Diss. ETH N° 19469

Computational Pathology
A Machine Learning Approach

DISSERTATION

submitted to

ETH ZURICH

for the degree of

DOCTOR OF SCIENCES

by

THOMAS JOSEF FUCHS
Dipl.-Ing., Technical University Graz
born April 13^{th}, 1979
citizen of Austria

accepted on the recommendation of

Prof. Dr. Joachim M. Buhmann examiner
Prof. Dr. Peter Bühlmann co-examiner
Prof. Dr. Holger Moch co-examiner
Prof. Dr. Pietro Perona co-examiner

2010

For Barbara

Abstract

Computational Pathology is a novel field comprising aspects of machine learning, computer vision, clinical statistics and general pathology. The principal focus of this thesis is to define this new field, develop and investigate statistical methods which can be combined within an unified framework to answer scientific and clinical questions in pathology.

Computational Pathology

The histological assessment of human tissue has emerged as the key challenge for detection and treatment of cancer. A plethora of different data sources ranging from tissue microarray data to gene expression, proteomics or metabolomics data provide a detailed overview of the health status of a patient. Medical doctors need to assess these information sources and they rely on data driven automatic analysis tools. Methods for classification, grouping and segmentation of heterogeneous data sources as well as regression of noisy dependencies and estimation of survival probabilities enter the processing workflow of a pathology diagnosis system at various stages.

The work presented in this thesis is structured in the following parts.

Data: First, we describe the data which forms the predictors or covariates in the models presented in this thesis. This part comprises the biomedical properties of renal cell carcinoma (RCC), the production of tissue microarrays and finally whole slide scanning of histological specimens. Second, we investigate in detail how the labeling data for modeling are procured, i.e., the lack of ground truth and the difficulties of producing a gold standard. To this end we report on studies which we conducted in collaboration with the University Hospital Zürich to quantify the inter and intra observer variability for three crucial tasks in computational pathology: (i) nuclei detection, (ii) nuclei classification, and (iii) staining estimation. The results underline the need for decision support systems, which not only automate cumbersome tasks in histopathology but also provide an objective view for the challenges at hand.

Data

Imaging **Imaging:** This chapter describes the development from classical image processing to statistical pattern recognition in the domain of computational pathology. Approaches based on mathematical morphology are compared to methods in machine learning. A novel framework for multiple object detection, *Relational Detection Forests*, is introduced and subsequently extended for inter-active and online learning of ensembles. This methodology is successfully employed for micrometastases detection and tissue classification in sentinel lymph nodes and robust pancreatic islet segmentation.

Statistics **Statistics:** For modeling time-to-event data in pathology we describe methods for survival analysis and the detection of subgroups of patients with different survival expectancy. Furthermore it is shown how random forest proximities and clustering can be employed for quality control within a computational pathology pipeline. Finally we developed two statistical learning approaches for (i) linear modeling for detection of urothelial bladder cancer cells and (ii) for learning a signature for clinical outcome prediction in malignant melanoma with implications for the therapy of this most common form of skin cancer.

Holistic View Concluding, the proposed holistic view is exemplified by a computational pathology pipeline for RCC, comprising methodology from all three described areas. This study demonstrates, for the first time, the feasibility of a system which performs equally well as a trained pathologist.

Zusammenfassung

Computational Pathology bezeichnet ein neues Feld an der Schnittstelle von maschinellem Lernen, Bildverarbeitung, klinischer Statistik und Pathologie. Der Schwerpunkt der vorliegenden Arbeit liegt zuallererst auf der Definition dieses neuartigen Feldes. Basierend darauf werden statistische Methoden entwickelt und analysiert, welche, kombiniert in einem einheitlichen System, zur Beantwortung wissenschaftlicher und klinischer Fragestellungen in der Pathologie dienen können.

Computational Pathology

Die histopathologische Analyse von menschlichem Gewebe stellt eine der primären Herausforderung für die Diagnose von kanzerogenen Krankheiten und deren Behandlung dar. Die moderne Pathologie greift dazu auf eine Vielzahl von medizinischen und biologischen Methoden zurück. Dazu zählen v.a. Schnitte von in Paraffinblöcken gesammelten Gewebezylindern, sogenannte *Tissue Microarrays* (TMA), deren rechnergestützte Analyse in dieser Arbeit im Detail beschrieben wird. Weitere Quellen stellen Expressionsdaten von Genen und Proteinen dar, die gemeinsam mit den immunhistochemischen Färbungen von TMAs einen überblick über einzelne Patienten, beziehungsweise ganze Patientenkohorten geben. Methoden zur Klassifikation, zur unüberwachten Gruppierung und zur Segmentierung dieser heterogenen Daten in Verbindung mit Regression auf oft verrauschte abhängige Variablen und das Schätzen der überlebenswahrscheinlichkeit von Patienten sind wichtige Bestandteile eines Diagnosesystems in der Pathologie.

Die wissenschaftlichen Beiträge dieser Arbeit gliedern sich in die folgenden drei Abschnitte:

Datenlage: Zuerst werden die Daten beschrieben, welche die unabhängigen Variablen in der Modellierung darstellen. Dazu gehört die Beschreibung klarzelliger Blasenkarzinome, die Herstellung von TMAs und abschließend die Digitalisierung ganzer histologischer Schnitte. Im Weiteren wird die überwachte Information, die in die abhängigen Variablen eingeht, detailliert untersucht. Das

Datenlage

Fehlen, bzw. die Unmöglichkeit der Messung eines objektiven Standards stellt eine der größten Herausforderungen in diesem Bereich dar. Zu diesem Zweck werden Ergebnisse von Studien vorgestellt, die wir am Universitätsspital Zürich durchgeführt haben, um den inter- und intrapathologen Fehler für drei der wichtigsten Aufgaben in dem Bereich zu analysieren: (i) die Detektion von Zellkernen, (ii) deren Klassifizierung in normal und kanzerogen, sowie (iii) das Schätzen der immunohistochemischen Färbung. Die Ergebnisse unterstreichen den Bedarf an einem System, das nicht nur mühsame Tätigkeiten im klinischen Alltag automatisiert, sondern auch objektive Schätzungen für wissenschaftliche Fragestellungen liefert.

Bildverarbeitung

Bildverarbeitung: In diesem Kapitel stellen wir die Entwicklung von klassischer Bildverarbeitung zu moderner, statistischer Mustererkennung im Bereich der Pathologie dar. Methoden basierend auf mathematischer Morphologie werden mit Ansätzen aus dem Bereich des maschinellen Lernens verglichen. Wir stellen ein neuartiges System zur Objektdetektion vor (*Relational Detection Forests*) und zeigen dessen Erweiterung für das interaktive Lernen von Klassifizierungs-Ensembles. Die vorgestellten Methoden werden dann dafür eingesetzt, um Gewebe zu klassifizieren, Micrometastasen in Lymphknoten zu finden und auf robuste Art und Weise Langerhans'che Inseln in der Bauchspeicheldrüse zu segmentieren.

Statistik

Statistik:Für die Modellierung zensierter überlebens-Daten in der Pathologie werden in dieser Arbeit Modelle vorgestellt, die dann dazu verwendet werden können, um Gruppen von Patienten zu beschreiben, die ein erhöhtes Risiko haben, an der Krebserkrankung zu sterben. Des Weiteren beschreiben wir die Verwendung von unüberwachter Gruppierung, basierend auf Nachbarschaftsbeziehung in Ensembles von randomisierten Entscheidungsbäumen, um die Qualität in einem pathologischen Analysesystem sicherzustellen. Zuletzt werden zwei Algorithmen aus dem Bereich der Statistik beschrieben, um (i) ein lineares Modell zur Erkennung von Harnblasenkrebszellen zu lernen und (ii) um eine Protein-Signatur zu konstruieren, mit deren Hilfe Patienten mit malignem Melanom in Risikogruppen eingeteilt werden können. Das vorgestellte Resultat hat auch Implikationen für zukünftige Behandlung von derartigen Krebspatienten.

Vereinheitlichte Sichtweise

Abschließend beschreiben wir ein System für die Analyse von histologischen Schnitten von klarzelligem Harnblasenkrebs, das Methoden aus den oben beschriebenen drei Bereichen in einheitlicher Weise zusammenführt. Wir konnten in dieser Studie, zum ersten Mal die grundsätzliche Möglichkeit aufzeigen, ein System zu entwerfen, das die medizinische Aufgabenstellung ähnlich gut wie ein ausgebildeter Pathologe durchführen kann.

Acknowledgement

It is a pleasure to thank the many people who made this thesis possible.

Joachim Buhmann

I owe my deepest gratitude to my Ph.D. advisor Prof. Joachim Buhmann. Without his sound advice and continuing support this thesis would not have been possible. I am especially grateful for all the opportunities Prof. Buhmann opened up not only by providing scientific ideas but also by encouraging interdisciplinary cooperation with researchers in diverse fields from all over the world. Prof. Buhmann always succeeded to guarantee a productive and cooperative environment for scientific research. One big part of that is his open door policy which allowed me to get advice whenever I needed it regardless of the time of day. Furthermore I want to highlight his willingness to not only listen to but to encourage and support extraordinary ideas which led to so many novel result and application of machine learning which form the basis of this work. I consider myself very lucky that I got the opportunity to accomplish this thesis under his guidance.

Holger Moch

I would like to thank Prof. Holger Moch for an extraordinary good collaboration with him personally and his institute of surgical pathology at the University Hospital Zürich in general. Without his vision and open mindedness the field of computational pathology would not exist today. His continuing support not only with human and technical resources but foremost with his expertise in pathology enabled the success of so many joint projects. The access to domain experts, microscopes, scanners and clinical practice, which Prof. Moch granted, was key for developing and implementing solutions which are not only scientifically interesting but which are of measurable benefit for pathologists and patients.

Volker Roth

My utmost gratitude goes to Volker Roth who is not only a very good friend but who guided me into the depth of statistical learning. With his clarity of thinking and his precise scientific questioning Volker will always be a role model for me.

Peter Wild — It is difficult to overstate my gratitude to Peter Wild, who taught me the best part I know about pathology. Looking through a microscope he patiently explained to me the differences between normal an abnormal nuclei on numerous cancer types. Peter spend hours and hours labeling tissue images without which a machine learning approach would simply not have been possible. I will always remember the long nights we spend in the university tavern pondering over cancer pathways and survival statistics which fruitfully yielded so many publications.

Johannes Haybäck — I sincerely thank Johannes Haybäck for the years of excellent collaboration. His knowledge and enthusiasm about pathology was enormously motivating. In Zürich and in Graz he took the time to explain to me the most sophisticated mouse models and he willingly annotated hundreds of images which lead to successful research projects on prions, liver cancer and randomized tree ensembles.

Stefanie Meyer — I would like to show my gratitude to Stefanie Meyer who was and is an excellent comrade in exploring the depth of skin cancer. Our research on predictive signatures for melanoma was one of the longest and most elaborate projects during my thesis and without her continuing support it would not have been nearly as successful.

Monika Bieri
Norbert Wey — I am indebted to Monika Bieri and Norbert Wey who prepared and scanned thousands of histological slides. The terabytes of data produced is the foundation of all our research in computational pathology. Only their vital engagement made it possible to integrate our research results in the clinical pipeline and is they who developed the framework which put these results to practical use in the clinic.

Xenofon Floros — Special thanks to Xenofon Floros for being an outstanding office mate, a prolific coauthor and one of my best friends. Fontas' exquisite Greek style café frappé was an indispensable pillar of our work.

Julia Vogt
Sudhir Raman — I am very thankful to Julia Vogt and Sudhir Raman for their exceptional work on Bayesian clustering and regression. It was an immense pleasure to work with them and I am very much looking forward to continue the collaboration in the years to come.

Cheng Soon Ong — It is an honor for me to thank Cheng Soon Ong. His profound knowledge of machine learning and his readiness to help whenever needed was of paramount importance for my work.

Rita Klute — I want to express my warmest gratitude to Rita Klute. With her kindness and

helpfulness she handled frictionless all administrative aspects of our work. Without her organizing talent the amount of work would not have been manageable.

I wish to thank Sara Abbasabadi and Georg Troxler whose master theses I had the pleasure to supervise. Both were exceptional students and the work of Sara together with the commitment of Daniela Mihic resulted in an impressive framework for micrometastasis detection which contributed to this work.

Sara Abbasabadi
Georg Troxler

I am immensely grateful to Verena Kaynig who was not only a superb colleague, a productive coauthor but first of all a good friend during our years of study.

Verena Kaynig

Philipp Fürnstahl was not only a formidable comrade during high school, at the university in Graz and during our PhD studies at ETH, he was and is a good friend. I owe him my deepest gratitude for his help and encouragement during good times and tough times.

Philipp Fürnstahl

It is my pleasure to thank Stefan Saur who was an outstanding collaborator and an excellent friend since our time in Princeton.

Stefan Saur

Special thanks to Christian Müller and Manfred Claassen who suffered and celebrated with me through all the years and to whom I will always be indebted for making the time so much fun.

Christian Müller
Manfred Claassen

Tilman Lange and Peter Orbanz were my dearest brothers in crime from the beginning of my PhD studies. I am very grateful for Peter's profound introduction into nonparametric Bayesian learning and I'm deep in debt for the entertaining evenings spend at Hot Pasta with Tilman discussing all good and bad aspects of academic life.

Tilman Lange
Peter Orbanz

My fellow PhD students have been a rich source of motivation and I thank all of them. In particular I thank Yvonne Moh for her moral support , Ludwig Busse for all the entertaining conversations, Mario Frank for all the entrepreneurial discussions and Peter Schüffler for contributing so much to the work on computational pathology.

Yvonne Moh
Ludwig Busse
Mario Frank
Peter Schüffler

I am very grateful to my co-examiners, Prof. Peter Bühlmann and Prof. Pietro Perona for reviewing my thesis. I feel honored by their interest in my work.

Peter Bühlmann
Pietro Perona

Finally, I am forever indebted to my family for their understanding, endless patience and support when it was most required. Most important of all, I thank my wife Barbara. Her love, encouragement and tolerance made possible everything. Thank you!

Family

Contents

CHAPTER 1

Introduction to Computational Pathology: The Systems View

Contents

1.1 Introduction

Modern pathology studies of biopsy tissue encompass multiple stainings of histological material, genomics and proteomics analyses as well as comparative statistical analyses of patient data. Pathology lays not only a scientific foundation for clinical medicine but also serves as a bridge between the fundamental sciences in natural science, medicine and patient care. Therefore, it can be viewed as one of the key hubs for translational research in the health and life sciences, subsequently facilitating translational medicine. In particular, the abundance of heterogeneous data sources with a substantial amount of randomness

and noise poses challenging problems for statistics and machine learning. Automatic processing of this wealth of data promises a standardized and hopefully more objective diagnosis of the disease state of a patient than manual inspection can provide today. An automatic computational pathology pipeline also enables the medical user to quantitatively benchmark the processing pipeline and to identify error sensitive processing steps which can substantially degrade the final prediction of survival times.

1.2 From Pixels to Proteins

The histological assessment of human tissue is the key point for the detection and the treatment of cancer. Furthermore it is an inevitable step for medical, biochemical and pharmaceutical research. At one point every biomarker which was identified by genomic, transcriptomic or proteomic research has to demonstrate its effect on human tissue.
Up until now this assessment is done manually by medical doctors with a microscope. Therefore pathologists use either whole microscope slides or tissue microarrays (TMA) which contain tissue samples from hundreds of patients. This manual procedure comprised largely tedious annotation work and it is prone to error. Studies presented in Section 2.3 show that the variability between pathologists for the estimation of stained cancerous cell nuclei can be as high as 20%. This discrepancy is even more critical taking into account, that a difference of a few percent can decide if a patient receives a chemotherapy or not.
Computer science faces the challenge to develop algorithms for large scale systems which are able to automatically and objectively analyze human tissue.

1.3 Problem Definition

For achieving such an ambitious goal a number of problems have to be solved, ranging from statistical pattern recognition over survival analysis to clinical applicability.

1.3.1 Image Understanding

For domain expert it takes years of training to detect, segment and recognize the various structures within human tissue, for example vessels, membranes and different types of cell nuclei. These problems have to be addressed by means of machine learning and computer vision. Current systems are only able to detect for example faces in photographs or single cells in cytology where they are easily detectable on a homogeneous background. The difficulties mainly arise from the following peculiarities of histopathological image processing: (a) The

objects which can be found are not only benign or malignant renal nuclei, but the tissue also comprises endothelial cells, lymphocytes, erythrocytes etc.. Endothelial cells for example are extremely elongated where on the other hand lymphocytes are nearly perfectly round and homogeneous. Neither of these cells must be counted in the estimation process. (b) A grave difficulty is the fact that nuclei as three dimensional structures are not always perfectly cut in their maximum dimension producing numerous cutting artifacts. For many of these artifacts it is not possible to determine if they are atypical or benign. (c) Variations in the production process of TMAs can lead to areas of different thickness within one section. This preprocessing artifact produces blurred regions in the image during the scanning process. The analysis and the understanding of complex tissue is just at its beginning.

1.3.2 Data Management

The great success of histology in general and tissue microarrays (TMA) in particular is closely related to experimental techniques which yield localized protein expression values in tissue. High resolution scanning technologies in recent years renders an automated analysis of histological slices feasible. The problem it poses is the enormous amount of data which is generated. It outnumbers classical medical imaging problems like CT and MR by a factor of 50 to 100. Therefore computer science has to provide systems which are able to store and handle these tera- or even petabytes of data.

1.3.3 Dealing with Uncertainty

One of the largest problems in pathology is the absence of an objective ground truth for the characterization of tissue. Therefore models have to be developed which are able to aggregate the knowledge of a cohort of domain experts. The expert information has to be gathered in various stages of the modeling process: Starting with object detection, followed by segmentation, classification and finally the estimation of the protein staining. Such a gold standard proved to be indispensable for training the algorithms in the framework and testing their performance.

1.3.4 Survival Statistics

An advance in survival statistics is of utmost importance to handle the data generated in these high throughput scenarios. Besides the aforementioned gold standard the survival time of patients is the only target which can be used to judge the performance of a biomarker. Classical methods like Cox Regression cannot deal with thousands or millions of features generated from an automated analysis system. Therefore grouping and regularization techniques have to be used to quantify the hazard ratio of the various covariates. In addition

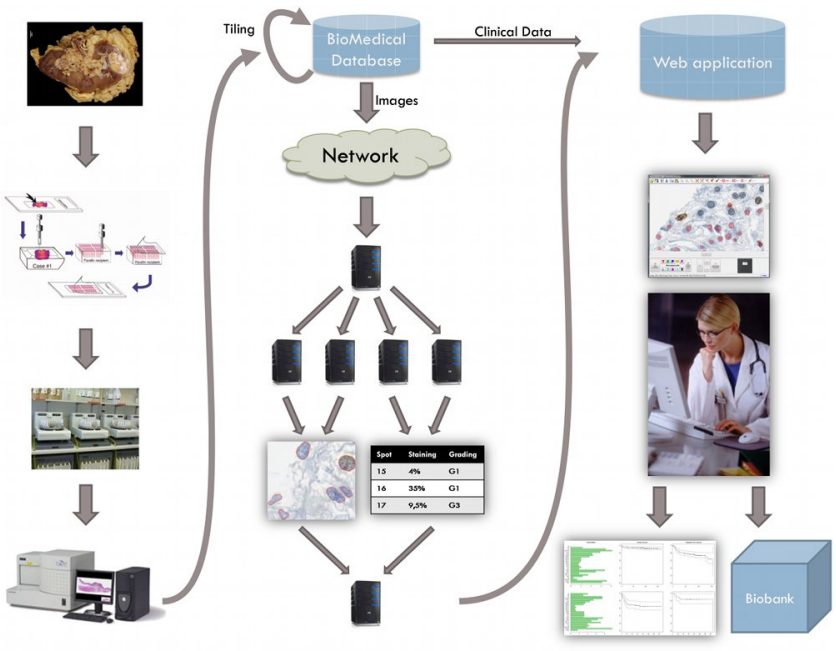


Figure 1.1
Schematic of a computational pathology workflow. First, the tissue is prepared in the hospital and microscopy slides are stained and scanned. The images are stored in a central database and then distributed on the machines of a computer cluster for automated analysis. The results are aggregated and presented to the pathologist by means of web applications.

specialized clustering methods can be used to find subgroups of patients with different survival expectation.

1.3.5 From Cluster to Clinic

The final challenge is the integration into clinical practice. A computational pathology system has to master all steps in the clinical pipeline (cf. Figure 1.1): Starting from patients' tissue, samples are either collected in tissue microarrays (TMA) or processed on whole microscope slides. The slides are stained with antibodies for the proteins of interest and subsequently digitized with a high resolution scanner for further processing. Due to the huge amount of imaging data, the analysis itself has to be performed in a distributed fashion on a cluster. The results from the imaging step are aggregated and used for survival statistics. The final estimations of the algorithms are presented to the medical doctors, together with the annotated images. Taking this plethora of information into account the pathologist has to decide on the final diagnosis.

1.4 Scientific Challenges

In order to tackle these challenges computer science has to develop solutions in various fields: Machine learning and computer vision algorithms have to be designed to analyze, characterize and understand the histopathological images. Databases, web applications and distributed systems are needed to store

and share the enormous amount of data. Parallel algorithms have to be implemented on a cluster infrastructure to analyze samples for hundreds of patients in reasonable time. Finally survival statistics has to be employed to find prognostic markers for the survival of patients.

1.5 Social Benefit

The social benefit of a solution of this challenge is twofold: First, every single cancer patient would benefit from an automated processing which would provide his doctor with an objective analysis of the tissue and therefore of the illness. A more precise diagnosis can lead to a better and more targeted application of the treatment. Second, biomedical science would enormously benefit from a consistent and high throughput analysis of tissue microarrays of hundreds of patients. Such a consistent data analysis process would advance the whole field of pathology from a qualitative to a quantitative science. An objective and standardized quantification of prognostic and predictive biomarkers promises a more detailed understanding of cancer and its causes. The results hopefully give rise to the discovery of new drugs and less invasive but more effective treatment procedures.

1.6 Interdisciplinary Endeavor

This multitude of problems can only be solved by a collaboration of scientists from different fields. On one hand it needs cooperation and mutual understanding of computer scientists and medical doctors. On the other hand, scientist from different branches of computer science have to work together frictionlessly. Such a collaborative effort has to comprise researchers in machine learning, statisticians, database engineers and experts for computer clusters and distributed systems.

1.7 Definition of Computational Pathology

Computational pathology as well as the medical discipline pathology is a wide and diverse field which encompass scientific research as well as day-to-day work in medical clinics. The following definition is an attempt for a concise and practical description of this novel field:

"Computational Pathology" A Definition

> Computational Pathology investigates a complete probabilistic treatment of scientific and clinical workflows in general pathology, i.e. it combines experimental design, statistical pattern recognition and survival analysis within a unified framework to answer scientific and clinical questions in pathology.

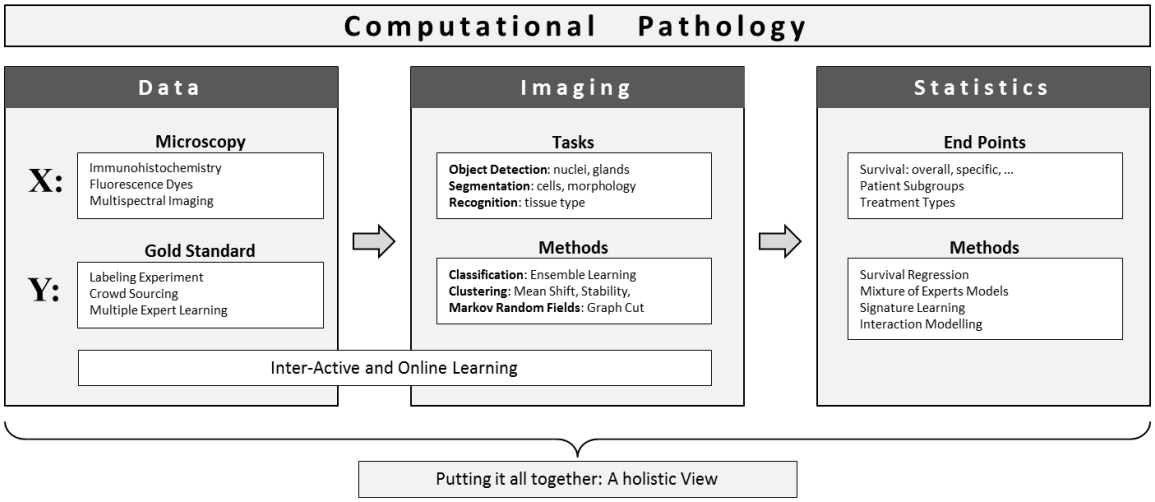


Figure 1.2

Schematic overview of the workflow in computational pathology comprising of three major parts: (i) the covariate data X is acquired via microscopy and the target data Y is generated in extensive labeling experiments; (ii) image analysis in terms of nuclei detection, cell segmentation or texture classification provides detailed information about the tissue; (iii) medical statistics, i.e. survival regression and mixture of expert models are used to investigate the clinical end point of interest using data from the previous two parts. The aim is to build a complete probabilistic workflow comprising all three parts.

Figure 1.2 depicts a schematic overview of the field and the three major parts it consists of: data generation, image analysis and medical statistics.

1.8 Original Contributions

Problem Formulation: The intellectual contribution of highest general interest is in the exact formulation of the problem. Although histopathological whole-slide imaging currently emerges as a new field, due to the just recent advent of powerful scanning technology, the methods employed by the community are predominantly based on classical image processing with all its deficits (cf. Section 3). We too, first tried to solve the nuclei detection and staining estimation problem by devising algorithms based on mathematical morphology (cf. Section 3.2), before realizing that an exact segmentation of nuclei boundaries is not mandatory for survival estimation at the end. The development of statistical learning methodology (cf. Section 3.5.3) tailored to the requirements of computational pathology was a big leap forward. Formulating the problem in a concise way, encompassing not only statistical pattern recognition but also survival statistics as the endpoint of the pipeline, paved the way for the development of combined systems, which solve problems in pathology with a holistic view in mind (cf. Section 5).

Analyzing Pathologists: To the best of our knowledge this study presents the first in-depth analysis of the detection and classification performance of trained pathologists on subcellular level. We conducted three studies in controlled environment with up to 14 trained pathologists per task. The studies were carried out at the University Hospital Zürich with dedicated software application for tablet-PC, which we developed for these special tasks. This thorough analysis is crucial for understanding the difficulties in this domain and for generating a gold standard which is the basis for training and evaluation in statistical learning. Section 2.3

Relational Detection Forests: We present a novel classification algorithm for object detection based on randomized trees. The insight is utilized that randomly selecting predictors from a enormously large pool of features (in the order of 10^{13}) leads to less correlation between the trees of the ensemble and hence decreases the variance of the combined forest. To solve this problem we propose to sample relational features at each node, which have the benefit that no threshold has to be fitted and that they are illumination invariant. We advocate a class-wise proportional sampling for tree induction in highly unbalanced settings. Section 3.5.3

Voronoi Sampling: In most object detection scenarios the expert labeling of the data consists only of the locations of the object in question. Before training a classifier, patches of the negative class have to be sampled from the background. This sampling is by far not trivial when multiple objects are densely packed in the image, e.g., in the case of cell nuclei detection. We Section 3.4

devised a novel sampling scheme for such problems, in which the background class is sampled from the vertices of a Voronoi tessellation, built on basis of the locations of the multiple foreground objects.

Section 3.6 *Inter-Active Ensemble Learning:* We extend the relational detection forest for interactive online learning. Therefore we take advantage of the fact that the proposed relational features are extremely fast to evaluate and hence trees can be induced and added to the ensemble in real time. Besides the experimental results and the comparison of various online update schemes applied to tree ensembles we contribute a software application which is used in practice by pathologists and biologists to inter-actively train object detectors to answer scientific questions.

Section 3.9 *Clinical Implementation and Validation:* Besides the statistical and scientific contribution we present a framework for micrometastases detection and tissue classification in sentinel lymph nodes. The system is currently validated in clinical practice at the University Hospital Zürich and hence represents the successful transfer of scientific research to a real world application with high social relevance.

Section 4.3 *Quality Control based on Clustering of Proximities:* Wishart-Dirichlet clustering based on sample proximities derived from relational detection forests is used to provide a means to differentiate between cell nuclei which are easy or difficult to classify. This contribution is an important step toward localized quality control on histological slides, which is currently the main advantage a human expert has over an automated system.

Section 4.5 *Signature Learning:* We present a novel feature selection scheme, based on proportional hazard models and false discovery rate (FDR), which is employed to device a signature for the progression in malignant melanoma. The proposed approach is not only able to implicitly handle missing values but provides also results for therapy of this severe cancer illness.

Robustness: Throughout the work conducted for this thesis,s special emphasis has been put on the robustness of the algorithms, and the ability to run in parallel on computer clusters and to handle terabytes of imaging data. The aim always was to develop solutions which *really work* in a robust manner, and which can be used in practice. This is illustrated not only by thousands lines of code in R and C# but first and foremost by the fact, that the software packages are used in practice by domain experts (cf. Section 3.9 and Section 3.11). They were employed in a dozen publications, where the methods and applications presented in this thesis were the enabling technique to gain scientific insight.

CHAPTER 2

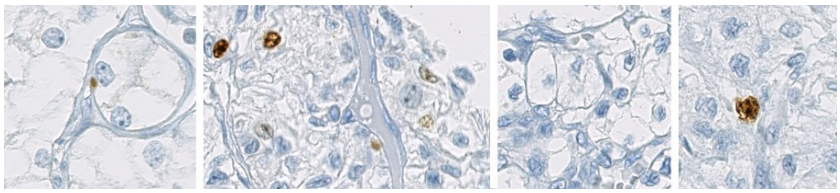
Data: Tissue and Ground Truth

Contents

2.1 Clear Cell Renal Cell Carcinoma

Throughout the thesis we use Renal cell carcinoma (RCC) as a disease case to design and optimize a computational pathology framework. RCC exhibits a number of properties which are highly relevant for computational pathology. Renal cell carcinoma figures as one of the ten most frequent malignancies in the statistics of Western societies (Grignon et al., 2004). The prognosis of renal cancer is poor since many patients suffer already from metastases at the time of first diagnosis. The identification of biomarkers for prediction of prognosis (prognostic marker) or response to therapy (predictive marker) is therefore of utmost importance to improve patient prognosis (Tannapfel et al., 1996). Various prognostic markers have been suggested in the past (Moch H, 1999; Su-

Figure 2.1
Biological variability of tissue within the subclass of clear cell renal cell carcinoma (ccRCC).



darshan S, 2006), but estimates of conventional morphological parameters still provide most valuable information for therapeutic decisions.

Clear cell RCC (ccRCC) emerged as the most common subtype of renal cancer and it is composed of cells with clear cytoplasm and typical vessel architecture. ccRCC exhibits an architecturally diverse histological structure, with solid, alveolar and acinar patterns. The carcinomas typically contain a regular network of small thin-walled blood vessels, a diagnostically helpful characteristic of this tumor. Most ccRCC show areas with hemorrhage or necrosis (Fig. 2.2d), whereas an inflammatory response is infrequently observed. Nuclei tend to be round and uniform with finely granular and evenly distributed chromatin. Depending upon the grade of malignancy, nucleoli may be inconspicuous and small, or large and prominent, with possibly very large nuclei or bizarre nuclei occurring (Grignon et al., 2004).

The prognosis for patients with RCC depends mainly on the pathological stage and the grade of the tumor at the time of surgery. Other prognostic parameters include proliferation rate of tumor cells and different gene expression patterns. Tannapfel et al. (1996) have shown that cellular proliferation potentially serves as another measure for predicting biological aggressiveness and, therefore, for estimating the prognosis. Immunohistochemical assessment of the MIB-1 (Ki-67) antigen indicates that MIB-1 immunostaining (Figure 2.2d) serves as an additional prognostic parameter for patient outcome. TMAs were highly representative of proliferation index and histological grade using bladder cancer tissue (Nocito A, 2001).

The TNM staging system specifies the local extension of the primary tumor (T), the involvement of regional lymph nodes (N), and the presence of distant metastases (M) as indicators of the disease state. Moch et al. (2009) focuses on reassessing the current TNM staging system for RCC and concludes that outcome prediction for RCC remains controversial. Although many parameters have been tested for prognostic significance, only a few were accepted in clinical practice. An especially interesting observation of Moch et al. (2009) is that multivariate Cox proportional hazards regression models including multiple clinical and pathologic covariates were more accurate in predicting patient outcome than the TNM staging system. On one hand this finding demonstrates the substantial difficulty of the task and on the other hand it motivates research in computational pathology to develop robust machine learning frameworks for reliable and objective prediction of disease progression.

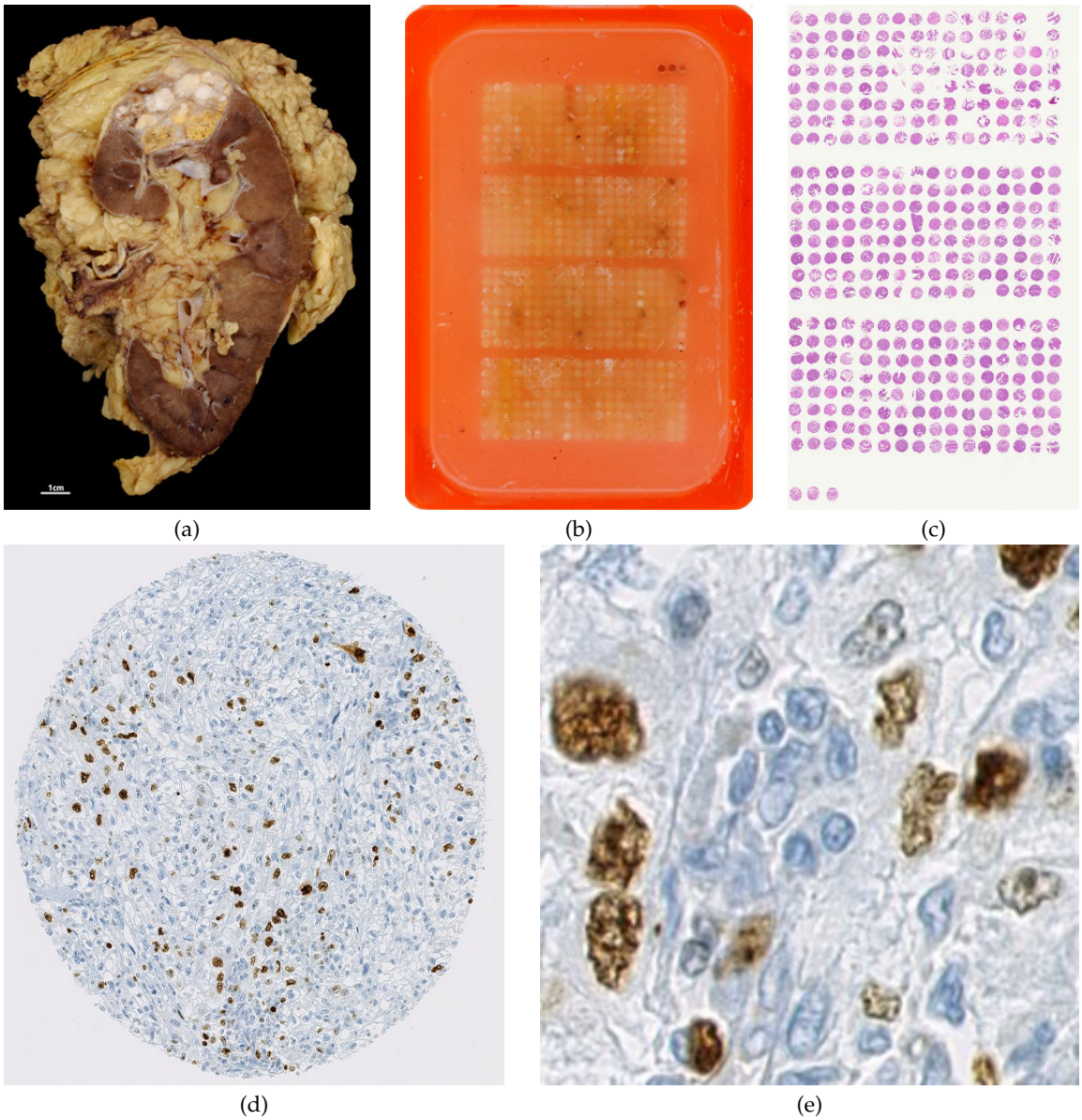


Figure 2.2

Tissue Microarray Analysis (TMA): Primary tissue samples are taken from a cancerous kidney (a). Then $0.6mm$ tissue cylinders are extracted from the primary tumor material of different patients and arrayed in a recipient paraffin block (b). Slices of $0.6\mu m$ are cut off the paraffin block and are immunohistochemically stained (c). These slices are scanned and each spot, represents a different patient. Image (d) depicts a TMA spot of clear cell renal cell carcinoma stained with MIB-1 (Ki-67) antigen. (e) shows details of the same spot containing stained and non-stained nuclei of normal as well as abnormal cells.

2.2 Tissue Microarrays

The tissue microarray (TMA) technology significantly accelerated studies seeking for associations between molecular changes and clinical endpoints (Kononen

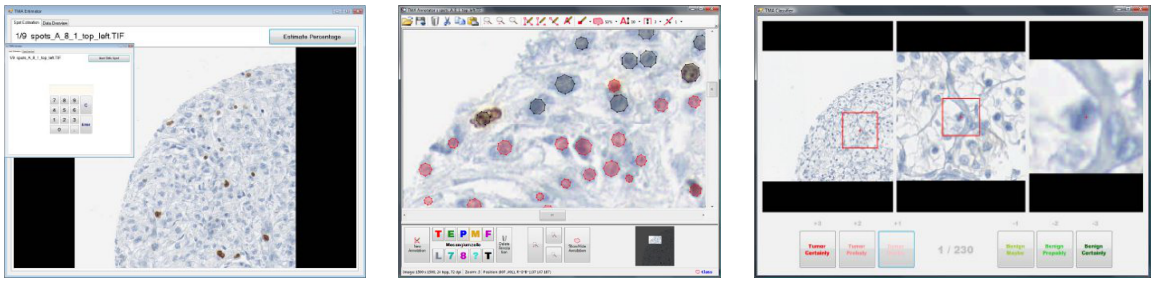
et al., 1998). In this technology, 0.6*mm* tissue cylinders are extracted from primary tumor material of hundreds of different patients and these cylinders are subsequently embedded into a recipient tissue block. Sections from such array blocks can then be used for simultaneous in situ analysis of hundreds or thousands of primary tumors on DNA, RNA, and protein level (cf. 2.2). These results can then be correlated with expression profile data which is expected to enhance the diagnosis and prognosis of ccRCC (Takahashi M, 2001; Moch H, 1999; Young AN, 2001). The high speed of arraying, the lack of a significant damage to donor blocks, and the regular arrangement of arrayed specimens substantially facilitates automated analysis.

Although tissue microarrays are produced by an almost routine process for most laboratories, the evaluation of stained tissue microarray slides remains tedious human annotation work, it is time consuming and prone to error. Furthermore, the significant intratumoral heterogeneity of RCC results in high interobserver variability. The variable architecture of RCC also results in a difficult assessment of prognostic parameters. Current image analysis software requires extensive user interaction to properly identify cell populations, to select regions of interest for scoring, to optimize analysis parameters and to organize the resulting raw data. Because of these drawbacks in current software, pathologists typically collect tissue microarray data by manually assigning a composite staining score for each spot - often during multiple microscopy sessions over a period of days. Such manual scoring can result in serious inconsistencies between data collected during different microscopy sessions. Manual scoring also introduces a significant bottleneck that hinders the use of tissue microarrays in high-throughput analysis.

2.3 Analyzing Pathologists

2.3.1 Motivation

To assess the inter and intra variability of pathologists we designed three different labeling experiments for the major tasks involved in TMA analysis. To facilitate the labeling process for trained pathologists we developed software suite which allows the user to view single TMA spots and which provides zooming and scrolling capabilities. The expert can annotate the image with vectorial data in SVG (support vector graphics) format and he/she can mark cell nuclei, vessels and other biological structures. In addition each structure can be labeled with a class which is encoded by its color. To increase usability and the adoption in hospitals we specifically designed the software for tablet PC so that a pathologist can perform all operations with a pen alone in a simple and efficient manner. Figure 2.3 depicts the graphical user interfaces of the three applications.



(a) (b) (c)

Figure 2.3

Three labeling applications for Tablet PC, which where used to construct a gold standard. Pathologists from the University Hospital Zürich conducted global staining estimation (a), nuclei detection (b) and nuclei classification (c).

2.3.2 Nuclei Detection

The most tedious labeling task is the detection of cell nuclei. In this experiment two experts on renal cell carcinoma exhaustively labeled a quarter of each of the 9 spots from the previous experiment. Overall each expert independently marked the center, the approximate radius and the class of more than 2000 nuclei. Again a tablet PC was used so it was possible to split up the work into several sessions and the experts could use the machine at their convenience. The user detects nuclei by marking the location with the pen on the tablet and indicates the diameter by moving the pen. A circular semi-transparent polygon is then drawn to mark the nucleus. The final step consists of choosing a class for the nucleus. In this setting it was either black for cancerous nuclei or red for normal ones. This task has to be repeated for each nucleus on each spot. Finally it is possible to show and hide the annotation to gain an overview over the original tissue. Figure 2.4 depicts a quarter of one of the RCC TMA spots together with the annotation and the inter expert disagreement.

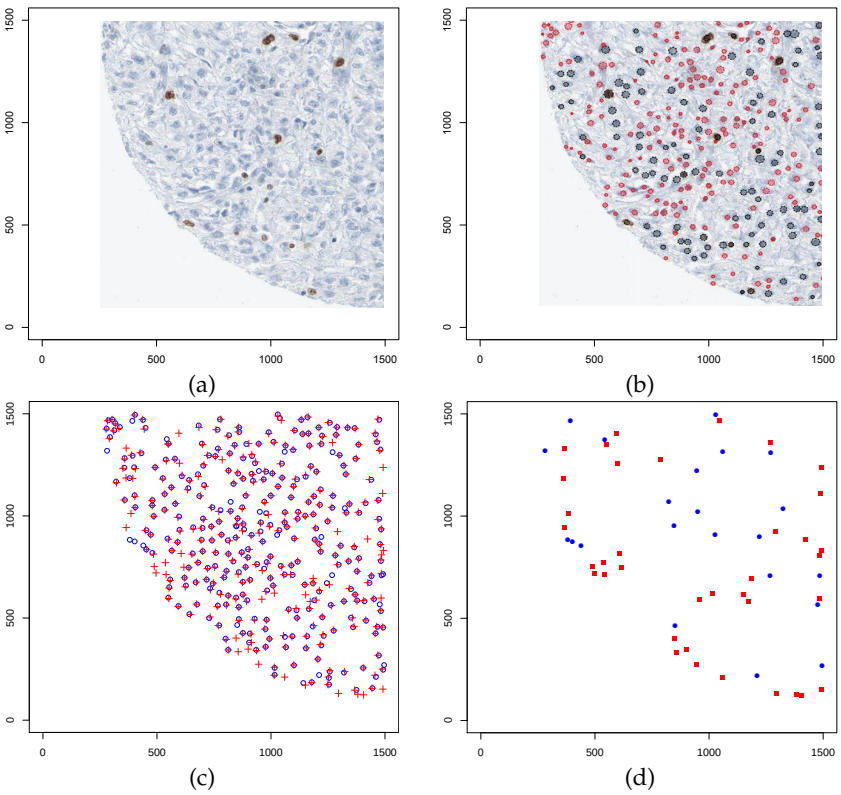
The average precision of one pathologist compared to the other is 0.92 and the average recall amounts to 0.91. These performance numbers show that even detecting nuclei on an histological slide is by far not an easy or undisputed task.

2.3.3 Nuclei Classification

The third experiment was designed to evaluate the inter and intra pathologist variability for nuclei classification, i.e. determining if a nucleus is normal/benign or abnormal/malignant. This step crucially influences the final outcome due to the fact that the percentage of staining is only estimated on the subset of cancerous nuclei. In the experiment, 180 randomly selected nuclei are sequentially presented in three different views of varying zoom stage. The query nucleus is indicated in each view with a red cross and the area which

Figure 2.4

(a) A quarter of an RCC TMA spot used for the nuclei detection experiment. (b) Annotations of one expert, indicating abnormal nuclei in black and normal ones in red. (c) Overlay of detected nuclei from expert one (blue circles) and expert two (red crosses). (d) Disagreement between the two domain experts regarding the detection task. Nuclei which were labeled only by pathologist one are shown in blue and the nuclei found only by expert two are depicted in red.



comprises the next zoom view is marked with a red bounding box (cf. Figure 2.3). During the setup phase the user can adjust these views to simulate his usual workflow as good as possible. During the experiment the expert has to select a class for each nucleus and rate his confidence. Thus, he has the choice between six buttons: tumor certainly, tumor probably, tumor maybe, benign certainly, benign probably and benign maybe. After classifying all nuclei, which have been classified as tumor, are displayed again and the pathologist has to estimate if the nucleus is stained or not. Again he has to rate his confidence in his own decision on a scale of three levels. To test the intra pathologist's variability a subset of nuclei was queried twice but the images were flipped and rotated by 90 degree at the second display to hamper recognition.

The results for inter-pathologist variability for the binary classification task are plotted in Figure 2.5a. Out of 180 nuclei all five experts agreed on 24 nuclei to be normal and 81 nuclei to be ab-normal, respectively cancerous. For the other 75 nuclei (42%) the pathologists disagreed.

The analysis of the intra-pathologist error is shown in Figure 2.5b. The overall intra classification error is 21.2% This means that every fifth nucleus was classified by an expert first as cancerous and the second time as normal or vice versa. The self-assessment of confidence allows us also to analyze single pathologists.

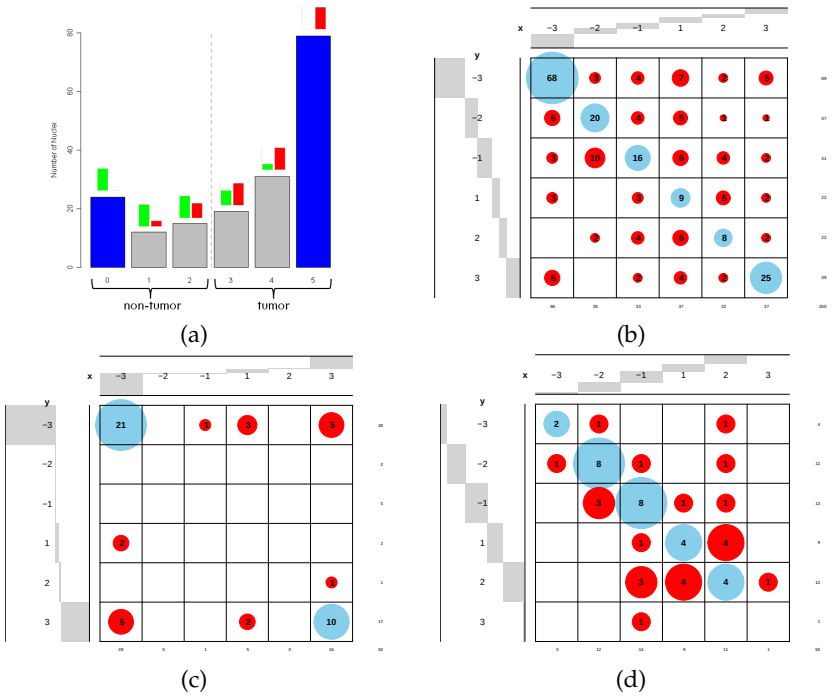


Figure 2.5

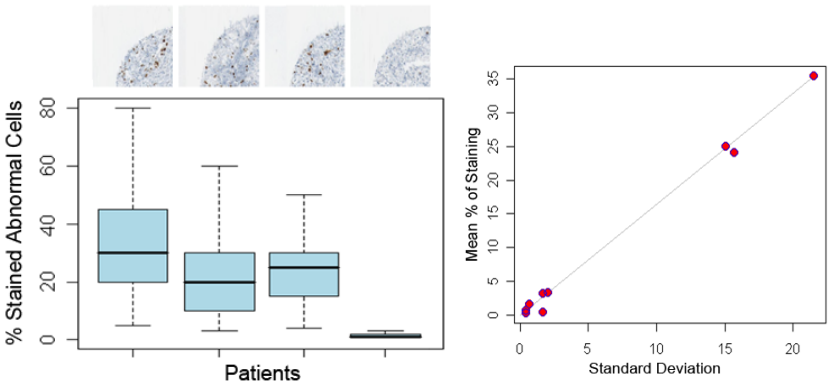
(a) Inter-pathologist classification variability based on 180 nuclei labeled by five domain experts. The experts agree on 105 out of 180 nuclei (blue bars: 24 normal, 81 cancerous). (b-d) Confusion matrices including reader confidence for intra-observer variability in nuclei classification: (b) The combined result of all five experts yields a intra pathologist classification error of 21.2%. (c) Example of an extremely self-confident pathologist with 30% error. (d) A very cautions pathologist with a misclassification error of 18%.

For example Figure 2.5c shows the results of a very self-confident pathologist who is always very certain of his decisions but ends up with an error of 30% in the replication experiment. Figure 2.5d on the other hand is the result of a very cautious expert who is rather unsure of his decision, but with a misclassification error of 18% he performs significantly better than the previous one. The important lesson learned is, that self-assessment is not a reliable information to learn from. The intuitive notion, to use only training samples which were classified with high confidence by domain experts is not valid.

In defense of human pathologist it has to be mentioned that these experiments represent the most general way to conduct a TMA analysis and analogous studies in radiology report similar results (Saur et al., 2009, 2010). In practice, domain experts focus only on regions of TMA spots which are very well processed, which have no staining artifacts or which are not blurred. The nuclei analyzed in this experiment were randomly sampled from the whole set of detected nuclei to mimic the same precondition which an algorithm would encounter in routine work. Reducing the analysis to perfectly processes regions would most probably decrease the intra-pathologist error.

2.3.4 Staining Estimation

The most common task in manual TMA analysis requires to estimate the staining. To this end a domain expert views the spot of a patients for several seconds



(a) (b)

Figure 2.6
(a) Results for 4 TMA spots from the labeling experiment conducted to investigate the inter pathologist variability for estimating nuclear staining. 14 trained pathologists estimated MIB-1 staining on 9 TMA spots. The boxplots show a large disagreement between pathologist on spots with an averages staining of more than 10%. The absolute estimated percentage is plotted on the y-axis. On the first spot the estimates from domain experts show a standard deviation of more than 20%. (b) The standard deviation grows linearly with the average estimated staining.

and estimates the number of stained abnormal cells without resorting to actual nuclei counting. This procedure is iterated for each spot on a TMA-slide to get an estimate for each patient in the study. It is important to note that, due to the lack of competitive algorithms, the results of nearly all TMA studies are based on this kind of subjective estimations.

To investigate estimation consistency we presented 9 randomly selected TMA spots to 14 trained pathologists of the University Hospital Zürich. The estimations of the experts varied by up to 20% as shown in Figure 2.6b. As depicted in Figure 2.6b the standard deviation between the experts grows linearly with the average estimated amount of staining. The high variability demonstrates the subjectivity of the estimation process. This uncertainty is especially critical for types of cancer for which the clinician chooses the therapy based on the estimated staining percentage. This finding not only motivates but emphasizes the need for more objective estimation procedures than current practice. Our research is stimulated by the hope, that computational pathology approaches do not only automated such estimation processes but also produce better reproduceable and more objective results than human judgment.

