosis factor treatment restores the gut barrier in Crohn's disease." Review of. *Am J Gastroenterol* 97 (8):2000-4.
- Talley, N.J., M.T. Abreu, J.P. Achkar, C.N. Bernstein, M.C. Dubinsky, S.B. Hanauer, S.V. Kane, et al. 2011. "An Evidence-Based Systematic Review on Medical Therapies for Inflammatory Bowel Disease." Review of. *American Journal of Gastroenterology* 106:S2-S25.
- Tannapfel, A., H.A. Hahn, A. Katalinic, R.J. Fietkau, R. Kuhn, and C.W. Wittekind. 1996. "Prognostic value of ploidy and proliferation markers in renal cell carcinoma." Review of. *Cancer* 77 (1):164-71.
- Team, R.D.C. "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing. <http://www.R-project.org>.
- Theurillat, J.P., U. Zurrer-Hardi, Z. Varga, M. Storz, N.M. Probst-Hensch, B. Seifert, M.K. Fehr, et al. 2007. "NY-BR-1 protein expression in breast carcinoma: a mammary gland differentiation antigen as target for cancer immunotherapy." Review of. *Cancer Immunology Immunotherapy* 56 (11):1723-31.
- Tibble, J.A., G. Sigthorsson, S. Bridger, M.K. Fagerhol, and I. Bjarnason. 2000. "Surrogate markers of intestinal inflammation are predictive of relapse in patients with inflammatory bowel disease." Review of. *Gastroenterology* 119 (1):15-22.
- Tielbeek, J.A.W., J.C. Makanyanga, S. Bipat, D.A. Pendse, C.Y. Nio, F.M. Vos, S.A. Taylor, et al. 2013. "Grading Crohn Disease Activity With MRI: Interobserver Variability of MRI Features, MRI Scoring of Severity, and Correlation With Crohn Disease Endoscopic Index of Severity." Review of. *American Journal of Roentgenology* 201 (6):1220-8.

- Tielbeek, J.A.W., F.M. Vos, and J. Stoker. 2012. "A computer-assisted model for detection of MRI signs of Crohn's disease activity: future or fiction?" Review of. *Abdom Imaging* 37 (6):967-73.
- Tsurui, H., H. Nishimura, S. Hattori, S. Hirose, K. Okumura, and T. Shirai. 2000. "Seven-color fluorescence imaging of tissue samples based on Fourier spectroscopy and singular value decomposition." Review of. *Journal of Histochemistry & Cytochemistry* 48 (5):653-62.
- Tu, Z.W., and X.A. Bai. 2010. "Auto-Context and Its Application to High-Level Vision Tasks and 3D Brain Image Segmentation." Review of. *Ieee Transactions on Pattern Analysis and Machine Intelligence* 32 (10):1744-57.
- Ulaş, A., M. Semerci, O.T. Yıldız, and E. Alpaydın. 2009. "Incremental construction of classifier and discriminant ensembles." Review of. *Information Sciences* 179 (9):1298-318.
- Van Assche, G., A. Dignass, B. Bokemeyer, S. Danese, P. Gionchetti, G. Moser, L. Beaugerie, et al. 2013. "Second European evidence-based consensus on the diagnosis and management of ulcerative colitis Part 3: Special situations." Review of. *Journal of Crohns & Colitis* 7 (1):1-33.
- van Hees, P.A., P.H. van Elteren, H.J. van Lier, and J.H. van Tongeren. 1980. "An index of inflammatory activity in patients with Crohn's disease." Review of. *Gut* 21 (4):279-86.
- Vos, F.M., J.A.W. Tielbeek, R.E. Naziroglu, Z. Li, P.J. Schüffler, D. Mahapatra, A. Wiebel, et al. 2012. Computational modeling for assessment of IBD: To be or not to be? Paper presented at the Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, Aug. 28 2012-Sept. 1 2012.
- Wang, H.A.O., D. Grolimund, C. Giesen, C.N. Borca, J.R.H. Shaw-Stewart, B. Bodenmiller, and D. Gunther. 2013. "Fast Chemical Imaging at High Spatial Resolution by Laser Ablation Inductively Coupled Plasma Mass Spectrometry." Review of. *Analytical Chemistry* 85 (21):10107-16.
- Wright, J.P., I.N. Marks, and A. Parfitt. 1985. "A simple clinical index of Crohn's disease activity--the Cape Town index." Review of. *S Afr Med J* 68 (7):502-3.
- Xu, Z., R. Jin, H. Yang, I. King, and M.R. Lyu. 2010. Simple and Efficient Multiple Kernel Learning by Group Lasso. Paper presented at the 27th International Conference on Machine Learning.
- Yankeelov, T.E., and J.C. Gore. 2009. "Dynamic Contrast Enhanced Magnetic Resonance Imaging in Oncology: Theory, Data Acquisition, Analysis, and Examples." Review of. *Curr Med Imaging Rev* 3 (2):91-107.

- Ziech, M.L., C. Lavini, M.W. Caan, C.Y. Nio, P.C. Stokkers, S. Bipat, C.Y. Ponsioen, et al. 2012a. "Dynamic contrast-enhanced MRI in patients with luminal Crohn's disease." Review of. *Eur J Radiol* 81 (11):3019-27.
- Ziech, M.L.W., P.M.M. Bossuyt, A. Laghi, T.C. Lauenstein, S.A. Taylor, and J. Stoker. 2012b. "Grading luminal Crohn's disease: Which MRI features are considered as important?" Review of. *Eur J Radiol* 81 (4):E467-E72.
- Zollner, F.G., E. Svarstad, A.Z. Munthe-Kaas, L.R. Schad, A. Lundervold, and J. Rorvik. 2012. "Assessment of Kidney Volumes From MRI: Acquisition and Segmentation Techniques." Review of. *American Journal of Roentgenology* 199 (5):1060-9.



## ABBREVIATIONS

ABWT	Automated Bowel Wall Thickness Measurement
AC1	Agreement Coefficient 1
AIS	Transmural Histopathological Scoring of Acute Inflammation
AMC	Academic Medical Center
CD	Crohn's Disease
CDAI	Crohn's Disease Activity Index
CDEIS	Crohn's Disease Endoscopic Index of Severity
DCE	Dynamic Contrast Enhancement
DM	Dice Metric
eAIS	Endoscopic biopsy Acute Inflammatory Score
e.g.	Example given
ETH	Eidgenössische Technische Hochschule
HD	Hausdorff Distance
IBD	Inflammatory Bowel Disease
ICC	Intra-Class-Statistics
IT	Information Theory
LOPO-CV	Leave-One-Patient-Out Cross-Validation
MaRIA	Magnetic Resonance Index of Activity
MKL	Multiple Kernel Learning
MRE	Magnetic Resonance Enterography
MRI	Magnetic Resonance Imaging
PCa	Prostate Cancer
pLSA	Probabilistic Latent Semantic Analysis
pP	Per Patient
RCC	Renal Cell Carcinoma
RCE	Relative Contrast Enhancement
ROI	Region Of Interest
SD	Standard Deviation
SLIC	Simple Linear Iterative Clustering
SMO	Sequential Minimal Optimization
SPGE	Spoiled Gradient Echo, a MRI technique
SSFSE	Single Shot Fast Spin Echo, a MRI technique
THRIVE	T1-weighted high-resolution isotropic volume examination
TIC	Time Intensity Curve
TMA	Tissue Micro Array
WSI	Wall Signal Intensity
wt	Wall Thickness