2.3.5 Discussion

The main challenge in this setting is posed by the vast heterogeneity of tissue in general and of cell nuclei in special. The difficulties mainly arise from the following peculiarities of histopathological image processing: (i) The objects which can be found are not only benign or malignant renal nuclei, but the tissue also comprises endothelial cells, lymphocytes, erythrocytes etc.. Endothelial cells for example are extremely elongated where on the other hand lymphocytes are nearly perfectly round and homogeneous. Neither of these cells must be counted in the estimation process. (ii) A grave difficulty is the fact that nuclei as three dimensional structures are not always perfectly cut in their maximum dimension producing numerous cutting artifacts. For many of these artifacts

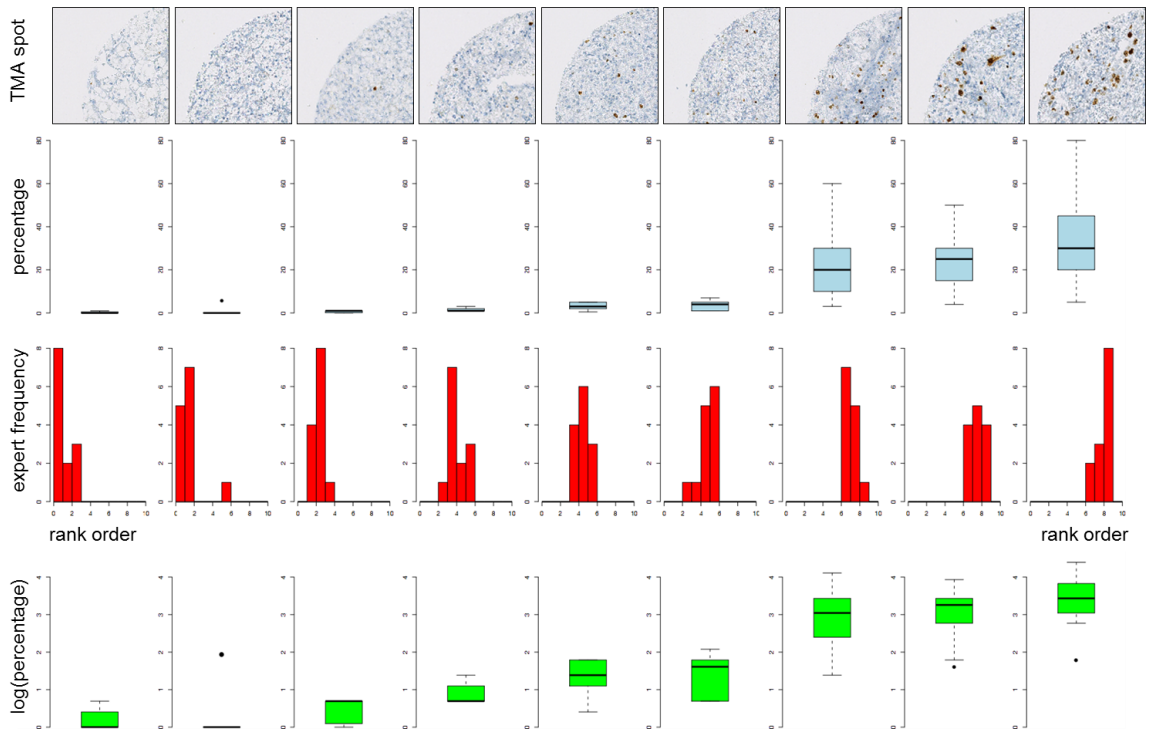


Figure 2.7

Comparison between the ranking and the continuous staining estimation of nine TMA spots (first row). The experiment was conducted by 14 trained pathologists and demonstrates the higher consistency of the rank order (third row) compared to the continuous estimation of stained abnormal nuclei in percentage (second row).

it is not possible to determine if they are atypical or benign. (iii) Variations in the production process of TMAs can lead to areas of different thickness within one section. This preprocessing artifact produces blurred regions in the image during the scanning process.

One important difference between the controlled labeling experiments and clinical practice is that pathologists don't assess the whole image, but only regions, which are of high quality in terms of imaging and biological relevance. To provide a similar functionality in an automated system there is a need for quality assessment procedures that are able to extract "relevant" regions which are then passed to detection and classification algorithms.

Hence it is reasonable to assume that the classification error of pathologist (25% intra pathologist error) is significantly lower on these "high quality" sub regions, but this hypothesis was not tested yet. The aim of this first study was to show if fully automated analysis on whole TMA images is even possible.

2.4 Expert Variability in Fluorescence Microscopy

Complementary to immunohistochemical TMA analysis, fluorescence microscopy is applied often for high-throughput screening of molecular phenotypes. A comprehensive study evaluating the performance of domain experts regarding the detection of lymphocytes is presented by Nattkemper et al. (2003). In a best case, a medium-skilled expert needs on average one hour for analyzing a fluorescence micrograph. Each micrograph contains between 100 and 400 cells and is of size 658×517 pixel. Four exemplary micrographs were blindly evaluated by five experts. To evaluate the inter-observer variability Nattkemper et al. (2003) define a gold standard comprising all cell positions in a micrograph that were detected by at least two experts.

Averaged over of CD3, CD4, CD7 and CD8 the sensitivity of the four biomedical experts is varying between 67.5% and 91.2% and the positive predictive value (PPV) between 75% and 100%. Thus the average detection error over all biomedical experts and micrographs is approximately 17%. Although fluorescence images appear to be easier to analyze due to their homogeneous background, this high detection error indicates the difficulty of this analysis task. These results corroborates the findings in the ccRCC detection experiment described in Section 2.1.

Positive Predictive Value:

$$PPV = \frac{TP}{TP + FP}$$

2.5 Generating a Gold Standard

The main benefit of labeling experiments like the ones described before, is not to point out the high inter and intra variability between pathologists, but to generate a gold standard. In absence of an objective ground truth measurement process, a gold standard is crucial for the use of statistical learning, first for learning a classifier or regressor and second for validating the statistical model. Section 5 shows an example how the information gathered in the experiments of Section 2.3 can be used to train a computational pathology system.

Besides labeling application which are developed for specific scenarios as the one described in Section 2.3 several other possibilities exist to acquire data in pathology in a structured manner. Although software for tablet PCs is the most convenient approach to gather information directly in the hospital it is limited by the low number of test subjects which can complete an experiment. To overcome this limitation the number of labelers can be extended significantly by the use of web-based technologies.

Crowd-sourcing services like Amazon Mechanical Turk can be used to gather large numbers of labels at a low cost. Applications in pathology suffer from the main problem, that the labelers are all non-experts. While crowd-sourcing works well for task based on natural images (Welinder and Perona, 2010), it poses a considerable problems in pathology where for example the decision if a nucleus is normal or cancerous is based on complicated criteria (Eble et al.,

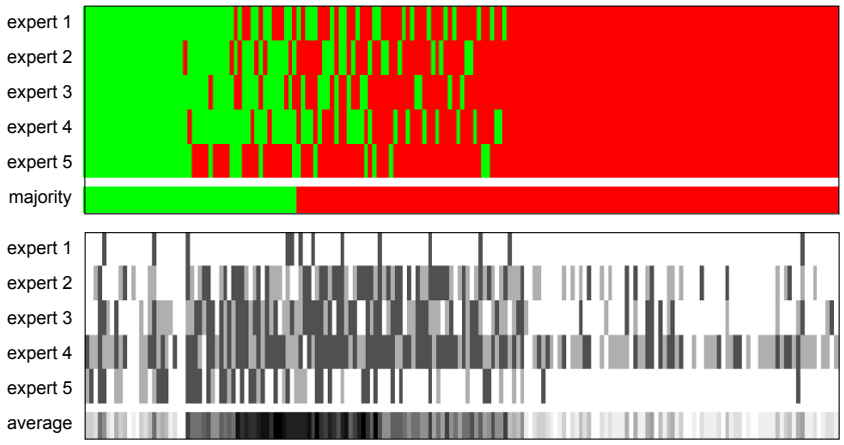


Figure 2.8
Labeling matrix with majority vote (top) and confidence matrix with confidence average (bottom) of five domain experts classifying 180 ccRCC nuclei into cancerous (red) and benign (green).

2004) which require medical training and knowledge. Likewise the recognition of some super cellular morphological structures requires years of training and supervision. Nevertheless crowd-sourcing could be useful in simple detection tasks like finding nuclei in histological slides.

2.6 Multiple Expert Learning

In classical supervised learning, a set of training data $\{(x_i, y_i)\}_{i=1,\dots,n}$ is available which consists of objects x_i and their corresponding labels y_i. The task is to predict the label y for a new test object x. This approach is valid as long as the target variable $Y = \{y_1, \dots, y_n\}$ denotes the ground truth of the application. If this condition is met, Y and $X = \{x_1, \dots, x_n\}$ can be used for classifier learning and evaluation.

Unfortunately, for a large number of real word application ground truth is either not available or very expensive to acquire. In practice, as a last resort, one would ask several domain experts for their opinion about each object x_i in question to generate a gold standard as described in Section 2.5. Depending on the difficulty of the task and the experience of the experts this questioning often results in an ambiguous labeling due to disagreement between experts. In pathology, very challenging scenarios, like assessing the malignancy of a cell, not only the inter, but also the intra expert variability is quite large (cf. Section 2.3). Moreover, restricting the dataset to the subset of consistently labeled samples results in loss of the majority of data in these scenarios.

Consequently, such a data acquisition procedure poses a fundamental problem for supervised learning. Especially in computational pathology there is clearly a need for novel algorithms to address the labeling problem and to provide methods to validate models under such circumstances.

More formally, each y_i is replaced by a D dimensional vector $\bar{y_i} = \{y_i^1, \dots, y_i^D\}$,

where y_i^d represents the i-th label of domain expert d. To this end one is interested in learning a classifier $\Phi(X, \bar{Y})$ from the design matrix X and the labeling matrix $\bar{Y}$. To date it is an open research question, how such classifier $\Phi(X, \bar{Y})$ should be formulated.

Recently Smyth et al. (1994) and Raykar et al. (2009) presented promising results based on expectation maximization where the hidden ground truth is estimated in turn with the confidence in the experts. Also along this lines Whitehill et al. (2009) introduced a probabilistic model to simultaneously infer the label of images, the expertise of each labeler, and the difficulty of each image. An application for diabetes especially for detecting hard exudates in eye fundus images was published by Kauppi et al. (2009).

Although a number of theoretical results exist (Lugosi, 1992; Smyth, 1996; Dekel and Shamir, 2009), empirical evidence is still lacking to establish that these approaches are able to improve over simple majority voting (Tullock, 1959; Downs, 1961) in real world applications.

A further, promising line of research investigates the question if such a classifier $\Phi(X, \bar{Y})$ can be learned in an on-line fashion, especially when the new labels come from a different domain expert. An affirmative answer would show a high impact in domains where specific models can be trained for years by a large number of experts, e.g. medical decision support.

In summary, extending supervised learning to handle domain expert variability is an exciting challenge and promises direct impact on applications not only in pathology but in a variety of computer vision tasks where generating a gold standard poses a highly non trivial challenge.

2.7 Public Datasets with Labeling Information

The available of public datasets with labeling information is crucial for the advance of an empirical science. Although a comprehensive archive like the UCI machine learning repository (Frank and Asuncion, 2010) does not exist for computational pathology, there are a number of datasets and initiatives which disseminate various kinds of data.

Immunohistochemistry: The most comprehensive database for antibodies and human tissue is by far the Human Protein Atlas (Berglund et al., 2008; Pontén et al., 2008). Comprising spots from tissue micro arrays of 45 normal human tissue types, it contains anywhere from $0 - 6$ images for each protein in each tissue type. The images are roughly 3000×3000 pixels in size, with each pixel approximately representing a $0.5 \times 0.5 \mu m$ region on the microscopy slide.

A segmentation benchmark for various tissue types in bioimaging was compiled by Drelie Gelasca et al. (2009), including 58 histopathological H&E stained images of breast cancer. The dataset provides labels from a single expert for the tasks of segmentation and cell counting.

Cytology: Automation in cytology is the oldest and most advanced branch in the field of medical imaging for pathology. This is mostly due to the fact that single cells are imaged on a homogeneous background and hence are more easily detected and segmented than in tissue. As a result commercial solutions are available since decades. Nevertheless especially the classification of detected and segmented nuclei still poses large difficulties for computational approaches. Lezoray and Cardot (2002) published ten color microscopic images from serous cytology with hand segmentation labels.
For bronchial cytology Meurie et al. (2005) provide eight color microscopic images. Ground truth information for three classes (nucleus, cytoplasm, and background pixels) is also available for each image. Pixels have a label specifying their classes (2: nucleus, 1: cytoplasm, 0: background).
A dataset of 3900 cells has been extracted from microscopical image (serous cytology) by Lezoray et al. (2003). This database has been classified into 18 cellular categories by experts.
Fluorescence Microscopy: A hand-segmented set of 97 fluorescence microscopy images with a total of 4009 cells has been published by Coelho et al. (2009). For fluorescence microscopy, the simulation of cell population images is an interesting addition to validation with manual labels of domain experts. Nattkemper et al. (2003); Lehmussola et al. (2007) present simulation frameworks for synthetically generated cell population images. The advantage of these techniques is the possibility to control parameters like cell density, illumination and the probability of cells clustering together. Lehmussola et al. (2007) supports also the simulation of various cell textures and different error sources. The obvious disadvantage are (i) that the model can only simulate what it knows and therefore can not represent the whole variability of biological cell images and (ii) that these methods can only simulate cell cultures without morphological structure. The later disadvantage also prevents their use in tissue analysis. Although the thought of simulated tissue images in light microscopy is appealing, currently there does not exist any methods which could even remotely achieve this goal.

CHAPTER 3

Imaging: From Classical Image Processing to Statistical Pattern Recognition

Contents

3.1 Overview

In recent years, a shift from rule based expert system towards learned statistical models could be observed in medical information systems. The substantial influence that machine learning had on the computer vision community is also reflecting more and more on medical imaging in general and histopathology in particular. Classifiers for object detection and texture description in conjunction with various kinds of Markov random fields are continuously replacing traditional watershed based segmentation approaches and handcrafted rulesets. Just recently Monaco et al. (2010) successfully demonstrated the use of pairwise Markov models for high-throughput detection of prostate cancer in histological sections. An excellent review of state-of-the-art histopathological image analysis methodology was compiled by Gurcan et al. (2009).
As with most cutting edge technologies, commercial imaging solutions lag behind in development but the same trend is evident. Rojo et al. (2009) review commercial solutions for quantitative immunohistochemistry in the pathology daily practice.
Despite the general trend towards probabilistic models, very classical approaches like mathematical morphology (Soille, 2003) are still used with large success.

Recently, Lzoray and Charrier (2009) presented a framework for segmentation based on morphological clustering of bivariate color histograms and Fuchs et al. (2008a) devised an iterative morphological algorithm for nuclei segmentation. Besides common computer vision tasks like object detection, segmentation and recognition, histopathological imaging poses domain specific problems such as estimating staining of nuclei conglomerates (Halama et al., 2009) and differentiation nuclei by their shape (Arif and Rajpoot, 2007).

3.2 An Iterative Morphological Approach

3.2.1 Summary

Mathematical morphology (Soille, 2003) provides a simple but comprehensive toolkit for classical image processing. In this section we present a framework for weakly supervised cell nuclei detection and segmentation on tissue microarrays of renal cell carcinoma. The presented approach combines an iterative morphological algorithm with subsequent nuclei filtering based on support vector machines (SVM).
As shown in Section 3.2.7, the proposed method outperforms previous approaches based on clustering and SVMs but still it is a representative of classical image processing methodology and inherits all the problems inherent to such procedures as described in Section 3.2.8.
A modern machine learning solution to the problem is described in Section 3.5.3 and compared to the morphological framework discussed here in Section LearnTree.

Acknowledgments
Special thanks to Tilman Lange who contributed his knowledge on SVMs and the Hungarian algorithm, to Peter J. Wild for labeling thousands of nuclei and his help with all medical questions regarding RCC, and Nima Rezavi for implementing the method proposed by Glotsos et al. (2005).

3.2.2 Uneven Illumination Correction

Most of the TMA spots show an illumination gradient resulting from light variations during the scanning process or an uneven cut tissue slice, which leads to thicker or thinner and, therefore, to darker and lighter areas on the image. We use a top-hat transform $\gamma(I)$ for mitigating the illumination gradients as described by Soille (2003) and Sonka et al. (2007). A top-hat with a large isotropic structuring element acts as a high-pass filter. Therefore, it can remove the illumination gradient which lies within the low frequencies of the image. In practice, we open the image I with square B of size 25×25 and subtract the result from the original Image: $I_{even} = I_{original} - \gamma_B(I)$.

3.2.3 Edge Pruning

We apply the Canny edge detector (Canny, 1986) to get an edge map of the TMA spot. Besides edges on nuclei boundaries, this results in a large number of edge responses from undesired boundaries between cytoplasm, vessels, connecting tissue and background. To filter out these edges, the following edge pruning

algorithm is applied:

First, we run a self devised, simple and fast junction detector to find and remove junctions between edges. This task is solved by applying a series of hit-or-miss (HMT) transformations with all possible structuring elements B, representing junctions in a 3×3 neighborhood: $HMT_B(X) = \epsilon_{B_1}(X) \cap \epsilon_{B_2}(X^C)$, where $\epsilon_B(X)$ is the morphological opening with a structuring element B, X is the set of pixel and X^C its complement. To reduce the set of possible junctions we shrink the edge map to minimally connected strokes. The result of this procedure is a set of thinned edges without junctions.

Second, these edges are split at equidistant points into edgels with a length of approximately ten pixels. For each of the edgels, a neighboring region on each side of this edgel is considered. We then calculate the mean intensity in these regions and use the lower value of this intensity as a score for the edgel. The rational for this procedure is the observation that desired edges occur on boundaries between a nucleus membrane and the cytoplasm as well as between a nucleus and the background. Hence the difference between the edgel neighboring regions can vary significantly but the lower mean value has to be below a maximum threshold θ for the edgel to be considered as part of a nucleus boundary.

Third, we keep all edgels that either score lower than θ or which are neighbors to edgels with satisfying score.
This pruning procedure allows us to discard undesired edges that are not part of a nucleus boundary.

3.2.4 Morphological Object Segmentation

To segment nuclei in the pruned edge map we devise a novel iterative, algorithm, that applies morphological opening and closing operations to detect potential nuclei. These nuclei are than subtracted from the pruned edge image and the process is repeated with larger structuring elements. In detail, the algorithm works as follows:

1. Perform a morphological closing ϕ_{B_1} with the structuring element B_1 to close gaps of the size of B_1: $I_{boundaries} = \phi_{B_1}(I_{edges})$
2. Fill all holes in the image $I_{boundaries}$.
3. Perform a morphological opening γ_{B_2} with a structuring element B_2 to remove single edges, which do not belong to a closed blob, respectively a nucleus: $I_{blobs} = \gamma_{B_2}(I_{boundaries})$
4. Add the resulting blobs to the final segmentation map:
   $I_{segmented} = I_{segmented} | I_{blobs}$

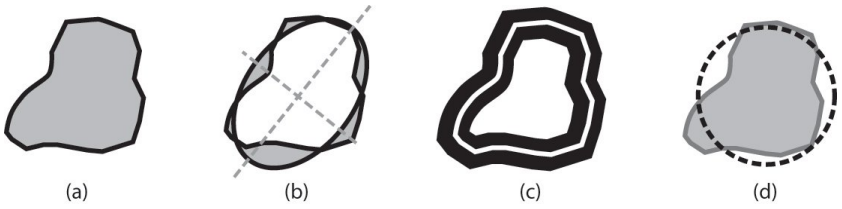


Figure 3.1
Illustration of the main nucleus features used for the classification process. From left to right: Schematic sketch of a nucleus. Overlap and extend of the nucleus and its ellipse with the same normalized second central moments. Outer and inner boundary regions of the nucleus membrane. Nucleus perimeter and its optimal circle with respect to the area.

5. Remove the edges, that belong to the found nuclei from the edge map: $I_{edges} = I_{edges} \setminus I_{blobs}$

6. Increase the size of the structuring element B_1 to close larger gaps in the next iteration: $B_1 = \delta_{B_3}(B_1)$

7. Start over at step 1 until a predefined number of iterations is performed.

The resulting segmentation contains all true positive nuclei and a number of false positive segmented blobs.

3.2.5 Nuclei Filtering

The morphological segmentation algorithm described above yields a large number of potential nuclei including many false positive. To solve this problem and to filter out these false positive detections a *soft-margin support vector machine* (Vapnik, 1998) is trained to classify between true nuclei and false positive. For this task we designed 21 features based on the expertise of the collaborating pathologists and these features are supposed to capture the properties of a nucleus in terms of shape, appearance and geometry.

The geometric features for a nucleus n (the set of pixels belonging to it) are defined as follows:

$$\begin{aligned}
\text{Size}(n) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(|n|-\mu)^2}{2\sigma^2}\right), \quad \mu = 600, \quad \sigma = 300 \\
\text{Ellipticity}(n) &= \frac{|n_{Ellipse}|}{|n_{Ellipse}| + |(n_{Ellipse} \cup n) \setminus (n_{Ellipse} \cap n)|} \\
\text{ShapeRegularity}(n) &= 2\pi\sqrt{\frac{n_{Area}}{\pi}} / n_{Perim}
\end{aligned}$$

where $|n|$ is its area and n_{Perim} its perimeter. $n_{Ellipse}$ is the ellipse that has the same normalized second central moments as the nucleus and $|n_{Ellipse}|$ is its

area. Color features are computed for each color channel separately:

$$
\begin{aligned}
\text{NucleusIntensity}(n) &= \frac{1}{|n|}\sum_{x\in n} x \\
\text{InnerIntensity}(n) &= \frac{1}{|n|}\sum_{x\in[n\setminus\epsilon_B(n)]} x \\
\text{OuterIntensity}(n) &= \frac{1}{|n|}\sum_{x\in[\delta_B(n)\setminus n]} x \\
\text{InnerHomogeneity}(n) &= \text{std}(x \in [n \setminus \epsilon_B(n)]) \\
\text{OuterHomogeneity}(n) &= \text{std}(x \in [\delta_B(n) \setminus n]) \\
\text{IntensityDifference}(n) &= \frac{1}{|n|}\sum_{x\in(\delta_B(n)\setminus n)} x \cdot \left(\frac{1}{|n|}\sum_{x\in(n\setminus\epsilon_B(n))} x\right)^{-1}
\end{aligned}
$$

where $\epsilon_B(n)$ is the morphological erosion of the nucleus n with the structuring element B and $\delta_B(n)$ is the equivalent morphological opening as described by Soille (2003). For the structuring element a disk with a radius of five is used. The outer and inner homogeneity is the standard deviation of the intensities in the corresponding regions. Figure 3.1 depicts the geometrical features and the inner and outer regions.

To train the support vector machine we used the labels from the generated gold standard, described in section 2.5. Therefore we used 300 cell nuclei from one patient for training and left out the remaining 1700 nuclei from the other 8 patients for validation. The training data consists of the 21 features for each nucleus candidate and the label y_n indicating whether the candidate nucleus is a true or a false positive. We used the support vector machine in conjunction with a Gaussian kernel function $k(x, y) = \exp(-\sigma\|x - y\|^2)$. The kernel width parameter σ as well the penalty for misclassified points C (cf. (Vapnik, 1998)) have been determined by K-fold cross-validation.

3.2.6 Performance Measure

One way to evaluate the quality of the segmentations/nuclei detection is to consider true positive (TP), false positive (FP) and false negative (FN) rates. The calculation of these quantities is based on a matching matrix where each boolean entry indicates if a machine extracted nucleus matches a hand labeled one or not. To quantify the number of correctly segmented nuclei, a strategy is required to uniquely match a machine detected nucleus to one identified by a pathologist. To this end we model this problem as bipartite matching problem, where the bijection between extracted and gold-standard nuclei is sought inducing the smallest detection error (Kuhn, 1955). This prevents overestimating the detection accuracy of the algorithms. To compare the performance of the algorithms we calculated precision and recall as well as the corresponding F-measure.

Precision:

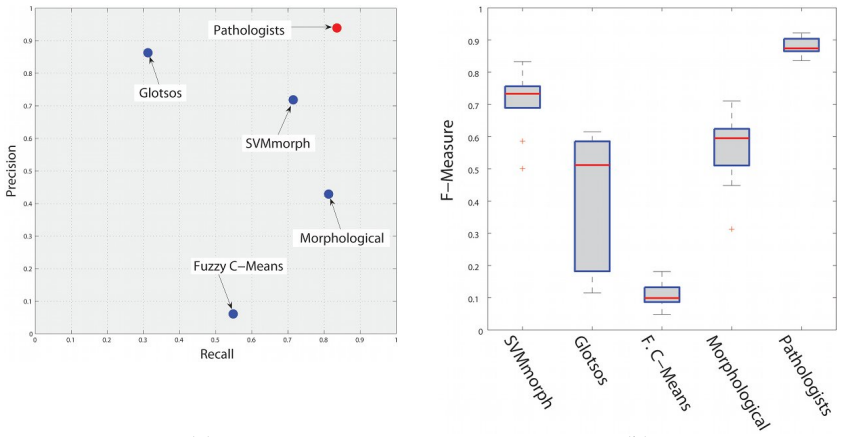
$$P = \frac{TP}{TP + FP}$$

Recall:

$$R = \frac{TP}{TP + FN}$$

F-measure:

$$F = 2 \cdot P \cdot \frac{R}{P + R}$$



(a) (b)

Figure 3.2
(a) Precision-Recall plot of the evaluated algorithms. (b) Boxplot of the F-Measure for all algorithms.

3.2.7 *Detection Accuracy*

At first, it is noteworthy that the annotations provided by the pathologists in terms of nuclei detection differ by roughly 20%, which is also depicted in Figure 3.3. This discrepancy is due to the fact that the pathologist did not agree if some structures in the image correspond to cell nuclei or not. In terms of the F-measure, the inter-pathologists variability is approximately 11%. Reaching this range of variability represents the ultimate goal of a computational approach to cell nuclei detection.

The morphological segmentation was applied with and without SVM classification to the 8 TMA test images. In order to relate the quality of our proposal to existing methods, we have additionally considered (i) a standard fuzzy c-means segmentation (Jain and Dubes, 1988) and (ii) a texture-based algorithm (Glotsos et al., 2005) relying on a support vector clustering algorithm. For the fuzzy c-means algorithm we started with four cluster and then merged the two darkest (which represent stained and not stained nuclei) and the two brightest (which represent cytoplasm and background). For the algorithm from Glotsos et al. (2005) we scaled down the images to the same resolution as described by Glotsos et al. (2005) and we used the recommended parameters.

Figure 3.2 summarizes the results of these experiments with *SVMmorph* representing our combined algorithm. The second best performing method in terms of the F-measure is the morphological segmentation. If the classification post-processing is additionally applied, the result is significantly improved. This quality jump demonstrates that the SVM-based method manages to significantly reduce the number of false positives without sacrificing too much accuracy, since the precision is higher while the recall is only marginally smaller than that of the pure morphological segmentation. The score achieved by this approach comes close to the inter-pathologists variability. Fuzzy c-means performs rather poorly on these images and it turned out to be the method with the

lowest performance in this study. This observation is primarily explained by the high number of false positives specifically in connecting tissue. Hence the low precision results in the worst value of F-measure in this experimental evaluation. Most of the nuclei actually detected by the proposed algorithm of Glotsos et al. (2005) can also be found in the annotation of the pathologists. Therefore, the results of their algorithm produces a fairly high precision value. This high precision, however, is achieved at the expense of a low recall: The algorithm misses many of the annotated cell nuclei due to its conservative search strategy. Hence, the method by Glotsos et al. (2005) is only the third best method in this study achieving a score of 55%. In summary, our method significantly outperforms the other methods under consideration.

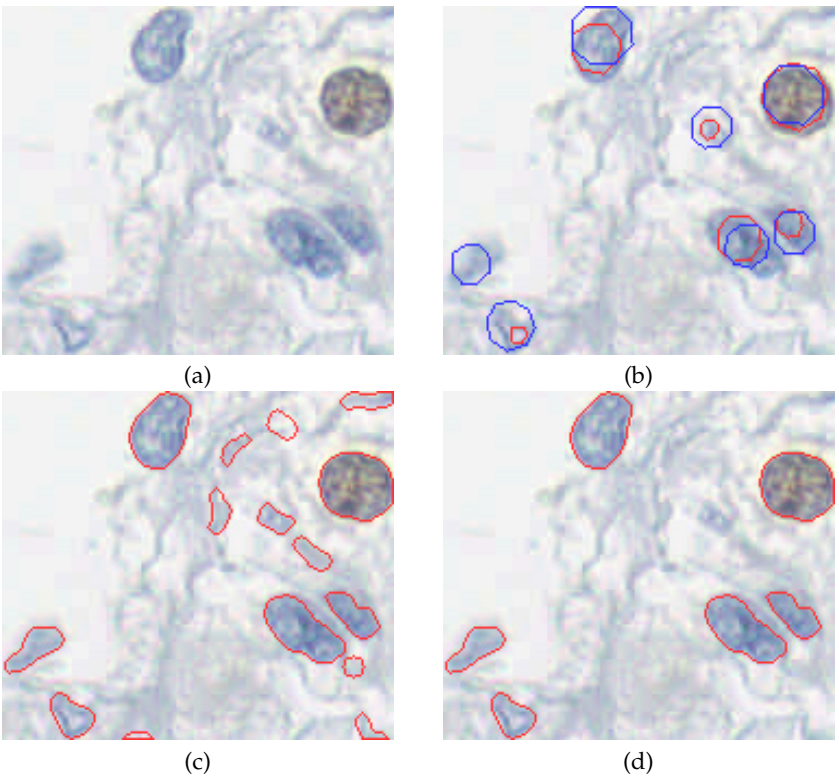


Figure 3.3
A detail of a TMA spot is shown in (a). The two independent annotations from the pathologists for this detail are depicted in (b). (c) shows the morphological segmentation and the final result after SVM nuclei classification is presented in (d). Comparing image (b) and (d) it can be seen that the algorithm misses one nucleus in the top right quadrant but segments another nucleus on the left border, which was detected by only one of the two experts.

3.2.8 Benefits and Disadvantages

Automatic, high throughput analysis of tissue microarrays promises new avenues for the discovery of biomarkers to detect and predict the progress of renal cell carcinoma.

Benefits: We have designed and evaluated an imaging pipeline for cell nuclei detection and segmentation which approaches the performance of working pathologists. Adaptive classification techniques with a soft-margin support

vector machine in combination with morphological image operations clearly outperform competing methods, such as the one proposed by Glotsos et al. (2005), by filtering out the overwhelming number of false positive detections of cell nuclei. To our knowledge, this is the first in depth study for tissue microarrays which incorporates expert labeling information down to the detail of single cell nuclei. The automated, quantitative analysis of tissue microarrays may in the future give rise to intelligent prognosis systems. Investigating the quality of such prediction approaches on the basis of TMA features and the correlation of survival time and automated prognosis is the subject of our current research.
Disadvantages: Although the iterative morphological system proposed in this section outperforms previous approaches it suffers from a number of drawbacks, like most classical image processing methods:

1. The number of parameters which have to be fitted is rather large. In addition to the morphological parameters the C of the SVM and the width of the RBF kernel have to be optimized in cross-validation experiments. Furthermore the system is rather sensitive to these parameters in contrast to algorithms like random forests (cf. Section 3.5.2).

2. The algorithm itself is not invariant to illumination changes in the scans of the histological slides. Hence an additional preprocessing step is necessary as proposed in Section 3.2.2. This problem is described in detail in the next section and an illumination invariant feature basis is defined.

3. Finally, the question arises if the presented framework does not solve a problem which is harder than necessary. I.e. for estimating the percentage of stained cancerous cells the exact segmentation of the nuclei would not be necessary if an object detector could be trained only for the malignant object class. In Section LearnTree such a system is described for multiple object detection.

3.3 Preprocessing vs. Algorithmic Invariance

Brightfield microscopic imaging often produces large differences of illumination within single slides or TMA spots. These variations are caused by the varying thickness of the slide or by imperfect staining. Such problems can be overcome either by preprocessing the image data or by designing and integrating invariance into the algorithmic processing to compensate for these variations.
Inconsistencies in the preparation of histology slides render it difficult to perform a quantitative analysis on their results. A normalization approach based on optical density and SVD projection is proposed by Macenko et al. (2009) for overcoming some of the known inconsistencies in the staining process. Slides which were processed or stored under very different conditions are projected into a common, normalized space to enable improved quantitative analysis.

Preprocessing and normalization methods usually not only reduce noise induced differences between samples but often also eliminate the biological signal of interest. As an alternative to such an irreparable information loss during data acquisition, algorithms with illumination invariance or with compensation of staining artifacts are designed which are robust to these uncontrollable experimental variations.

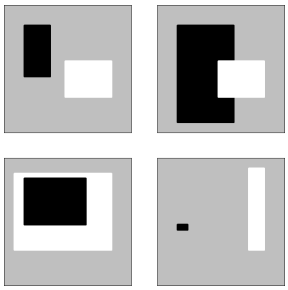


Figure 3.4

Examples of gray scale relational features. The coordinated of two rectangles are sampled uniformly without constraints within a training window of size $64 \times 64px$. The resulting binary feature is evaluated by considering only a relation (e.g. $I_{R_1} \leq I_{R_2}$) between the average intensities of the rectangles.

Relational Detection Forests (Section 3.5.3) provide one possibility to overcome this problem of information loss. Especially designed for detection of cell nuclei in histological slides, they are based on the concept of randomized trees (Breiman, 2001). The features, which are selected for this framework center around the idea that relation between features are more robust than thresholds on single features. A similar idea was applied by (Geman et al., 2004) to gene chip analysis where similar problems occur, due to the background noise of different labs. Contrary to the absolute values the relation between DNA expression is rather robust.

Object detection is commonly solved by training a classifier on patches centered at the objects of interest (Viola and Jones, 2001), e.g., the cell nuclei in medical image processing of histological slides. Considering only the relation between rectangles within these patches results in illumination invariant features which give the same response for high and low contrast patches as long as the shape of the object is preserved. It has to be noted, that due to the directionality of the relation they fail if the image is inverted. In general, illumination invariance speeds up the whole analysis process because neither image normalization nor histogram equalization are required.

The feature base is defined as follows: The coordinates of two rectangles R_1 and R_2 are sampled uniformly within a predefined window size w:

$$R_i = \{c_{x1}, c_{y1}, c_{x2}, c_{y2}\}, \quad c_i \sim U(x|0, w)$$

For each rectangle the intensities of the underlying gray scale image are summed up and normalized by the area of the rectangle. The feature $f(s, R_1, R_2)$ evaluates to a boolean value by comparing these quantities:

$$f(s, R_1, R_2) = \begin{cases} 1 & \text{if } \sum_{i|x_i \in R_1} \frac{x_i}{n_1} < \sum_{i|x_i \in R_2} \frac{x_i}{n_2} \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

where x_i is the gray value intensity of pixel i of sample $s = \{x_1, x_2, \ldots, x_n\}$ and n_1, n_2 denote the number of samples in R_1, R_2, respectively. From a general point of view this definition is similar to generalized Haar features but there are two main differences: (i) the quantity of interest is not the continuous difference between the rectangles but the boolean relation between them and hence (ii) it is not necessary to learn a threshold on the difference to binarize the feature.

For example, in the validation experiments a window size of 65×65 pixels was

chosen. Taking into account that rectangles are flipping invariant, this results in $\left((64^4)/4\right)^2 \approx 2 \cdot 10^{13}$ possible features.
Putting this into perspective, the restriction of the detector to windows of size 24×24 leads to $\sim 6.9 \cdot 10^9$ features which are significantly more than the 45,396 Haar features from classical object detection approaches (Viola and Jones, 2001). For such huge feature spaces it is currently not possible to exhaustively evaluate all features while training a classifier. Approaches like AdaBoost (Freund and Schapire, 1996) which yield very good results for up to hundreds of thousands of features are not applicable any more. These problems can be overcome by employing randomized classification algorithms (Fuchs et al., 2009; Geurts et al., 2006) where features are sampled randomly for learning classifiers on these random projections.
Considering relations between rectangles instead of cut-offs on differences leads to a number of benefits:

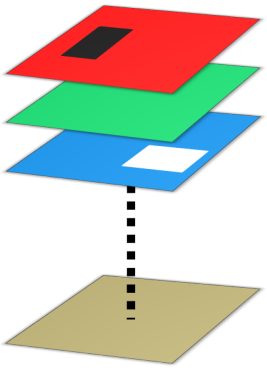


Figure 3.5
Relational color features are constructed equivalently to the gray scale feature with the addition of independently sampling the color channel for each rectangle.

Illumination Invariance: A major problem in multiple object detection in general and on microscopic images of human tissue in particular are vast differences of illumination within a single image. In natural images this can be due to shadows or fog and in histopathology it is mainly due to varying thickness of the slide or imperfect staining. Taking into account only the relation between rectangles lead to illumination invariant features which give the same response for high and low contrast patches as long as the shape of the object is preserved. It has to be noted, that due to the directionality of the relation they fail if the image is inverted. In general, illumination invariance speeds up the whole analysis process because neither image normalization nor histogram equalization are required.

Fast Evaluation: The most time intensive step in the induction of tree classifiers is the repeated evaluation of features. In the classical textbook approach as followed by (Breiman, 2001; Freund and Schapire, 1996; Viola and Jones, 2001) at each node all possible cut-offs on feature values are evaluated for every feature exhaustively. In the worst case the number of feature evaluations is as high as the number of samples at a given node. Recently (Geurts et al., 2006) proposed extremely randomized trees to overcome this problem. In their approach only a small number of cut-offs is randomly sampled to reduce the number of tests. Instead of ordering all feature values only the maximum and minimum have to be calculated and the cut-offs are drawn uniformly. The only assumption necessary for this approach is, that the feature values for all samples increase linearly. Given an exponential or logarithmic behavior a low number of sampled cut-offs are not able to capture the power of the feature.
In contrast to that, the proposed features require only one single evaluation. In practice this allows for testing hundreds or thousands of features per split with the same number of CPU cycles classical approaches

needed to test all thresholds of a single feature. Analyzing single features, the use of relations instead of cut-offs on differences seems to lead to lower variance and higher bias.

Ensemble Diversity: One of the main concerns in ensemble learning is to induce sufficiently diverse base classifiers. Increased diversity leads to increased performance. Randomized algorithms as described in Section 3.5.3 can benefit from the enormous size of the described feature space. Learning trees by randomly selecting features from this large set guarantees small correlation between base classifiers. Even if thousands of features are evaluated at each split, the chances that two trees are build of the same features are negligible.

3.4 Voronoi Sampling

It is common in object detection frameworks (Viola and Jones, 2001; Grabner et al., 2008; Dalal and Triggs, 2005) to randomly choose negative background samples. Therefore first, points are uniformly sampled from an interval for each coordinate and second, points which are closer to a positive sample than a predefined threshold are discarded. On images with dense packing of multiple object, as it is the case in most cell detection tasks, this procedure leads to two main problems: (i) Uniform sampling with rejection results on one hand in a higher percentage of negative samples in areas with sparse object distribution and on the other hand in very few samples in dense packed areas. Hence negative samples are lacking especially there, where differentiation between object and background is difficult. (ii) Without specific presorting or local sampling algorithms, uniform sampling can be cumbersome and slow in domains with densely packed objects.

To overcome these drawbacks we propose a simple sampling algorithm which we term *Voronoi Sampling*[1]. A Voronoi diagram or Dirichlet tessellation is a decomposition of a metric space determined by distances to a specified discrete set of objects in this space. In a multiple object detection scenario these mathematical objects are the centers of the queried objects.

To create a Voronoi diagram first a Delaunay triangulation is constructed, second the circumcircle centers of the triangles is determined, and third these points are connected according to the neighborhood relations between the triangles. In the experiments in Section 3.10.3 the Voronoi tessellation is created based on the joint set of nuclei from both pathologists to prevent the use of good and ambiguous nuclei as negative instances for learning. In contrast to uniform sampling, using a tessellation has the additional advantage that the negative samples are concentrated on the area of tissue and few samples are spent on

[1] The proposed algorithm should not be confused with Lloyd's algorithm (Lloyd, 1982) which is also known as "Voronoi iteration".

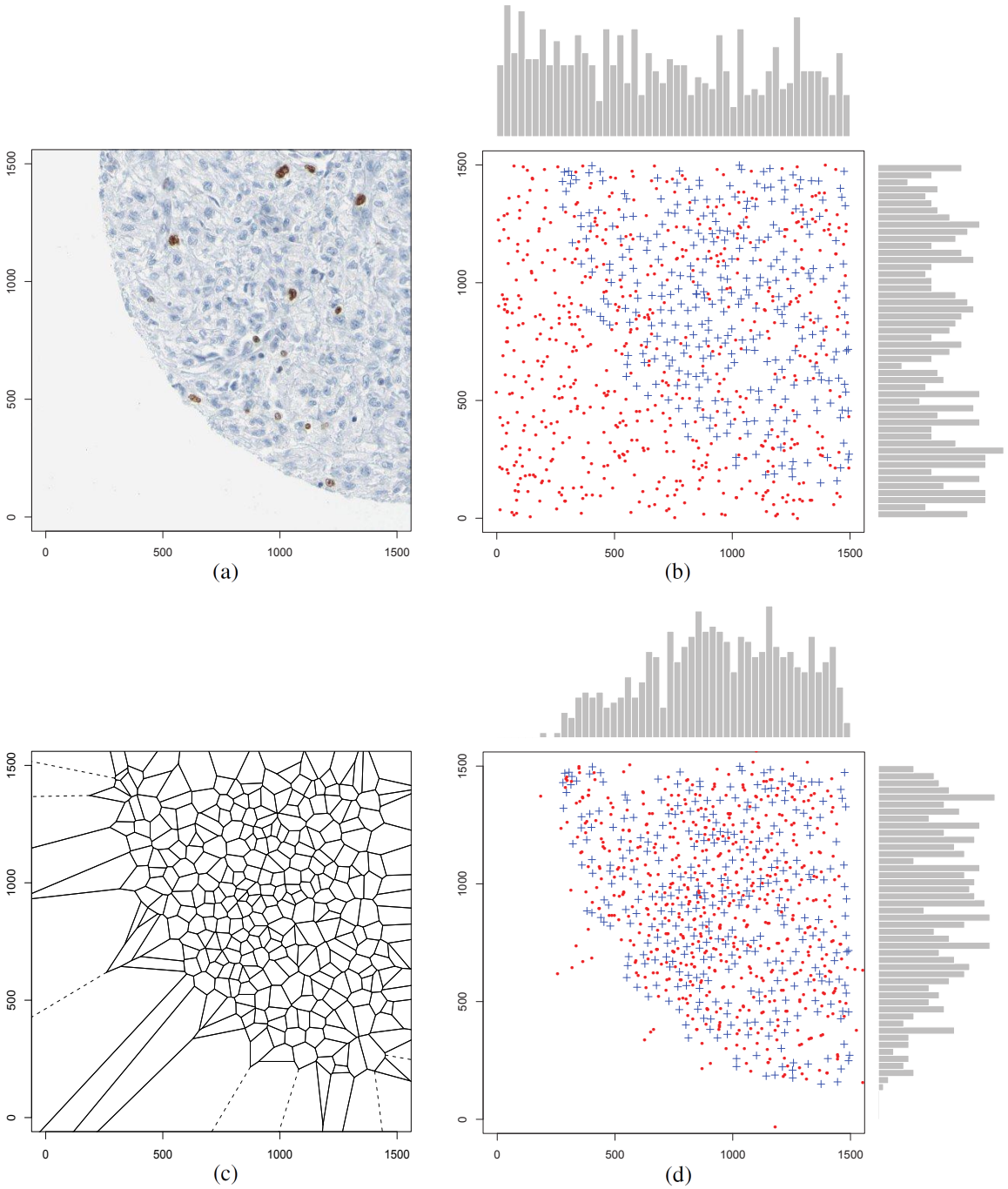


Figure 3.6

Voronoi Sampling. (c) Voronoi tessellation of the input space based on labeled nuclei. (d) Sampling negative instances from nodes of the Voronoi tessellation (red dots). Cell nuclei are marked with blue crosses. (b) Uniform rejection sampling in the whole input space. The marginal histograms show the frequency of negative samples. In contrast to uniform sampling, using a tessellation has the additional advantage that negative samples are concentrated on the area of tissue and few samples are spent on the homogeneous background.

the homogeneous background. Figure 3.6 shows a comparison of Voronoi and uniform rejection sampling.

3.5 Randomized Tree Ensembles

3.5.1 Historical Background

The idea of ensemble classifier emerged in the 70s with Tukey's twicing (Tukey, 1977), which can be viewed as a simple form of boosting without weighting (Buhlmann and Yu, 2003). The next important step towards random forests was bootstrap aggregation of decision trees, termed bagging (Breiman, 1996).

The history of ensemble classifiers based on randomized trees started in the mids of the 1990s under various names and implementations. Although a singular origin is difficult to pin point, two of the earliest publications seem to be "Randomized Decision Forests" by (Ho, 1995) and "Shape Quantization and Recognition with Randomized Trees" by (Amit and Geman, 1997). Although presented at well known machine learning conferences (Ho (1998) at ICPR[2]) and published in major machine learning journals (Amit and Geman (1997) in *Neural Computation*,(Dietterich, 2000) in Machine Learning, (Amit and Geman, 1997) and (Ho, 1998) in PAMI[3]) the breakthrough of randomized ensemble methods in respect to the general machine learning audience did not happen until 2001. That year the seminal paper of (Breiman, 2001) was published which is not only the most cited paper in this scope (cf. Section 3.1) but also provided the catchy name that stuck: *Random Forests*. One of the reasons for the delay of half a decade in public reception was probably the success of boosting (Freund and Schapire, 1996) which clearly dominated the discourse in the machine learning community at the end of the century. This is also demonstrated by the fact that already in the abstract of his 2001 paper Breiman explicitly positions random forests as a method, which "compares favorable" to Adaboost (Freund and Schapire, 1996).

With usual delay the computer vision community followed five years later with the application of boosting (Viola and Jones, 2001) for face detection as well as with the use of random forests for keypoint recognition (Lepetit and Fua, 2006).

Noteable improvements and extensions of random forests in recent years are *extremely randomized trees* by (Geurts et al., 2006) and *random survival forests* by (Ishwaran et al., 2008).

The name "random forest" itself is not an invention by Leo Breiman. The nomenclature is common in graph theory (cf. (Pitman, 1999; Pavlov, 2000)) where a "forest" is defined as a graph without cycles whereas a tree is a *connected* graph without cycles. Subsequently, a "random forest" is a forest consisting of "random trees", i.e. trees generated by a stochastic process.

[2]Proceedings of the 14th International Conference on Pattern Recognition
[3]IEEE Transactions on Pattern Analysis and Machine Intelligence

Reference	Ho (1995)	Amit and Geman (1997)	Breiman (2001)
Nomenclature	Randomized Decision Forests, Random Subspace Selection	Randomized Trees	Random Forests
Application	Handwriting Recognition	Handwriting Recognition	Classification and regression
Google Scholar Citation Count 2010-10-20	133	264	3718

Table 3.1

Historical overview and citation count of the first ensemble classifiers based on randomized trees together with the proposed nomenclature and the field of application.

3.5.2 Random Forests

Definition

A random forest with T trees is defined by (Breiman, 2001) as follows:

Definition 1. *A random forest is a classifier consisting of a collection of tree-structured classifiers* $\{h(\boldsymbol{x}, \theta_t), t = 1, \ldots, T\}$ *where* $\{\theta_t\}$ *are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input* $\boldsymbol{x}$.

The random vector θ_t for the t-th tree is generated independent of the past random vectors $\theta_1, \ldots, \theta_{t-1}$ but with the same distribution. Tree induction is then performed using the training set and θ_t, resulting in a classifier $h(\boldsymbol{x}, \theta_t)$ where $\boldsymbol{x}$ is the input vector.
In bagging (Breiman, 1996) the random vector θ are the indices of bootstrap samples for each tree. In random split selection θ consists of a number of independent random integers between 1 and the number of samples N. Hence it is important to note that the random vector θ can take various forms and the dimensionality may vary depending on its use during tree induction.

Tree Induction and Combination

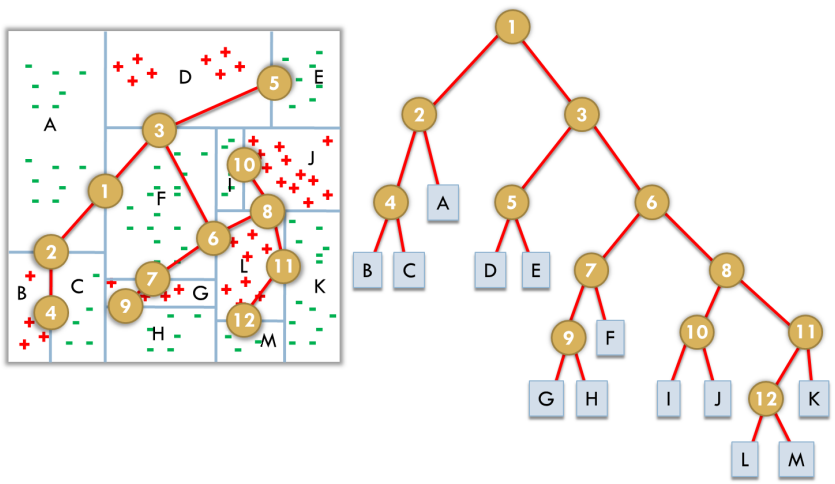


Figure 3.7

Toy example of an ordinary binary decision tree with 12 splitting rules for a binary classification task. The tree partitions the two dimensional feature space with axis-parallel splits into 13 regions. The stopping criterion for tree induction in this example was the label purity of all samples in a leave node. This example also shows that decision trees do not have to be balanced and that tree induction without subsequent pruning clearly overfits the data.

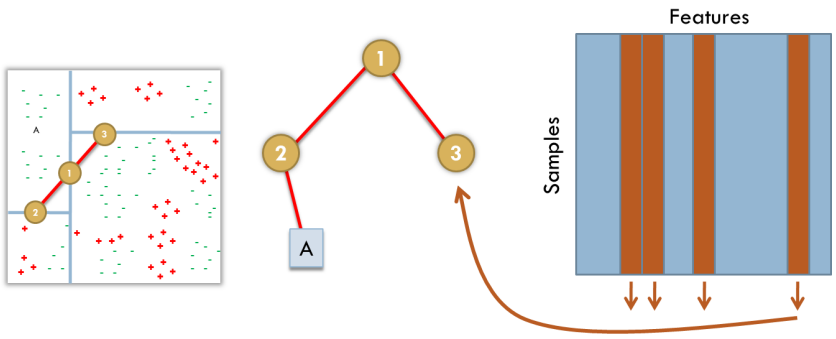


Figure 3.8
Tree induction in random forests. At each node a random subset of m features is drawn uniformly from the set of all p features. For each feature all possible splits are evaluated and then the feature with its split threshold is selected which minimized the cost function of choice.

One crucial advantage of random forests is, that it inherits all the positive properties of decision trees. Induction methods like CART (classification and regression trees Breiman et al. (1984)) or C4.5 (Quinlan, 1993) produce decision trees which are able to capture complex interaction structures in the data and therefore are inherently nonlinear. Trees can not only be used for classification and regression tasks but also for survival risk estimation (Hothorn et al., 2006). Furthermore it is possible to mix different data types like categorical, ordinal and continuous variables in the model which provides them with a significant advantage in real world applications. Along the same lines it is of great benefit that missing values can be handled implicitly without the use of imputation techniques (Breiman et al., 1984; Breiman, 2001). Furthermore decision tree induction and prediction can be implemented on GPUs(Sharp, 2008). Combined with the possibility to conduct tree induction in parallel this allows for realtime training and testing of random forests.
After the induction of T trees $\{h(x,\theta_t)\}_1^T$ the combined random forest predictor is

$$\hat{f}_{rf}^T = \frac{1}{T}\sum_{t=1}^{T} h(x_i,\theta_t),$$

where θ_t completely specifies the t-th tree in terms of nodes, split features, split thresholds and leave-node values. Also in classification problems, the average over all trees can be considered as an estimate of the class posterior probabilities which can be used to optimize the operation point on ROC curves for class unbalanced problems (Dahinden, 2006).

OOB Estimate

A very useful feature of random forests in practice is its ability to use out-of-bag (OOB) samples to estimate the generalization error. During the training of the ensemble bootstrap samples $Z^{*t}, t = 1 \dots T$ are drawn with replacement from the dataset $Z = \{(y_1, x_1), \dots, (y_N, x_N)\}$. From each bootstrap sample Z^{*t} a decision tree is induced and added to the ensemble. To estimate the generalization error each datum x_i is classified by all the trees for which the datum was out-of-bag, i.e. $x_i \notin Z^{*t}$.

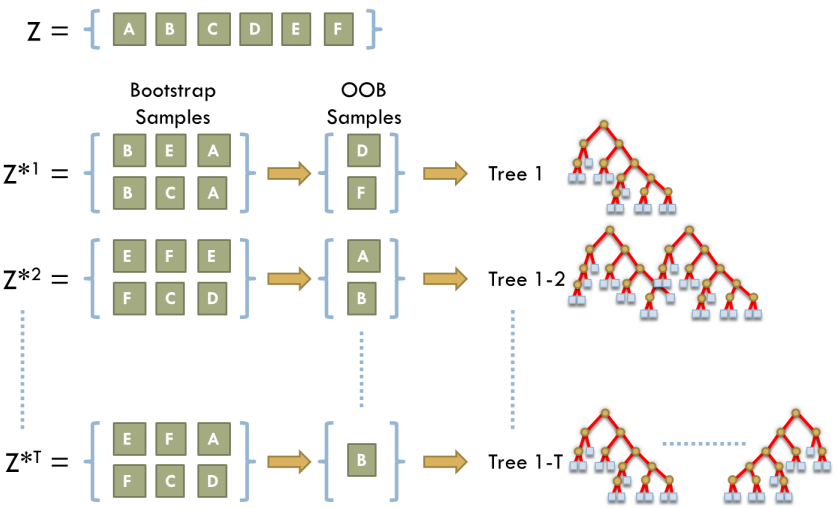


Figure 3.9
Ensemble construction in random forests. Bootstrap samples $Z^{*t}, t = 1 \dots T$ are drawn with replacement from the dataset $Z = \{(y_1, x_1), \dots, (y_N, x_N)\}$. From each bootstrap sample Z^{*t} a decision tree is induced and added to the ensemble. To estimate the generalization error each datum x_i is classified by all the trees for which the datum was out-of-bag, i.e. $x_i \notin Z^{*t}$.

Unlike many other nonlinear estimators, random forests can be fit in one single run, with cross-validation being performed along the way in the form of the OOB estimate (Hastie et al., 2009). As soon as the OOB error stabilizes, the training can be terminated. Hence the number of trees T is not to be regarded as a classical tuning parameter which has to be learned in separate cross-validation runs.

Correlation

Random forests, like bagging, is a variance reduction approach. A decision tree grown sufficiently deep has a low bias but high variance. Trees in an ensemble are grown identically distributed (i.d.). The variance over the average is

$$V[\hat{f}_{rf}^{T}] = \rho\sigma^2 + \frac{1-\rho}{T}\sigma^2,$$

where ρ is the positive pairwise correlation (Breiman, 2001). As the number of trees T increases the second term disappears. Therefore it is the correlation between pairs of bagged trees which limits the advantages of averaging (Hastie et al., 2009). To this end the main idea of random forests is to improve over bagging by reducing the correlation between trees, without increasing the variance too much. The correlation is reduced by introducing further randomness in the tree induction procedure by considering only a random subset of m features at each split of a tree. By reducing m the correlation between any pair of trees in the ensemble is reduced and thus reduces also the variance of the average. For datasets with p features (Breiman, 2001) proposed $m = \lfloor\sqrt{p}\rfloor$ and a minimum node size of one for classification and $m = \lfloor\frac{p}{3}\rfloor$ with a minimum node size of five for regression. Although m is a tuning parameter of random forests, it is interesting to note that for most practical problems the recommendations of

Breiman work surprisingly well.
Continuing along the same lines (Geurts et al., 2006) introduced "extremely randomized trees". Instead of learning the best threshold for a randomly selected feature, only a small number of thresholds is sampled and the best one is chosen. (Geurts et al., 2006) demonstrated that this approach improves over Breimans random forest on several datasets.
In Section 3.5.3 we present an approach for object detection with random forests where we go even a step further and sample features from an extremely large pool of binarized variables. The benefit of this approach is a low correlation between the trees of the ensemble.
(Amit and Geman, 1997) and (Breiman, 2001) provide further details how to measure strength and correlation in ensembles of randomized trees.

Variable Importance

The importance of variables within a random forest can be calculated with two different approaches. (Breiman, 2001) proposed to use the OOB samples, which are passed down the tree to calculate its prediction accuracy. Sequentially the values of each variable are randomly permuted in the OOB samples and the accuracy is calculated again. The decrease of accuracy is averaged over all trees of the ensemble and used as a measure for the importance of the variable.
An alternative approach is simply to sum up the improvement in the split criterion for each variable at each node of every tree. (Hastie et al., 2009) argue that this second approach yields more pronounced importance profiles compared to Breimans idea.

3.5.3 Relational Detection Forests

Motivation

The aim is to construct a classifier for object detection which is invariant under certain transformations and fast during training for the use in online learning settings. The proposed framework is analyzed within the context of invariance, generalization error and survival estimation.

Tree Induction

The base learners for the inter-active ensemble are binary decision trees. Tree learning follows loosely the original approach for random forests described by Breiman (2001). A recursive formulation of the learning algorithm is given in procedure `LearnTree`. The sub routine `SampleFeature` returns a feature consisting of two rectangles uniformly sampled within a predefined window as described in Section 3.3.
In accordance with (Breiman, 2001) the Gini Index is used as splitting criterion, i.e. the Gini gain is maximized. At a given node, the set $S = \{s_1, \ldots, s_n\}$ holds

Procedure LearnTree

Input: set of samples $S = \{s_1, s_2, \ldots, s_n\}$
Input: depth d
Input: max depth d_{max}
Input: features to sample $mTry$

Init: $\widehat{label} = null$**;** $g = -\inf$
Init: $N_{left} = null$**;** $N_{right} = null$
if $(d = d_{max})$ *OR* $(\texttt{isPure}(S))$ **then**
  $\widehat{label} = \arg\max_{l \in \{true, false\}} \#\{i | s_i = l\}$
else
  for $i = 0, i < mTry, i++$) **do**
    $f_i =$ **SampleFeature()**
    $S_L = \{s_j | f_i(s_j) = true\}$
    $S_R = \{s_j | f_i(s_j) = false\}$
    $g_i = \widehat{\Delta G}(S_L, S_R)$
    if $g_i > g$ **then**
      $f_{best} = f_i; \quad g = g_i$
    end
  end
  $N_{left} =$ **LearnTree**$(\{s_i | f_{best}(s_i) = true\})$
  $N_{right} =$ **LearnTree**$(\{s_i | f_{best}(s_i) = false\})$
end

the samples for feature f_j. For a binary response Y and a feature f_j the Gini Index of S is defined as:

$$\begin{aligned} \widehat{G}(S) &= 2\frac{N_{false}}{|S|}\left(1 - \frac{N_{false}}{|S|}\right), \\ N_{false} &= \sum_{s_i} I\left(f_j(s_i) = false\right), \end{aligned}$$

where $|S|$ is the number of all samples at the current node and N_{false} denotes the number of samples for which f_j evaluates to $false$. The Gini indices $\widehat{G}(S_L)$ and $\widehat{G}(S_R)$ for the left and right subset are defined similarly. The Gini gain resulting from splitting S into S_L and S_R with feature f_j is then defined as:

$$\widehat{\Delta G}(S_L, S_R) = \widehat{G}(S) - \left(\frac{|S_L|}{|S|}\widehat{G}(S_L) + \frac{|S_R|}{|S|}\widehat{G}(S_R)\right),$$

where $S = S_L \cup S_R$. From that follows, that the larger the Gini gain, the larger the impurity reduction. Recently (Strobl et al., 2007) showed that the use of Gini gain can lead to selection bias because categorical predictor variables with many categories are preferred over those with few categories. In the proposed framework this deficit is not harmful due to the fact that the features are relations between sampled rectangles and therefore evaluate always to binary predictor variables.

Multiple Object Detection

For multiple object detection in a gray scale image every location on a grid with step size δ is considered as an independent sample s which is classified by the ensemble. Therefore each tree casts a binary vote for s being and object or background. The whole ensemble estimates the posterior probability of being class 1: $RDF(s) = \#\{i|t_i(s) = 1\}/\#\{i\}$, where t_i is the ith tree. This procedure results in an accumulator or probability map for the whole image.

The final centroids of detected objects are retrieved by applying weighted mean shift clustering (Comaniciu and Meer, 2002) with a circular box kernel of radius r. During shifting, the coordinates are weighted by the probabilities of the accumulator map. While this leads to good results in most cases, homogeneous ridges in the accumulator can yield multiple centers with a pairwise distance smaller than r. To this end we run binary mean shift on the detection from the first run until convergence. The radius is predefined by the average object size. If the objects vary largely in size the whole procedure can be employed for different scales. To this end, in accordance with Viola and Jones (2001), not the image but the features respectively the rectangles are scaled.

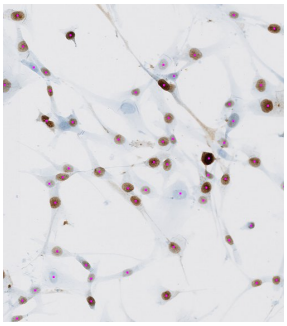


Figure 3.10
Nuclei detection on slides of paraffin embedded cellcultures of primary cancer of aggressive fibromatosis. Antigen identified by monoclonal antibody Ki-67. This gene encodes a nuclear protein that is associated with and may be necessary for cellular proliferation.

Detection Accuracy

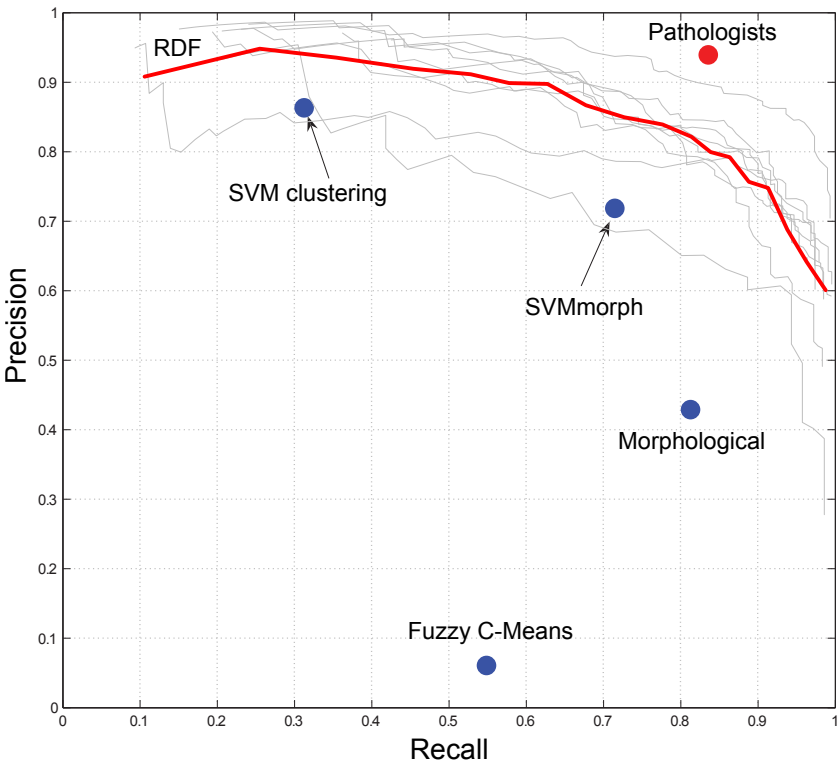
Three fold cross-validation was employed to analyze the detection accuracy of RDFs. The nine completely labeled patients were randomly split up into three sets. For each fold the ensemble classifier was learned on six patients and tested on the the other three. During tree induction, at each split 500 features were sampled from the feature generator. Trees were learned to a maximum depth of 10 and the minimum leave size was set to 1. The forest converges after 150 to an out of bag (OOB) error of approximately 2%. Finally, on the test images each pixel was classified and mean shift was run on a grid with $\delta = 5$.

Figure 3.11 shows precision/recall plot for single patients and the average result of the RDF object detector. The algorithm is compared to point estimates of several state of the art methods: SVM clustering was successfully employed to detect nuclei in H&E stained images of brain tissue by (Glotsos et al., 2005). SVMmorph (Fuchs et al., 2008a) is an unsupervised morphological (Soille, 2003) approach for detection combined with an supervised support vector machine for filtering. The marker for the pathologists shows the mean detection accuracy if alternately one expert is used as gold standard. On average the pathologists disagree on 15% of the nuclei.

Although only gray scale features were used for RDF it outperforms all previous approaches which also utilize texture and color. This observation can be a cue for further research that the shape information captured in this framework is crucial for good detection results.

Figure 3.11

Precision/Recall plot of cross-validation results on the renal clear cell cancer (RCC) dataset. For Relational Detection Forests (RDF) curves for the nine single patients and their average (bold) are depicted. RDF with the proposed feature base outperforms previous approaches based on SVM clustering (Glotsos et al., 2005), mathematical morphology and combined methods (Fuchs et al., 2008a). The inter pathologist performance is depicted in the top right corner.



Unbalanced Data and Limitations of Random Forests

In this section we present results on a publicly available dataset of volcanos on Venus imaged by JPL's Magellan space probe (UCI KDD Archive "volcanoes" data). As described by Burl et al. (1998), all examples were determined from a focus of attention (FOA) matched filter that was designed to quickly scan large images and select 15×15-pixel subimages with relatively high false alarm rates but low miss rates (about $20 : 1$). The number of example images vary from thousands to tens of thousands across the experiments.

We compare two random forests approaches with an SVM model described by DeCoste and Schölkopf (2002). Due to the much larger number of negative examples than positive examples (by factors ranging from 10 to 20), DeCoste and Schölkopf (2002) trained SVMs used SVM^{light}, employing its implemented ability to use two regularization parameters, C^- and C^+, instead of a single C. This was done to reflect the importance of a high detection rate despite the relatively scarcity of positive training examples. Specifically, they used $C^- = 1$ and $C^+ = 3C^-$ and the polynomial kernel $K(u,v) = \frac{1}{512}(u \cdot v + 1)^9$, based solely on manual tuning on the data set.

The images were normalized for SVM and RF as described by Burl et al. (1998). The mean pixel value of each example image is 0 and the standard deviation is 1.

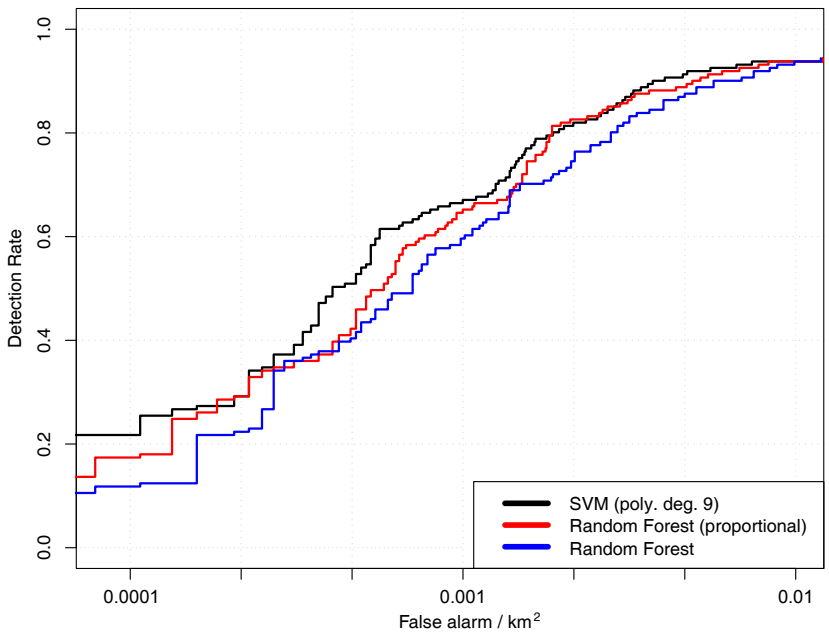


Figure 3.12

Free-response ROC (FROC) curves depicting the cross-validation performance of three classification methods on the Magellan data from Venus. The figure shows the trade-off between detection rate and the number of false alarms per area. The SMV was trained with a polynomial kernel of degree 9 and different Cs for the volcano and background class. "Random Forest (proportional)" comprised 1000 trees and was inferred with 25 randomly selected features per split, a minimum node size of 5, and both classes were sampled proportional for each tree. "Random Forest" is the out-of-the box implementation of Breiman (2001).

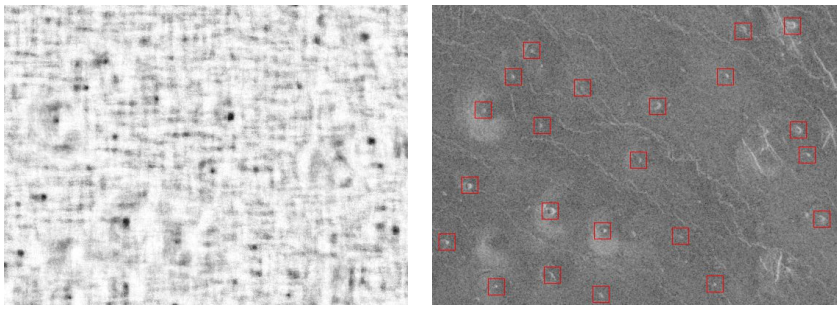
Results from the cross-validation (CV) experiment are depicted in Figure 3.12. The CV folds were pre-defined by Burl et al. (1998). First, it can be seen from these results that neither the out-of-the box implementation of random forests by Breiman (2001) nor the optimized RF model with class-wise proportional sampling can improve over the SVM results. One reason could be slight overfitting of the CV-experiment by optimizing $C^?$, C^+ and the degree of the polynomial kernel. Along these lines it is interesting to note, that further experiments with lower and higher degrees did not improve over the kernel with degree 9 as proposed by DeCoste and Schölkopf (2002).

Class-wise proportional subsampling for tree induction

Second, the class-wise proportional sampling clearly improves over the use of all samples for each tree. To this end, the bootstrap samples for tree induction were taken independently for the foreground class and the background class. For the much larger background class only the equivalent number of samples were drawn as contained in the smaller foreground class. As a result, each tree in the ensemble "sees" only a small part of the background and a different one than the other trees, which has the positive effect of de-correlating the trees in the forest. This reduces the variance of the whole ensemble which leads to the superior results.

Figure 3.13 depicts the object detection results of a RDF model trained with the interactive online learning framework. The qualitatively good results indicate that a larger window size and relational features as described in this Chapter could improve over the SVM model described by DeCoste and Schölkopf (2002).

Figure 3.13
Volcano detection on Venus. Interactive online learning fit of a RDF model to SAR data of Venus from the Magellan probe. (a) Accumulator of the classifier output. (b) Final detections after mean shift clustering.



(a) (b)

Final Remarks

We presented a framework for learning Relational Detection Forests (RDF) for object detection. A comprehensive study was conducted on two different species to investigate the performance of the algorithm in terms of detection accuracy and survival estimation.
The proposed framework is characterized by the following properties: (i) **Simplicity:** It can be used off-the-shelf to train object detectors in near real time for large variety of tasks. (ii) **Novel Feature Basis:** The introduced relational features are able to capture shape information, they are illumination invariant and extremely fast to evaluate. (iii) **Randomization:** The randomized tree induction algorithm is able to handle the intractable large feature space and to take advantage of it by increasing diversity of the ensemble. (iv) **Real World Applicability:** We successfully applied the proposed RDF algorithm to real world problems in computational pathology. We are convinced that the availability of an off-the-shelf object detection framework is of immense benefit for medical research where fast and accurate adaption to a large number of cancer types is indispensable.

3.6 Inter-Active and Online Learning for Clinical Application

3.6.1 Motivation

Day-to-day clinical application of computational pathology algorithms require adaptivity to a large variety of scenarios. Not only that staining protocols and slide scanners are constantly updated and changed but common algorithms like the quantification of proliferation factors have to work robustly on various tissue types. The detection of multiple objects like nuclei in noisy images without an explicit model is still one of the most challenging tasks in computer vision. Methods which can be applied in an plug-and-play manner are not available to date.

3.6.2 Introduction to Online Ensemble Learning

Incorporating the knowledge of domain experts into the process of learning statistical models poses one of the main challenges in machine learning (Vapnik, 1998) and computer vision. Data analysis applications in pathology share properties of online and active learning which can be termed inter-active learning. The domain expert interferes with the learning process by correcting falsely classified samples. Algorithm 9 sketches an overview of the inter-active learning process.

In recent years online learning has been of major interest to a large variety of scientific fields. From the viewpoint of machine learning (Blum, 1996) summarizes a comprehensive overview of existing methods and open challenges. In computer vision online boosting has been successfully applied to car detection (Nguyen et al., 2007), video surveillance (Celik et al., 2008) and visual tracking (Grabner et al., 2008). One of the first inter-active frameworks was developed by (Roth et al., 2008a) and applied to pedestrian detection.

Ensemble methods like boosting (Freund and Schapire, 1996) and random forests (Amit and Geman, 1997; Breiman, 2001) celebrated success in a large variety of tasks in statistical learning but in most cases they are only applied offline. Lately, online ensemble learning for boosting and bagging was investigated by (Oza, 2001) and (Fern and Givan, 2003). The online random forest as proposed by (Elgawi, 2008) incrementally adopts new features. Updating decision trees with new samples was described by (Utgoff, 1989, 1994) and extended by (Kalles and Morris, 1996; Pfahringer et al., 2007). Update schemes for pools of experts like the WINNOW and Weighted Majority Algorithm were introduced by (Littlestone, 1988; Littlestone and Warmuth, 1989) and successfully employed since then.

In many, not only medical domains, accurate and robust object detection specifies a crucial step in data analysis pipelines. In pathology for example, the de-

tection of cell nuclei on histological slides serves as the basis for a larger number of tasks such as immunohistochemical staining estimation and morphological grading. Results of medical interest such as survival prediction are sensitively influenced by the accuracy of the object detection algorithm. The diagnosis of the pathologist in turn leads to different treatment strategies and hence directly affects the patient. For most of these medical procedures the ground truth is not known (see Section 2.5) and for most problems biomedical science lacks orthogonal methods which could verify a considered hypotheses. Therefore, the subjective opinion of the medical doctor is the only gold standard available for training such decision support systems.

(Fuchs and Buhmann, 2009) present an inter-active ensemble learning algorithm based on randomized trees, which can be employed to learn an object detector in an inter-active fashion. In addition this learning method can cope with high dimensional feature spaces in an efficient manner and in contrast to classical approaches, subspaces are not split based on thresholds but by learning relations between features.

Algorithm 1: Schematic workflow of an inter-active ensemble learning framework. The domain expert interacts with the algorithm to produce a classifier (object detector) which satisfies the conditions based on the experts domain knowledge.

Data: Unlabeled Instances $U = \{u_1, \ldots, u_n\}$
%(e.g. image) **Input**: Domain Expert E
Output: Ensemble Classifier C

while *(expert is unsatisfied with current result)* **do**
  classify all samples u_i;
  while *(expert corrects falsely predicted sample u_i with label l_i)* **do**
    update weights of the base classifiers
    learn new base classifiers
  end
end
return C

In such scenarios the subjective influence of a single human can be mitigated by combining the opinions of several experts. In practice consolidating expert judgments is a cumbersome and expensive process and often additional experts are not available at a given time. To overcome these problems online learning algorithms are capable of incorporating additional knowledge, so-called side-information, when it is available.

In an ideal clinical setting, a specialized algorithm for cell nuclei detection should be available for each subtype of cancer. By using and correcting the algorithm several domain experts as its users continuously train and update the method. Thereby, the combined knowledge of a large number of experts and repeated

	promote	demote	promotion op. op_{prom}	demotion op. op_{dem}	drop tree θ_{drop}	update factor β
WM	no	yes		*	0	$0 \leq \beta < 1$
WINNOW1	yes	yes	*	$\leftarrow 0$	0	> 1
WINNOW2	yes	yes	*	/	0	> 0
EWM	yes	yes	*	/	> 0	> 1
ADD	yes	yes	+	-	> 0	> 0

Table 3.2

Parametrization of the update procedure in Algorithm 28 for inter-active ensemble learning. Specific choices lead to known update methods like WINNOW1, WINNOW2 and Weighted Majority (WM) as well as to the two novel approaches Extended Weighted Majority (EWM) and Additive Update (ADD).

training over a longer period of time yields more accurate and more robust classifiers than batch learning techniques.

The described setting differs from the conventional views of online learning and active learning insofar that new samples are neither chosen at random nor proposed for labeling by the algorithm itself. In addition, the adversary is not considered malicious but also not completely trustworthy. The domain expert reacts to the classification of unlabeled data and corrects wrongly classified instances. These preconditions lead to the success or failure of different combination rules.

It has to be noted, that these kind of machine learning approaches are in sharp contrast to classical rule bases expert systems (Hayes-Roth et al., 1983) which are still used by a number of commercial medical imaging companies. For these applications the user has to be an image processing experts who choses dozens of features and thresholds by hand to create a rule set adapted to the data. Contrary to that strategy, in an inter-active learning framework the user has to be a domain expert, in our case a trained pathologists. Feature extraction and learning of statistical models is performed by the algorithms so that the expert can concentrate on the biomedical problem at hand. Inter-active learning frameworks like (Nguyen et al., 2007; Fuchs and Buhmann, 2009) show promising results, but further research especially on long term learning and robustness is mandatory to estimate the reliability of these methods prior to an application in clinical practice.

3.6.3 Ensemble Online Updates

Following online learning notation (Littlestone and Warmuth, 1989) we consider the base classifiers of the ensemble as experts. In addition trees are only induced from a bootstrap Z of recently labeled samples which are collected in a buffer B of fixed size. This restriction avoids retraining of the ensemble on the whole dataset after each iteration, but relaxes the condition that the classifier can only access the last sample.

The proposed update scheme differs from previous approaches mainly in two

points: (i) in contrast to (Utgoff, 1989, 1994; Kalles and Morris, 1996) no incremental induction of decision trees is employed. After a base classifier is induced it is kept fixed and only its weight in the ensemble is changed. The rational for this design choice was the fact, that tree learning on the basis of relational features is extremely fast as described in Section 3.6.5. Learning additional trees is likely to be faster than incrementally updating the old ones. In addition, due to the randomized design, it is desirable to introduce diversity in the ensemble. Inducing trees from new samples in the buffer B is likely to result in less correlation between the base classifiers in the ensemble. (ii) Classical update schemes (Littlestone, 1988; Littlestone and Warmuth, 1989; Elgawi, 2008) for online learning assume the number of experts to be fixed. Compared to these we consider an intractable large pool of experts, namely all possible decision trees induced from relational features. The goal is to design an update procedure which is able to improve the ensemble accuracy by adding new experts and dropping poorly performing ones.

Algorithm 2 provides a general framework for online update schemata. Specific parameterizations lead to known update methods like WINNOW1, WINNOW2 and Weighted Majority (WM) as well as to two novel approaches termed Extended Weighted Majority (EWM) and Additive Updates (ADD) respectively. Table 3.2 summarizes parameter settings for the different algorithms.

The proposed algorithm mainly consists of the iteration between classification of all samples, correction by the domain expert and update of the tree ensemble. As long as the expert is not satisfied with the current result, the ITE algorithm classifies all samples s_i in the focus of the expert and waits for corrections. This focus could be a whole image and the result were the detected objects in the image. In this context a false positive (FP) hit would be background classified as object and a false negative (FN) hit were a missed object classified as background. Next, the domain expert is supposed to correct falsely classified samples. In practice the user clicks on the image, labeling a location either as object or background. This corrected sample s_c is classified by the algorithm and added to the sample buffer B. Then, depending on the update scheme the weights w_i of the trees t_i are updated either always or only when the ensemble has misclassified s_c. Depending on the method, trees get promoted if they were correct or demoted if they were wrong. The promotion operator op_{prom} or demotion operator op_{dem} can either be a multiplication and a division or an addition and a subtraction with a predefined factor β. In addition if a tree gets demoted and its weight drops below a given threshold θ_{drop} it is permanently removed from the ensemble. Finally for each corrected sample s_c a new tree is learned and added to the ensemble by drawing a bootstrap Z of size $|B|$ from B and running the induction procedure `LearnTree`$(Z, 0)$.

Algorithm 2: Inter-Active Tree Ensemble (ITE)

Input: $S = \{s_1, s_2, \ldots\}$
Init: labels $y_i = false$
Init: buffer $B = \{\emptyset\}$
Init: ensemble $ITE = \{\emptyset\}$
Init: tree weights $W = \{\emptyset\}$

while *expert is unsatisfied with current result* **do**
  $\widehat{y_i} = ITE(s_i) \quad \forall i \in$ **expert focus**
  while *expert corrects sample* s_c *with label* y_c **do**
    $\widehat{y_c} = ITE(s_c)$
    $B = B \cup s_c$
    if $(\widehat{y_c} \neq y_c)$ **then**
      foreach *tree* t_j *in* ITE **do**
        if $(t_j(s_c) = y_c)$ *AND (promote)* **then**
          $w_j = op_{prom}(w_j, \beta)$
        end
        if $(t_j(s_c) \neq y_c)$ *AND (demote)* **then**
          $w_j = op_{dem}(w_j, \beta)$
          if $w_j < \theta_{drop}$ **then**
            $ITE = ITE \setminus t_j$
          end
        end
      end
    end
    draw a bootstrap Z **of size** $|B|$ **from** B**.**
    $Z_i = (z_1, z_2, \ldots z_{N_Z})$ **where** $z_i = (x_i, y_i)$
    $tree =$ **`LearnTree`**$(Z,0)$
    $ITE = ITE \cup t$
  end
end

3.6.4 Online Multiple Object Detection

For multiple object detection in a gray scale image every location on a grid with step width δ is considered as an independent sample s which is classified by the ensemble ITE. Therefore, each tree casts a binary vote for s being an object or background. In contrast to WINNOW and WM the algorithm predicts not a binary class label but the probability of being class 1:

$$ITE(s) = \frac{\sum_{i|t_i(s)=1} w_i}{\sum_i w_i},$$

where w_i is the weight for the ith tree t_i. This procedure results in an accumulator or probability map for the whole image. The final centroids of detected objects are retrieved by first thresholding the accumulator at 0.5 and second by employing mean shift clustering with (Comaniciu and Meer, 2002) a circular box kernel of radius r. The radius is predefined by the average object size. If

the objects vary largely in size the whole procedure is employed for different scales. To this end, similarly to (Viola and Jones, 2001), not the image but the features respectively the rectangles are scaled.

3.6.5 Implementation Details

The ensemble learning framework and the graphical user interface were implemented in C# and the statistical analysis was conducted in R (R Development Core Team, 2009). Rectangle intensities were calculated using integral images as described by Viola and Jones (2001). Employing a multi threaded architecture tree ensembles are learned in real time on a standard dual core processor with 2.13 GHz. Inducing a tree for 1000 samples with a maximum depth of 10 and sampling 500 features at each split takes on average less than $500ms$ and can be performed in background while the domain expert corrects false classifications in the image.
To guarantee invariance of the classifier to rotation and flipping for each sample three rotated and two flipped versions of the original patch were added to the sample buffer.
Classifying an image of 1500×1500 pixels on a grid with $\delta = 4$ takes approximately three seconds using the non optimized C# code.

3.6.6 Experimental Setup

To quantitatively analyze the inter-active training process the online handling of the system by the user was simulated. To this end, after each classification step the sets of FP S_{FP} and FN S_{FN} samples were generated. Then 10 samples were randomly drawn from the larger set and a proportional number of samples was drawn from the smaller set. This procedure mimics quite closely the behavior of users who intuitively correct the class with the higher amount of errors. For example if the classifier is to sensitive and outputs a huge number of false positive hits, most expert start correcting FPs and labeling them as background, ignoring the few FN. The remaining procedure was the same as described in Algorithm 28.

General learning parameters were set as follows: buffer size $|B| = 100$ samples; number of sampled features per split $mTry = 500$; maximum tree depth $d_{max} = 10$;

To explore various setting for different update schemes and parameterizations from Table 3.2 the experiments were run in parallel on 100 cores of a $64bit$ UNIX cluster. The performance of a model was evaluated by calculating the area under the F-measure curve over all iterations.

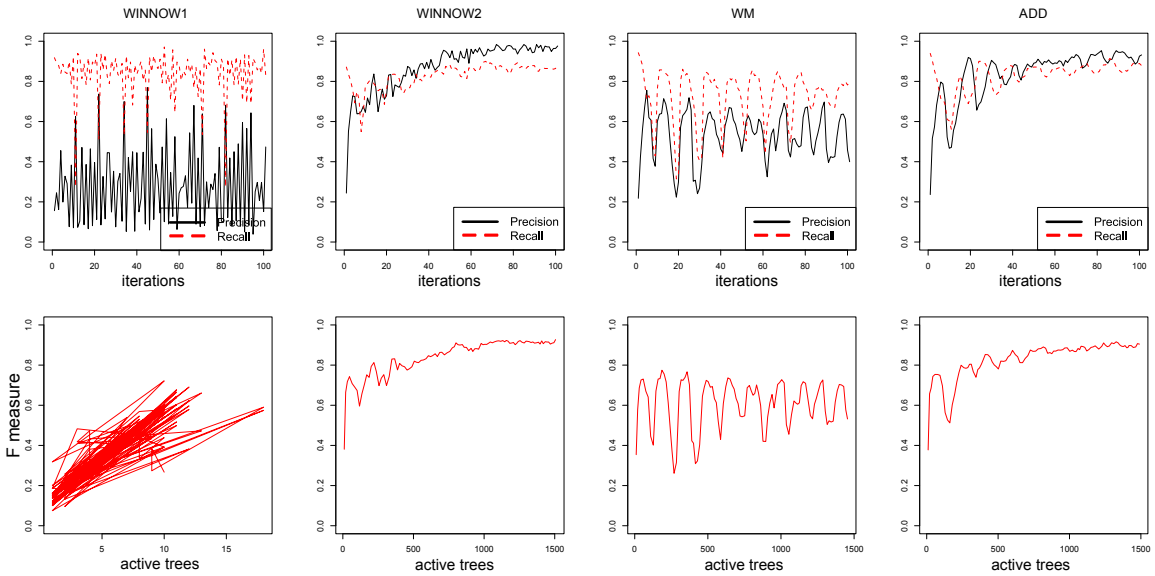


Figure 3.14

The best models for WINNOW1, WINNOW2, Weighted Majority (WM) and ADD. **Top row:** Precision and recall for increasing number of labeling/correction interventions of the domain expert. The rigorous demotion steps in the update rules of WINNOW1 and WM lead to oscillating behavior while the domain expert is trying to compensate for the demoted trees either for negative or positive samples. **Bottom row:** F-measure plotted against the number of active trees. After an heating up phase for WINNOW2 and ADD the accuracy increases with the number of trees in the ensemble. In both methods, poorly performing trees are only weighted down but not removed. In contrast to that, the binary demotion rule of WINNOW1 removes trees from the ensemble if they fail just once. This leads to a behavior where frequently all trees are removed before new ones are learned as depicted in the bottom left plot.

3.6.7 Online Ensemble Learning Results

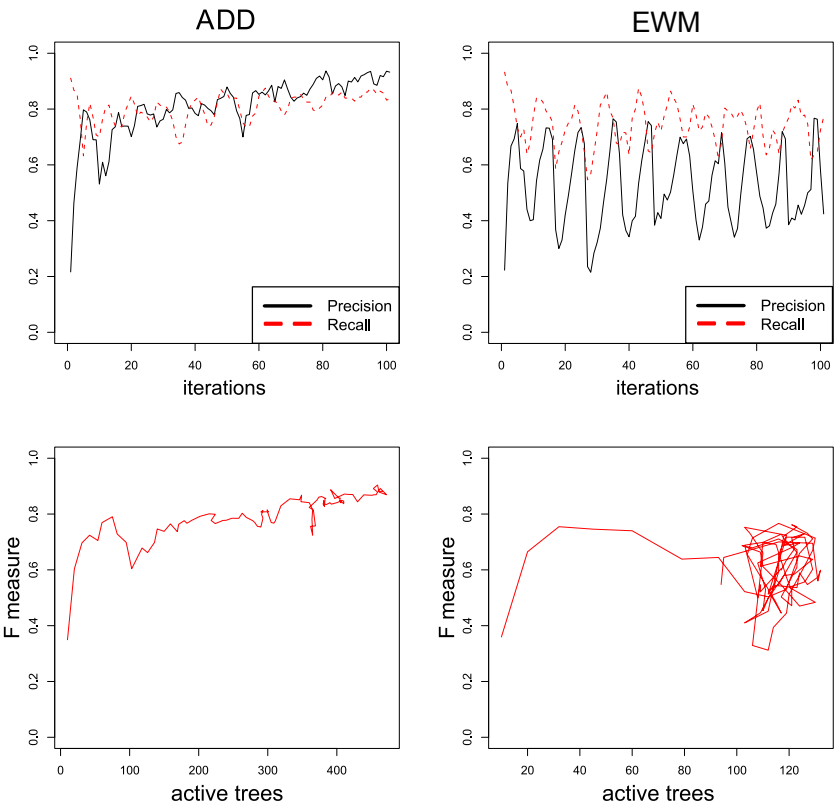
The experimental results for cell nuclei detection on renal cell carcinoma TMA spots are discussed in the following. Figure 3.14 depicts the best models of the parameter search for the different update schemes. Figure 3.15 shows Additive Update (ADD) and Weighted Majority (WM) with ensemble pruning.

WINNOW1 is characterized by very rigorous demotion steps in the update rules. On one hand this leads to extremely small ensembles especially for WINNOW1 but on the other hand it results in oscillating behavior due to the fact that the domain expert is trying to compensate for the demoted trees. For example, if the current ensemble has low sensitivity and therefore a large FN rate, the user starts labeling cell nuclei. At each click the tree weights are updated and if a tree errs once its weight is set to 0 for WINNOW1.

In general Weighted Majority (WM) exhibits the same behavior as WINNOW1. Therefore, we tried to improve WM by activating ensemble pruning, i.e. $\theta_{drop} > 0$ and terming it Enhancing WM (EWM). After a heating up phase of approximately 100 trees EWM is locked in an oscillating state where tree adding and removal cancel each other out. This behavior can be observed in Figure 3.15.

Figure 3.15

Online ensemble pruning for Additive Updates (ADD) and Extended Weighted Majority (EWM). **Top row:** Precision and recall for increasing number of labeling/correction interventions of the domain expert. **Bottom row:** F-measure plotted against the number of active trees. Additive Update (ADD) with ensemble pruning performs as well as a complete ensemble (1500 trees learned by a batch process) but uses less than a third as many trees (470 trees). In comparison Extended Weighted Majority (EWM) performs much worse and exhibits oscillating behavior due to rigorous pruning. The bottom right plot shows that after a heating up phase of approximately 100 trees EWM is locked in an oscillating state where tree adding and removal cancel each other out.



In contrast to the previous three methods WINNOW2 shows a slow but continuous improvement of the combined classifier. The drawback is an ever growing ensemble due to the fact, that no trees are removed. Especially in an object detection scenario where the goal is to exhaustively classify all possible locations in an image the time demand grows linearly with the number of trees.

Finally we examined the performance of Additive Update (ADD) which aims to combine the benefits of all update schemes. Similar to WINNOW2 it uses a demotion and a promotion step but softens the rigorous punishment of the classical approaches by replacing the multiplication and division for op_{prom} and op_{dem} by an addition and a subtraction. This leads to excellent performance and continuous improvement per iteration as depicted in Figure 3.14. In addition if θ_{drop} is larger than 0, Additive Updating leads to much smaller ensembles than WINNOW2.

An insight gained from this experimental study is the notion that cautious update schemes are of benefit for ensembles with high diversity. Methods like the WINNOW algorithm were originally developed for boolean formulas and their promotion and demotion steps are much too severe for ensembles which relay on average performance of their randomized base classifiers.

Qualitative results from the application of the system to a nuclei detection task

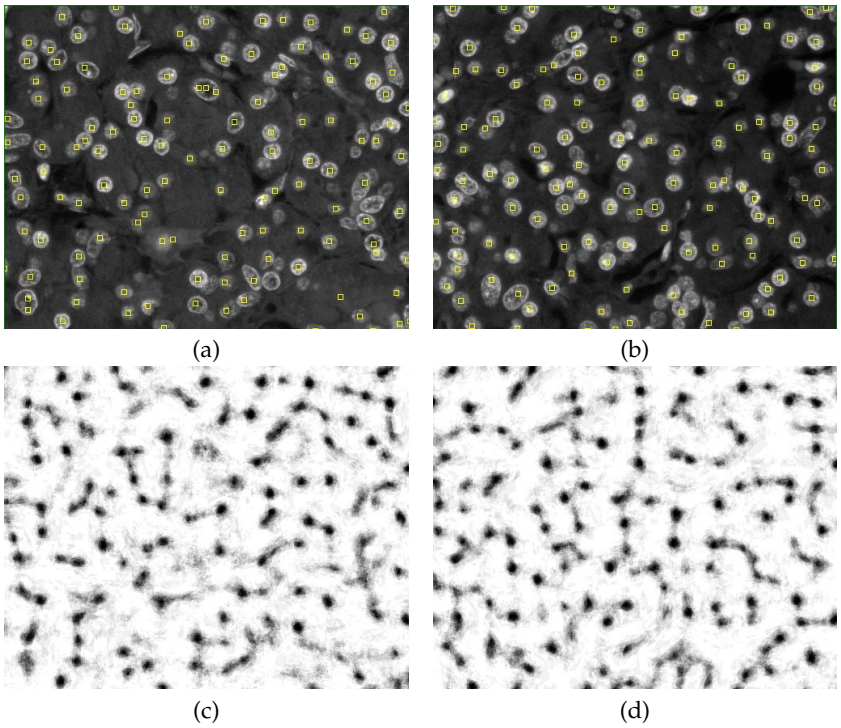


Figure 3.16
Qualitative results from a real world application of interactive learning of relational detection forests (RDF). The domain expert trained the object detector by iteratively clicking on nuclei in the training image (a) and correcting false detections and missed out nuclei. The detection result on an unlabeled test image is sown in (b). The accumulators or confidence maps are depicted for training (c) and testing (d). Final detections after mean shift clustering are shown as yellow rectangles.

in fluorescence imaging are shown in Figure 3.16. The domain expert was able to interactively train a relational detection forest (RDF) which satisfied his demand and the final nuclei counting results were used as read-out to answer the underlying biological question.

3.6.8 Concluding Remarks on Online Ensemble Learning

We presented a framework for learning inter-active tree ensembles (ITE) for object detection. A comprehensive study was conducted to investigate the performance of various ensemble update schemes and to test their compatibility to randomized tree ensembles.

The proposed framework is characterized by the following properties: (i) **Inter-Active:** It can be used off-the-shelf by domain experts to train object detectors in real time in an inter-active manner. (ii) **Novel Feature Basis:** The introduced relational features are able to capture shape information, they are illumination invariant and extremely fast to evaluate. (iii) **Randomization:** The randomized tree induction algorithm is able to handle the intractable large feature space and to take advantage of it by increasing diversity in the ensemble. (iv) **Flexibility:** The framework can be parametrized to use classical update schemes as well as novel approaches such as Additive Updating. (v) **Real World Applicability:** We successfully applied the proposed ITE algorithm to a real world problem

in computational pathology. We are convinced that the availability of an interactive object detection framework is of immense benefit for medical research where fast and accurate adaption to a large number of cancer types is indispensable.

3.7 Multispectral Imaging and Source Separation

Multispectral imaging (Levenson and Mansfield, 2006; van der Loos, 2008) for immunohistochemically stained tissue and brightfield microscopy seems to be a promising technology although a number of limitations have to be kept in mind.

To date, double- or triple-staining of tissue samples on a single slide in brightfield (non-fluorescence) microscopy poses still a major challenge. Traditionally, double staining relied on chromogens, which have been selected to provide maximum color contrast for observation with the unaided eye. For visually good color combinations, however, technically feasible choices always include at least one diffuse chromogen, due to the lack of appropriate chromogen colors. Additional problems arise from spatial overlapping and from unclear mixing of colors. Currently, these problem are addressed by cutting serial sections and by staining each one with a different antibody and a single colored label. Unfortunately, localized information on a cell-by-cell basis is lost with this approach. In the absence of larger structures like glands, registration of sequential slices proved to be highly unreliable and often not feasible at all. Multispectral imaging yield single-cell-level multiplexed imaging of standard Immunohistochemistry in the same cellular compartment. This technique even works in the presence of a counterstain and each label can be unmixed into separate channels without bleed-through.

Computational pathology algorithms would profit from multispectral imaging also is experiments with single stains, due to the possibility to accurately separate the specific label signals from the background counterstain.

Practical suggestions for immunoenzyme double staining procedures for frequently encountered antibody combinations like rabbit-mouse, goat-mouse, mouse-mouse, and rabbit-rabbit are discussed by van der Loos (2008). The suggested protocols are all suitable for a classical red-brown color combination plus blue nuclear counterstain. Although the red and brown chromogens do not contrast very well visually, they both show a crisp localization and can be unmixed by spectral imaging.

Detection and segmentation of nuclei, glands or other structures constitute a crucial steps in various computational pathology frameworks. With the use of supervised machine learning techniques these tasks are often performed by trained classifiers which assign labels to single pixels. Naturally one can ask if MSI could improve this classification process and if the additional spectral

bands contain additional information? A study conducted by (Boucheron et al., 2007) set out to answer this question in the scope of routine clinical histopathology imagery. They compared MSI stacks with RGB imagery with the use of several classifier ranging from linear discriminant analysis (LDA) to support vector machines (SVM). For H&E slide the results indicate performance differences of less than 1% using multispectral imagery as opposed to preprocessed RGB imagery. Using only single image bands for classification showed that the single best multispectral band (in the red portion of the spectrum) resulted in a performance increase of 0.57%, compared to the performance of the single best RGB band (red). Principal components analysis (PCA) of the multispectral imagery indicated only two significant image bands, which is not surprising given the presence of two stains. The results of (Boucheron et al., 2007) indicate that MSI provides minimal additional spectral information than would standard RGB imagery for routine H&E stained histopathology.

Although the results of this study are convincing it has to be noted that only slides with two channels were analyzed. For triple and quadruple staining as described by van der Loos (2008) MSI could still encode additional information which should lead to a higher classification performance. Similar conclusions are drawn by Cukierski et al. (2009), stating that MSI has significant potential to improve segmentation and classification accuracy either by incorporation of features computed across multiple wavelengths or by the addition of spectral unmixing algorithms.

Complementary to supervised learning as described before, Rabinovich et al. (2003) proposed unsupervised blind source separation for extracting the contributions of various histological stains to the overall spectral composition throughout a tissue sample. As a preprocessing step all images of the multispectral stack were registered to each other considering affine transformations. Subsequently it was shown that Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999) and Independent Component Analysis (ICA) (Hyvarinen, 2001) compare favorable to Color Deconvolution (Ruifrok and Johnston, 2001). Along the same lines Begelman et al. (2009) advocate principal component analysis (PCA) and blind source separation (BSS) to decompose hyperspectral images into spectrally homogeneous compounds.

In the domain of fluorescence imaging Zimmermann (2005) give an overview of several source separation methods. The main difficulty stems from the significant overlap of the emission spectra even with the use of fluorescent dyes. To this end Newberg et al. (2009) conduct a study on more than 3500 images from the Human Protein Atlas (Berglund et al., 2008; Pontén et al., 2008). They concluded that subcellular locations can be determined with an accuracy of 87.5% by the use of support vector machines and random forests (Amit and Geman, 1997; Breiman, 2001). Due to the spread of Type-2 diabetes there is growing interest in pancreatic islet segmentation and cell counting of α and β-cells (Herold et al., 2009). An approach which is based on the strategies described in Section

3.3 and Section 3.6 is described by Floros et al. (2009).
It is an appealing idea to apply source separation techniques not only to multispectral imaging but also to standard RGB images. This approach could be useful for a global staining estimation of the separate channels or as a preprocessing step for training a classifier. Unfortunately, antigen-antibody reactions are not stoichiometric. Hence the intensity/darkness of a stain does not necessarily correlate with the amount of reaction products. With the exception of Feulgen staining also most histological stains are not stoichiometric. van der Loos (2008) also state that the brown DAB reaction product is not a true absorber of light, but a scatterer of light, and has a very broad, featureless spectrum. This optical behavior implies that DAB does not follow the Beer-Lambert law, which describes the linear relationship between the concentration of a compound and its absorbance, or optical density. As a consequence, darkly stained DAB has a different spectral shape than lightly stained DAB. Therefore attempting to quantify DAB intensity using source separation techniques is not advisable. Contrary to this observation, employing a non-linear convolution algorithm as preprocessing for a linear classifier, e.g. for segmentation could be of benefit.
Finally, multispectral imaging is not available for automated whole slide scanning which constrains its applicability. Imaging a TMA manually with a microscope and a MSI adapter is too tedious and time consuming.

3.8 Software Engineering Aspects

One of the earliest approaches for high performance computing in pathology used image matching algorithms based on decision trees to retrieve images from a database (Wetzel, 1997). The approach was applied to Gleason grading in prostate cancer. Web-based data management frameworks for TMAs (Thallinger et al., 2007) facilitate not only storage of image data but also storage of experimental and production parameters throughout the TMA workflow.
A crucial demand on software engineering is the ability to scale automated analysis to multiple spots on a TMA slide and even multiple whole microscopy slides. Besides cloud computing one possibility to achieve that goal is grid computing. (Foran et al., 2009) demonstrated the feasibility of such a system by using the caGrid infrastructure (Oster et al., 2008) for Grid-enabled deployment of an automated cancer tissue segmentation algorithm for TMAs.
A comprehensive list of open source and public domain software for image analysis in pathology is available at www.computational-pathology.org.

3.9 Micrometastases Detection in Sentinel Lymph Nodes

3.9.1 Introduction

Acknowledgments
I would like to show my gratitude to Sara Abbasabadi, whose help in conducting this large scale project during her master thesis was indispensable. Sara also conducted most of the experiments in this section and prepared the manuscript for the accompanying publication. Special thanks to Daniela Mihic who formulated the medical questions, labeled numerous lymph nodes and is conducted the clinical study in the University Hospital Zürich (USZ). Furthermore I am indebted to Monika Bieri and Norbert Wey, who digitalized hundreds of slides and produces terabytes of data is the basis for this work.

In this study, we present a fully automated machine learning framework for the analysis of immunohistochemically stained sentinel lymph node tissue specimens of primary melanoma patients. The purpose of our system is to alleviate the tedious, time consuming and error-prone screening of slides which is performed manually by trained pathologists. Therefore we propose a high-throughput automatic decision support system which performs this task in an automated and objective manner with minimal user intervention.

Cutaneous malignant melanoma remains the leading cause of skin cancer-related death in industrialized countries. Over the last decade, it has been one of the most rapidly increasing malignancies in humans (Jemal et al., 2005) and its incidence is predicted to continue rising with the further depletion of the ozone layer. Malignant melanoma is a highly aggressive neoplasm and very small primary tumors (Breslow tumor thickness $< 1mm$) may already produce metastatic disease, leading rapidly to death (Kalady et al., 2003; McKinnon et al., 2003) Metastatic disease, developed by about one third of melanoma patients, is the most powerful adverse prognostic sign. The lymphatic system is involved in about 70% of tumorprogression. 30% of patients with progressive disease develop distant metastases without involving the lymphatic system. The reported 5-year survival rate of melanoma patients with metastatic disease is below 20% (Meier et al., 2002; Reintgen et al., 1994; Balch et al., 2001b). Therefore tumor staging, meaning identifying patients with metastatic disease, is one of the most critical issues in the management of melanoma.

Morton et al. (1992) introduced the method of sentinel lymph node (SLN) biopsy in 1992. SLN is defined as the first draining lymph node in a lymph node area. Morton demonstrated that a negative SLN biopsy has a very high negative predictive value (0.99) for the remaining lymph nodes of an examined basin. As a consequence there is no need for lymphadenectomy, which is associated with a high morbidity, in SLN negative patients. The method of SLN biopsy has been validated and refined in the following years (Reintgen et al., 1994; Balch et al., 2001b). Today, SLN biopsy is the gold standard for the initial staging of melanoma. The presence of metastatic disease in a SLN has the therapeutic consequence of lymphadenectomy. Therefore the working up protocol of a SLN biopsy is extensive and time consuming in order not to miss metastatic disease.

Figure 3.17
Sentinel lymph node tissue slides preparation: Extracted lymph nodes (a) are stained (b) after embedding in paraffin blocks and sectioning at $50\mu m$ intervals. Lymph node tissue slides are depicted in (c), which were stained with haematoxylin and eosin (left) and pan-melanoma (right).

(a) (b) (c)

3.9.2 Tissue Sample Preparation and Scanning

All data described in this study is from patients of the University Hospital Zürich who had a primary melanoma with Breslow tumor thickness $\geq 1mm$. All patients therefore got a sentinel lymph node biopsy. The protocol used is based on the recommendations of the European Organization for Research and the Treatment of Cancer (EORTC) melanoma cooperative group, pathology subgroup: The SLN is fixed in 5% formaldehyde. After fixation the hilar region is identified and the node is divided in two halves along to the long axis to provide a maximum surface area for sectioning. Depending on its size, the bisected node is embedded in one or two paraffin blocks. Six step sections at intervals of $50\mu m$ are done from each paraffin block. From each step section three slides are prepared. The first slide is stained with haematoxylin and eosin, the second is immunostained with Pan-melanoma and the third is stored for potential further investigations if necessary (cf. Figure 3.17). Pan-melanoma is a cocktail, composed of antibodies against melanocytic differentiation antigens HMB 45, MART-1 and tyrosinase.

All specimens are examined by an experienced pathologist. Four different diagnoses are possible, based on the recommendations of the International Union against Cancer: (i) no tumor, (ii) isolated tumor cells, (iii) micrometastasis ($< 2mm$), and (iv) metastasis ($> 2mm$). In addition there may be the diagnosis of an intracapsular nevus cell (cf. Figure 3.18).

For offline analysis, the tissue specimens were scanned with a Nanozoomer C9600 virtual slide light microscope scanner from HAMAMATSU Photonics K.K.. The magnification of $40x$ resulted in a per pixel resolution of $0.23\mu m$. Due to the large size of each tissue slide (more than 20 GB on average), it was tiled into sub images of $2000 \times 2000 \times 3$ pixels size with 200 pixels overlap in the horizontal and vertical direction.

3.9.3 Background and Overview

Computer-assisted object detection has been widely employed in medical imaging, especially in the analysis of mammographic images for breast cancer detection (Helvie et al., 2004) and lung nodules (Reeves and Kostis, 2000; Schilham

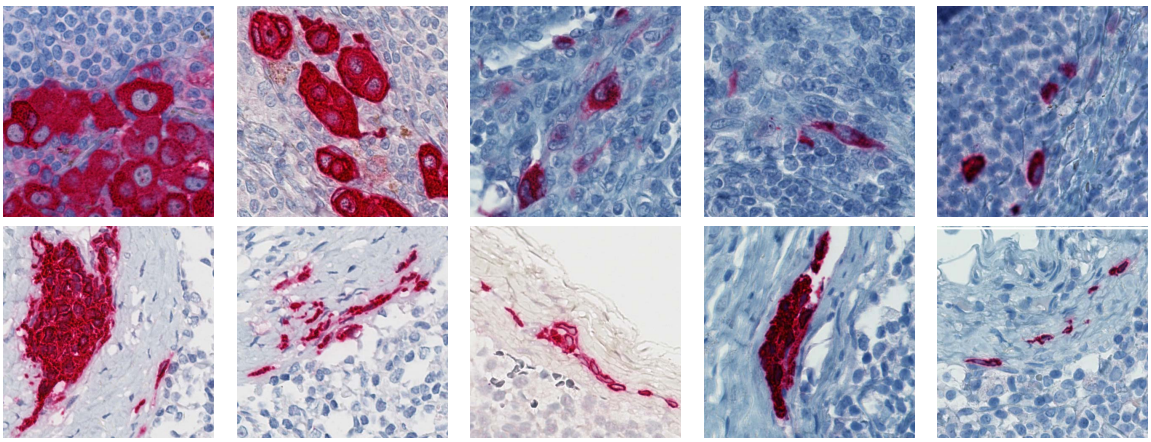


Figure 3.18

Representative detail images of various the four diagnosis types on pan-melanoma stained slides. metastasis (a), micrometastasis (b), isolated tumour cells (c,d,e) and benign nevus cells (second row)

et al., 2006). In recent years, high resolution scanning technologies have made it possible to automatically analysis histological tissue slides and to utilize such systems in routine clinical applications. Recently, it was demonstrated that a classifier can be trained with global features from histological slices to differentiate between breast cancer and normal tissue (Hall et al., 2007; Yang et al., 2007). The authors used filter banks (Helvie et al., 2004) and extracted a texton codebook to derive texture signatures for the tissue microarray (TMA) spots. After dimensionality reduction, the classifier was trained by boosting (Freund and Schapire, 1995). A comparable approach was used by Shotton et al. (2006) for object recognition but extends the setting to a multi-class scenario. Classification using random forest algorithm has been of high interest in the last few years for medical applications (Fuchs et al., 2008b; Maree et al., 2007). It was successfully applied to survival prediction of renal cell carcinoma (RCC) patients using local binary patterns features on immunohistochemically stained tissue microarrays of RCC patients.

Due to the high user dependence of current software for evaluation of lymph nodes (Mesker et al., 2004; Weaver et al., 2003), pathologists typically analyze the tissue slides manually using a light microscope at a low magnification. To our knowledge, this is the first study aiming for automated detection and localization of metastatic melanoma (metastasis, micrometastases and isolated tumor cells) and benign nevus cells on IHC stained lymph node tissue slides. For development and validation, the domain knowledge of pathologists was integrated in two different ways. First pathologist exhaustively labeled a representative set of images, which was used as gold standard to train the algorithms. Second the clinical diagnosis of melanoma patients was used for the large scale evaluation study of the framework.

In the first stage of our pipeline, manually labeled images that contain metastatic melanoma, nevus and healthy tissues, are used. Based on these images, our algorithm learns a model that is capable of screening histological images of the lymph node. The output consists of regions of interest where presumed metastases and nevus cells are located. The object detection task is solved by employing a multiple discriminative analysis (MDA) classifier based on color features for the IHC reaction.

Despite the high sensitivity of Pan-melanoma staining as a marker (Sheffield et al., 2002), it can not discriminate between lymph node nevus cells and metastatic melanoma resulting in low specificity, similar to other recommended markers for diagnosis of melanoma (Pinto, 1986). Nevertheless, this differentiation is highly critical since it affects the treatment planning of the cancer patients. Benign nevus cells are present mainly within the fibrous capsule or trabeculae (extension of the capsule inside the node) (Carson et al., 1996; Murray et al., 2004). However, melanoma cells are located in the not fibrous parenchymal tissue in the sub-capsular region of the lymph node. Using this localization as cue, in the second step of our work, an algorithm has been developed which can discriminate fibrous capsule/trabeculae from the parenchymal tissue of the lymph node based on the tissue architecture. This determines localization of the candidate detections within the lymphatic tissue which in turn categorizes them into two groups of metastatic melanoma or benign nevus. In order to represent the tissue architecture, several textual feature metrics including Local Binary Patterns (LBP), Haar wavelets and filter banks together with color information were examined and the optimal technique was selected. At the local optimization level, random forests were utilized to train a model for classification of the two tissue types. This results in a heterogeneous probability map due to the high intra class tissue variability. To smooth the decision boundaries in the global optimization level, a graph cut segmentation approach was employed.

Three different kinds of validation were performed. First the algorithms were evaluated on an independent validation set to select the optimum feature base. Second, the results of the algorithm were compared against the gold standard labels generated by an expert pathologist. In the third evaluation setting, the performance of the automated framework was tested on a clinical data set of 53 patients by comparing the algorithmic results to the diagnosis made by expert pathologists. Finally, the developed system is integrated into a joint processing pipeline with the University Hospital Zürich and the results are represented in the web viewer software which can be employed by pathologist to compile clinical reports.

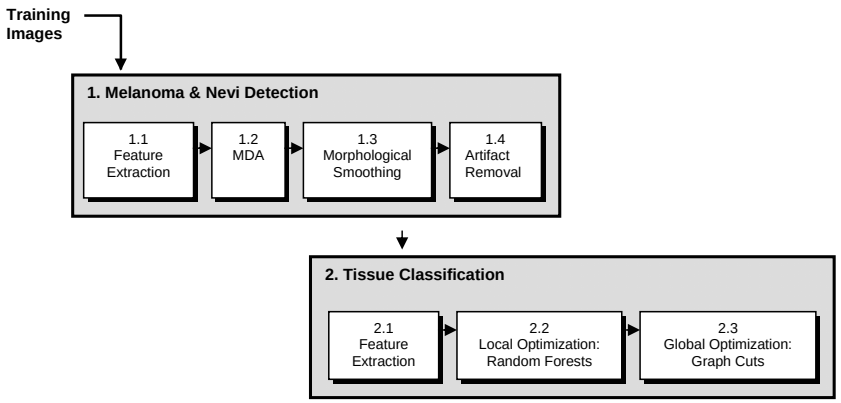


Figure 3.19
Block diagram of the processing pipeline for micrometastases detection and tissue classification in lymph node tissue slides.

3.9.4 Methods

The processing pipeline adopted in this work is subdivided into two main sequential stages: "Melanoma and nevi detection" and "tissue classification" as shown in Figure 3.19.

Melanoma metastases and benign nevus detection

For training, image patches were manually extracted by an expert pathologist from containing metastatic melanoma, lymph node nevus and healthy tissue regions. Three channel color information was employed in pixel wise feature extraction due to the colored IHC reaction at locations where melanoma or nevi cells are present (block 1.1, Figure 3.19). Multiple Discriminant Analysis (MDA) was used to train a classifier to distinguish between normal tissue (lymphocytes, connective tissue, background) and abnormal tissue (melanoma metastases and lymph node nevi). (block 1.2, Figure 3.19). MDA finds the linear transformation that maximizes the ratio of the inter-class scatter matrix determinant to the intra-class scatter matrix determinant of the transformed samples (McLachlan, 2004). This classifier is then employed to classify each pixel of the image. In order to get closed regions for detection and to smooth the results morphological opening and closing operations were applied (block 1.3, Figure 3.19). Next, False positive detections located outside the lymphatic tissue on the slide were considered artifacts and were filtered by intensity thresholding the surrounding pixels of the detected object (block 1.4, Figure 3.19). Resulting from this procedure are bounding boxes enclosing detections of malign melanoma metastases or benign lymph node nevus cells.

Classification of melanoma metastases and benign nevus

The immunohistochemical pan-melanoma reaction is not able to discriminate between malign melanoma metastases and benign nevus. However, this differentiation is highly important in patient's diagnosis and subsequent treat-

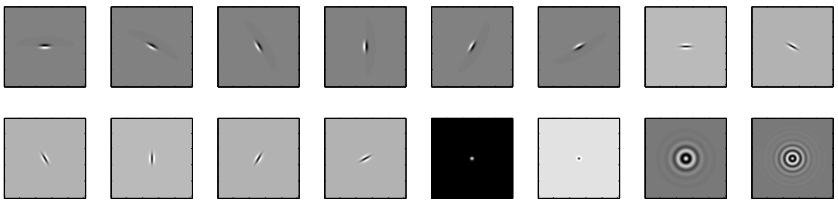


Figure 3.20
Filters used in the presented framework. One edge and one bar filter with $(\delta_x, \delta_y) = (\sqrt{2}, 3\sqrt{2})$ at 6 orientations, a Gaussian and a LOG filter with $(\delta_x, \delta_y) = (\sqrt{2}, 3\sqrt{2})$ from the LM filter bank and two Schmid filters $(\delta, \tau) = \{(10,2),(10,3)\}$ were used leading to 16 filters in total.

ment planning. In clinical practice the localization of the nodal nevi within the lymphatic capsule or fibrous trabeculae is used to discriminate nevi cells from metastases melanoma that occur within the nodal parenchyma. We utilize local information in terms of texture and color for this classification. A dataset of 22713 sample images of size $202 \times 202 \times 3$ taken on a grid with $\Delta = 40$ from each tissue structure (fibrous capsule/trabeculae and parenchymal tissue) was extracted from the images. 80% of this dataset were used for training and testing within a machine learning setting and the remaining 20% were kept as independent validation set. Lymph node tissue slides from nine different patients were used to extract the training samples. During preprocessing, images were resized to half their resolution to increase the performance of the algorithm. Histogram equalization was applied to improve the contrast of the images.

Texture and color feature extraction

The fibrous tissue is mainly composed of elongated cells whereas the lymphocytes are mainly circular and homogeneous . A major challenge is the very high inner class variability of lymphatic tissue is. Three promising candidates for textual feature extraction were examined: linear filter banks, Local Binary Patterns and Haar wavelets. All three were combined with color features and evaluated with respect to their classification performance and efficiency (block 2.1, Figure 3.19). Associated parameters for these techniques were optimized via cross-validation and validated on an additional hold-out set. The optimal technique in respect to accuracy and performance was used for the clinical application.

Filter Banks: Texture is represented by its response to a set of linear filters. A mixture of edge, bar and spot filters at various scales and orientations were applied to exploit region-based texture features due to presence of edge like and circular structures in the lymphatic tissue architecture. The filters were adapted from the Leung and Malik (LM) filter bank (Leung and Malik, 2001) that contains edge, bar, isotropic Gaussian and Laplacian of Gaussian (LOG) filters and the Schmid filter bank (Schmid, 2001) (Figure 3.20). Color information was incorporated by applying the filters to each of the tree channel color spaces: HSV, RGB and Lab. The best performing color space was selected according to the validation experiments.

Local Binary Patterns: Local Binary Patterns (LBP) (Ojala et al., 1996) are reported to have high discrimination rates in texture classification (Ojala et al., 1996; Maenpaa et al., 2000). The LBP operator is resistant to monotonic gray-level transformations making it invariant against lighting variations which often occur in this context. It encodes the fine detail information in the image by considering each pixel in the neighborhood separately. It is fast, rich in information and computationally rather cheap. LBP has also been used successfully in face recognition generating a global facial descriptor (Ahonen et al., 2004). A binary LBP code is produced for each pixel in the image considering the gray values of pixels in a circularly symmetric neighborhood:

$$LBP(x_c, y_c) = \sum_{p=1}^{P-1} s(i_p - i_c)2^p \tag{3.2}$$

where x_c and y_c are the coordinates of the central pixel, c. P is the number of equally spaced pixels forming the circular symmetric neighborhood. i_c and i_p are the gray values at c and P and $s(x)$ is the sign function being 1 if $x \geq 0$ and 0 otherwise. Rotational invariance is achieved by circularly rotating the LBP code until it has the minimum value (McLachlan, 2004).

$$LBP_{ri} = \min\{rot(LBP(x_c, y_c), n) \mid n = 0, 1, ..., P-1\} \tag{3.3}$$

Where LBP_{ri} is the rotation invariant code, $rot(x, n)$ is the function that rotates the LBP code times to the right. To produce uniform LBP codes, a uniformity measure is defined for each neighborhood (Maenpaa et al., 2000):

$$U(I_P) = |s(i_{p-1} - i_c) - s(i_0 - i_c)| + \sum_{p=1}^{P-1} |s(i_p - i_c) - s(i_{p-1} - i_c)| \tag{3.4}$$

The rotation invariant uniform codes LBP_{riu} which have at most two $0-1$ or $1-0$ transitions in the circular binary form are then calculated as

$$LBP_{riu} = \begin{cases} \sum_{p=0}^{P-1} s(i_p - i_c) & U(I_P) \leq 2 \\ P+1 & \text{otherwise} \end{cases} \tag{3.5}$$

Uniformity is an important factor in using the LBP operation, reducing the number of bins in the histogram LBP while preserving most of the information. The resulting feature vector has 14 dimensions.

Haar Wavelets: The wavelet transform technique produces a compact representation of the images. It encodes the shape and edge information from multiple scales. Wavelet features are employed successfully in object detection (Papageorgiou and Poggio, 2000) and face recognition (Garcia et al., 2000; Schneiderman, 2000). Especially the Haar Wavelet Transform (HWT) is a simple, computationally efficient, local image descriptor that generates a non-redundant representation of the image due to its orthogonal basis. Given as input the

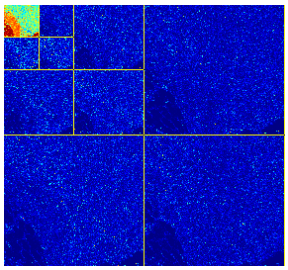


Figure 3.21
Three scale 2-D Haar wavelet transform of a lymphatic tissue image producing four sub-bands. Low frequency coefficients of the third level decomposition is employed in constructing the feature vector.

resized tiles of the size 1000×1000 pixels, wavelets of the same scale were calculated. Figure 3.21 illustrates a 2-D three scale Haar wavelet decomposition of a 2000×2000 lymphatic tissue sub-image containing tissues from both the fibrous capsule and the nodal parenchyma. Four sub-band images are produced: low frequency (approximation) component and horizontal, vertical and diagonal high frequency components. It can be observed that the approximation component of the third level decomposition contains most of the energy and is used for the feature calculation producing a 169 dimensional feature vector.

Classification Framework: Random Forest

Random forests (RF) (Breiman, 2001) have proven to be successful for high-dimensional classification problems in pattern recognition. In this study, RF are used to classify fibrous capsule/trabeculae from the parenchymal tissue based on the feature sets described in the previous section (block 2.2, Figure 3.19). Random forests are an ensemble method consisting of a set of decision trees. Each of these trees is learned from a bootstrap sample of the training data. For each split only a random subset of features is taken into consideration and the one is chosen which maximizes the GINI index (Breiman, 2001). The output of the classifier is the majority vote over all decision trees or the proportion of trees in the forest voting for a class. In this two class scenario this is the probability of a sample belonging to the one class of tissue or the other. Random Forests offer a number of beneficial properties for this problem setting. Besides fast learning and high classification accuracy, they are able to efficiently handle large numbers of samples and input variables. In addition they provide an internal unbiased estimate of the generalization error based on the out of bag (OOB) error. In this work the R package implementation of random forests (Liaw and Wiener, 2002) was used, which is based on the original Fortran code written by Leo Breiman.

Various parameter configurations were examined for the number of trees to build (*ntree*) and the number of variables for the best split at each node (*mtry*). Optimal parameters were selected based on the internal generalization error and their performance in the validation experiments. Accordingly, the recommended parameters were $mtry = 7, 22, 4$ and $ntree = 100, 60, 100$ for the filter bank, LBP and Haar wavelet feature measures respectively. *mtry* is equal to $(\sqrt{number \cdot of \cdot features})$ suggested by Breiman. The classifier is then applied to the tiles containing metastatic disease or benign nevus regions in a sliding window manner, classifying every pixel on a grid with step width $\triangle = 10$ to decrease the computational cost. Finally, the results are resized to the resolution of the original image tile.

Global Optimization: Graph Cut (Max Flow/Min Cut) Algorithm

The local classification described in Section 3.9.4 results in a rather heterogeneous probability map. This is mostly due to the high variability within the different tissue types. In order to get smooth decision boundaries we formulate this global optimization problem as an energy minimization problem on a graph where each node represents a pixel of the tissue image (block 2.3, Figure 3.19). Optimization problems of this kind can be efficiently solved with graph cut algorithms (Boykov and Kolmogorov, 2004). A directed graph is constructed based on the input image, where every pixel is represented by a node in the graph. The nodes are connected by a set of directed edges denoted as n-links. The set of nodes contains two additional nodes for the source and the sink which represent the foreground and the background of the image. Every node in the graph is connected to the terminal source and sink through edges denoted as t-links. The goal is to find a labeling function f that minimizes the energy function:

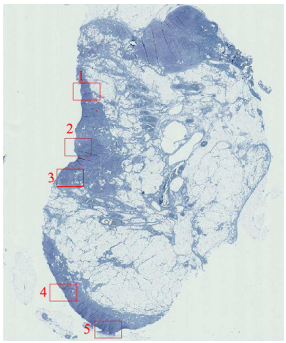


Figure 3.22
Overview of a lymphnode section with five regions selected for expert labeling. Each of the rectangular regions has a size of approximately 6000 × 4000 pixels.

$$E(f) = E_d(f) + E_s(f) \tag{3.6}$$

The data term $E_d(f)$ is the cost of assigning each tissue class to the pixels.

$$E_d(f) = \sum_{x \in I} C(x, f(x)), \tag{3.7}$$

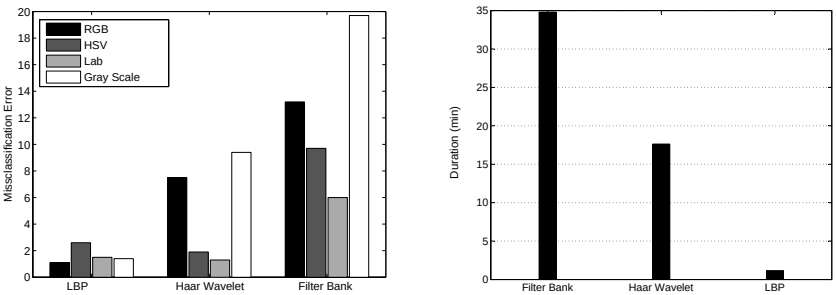
where x represents each pixel in the reference image I and C(x,f(x)) represents the probability of belonging to each tissue class defined by the random forest classifier. The smoothness term $E_s(f)$ controls the smoothness of the results by assigning a penalty for discontinuity between objects. This term is defined based on the intensity difference between two neighboring pixels, x and x':

$$E_s(f) = \sum_{(x,x')} d(I(x), I(x')) \tag{3.8}$$

Maximum flow/minimum cut algorithm (Ford and Fulkerson, 1962) was employed to find the global minimum of the energy function. Graph nodes represent pixels in the image and foreground and background are the two lymphatic tissue classes: fibrous capsule/trabeculae and the lymphatic parenchyma. The optimum cut with minimum energy through the graph is equivalent to finding the maximum flow from foreground to the background. Resulting from this algorithm each pixel in the image is labeled as either object or background with one of the lymphatic tissue classes.

3.9.5 Results

To evaluate the performance and robustness of the detection and classification of the proposed framework, three validation experiments were conducted in addition to the internal generalization error estimated by the random forest



(a) (b)

Figure 3.23
(a) Performance of the proposed feature bases on four different color spaces. Plotted is the misclassification error on the validation set. (b) Time to process a single tile $2000 \times 2000 \times 3$ for all three feature sets.

classifier. In the first validation method, performance of the algorithm was evaluated on an independent validation set. In the second validation experiment the performance of the algorithm was evaluated with respect to the ground truth labels produced by an expert pathologist. The third validation method involves assessment of the framework on a clinical dataset of 53 pre-diagnosed patients from the University Hospital Zurich.

Performance Evaluation on the Validation Dataset

The validation dataset included 4317 image patches of size $202 \times 202 \times 3$ pixels from each tissue class. Classification performance and timing requirements of the three feature measures on the validation set were compared (Figure 3.23a) and the misclassification error rates were calculated. Filter banks combined with Lab color features resulted in a classification accuracy of 94%. Using approximation coefficients of the third level Haar wavelet decomposition combined with lab color information, the correct classification rate increased to 98.5%. LBP features with 12 border pixels on a neighborhood of radius 2.5 used with gray scale information led to 98.6% accuracy. Figure 3.23b shows the duration in minutes to process one image tile of $2000 \times 2000 \times 3$ for each of the feature extraction approaches.
Accordingly, LBP was the least expensive method with high classification accuracy. Consequently LBP features used with gray scale images were chosen for the final model and for further validation.

Comparison to Domain Expert

Generation of the Gold Standard Labels A gold standard was produced by using the knowledge of an expert pathologist. For this purpose, a special labeling tool for lymph node tissue slides was developed (Figure 3.24). Using this software, pathologists can annotate images in SVG (support vector graphics) format and mark melanoma metastases, benign nevi, unspecific staining, dirt and undefined regions. The tool is used on a tablet PC and annotations are performed by the pathologist using a pen.

		Gold Standard	
		M+N+UD	US
Predicted Classes	M+N+UD	TP=146	FP=23
	US	FN=14	TN

Table 3.3
Confusion matrix for the detection algorithm

A trained pathologist from the University Hospital Zürich annotated 33 image patches of size $5000 \times 5000 \times 3$ extracted from lymph node tissue slides of 8 patients. The labels included micrometastases, isolated tumor cell, nevi, unspecific staining, dirt, and undefined immunohistochemical response. In total 160 objects were annotated by the pathologist, 91 of which were melanoma and 32 were nevi. The validations were performed in two steps: first for the detection algorithm and then for the classification algorithm. A confusion matrix was built based on the performance of the classifier which indicates the relationships between the predicted output classes and the gold standard classes. Based on the confusion matrix, true positive (TP), false positive (FP), true negative (TN) and false negative (FN) rates were computed. Basic performance measures used include precision, recall and the F-measure.

Detection Validation In the detection step, Melanoma (M), nevi (N) and undefined (UD) pixels represent true positives and unspecific staining (US) together with the stainless background pixels represent true negatives. The confusion matrix for the detection pipeline is indicated in Table 3.3.
The experimental results of the detection pipeline showed a precision of 86%, a recall of 91% and the F-measure amounted to 88%. In order to interpret these results, it is note worthy that context is an important factor that affects the diagnosis of a pathologist. Based on this factor a faint staining in context of metastases is considered melanoma. However the same degree of staining is not considered melanoma in a healthy tissue context. According to this factor, precision and recall values can be better understood. In computing the recall value some of the false negative objects missed by the algorithm are labeled as positive by the pathologist because of their location in a metastatic context which is also highlighted by the algorithm. The objects with the same degree of staining located in a healthy context are not labeled as metastases by the pathologist. On the other hand, the precision value is also affected by the context factor. Some of the false positive detections would have been labeled as positive by the pathologist when present in a metastatic context.

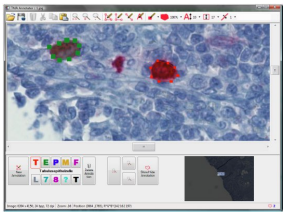


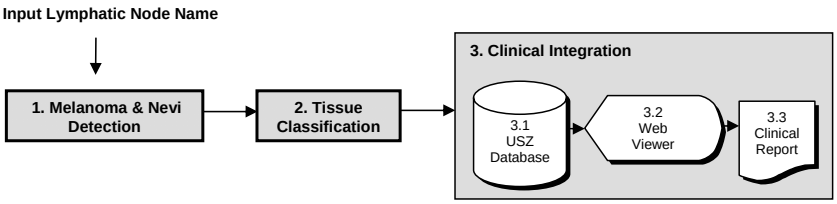
Figure 3.24
Screenshot of the tissue annotation tool used by the pathologist to label image patches. These labels were utilized as gold standard for the comparison of the algorithm to the domain expert.

Classification Validation In the tissue classification step, detections located in the lymphatic parenchymal tissue are considered melanoma and generate true positives. Besides, detections located in the fibrous capsule/trabeculae are considered nevus and represent true negatives. Table 3.4 illustrates the confusion matrix for the classification validation approach.
The classification pipeline gains a precision value of 98%, given a recall of 96%

Table 3.4
Confusion matrix for the classification algorithm

		Gold Standard	
		M	N
Predicted Classes	M	TP=88	FP=1
	N	FN=3	TN=31

Figure 3.25
Block diagram for the integration of the presented object detection and classification framework with the hospital database and web viewer.



and a F-measure of 97%.

Clinical Validation of 53 *Melanoma Patients*

In the third validation scenario, the performance of the pipeline was evaluated in a real world application in order to assess the framework's usefulness in clinical practice. The same parameter settings are applied as in the previous validation experiments. These parameters were tuned during the training phase as described above. In this experiment a dataset of pan-melanoma stained lymph node tissue slides from 53 patients with a size of 805.8GB was used. Half of these patients had a previous diagnosis of melanoma micrometastases and the rest had benign nevi or healthy diagnosis. Each node included 6 tissue slides on average, each slide having a size of approximately 1.8GB. These slides were processed independently. LBP measures were utilized with the random forest detection model on a high performance 64 bit cluster. 100 cores were used to run the computations in parallel. The processing time varies according to the number of tiles containing melanoma or nevus cells in a tissue slide. The processing of a single slide containing 23 to 138 tiles took between 187 to 300 minutes.

By comparing the results of our algorithm and the ground truth (patients' diagnosis), it is revealed that the algorithm accurately predicts metastases on all the 25 nodes from patients with metastatic melanoma. The 7 nodes containing benign nevi cells are also accurately detected. Our framework detected additional melanoma and nevi, undetected by the pathologist in the light microscopy screening. In the review process, during which these cases were presented to the pathologist on the computer screen, it was assessed that the additional findings by the algorithm are of high importance in the follow up procedure. During this procedure, they can be examined by the pathologist according to their morphology, degree of staining, context, similarity to primary tumor and number of occurrence in the slide. According to these criteria they are classified into melanophages, mast cells or tumor cells. In a detailed review effort, the additional findings in eight patients' lymph node slides, which originally were

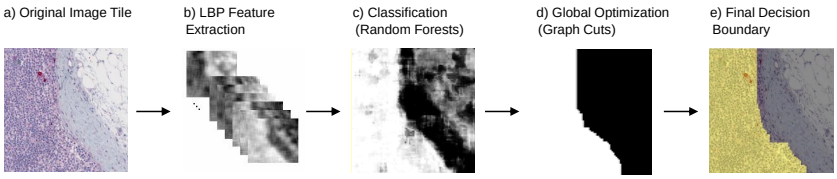


Figure 3.26
Tissue classification stage for an image tile (a), local optimization by random forests following LBP feature extraction (c), global optimization via graph cuts (d) and the overlaid results with the original image (e).

diagnosed as benign were throughly inspected by the expert pathologist. This revision discovered that in five of these patients several of the findings represented malignant regions in the form of metastasis or isolated tumor cells and in one patient represented benign nevi.

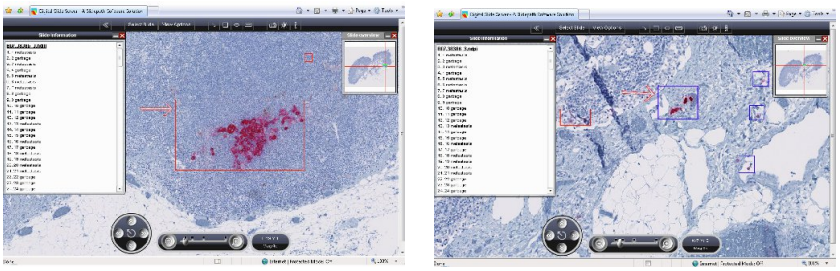
The results demonstrated that the proposed framework is more sensitive than a human expert in accurately detecting malign metastases and benign nevi cells when employed in clinical practice. The specificity on the other hand leaves room for improvement in future research. In addition, the framework speeds up the analysis process by assisting the pathologist in diagnosing melanoma patients.

3.9.6 Clinical Integration

Our system was integrated into a joint processing pipeline of ETH Zürich and the University Hospital Zürich (USZ), as depicted in Figure 3.25.

After extraction of the sentinel lymph nodes, they are sectioned and stained using pan-melanoma IHC staining as described in Section 3.9.1. Subsequently, the tissue slides are scanned using high resolution light microscopy and tiled. In the "Melanoma & Nevi Detection" step (block 1, Figure 3.25), all image tiles of the current lymph node slide are searched for melanoma and benign nevus cells as illustrated in Section 3.9.4. The surrounding tissue is classified on the image tiles which contain melanoma or nevi (block 2, Figure 3.25). Thus, each detection is assigned a tissue label (fibrous capsule/trabeculae or parenchymal tissue) (Figure 3.26) .

Finally, a result file is produced including information about coordinates of the detections regarding the origin in the overview image, tissue class of the detections, their size and color intensity. Accounting for the tiling sequence, the size of each tile and the overlap between the tiles, the exact coordinates are calculated with respect to the coordinate system in the overview image. Size and color intensity of the detections are the statistics of interest for the pathologist to evaluate the state of the lymph node. During post-processing, the detections appearing at the borders of the tiles were merged with their complementary parts in the neighboring tile or tiles and were ranked based on their size and color intensity. The results are then integrated into the hospital database (block 3.1, Figure 3.25) and can be viewed by the pathologist using a software which is able to stream the images via the Internet HTTP (Digital Slide Server, Slide Path Inc.) (block 3.2, Figure 3.25). Two screen shots with annotations resulting



(a) (b)

Figure 3.27
Representative annotations displayed in the web viewer software during integration of the analysis results. (a) The primary detection of the algorithm which is labeled as melanoma metastasis (b) The second detection labeled as nevus.

from our algorithm are illustrated in Figure 3.27 representing examples from both types of tissues.

3.9.7 Discussion and Conclusion

In this work, an automated image analysis system has been developed, evaluated and clinically implemented for the joint detection and classification of metastatic melanoma and benign nevus cells on immunohistochmically stained sentinel lymph node tissue specimens.
In summary, the framework proposed is characterized by the following properties: (i) *High Accuracy:* Extended evaluation demonstrated that the algorithm achieves high precision and recall with respect to the gold standard generated by trained pathologists. (ii) *Robustness:* The application of the framework to 411 tissue slides from different patients without any previous data filtering confirmed the robustness of the system and showed that it is able to cope with biological and technical variability in real world data. (iii) *Autonomy:* Due to extensive usage of machine learning techniques throughout the pipeline, user intervention is only required during the training phase to learn the parameters for feature extraction and classification. (iv) *High Throughput Capability:* to the best of our knowledge this is the first study, that demonstrated that automated micrometastases detection and nevus classification is possible in a high throughput manner. To that end more than $800GB$ of image data was processed in parallel to achieve that goal. (v) *Clinical Applicability:* one of the main contributions of this study is the incorporation of the described decision support system into the clinical workflow. Analyzing hundreds of slides from 53 patients, the system advised the correct diagnosis for all cases with melanoma metastases and presented unexplained regions of tissue for ambiguous cases. We are convinced that the proposed system not only accelerates lymph node analysis by assisting pathologist, but also increases the accuracy of the whole clinical workflow and thus directly benefit melanoma patients.

3.10 Nuclei Detection as Precursor for Robust Pancreatic Islet Segmentation

3.10.1 Introduction

Acknowledgments
I am very thankful to Xenofon Floros, who conducted the segmentation experiments based on the nuclei detection results and who prepared the larger part of the published manuscript. I also want to thank Markus Rechsteiner, who prepared the biological samples and labeled the pancreatic islets.

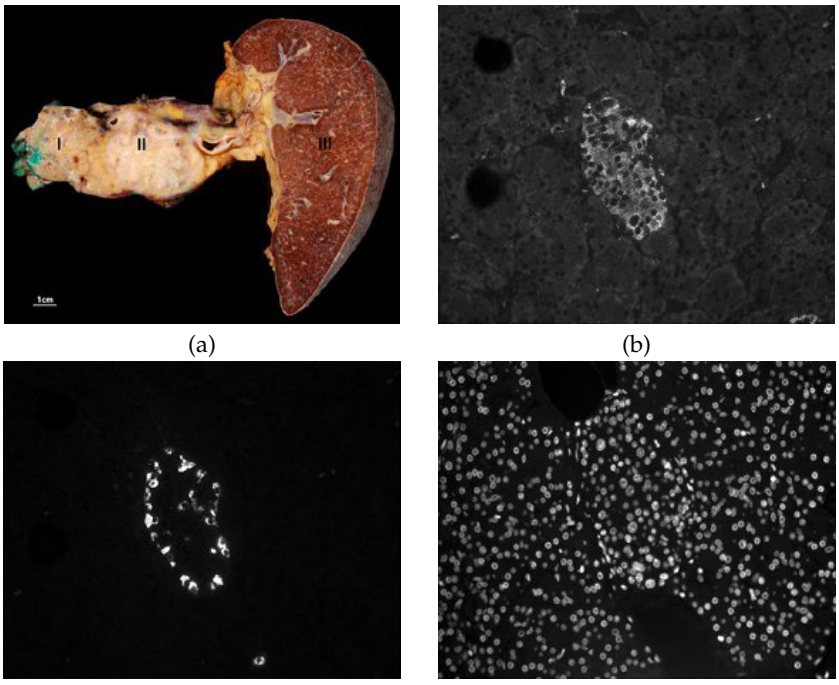
The computational pathology framework presented in this work aims at automated segmentation of type 2 diabetes mellitus (T2DM) islets. T2DM is a chronically progressive disease which is characterized by hyperglycaemia, insulin resistance, and insulin deficiency. Taken together, these factors lead to organ failure and the increased risk for cardiovascular diseases (Kasuga, 2006). It is estimated that by 2010 there will be more than 220 million patients suffering from T2DM (Zimmet et al., 2001). Thus, the search for diagnostic and treatment of this disease is pushing forward at tremendous speed not only in academia but in pharmaceutical industry as well.
Currently, the diagnosis of T2DM includes the measurement of Hemoglobin (Hb) A1c (normal: 4-6%, metabolic syndrome: 6-7%, T2DM: $> 7\%$) which reflects the blood sugar levels (Vijan et al., 1997). As hyperglycaemia is a marker for the progressed disease status most of the patients are diagnosed as T2DM when the disease has already manifested. Therefore, the search for early T2DM and pre-diabetic markers is of urgent need. From mouse and rat models it is known that the pancreatic islets of Langerhans, which consists of beta cells for insulin production, increase in size to compensate the additional demand for insulin to maintain normoglycaemic blood levels (Maedler et al., 2006). The same situation seems to be true in humans as summarized in (Bonner-Weir and O'Brien, 2008), hence indicating that pancreatic islet features could be used as early T2DM markers.

Motivation:

New targets for T2DM prediction, prevention, and treatment generated by the available in vitro and in vivo animal models need to be quantitatively analyzed by high-throughput screening and later verified in human tissue. Therefore a computational pathology approach is necessary to be adopted. The aim of an automated analysis pipeline is first the detection of cell nuclei and second the segmentation of human pancreatic islets based on specific staining for α and β-cells. The robust segmentation of islets is the basis for further quantification of biomarkers regarding T2DM in human patients. Having correctly isolated the islets, it is possible to extract features (e.g. area of the islet) and test their ability to differentiate early T2DM patients and control cases.
Automated analysis of fluorescence images of human tissue poses two main difficulties which are going to be addressed in this work. (i) The 3D structure of the tissue leads to the problem that cell nuclei are not always perfectly cut in their maximum dimension producing numerous cutting artifacts. It has to be noted

Figure 3.28

Human pancreatic tissue: Primary tissue is taken from either whole pancreas from autopsies (I: normal pancreas, II: adenocarcinoma, III: spleen) (a) or biopsy resectates. Sections from fixed and embedded tissue blocks are stained with specific antibodies for β-cells (b), α-cells (c) and DAPI (d). Pictures b-d were taken with a magnification of 20x.



(a) (b)

(c) (d)

that this does not happen in applications with cell cultures or on blood smears for which the large majority of image processing tools in this field is developed. This problem is addressed in the presented work by training a robust classifier for object detection in contrary to using morphological or watershed based approaches. (ii) Variations in the production process of the histological slices can lead to areas of different thickness within one section. This preprocessing artifact produces not only blurred regions in the image but also illumination variation which are even worsened by variations in the fluorescent staining process. These problems are tackled by first using illumination invariant features for the classifier, second by employing clustering for the α and β-cell classification and third by facilitating a graph-based approach for the islet detection.

Tissue Preparation and Imaging:

Human pancreatic tissue from either autopsies or biopsies were formalin fixed and paraffin embedded. Sections were cut at a thickness of $2\mu m$ and stored at 4^o Celsius till use.

For immunofluorescence, sections were deparaffinized and stained with antibodies specific for α and β-cells. Furthermore DAPI staining (DAKO, Carpenteria, CA) was used to label the cell nuclei. Fluorescence pictures were taken with a resolution of $1376 \times 1032 \times 3$ pixels and 20x magnification. Raw unedited material was used in the analysis (Figure 3.28).

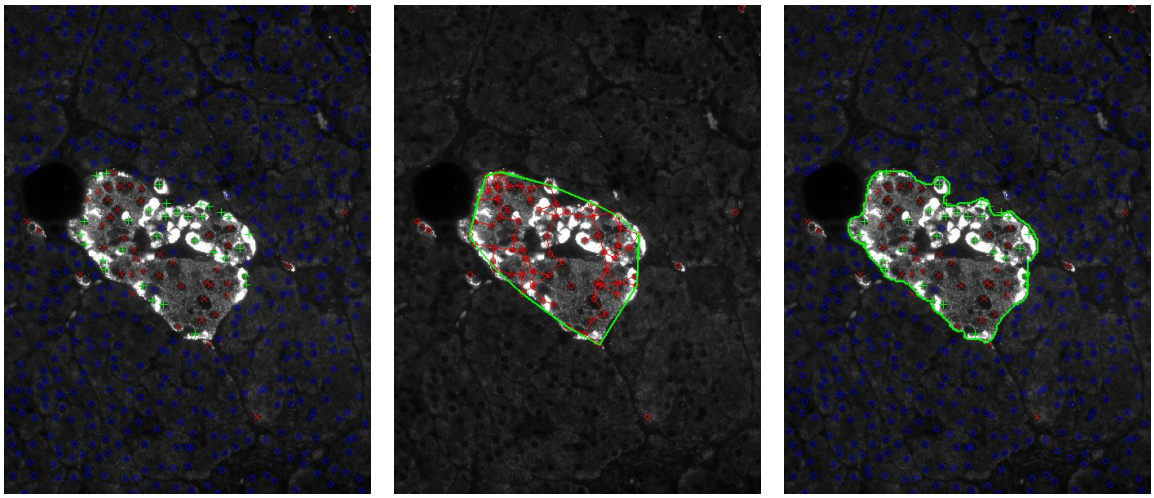
Problem Formulation:

From a computational viewpoint the input to the pipeline consists of three fluorescence images: (i) DAPI-Channel $\rightarrow I_D$ (staining specific for cell nuclei detection), (ii) Alpha-Channel $\rightarrow I_\alpha$ (staining specific for α-cells) and (iii) Beta-Channel $\rightarrow I_\beta$ (staining specific for β-cells).
The output of the algorithm is the segmented area that the pancreatic islet of interest occupies. Prior information from expert pathologists is incorporated in order to guide the search for a meaningful extraction of the islet area. The domain knowledge can be summarized in two main hypotheses:
$\mathcal{H}_1$: *The islets are defined as an area with high density of α and β-cells*, with the α-cells being more specific in specifying the islet area.
$\mathcal{H}_2$: *There is only one islet of interest per image*. In most of the images additional structures are observed, such as smaller islets, disrupted islets or outliers due to staining failures. The main goal is to extract only the dominant islet in each image while excluding artifacts.
The distinct steps of the computational pathology pipeline are described in detail in the next Section.

3.10.2 Methods

Cell Nuclei Detection

Cell nuclei on DAPI stained images, I_D, are detected by following the approach in (Fuchs et al., 2008b) which showed excellent results on histopathological tissue with immunohistochemical staining. To generate a set of positive and negative training patches of size 65×65, a domain expert labeled two images from different patients. In addition to the selected cell nuclei their rotated and flipped counterparts were added to the positive training set. The negative class was down sampled to have a balanced training set. In total 1214 positive and 1214 negative samples were used for training.
For each of these samples a feature vector of length 281 was generated consisting of local binary patterns (LBP) (Ahonen et al., 2004) and a histogram of gray scale values. A great advantage of LBPs for this application is that they are illumination invariant, i.e. invariant with respect to monotonic gray scale-changes and therefore no gray-scale normalization or histogram equalization is needed. Based on these features a random forests classifier (Breiman, 2001) was learned to differentiate between cell nuclei and background. A random forest classifier consists of a collection of tree-structured classifiers $\{h(\mathbf{x}, \Theta_k), k = 1, \ldots\}$ where $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\mathbf{x}$. Random forests posses a number of advantages over classical boosting approaches to object detection as described by Breiman (2001) and (Fuchs et al., 2008b). One of them is the internal out of bag (OOB) error which provides an unbiased estimate of



(a) (b) (c)

Figure 3.29

(a) Nuclei detection results with subsequent classification (green crosses correspond to alpha cells, red x's to beta cells and the blue circles to the normal cells). (b) Graph construction and initialization of the convex hull. (c) Islet segmentation with active contours.

the generalization error. For this application the classifier converges to an OOB error of about 3% after 25 trees.

Finally, to detect the nuclei we classified each pixel of the DAPI test images to generate an accumulator map with a probability at each pixel for being a cell nucleus or not. After non maxima suppression the detections within a range of 20 pixels were clustered to one final hit which is approximately the size of an average nucleus. The output of this step consists of a list of the coordinates of all detected cell-nuclei, $\mathbf{x_i} \in \mathbf{R}^2,\ i = 1, \ldots, N$.

Cell Nuclei Classification

The two channels accounting for the staining of α and β cells, I_α and I_β respectively, are segmented into background and staining using k-means clustering (with k = 2 classes) on the intensity histograms. In order to classify each detected nucleus from step 3.10.2, a neighborhood of 10×10 pixels at the nucleus center is considered. The nucleus is classified based on a majority voting scheme of the segmented binary pixels in the patch of each channel I_α, I_β. If there is strong evidence provided from the segmented staining of channel I_α (I_β) then the nucleus is classified as α-cell (β-cell), otherwise we characterize it as "normal" cell. Thus, tuples of coordinates plus labels for all detected cells of the previous step are obtained : $(\mathbf{x_i}, y_i),\quad y_i = \{\alpha, \beta, n\}$ This approach mimics the workflow of the pathologists by first detecting all cell nuclei and then classifying them to their respective classes based on the intensity of the class-specific

staining around each nuclei.

Graph Construction

Based on the main hypothesis $\mathcal{H}_1$, a neighborhood graph on the identified α and β-cells is constructed, in such a way that clusters of cells correspond to connected components of the graph. Regions of the image with high cell density will be represented by a unique connected component in the graph. This construction is motivated in (Maier et al., 2007), where theoretical aspects of clustering with nearest-neighbor (NN) graphs are explored. In general this task can be solved either by constructing a knn graph or an ϵ-neighborhood graph. Empirical results showed that for this specific task of islet detection, the latter graph performed better, mainly because $\epsilon \in \mathbf{R}^+$ allows for more flexible structures.
Hence, given the set V of α and β-cells detected in the previous steps the ϵ-neighborhood graph $G = G_{eps}(V, \epsilon)$ is constructed, such that two nodes $\mathbf{x_i}, \mathbf{x_j}$ are connected with an edge iff $||\mathbf{x_i} - \mathbf{x_j}||_2 \leq \epsilon$. The euclidean distance is an intuitive choice for this problem setting, because it captures the local structure of cell proximities in the images.

Islet Selection and Segmentation

Given the constructed graph G_{eps}, clusters of α and β-cells are identified by isolating the connected components of the graph. Under the hypotheses $\mathcal{H}_1$ and $\mathcal{H}_2$, the largest cluster corresponds to the islet of interest. Therefore the largest connected component G^{islet} of graph G is extracted, as a first crude approximation of the islet area. This first approximation depends on the parameter ϵ of the graph construction, and can be viewed as the computational equivalent of an expert focusing in the densely stained regions of the image trying to get a first impression of the islet location. Furthermore, it acts on the nuclei level and not on the staining intensities, thus resulting in higher robustness.
Based on the crude estimation of the islet boundaries, an active contour scheme is employed, in order to refine the detected islet area. The basic idea in active contours (Kass et al., 1988), is to evolve a curve, subject to constraints based on the given image, in order to detect objects in the image. In the proposed pipeline, we apply the model described by Chan and Vese (2001), which does not use an edge-detector to stop the evolving curve in the boundary, hence does not depend on the gradient of the image. Furthermore it is shown to be quite effective under the presence of noise and does not require any preprocessing (e.g. smoothing) of the initial image (Chan and Vese, 2001). As motivated above, we initialize the curve on the convex hull of G^{islet} and apply it on the superposition of the two stained channels $I_\alpha + I_\beta$ in order to refine the boundary of the islet. The active contour model used, is governed by two parameters, (s, r), with $s \in \mathbf{R}^+$ controlling the smoothness of the active contour and $r \in \mathbf{N}$ the

number of iterations. The proposed initialization of the active contour is beneficial in two ways: (i) Active contours schemes are known to be sensitive in the curve initialization. A meaningful initialization is provided, tailored to the specific problem and (ii) starting close to the islet boundary also reduces the computation time needed. Figure 3.29 depicts the major stages of the workflow for a single islet.
The algorithm outputs a binary mask, I_{islet}^{seg} (of the same size as the input channels), which corresponds to the detected area of the human islet. The whole pipeline is governed by a tuple of parameters $` = (\epsilon, s, r)$. Based on this segmentation it is possible to automatically extract all biologically meaningful features that can be used as predictive markers for early T2DM, e.g. islet area, staining intensities, fractions of α and β-cells in the islet. Furthermore, the automatically extracted segmentation results are compared with manually segmented islets from expert pathologists in order to assess the algorithm performance against an objective ground truth.

Baseline method

According to our knowledge, there are no published approaches to the specific problem of pancreatic islet segmentation on histopathological tissue. The absence of a competing method was partially compensated by the construction of a baseline morphological approach which also exploits the prior knowledge on the islet segmentation, captured by hypotheses $\mathcal{H}_1$ and $\mathcal{H}_1$. The steps of the baseline method can be summarized as follows: (i) smooth the input staining $I_1 = I_\alpha + I_\beta$ using a Gaussian filter, (ii) globally threshold I_1 to obtain a coarse segmentation, thus $I_2 = I_1 \geq t$, (iii) remove small holes by calculating the closing of $I_2 \rightarrow I_3$, (iv) extract the biggest contiguous region from I_3 and return this as the detected islet I_{islet}^{seg}. An alternative version has an extra step (v) where in addition an active contour is initialized with I_{islet}^{seg} as in step 3.10.2 of the proposed pipeline. Similarly to the proposed approach the main parameters form a tuple $` = (t, s, r)$.

Statistical Evaluation

Given the binary mask of the algorithmic segmentation (I_{islet}^{seg}) and the manually segmented islet from the expert pathologist (I_{islet}^{man}) the cell nuclei agreement between the two masks is calculated. For example, $TP = \#cells \in \{I_{islet}^{seg} \wedge I_{islet}^{man}\}$, $FP = \#cells \in \{I_{islet}^{seg} \wedge \neg I_{islet}^{man}\}$ etc. From the error counts we extract common evaluation metrics, such as precision, recall and F-measure.

3.10.3 Results

The training set consists of 18 triplets of images (three stained channels) corresponding to two patients with T2DM and one control case. Four expert pathol-

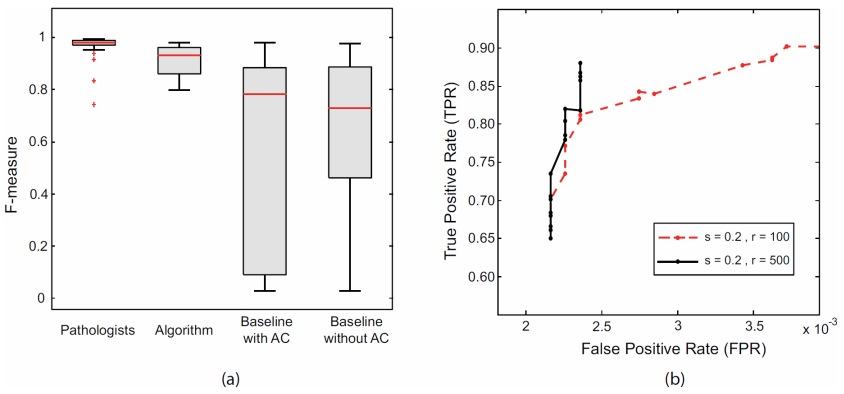


Figure 3.30
(a) F-measure box plots (from left to right: pathologists, proposed algorithm, baseline with active contours (AC) and baseline without AC). The proposed pipeline performs comparably to the pathologists in terms of F-measure when compared to the expert consensus. Furthermore, the test error variance is low. Both baseline methods fail to achieve consistent segmentations as they perform well in some instances, but fail to segment properly a large number of cases in each cross-validation fold. (b) ROC curves for parameter ϵ of the graph construction, keeping s fixed and equal to 0.2 and setting r equal to 100 (dashed line) and 500 iterations (continuous line).

ogists independently segmented the islet of interest for each of the images in the training set. For each one of the 18 training cases we calculated the consensus over the 4 experts, thus obtaining a consensus ground truth. This enables us to compare the performance of the algorithm against a "consensus" expert, but also to estimate the intra-pathologist labeling agreement.

A cross-validation scheme was employed to compare the algorithmic approaches to the gold standard, i.e. the consensus of the pathologists. More specifically, a 3-fold cross-validation is used, where in each fold the algorithm is trained on 12 cases (choosing the parameter values that minimize the error) and then the generalization performance is tested on the other 6 that are left aside as a validation set. For the pathologists, each one of them is compared against the consensus segmentations. It has to be noted that since the pathologists' segmentations are considered as the gold standard, the computational pathology approaches cannot perform better than the experts. The results are depicted in Figure 3.30a.

Regarding the individual experts' annotations we observe that they are very close to the consensus ground truth (with an average F-measure of 0.97) and exhibit quite low variance. Such a high performance is expected since by construction the ground truth labels are computed by averaging the individual ones. The proposed algorithm performs comparably to the pathologists (with an average F-measure of 0.92 across folds) keeping also the variance in a reasonable range. On the other hand both baseline methods (with and without the active contour module) are outperformed by the graph-based segmentation in terms of the F-measure. Furthermore, we observe that the baseline segmentations exhibit high variance, which indicates also a tendency to generalize poorly to new data.

In Figure 3.30b a specific instance of a ROC curve is plotted to evaluate the performance of the proposed algorithm with respect to parameter ϵ which controls the graph construction and thus the initial key step of islet segmentation. More specifically parameter s, which controls the smoothness of the boundary, is kept fixed and for two values of parameter r, the number of iterations the active con-

tour is updated ($r = 100, 500$), the true positive rate (TPR) is plotted against the false positive rate (FPR) over a wide range of parameter ϵ. At a first glance a complex behavior is observed for the large number of iterations in the active contour evolution ($r = 500$). For increasing values of ϵ, vertical ascents are observed in the plot where FPR stays the same and TPR increases. Furthermore, for some sequential increases of ϵ the FPR increases while TPR decreases, a behavior which is not usually observed in ROC curves. Both phenomena can be explained if we keep in mind that parameter ϵ does not directly affect the final segmentation, since the active contour based boundary refinement is applied in between. Increasing ϵ adds more nodes to the graph, thus increasing the initialization area. However if the active contour algorithm performs an adequate number of iterations it will dominate and converge to the islet, hence filtering out the false positive cells. A more balanced behavior is observed for a smaller number of iterations, where for increasing values of ϵ, in most of the times, both TPR and FPR are increased.

3.10.4 Conclusion

The computational pathology system presented in this work is able to, objectively and automatically, estimate the boundaries of human pancreatic islets. The whole pipeline is transparent, modular and based on explicit hypotheses describing the domain knowledge. To the best of our knowledge this is the first framework that successfully tackles this specific segmentation problem. Cross-validation results indicate that the algorithm performs competitively to human experts. Having a reliable pipeline to detect and isolate pancreatic islets from human histological tissue, enables researchers to test specific hypotheses regarding T2DM. We are convinced that the proposed framework can be the basis for further research regarding T2DM and that it can significantly assist the search for diagnostic and therapeutic markers.

3.11 Proliferation in Murine Liver Tissue

3.11.1 Introduction

Acknowledgments
I would like to show my gratitude to Johannes Haybäck, who was not only the key collaborator for the experiments presented in this section, but who was also the first beta tester of the inter-active learning application used in this study.

Inflammatory disorders of the liver can be classified by histological analysis where exact counts of proliferating organ specific cells is important for various studies. We propose a completely automated image analysis pipeline to count cell nuclei that are indicated by a proliferation marker (MIB-1) based on the analysis of immunohistochemical staining of mouse tissues. Most laboratories that are dealing with evaluation of immunohistochemically stained tissue specimens are confronted with very tedious, time consuming and thereby prone to error analysis. Current image analysis software requires extensive user interaction to properly identify cell populations, to select regions of interest for scoring, to optimize analysis parameters, and to organize the resulting raw data. Due to these facts in current software, typically pathologists manually assign a composite staining score for each spot during many microscopy sessions over a period of several days.

3.11.2 Sample Preparation and Data Generation

The tissue blocks were generated in a trial at the Department of Pathology from the University Hospital Zürich. Murine hepatic tissue from independent experiments were formalin fixed and paraffin embedded. Sections were cut at a thickness of $2\mu m$, stained with the MIB-1 (Ki-67) antigen and stored at 4^o Celsius till use. Slices from the murine tissue block and the RCC TMA were scanned on a Nanozoomer C9600 virtual slide light microscope scanner from HAMAMATSU. The magnification of $40\times$ resulted in a per pixel resolution of $0.23\mu m$. Finally 11 patches of size 2000×2000 pixel were randomly sampled from whole tissue slides of each of the 8 mice.

3.11.3 Results

To validate the performance of the Relational Detection Forest presented in Section 3.5.3 on diverse tissue samples and different species we conducted an experiment with liver tissue from eight mice. In addition this experiment was designed to learn the difference between non parenchymal cells and parenchymal, organ specific cell types in order to distinguish between organ prone proliferation and inflammatory cell derived positivity for MIB-1. Figure 3.31 shows a comparison of MIB-1 staining in 88 images of murine liver tissue between case and control groups. Although the algorithm is as well capable as the pathologist in differentiating between case and control ($p < 0.001$) the absolute estimates of stained nuclei are much higher. This is mostly due to wrong hits in area of high lymphocyte density as shown in the tissue image in Figure 3.31.

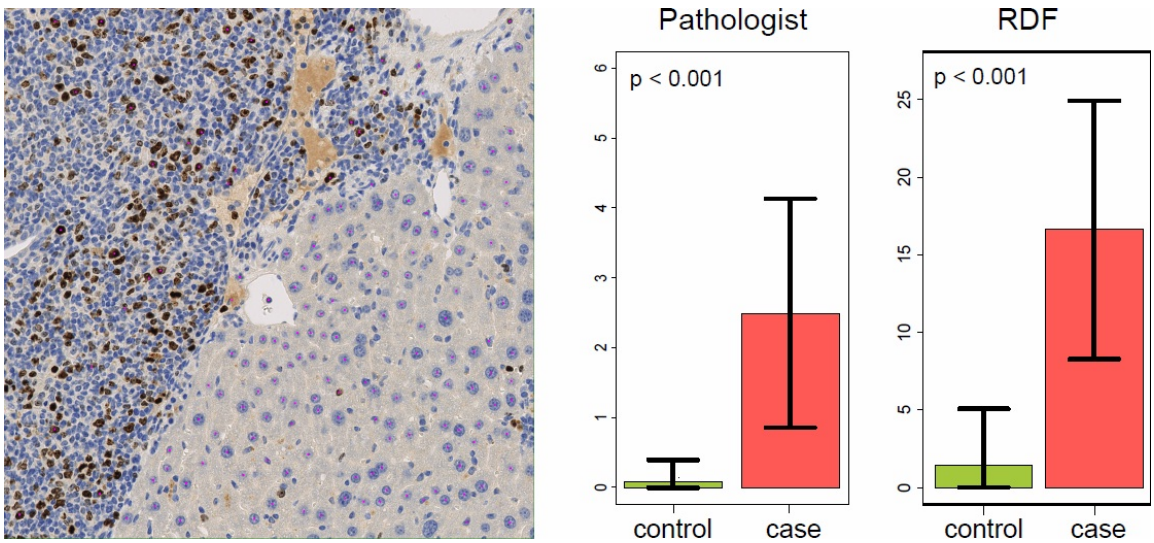


Figure 3.31

Left: Murine liver tissue with parenchymal nuclei of interest in the bottom right part of the image. The goal is to estimate the number of stained nuclei of this specific type. Detections are shown as pink dots. Most of the inflammatory cells in the top left part are correctly not detected.
Right: Comparison of MIB-1 staining in 88 images of mouse liver tissue between case and control groups. Although the algorithm is as well capable as the pathologist in differentiating between case and control ($p < 0.001$) the absolute estimates of stained nuclei are much higher.

CHAPTER 4

Statistics: Survival Analysis and Machine Learning in Medical Statistics

Contents

4.1 Overview

The main thrust of research in computational pathology is to build completely probabilistic models of the complete processing pipelines for histological and medical data. In medical research this nearly always also includes time to event data, where the event is either overall survival, specific survival, event free survival or recurrence free survival of patients. Statistics and machine learning within this scope is defined as Survival Analysis.

4.2 Survival Analysis

4.2.1 Censoring and Descriptive Statistics

Acknowledgments
I want to thank my co-author Sudhir Raman and especially Volker Roth for the introduction into Bayesian survival regression. Edgar Dahl provided the data for the experiments in this section.

Most difficulties in survival statistics arise from the fact, that nearly all clinical datasets contain patients with censored survival times. The most common form of censoring is right censored data which means that the death of the patient is not observed during the runtime of the study or that the patient withdrew from the study, e.g. because he moved to another location.

The nonparametric Kaplan-Meier estimator (Kaplan and Meier, 1958) is frequently used to estimate the survival function from right censored data. This procedure requires first to order the survival times from the smallest to the largest such that $t_1 \leq t_2 \leq t_3 \leq \ldots \leq t_n$, where t_j is the jth largest unique survival time. The Kaplan-Meier estimate of the survival function is then obtained as

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{r_j}\right) \tag{4.1}$$

where r_j is the number of individuals at risk just before t_j, and d_j is the number of individuals who die at time t_j.

To measure the goodness of separation between two or more groups, the log-rank test (Mantel-Haenszel test) (Mantel and Haenszel, 1959) is employed to assesses the null hypothesis that there is no difference in the survival experience of the individuals in the different groups. The test statistic of the log-rank test (LRT) is χ^2 distributed:

$$\hat{\chi^2} = \frac{\left(\sum_{i=1}^{m} \left(d_{1i} - \hat{e}_{1i}\right)\right)^2}{\sum_{i=1}^{m} \hat{v}_{1i}} \tag{4.2}$$

where d_{1i} is the number of deaths in the first group at t_i and $e_{1i} = r_{1j}\frac{d_i}{r_i}$ where d_i is the total number of deaths at time $t_{(i)}$, r_j is the total number of individuals at risk at this time, and r_{1i} the number of individuals at risk in the first group. Figure 5.2 depicts Kaplan-Meier plots for two subgroups each and the LRT p-values. The associated data is described in detail in Section 5.

4.2.2 Modeling of Time-to-Event Data

Survival Analysis as a branch of statistics is not restricted to medicine but analyses time to failure or event data and is also applicable to biology, engineering, economics etc. Particularly in the context of medical statistics, it is a powerful tool for understanding the effect of patient features on survival patterns within specific groups (Klein and Moeschberger, 1997). A parametric approach to such an analysis involves the estimation of parameters of a probability density function which models time.
In general the distribution of a random variable T (representing time) is defined over the interval $[0, \infty)$. Furthermore, a standard survival function is specified based on the cumulative distribution over T as follows:

$$S(t) = 1 - p(T \leq t_0) = 1 - \int_0^{t_0} p(t)dt, \tag{4.3}$$

which models the probability of an individual surviving up to time t_0. The hazard function $h(t)$, the instantaneous rate of failure at time t, is defined by:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{p(T = t)}{S(t)}. \tag{4.4}$$

The model is further extended by considering the effect of covariates X on time via a regression component. In medical statistics the most popular method for modeling such effects is Cox's proportionality hazards model (Cox, 1972):

$$h(t|\boldsymbol{x}) = h_0(t) \exp(\boldsymbol{x}^T \boldsymbol{\beta}), \tag{4.5}$$

where $h_0(t)$ is the baseline hazard function, which is the chance of instant death given survival uptil time t, $\boldsymbol{x}$ is the vector of covariates and $\boldsymbol{\beta}$ are the regression coefficients.

4.2.3 A Bayesian View of Survival Regression

Bayesian methods are gaining more and more popularity in machine learning in general and in medical statistics in special. A big advantage in survival analysis is the possibility to investigate the posterior distribution of a model. Especially in regularized survival regression models (Roth et al., 2008b) it is possible to get a posterior distribution also on zero coefficients, i.e. for biomarkers which hence were not included in the model.
A common choice of distribution for modeling time is the Weibull distribution which is flexible in terms of being able to model a variety of survival functions and hazard rates. Apart from flexibility, it is also the only distribution which captures both the accelerated time model and the proportionality hazards model (Joseph G.Ibrahim, 2001). The Weibull distribution is defined as follows:

$$p(t|\alpha_w, \lambda_w) = \alpha_w \frac{1}{\lambda_w} t^{\alpha_w - 1} \exp\left(-\frac{1}{\lambda_w} t^{\alpha_w}\right), \tag{4.6}$$

where α_w and λ_w are the shape and scale parameters, respectively. Based on the above definition and assuming right-censored data (Klein and Moeschberger, 1997), the likelihood assumes the form

$$p(\{t_i\}_{i=0}^{N}|\alpha_w, \lambda_w) = \prod_{i=1}^{N} \left(\frac{\alpha_w}{\lambda_w} t_i^{\alpha_w - 1}\right)^{\delta_i} \exp\left(-\frac{1}{\lambda_w} t_i^{\alpha_w}\right), \tag{4.7}$$

where $\delta_i = 0$ when the i^{th} observation is censored and 1 otherwise. Further, to model the effect of covariates $\boldsymbol{x}$ on the distribution over time, Cox's proportional hazards model can be applied. Under this model, the covariates are assumed to have a multiplicative effect on the hazard function.

4.2.4 Higher Order Interactions

A reoccurring question in biomedical research projects and especially in TMA analysis studies interactions of markers and their influence on the target. Two modern approaches within the scope of computational pathology try to solve this question from a frequentist (Dahinden et al., 2010) and a Bayesian (Roth et al., 2008b) point of view.
The most frequent approach for modeling higher-order interactions (like pairs or triplets of features etc.) instead of modeling just the main effects (individual features) are polynomial expansions of features. For example the vector $\boldsymbol{x} = \{x_1, x_2, x_3\}$ can be expanded up to order 2 as follows:

$$\boldsymbol{x}' = \{x_1, x_2, x_3, x_1 : x_2, x_1 : x_3, x_2 : x_3, x_1 : x_2 : x_3\}$$

, where $x_1 : x_2$ denotes an interaction between covariates x_1 and x_2. Additional flexibility is built into this model by including a random effect in η in the following manner:

$$\eta = \boldsymbol{x}^t \boldsymbol{\beta} + \epsilon, \qquad \text{where } \epsilon \sim N(0, \sigma^2). \tag{4.8}$$

To include the covariate effect the likelihood of Equation 4.7 is modified as follows:

$$\begin{aligned} p(\{t_i\}_{i=0}^{N}|\boldsymbol{x}_i, \alpha_w, \lambda_w) \quad = \quad & \prod_{i=1}^{N} \left[\frac{\alpha_w}{\lambda_w} t_i^{\alpha_w - 1} \exp(\eta_i)\right]^{\delta_i} \cdot \\ & \exp\left(-\frac{1}{\lambda_w} t_i^{\alpha_w} \exp(\eta_i)\right) \end{aligned}$$

These kind of models can be seen as enhancement of generalized linear models (McCullaghand and Nelder, 1983) and are called random-intercept models. For a full Bayesian treatment of the model, suitable priors have to be defined for the parameters of the model, namely α_w, λ_w, σ and $\boldsymbol{\beta}$. Useful priors for this model are described in (Roth et al., 2008b).

4.2.5 Mixtures of Survival Experts

Frequently, sub-groups of patients specified by characteristic survival times have to be identified together with the effects of covariates within each sub-group. Such information might hint at the disease mechanisms. Statistically this grouping is represented by a mixture model or specifically by a mixture of survival experts.

To this end, (Rosen and Tanner, 1999) define a *finite* mixture-of-experts model by maximizing the partial likelihood for the regression coefficients and by using some heuristics to resolve the number of experts in the model. More recently (Ando et al., 2004) use a maximum likelihood approach to infer the parameters of the model and the Akaike information criterion (AIC) to determine the number of mixture components.

A Bayesian version of the mixture model (Kottas, 2006) analyzes the model with respect to time but does not capture the effect of covariates. On the other hand the work by (Ibrahim et al., 1996) performs variable selection based on the covariates but ignores the clustering aspect of the modeling. Similarly, (Paserman, 2004) defines an infinite mixture model but does not include a mixture of experts, hence implicitly assuming that all the covariates are generated by the same distribution with a common shape parameter for the Weibull distribution. (Roth et al., 2008b) unify the various important elements of this analysis into a Bayesian mixture-of-experts (MOE) framework to model survival time, while capturing the effect of covariates and also dealing with an unknown number of mixing components. To infer the number of experts a Dirichlet process prior on the mixing proportions is applied, which solves the issue of determining the number of mixture components beforehand (Rasmussen and Ghahramani, 2001). Due to the lack of fixed-length sufficient statistics, the Weibull distribution is not part of the exponential family of distributions and hence the regression component, introduced via the proportionality hazards model, is non-standard. Furthermore, the framework of (Roth et al., 2008b) includes sparsity constraints to the regression coefficients in order to determine the key explanatory factors (biomarkers) for each mixture component. Sparseness is achieved by utilizing a Bayesian version of the Group-Lasso (Raman et al., 2009; Raman and Roth, 2009) which is a sparse constraint for grouped coefficients (Yuan and Lin, 2006).

4.3 Wishart-Dirichlet Partitioning for Quality Control

4.3.1 Motivation

Acknowledgments
Special thanks to Julia Vogt and Sandhya Prabhakaran, who are co-authors on the Wishard-Dirichlet manuscript and to Volker Roth who proficiently lead the research efforts described in this section.

A main challenge in computational pathology is the automated analysis of tissue microarrays (TMA). TMAs consist of tissue samples from hundreds of patients which can be stained with various antibodies for protein expression analysis. An automated analysis pipeline consists of three major steps: (i) cell nuclei detection, (ii) nuclei classification into malignant and benign, and (iii) staining estimation. The resulting estimation per patient can then be used to correlate marker expression with the survival times or other clinical variables.

The most crucial step in the analysis pipeline is the classification of nuclei into malignant and benign, because the subsequent staining estimation has to be performed only on the subgroup of cancerous nuclei. Proliferation markers like MIB-1 (Ki-67) stain cell nuclei shortly before and after mitosis. The percentage of stained cancerous nuclei is one of the best prognostic factors for the survival of cancer patients (Tannapfel et al., 1996), due to the fact that it directly relates to aggressiveness of the disease. As a consequence the differentiation between malignant and benign nuclei directly affects the final survival model for cancer patients. Stained benign nuclei, which were falsely classified, can considerably worsen the survival prediction in a domain, where small differences in the low percentage regime have a large impact on the progression of the disease.

Previous approaches (Fuchs et al., 2008b) to automatic TMA analysis demonstrated that reasonable nuclei detection and staining estimation is possible and approaches the performance of trained pathologists. The main drawback of these models is the requirement for (almost) perfectly processed TMA spots. The predominant problem in clinical practice is the high variability between and within single spots, respectively patients. Noise and variations are not only imposed by biology but also by technical preprocessing which comprises error prone steps like micro-cutting, punching of TMA spots and staining, which involves applying antibodies and microwaving of the tissue. The final step comprises scanning of the microscope slides and tiling of the TMA into single spots. All these steps lead to biological, technical and digital artifacts in the images resulting in distorted, blurred or obfuscated regions. Thus, trained pathologists do not take into account the whole spot when manually analyzing TMAs, but restrict themselves to regions of high quality only. This preference could also be observed during extensive labeling experiments for generating a "gold standard". Forcing pathologist to classify randomly drawn nuclei led not only to high inter-pathologist variability but also to high intra-pathologist variability ($\sim$ 25%). To this end, the main goal in this application scenario is to create an algorithm which is robust to tissue variations by mimicking the work-flow of trained domain experts.

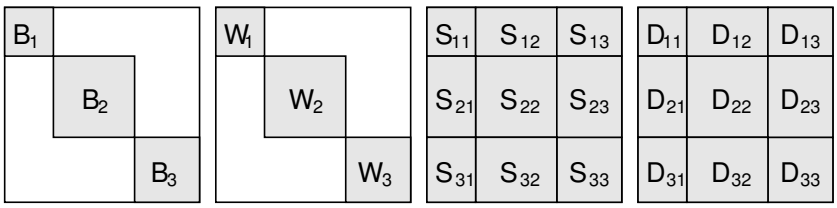


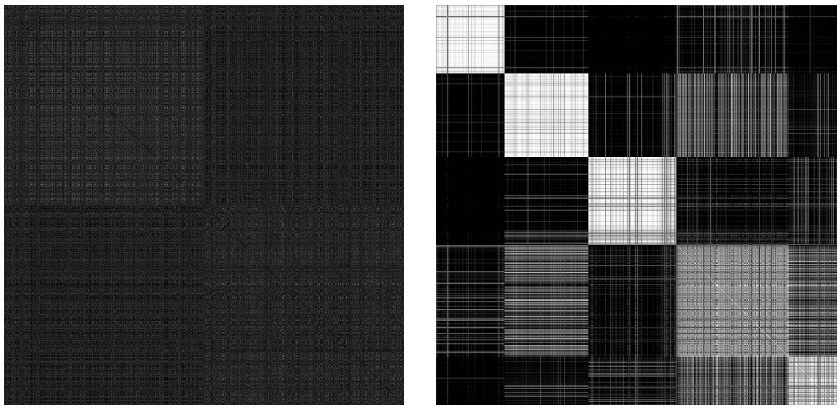
Figure 4.1
Example of the block structure of B and W (left) and the definition of sub-matrices in S and D (right) for $k_B = 3$.

4.3.2 Wishart-Dirichlet Models for Partitioning Matrices

The Bayesian clustering approach presented in this work aims at identifying subsets (or "clusters") of objects represented as columns/rows in a dissimilarity matrix. The underlying idea is that objects grouped together in such a cluster can be reasonably well described as a homogeneous sub-population. Our focus on dissimilarity matrices implies that we do not have access to a vectorial representation of the objects. Such underlying vectorial representation may or may not exist, depending on whether the dissimilarity matrix can be embedded (without distortion) in a vector space. One way of dealing with such clustering problems would be to explicitly construct an Euclidean embedding (or possibly a distorted embedding), and to apply some more traditional clustering methods in the resulting Euclidean space. We argue, however, that even under the assumption that there exists an Euclidean embedding it is better *not* to explicitly embed the data, since any such choice might induce an unnecessary bias in the further clustering process. Technically speaking, such embeddings break the symmetry induced by the translation- and rotation-invariance which reflects the information loss incurred when moving from vectors to pairwise dissimilarities. We propose a clustering model which works directly on dissimilarity matrices. It is invariant against label- and object permutations and against scale transformations. Since the model is fully probabilistic in nature, as output we are not given a single clustering solution (if desired, a "representative" solution can be computed, though), but samples from a probability distribution over partitions. Further, the use of a Dirichlet process prior unburdens the user from explicitly fixing the number of clusters. On the algorithmic side we present a highly efficient sampling algorithm which avoids costly matrix operations by carefully exploiting the structure of the clustering problem. Invariance against label permutations is a common cause of the so-called "label switching" problem in mixture models. By formulating the model as a partition process this switching problem is circumvented.

The *Dirichlet cluster process for Gaussian mixtures* (McCullagh and Yang, 2008) is employed, which basically is a Gaussian mixture model for vectorial data which is invariant to label- and object permutations, but still avoids the "label switching" problem. This model is generalized to relational data by additionally enforcing translation- and scale invariance. We call this new model the *Translation-invariant Wishart-Dirichlet (WD) cluster process*.

We extend the Gauss-Dirichlet cluster process to a sequence of inner-product



(a) (b)

Figure 4.2
(a) Proximity matrix of the nuclei in the tree ensemble. The matrix is generated by putting all training samples and OOB samples down each tree of the ensemble. If two samples end up in the same terminal node, their proximity is increased by one. Finally, the similarities are normalized by the number of trees. The resulting matrix shows some negative eigenvalues, which are dealt with by using the shift-embedding trick, followed by PCA-denoising.
(b) Matrix of estimated co-membership probabilities between cell nuclei, computed by averaging the partitions obtained from 10000 Gibbs sweeps. White entries mean that two nuclei are in the same cluster with (estimated) probability one. The two clusters which are highly indicative for the malignant/benign-label refer to the second and third block (from top left).

and distance matrices. Assume that the random matrix $X_{n\times d}$ follows the zero-mean Gaussian distribution specified in (ref coveq), with $\Sigma_0 = \alpha I_d, \Sigma_1 = \beta I_d$. Then, conditioned on the partition B, the inner product matrix $S = XX^t/d$ follows a (possibly singular) Wishart distribution in d degrees of freedom, $S \sim \mathcal{W}_d(\Sigma_B)$, (Srivastava, 2003). If we directly observe the dot products S, it suffices to consider the conditional probability of partitions, $P_n(B|S)$, which has the same functional form for ordinary and singular Wishart distributions:

$$\begin{aligned} P_n(B|S,\alpha,\beta,\xi,k) &\propto \mathcal{W}_d(S|\Sigma_B)\cdot P_n(B|\xi,k) \\ &\propto |\Sigma_B|^{-\frac{d}{2}} \exp\left(-\tfrac{d}{2}\mathrm{tr}(\Sigma_B^{-1}S)\right)\cdot P_n(B|\xi,k), \end{aligned} \tag{4.9}$$

For the following derivation it is suitable to re-parametrize the model in terms of $(\alpha, \theta := \beta/\alpha)$ instead of (α,β), and in terms of $W := \Sigma_B^{-1}$. Due to the block structure in B, $P_n(B|S)$ factorizes over the blocks $b \in B$:

$$\begin{aligned} &P_n(B|S,\alpha,\theta,\xi,k) \propto P_n(B|\xi,k) \\ &\cdot \left[\textstyle\prod_{b\in B}|W_b|^{\frac{d}{2}}\right] \exp\left(-\textstyle\sum_{b\in B}\frac{d}{2}\mathrm{tr}(W_b S_{bb})\right), \end{aligned} \tag{4.10}$$

where W_b, S_{bb} denote the submatrices corresponding to the b-th diagonal block in B or W, see Figure 4.1.
The above factorization property can be exploited to derive an efficient inference algorithm for this model. The key observation is that the inverse matrix $W_b = \Sigma_b^{-1}$ can be analytically computed as

$$\begin{aligned} W_b &= (\alpha I_b + \beta \mathbf{1}_b\mathbf{1}_b^t)^{-1} = \left[\alpha(I_b + \theta\mathbf{1}_b\mathbf{1}_b^t)\right]^{-1} \\ &= \tfrac{1}{\alpha}\left[I_b - \tfrac{\theta}{1+n_b\theta}\mathbf{1}_b\mathbf{1}_b^t\right]. \end{aligned} \tag{4.11}$$

Thus, the contribution of block b to the trace is

$$\mathrm{tr}(W_b S_{bb}) = \tfrac{1}{\alpha}\left[\mathrm{tr}(S_{bb}) - \tfrac{\theta}{1+n_b\theta}\bar{S}_{bb}\right], \tag{4.12}$$

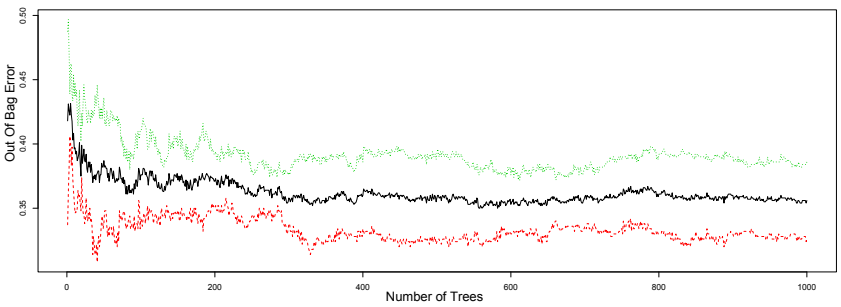


Figure 4.3
Out of bag (OOB) error of the random forest classifier for the whole dataset. The error converges after approximately 300 trees to an average classification error (black) of approximately 36%. *red*: OOB error of the malignant class. *green*: OOB error of the benign class.

where $\bar{S}_{bb} = \mathbf{1}_b^t S_{bb} \mathbf{1}_b$ denotes the sum of the b-th diagonal block of S. A similar trick can be used for the determinant which is the product of the eigenvalues: the k_B smallest eigenvalues of W are given by $\lambda_b = \alpha^{-1}(1 + \theta n_b)^{-1}$. The remaining $n - k_B$ eigenvalues are equal to α^{-1}. Thus, the determinant reads

$$|W| = \prod_{b \in B} \lambda_b = \alpha^{-n} \prod_{b \in B} (1 + \theta n_b)^{-1}. \tag{4.13}$$

4.3.3 Quality Control in Computational Pathology

The dataset consists of 500 cancerous nuclei and 500 normal nuclei sampled from TMA spots of 9 clear cell renal cell carcinoma patients (ccRCC). The spots were exhaustively labeled by a trained pathologist to generate a gold standard. To differentiate between malignant and benign cell nuclei a Random Forest (RF) classifier (Breiman, 2001) is trained. Each sample consists of a patch of size 65×65 pixels, centered at the nucleus. Local Binary Patterns (LBP) (Ahonen et al., 2004) are extracted from the gray scale images to form a feature vector of size 256 for each sample. LBPs have the advantage of illumination invariance, i.e. they are invariant with respect to monotonic gray-scale changes. A random forest classifier consists of a collection of tree-structured classifiers $\{h(\boldsymbol{x}, \Theta_k), k = 1, \ldots\}$ where $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\boldsymbol{x}$. One beneficial property of RFs is the internal out of bag (OOB) error which provides an unbiased estimate of the generalization error and which is reported in the following evaluation.

Learning a RF on the whole data set leads to an OOB error of 36% (Figure 4.3). Hence every third subsequent staining estimation is performed on a falsely classified nucleus. To enhance these results we follow the analysis strategy of pathologists by excluding detection regions with poor quality. To this end we use our matrix partitioning model to find subgroups of cell nuclei. A 1000×1000 similarity matrix is generated by measuring the proximity of nuclei in the tree ensemble (Breiman, 2001). All training samples and OOB samples are put down each tree of the ensemble. If two samples end up in the same terminal node, their proximity is increased by one. Finally, the similarities are normal-

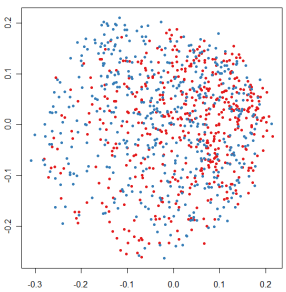


Figure 4.5
Multidimensional scaling of the random forest proximity matrix. Visually there are no clusters decernable of cancerous (red) or normal (blue) nuclei.

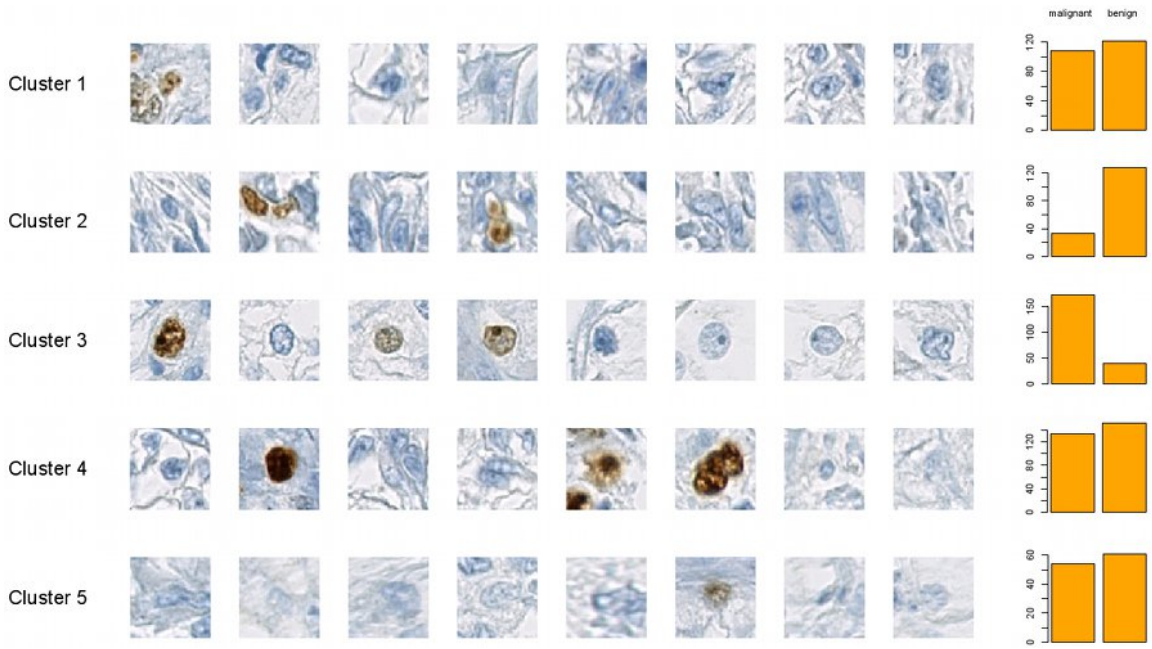


Figure 4.4

Cluster of cell nuclei from renal cell carcinoma patients. **Left:** Randomly drawn nuclei from the five cluster revealed by W-D-clustering.
Right: Within-cluster distribution of cancerous and benign nuclei.
For most cluster the semantic interpretation of pathologists is in agreement with the class distribution: e.g. cluster 3 consists mainly of large nuclei which are clearly distinguishable from background. This kind of morphology is articulated in cancerous cells which are no longer embedded in cohesive tissue. Cluster 2 on the other hand comprises small and elongated nuclei on cluttered background. This is characteristic for benign nuclei in healthy tissue and for endothelial cells in connective tissue. In contrast the patches in cluster 5 are either blurred, distorted or show no clear structure which is mainly due to technical processing flaws which lead to image regions of poor quality. In addition cluster 1 and 4 show no interpretable pattern and vary largely in tissue morphology and image quality. These three cluster also exhibit a uniform distribution of malignant and benign cells.

ized by the number of trees. The resulting matrix shows some negative eigenvalues, which are dealt with by using the shift-embedding trick (Vogt et al., 2010), followed by PCA-denoising.

Our clustering model reveals five stable clusters shown in Figure 4.2 and Figure 4.4. Most of the clusters can be interpreted semantically. For instance, cluster 3 contains large nuclei which are clearly distinguishable from background. Such morphology is articulated in cancerous nuclei which are no longer embedded in cohesive tissue. Cluster 2, on the other hand, comprises small and elongated nuclei on cluttered background. This is characteristic for benign nuclei in healthy tissue and for endothelial cells in connective tissue. These observation are consistent with the observed distribution of labels (Figure 4.4). In contrast, the patches in cluster 5 are either blurred, distorted or show no clear structure. This lack of sematical meaning is mainly caused by technical process-

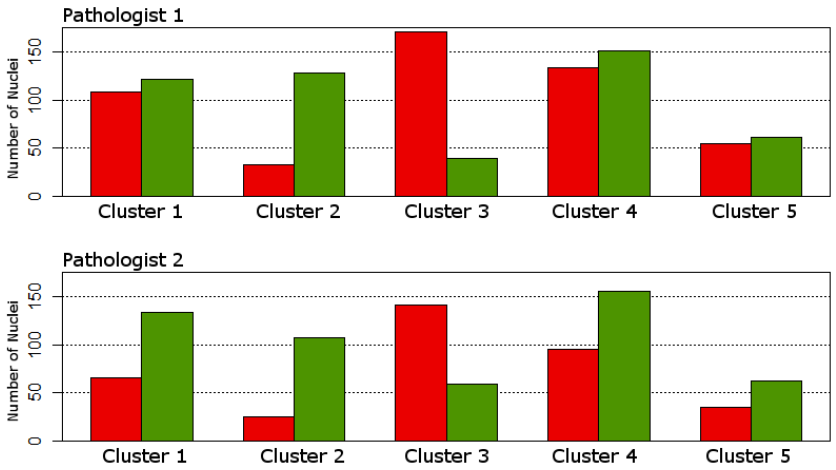


Figure 4.6

Label distribution of each cluster per pathologist. Both experts agree that cluster 2 consists predominantly of normal nuclei, while cluster 3 comprises mostly cancerous nuclei.

ing flaws which result in regions, which cannot be classified reliably, although the pathologists detected remnants of nuclei. Cluster 1 and 4 vary largely in tissue morphology and image quality. Consequently, these three clusters show an almost uniform distribution of malignant and benign cells. A computational TMA analysis tool should reject such nuclei, in the same manner as a domain expert would go about it to avoid contamination of the whole patient sample. Proceeding at these lines, a RF classifier is trained on the subset of nuclei from cluster 2 and 3, resulting in an OOB error of 19.4%. This significant reduction nicely demonstrates the importance of quality assessment preceeding classification. We are convinced that this data curation approach is the key to solving one of the most severe problems in the design of computational TMA analysis tools.

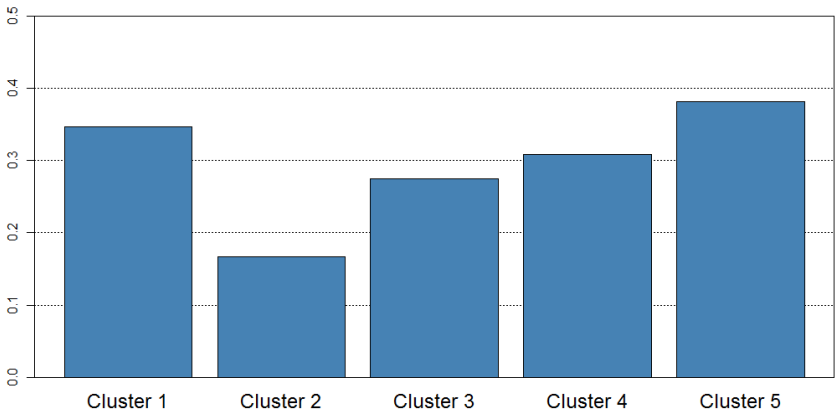


Figure 4.7

Inter observer misclassification error of two pathologists for each of the five cluster determined by Wishart-Dirichlet partitioning.

4.4 Linear Modeling for Detection of Urothelial Bladder Cancer Cells

4.4.1 Overview

Acknowledgments
I want to thank Peter Wild for the long and fruitful cooperation described in this section. In addition I like to thank all our co-authors: Robert Stoehr, Dieter Zimmermann, Simona Frigerio, Barbara Padberg, Inbal Steiner, Ellen C. Zwarthoff, Maximilian Burger, Stefan Denzinger, Ferdinand Hofstaedter, Glen Kristiansen, Thomas Hermanns, Hans-Helge Seifert, Maurizio Provenzano, Tullio Sulser, Volker Roth, Joachim M. Buhmann, Holger Moch, and Arndt Hartmann.

Purpose: We evaluate molecular and immunohistochemical markers and develop statistical model for molecular grading of urothelial bladder cancer. Finally we test these markers in voided urine samples.

Experimental Design: 255 consecutive biopsies from primary bladder cancer patients were evaluated on a tissue microarray. The following clinical parameters were collected: *gender*, *age*, *adjacent carcinoma in situ*, and *multifocality*. UroVysion fluorescence in situ hybridization (FISH) was performed. Expression of cytokeratin 20, MIB1, and TP53 was analyzed by immunohistochemistry. Fibroblast growth factor receptor 3 (FGFR3) status was studied by SNaPshot mutation detection. Results were correlated with clinical outcome by Cox regression analysis. To assess the predictive power of different predictor subsets to detect high grade and tumor invasion, logistic regression models were learned. Additionally, voided urine samples of 119 patients were investigated. After cytologic examination, urine samples were matched with their biopsies and analyzed for loss of heterozygosity (LOH), FGFR3 mutation, polysomy, and p16 deletion using Uro- Vysion FISH. Receiver operator characteristic curves for various predictor subsets were plotted.

Results: In biopsies, high grade and solid growth pattern were independent prognostic factors for overall survival. A model consisting of UroVysion FISH and FGFR3 status (FISH + FGFR3) predicted high grade significantly better compared with a recently proposed molecular grade (MIB1 + FGFR3). In voided urine, the combination of cytology with LOH analysis (CYTO + LOH) reached the highest diagnostic accuracy for the detection of bladder cancer cells and performed better than cytology alone (sensitivity of 88.2% and specificity of 97.1%).

Conclusions: The combination of cytology with LOH analysis could reduce unpleasant cystoscopies for bladder cancer patients.

4.4.2 Introduction

At the time of first diagnosis, $\approx$ 70% of bladder tumors are noninvasive papillary low-grade tumors (pT_a). Despite the fact that the majority of urothelial bladder tumors are clinically benign, regular cystoscopic follow-up at intervals is performed in all patients with non-muscle-invasive bladder cancer after complete transurethral resection to detect recurrence and progression.
Mutations of the tumor suppressor genes TP53 and RB1 are common and have

predictive value in clinical studies of invasive bladder cancer (Cordon-Cardo et al., 1997; Cote et al., 1998; Masters et al., 2003). Although TP53 alterations have been suggested as prognostic marker in pTa tumors (Sarkis et al., 1993), the prognostic value of both TP53 and RB1 is restricted to invasive tumors. In non-muscle-invasive bladder cancer, homogeneous expression of cytokeratin 20 (CK20; (Harnden et al., 1999)), lack of fibroblast growth factor receptor 3 (FGFR3) mutations (van Rhijn et al., 2001, 2003b), and high nuclear Ki-67 labeling index (van Rhijn et al., 2003b) show promise in predicting recurrence. Mutations in the FGFR3 gene are very frequent in pTa bladder tumors ($\approx$ 70%; refs. (van Rhijn et al., 2003b; Billerey et al., 2001; van Rhijn et al., 2002; van Oers et al., 2007). Hernandez et al. have determined the frequency and the prognostic value of FGFR3 mutations in patients with primary non-muscle-invasive bladder cancer in a large prospective study ($n = 772$; (Hernandez et al., 2006)). In analogy to the data presented by van Rhijn et al. (2001), their findings strongly support the notion that FGFR3 mutations characterize a subgroup of bladder cancers with good prognosis. However, there is no prospectively evaluated set of molecular markers with sufficient predictive power to select patients for a differential therapeutic approach.

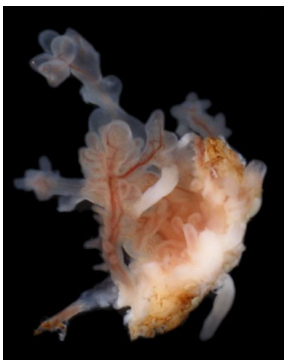


Figure 4.8
Bladder resectate (TUR-B). A low-grade papillary non-invasive urothelial carcinoma. Macroscopic photography taken under water.

Conventionalurine cytology is used as a complement to cystoscopy for the detection of new bladder carcinomas and recurrences. However, application of cystoscopy every 3 to 6 months is very unpleasant for the patient. In the past, the low sensitivity of urine cytology reported in diagnosing low-grade papillary tumors has limited its use and prevented cytology from replacing cystoscopy (Brown, 2000). Fluorescence in situ hybridization (FISH)-based detection systems are currently used in conjunction with cystoscopy for the examination of bladder washings and voided urine samples. The use of UroVysion Multicolor FISH (Vysis), a FISH assay for detection of bladder cancer, based on the use of voided urine samples, has been evaluated in multiple studies (Placer et al., 2002; Halling et al., 2000; Bubendorf et al., 2001). The UroVysion FISH test was the first and only test approved by the U.S. Food and Drug Administration, which uses DNA probes to identify aneuploidy for chromosomes 3, 7, and 17 and loss of the 9p21 locus in urine specimens from subjects with urothelial bladder cancer. van Rhijn et al. (van Rhijn et al., 2005) have systematically reviewed urine markers for bladder cancer surveillance. Urinary cytology afforded a median sensitivity of 35% (range, $13 - 75\%$), whereas the median sensitivity for FISH and microsatellite analysis was 79% (range, 70-86) and 82% (range, $75 - 92\%$). Microsatellite analysis of loss of heterozygosity (LOH) and FISH were among the most promising markers for surveillance (van Rhijn et al., 2005).We and others have previously shown that the detection of bladder carcinoma cells can be improved by standardized microsatellite analysis (Frigerio et al., 2007; van der Aa et al., 2009). Over 93% of patients with recurrent bladder cancer disease were identified by a combination of microsatellite (LOH) analyses and cytology of their voided urine samples.

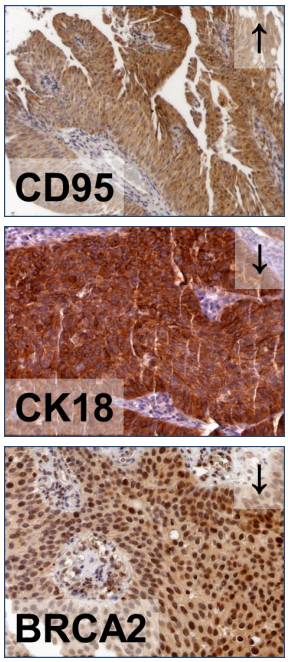


Figure 4.9
IHC reactions from top to bottom: CD95, CK18, BRCA2

High histologic grade as a marker for chromosomal instability is the clinically most important marker for increased risk of progression to muscle-invasive disease. However, histologic grading has a high inter-observer variability with varying prognostic implications (Tosoni et al., 2000). The aim of the current study was to systematically evaluate a set of molecular and immunohistochemical markers to (a) develop a reliable molecular grading of urothelial bladder cancer and (b) evaluate the usefulness of these markers to detect bladder cancer cells in voided urine.

4.4.3 Materials and Methods

Bladder Cancer Tissue Microarray. A tissue microarray was constructed as described previously (van Oers et al., 2007) from 255 consecutive, formalin-fixed, paraffin-embedded, primary urothelial bladder cancer tissues (Institute of Pathology, University of Regensburg). Clinical data were obtained from the Central Tumor Registry Regensburg and by telephone interviews (M.B. and S.D.) in case of missing data. The tissue microarray contained two tissue cores of each tumor specimen. The Institutional Review Board of the University of Regensburg approved analysis of tissues from human subjects. H&E-stained slides of all tumors were evaluated by a single surgical pathologist (A.H.). Tumor stage and grade were assigned according to International Union Against Cancer and WHO criteria (Grignon et al., 2004). Invasive bladder carcinomas were graded as either low grade (G2) or high grade (G3). Growth pattern was determined for all invasive tumors ($\geq$pT1). Papillary growth was defined by the presence of a papillary tumor component ($\geq$ 20%) with a histological grade identical to the invasive tumor. All other tumors were considered to have a solid growth pattern. Clinicopathologic data are summarized in Table 5.1. Retrospective clinical follow-up data were available regarding the endpoints recurrence-free and overall survival. The median follow-up period was 77 months (range, $0-166$ months). Thirty-eight of 215 (15%) analyzable patients died during follow-up. The median follow-up for censored patients was 84 months. Recurrences were defined as cystoscopically visible tumors (using photodynamic diagnosis with 5-aminolevulinic acid) with histological verification.

Urine Samples. As described previously (Frigerio et al., 2007), voided urine samples of 119 patients scheduled for transurethral resection were prospectively collected over a period of 20 months and matched with their corresponding biopsies. Of these, 81 biopsies proved to be neoplastic on histologic examination. Characteristics are given in Table 5.2. Additional 38 urine samples were collected from patients whose biopsies turned out to be histologically normal or displayed inflamed urothelium without presence of neoplastic cells. Half of these tumor-negative samples were derived from patients without previous history of bladder cancer. All urine samples ($15mL$) were directly collected at the Department of Urology, University Hospital Zürich, shortly before

transurethral resection. The urine samples were centrifuged at $1,300 \times g$ for 10 min and sediments were immediately processed for cytologic examination and FISH analysis. This study has been approved by the local ethics committee (StV-14/2003; July 30, 2003) and informed consent was obtained from all patients.

Cytologic Examination of Urine Sediments. Urine sediments were resuspended in PBS and one to three cytospin slides were prepared from an aliquot. The slides were fixed with Cytostat 400 solution (Simat) and stained with standard Papanicolaou. A cell density between 25 and 50 cells per visual field using a $\times 20$ objective was regarded as sufficient for analysis. Slides were reviewed in a blinded fashion by a cytopathologist (B.P.) and classified according to the following morphologic criteria: cells with severe atypia diagnostic of neoplasia (P), moderately atypical cells suspicious of neoplasia (S), cells with reactive alterations (NR), and cells with normal morphology (N). Immunohistochemistry. Immunohistochemical studies were done as described previously (van Oers et al., 2007) One surgical pathologist (A.H.) performed a blinded evaluation of the slides. Positive TP53 immunoreactivity was defined as strong nuclear staining in $>$ 10% of the tumor cells. The percentage of MIB1-positive cells of each specimen was determined as described previously (Nocito et al., 2001). High MIB1 labeling index was defined if $>$ 25% of the tumor cells were positive (van Rhijn et al., 2003b). CK20 staining was defined as normal (superficial staining pattern) or abnormal (negative or $>$ 10% stained) according to Harnden et al. (1999).

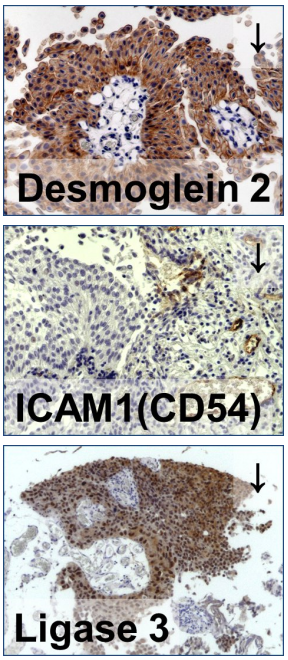


Figure 4.10
IHC reactions from top to bottom: Desmoglein 2, ICAM1(CD54), Ligase 3

DNA Isolation. Genomic DNA of paraffin-embedded tumors on the tissue microarray was isolated from $1.5mm$ punch biopsies of the paraffin blocks (one tissue core per case). Tumor areas were marked by a surgical pathologist (A.H.) to ensure a tumor cell content of at least 80%. DNA isolation was done using the Magna Pure DNA isolation kit (Roche) according to the manufacturer's instructions. DNA from urine samples was extracted and purified with a DNA Blood Mini-Kit (Qiagen) following instructions of the manufacturer. For the few samples containing only little DNA, at least $2ng$ DNA was applied.

FGFR3 Mutation Analysis. FGFR3 mutation analysis was done using the SNaPshot method (van Oers et al., 2005). All mutations were verified by a second and independent SNaPshot analysis.

FISH Analysis of Paraffin Specimens. Multicolor FISH was performed using the UroVysion probe set (Abbott Laboratories) according to the manufacturer's instructions to assess aberrations of chromosomes 3, 7, and 17 by centromeric probes and to detect relative deletions of p16 on locus 9p21 (Schwarz et al., 2008). For each case, 50 nuclei were selected for scoring according to morphologic criteria using 4',6- diamidino-2-phenylindole staining. Only nonoverlapping intact nuclei were scored. Clearly distinguishable nonurothelial cells were discarded. All hybridizations were evaluated by two investigators (R.S. and I.S.) with random quality control checks (A.H.). Each cell was simultaneously

analyzed for centromeric signals of chromosomes 3, 7, and 17 and the p16 locus on 9p21. A cell was considered aberrant if at least one of three centromeric signals was amplified ($>$ 2 signals per cell) or if 9p21 was deleted. Polyploid cells (4 signals of all the three probes) were regarded normal (euploid). A relative deletion of the p16 locus (CDKN2A) was recognized if the signal number of 9p21 was $>$ 1 unit lower than the mean value of the centromeric signals. Based on the occurrence of polysomy and deletions of 9p21 in non-tumor-associated bladder urothelium of patients with benign prostatic hyperplasia ($n = 10$), a cutoff was defined using three times the SD (Schwarz et al., 2008). Accordingly, a case was considered aberrant if $>$ 9 cells of 50 showed polysomy ($>$ 18% of the cells). A sample was considered carrying a deletion of p16 if more than 7 of 50 cells ($>$ 14% of the cells) showed a relative deletion of 9p21.

FISH Analysis of Voided Urine Samples. In each case, 25 selected cells were analyzed. The cell selection criteria included patchy and lighter nuclear 4',6-diamidino-2-phenylindole staining, nuclear enlargement, irregular nuclear contour, and presence in a small cluster. Overlapping cells were not analyzed. Samples were scored as FISH positive, if $\geq$ 4 cells showed at least 3 copies of any of the centromeric signals for chromosomes 3, 7, and 17 and if $\geq$ 12 cells displayed a homozygous loss of 9p21 (Frigerio et al., 2007). Due to technical reasons, only 2 of the 38 (5%) nonneoplastic urine samples were analyzable with FISH.

Statistical Analysis. Statistical analyses were completed using SPSS version 16.0 (SPSS) and R (R Development Core Team, 2009). Differences were considered significant if $P < 0.05$. All samples were considered independent.

Associations between measured parameters were obtained by applying X^2 and two-sided Fisher's exact tests. The Kaplan-Meier method was used to compare curves for the different variables with regard to recurrence-free and overall survival, with significance evaluated by twosided log-rank statistics. For the analysis of recurrence-free survival, patients were censored at the date when cystectomy was done or at the time of their last tumor-free clinical follow-up appointment. For survival analysis, patients were censored at the time of their last clinical follow-up appointment. Cox proportional hazard ratios were estimated to obtain risks of death and to find independent prognostic factors in a multivariate model. Limit for reverse selection procedures was $P = 0.1$.

To assess the predictive power of different predictor subsets, a logistic regression model was learned for each set. Cross-validation was used to validate the predictive power of the models. Therefore, 70% of the samples were drawn at random to form the training set on which a model was learned, which then was tested on the 30% out of bag samples. This procedure was repeated 100 times for each model to get estimates for the prediction error. Student's t test was employed to quantify differences between the error distributions of different models. Data from voided urine samples were described by plotting receiver operator characteristic (ROC) curves of the posterior probability for various predictor subsets. The best operation point was highlighted and the corresponding sen-

Variable	Categorization	Polysomy			Relative *p16* deletion		
		≤18%	>18%	*P**	≤14%	>14%	*P**
Clinicopathologic data							
Tumor stage	pT_a	62	75	**<0.001**	90	47	**<0.001**
	pT_1	5	39	17	27		
	pT_2	4	46	20	30		
	pT_3	0	2	0	2		
	pT_4	0	3	1	2		
Histologic grade	Low	64	76	**<0.001**	93	47	**<0.001**
	High	7	89	35	61		
Adjacent carcinoma *in situ*	No	69	139	**0.004**	116	92	0.228
	Yes	2	26	12	16		
Multifocality	Solitary	12	39	0.302	16	35	**<0.001**
	Multifocal	59	126	112	73		
Growth pattern	Papillary	69	124	**<0.001**	114	79	**0.003**
	Solid	2	40	14	28		
Molecular data							
FGFR3 gene	Wild-type	12	88	**<0.001**	46	54	**0.014**
	Mutation	44	49	60	33		
Immunohistochemistry							
MIB1 immunohistochemistry	≤25%	62	100	**<0.001**	94	68	0.056
	>25%	7	57	28	36		
TP53 immunohistochemistry	≤10%	64	103	**<0.001**	101	66	**0.002**
	>10%	4	59	23	40		
CK20 immunohistochemistry	Superficial staining pattern	22	25	**0.013**	35	12	**0.002**
	Negative or >10%	49	134		90	93	

Table 4.1

Polysomy and relative *p16* deletion in relation to clinicopathologic, molecular, and immunohistochemical markers. (Boldface representing P values < 0.05)

sitivity, specificity, and positive and negative predictive values (PV+, PV-) were reported.

4.4.4 Results

Tissue Microarray Study of Urinary Bladder Cancer.

Immunohistochemical and Molecular Markers. The prognostic effect of the UroVysion kit in concert with four previously described molecular markers (FGFR3, C K20, MIB1, and TP53) was investigated retrospectively. Investigation of UroVysion FISH in a series of 255 primary urothelial bladder cancers using tissue microarray technology was informative in 92.5% (236 of 255) of the cases. Cases of noninterpretable results were due to poor technical quality or lack of epithelial cell content. Polysomy of at least one chromosome was found in 69.9% (165 of 236) and a relative deletion of 9p21 in 45.8% (108 of 236) of urothelial neoplasms. Results of FGFR3 mutation analysis and MIB1, TP53, and CK20 immunohistochemistry have been published previously (van Oers et al., 2007) and are given in Tables 4.1 and 4.3.

Table 4.1 shows the association of UroVysion FISH results with clinicopathologic, immunohistochemical, and molecular parameters. Polysomy and relative p16 deletion was significantly associated with high tumor stage, high grade, and solid growth pattern. Almost all cases with adjacent carcinoma in situ showed polysomy in at least one chromosome ($P = 0.004$). These data confirm that polysomy and relative p16 deletions are associated with adverse

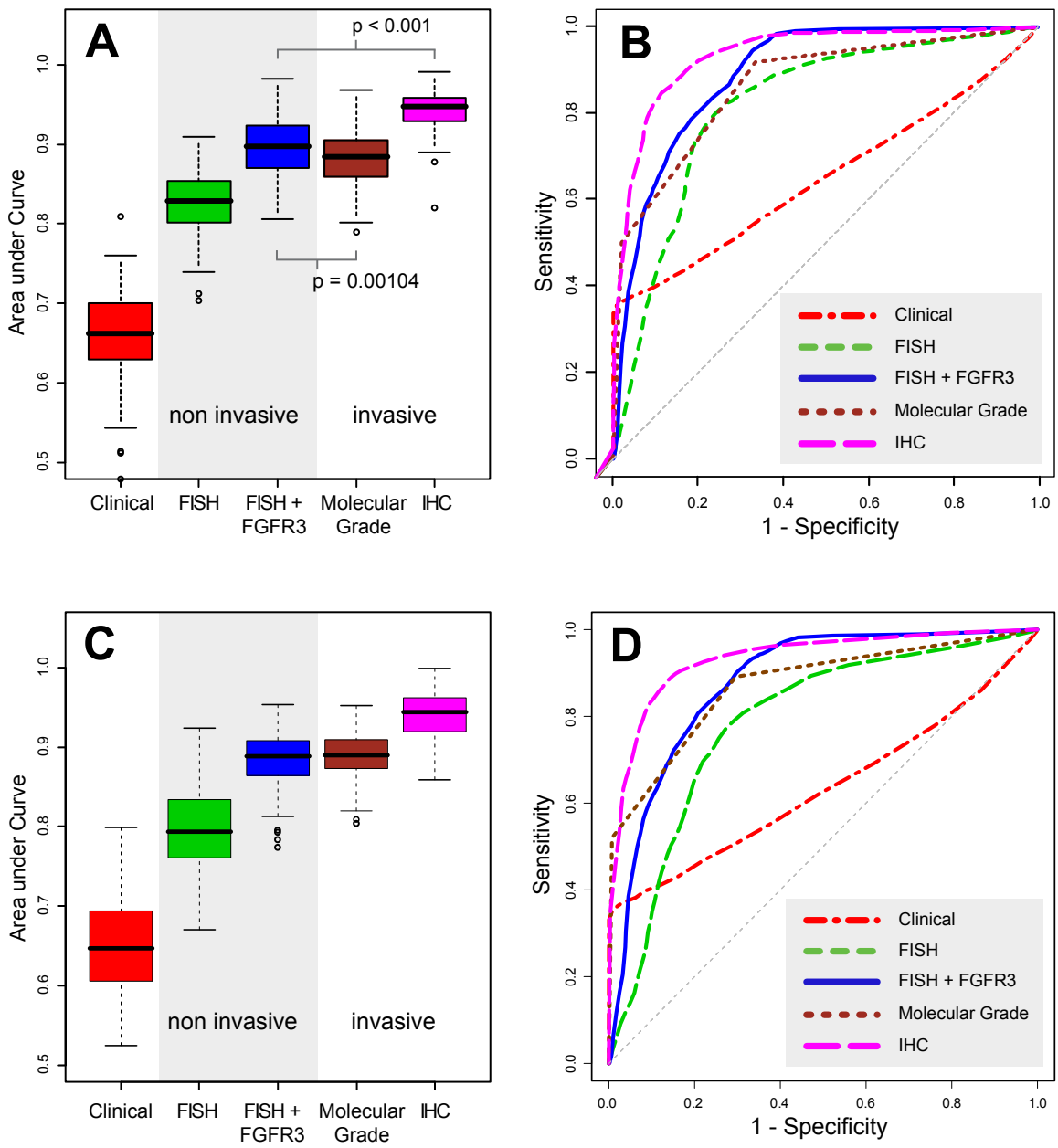


Figure 4.11

A to **D**, analysis of the prediction performance of the three models based on noninvasive predictors (CLINICAL, FISH, and FISH + FGFR3) and the two models based on invasive predictors (molecular grade and immunohistochemistry). The box plots show the area under the ROC curve to predict high histologic grade (**A** and **B**) and infiltrative growth (stage $\geq$pT1; **C** and **D**) based on 100 cross-validation experiments. Corresponding ROC curves for the three models based on noninvasive predictors (CLINICAL, FISH, and FISH + FGFR3) and the two models based on invasive predictors (molecular grade and immunohistochemistry). The ROC curves plotted are generated by varying the threshold of the logistic regression model $log(\frac{p(x)}{1-p(x)}) = b_0 + \sum_{i=1}^{k} b_i * x_i$ and are the average curves based on 100 cross-validation experiments.

Variable	Categorization	Tumor recurrence			Overall survival		
		n^*	Events	$P^\dagger$	n^*	Events	$P^\dagger$
Pathologic data							
Tumor stage	pT_a	146	72	0.7534	146	4	**<0.0001**
	pT_1	48	18	48	3		
	pT_2	56	15	56	27		
	pT_3	2	1	2	2		
	pT_4	3	0	3	2		
Histologic grade	Low	150	49	0.176	150	5	**<0.0001**
	High	105	32	105	33		
Adjacent carcinoma *in situ*	No	222	95	0.6429	222	26	**0.0001**
	Yes	33	11	33	12		
Multifocality	Unifocal tumor	53	19	0.7129	53	14	**0.0029**
	Multifocal tumor	202	87	202	24		
Growth pattern	Papillary	207	95	0.3254	207	13	**<0.0001**
	Solid	47	10	47	24		
Immunohistochemistry							
MIB1	≤25%	168	76	0.7484	168	13	**<0.0001**
	>25%	68	23	68	24		
TP53	≤10%	179	80	0.5483	179	22	**0.0161**
	>10%	66	22	66	16		
CK20	Superficial staining pattern	49	23	0.6535	49	2	**0.0155**
	Negative or >10%	192	74	192	35		
Molecular data							
FGFR3 mutational status	Wild-type	110	38	0.1382	110	24	**0.0026**
	Mutation	98	50	98	7		
Relative p16 deletion	≤14%	128	56	0.881	128	12	**0.009**
	>14%	108	40	108	22		
Polysomy	≤18%	71	31	0.958	71	3	**0.004**
	>18%	165	65		165	31	

Table 4.2

Univariate analyses of factors possibly influencing recurrence-free and overall survival. Only the initial biopsy of each patient is included. (Log-rank test (two-sided); boldface representing P values < 0.05.)

histopathologic characteristics. Interestingly, a relative p16 deletion was predominantly found in solitary compared with multifocal urothelial bladder tumors ($P < 0.001$). As expected, polysomy and relative p16 deletion were significantly associated with wild-type FGFR3 status, high proliferation, high TP53 immunoreactivity, and abnormal CK20 staining pattern (Table 4.1).

Molecular and Immunohistochemical Markers and Disease Course. The end points in the retrospective tissue microarray study were recurrence-free and overall survival. Patients with bladder tumors showing relative p16 deletion or polysomy had a significantly shorter overall survival. Table 4.2 shows univariate P values for all variables investigated. None of the parameters showed a prognostic effect for tumor recurrence.

We investigated the association between the molecular markers and overall survival more closely by adjusting a Cox regression model (Table 4.3). In the global model (including tumor stage, grade, adjacent carcinoma in situ, multifocality, growth pattern, FGFR3 status, MIB1 immunohistochemistry, TP53 immunohistochemistry, CK20 staining pattern, polysomy, and relative p16 deletion), only solid growth pattern proved to be an independent predictor of shorter overall survival ($P = 0.002$). After stepwise reverse selection (limit $P = 0.1$), histologic grade, growth pattern, and TP53 immunohistochemistry remained in the model. The estimated probability of death was at least six times higher in

patients with high-grade bladder cancer than that in patients with low-grade cancers ($P = 0.003$).

Assuming different model constructs, the following conclusions can be drawn: (a) histologic grade (high versus low grade) and growth pattern (solid versus papillary) are independent prognostic factors for overall survival and (b) relative p16 deletion and polysomy cannot be considered independent prognostic factors for the survival probability of bladder cancer patients.

Model Comparison. Given the prognostic effect of histologic grade, sensitivity and specificity for the detection of high-grade tumors were calculated. In this study, we compared five different models for their power to predict the surrogate markers high grade and infiltrative tumor growth (stage $\geq$pT1) using a logistic regression model. The first model consisted of the clinical parameters *sex*, *age*, *adjacent carcinoma in situ*, and *multifocality* (CLINICAL). The second model comprised *polysomy* and *relative p16 deletion* (FISH). The third model extends the FISH model for the *FGFR3 mutational status* (FISH + FGFR3). The latter two models were constructed with markers, which could also be measured noninvasively using urine. The fourth model was the molecular grading model from van Rhijn et al. (2003b) and consisted of MIB1 immunohistochemistry and FGFR3 mutational status (molecular grade). The last model consisted of the immunohistochemical markers MIB1, TP53, and the CK20 pattern.

Targeting high tumor grade, the (noninvasive) FISH + FGFR3 model performs slightly better ($P = 0.001$) than the molecular grading model from van Rhijn et al. (2003b) with respect to the area under curve (AUC). Both have an AUCof ≈ 0.9 as shown in Figure 4.11A and B. Observing the ROCcurves in Figure 4.11B in detail, it can be seen that on average FISH + FGFR3 is more sensitive and that van Rhijn et al. molecular grading is more specific regarding high grade. The classic immunohistochemical markers were superior to FISH + FGFR3 ($P < 0.001$) and the molecular grade ($P < 0.001$). The clinical markers alone (CLINICAL) perform worse than all other discussed models. FISH analysis alone failed only in 4 of 96 high-grade bladder cancer cases (3 pT1 and 1 pT2 tumor). All other high-grade tumors were FISH positive.

Targeting infiltrative tumor growth (stage $\geq$pT1), the FISH + FGFR3 model and the molecular grading model performed equally well with respect to the AUC(0.9; Figure 4.11C and D). Again, classic immunohistochemistry markers are superior to FISH + FGFR3 and the molecular grade.

Our results show that a model consisting of UroVysion FISH and FGFR3 status (FISH + FGFR3) can predict high grade just as well as the molecular grade proposed by (van Rhijn et al., 2003b) and feature a higher sensitivity. Although a model based on the classic immunohistochemical markers is still the most powerful regarding the target high tumor grade, our FISH + FGFR3 model reaches nearly the same accuracy (0.9 versus 0.95 AUC) only with markers that could be measured with noninvasive techniques using urine samples instead of paraffin-embedded specimens. Parameters and P values of the FISH + FGFR3 logistic re-

Variable	Categorization		Global P	Reverse selection (limit P = 0.1)	
				Hazard ratio (95% confidence interval)	P
Pathologic data					
Tumor stage	pT_a	0	0.558	—	
	pT_{1-4}	1			
Histologic grade	Low	0	0.199	6.608 (1.929-22.633)	**0.003**
	High	1			
Adjacent carcinoma *in situ*	No	0	0.516	—	
	Yes	1			
Multifocality	Unifocal tumor	0	0.304	—	
	Multifocal tumor	1			
Growth pattern	Papillary	0	**0.002***	4.804 (1.959-11.783)	**0.001**
	Solid	1			
MIB1 immunohistochemistry	≤25%	0	0.823	—	
	>25%	1			
TP53 immunohistochemistry	≤10%	0	0.106	0.488 (0.213-1.114)	0.088
	>10%	1			
FGFR3 gene	Wild-type	0	0.199	—	
	Mutation	1			
CK20 immunohistochemistry	Superficial staining pattern	0	0.585	—	
	Negative or >10%	1			
Relative *p16* deletion	≤14%	0	0.945	—	
	>14%	1			
Polysomy	≤18%	0	0.918	—	
	>18%	1			

Table 4.3

Multivariate analysis of factors possibly influencing overall survival ($n = 186$). (Boldface representing P values < 0.05)

gression model are given in Table 5.5. The model contained polysomy, relative p16 deletion, and FGFR3 mutational status as predictors from which polysomy and FGFR3 status are significant for the prediction of high grade.

Voided Urine Study. An independent set of 119 voided urine samples were investigated to estimate the diagnostic power of the different assays. After cytologic examination, voided urine samples were matched with corresponding biopsies and analyzed for LOH, FGFR3 mutation, polysomy, and relative p16 deletion using UroVysion FISH. A mutated FGFR3 gene was found significantly more frequent in patients with malignant disease ($P = 0.005$; Table 4.4). Only one patient with a normal biopsy was found to have a urine specimen with mutated FGFR3. Of note is that the same patient developed a pTa G2 bladder tumor with mutated FGFR3 within the same year of follow-up. Mutated FGFR3 was significantly associated with a positive cytologic result ($P < 0.001$; Table 4.4). None of the cytologically negative cases displayed a FGFR3 mutation. Separate results of the analysis of voided urine samples for patients with high-grade ($n = 40$) and low-grade ($n = 41$) bladder tumors are given in Table 5.4.

Sensitivity, specificity, and positive and negative predictive values for the detection of bladder cancer cells in voided urine were calculated (Figure 4.12A-F). Six different models were investigated for their power to predict bladder cancer cells in urine. The three methods (CYTO, LOH, and FGFR3) were tested individually (Figure 4.12A-C). Additionally, FGFR3 and LOH analysis were tested in combination with cytology (CYTO + FGFR3, and CYTO + LOH). Targeting neoplastic cells, the combination of cytology with LOH and FGFR3 analysis per-

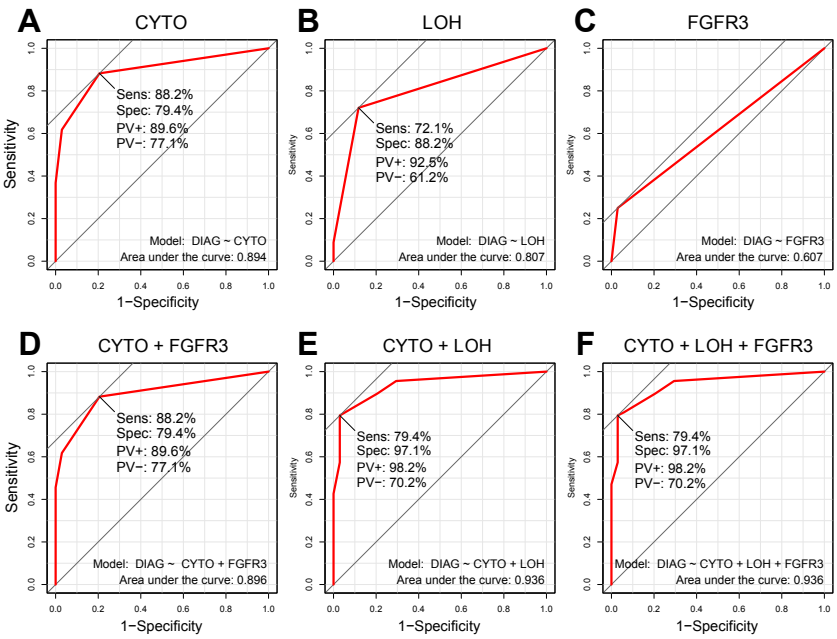


Figure 4.12

A to **D**, ROC curves for six different models to predict bladder cancer cells in a set of prospectively collected voided urine samples. The three methods (CYTO, LOH, and FGFR3) were tested individually for their ability to detect bladder cancer cells in urine (**A**-**C**). Additionally, FGFR3 and LOH analysis were tested in combination with cytology (**D** and **E**). The combination of the three techniques (CYTO + LOH + FGFR3) did not add significant diagnostic advantage (**F**).

formed equally well, respectively. Observing the ROCcurves in Figure 4.12D and E in detail, it can be seen that on average CYTO + FGFR3 is slightly more sensitive and CYTO + LOH is more specific for the detection of bladder cancer cells. The combination of the three techniques (CYTO + LOH + FGFR3) did not add significant diagnostic advantage (Figure 4.12F) when compared with the dual models. Results of the FISH assay could not be included in the logistic regression analysis for cancer cell detection because only 2 of the 38 (5%) non-neoplastic urine samples were analyzable with FISH. Sixty-nine percent of the urine samples from patients with a malignant bladder biopsy were UroVysion FISH positive. Measures of the performance of the various assays for the detection of highgrade versus non-high-grade tumors are provided in Figure 5.3A to F.

4.4.5 Discussion

In this study, we show that the combination of classic cytology with LOH analysis reaches the highest diagnostic accuracy for the detection of urothelial bladder cancer cells in voided urine samples.

To analyze large numbers of bladder cancer specimens, we first evaluated a tissue microarray comprising 255 consecutive, primary bladder cancers. In our survey, solid growth pattern and high histologic grade were the most important prognostic factor for overall survival. However, histologic grading has a high interobserver variability with varying prognostic implications (Tosoni et al., 2000). Burger et al. (2008) have shown that the current WHO classifica-

Variable	Categorization	*FGFR3* analysis		
		Wild-type	Mutation	*P**
Cystoscopically obtained biopsies				
Histologic diagnosis	Nonmalignant	33	1	**0.005**
	Malignant	51	17	
Voided urine samples				
Cytologic diagnosis	Negative cytology (N)	35	0	**<0.001**
	Atypical cytology, not specified (NR)	20	4	
	Atypical cytology, suspicious (S)	12	6	
	Positive cytology (P)	17	8	
LOH	No LOH	45	4	**0.032**
	Suspected LOH	5	1	
	LOH	34	13	
UroVysion FISH	No polysomy or loss of p16	11	2	0.290
	Polysomy or loss of p16	20	10	

Table 4.4

FGFR3 analysis of voided urine samples in relation to histology, cytology, and molecular analyses. (Fisher's exact test, two-sided; boldface representing P values < 0.05.)

tion (Grignon et al., 2004) reflects the outcome of bladder cancer patients more accurately than the 1973 classification system (Mostofi et al., 1973). The authors concluded that novel methods including molecular markers need to be evaluated for clinical use. In a second study on urothelial bladder cancer from the same group ($n = 221$), Burger et al. (2008) prospectively investigated the prognostic value of the WHO 1973 and 2004 grading systems and biomarkers FGFR3, CK20, and Ki-67. They found that both grading systems contribute valuable independent information. Interestingly, combining WHO2004 grading with FGFR3 status allowed a better risk stratification for patients with high-grade non-muscle invasive urothelial bladder cancer.

A set of molecular and immunohistochemical markers was evaluated to develop a reliable and objective grading systems of urothelial bladder cancer. A model consisting of UroVysion FISH and FGFR3 status (FISH + FGFR3) predicted high grade significantly better compared with the molecular grade proposed by van Rhijn et al. (2003b) (Figure 4.11A and B).

In general, urethrocystoscopy (every 3-4 months for the first 2 years and longer intervals in subsequent years) remains the standard of care for the detection and follow-up of urothelial bladder cancer. Interestingly, van der Aa et al. (2008) have assessed the discomfort and pain reported during follow-up of patients ($n = 220$) with non-muscleinvasive low-grade urothelial bladder cancer comparing urethrocystoscopy and surveillance by microsatellite analysis. According to van der Aa et al., periodic urethrocystoscopy caused pain and discomfort in about a third of the patients, whereas the burden of microsatellite analysis appeared fully attributable to the waiting time for the test result. The authors concluded that less invasive surveillance tests are urgently needed (van der Aa et al., 2008).

But can the results of our aforementioned tissue microarray study be used for the detection of neoplastic cells in voided urine? To address this question, we

estimated the diagnostic power for the detection of bladder cancer cells in 119 voided urine samples using LOH and FGFR3 analysis, UroVysion FISH, and cytology as predictors. We could show that the combination of classic cytology with LOH analysis (CYTO + LOH) significantly increased the accuracy to detect malignant urothelial cells in voided urine (Figure 4.12E). In our study, sensitivity and specificity of conventional cytology was already 88.2% and 79.4% ($AUC = 0.894$; Figure 4.12A). Using a combination of cytology and FGFR3 analysis (CYTO + FGFR3), sensitivity and specificity could not be increased (Figure 4.12D). The combination of cytology with microsatellite analysis (CYTO + LOH) was able to increase specificity (97.1%) and the area under the ROCcurve (Figure 4.12D). However, sensitivity slightly decreased to 79.4%. Combination of the three techniques (CYTO + LOH + FGFR3) did not add significant diagnostic advantage (Figure 4.12F).

These results are contrary to data published by van Rhijn et al. (2003a) who have also combined LOH and FGFR3 mutation analysis (molecular grade) for the detection of urothelial cancer cells in voided urine. After cytologic examination, an independent set of voided urine samples was matched with corresponding biopsies and analyzed for LOH, FGFR3 mutation, polysomy, and p16 deletion using UroVysion FISH. Combining results of LOH and FGFR3 mutation analysis, the sensitivity of the combined approach increased to 89% and was superior to the sensitivity of conventional cytology for every clinical subdivision (van Rhijn et al., 2003a). In our study, however, sensitivity and specificity of conventional cytology were already very high (88.2% and 79.4%). The area under the ROC curve in Figure 4.12A (AUC0.894) could only be increased by adding the results of the microsatellite analysis ($AUC = 0.936$; Figure 4.12E).

Recently, (van der Aa et al., 2009) have reported the results of a longitudinal prospective multicenter trial for surveillance of patients with low-grade non-muscle invasive urothelial cancer using microsatellite analysis ($n = 228$). The authors concluded that microsatellite analysis on voided urine samples is not sufficiently sensitive to recommend implementation in routine clinical practice. Cytologic examination of voided urine is unexpensive, established in almost every pathology department, and should be part of any bladder cancer surveillance protocol. However, application of urine cytology is operator-dependent and can be hampered by the low sensitivity for low-grade lesions (Sherman et al., 1984). In contrast to our study, simultaneous cytologic examinations were not taken into account by van der Aa et al. when calculating sensitivity and specificity of the various tests (van der Aa et al., 2009; van Rhijn et al., 2003a).

The combination of cytology with LOH analysis reached the highest diagnostic accuracy for the detection of urothelial bladder cancer cells in voided urine samples. A monitoring scheme alternating invasive cystoscopy with a combination of noninvasive techniques (including classic urine cytology and LOH analysis) could reduce unpleasant interventions and improve follow-up compliance of patients with recurrent urothelial bladder cancer.

4.5 Learning a Siganture for Clinical Outcome Prediction in Malignant Melanoma.

4.5.1 Overview

Background: Current staging methods such as tumor thickness, ulceration and invasion of the sentinel node are known to be prognostic parameters in patients with malignant melanoma (MM). However, predictive molecular marker profiles for risk stratification and therapy optimization are not yet available for routine clinical assessment.

Methods: Using tissue microarrays, we retrospectively analyzed samples of 364 patients with primary MM. We investigated a panel of 70 immunohistochemical (IHC) antibodies for cell cycle, apoptosis, DNA mismatch repair, differentiation, proliferation, cell adhesion, signaling and metabolism. A marker selection procedure based on univariate Cox regression and multiple testing correction was employed to correlate the IHC expression data with the clinical follow-up (overall and recurrence-free survival). The model was thoroughly evaluated with two different cross-validation experiments, a permutation test and a multivariate Cox regression analysis. The predictive power of the identified marker signature was validated on a second independent external test cohort ($n = 225$).

Results: A signature of seven biomarkers (Bax, Bcl-X, β-Catenin, CD20, COX-2, MTAP, PTEN) was found to be an independent predictor for overall and recurrence-free survival in patients with MM. The seven-marker signature could also predict those patients with worse prognosis despite small tumor thickness ($\leq 2.00mm$). In particular, three of these markers (CD20, COX-2, MTAP) were shown to offer direct therapeutic implications.

Conclusions: Our seven-marker signature is closely associated with the prognosis of patients with MM and offers direct therapeutic implications.

Acknowledgments

I want to express my sincere gratitude to Stefanie Meyer, who is an excellent collaborator and the nearly two years we worked on the project described here are a very rare example of frictionless interdisciplinary teamwork. Furthermore I want to thank our co-authors: Anja K. Bosserhoff, Ferdinand Hofstdter, Dirk Schadendorf, Volker Roth, Joachim M. Buhmann, Ingrid Moll, Nikos Anagnostou, Johanna Brandner, Holger Moch, Michael Landthaler, Thomas Vogt and Peter J. Wild

4.5.2 Introduction

Cutaneous malignant melanoma (MM) represents the most common cause of death from skin cancer, and, apart from female lung cancer, it is the tumor entity with the highest increase of incidence worldwide (Jemal et al., 2008). MM is characterized by a multi-factorial etiology. Sun exposure and genetic susceptibility have been proposed as major etiological and predisposing factors and may explain the reported increase of incidence to some degree (Lens and Dawes, 2004). The metastatic stage IV of MM with an average 10-year survival rate ranging from 3% to 16% (depending on its pattern of metastasis) (Balch et al., 2001a) cannot yet be cured and improvement in overall survival among these patients remains an elusive goal. Despite novel therapeutic approaches the prognosis of patients suffering from metastatic stage IV MM remains infaust (Agarwala, 2009). De facto, the prognosis of patients with MM may only

be conditionally derived from clinical and histological parameters. According to the AJCC 2009 classification, the findings of vertical tumor thickness, tumor ulceration and sentinel node biopsy (Morton et al., 2006) represent the most dominant prognostic factors. In stage pT1 melanomas ($\leq$ 1.00 thickness), the mitotic rate (histologically defined as mitoses/mm2) has to be considered as additional prognostic parameter (Balch et al., 2009).
In MM, multiple cellular factors are known to be deregulated in the initiation and progression phase of the tumor; among these are protein regulators of the cell cycle, apoptosis, signal transduction, cell adhesion and matrix digestion. Despite the fact that hundreds of studies sought to assess the potential prognostic value of molecular markers in predicting the course of cutaneous MM, according to the latest review meta-analyses (Gould Rothberg et al., 2009; Gould Rothberg and Rimm, 2010), there are no predictive molecular profiles for risk stratification or therapy optimization applicable for routine clinical assessment of MM. To this end, we examined the immunohistochemical (IHC) expression of 70 candidate biomarkers of MM including regulating proteins of the cell cycle and apoptosis control, factors of signal transduction, cell adhesion, transcription factors, differentiation, and melanoma-specific antigens using tissue microarrays (TMA). The study was based on extensive follow-up investigation of a total of 589 patients with primary MM from two independent cohorts, and initiated to identify a clear set of reliable IHC markers for routine clinical assessment of patients with primary MM. Accordingly, and as recently requested by Gould Rothberg et al. (2009); Gould Rothberg and Rimm (2010), this biomarker study aimed at identifying an independent prediction model for clinical outcome and individualized targeted therapy options in patients with MM.

4.5.3 Material and Methods

The medical ethical committee of the University of Regensburg, Germany, approved the reported experiments. The retrospective study was conducted according to the Declaration of Helsinki Principles.

Tissue Microarrays

TMAs were constructed as described previously (Simon et al., 2003) and based on primary melanoma material. TMA 1, the primary cohort, contained tissue punch samples from 364 consecutive (non-selected), formalin-fixed, paraffin-embedded MMs of 364 different patients, and were from the Department of Dermatology, University Hospital of Regensburg, Germany. TMA 2, the secondary cohort, which was used as independent external validation cohort, consisted of consecutive (non-selected) melanoma samples from 235 patients of the Department of Dermatology, University Hospital Hamburg-Eppendorf, Germany. For patients with multiple subsequent neoplasms, only initial and single

primary MMs were included. H&E-stained slides of all MMs were evaluated by two histopathologists (TV, PJW). The clinicopathological characteristics of the two independent cohorts of melanoma patients are given in Tables 5.6 and 5.7. Clinical follow-up data, provided by the local tumor registries, were available for all patients of the primary cohort (n=364) and 231 patients of the secondary cohort. Patients were censored at 120 months, if their follow-up exceeded the 10-year scope of the study.

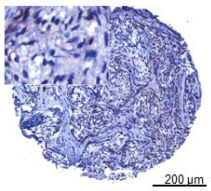
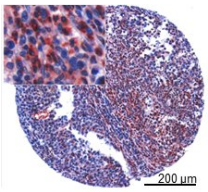
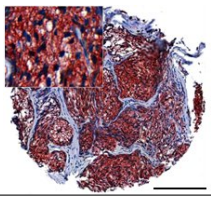
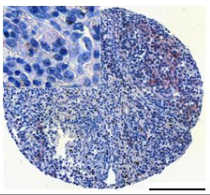
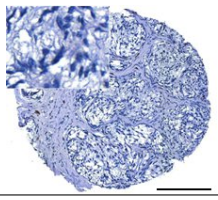
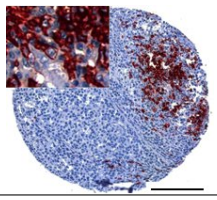
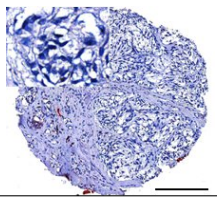
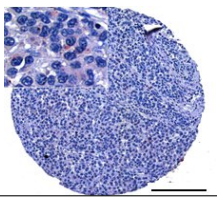
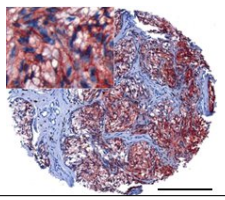
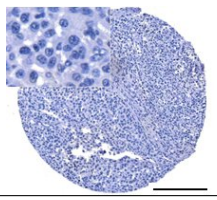
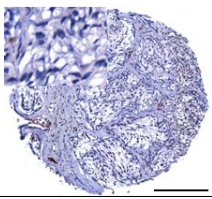
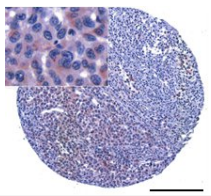
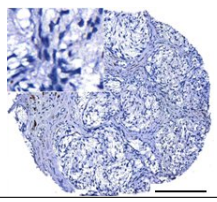
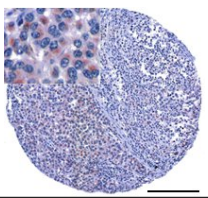
Immunohistochemical Analysis

Paraffin-embedded preparations of melanoma tissues were screened for protein expression according to standardized IHC protocols as described previously (Alonso et al., 2004; Wild et al., 2006; Meyer et al., 2009). The primary antibodies used in this study were selected for reporting on key aspects of apoptosis, cell cycle, signal transduction, cell adhesion, melanoma differentiation and proliferation, and tumor metabolism. The candidate markers were chosen because of their described role in MM in the literature (Gould Rothberg et al., 2009; Gould Rothberg and Rimm, 2010; Alonso et al., 2004) or on the basis of previous studies by Wild et al. (2006); Meyer et al. (2009).

All IHC investigations were based on an avidin-biotin peroxidase method with a 3-amino-9-ethylcarbazole (AEC) chromatogen. After antigen retrieval (steam boiler with citrate-buffer, pH 6.0 or with Tris-EDTA-buffer, pH 9.0 for 20 min) immunohistochemistry was carried out applying the ZytoChemPlus HRP Broad Spectrum Kit (Zytomed Systems, Berlin, Germany) according to the manufacturers instructions. IHC stainings were performed for 70 different primary antibodies. Cytoplasmic and nuclear markers were visualized with AEC solution (AEC+ High Sensitivity Substrate Chromogen, ready-to-use, DAKO, Glostrup, Denmark). The red color of the AEC substrate chromogen (3-amino-9-ethylcarbazole) is very beneficial to rule out the possibility of a role of endogenous melanin in the observed reactivity. All sections were counterstained with hematoxylin (DAKO). Negative controls were obtained by omitting the primary antibody. Two dermatohistopathologists (TV, SM) performed a blinded evaluation of the stained slides without knowledge of clinical data. As previously described by (Wild et al., 2006; Meyer et al., 2009), cytoplasmic and nuclear immunoreactivity was estimated using a semiquantitative scoring system (0 to 4+), depending on the intensity of cytoplasmic staining and the percentage of cell nuclei with positive staining, respectively: 0 (negative): no cytoplasmic staining or 0% of cell nuclei stained; 1+: weak cytoplasmic staining or less than 20% of cell nuclei stained; 2+: moderate cytoplasmic staining or 21 to 50% of cell nuclei stained; 3+: strong cytoplasmic staining or 51 to 90% of cell nuclei stained; 4+:very strong cytoplasmic staining or nuclear staining greater than 90%. Causes of not interpretable results included lack of tumor tissue and presence of necrosis or crush artifact.

Table 4.5

Immunohistochemically stained TMA Specimens illustrating the Seven-Marker Signature for one Patient with High-Risk and one Patient with Low-Risk Melanoma. The low-risk melanoma (Column C) shows a strong cytoplasmic staining for β-Catenin and MTAP, respectively. Immunoreactivity of these two protective markers was not found in the high-risk melanoma (Column D). In contrast, the high-risk melanoma demonstrated features a moderate to strong cytoplasmic staining for Bax, CD20, Bcl-X, PTEN and COX-2.

Protein	Low-risk signature	High-risk signature
Bax: Bcl-2 family protein, proapoptotic *Cytoplasmic*		
fi-Catenin: Key downstream effector in Wnt signaling pathway, implicated in two major biological processes: embryonic development and tumorigenesis *Cytoplasmic and nuclear*		
CD 20: Pan-B-cell marker, stem cell marker candidate *Cell membrane and cytoplasmic*		
Bcl-X: Bcl-2 family protein, antiapoptotic *Cytoplasmic*		
MTAP: Constitutive enzyme of the polyamine metabolism; modulator of interferon-dependent signaling *Cytoplasmic*		
PTEN: Tumor suppressor implicated in a variety of human cancers; major negative regulator of the PI3K/Akt signaling pathway *Cytoplasmic and nuclear*		
Cox-2: Cyclooxygenase for biosynthesis of prostaglandins under inflammatory conditions; overexpressed in a variety of tumors *Cytoplasmic*		

	High risk (N=181)	Low risk (N=181)	p-Value: high vs. low risk	Multivariate Cox Regression Analysis Hazard ratio(95% CI)	p-Value
7-Marker risk score	0.267 ± 0.092	0.0017 ± 0.12	<< 0.0001 [#1]	13.54 (4.27–42.97)	0.0000098 ***
Age — yr	59.5 ± 15.0	57.7 ± 14.9	0.263 [#1]	1.03 (1.02–1.05)	0.0000027 ***
Sex — no. of patients (%)					
Male	105 (58)	89 (49.2)	1 [#2]	1.98 (1.36–2.89)	0.00034 ***
Female	76 (42)	92 (50.8)			
Tumor thickness — mm	2.52 ± 2.38	1.40 ± 2.21	0.00000646 [#1]	1.17 (1.12–1.23)	0.00000000023 ***

Table 4.6

Clinical Characteristics of the Primary Cohort of Patients with MM (TMA 1). Comparing high-risk patients (first column) with low-risk patients (second column) based on their seven-marker risk score shows significant difference in tumor thickness ($p < 0.001$) and no difference in sex ($p = 1$) and age ($p = 0.263$). Furthermore, hazard ratios and p-values are reported for a multivariate Cox regression model comprising all listed variables. Regarding overall survival the seven-marker risk score is statistically significant ($p < 0.001$) independent of sex, age and tumor thickness. Continuous variables are reported with mean and standard deviation and categorical variables are listed with number of counts and percentages.

Statistical Analysis

One of the main statistical problems in large scale IHC studies are missing values in the design matrix due to missing or corrupt spots on the TMA. The more markers are investigated the higher the chance that at least one value is missing per patient. Frequently this problem is addressed by either sacrificing a larger number of patient records or by employing volatile multiple imputation techniques. In this study 9.3% of values are missing which would reduce the set of patients with all IHC measurements from 364 to 170. Algorithms like random survival forests (Ishwaran et al., 2008) and ensemble learning with gradient boosting (Hothorn et al., 2006) are capable of dealing with missing values, but lead to models, which are not intuitively interpretable and difficult to implement in clinical practice. To overcome these problems we employed the following learning procedure which is invariant to missing values and results in an easily interpretable and practically applicable linear model.

Prognostic power of the 70 markers was assessed by learning univariate proportional hazard models (Cox, 1972), yielding 11 markers significantly associated with overall survival. To correct for multiple testing, the false discovery rate (FDR) procedure (Benjamini and Hochberg, 1995) was applied with a FDR of 0.15 reducing the set of significantly associated markers to 9. A risk score is calculated for each patient by a linear combination of the univariate Cox regression coefficients β and the corresponding IHC measurements X. Finally, the score is normalized by the number of markers measured:

$$\text{score}(x) = \left(\sum_{i=1}^{D}(\beta_i x_i)\alpha_i\right) / \left(\sum_{i=1}^{D}\alpha_i\right), \quad \alpha_i = \begin{cases} 1, & \text{if } x_i \text{ exists} \\ 0, & \text{if } x_i \text{ is missing} \end{cases} \tag{4.14}$$

where D is the number of markers in the signature. Based on this risk score, patients were assigned to a high risk group and a low risk group, split at the 50^{th} percentile (median) of all scores. Thus, the final model consists of the coefficient vector β and the median threshold θ.
Nonparametric Kaplan-Meier estimators (Kaplan and Meier, 1958) were used to analyze overall survival and recurrence-free survival. Differences between survival estimates were assessed with the log-rank test (LRT) (Mantel and Haenszel, 1959) P-values below 0.05 were considered to indicate statistical significance. Statistical analyses were conducted using R version 2.11 (R Development Core Team, 2009).

Statistical Validation

The validity of the learning procedure and hence the accuracy of the signature was assessed in three different validation experiments.
The cross-validation experiments were conducted as follows:

1. Divide the patients into K cross-validation folds (groups) at random.
2. For each fold $k = 1, 2, \ldots, K$
   a) Find a subset of univariate statistical significant (LRT $p < 0.05$) predictors for the overall survival, using all of the patients except those in fold k.
   b) Filter the selected predictors based on a FDR of 0.15.
   c) Using just this subset of predictors, build a multivariate linear model using the formulation in Equation 4.14, using all of the patients except those in fold k.
   d) Use the model to predict the score for the patients in fold k.
3. Aggregate the out-of-bag predictions of all patient and split them in two groups based on the median predicted score.
4. Calculate the Kaplan-Meier estimator for each group and report the LRT p-value of their difference in survival expectation.

First, leave-one-out cross-validation was employed by excluding one patient at a time from the training set and subsequently scoring the left out patient with the signature learned from the rest of the patients. Repeating this procedure 364 times yields a leave-one-out score estimate for each patient in the study. The resulting difference between high risk patients and low risk patients was highly significant ($p < 0.001$) and is depicted in Figure 4.15 C and D.
Second, 10-fold cross-validation (Hastie et al., 2009) was conducted by partitioning the dataset into 10 parts of equal size using 90% of the patients for learning and 10% for testing. The procedure was repeated 10 times resulting in a 10-fold score for each patient. The resulting differentiation between high

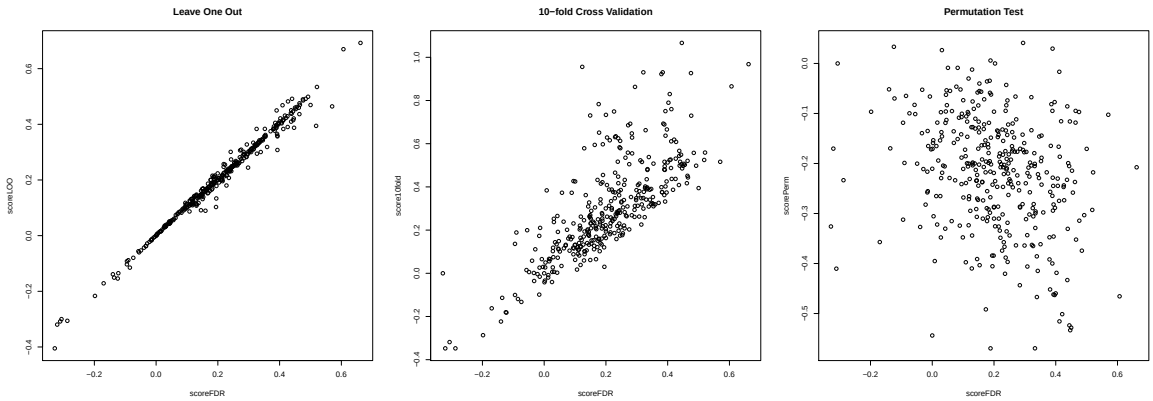


Figure 4.13

Cross-validation scores plotted against training fit scores. From left to right: 10-fold cross-validation, leave one out (LOO) cross-validation and permutation test. It can be seen that LOO has a clearly lower bias than 10-fold cross-validation. The permutation scores do not correlate with the training scores indicating that the proposed selection scheme is not able to learn a signature for the permuted survival times and hence does not overfit.

risk and low risk patient was worse in terms of the LRT p-value but was still highly significant ($p < 0.001$) as shown in Figure 4.16C and D.
The third validation experiment was conducted to assess, if the proposed marker selection procedure is prone to over fitting. To this end the target variable was randomly permuted and a model was learned to predict the risk score based on this distorted data. Figure 4.15E illustrates that it was impossible to learn a meaningful score ($p > 0.5$) based on the permuted labels. This indicates that the proposed algorithm does not over fit, although a large number of markers were tested to learn the signature.

4.5.4 Results

The Nine-Marker Signature and Survival

The proposed learning procedure based on the Cox regression coefficients and multiple testing correction with FDR yielded 9 markers which were correlated with death from any cause: two were protective markers (associated with a hazard ratio of less than 1.00) and seven were risk markers (associated with a hazard ratio of more than 1.00) (Figure 4.18).
Among the nine markers were Bax and Bcl-X, two major regulators of the intrinsic mitochondrial apoptosis pathway (Lowe et al., 2004), and β-Catenin, a key downstream effector in the *Wnt* signaling pathway (Delmas et al., 2007). Moreover, CD20, a known B-cell marker recently suggested as candidate marker for melanoma stem cells (Zabierowski and Herlyn, 2008) and CD49d, an α4-integrin (ITGA4) participating in cell-surface mediated signaling and adhesion, were included (Kuphal et al., 2005). Apart from this, COX-2, a cyclooxygenase also referred to as Prostaglandin H Synthase 2 which was shown to be overex-

	High risk (N=181)	Low risk (N=181)	p-Value: high vs. low risk	Multivariate Cox Regression Analysis Hazard ratio(95% CI)	p-Value
7-Marker risk score	0.270 ± 0.098	0.04 ± 0.071	<< 0.0001 [#1]	24.91 (3.84–161.71)	0.00076 ***
Age — yr	55.2 ± 16	52.6 ± 6.6	0.267 [#1]	1.02 (1.00–1.04)	0.016 *
Sex — no. of patients (%)					
Male	81 (55.1)	41 (54.7)	1 [#2]	0.67 (0.39–1.14)	0.14
Female	66 (44.9)	34 (45.3)			
Tumor thickness — mm	2.55 ± 2.67	1.08 ± 1.17	0.0000000512 [#1]	1.28 (1.19–1.38)	0.000000000028 ***

Table 4.7

Clinical Characteristics of the the External Test Cohort of Patients with MM (TMA 2). Comparing high-risk patients (first column) with low-risk patients (second column) based on their seven-marker risk score shows significant difference in tumor thickness ($p < 0.001$) and no difference in sex ($p = 1$) and age ($p = 0.267$). Furthermore, hazard ratios and p-values are reported for a multivariate Cox regression model comprising all listed variables. Regarding overall survival the seven-marker risk score is statistically significant ($p < 0.001$) independent of sex, age and tumor thickness. Continuous variables are reported with mean and standard deviation and categorical variables are listed with number of counts and percentages.

pressed in a variety of tumors including MMs was found (Meyer et al., 2009). MLH1, a DNA mismatch repair protein (Korabiowska et al., 2006), MTAP, a housekeeping enzyme" in polyamine metabolism which may modulate interferon response mechanisms (Wild et al., 2006; Behrmann et al., 2003), and the tumor suppressor phosphatase and tensin homolog PTEN were part of the signature. PTEN counteracts one of the most critical cancer promoting pathways (Zhang and Yu, 2010), the phosphatidylinositol 3-kinase (PI3K)/Akt signaling pathway. Clinically, PTEN mutations and deficiencies are prevalent in many types of human cancers and loss of functional PTEN has substantial impact on multiple aspects of cancer development. MTAP and β-Catenin were the only protective markers, whereas the other seven markers (Bax, Bcl-X, CD20, CD49d, COX-2, MLH1, PTEN) were assigned risk markers.

Table 4.6 lists the characteristics of 362 patients in the study (two patients were removed due to lack of all nine-markers from the signature). Among these 362 patients of the primary cohort tumors with high risk scores expressed risk markers, whereas tumors with low risk scores expressed protective markers (Figure 4.14A). Patients with a high-risk nine-marker signature had a lower median overall survival than those with a low-risk nine-marker signature (90 months versus not reached) (Figure 4.14B). Patients with tumors with a high-risk marker signature were associated with a lower median recurrence-free survival than tumors with a low-risk gene signature (36 months versus 88) (Figure 4.14C).

The cross-validation experiments showed comparable results and demonstrated that learning a marker signature for overall survival is feasible and reproducible (Figure 4.16C, D). For leave-one-out cross-validation, patients with high risk

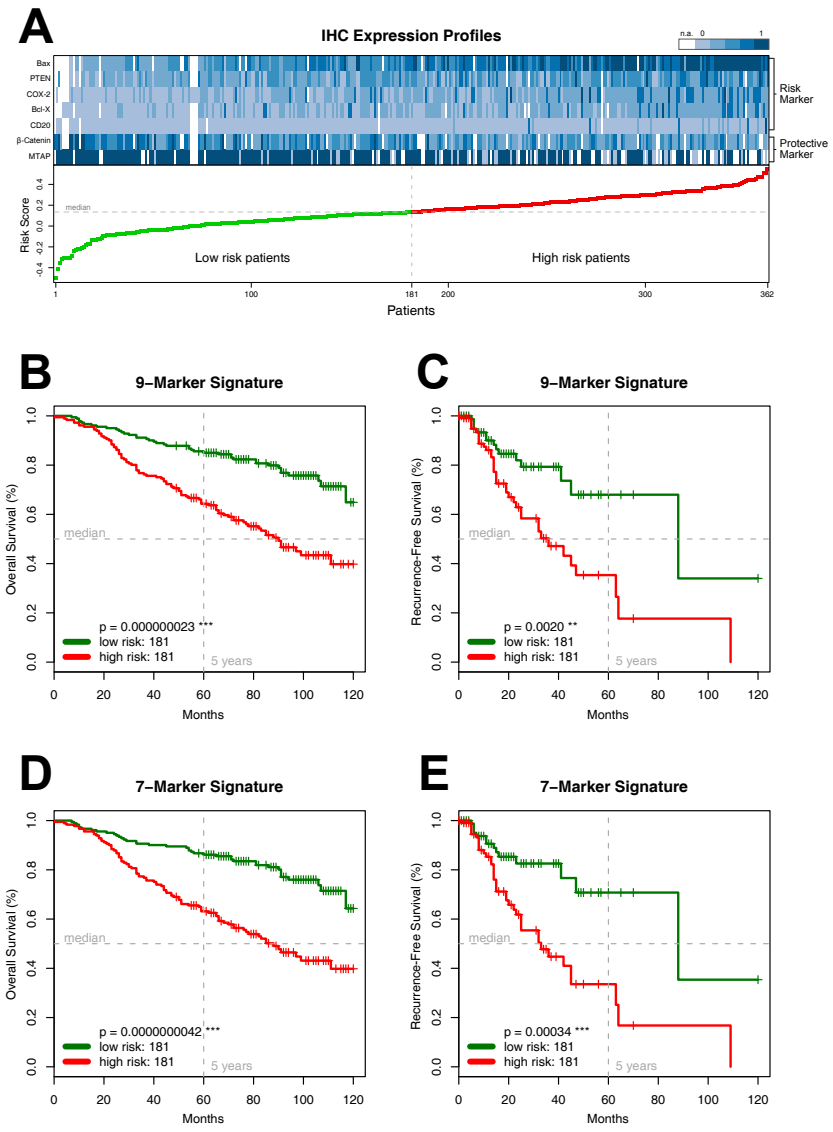


Figure 4.14

The Seven-Marker Signature and Survival of 362 Patients with Primary MM.

Panel A shows the IHC expression profiles of 362 tumor specimens from the primary cohort ordered by their predicted risk score. Each column represents an individual patient consisting of the expression values of the seven-marker signature (5 risk markers and 2 protective markers). The magnitude of the corresponding risk score is plotted below for 181 low risk patients (green) and 181 high risk patients (red). IHC expression values were scaled between 0 (light blue) and 1 (dark blue) for plotting only. W and white cells represent missing values (n.a.).

Panels B E show Kaplan-Meier estimates of overall and recurrence-free survival for high risk patients (red) and low risk patients (green) from the primary cohort according to the nine-marker signature (**Panels B, C**) and its reduced version, the seven-marker signature (**Panels D, E**), respectively. Equality in survival expectancy of the subgroups is assessed by the log-rank test. Removing the two "unspecific" markers (MLH1 and CD49d) from the signature does not reduce the statistical power of the predicted risk score. The difference between high risk patients and low risk patients is highly significant ($p < 0.001$) for the seven-marker signature.

scores had a median survival of 94 month whereas median survival for patients with low risk signature was not reached (Figure 4.15C). The difference in survival expectancy between patients with high-risk score and low-risk score was highly significant ($p = 0.000067$). Although 10-fold cross-validation has lower bias and higher variance the difference between the high risk and low risk group (94 month versus not reached) was still significant ($p = 0.00017$) as shown in Figure 4.16C. In contrast to the cross-validation experiments it was not possible to learn a signature to predict permuted labels ($p > 0.5$), which indicates that the proposed learning procedure is not over fitting. In the permutation test median survival was not reached by any risk group (Figure 4.15E).

The Seven-Marker Signature and Survival

The aim of this study was to provide a maximum of prognostic and therapeutically relevant information by a minimum of markers combined in a clear signature. For the sake of clinical feasibility and cost effectiveness, an IHC marker set suitable for routine clinical assessment should be based on a limited number of antibodies. Accordingly, the 9-marker signature was reduced by the risk marker with the lowest Cox regression coefficients $\boldsymbol{\beta}$, i.e. MLH1 ($\beta = 0.254$). Subsequently, the remaining six risk markers were evaluated regarding their impact on cancer development and progression and potential therapeutic implications. In this setting, CD49d, an α4-integrin (ITGA4) participating in cell-surface mediated signaling and adhesion, was considered to be the most dispensable marker.

Among the 362 patients of the primary cohort, patients with a high-risk seven-marker signature (Bax, Bcl-X, β-Catenin, CD20, COX-2, MTAP, PTEN) had a shorter median overall survival than the patients with a low-risk seven-marker signature (88 months versus not reached) and the difference between the two patient groups was highly significant ($p = 4.2 \cdot 10^{-9}$) (Figure 4.14D). The high-risk seven-marker signature was associated with a median recurrence-free survival of 33 months, whereas the low-risk seven-marker signature was associated with a median recurrence-free survival of 88 months (LRT $p = 0.00034$) (Figure 4.14E).

According to multivariate Cox regression analysis, the *seven-marker risk score*, *tumor thickness*, *sex*, and *age* were significantly associated with death from any cause among the 356 patients (6 observations were deleted due to missing values) (Table 4.6).

A subgroup analysis of 253 patients with a tumor depth of $\geq 2mm$ revealed that those 148 patients with a high-risk marker signature had a significant ($p = 0.0053$) shorter overall survival (Figure 4.15A) and recurrence-free survival ($p = 0.008$) than the 105 patients with a low-risk marker signature (Figure 4.15B).

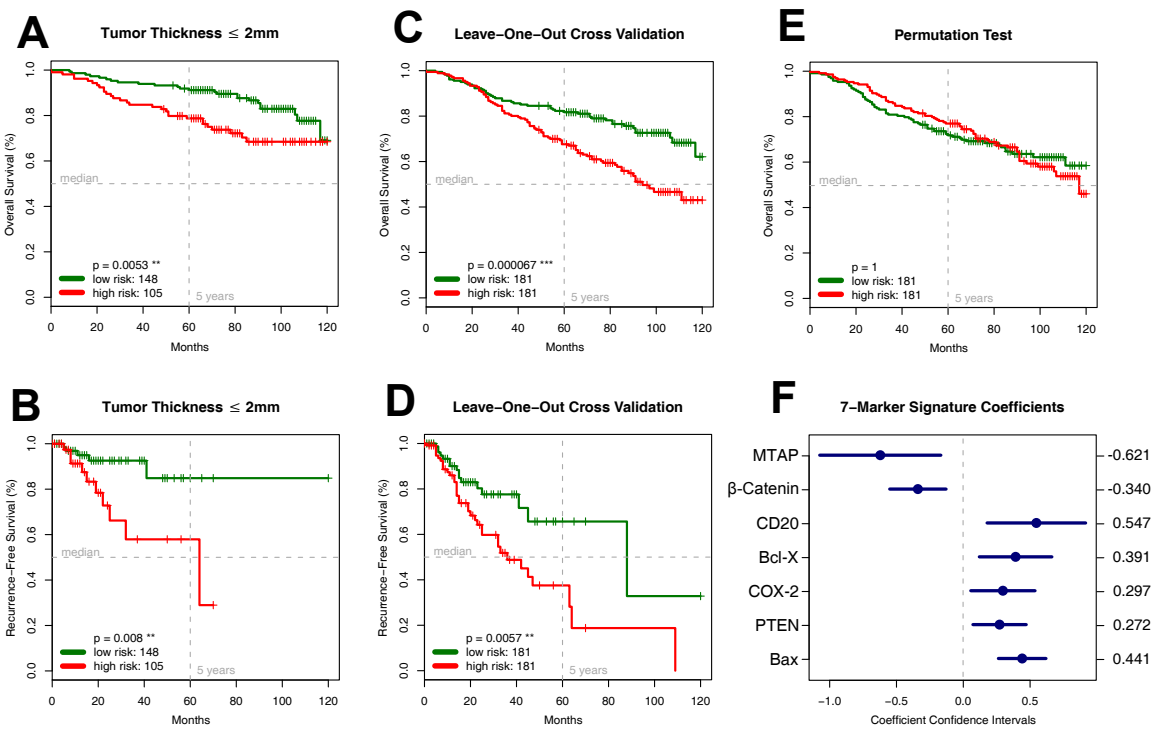


Figure 4.15

Panel A, B: The Seven-Marker Signature and Survival of Patients with a Tumor Thickness $\leq 2.0mm$. Kaplan-Meier estimates show a significantly lower overall ($p = 0.0053$, Panel A) and recurrence-free survival ($p = 0.008$, Panel B) for patients with a comparatively low tumor thickness $\leq 2.0mm$ but high-risk score.
Panel C, D: Leave-One-Out (LOO) Cross-Validation (CV). To investigate the generalization error of the models produced by the FDR signature learning procedure LOO CV was conducted on the primary cohort of 362 MM patients. The resulting risk score could significantly ($p < 0.001$) differentiate between patients with higher or lower overall survival expectancy. The two patient groups also significantly ($p = 0.0057$) differ in recurrence-free survival.
Panel E: Permutation Test. In addition to the CV experiments a permutation test was conducted to assess if the signature learning procedure is over fitting the data set. The resulting signature, which was learned on permuted overall survival data, was not able ($p = 1$) to discriminate between patients with differing survival expectancy. This result indicates that the proposed learning procedure does not over fit the data.
Panel F: Coefficients and Confidence Intervals of the Seven-Marker Signature. The coefficients from the univariate Cox proportional hazard models are used in a weighted linear combination to predict the risk score for each patient. Markers with negative coefficients represent protective markers (MTAP, β-Catenin); those with positive coefficients risk markers (Bax, CD20, Bcl-X, PTEN and COX-2).

Validation of the Seven-Marker Signature on an External Test Cohort

The clinical characteristics of the 225 patients in the external test cohort are listed in Table 4.7. Patients with a high-risk marker signature had a significantly ($p = 0.000017$) different survival expectancy and shorter median overall survival compared to patients with a low-risk signature (95 months versus not reached) (Figure 4.16A). According to multivariate Cox regression including sex, age and tumor thickness, the seven-marker signature was significantly associated with overall survival ($p = 0.0000098$, Table 4.6). Additionally, the recurrence-free survival differed significantly between the two risk groups ($p = 0.004$; Figure 4.16B).

4.5.5 Discussion

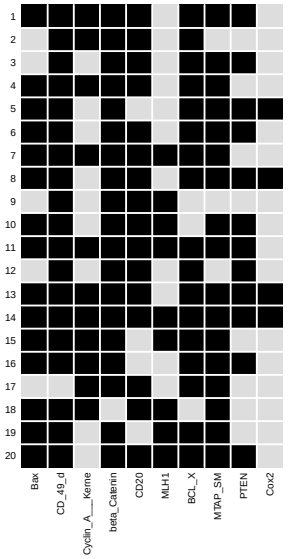


Figure 4.17

Results from an exhaustive subset search of 1023 models. Depicted is the selection matrix for the first 20 highest ranked models based on the LRT P-value. (Markers which are included in a model are indicated with a black squares.)

In this retrospective study including melanoma samples from a primary test cohort of 364 patients, we identified a combination of seven biomarkers representing an independent significant seven-marker signature of prognosis. Notably, the predictive power of the signature was carefully validated and confirmed on a secondary independent external test cohort including melanoma samples of 225 patients from a different hospital. With a total of 24.674 punch specimens of primary MM analyzed by IHC, this TMA study is unparalleled in the literature. The designated seven-marker signature includes two protective markers (β-Catenin, MTAP) and five risk markers (Bax, Bcl-X, CD20, COX-2, PTEN). The defined signature is of prognostic and therapeutic relevance, and provides interesting therapeutic implications. For the practitioner, the assessment of this set of seven IHC markers promises to be a helpful tool to answer the crucial question whom to treat, and how to treat, especially in the adjuvant setting after surgical excision of early-stage and localized primary MM (Stage I to IIa). With CD20, COX-2 and MTAP, three markers of the seven-marker signature offer direct therapeutic implications, since the corresponding drugs have already been approved by the FDA:

The CD20-antigen is known to be an effective therapeutic target in the treatment of patients with CD20-positive B-Cell-Non-Hodgkin-Lymphomas. The monoclonal chimeric antibody Rituximab is indicated for alternative immunotherapy (Avivi et al., 2003). The antibody binds specifically with CD20-antigen presented on the surface of normal and malignant B-lymphocytes and causes a cell- and complement-mediated cytotoxic death of these cells. Considering this effect, the anti-CD20-antibody Rituximab or corresponding radioimmunoconjugates could be used for the treatment of CD20-positive melanoma cells, as well.

Cyclooxygenase 2 may represent another promising therapeutic target. Cyclooxygenases (COXs) catalyze the first rate-limiting step in the conversion of arachidonic acid to prosta-glandins. In contrast to COX-1, the COX-2 isoenzyme is not detectable in most normal tissues but is rapidly induced by various

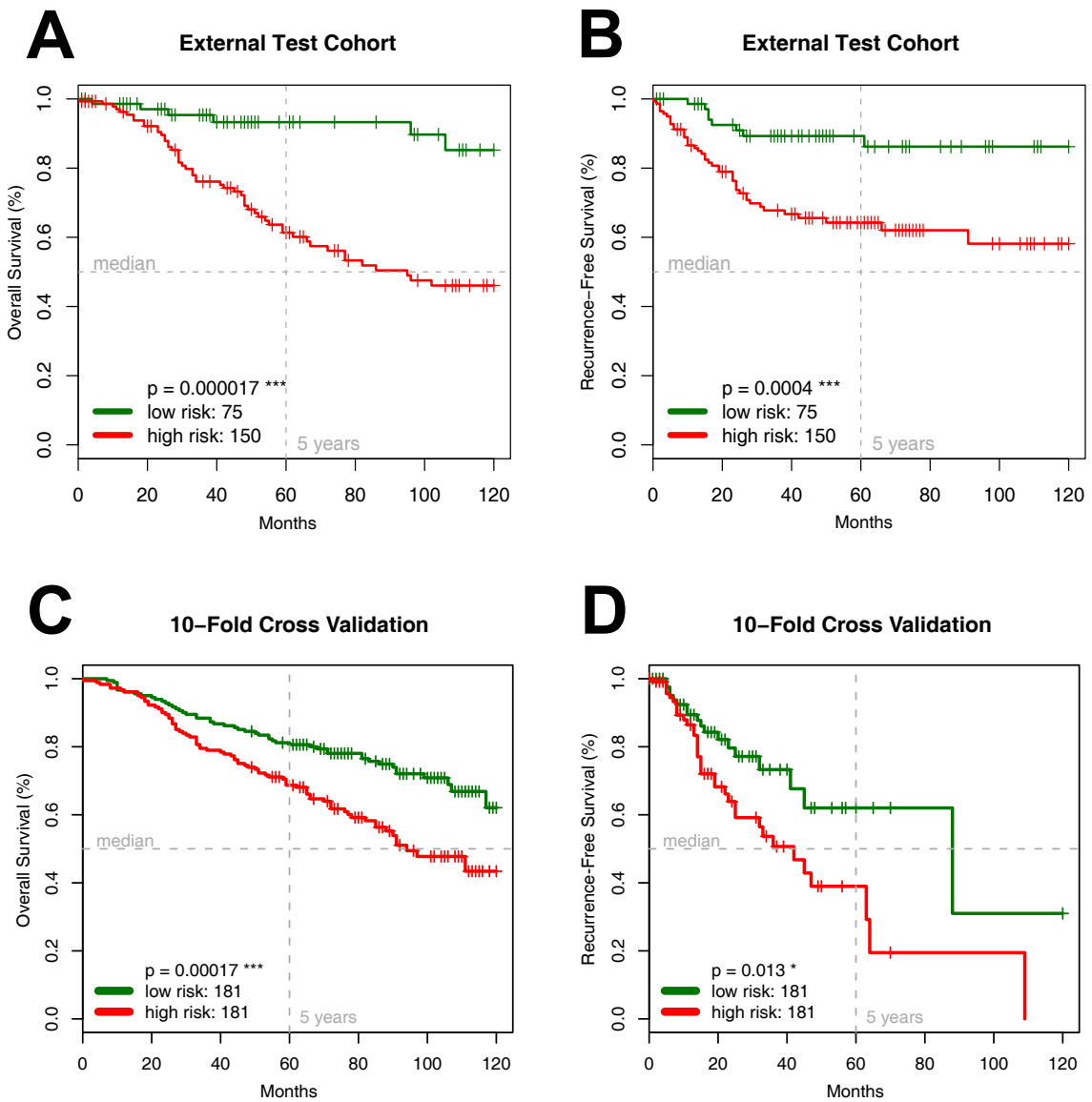


Figure 4.16

Panel A-D: Validation of the Seven-Marker Signature and the FDR Marker Selection Procedure. Kaplan-Meier estimates of overall (Panel A) and recurrence-free survival (Panel B) for the independent external test cohort of 225 patients (TMA 2) confirm the predictive prognostic power of the signature ($p < 0.001$). In addition, the FDR marker selection procedure was tested by a 10-fold cross-validation experiment on the 362 patients of the primary cohort (TMA 1) resulting in still significant estimates for overall survival ($p < 0.001$; Panel C) and recurrence-free survival ($p = 0.013$; Panel D).

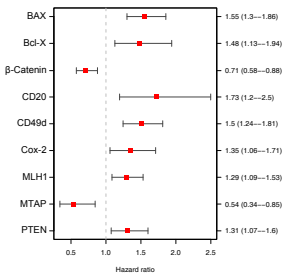


Figure 4.18

Hazard Ratios of the Nine-Marker Signature learned by the FDR selection procedure. Markers with a hazard ratio smaller than 1.00 represent protective markers (MTAP, β-Catenin). Those with hazard ratios larger than 1.00 represent risk markers (Bax, Bcl-X, CD20, CD49d, COX-2, MLH1 and PTEN).

stimuli such as inflammatory reactions (Hla and Neilson, 1992). COX-2 is also expressed in various tumor types and levels of expression have been shown to correlate with invasiveness and prognosis in some tumor entities, including epithelial and melanocytic skin cancer (Meyer et al., 2009; Denkert et al., 2001) suggesting an important role of COX-2 in tumor development and progression. The mechanism by which COX-2 expression accelerates tumorigenesis is poorly understood. Epidemiological studies showed that prolonged COX-2 inhibition through acetylsalicylic acid or other nonsteroidal anti-inflammatory drugs (NSAIDs) might offer some protection against colon cancer and some other malignancies (Thun et al., 2004). So far the benefit of COX-2 inhibitors has not been studied in the adjuvant treatment of early-stage melanomas to prevent metastasis.

In the second-line treatment of advanced metastatic melanoma disease, however, a survival benefit was shown for targeted combined therapy using COX-2 inhibitors and PPARG-agonists for anti-inflammatory treatment together with low-dose metronomic chemotherapy. This observation refers to a randomized multi-institutional phase II trial including 76 mostly chemorefractory patients with progression of metastatic melanoma (stage IV melanoma according to AJCC criteria) (Reichle et al., 2007). In this study, patients that received angiostatically scheduled low-dose metronomic chemotherapy (trofosfamide) in combination with a COX-2 inhibitor (rofecoxib) and a PPARG antagonist (pioglitazone) showed a significantly prolonged progression-free survival compared to the group of patients who received metronomic chemotherapy alone. Accordingly, tumor associated inflammatory and angiogenic processes mediated by COX-2 overexpression are very likely to play a pivotal role in the biology of melanoma progression. Considering this aspect and the observation that melanoma patients with COX-2-positive primary tumors bear a significantly higher risk of tumor recurrence (Meyer et al., 2009), it would be conclusive to introduce COX-2 inhibitors for primary adjuvant treatment of these patients.

Currently, in the adjuvant treatment of MM, interferon alpha is the only clinically accepted therapeutic agent providing a significant benefit regarding recurrence-free survival within a small but distinct percentage of patients (Ascierto and Kirkwood, 2008). As latest meta-analyses of survival data show, the difference between treatment and non-treatment with interferon is only about 3% at 5 years concerning overall survival. The survival benefit of interferon treatment seems to be limited to a small subset of patients.

On account of the serious side effects and the high costs of the therapy, only those patients with a realistic chance to benefit from interferon should receive this treatment. We have recently shown that interferon response may be correlated with the expression of interferon response genes such as MTAP, using cell culture experiments (Zhang and Yu, 2010) and TMA analyses (Wild et al., 2006; Meyer et al., 2010), respectively. According to multivariate Cox regression analysis, MTAP was the strongest independent positive prognostic marker

for recurrence-free and overall survival; patients with MTAP-positive MMs showed a significantly prolonged recurrence-free and overall survival (Meyer et al., 2010). Most interestingly, further subgroup analysis within the group of patients receiving adjuvant interferon therapy revealed a significant survival benefit (recurrence-free survival) in the patients with MTAP-positive MM compared to those with MTAP-negative disease; i.e., individuals with MTAP-negative MM did not benefit from adjuvant IFN treatment. As recently summarized by Ascierto and Kirkwood (2008) there is a compelling rationale for new research upon interferon response, especially in adjuvant settings. Scientific approaches that may enable practitioners to determine which patients may benefit from interferon therapy are essential for a patient-oriented therapy of malignant melanoma and indispensable from the health economic point of view. According to our data there is a clear association between MTAP expression in the primary melanoma and melanoma progression and, even more importantly, response to interferon treatment. Now, after this retrospective investigation prospective clinical trials are necessary to validate the predictive therapeutic relevance of MTAP expression regarding interferon response. This could provide a new basis for a more differentiated application of interferon therapy in melanoma patients in the future.

One additional marker, Bcl-X, has been targeted in preclinical tests and several targeting agents are in the clinical testing phase by now (Azmi and Mohammad, 2009): Bcl-X is related to the anti-apoptotic Bcl-2 protein family. Overexpression of these anti-apoptotic proteins protects cancer cells against death signals of apoptosis. Interestingly, tumors expressing high levels of Bcl-2 or Bcl-X are often found to be resistant to chemotherapeutic agents or radiation therapy (Heere-Ress et al., 2002). In the recent years there has been an exponential growth in the identification and synthesis of non-peptidic cell permeable "small molecule inhibitors" (SMIs) against antiapoptotic proteins like Bcl-2 or Bcl-X. SMIs inhibit distinct protein-protein interactions by blocking specific binding sites of the target molecule, thus supporting the apoptotic machinery (Azmi and Mohammad, 2009). Inhibition of Bcl-X may exert a synergistic effect with conventional treatments like chemo- or radiation therapy. Regarding melanoma therapy, this effect would be a decisive therapeutic success.

According to the data presented here, the seven-marker signature represents a highly promising clinical tool to predict a patient's prognosis. Most importantly, the seven-marker signature might improve the clinical management and adjuvant treatment of early-stage MM with a high risk of recurrence. In the treatment of advanced metastatic melanoma, novel immune-based antitumor therapies targeting signal transduction pathways or tumor immunity barriers by monoclonal antibodies like selective BRAF inhibitors (Hauschild et al., 2009) or anti-cytotoxic T-lymphocyte antigen 4 (CTLA-4) antibodies (Hodi et al., 2010) have already entered clinical studies. This promising therapeutic option in the treatment of advanced metastatic MM development, together with the set of

molecular markers identified in this study may provide new risk-oriented indications for an individualized targeted antitumor therapy of MM.
Additional prospective clinical trials are necessary to validate the prognostic and therapeutic value of this seven-marker signature and its benefit for routine clinical assessment of MM. The detected signature might serve as a prognostic tool enabling physicians to selectively triage, at the time of diagnosis and initial surgery, the subset of high recurrence risk Stage III patients for adjuvant therapy. Selective treatment of those patients that are more likely to develop distant metastatic disease could potentially lower the burden of untreatable metastatic melanoma and revolutionize the therapeutic management of MM.

CHAPTER 5

The Computational Pathology Pipeline: A holistic View

Contents

5.1 Overview

This section describes an genuine computational pathology project, which has been designed following the principles described in the previous chapter of this thesis. It is an ongoing project in kidney cancer research conducted at the University Hospital Zürich and ETH Zürich. Parts of it were published in (Fuchs et al., 2008b) and (Fuchs et al., 2009). Figure 5.1 depicts a schematic overview of the project, subdivided into the three main parts which are discussed in the following.

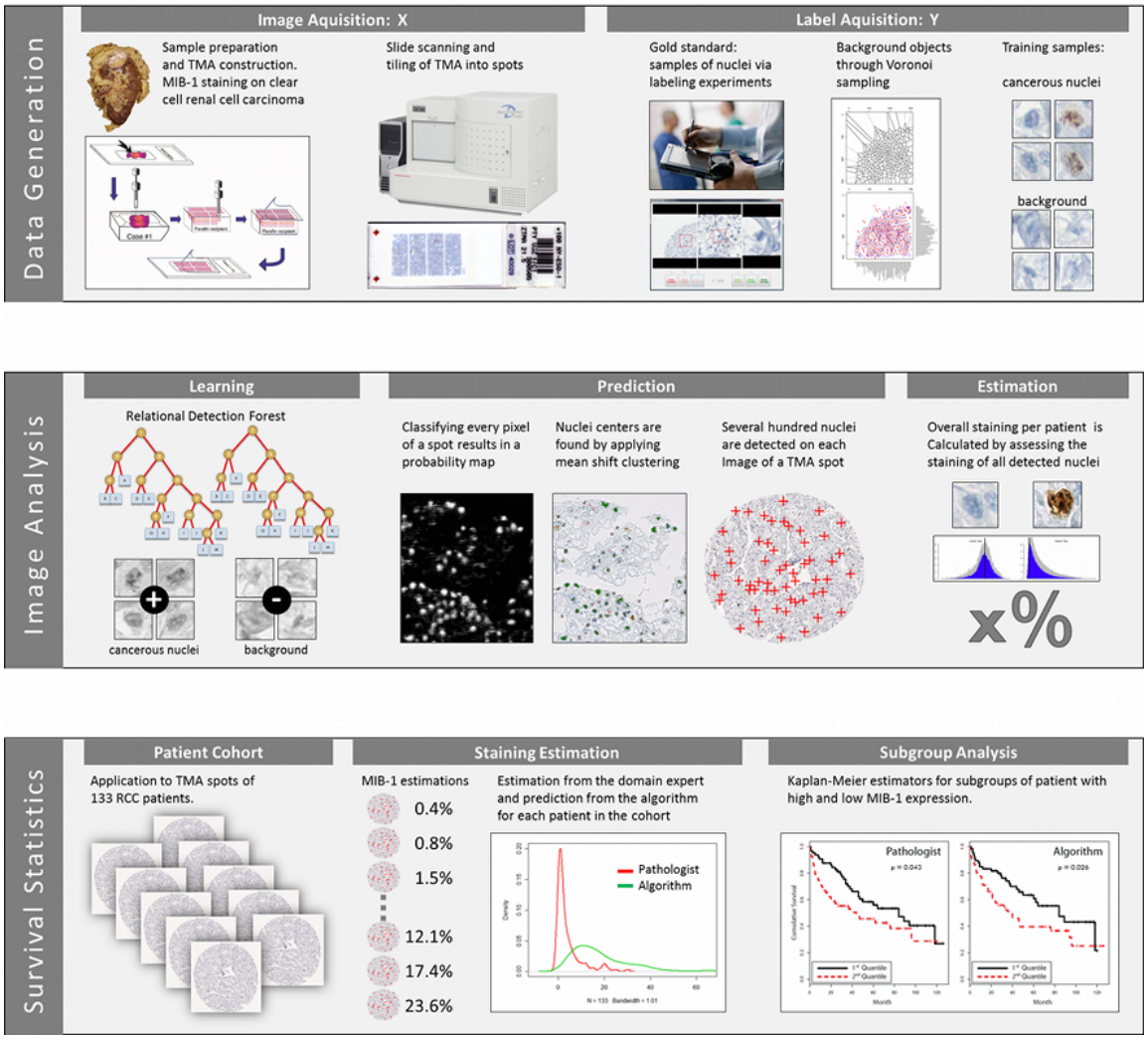


Figure 5.1

A computational pathology framework for investigating the proliferation marker MIB-1 in clear cell renal cell carcinoma. Following the definition in Section 1.7 the framework consists of three parts: (i) The covariate data X existing of images of TMA spots was generated in a trial at the University Hospital Zürich. Extensive labeling experiments were conducted to generate a gold standard comprising cancerous cell nuclei and background samples. (ii) Image analysis consisted of learning a relational detection forest (RDF) and conducting mean shift clustering for nuclei detection. Subsequently, the staining of detected nuclei was determined based on their color histograms. (iii) Using this system, TMA spots of 133 RCC patients were analyzed. Finally, the subgroup of patients with high expression of the proliferation marker was compared to the group with low expression using the Kaplan-Meier estimator.

5.2 Data Generation

The data generation process consists of acquiring images of the TMA spots representing the covariates X in the statistical model and the target variable Y which comprises detection and classification labels for nuclei.
The tissue microarray block was generated in a trial at the University Hospital Zürich. TMA slides were immunohistochemically stained with the MIB-1 (Ki-67) antigen and scanned on a Nanozoomer C9600 virtual slide light microscope scanner from HAMAMATSU. The magnification of $40\times$ resulted in a per pixel resolution of $0.23\mu m$. The tissue microarry was tiled into single spots of size 3000×3000 pixel, representing one patient each.
Various strategies can be devised to estimate the progression status of cancerous tissue: (i) we could first detect cell nuclei and then classify the detected nuclei as cancerous or benign (Schüffler et al., 2010); (ii) the nucleus detection phase could be merged with the malignant/benign classification to simultaneously train a sliding window detector for cancerous nuclei only. To this end samples of cancerous nuclei were collected using the labeling experiments described in Section 2.3. Voronoi Sampling (cf. 3.4) was used to generate a set of negative background patches which are spatially well distributed in the training images. Hence a Voronoi tessellation is created based on the locations of the positive samples and background patches are sampled at the vertices of the Voronoi diagram. In contrast to uniform rejection sampling, using a tessellation has the advantage that the negative samples are concentrated on the area of tissue close to the nuclei and few samples are spent on the homogeneous background. The result of the data generation process is a labeled set of image patches of size 65×65 pixel.

5.3 Image Analysis

The image analysis part of the pipeline consists of learning a relational detection forest (Fuchs et al., 2009) based on the samples extracted in the previous step (cf. Section 3.5.3). To guarantee illumination invariance, the feature basis described in Section 3.3 is used.
The strong class imbalance in the training set is accounted for by randomly subsampling the background class for each tree of the ensemble (cf. Section LearnTree). The model parameters are adjusted by optimizing the out of bag (OOB) error (Breiman, 2001) and they consist of the number of trees, the maximum tree depth and the number of features sampled at each node in a tree.
For prediction each pixel of a TMA spot is classified by the relation detection forest. This results in a probability map in the size of the image where the gray value at each position indicates the probability of being the location of a cancerous nucleus. Finally, weighted mean shift clustering (Comaniciu and Meer, 2002) is conducted with a circular box kernel based on the average radius

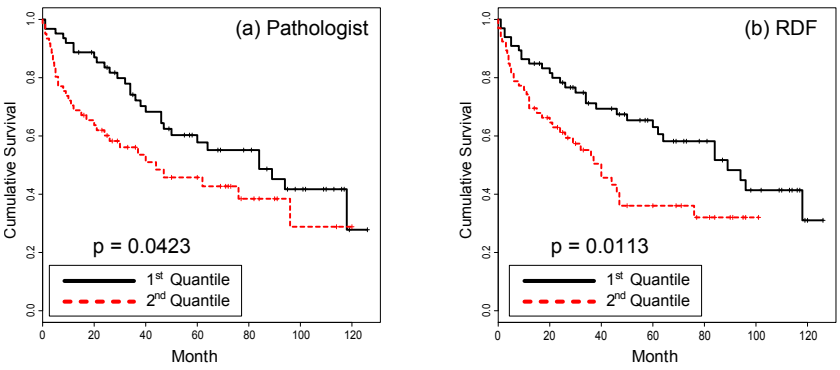


Figure 5.2
Kaplan-Meier estimators show significantly different survival times for renal cell carcinoma patients with high and low proliferating tumors. Compared to the manual estimation from the pathologist (a) ($p = 0.04$), the fully automatic estimation from the algorithm (b) compares favorable ($p = 0.01$) in terms of survival differences (log rank test) for the partitioning of patients into two groups of equal size (Fuchs et al., 2009).

r of the nuclei in the training set. This process yields the final coordinates of the detected cancerous nuclei.

To differentiate a stained nucleus from a non-stained nucleus a simple color model is learned. Based on the labeled nuclei, color histograms are generated for both classes based on the pixels within a radius r. A test nucleus is then classified based on the distance to the centroid histograms of both classes.

The final staining estimation per patient is achieved by calculating the percentage of stained cancerous nuclei.

5.4 Survival Statistics

The only objective endpoint in the majority of TMA studies is the time of survival of each patient. The experiments described in Section 2.3 document the large disagreement between pathologists in the estimation of staining. Hence, fitting an algorithm to the estimates of a single pathologist or to a consensus voting of a committee of pathologist is not desirable.

To this end the proposed computational pathology framework is validated against the right censored clinical survival data of the 133 ccRCC patients. In addition these results were compared to the estimations of an expert pathologist specialized on renal cell carcinoma. He analyzed all spots in an exceptional thorough manner which required him more than two hours. This time consuming annotation exceeds the standard clinical practice significantly by a factor of $10 - 20$ and, therefore, the results can be viewed as an excellent human estimate for this dataset.

Figure 5.2 shows Kaplan-Meier plots of the estimated cumulative survival for the pathologist and the computational pathology framework. The farther the survival estimates of the two groups are separated the better the estimation. Quantifying this difference with a log-rank test shows that the proposed framework performs favorable ($p = 0.0113$) to the trained pathologist ($p = 0.0423$) and it can differentiate between the survival expectancy of the two groups of patients.

5.5 Conclusion

The presented computational pathology framework can be characterized by the following properties:

Simplicity: It can be used off-the-shelf to train object detectors in near real time for large variety of tasks.

Novel Feature Basis: The introduced relational features are able to capture shape information, they are illumination invariant and extremely fast to evaluate.

Randomization: The randomized tree induction algorithm is able to handle an immensely large feature space and to take advantage of it by increasing the diversity in the ensemble.

Real World Applicability: The proposed algorithms perform well not only on renal cancer tissue but also in fluorescent imaging of pancreatic islets (cf. Section 3.10, Floros et al. (2009)) and in quantifying staining in murine samples (cf. Section 3.11, (Bettermann et al., 2010)).

5.6 Criticism and Limitations

Computational Pathology as a field is still in its infancy. It is heavily depending on staining and scanning technology which are advancing rapidly. This progress in biochemistry and digitalization hardware is constantly challenging researchers in this field but also provides numerous possibilities for advancement in computer vision and machine learning.
A limiting factor in practice is still the enormous amount of imaging data produced in histopathology. Even extremely well equipped hospitals are not able to store every microscopic slide in digital form, which would mount up to petabytes of data. For the analysis this data volume poses also numerous problems regarding parallelization of algorithms and the use of computer clusters. Although solutions exist for implementing random forests on GPUs (Sharp, 2008) this research direction was not explored for relational detection forests in this thesis. Nevertheless, the enormous amount of weakly labeled data and the medical significance of the problems provide exactly the kind of playground machine learning researchers are longing for.
From a technical perspective it was demonstrated in Section LearnTree that randomized tree ensembles also have their limitations. The experiments on the Magellan data showed that it was not possible to improve over a well tuned SVM for this specific problem. This failure serves as a reminder that machine learning researchers constantly have to compare their algorithms to competing methods. However, it has to be noted that a SVM could not handle the proposed relational feature basis (cf. Section 3.3) and it is also unclear how a SVM

could be integrated in the inter-active learning framework presented in Section 3.6.
Finally, from a software engineering view it would be desirable to implement the system in form of an AJAX based web application to provide simple access for pathologists worldwide. This approach would also provide constant inflow of new labeling information which could be utilized to constantly update and improve the models presented in this thesis.

5.7 Future Directions

5.7.1 Histopathological Imaging

One promising research direction in medical image analysis points to online learning and interactive learning of computer vision models. Not only covers histopathology a broad and heterogeneous field but new biomarkers, antibodies and stainings are developed on a daily basis. To this end, real world applications have to quickly adapt to changing tissue types and staining modalities. Domain experts should be able to train these models in an interactive fashion to accustom novel data. For example, a classifier for object detection can be trained by clicking on novel objects or correcting for false detections.
A necessary prerequisite for research in computational pathology proved to be the scanning of whole slides and TMAs. (Huisman et al., 2010) describe a fully digital pathology slide archive which has been assembled by high-volume tissue slide scanning. The Peta bytes of histological data which will be available in the near future pose also a number of software engineering challenges, including distributed processing of whole slides and TMAs in clusters or the cloud, multiprocessor and multicore implementations of pattern analysis algorithms and facilitating real time image processing on GPUs.

5.7.2 Clinical Application and Decision Support

In today's patient care we observe the interesting trend to integrate pathological diagnoses in web based patient files. Avatar based visualization proved to be useful not only for medical experts but also for a new generation of patients who are better informed and demand online updated and appropriately visualized informations about their own disease state and treatment procedures.
Furthermore this approach can be extended for decision support by statistical models which are able to utilize this unified view of patients incorporating data from a large variety of clinical sources, e.g. pathology, cytology, radiology, etc.

5.7.3 Pathology@home

Real-time, in vivo cancer detection on cellular level appears as a futuristic dream in patient care but could be a reality in a few years. (Shin et al., 2010) con-

structed a fiber-optic fluorescence microscope using a consumer-grade camera for in vivo cellular imaging. The fiber-optic fluorescence microscope includes an LED light, an objective lens, a fiber-optic bundle, and a consumer-grade DSLR. The system was used to image an oral cancer cell line, a human tissue specimen and the oral mucosa of a healthy human subject in vivo, following topical application of 0.01% proflavine. The fiber-optic microscope resolved individual nuclei in all specimens and tissues imaged. This capability enabled qualitative and quantitative differences between normal and precancerous or cancerous tissues to be identified. In combination with a computational pathology framework, this technique would allow the real time classification of cancerous cells in epithelial tissues. Such a portable and inexpensive system is especially interesting for patient care in low-resource settings like the developing world.
Constructing a microscope for mobile phones defines the future of patient care in remote sites with centralized analysis support. (Breslauer et al., 2009) built a mobile phone-mounted light microscope and demonstrated its potential for clinical use by imaging sickle and P. falciparum-infected red blood cells in bright-field and M. tuberculosis-infected sputum samples in fluorescence with LED excitation. In all cases the resolution exceeded the critical level that is necessary to detect blood cell and microorganism morphology. This concept could provide an interesting tool for disease diagnosis and screening, especially in the developing world and rural areas where laboratory facilities are scarce but mobile phone infrastructure is available.

5.7.4 Standards and Exchange Formats

One of the major obstacles for wide spread use of computational pathology remains the absence of generally agreed upon standards and exchange formats. This deficit not only handicaps slide processing management and whole slide digital imaging (Daniel et al., 2009), but it also extends to statistical models and analysis software. Standardized exchange formats would support project specific combinations of object detectors, staining estimation algorithms and medical statistics. It would be very highly desirable if at least the research community would agree on a few simple interfaces for data and model exchange.

Appendix

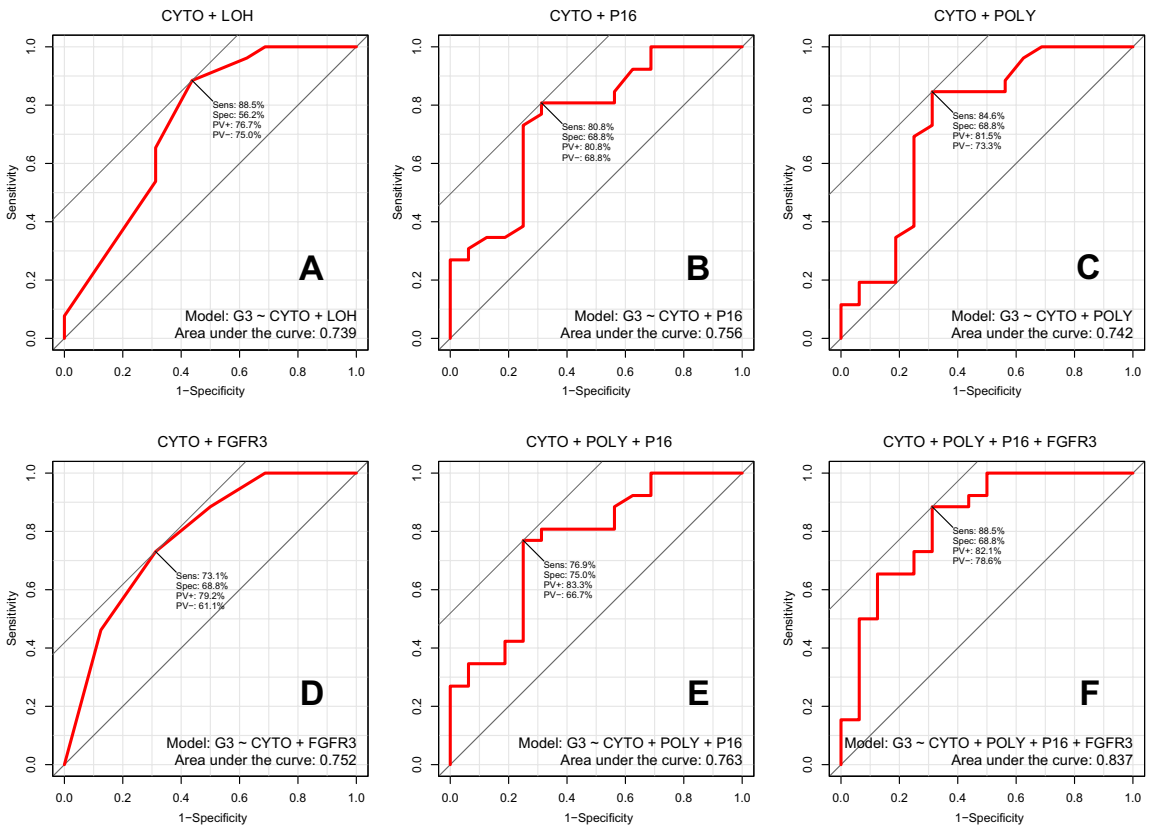


Figure 5.3

ROC curves of the performance of the various assays for the detection of high-grade versus non-high-grade tumors.

Table 5.1
Patient and tumor characteristics and results of molecular and immunohistochemical analyses.

Variable	Categorisation	TMA Study n analyzable	%
Clinico-pathologic data:			
Age at diagnosis			
	<70 years	141	55.3
	≥70 years	114	44.7
Sex			
	female	64	25.1
	male	191	74.9
Tumour stage			
	pTa	146	57.3
	pT1	48	18.8
	pT2	56	22.0
	pT3	2	0.8
	pT4	3	1.2
Histologic grade			
	low grade	150	58.8
	high grade	105	41.2
Adjacent carcinoma in situ			
	no	222	87.1
	yes	33	12.9
Multifocality			
	solitary	53	20.8
	multifocal	202	79.2
Growth pattern			
	papillary	207	81.5
	solid	47	18.5
Molecular data:			
FGFR3 gene			
	wild-type	110	52.9
	mutation	98	47.1
Immunohistochemistry (IHC):			
MIB1 IHC			
	≤25%	168	71.2
	>25%	68	28.8
TP53 IHC			
	≤10%	179	73.1
	>10%	66	26.9
CK20 IHC			
	superficial staining pattern	49	20.3
	negative or >10%	192	79.7
Fluorescence in situ hybridization (FISH):			
Relative *p16* deletion (9p21)			
	≤14%	128	54.2
	>14%	108	45.8
Polysomy			
	≤18%	71	30.1
	>18%	165	69.9

Table 5.2
Characteristics of voided urine samples and corresponding biopsies.

Variable	Categorisation	Voided urine study n	%
Cystoscopically obtained biopsies:			
Histologic diagnosis			
	normal urothelium	27	22.7
	inflammatory urothelium	11	9.2
	pTa	43	36.1
	pT1	18	15.1
	pT2	11	9.2
	pTis	9	7.6
Histologic grade			
	low grade	41	50.6
	high grade	40	49.4
Voided urine:			
Cytologic diagnosis			
	Negative cytology [N]	40	33.6
	Atypical cytology, not specified [NR]	25	21.0
	Atypical cytology, suspicious [S]	24	20.2
	Positive cytology [P]	30	25.2
FGFR3 analysis			
	wild-type	84	70.6
	mutation	18	15.1
	unknown	17	14.3
Loss of heterozygosity (LOH)			
	no LOH	57	47.9
	suspected LOH	7	5.9
	LOH	55	46.2
UroVysion Fluorescence in situ Hybridization			
	no polysomy or loss of p16	15	12.6
	polysomy or loss of p16	35	29.4
	unknown	69	58.0

Variable	Categorisation	*FGFR3* gene			MIB1 expression			TP53 expression			CK20 IHC staining pattern		
		wild type	mutated	p*	≤25%	>25%	p*	≤10%	>10%	p*	superficial	negative or >10%	p*
Tumour stage													
	pTa	31	85	**<0.001**	127	9	**<0.001**	130	10	**<0.001**	48	90	**<0.001**
	pT1	29	9		24	19		20	26		1	44	
	pT2	46	3		16	36		26	28		0	53	
	pT3	2	0		1	1		0	2		0	2	
	pT4	2	1		0	3		3	0		0	3	
Histologic grade													
	low grade	31	87	**<0.001**	132	8	**<0.001**	131	12	**<0.001**	48	94	**<0.001**
	high grade	79	11		36	16		48	54		1	98	
Adjacent carcinoma in situ													
	no	81	95	**<0.001**	155	51	**0.001**	165	48	**<0.001**	48	164	**0.013**
	yes	29	3		13	17		14	18		1	28	
Multifocality													
	solitary	25	14	0.2	30	18	1.0	33	19	1.0	10	40	1.0
	multifocal	85	84		138	50		146	47		39	152	
Growth pattern													
	papillary	71	96	**<0.001**	156	36	**<0.001**	159	40	**<0.001**	49	147	**<0.001**
	solid	39	2		12	31		20	25		0	44	

Table 5.3

Comparison of molecular and immunohistochemical data with pathologic characteristics. (* Fisher's exact test (2-sided); bold face representing p-values < 0.05.)

Table 5.4

Analysis of voided urine samples, separate results for patients with high grade ($n = 40$) and low grade ($n = 41$) bladder tumors.

Variable	Categorisation	low grade tumours		high grade tumours	
		n	valid %	n	valid %
Voided urine samples:					
Cytologic diagnosis					
	Negative cytology [N]	8	19.5	1	2.5
	Atypical cytology, not specified [NR]	17	41.5	2	5.0
	Atypical cytology, suspicious [S]	10	24.4	13	32.5
	Positive cytology [P]	6	14.6	24	60.0
	missing	0		0	
Loss of heterozygosity (LOH)					
	no LOH	17	41.5	7	17.5
	suspected LOH	2	4.9	4	10.0
	LOH	22	53.7	29	72.5
	missing	0		0	
FGFR3 gene					
	wild-type	25	71.4	26	78.8
	mutated	10	28.6	7	21.2
	missing	6		7	
UroVysion Fluorescence in situ Hybridization					
	no polysomy or loss of p16	10	52.6	5	16.7
	polysomy or loss of p16	9	47.4	25	83.3
	missing	22		10	

	High Grade		Stage ≥pT1	
	Estimate	**P**	**Estimate**	**P**
Intercept	-0.5319	0.116	-0.3623	0.264
polysomy	4.0808	**< 0.001**	3.1041	**< 0.001**
p16	0.6360	0.359	0.6877	0.286
FGFR3	-2.9857	**< 0.001**	-2.5355	**< 0.001**

Table 5.5

Parameters and their P-values of the FISH+FGFR3 logistic regression model. The model contains polysomy, p16 and FGFR3 as predictors from which polysomy and FGFR3 are significant for the prediction of high grade as well as infiltrative tumor growth (stage $\geq pT1$).

	Primary cohort		Ext. Test cohort	
TMA characteristics	**N**	**%**	**N**	**%**
Origin	Regensburg		Hamburg	
Patients				
No. of patients	364		235	
No. follow-up	364	100.0	231	98.3
No. of patients with at least 1 signature marker	362	99.5	225	95.7
TMA Spots				
No. of biomarkers	70		7	
Valid spots	23106	90.7	1541	93.7
Missing spots	2374	9.3	104	6.3
Clinicopathological characteristics	**N**	**%**	**N**	**%**
Age				
<=60	180	49.5	139	59.1
>60	184	50.5	92	39.1
unknown			4	1.7
Sex				
Male	195	53.6	126	53.6
Female	169	46.4	105	44.7
unknown			4	1.7
Tumor thickness				
<= 1mm	163	44.8	110	46.8
1.01--2mm	92	25.3	47	20.0
2.01--4mm	61	16.8	36	15.3
> 4mm	42	11.5	36	15.3
unknown	6	1.6	6	2.6
Clark level				
1	2	0.5	1	0.4
2	75	20.6	39	16.6
3	106	29.1	80	34.0
4	149	40.9	90	38.3
5	14	3.8	19	8.1
unknown	18	4.9	6	2.6

Table 5.6

Clinical Data: Characterization and Comparison of the Primary Cohort (TMA 1) and the External Test Cohort (TMA 2). Reported are the number of counts and the associated percentages for all specimens on the tissue microarrays. Missing values are listed as "unknown".

Table 5.7

Biomarker: Characterization and Comparison of the Primary Cohort (TMA 1) and the External Test Cohort (TMA 2). Reported are the number of counts and the associated percentages for all specimens on the tissue microarrays. CD49d and MLH1 are not contained in the final seven-marker signature and therefore were not analyzed on the external test TMA 2. Missing values are listed as "unknown".

Immunohistochemical data	N	%	N	%
Bax				
0	5	1.4	5	2.1
1	60	16.5	49	20.9
2	95	26.1	81	34.5
3	88	24.2	56	23.8
4	88	24.2	29	12.3
unknown	28	7.7	15	6.4
b-Catenin				
0	10	2.7	4	1.7
1	121	33.2	43	18.3
2	106	29.1	108	46.0
3	71	19.5	53	22.6
4	17	4.7	11	4.7
unknown	39	10.7	16	6.8
CD20				
0	333	91.5	154	65.5
1	12	3.3	48	20.4
2	4	1.1	16	6.8
3	1	0.3	1	0.4
unknown	14	3.8	16	6.8
BCL-X				
0	151	41.5	66	28.1
1	167	45.9	117	49.8
2	26	7.1	38	16.2
3	1	0.3	4	1.7
unknown	19	5.2	10	4.3
MTAP				
0	56	15.4	103	43.8
1	245	67.3	101	43.0
2			16	6.8
unknown	63	17.3	15	6.4
PTEN				
0	57	15.7	23	9.8
1	140	38.5	87	37.0
2	116	31.9	76	32.3
3	28	7.7	25	10.6
4	4	1.1	8	3.4
unknown	19	5.2	16	6.8
Cox-2				
0	121	33.2	20	8.5
1	188	51.6	103	43.8
2	39	10.7	73	31.1
3	5	1.4	21	8.9
4			2	0.9
unknown	11	3.0	16	6.8
CD49d				
0	63	17.3	n.a	
1	137	37.6		
2	78	21.4		
3	33	9.1		
4	3	0.8		
unknown	50	13.7		
MLH1				
0	65	17.9	n.a	
1	130	35.7		
2	99	27.2		
3	37	10.2		
4	13	3.6		
unknown	20	5.5		

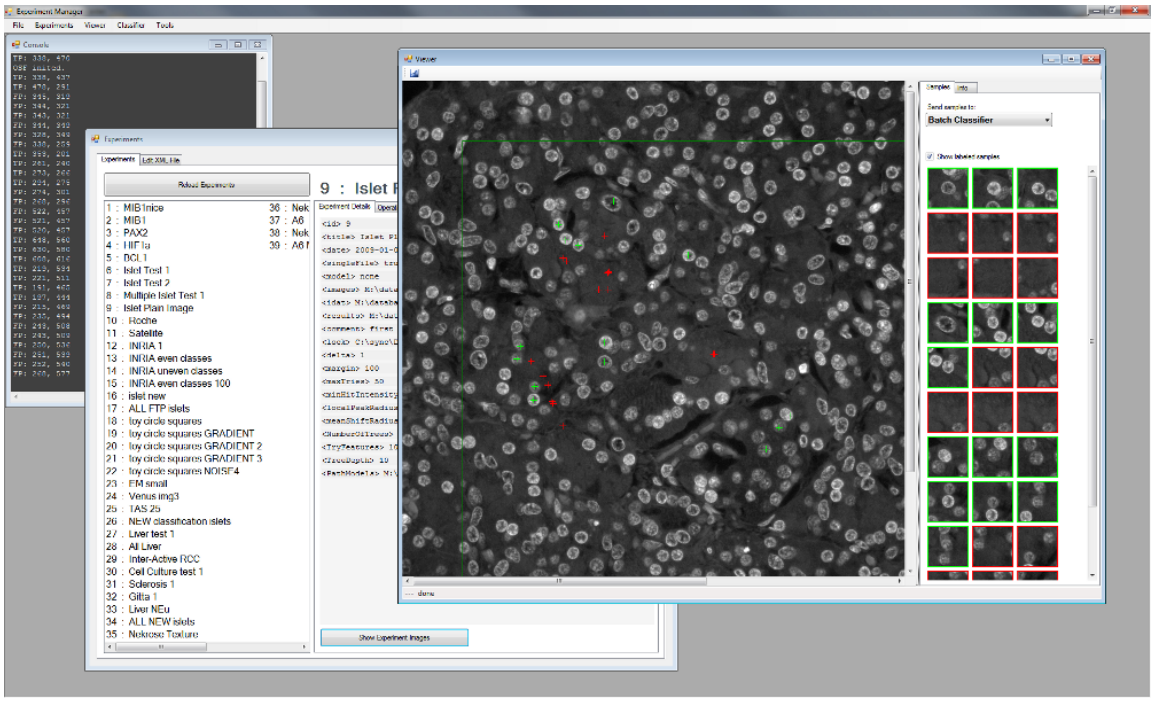


Figure 5.4

Screenshot of the "Experiment Manager" application for labeling and multiple object detection. It comprises inter-active and batch learning implementations of relational detection forests. Additionally, large scale analyses can be conducted in parallel utilizing the ETH computer cluster.

Bibliography

Agarwala, S. S. 2009. Current systemic therapy for metastatic melanoma. Expert Rev Anticancer Ther **9**:587–595.

Ahonen, T., A. Hadid, and M. Pietikainen. 2004. Face Recognition with Local Binary Patterns. ECCV 2004 **2**:469–481.

Alonso, S. R., P. Ortiz, M. Pollan, et al. 2004. Progression in cutaneous malignant melanoma is associated with distinct expression profiles: a tissue microarray-based study. Am. J. Pathol. **164**:193–203.

Amit, Y., and D. Geman. 1997. Shape quantization and recognition with randomized trees. Neural Computation **9**:1545–1588.

Ando, T., S. Imoto, and S. Miyano. 2004. Kernel Mixture Survival Models for Identifying Cancer Subtypes, Predicting Patient's Cancer Types and Survival Probabilities.

Arif, M., and N. Rajpoot. 2007. Classification of potential nuclei in prostate histology images using shape manifold learning. In: Machine Vision, 2007. ICMV 2007. International Conference on. 113 –118.

Ascierto, P. A., and J. M. Kirkwood. 2008. Adjuvant therapy of melanoma with interferon: lessons of the past decade. J Transl Med **6**:62.

Avivi, I., S. Robinson, and A. Goldstone. 2003. Clinical use of rituximab in haematological malignancies. Br. J. Cancer **89**:1389–1394.

Azmi, A. S., and R. M. Mohammad. 2009. Non-peptidic small molecule inhibitors against Bcl-2 for cancer therapy. J. Cell. Physiol. **218**:13–21.

Balch, C. M., A. C. Buzaid, S. J. Soong, et al. 2001a. Final version of the American Joint Committee on Cancer staging system for cutaneous melanoma. J. Clin. Oncol. **19**:3635–3648.

Balch, C. M., J. E. Gershenwald, S. J. Soong, et al. 2009. Final version of 2009 AJCC melanoma staging and classification. J. Clin. Oncol. **27**:6199–6206.

Balch, C. M., S.-J. Soong, J. E. Gershenwald, et al. 2001b. Prognostic Factors Analysis of 17,600 Melanoma Patients: Validation of the American Joint Committee on Cancer Melanoma Staging System. J Clin Oncol **19**:3622–3634.

Begelman, G., M. Zibulevsky, E. Rivlin, and T. Kolatt. 2009. Blind Decomposition of Transmission Light Microscopic Hyperspectral Cube Using Sparse Representation. Medical Imaging, IEEE Transactions on **28**:1317 –1324.

Behrmann, I., S. Wallner, W. Komyod, P. C. Heinrich, M. Schuierer, R. Buettner, and A. K. Bosserhoff. 2003. Characterization of methylthioadenosin phosphorylase (MTAP) expression in malignant melanoma. Am. J. Pathol. **163**:683–690.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B Methodological **57**:289–300.

Berglund, L., E. Bjrling, P. Oksvold, et al. 2008. A Genecentric Human Protein Atlas for Expression Profiles Based on Antibodies. Molecular & Cellular Proteomics **7**:2019–2027.

Bettermann, K., M. Vucur, J. Haybaeck, et al. 2010. TAK1 Suppresses a NEMO-Dependent but NF-[kappa]B-Independent Pathway to Liver Cancer. Cancer Cell **17**:481 – 496.

Billerey, C., D. Chopin, M. H. Aubriot-Lorton, et al. 2001. Frequent FGFR3 mutations in papillary non-invasive bladder (pTa) tumors. Am. J. Pathol. **158**:1955–1959.

Blum, A. 1996. On-line Algorithms in Machine Learning. In: Online Algorithms. 306–325.

Bonner-Weir, S., and T. D. O'Brien. 2008. Islets in type 2 diabetes: in honor of Dr. Robert C. Turner. Diabetes **57**:2899–2904.

Boucheron, L., Z. Bi, N. Harvey, B. Manjunath, and D. Rimm. 2007. Utility of multispectral imaging for nuclear classification of routine clinical histopathology imagery. BMC Cell Biology **8**:S8.

Boykov, Y., and V. Kolmogorov. 2004. An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. IEEE Trans. Pattern Anal. Mach. Intell. **26**:1124–1137.

Breiman, L. 1996. Bagging Predictors. Machine Learning **24**:123–140.

———. 2001. Random Forests. Machine Learning **45**:5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. Classification and Regression Trees. Chapman & Hall, New York, NY.

Breslauer, D. N., R. N. Maamari, N. A. Switz, W. A. Lam, and D. A. Fletcher. 2009. Mobile Phone Based Clinical Microscopy for Global Health Applications. PLoS ONE **4**:e6320.

Brown, F. M. 2000. Urine cytology. It is still the gold standard for screening? Urol. Clin. North Am. **27**:25–37.

Bubendorf, L., B. Grilli, G. Sauter, M. J. Mihatsch, T. C. Gasser, and P. Dalquen. 2001. Multiprobe FISH for enhanced detection of bladder cancer in voided urine specimens and bladder washings. Am. J. Clin. Pathol. **116**:79–86.

Buhlmann, P., and B. Yu. 2003. Boosting With the L2 Loss: Regression and Classification. Journal of the American Statistical Association **98**:324–339.

Burger, M., S. Denzinger, W. F. Wieland, C. G. Stief, A. Hartmann, and D. Zaak. 2008. Does the current World Health Organization classification predict the outcome better in patients with noninvasive bladder cancer of early or regular onset? BJU Int. **102**:194–197.

Burl, M. C., L. Asker, P. Smyth, U. Fayyad, P. Perona, L. Crumpler, and J. Aubele. 1998. Learning to Recognize Volcanoes on Venus. Mach. Learn. **30**:165–194.

Canny, J. 1986. A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. **8**:679–698.

Carson, K. F., D. R. Wen, P. X. Li, A. M. Lana, C. Bailly, D. L. Morton, and A. J. Cochran. 1996. Nodal nevi and cutaneous melanomas. Am. J. Surg. Pathol. **20**:834–840.

Celik, H., A. Hanjalic, E. Hendriks, and S. Boughorbel. 2008. Online training of object detectors from unlabeled surveillance video. In: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on. 1–7.

Chan, T. F., and L. A. Vese. 2001. Active contours without edges. Image Processing, IEEE Transactions on **10**:266–277.

Coelho, L. P., A. Shariff, and R. F. Murphy. 2009. Nuclear segmentation in microscope cell images: a hand-segmented dataset and comparison of algorithms. In: ISBI'09: Proceedings of the Sixth IEEE international conference on Symposium on Biomedical Imaging. Piscataway, NJ, USA: IEEE Press, 518–521.

Comaniciu, D., and P. Meer. 2002. Mean shift: a robust approach toward feature space analysis. Pattern Analysis and Machine Intelligence, IEEE Transactions on **24**:603 –619.

Cordon-Cardo, C., Z. F. Zhang, G. Dalbagni, M. Drobnjak, E. Charytonowicz, S. X. Hu, H. J. Xu, V. E. Reuter, and W. F. Benedict. 1997. Cooperative effects of p53 and pRB alterations in primary superficial bladder tumors. Cancer Res. **57**:1217–1221.

Cote, R. J., M. D. Dunn, S. J. Chatterjee, et al. 1998. Elevated and absent pRb expression is associated with bladder cancer progression and has cooperative effects with p53. Cancer Res. **58**:1090–1094.

Cox, D. R. 1972. Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological) **34**:187–220.

Cukierski, W. J., X. Qi, and D. J. Foran. 2009. Moving beyond color: the case for multispectral imaging in brightfield pathology. In: ISBI'09: Proceedings of the Sixth IEEE international conference on Symposium on Biomedical Imaging. Piscataway, NJ, USA: IEEE Press, 1111–1114.

Dahinden, C. 2006. Classification with Tree-Based Ensembles Applied to the WCCI 2006 Performance Prediction Challenge Datasets. In: Neural Networks, 2006. IJCNN '06. International Joint Conference on. 1669 –1672.

Dahinden, C., B. Ingold, P. Wild, et al. 2010. Mining Tissue Microarray Data to Uncover Combinations of Biomarker Expression Patterns that Improve Intermediate Staging and Grading of Clear Cell Renal Cell Cancer. Clinical Cancer Research **16**:88–98.

Dalal, N., and B. Triggs. 2005. Histograms of Oriented Gradients for Human Detection. In: CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1. Washington, DC, USA: IEEE Computer Society, 886–893.

Daniel, C., M. Garca Rojo, K. Bourquard, D. Henin, T. Schrader, V. D. Mea, J. Gilbertson, and B. A. Beckwith. 2009. Standards to Support Information Systems Integration in Anatomic Pathology. Archives of Pathology & Laboratory Medicine **133**:1841–1849.

DeCoste, D., and B. Schölkopf. 2002. Training Invariant Support Vector Machines. Machine Learning **46**:161–190.

Dekel, O., and O. Shamir. 2009. Vox Populi: Collecting High-Quality Labels from a Crowd. In: Proceedings of the 22nd Annual Conference on Learning Theory (COLT 2009).

Delmas, V., F. Beermann, S. Martinozzi, et al. 2007. Beta-catenin induces immortalization of melanocytes by suppressing p16INK4a expression and cooperates with N-Ras in melanoma development. Genes Dev. **21**:2923–2935.

Denkert, C., M. Kobel, S. Berger, A. Siegert, A. Leclere, U. Trefzer, and S. Hauptmann. 2001. Expression of cyclooxygenase 2 in human malignant melanoma. Cancer Res. **61**:303–308.

Dietterich, T. G. 2000. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Machine Learning **40**:139–157.

Downs, A. 1961. Problems of Majority Voting: In Defense of Majority Voting. The Journal of Political Economy **69**:192–199.

Drelie Gelasca, E., B. Obara, D. Fedorov, K. Kvilekval, and B. Manjunath. 2009. A biosegmentation benchmark for evaluation of bioimage analysis methods. BMC Bioinformatics **10**:368.

Eble, J. N., G. Sauter, J. I. Epstein, and I. A. Sesterhenn. 2004. Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs, volume 7 of *Classification of Tumours*. World Health Organization.

Elgawi, O. H. 2008. Online random forests based on CorrFS and CorrBE. In: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on. 1–7.

Fern, A., and R. Givan. 2003. Online Ensemble Learning: An Empirical Study. Machine Learning **53**:71–109.

Floros, X. E., T. J. Fuchs, M. P. Rechsteiner, G. Spinas, H. Moch, and J. M. Buhmann. 2009. Graph-Based Pancreatic Islet Segmentation for Early Type 2 Diabetes Mellitus on Histopathological Tissue. In: MICCAI (1). 633–640.

Foran, D. J., L. Yang, O. Tuzel, W. Chen, J. Hu, T. M. Kurc, R. Ferreira, and J. H. Saltz. 2009. A caGRID-enabled, learning based image segmentation method for histopathology specimens. In: ISBI'09: Proceedings of the Sixth IEEE international conference on Symposium on Biomedical Imaging. Piscataway, NJ, USA: IEEE Press, 1306–1309.

Ford, L. R., and D. R. Fulkerson. 1962. Flows in networks, by L.R. Ford, Jr. [and] D.R. Fulkerson. Princeton University Press, Princeton, N.J.,.

Frank, A., and A. Asuncion. 2010. UCI Machine Learning Repository.

Freund, Y., and R. Schapire. 1996. Experiments with a new boosting algorithm. Machine Learning: Proceedings of the Thirteenth International Conference :148–156.

Freund, Y., and R. E. Schapire. 1995. A decision-theoretic generalization of on-line learning and an application to boosting. In: EuroCOLT '95: Proceedings of the Second European Conference on Computational Learning Theory. London, UK: Springer-Verlag, 23–37.

Frigerio, S., B. C. Padberg, R. T. Strebel, D. M. Lenggenhager, A. Messthaler, M. T. Abdou, H. Moch, and D. R. Zimmermann. 2007. Improved detection of bladder carcinoma cells in voided urine by standardized microsatellite analysis. Int. J. Cancer **121**:329–338.

Fuchs, T. J., and J. M. Buhmann. 2009. Inter-Active Learning of Randomized Tree Ensembles for Object Detection. In: ICCV Workshop on On-line Learning for Computer Vision, 2009. IEEE.

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. In: Bebis, G., R. D. Boyle, B. Parvin, et al., editors, ISVC (1), volume 5875 of *Lecture Notes in Computer Science*. Springer, 367–378.

Fuchs, T. J., T. Lange, P. J. Wild, H. Moch, and J. M. Buhmann. 2008a. Weakly Supervised Cell Nuclei Detection and Segmentation on Tissue Microarrays of Renal Cell Carcinoma. In: Pattern Recognition. DAGM 2008, volume 5096 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 173–182.

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. In: Medical Image Computing and Computer-Assisted Intervention. MICCAI 2008, volume 5242 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 1–8.

Garcia, C., G. Zikos, and G. Tziritas. 2000. Wavelet packet analysis for face recognition. Image and Vision Computing **18**:289 – 297.

Geman, D., C. d'Avignon, D. Q. Naiman, and R. L. Winslow. 2004. Classifying Gene Expression Profiles from Pairwise mRNA Comparisons. Statistical Applications in Genetics and Molecular Biology **3**:19.

Geurts, P., D. Ernst, and L. Wehenkel. 2006. Extremely randomized trees. Machine Learning **63**:3–42.

Glotsos, D., P. Spyridonos, D. Cavouras, P. Ravazoula, P. A. Dadioti, and G. Nikiforidis. 2005. An image-analysis system based on support vector machines for automatic grade diagnosis of brain-tumour astrocytomas in clinical routine. Medical Informatics and the Internet in Medicine **30**:179–193(15).

Gould Rothberg, B. E., M. B. Bracken, and D. L. Rimm. 2009. Tissue biomarkers for prognosis in cutaneous melanoma: a systematic review and meta-analysis. J. Natl. Cancer Inst. **101**:452–474.

Gould Rothberg, B. E., and D. L. Rimm. 2010. Biomarkers: the useful and the not so useful–an assessment of molecular prognostic markers for cutaneous melanoma. J. Invest. Dermatol. **130**:1971–1987.

Grabner, H., C. Leistner, and H. Bischof. 2008. Semi-supervised On-Line Boosting for Robust Tracking. In: ECCV (1). 234–247.

Grignon, D., J. Eble, S. Bonsib, and H. Moch. 2004. Clear cell renal cell carcinoma. World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs. **IARC Press**.

Gurcan, M., L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener. 2009. Histopathological Image Analysis: A Review. Biomedical Engineering, IEEE Reviews in **2**:147 –171.

Halama, N., I. Zoernig, A. Spille, K. Westphal, P. Schirmacher, D. Jaeger, and N. Grabe. 2009. Estimation of Immune Cell Densities in Immune Cell Conglomerates: An Approach for High-Throughput Quantification. PLoS ONE **4**:e7847.

Hall, B., W. Chen, M. Reiss, and D. J. Foran. 2007. A clinically motivated 2-fold framework for quantifying and classifying immunohistochemically stained specimens. In: MICCAI'07: Proceedings of the 10th international conference on Medical image computing and computer-assisted intervention. Berlin, Heidelberg: Springer-Verlag, 287–294.

Halling, K. C., W. King, I. A. Sokolova, et al. 2000. A comparison of cytology and fluorescence in situ hybridization for the detection of urothelial carcinoma. J. Urol. **164**:1768–1775.

Harnden, P., N. Mahmood, and J. Southgate. 1999. Expression of cytokeratin 20 redefines urothelial papillomas of the bladder. Lancet **353**:974–977.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning: data mining, inference and prediction. Springer, 2 edition.

Hauschild, A., S. S. Agarwala, U. Trefzer, et al. 2009. Results of a phase III, randomized, placebo-controlled study of sorafenib in combination with carboplatin and paclitaxel as second-line treatment in patients with unresectable stage III or stage IV melanoma. J. Clin. Oncol. **27**:2823–2830.

Hayes-Roth, F., D. A. Waterman, and D. B. Lenat. 1983. Building expert systems. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Heere-Ress, E., C. Thallinger, T. Lucas, H. Schlagbauer-Wadl, V. Wacheck, B. P. Monia, K. Wolff, H. Pehamberger, and B. Jansen. 2002. Bcl-X(L) is a chemoresistance factor in human melanoma cells that can be inhibited by antisense therapy. Int. J. Cancer **99**:29–34.

Helvie, M. A., L. Hadjiiski, E. Makariou, et al. 2004. Sensitivity of Noncommercial Computer-aided Detection System for Mammographic Breast Cancer Detection: Pilot Clinical Trial1. Radiology **231**:208–214.

Hernandez, S., E. Lopez-Knowles, J. Lloreta, M. Kogevinas, A. Amoros, A. Tardon, A. Carrato, C. Serra, N. Malats, and F. X. Real. 2006. Prospective study of FGFR3 mutations as a prognostic factor in nonmuscle invasive urothelial bladder carcinomas. J. Clin. Oncol. **24**:3664–3671.

Herold, J., L. Zhou, S. Abouna, S. Pelengaris, D. Epstein, M. Khan, and T. W. Nattkemper. 2009. Integrating semantic annotation and information visualization for the analysis of multichannel fluorescence micrographs from pancreatic tissue. Computerized Medical Imaging and Graphics **In Press, Corrected Proof**:–.

Hla, T., and K. Neilson. 1992. Human cyclooxygenase-2 cDNA. Proc. Natl. Acad. Sci. U.S.A. **89**:7384–7388.

Ho, T. K. 1995. Random decision forests. In: ICDAR. 278–.

———. 1998. The Random Subspace Method for Constructing Decision Forests. IEEE Trans. Pattern Anal. Mach. Intell. **20**:832–844.

Hodi, F. S., S. J. O'Day, D. F. McDermott, et al. 2010. Improved survival with ipilimumab in patients with metastatic melanoma. N. Engl. J. Med. **363**:711–723.

Hothorn, T., P. Buehlmann, S. Dudoit, A. Molinaro, and M. V. D. Laan. 2006. Survival Ensembles. Biostatistics **7**:355–373.

Huisman, A., A. Looijen, S. M. van den Brink, and P. J. van Diest. 2010. Creation of a fully digital pathology slide archive by high-volume tissue slide scanning. Human Pathology **41**:751 – 757.

Hyvarinen, A. 2001. Independent Component Analysis. Neural Computing Surveys **2**.

Ibrahim, J. G., M. hui Chen, and S. N. Maceachern. 1996. Bayesian Variable Selection for Proportional Hazards Models.

Ishwaran, H., U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. 2008. Random survival forests. Ann. Appl. Stat. **2**:841–860.

Jain, A. K., and R. C. Dubes. 1988. Algorithms for clustering data. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Jemal, A., T. Murray, E. Ward, A. Samuels, R. C. Tiwari, A. Ghafoor, E. J. Feuer, and M. J. Thun. 2005. Cancer Statistics, 2005. CA Cancer J Clin **55**:10–30.

Jemal, A., R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M. J. Thun. 2008. Cancer statistics, 2008. CA Cancer J Clin **58**:71–96.

Joseph G.Ibrahim, D. S., Ming-Hui Chen. 2001. Bayesian Survival Analysis. Springer-Verlag:New York Inc.

Kalady, M. F., R. R. White, J. L. Johnson, D. S. Tyler, and H. F. Seigler. 2003. Thin Melanomas: Predictive Lethal Characteristics From a 30-Year Clinical Experience. Annals of Surgery **238**.

Kalles, D., and T. Morris. 1996. Efficient Incremental Induction of Decision Trees. Machine Learning **24**:231–242.

Kaplan, E. L., and P. Meier. 1958. Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association **53**:457–481.

Kass, M., A. Witkin, and D. Terzopoulos. 1988. Snakes: Active contour models. International Journal of Computer Vision **V1**:321–331.

Kasuga, M. 2006. Insulin resistance and pancreatic beta cell failure. J Clin Invest **116**:1756–1760.

Kauppi, T., J.-K. Kamarainen, L. Lensu, V. Kalesnykiene, I. Sorri, H. Kälviäinen, H. Uusitalo, and J. Pietilä. 2009. Fusion of Multiple Expert Annotations and Overall Score Selection for Medical Image Diagnosis. In: SCIA '09: Proceedings of the 16th Scandinavian Conference on Image Analysis. Berlin, Heidelberg: Springer-Verlag, 760–769.

Klein, J. P., and M. L. Moeschberger. 1997. Survival Analysis: Techniques for Censored and Truncated Data. Springer-Verlag:New York Inc.

Kononen, J., L. Bubendorf, A. Kallionimeni, M. Barlund, P. Schraml, S. Leighton, J. Torhorst, M. J. Mihatsch, G. Sauter, and O.-P. Kallionimeni. 1998. Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nat Med **4**:844–847.

Korabiowska, M., U. Brinck, J. Stachura, J. Jawien, F. M. Hasse, C. Cordon-Cardos, and G. Fischer. 2006. Exonic deletions of mismatch repair genes MLH1 and MSH2 correlate with prognosis and protein expression levels in malignant melanomas. Anticancer Res. **26**:1231–1235.

Kottas, A. 2006. Nonparametric Bayesian survival analysis using mixtures of Weibull distributions. Journal of Statistical Planning and Inference **136**:578 – 596.

Kuhn, H. W. 1955. The Hungarian Method for the assignment problem:. Naval Research Logistic Quarterly :2:83–97.

Kuphal, S., R. Bauer, and A. K. Bosserhoff. 2005. Integrin signaling in malignant melanoma. Cancer Metastasis Rev. **24**:195–222.

Lee, D. D., and H. S. Seung. 1999. Learning the parts of objects by non-negative matrix factorization. Nature **401**:788–791.

Lehmussola, A., P. Ruusuvuori, J. Selinummi, H. Huttunen, and O. Yli-Harja. 2007. Computational Framework for Simulating Fluorescence Microscope Images With Cell Populations. Medical Imaging, IEEE Transactions on **26**:1010 –1016.

Lens, M. B., and M. Dawes. 2004. Global perspectives of contemporary epidemiological trends of cutaneous malignant melanoma. Br. J. Dermatol. **150**:179–185.

Lepetit, V., and P. Fua. 2006. Keypoint Recognition Using Randomized Trees. IEEE Trans. Pattern Anal. Mach. Intell. **28**:1465–1479.

Leung, T., and J. Malik. 2001. Representing and recognizing the visual appearance of materials using threedimensional textons. In: International Journal on Computer Vision, volume 43-1. 29–44.

Levenson, R. M., and J. R. Mansfield. 2006. Multispectral imaging in biology and medicine: Slices of life. Cytometry Part A **69A**:748–758.

Lezoray, O., and H. Cardot. 2002. Cooperation of color pixel classification schemes and color watershed: a study for microscopical images. IEEE transactions on Image Processing **11**:783–789.

Lezoray, O., A. Elmoataz, and H. Cardot. 2003. A Color object recognition scheme: application to cellular sorting. Machine Vision and Applications **14**:166–171.

Liaw, A., and M. Wiener. 2002. Classification and Regression by randomForest. R News **2**:18–22.

Littlestone, N. 1988. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. In: Machine Learning. 285–318.

Littlestone, N., and M. K. Warmuth. 1989. The Weighted Majority Algorithm. In: FOCS. 256–261.

Lloyd, S. P. 1982. Least squares quantization in PCM. Information Theory, IEEE Transactions on **28**:129–137.

Lowe, S. W., E. Cepero, and G. Evan. 2004. Intrinsic tumour suppression. Nature **432**:307–315.

Lugosi, G. 1992. Learning with an unreliable teacher. Pattern Recogn. **25**:79–87.

Lzoray, O., and C. Charrier. 2009. Color image segmentation using morphological clustering and fusion with automatic scale selection. Pattern Recognition Letters **30**:397 – 406.

Macenko, M., M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas. 2009. A method for normalizing histology slides for quantitative analysis. In: ISBI'09: Proceedings of the Sixth IEEE international conference on Symposium on Biomedical Imaging. Piscataway, NJ, USA: IEEE Press, 1107–1110.

Maedler, K., D. M. Schumann, F. Schulthess, J. Oberholzer, D. Bosco, T. Berney, and M. Y. Donath. 2006. Aging correlates with decreased beta-cell proliferative capacity and enhanced sensitivity to apoptosis: a potential role for Fas and pancreatic duodenal homeobox-1. Diabetes **55**:2455–2462.

Maenpaa, T., T. Ojala, M. Pietikainen, and M. Soriano. 2000. Robust Texture Classification by Subsets of Local Binary Patterns. In: ICPR '00: Proceedings of the International Conference on Pattern Recognition. Washington, DC, USA: IEEE Computer Society, 3947.

Maier, M., M. Hein, and U. von Luxburg. 2007. Cluster Identification in Nearest-Neighbor Graphs. In: M.Hutter, R. Servedio, and E. Takimoto, editors, Proceedings of the 18th Conference on Algorithmic Learning Theory, volume 4754 of *Lecture Notes in Artificial Intelligence*. Springer, Berlin, 196–210.

Mantel, N., and W. Haenszel. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst **4**:719–748.

Maree, R., P. Geurts, and L. Wehenkel. 2007. Random subwindows and extremely randomized trees for image classification in cell biology. BMC Cell Biology **8**:S2.

Masters, J. R., U. D. Vani, K. M. Grigor, G. O. Griffiths, A. Crook, M. K. Parmar, and M. A. Knowles. 2003. Can p53 staining be used to identify patients with aggressive superficial bladder cancer? J. Pathol. **200**:74–81.

McCullagh, P., and J. Yang. 2008. How many clusters? Bayesian Analysis **3**:101–120.

McCullaghand, P., and J. Nelder. 1983. Generalized Linear Models. Chapman & Hall.

McKinnon, J. G., X. Q. Yu, W. H. McCarthy, and J. F. Thompson. 2003. Prognosis for patients with thin cutaneous melanoma. Cancer **98**:1223–1231.

McLachlan, G. J. 2004. Discriminant analysis and statistical pattern recognition. Wiley-Interscience, Hoboken, N.J.

Meier, F., S. Will, U. Ellwanger, B. Schlagenhauff, B. Schittek, G. Rassner, and C. Garbe. 2002. Metastatic pathways and time courses in the orderly progression of cutaneous melanoma. British Journal of Dermatology **147**:62–70.

Mesker, W. E., H. Torrenga, W. C. R. Sloos, H. Vrolijk, R. A. E. M. Tollenaar, P. C. de Bruin, P. J. van Diest, and H. J. Tanke. 2004. Supervised automated microscopy increases sensitivity and efficiency of detection of sentinel node micrometastases in patients with breast cancer. Journal of Clinical Pathology **57**:960–964.

Meurie, C., O. Lezoray, C. Charrier, and A. Elmoataz. 2005. Combination of multiple pixel classifiers for microscopic image segmentation. IJRA (IASTED International Journal of Robotics and Automation) **20**:63–69. Special issue on Colour Image Processing and Analysis for Machine Vision, ISSN 0826-8185.

Meyer, S., T. Vogt, M. Landthaler, et al. 2009. Cyclooxygenase 2 (COX2) and Peroxisome Proliferator-Activated Receptor Gamma (PPARG) Are Stage-Dependent Prognostic Markers of Malignant Melanoma. PPAR Res **2009**:848645.

Meyer, S., P. J. Wild, T. Vogt, F. Bataille, C. Ehret, S. Gantner, M. Landthaler, M. Klinkhammer-Schalke, F. Hofstaedter, and A. K. Bosserhoff. 2010. Methylthioadenosine phosphorylase represents a predictive marker for response to adjuvant interferon therapy in patients with malignant melanoma. Exp. Dermatol. **19**:e251–257.

Moch, H., W. Artibani, B. Delahunt, V. Ficarra, R. Knuechel, F. Montorsi, J.-J. Patard, C. G. Stief, T. Sulser, and P. J. Wild. 2009. Reassessing the Current UICC/AJCC TNM Staging for Renal Cell Carcinoma. European Urology **56**:636 – 643.

Moch H, e. a., Schraml P. 1999. High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma. Am J Pathol. **Apr;154(4)**:981–6.

Monaco, J. P., J. E. Tomaszewski, M. D. Feldman, I. Hagemann, M. Moradi, P. Mousavi, A. Boag, C. Davidson, P. Abolmaesumi, and A. Madabhushi. 2010. High-throughput detection of prostate cancer in histological sections using probabilistic pairwise Markov models. Medical Image Analysis **14**:617 – 629.

Morton, D. L., J. F. Thompson, A. J. Cochran, et al. 2006. Sentinel-node biopsy or nodal observation in melanoma. N. Engl. J. Med. **355**:1307–1317.

Morton, D. L., D.-R. Wen, J. H. Wong, J. S. Economou, L. A. Cagle, F. K. Storm, L. J. Foshag, and A. J. Cochran. 1992. Technical Details of Intraoperative Lymphatic Mapping for Early Stage Melanoma. Arch Surg **127**:392–399.

Mostofi, F. K., L. H. Sobin, and H. Torloni. 1973. Histological typing of urinary bladder tumours. International classification of tumours. World Health Organization, Geneva :.

Murray, C. A., W. L. Leong, D. R. McCready, and D. M. Ghazarian. 2004. Histopathological patterns of melanoma metastases in sentinel lymph nodes. J. Clin. Pathol. **57**:64–67.

Nattkemper, T. W., T. Twellmann, H. Ritter, and W. Schubert. 2003. Human vs. machine: evaluation of fluorescence micrographs. Computers in Biology and Medicine **33**:31 – 43.

Newberg, J. Y., J. Li, A. Rao, F. Pontén, M. Uhlén, E. Lundberg, and R. F. Murphy. 2009. Automated analysis of human protein atlas immunofluorescence images. In: ISBI'09: Proceedings of the Sixth IEEE international conference on Symposium on Biomedical Imaging. Piscataway, NJ, USA: IEEE Press, 1023–1026.

Nguyen, T. T., H. Grabner, H. Bischof, and B. Gruber. 2007. On-line Boosting for Car Detection from Aerial Images. In: RIVF. IEEE, 87–95.

Nocito, A., L. Bubendorf, E. M. Tinner, et al. 2001. Microarrays of bladder cancer tissue are highly representative of proliferation index and histological grade. J. Pathol. **194**:349–357.

Nocito A, e. a., Bubendorf L. 2001. Microarrays of bladder cancer tissue are highly representative of proliferation index and histological grade. J Pathol. **Jul;194(3):**:349–57.

Ojala, T., M. Pietikainen, and D. Harwood. 1996. A comparative study of texture measures with classification based on featured distributions. Pattern Recognition **29**:51 – 59.

Oster, S., S. Langella, S. Hastings, et al. 2008. caGrid 1.0: An Enterprise Grid Infrastructure for Biomedical Research. Journal of the American Medical Informatics Association **15**:138–149.

Oza, N. C. 2001. Online Ensemble Learning. Ph.D. thesis, The University of California, Berkeley, CA.

Papageorgiou, C., and T. Poggio. 2000. A Trainable System for Object Detection. Int. J. Comput. Vision **38**:15–33.

Paserman, M. D. 2004. Bayesian Inference for Duration Data with Unobserved and Unknown Heterogeneity: Monte Carlo Evidence and an Application. IZA Discussion Papers 996, Institute for the Study of Labor (IZA).

Pavlov, Y. 2000. Random forests. VSP.

Pfahringer, B., G. Holmes, and R. Kirkby. 2007. New Options for Hoeffding Trees. In: Orgun, M. A., and J. Thornton, editors, Australian Conference on Artificial Intelligence, volume 4830 of *Lecture Notes in Computer Science*. Springer, 90–99.

Pinto, M. M. 1986. An immunoperoxidase study of S-100 protein in neoplastic cells in serous effusions. Use as a marker for melanoma. Acta Cytol. **30**:240–244.

Pitman, J. 1999. Coalescent random forests. J. Combin. Theory Ser. A **85**:165–193.

Placer, J., B. Espinet, M. Salido, F. Sole, and A. Gelabert-Mas. 2002. Clinical utility of a multiprobe FISH assay in voided urine specimens for the detection of bladder cancer and its recurrences, compared with urinary cytology. Eur. Urol. **42**:547–552.

Pontén, F., K. Jirström, and M. Uhlen. 2008. The Human Protein Atlas?-?a tool for pathology. The Journal of Pathology **216**:387–393.

Quinlan, J. R. 1993. C4.5: programs for machine learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

R Development Core Team. 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rabinovich, A., S. Agarwal, C. A. Laris, J. H. Price, and S. Belongie. 2003. Unsupervised Color Decomposition of Histologically Stained Tissue Samples. In: in Advances in Neural Information Processing Systems. MIT Press.

Raman, S., T. J. Fuchs, P. J. Wild, E. Dahl, and V. Roth. 2009. The Bayesian group-Lasso for analyzing contingency tables. In: ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning. New York, NY, USA: ACM, 881–888.

Raman, S., and V. Roth. 2009. Sparse Bayesian Regression for Grouped Variables in Generalized Linear Models. In: Proceedings of the 31st DAGM Symposium on Pattern Recognition. Berlin, Heidelberg: Springer-Verlag, 242–251.

Rasmussen, C. E., and Z. Ghahramani. 2001. Infinite Mixtures of Gaussian Process Experts. In: In Advances in Neural Information Processing Systems 14. MIT Press, 881–888.

Raykar, V. C., S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy. 2009. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In: ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning. New York, NY, USA: ACM, 889–896.

Reeves, A. P., and W. J. Kostis. 2000. Computer-aided diagnosis for lung cancer. Radiol. Clin. North Am. **38**:497–509.

Reichle, A., T. Vogt, B. Coras, et al. 2007. Targeted combined anti-inflammatory and angiostatic therapy in advanced melanoma: a randomized phase II trial. Melanoma Res. **17**:360–364.

Reintgen, D., C. W. Cruse, K. Wells, et al. 1994. The Orderly Progression of Melanoma Nodal Metastases. Annals of Surgery 00000658-199412000-00009 **220**:759–767.

Rojo, M., G. Bueno, and J. Slodkowska. 2009. Review of imaging solutions for integrated quantitative immunohistochemis try in the Pathology daily practice. Folia Histochemica et Cytobiologica **47**:349–354.

Rosen, O., and M. Tanner. 1999. Mixtures of Proportional Hazards Regression models. Statistics in Medicine **18**:1119–1131.

Roth, P. M., H. Grabner, C. Leistner, M. Winter, and H. Bischof. 2008a. InterActive Learning a Person Detector: Fewer Clicks - Less Frustration. In: Proc. Workshop of the Austrian Association for Pattern Recognition.

Roth, V., T. J. Fuchs, S. Raman, P. Wild, E. Dahl, and J. M. Buhmann. 2008b. Full Bayesian Survival Models for Analyzing Human Breast Tumors. In: NIPS Workshop: MLCP–Machine Learning in Computational Biology.

Ruifrok, A. C., and D. A. Johnston. 2001. Quantification of histochemical staining by color deconvolution. Analyt Quant Cytol Histol **23**:291299.

Sarkis, A. S., G. Dalbagni, C. Cordon-Cardo, Z. F. Zhang, J. Sheinfeld, W. R. Fair, H. W. Herr, and V. E. Reuter. 1993. Nuclear overexpression of p53 protein in transitional cell bladder carcinoma: a marker for disease progression. J. Natl. Cancer Inst. **85**:53–59.

Saur, S. C., H. Alkadhi, L. Desbiolles, T. J. Fuchs, G. Szkely, and P. C. Cattin. 2009. Guided review by frequent itemset mining: additional evidence for plaque detection. Int J Comput Assist Radiol Surg **4**:263–271.

Saur, S. C., H. Alkadhi, P. Stolzmann, S. Baumller, S. Leschka, H. Scheffel, L. Desbiolles, T. J. Fuchs, G. Szkely, and P. C. Cattin. 2010. Effect of reader experience on variability, evaluation time and accuracy of coronary plaque detection with computed tomography coronary angiography. Eur Radiol **20**:1599–1606.

Schilham, A. M., B. van Ginneken, and M. Loog. 2006. A computer-aided diagnosis system for detection of lung nodules in chest radiographs with an evaluation on a public database. Med Image Anal **10**:247–258.

Schmid, C. 2001. Constructing models for content-based image retrieval. In: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, volume 2. II–39 – II–45 vol.2.

Schneiderman, H. W. 2000. A statistical approach to 3d object detection applied to faces and cars. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, USA. Chair-Kanade, Takeo.

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: DAGM-Symposium. 202–211.

Schwarz, S., M. Rechenmacher, T. Filbeck, R. Knuechel, H. Blaszyk, A. Hartmann, and G. Brockhoff. 2008. Value of multicolour fluorescence in situ hybridisation (UroVysion) in the differential diagnosis of flat urothelial lesions. J. Clin. Pathol. **61**:272–277.

Sharp, T. 2008. Implementing Decision Trees and Forests on a GPU. In: Forsyth, D., P. Torr, and A. Zisserman, editors, Computer Vision ECCV 2008, volume 5305 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 595–608.

Sheffield, M. V., H. Yee, C. C. Dorvault, K. N. Weilbaecher, I. A. Eltoum, G. P. Siegal, D. E. Fisher, and D. C. Chhieng. 2002. Comparison of five antibodies as markers in the diagnosis of melanoma in cytologic preparations. Anglais.

Sherman, A. B., L. G. Koss, and S. E. Adams. 1984. Interobserver and intraobserver differences in the diagnosis of urothelial cells. Comparison with classification by computer. Anal Quant Cytol **6**:112–120.

Shin, D., M. C. Pierce, A. M. Gillenwater, M. D. Williams, and R. R. Richards-Kortum. 2010. A Fiber-Optic Fluorescence Microscope Using a Consumer-Grade Digital Camera for ¡italic¿In Vivo¡/italic¿ Cellular Imaging. PLoS ONE **5**:e11218.

Shotton, J., J. M. Winn, C. Rother, and A. Criminisi. 2006. TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation. In: ECCV (1). 1–15.

Simon, R., M. Mirlacher, and G. Sauter. 2003. Tissue microarrays in cancer diagnosis. Expert Rev. Mol. Diagn. **3**:421–430.

Smyth, P. 1996. Bounds on the mean classification error rate of multiple experts. Pattern Recognition Letters **17**:1253 – 1257.

Smyth, P., U. M. Fayyad, M. C. Burl, P. Perona, and P. Baldi. 1994. Inferring Ground Truth from Subjective Labelling of Venus Images. In: NIPS. 1085–1092.

Soille, P. 2003. Morphological Image Analysis: Principles and Applications. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Sonka, M., V. Hlavac, and R. Boyle. 2007. Image Processing, Analysis, and Machine Vision. Thomson-Engineering.

Srivastava, M. 2003. Singular Wishart and multivariate beta distributions. Annals of Statistics **31**:1537–1560.

Strobl, C., A.-L. Boulesteix, and T. Augustin. 2007. Unbiased split selection for classification trees based on the Gini Index. Computational Statistics & Data Analysis **52**:483–501.

Sudarshan S, L. W. 2006. Genetic basis of cancer of the kidney. Semin Oncol. **Oct;33(5)**:544–51.

Takahashi M, e. a., Rhodes DR. 2001. Gene expression profiling of clear cell renal cell carcinoma: gene identification and prognostic classification. In: Proc Natl Acad Sci U S A., volume Aug 14;98(17). 9754–9.

Tannapfel, A., H. Hahn, A. Katalinic, R. Fietkau, R. Khn, and C. Wittekind. 1996. Prognostic value of ploidy and proliferation markers in renal cell carcinoma. Cancer **Jan 1;77(1)**:164–71.

Thallinger, G., K. Baumgartner, M. Pirklbauer, M. Uray, E. Pauritsch, G. Mehes, C. Buck, K. Zatloukal, and Z. Trajanoski. 2007. TAMEE: data management and analysis for tissue microarrays. BMC Bioinformatics **8**:81.

Thun, M. J., S. J. Henley, and T. Gansler. 2004. Inflammation and cancer: an epidemiological perspective. Novartis Found. Symp. **256**:6–21.

Tosoni, I., U. Wagner, G. Sauter, M. Egloff, H. Knonagel, G. Alund, F. Bannwart, M. J. Mihatsch, T. C. Gasser, and R. Maurer. 2000. Clinical significance of interobserver differences in the staging and grading of superficial bladder cancer. BJU Int. **85**:48–53.

Tukey, J. W. 1977. Exploratory Data Analysis. Addison-Wesley.

Tullock, G. 1959. Problems of Majority Voting. The Journal of Political Economy **67**:571–579.

Utgoff, P. E. 1989. Incremental Induction of Decision Trees. Machine Learning **4**:161–186.

———. 1994. An Improved Algorithm for Incremental Induction of Decision Trees. In: ICML. 318–325.

van der Aa, M. N., E. W. Steyerberg, E. F. Sen, E. C. Zwarthoff, W. J. Kirkels, T. H. van der Kwast, and M. L. Essink-Bot. 2008. Patients' perceived burden of cystoscopic and urinary surveillance of bladder cancer: a randomized comparison. BJU Int. **101**:1106–1110.

van der Aa, M. N., E. C. Zwarthoff, E. W. Steyerberg, M. W. Boogaard, Y. Nijsen, K. A. van der Keur, A. J. van Exsel, W. J. Kirkels, C. Bangma, and T. H. van der Kwast. 2009. Microsatellite analysis of voided-urine samples for surveillance of low-grade non-muscle-invasive urothelial carcinoma: feasibility and clinical utility in a prospective multicenter study (Cost-Effectiveness of Follow-Up of Urinary Bladder Cancer trial [CEFUB]). Eur. Urol. **55**:659–667.

van der Loos, C. M. 2008. Multiple Immunoenzyme Staining: Methods and Visualizations for the Observation With Spectral Imaging. J. Histochem. Cytochem. **56**:313–328.

van Oers, J. M., I. Lurkin, A. J. van Exsel, Y. Nijsen, B. W. van Rhijn, M. N. van der Aa, and E. C. Zwarthoff. 2005. A simple and fast method for the simultaneous detection of nine fibroblast growth factor receptor 3 mutations in bladder cancer and voided urine. Clin. Cancer Res. **11**:7743–7748.

van Oers, J. M., P. J. Wild, M. Burger, et al. 2007. FGFR3 mutations and a normal CK20 staining pattern define low-grade noninvasive urothelial bladder tumours. Eur. Urol. **52**:760–768.

van Rhijn, B. W., I. Lurkin, D. K. Chopin, W. J. Kirkels, J. P. Thiery, T. H. van der Kwast, F. Radvanyi, and E. C. Zwarthoff. 2003a. Combined microsatellite and FGFR3 mutation analysis enables a highly sensitive detection of urothelial cell carcinoma in voided urine. Clin. Cancer Res. **9**:257–263.

van Rhijn, B. W., I. Lurkin, F. Radvanyi, W. J. Kirkels, T. H. van der Kwast, and E. C. Zwarthoff. 2001. The fibroblast growth factor receptor 3 (FGFR3) mutation is a strong indicator of superficial bladder cancer with low recurrence rate. Cancer Res. **61**:1265–1268.

van Rhijn, B. W., R. Montironi, E. C. Zwarthoff, A. C. Jobsis, and T. H. van der Kwast. 2002. Frequent FGFR3 mutations in urothelial papilloma. J. Pathol. **198**:245–251.

van Rhijn, B. W., H. G. van der Poel, and T. H. van der Kwast. 2005. Urine markers for bladder cancer surveillance: a systematic review. Eur. Urol. **47**:736–748.

van Rhijn, B. W., A. N. Vis, T. H. van der Kwast, W. J. Kirkels, F. Radvanyi, E. C. Ooms, D. K. Chopin, E. R. Boeve, A. C. Jobsis, and E. C. Zwarthoff. 2003b. Molecular grading of urothelial cell carcinoma with fibroblast growth factor receptor 3 and MIB-1 is superior to pathologic grade for the prediction of clinical outcome. J. Clin. Oncol. **21**:1912–1921.

Vapnik, V. 1998. Statistical Learning Theory. New York: Wiley. Forthcoming.

Vijan, S., D. L. Stevens, W. H. Herman, M. M. Funnell, and C. J. Standiford. 1997. Screening, prevention, counseling, and treatment for the complications of type II diabetes mellitus. Putting evidence into practice. J Gen Intern Med **12**:567–580.

Viola, P., and M. Jones. 2001. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, volume 1. Los Alamitos, CA, USA: IEEE Comput. Soc, I–511–I–518.

Vogt, J. E., S. Prabhakaran, T. J. Fuchs, and V. Roth. 2010. The Translation-invariant Wishart-Dirichlet Process for Clustering Distance Data. In: ICML 2010: Proceedings of the 27th Annual International Conference on Machine Learning. 1111–1118.

Weaver, D. L., D. N. Krag, E. A. Manna, T. Ashikaga, S. P. Harlow, and K. D. Bauer. 2003. Comparison of Pathologist-Detected and Automated Computer-Assisted Image Analysis Detected Sentinel Lymph Node Micrometastases in Breast Cancer. Mod Pathol **16**:1159–1163.

Welinder, P., and P. Perona. 2010. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In: Workshop on Advancing Computer Vision with Humans in the Loop at CVPR'10.

Wetzel, A. 1997. Computational Aspects of Pathology Image Classification and Retrieval. J. Supercomput. **11**:279–293.

Whitehill, J., P. Ruvolo, T. fan Wu, J. Bergsma, and J. Movellan. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In: Bengio, Y., D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, Advances in Neural Information Processing Systems 22. 2035–2043.

Wild, P. J., S. Meyer, F. Bataille, et al. 2006. Tissue microarray analysis of methylthioadenosine phosphorylase protein expression in melanocytic skin tumors. Arch Dermatol **142**:471–476.

Yang, L., W. Chen, P. Meer, G. Salaru, M. D. Feldman, and D. J. Foran. 2007. High throughput analysis of breast cancer specimens on the grid. In: MICCAI'07: Proceedings of the 10th international conference on Medical image computing and computer-assisted intervention. Berlin, Heidelberg: Springer-Verlag, 617–625.

Young AN, e. a., Amin MB. 2001. Expression profiling of renal epithelial neoplasms: a method for tumor classification and discovery of diagnostic molecular markers. Am J Pathol. **May;158(5)**:1639–51.

Yuan, M., and Y. Lin. 2006. Model Selection and Estimation in Regression with Grouped Variables. J. Roy. Stat. Soc. B :49–67.

Zabierowski, S. E., and M. Herlyn. 2008. Melanoma stem cells: the dark seed of melanoma. J. Clin. Oncol. **26**:2890–2894.

Zhang, S., and D. Yu. 2010. PI(3)king apart PTEN's role in cancer. Clin. Cancer Res. **16**:4325–4330.

Zimmermann, T. 2005. Spectral Imaging and Linear Unmixing in Light Microscopy. In: Rietdorf, J., editor, Microscopy Techniques, volume 95 of *Advances in Biochemical Engineering/Biotechnology*. Springer Berlin / Heidelberg, 245–265. 10.1007/b102216.

Zimmet, P., K. G. Alberti, and J. Shaw. 2001. Global and societal implications of the diabetes epidemic. Nature **414**:782–787.

List of own Publications

Parts of this thesis have been published in the following papers:

1. Thomas J. Fuchs and Joachim M. Buhmann, Computational Pathology: Challenges and Promises for Tissue Analysis, Journal of Computerized Medical Imaging and Graphics, Special Issue on Whole Slide Microscopic Image Processing, 2011 (Article in Press)

2. Johannes Haybaeck, Mathias Heikenwalder, Britta Klevenz, Petra Schwarz, Ilan Margalith, Claire Bridel, Kirsten Mertz, Elizabeta Zirdum, Benjamin Petsch, Thomas J. Fuchs, Lothar Stitz, Adriano Aguzzi, Aerosols Transmit Prions to Immunocompetent and Immunodeficient Mice, PLoS Pathog 7(1): e1001257. 2011

3. Sudhir Raman, Thomas J. Fuchs, Peter J. Wild , Edgar Dahl, Joachim M. Buhmann and Volker Roth, Infinite mixture-of-experts model for sparse survival regression with application to breast cancer, BMC Bioinformatics 2010

4. Peter J. Schuefler, Thomas J. Fuchs, Cheng Soon Ong, Joachim M. Buhmann, Computational TMA Analysis and Cell Nucleus Classification of Renal Cell Carcinoma, DAGM 2010

5. Kira Bettermann, Mihael Vucur, Johannes Haybaeck, Christiane Koppe, Jrn Janssen, Felix Heymann, Achim Weber, Ralf Weiskirchen, Christian Liedtke, Nikolaus Gassler, Michael Mller, Rita de Vos, Monika Julia Wolf, Yannick Boege, Gitta Maria Seleznik, Nicolas Zeller, Daniel Erny, Thomas Fuchs, Stefan Zoller, Stefano Cairo, Marie-Annick Buendia, Marco Prinz, Shizuo Akira, Frank Tacke, Mathias Heikenwalder, Christian Trautwein and Tom Luedde, TAK1 suppresses a NEMO-dependent, but NF-kB-independent pathway to liver cancer, Cancer CELL 17, 481-496, May 18, 2010

6. Verena Kaynig, Thomas J. Fuchs, Joachim M. Buhmann, Geometrical Consistent 3D Tracing of Neuronal Processes in ssTEM Data, MICCAI 2010

7. Julia E. Vogt, Sandhya Prabhakaran, Thomas J. Fuchs and Volker Roth, The Translation-invariant Wishart-Dirichlet Process for Clustering Distance Data, ICML 2010, runner-up Best Paper Award

8. Verena Kaynig, Thomas J. Fuchs, Joachim M. Buhmann, Neuron Geometry Extraction by Perceptual Grouping in ssTEM images, CVPR 2010

9. Stefan C. Saur, Hatem Alkadhi, Paul Stolzmann, Stephan Baumller, Sebastian Leschka, Hans Scheffel, Lotus Desbiolles, Thomas J. Fuchs, Gbor Szkely, Philippe C. Cattin, Effect of reader experience on variability, evaluation time and accuracy of coronary plaque detection with computed tomography coronary angiography, European Radiology, Volume 20, Number 7 / July, 2010

10. Thomas J. Fuchs, Johannes Haybaeck, Peter Wild, Mathias Heikenwalder, Holger Moch, Adriano Aguzzi, Joachim M. Buhmann, Randomized Tree Ensembles for Object Detection in Computational Pathology, ISVC 2009

11. Thomas J. Fuchs and Joachim M. Buhmann, Inter-Active Learning of Randomized Tree Ensembles for Object Detection, Workshop on On-line Learning for Computer Vision, ICCV 2009

12. Wild PJ, Fuchs TJ, Stoehr R, Zimmermann D, Frigerio S, Padberg B, Steiner I, Zwarthoff EC, Burger M, Denzinger S, Hofstaedter F, Kristiansen G, Hermanns T, Seifert HH, Provenzano M, Sulser T, Roth V, Buhmann JM, Moch H, Hartmann A., Detection of urothelial bladder cancer cells in voided urine can be improved by a combination of cytology and standardized microsatellite analysis., Cancer Epidemiol Biomarkers Prev. 2009 Jun;18(6):1798-806. Epub 2009 May 19.

13. Juergen Veeck, Peter J. Wild, Thomas J. Fuchs, Peter J. Schueffler, Arndt Hartmann, Ruth Knchel, Edgar Dahl, Prognostic relevance of Wnt-inhibitory factor-1 (WIF1) and Dickkopf-3 (DKK3) promoter methylation in human breast cancer, BMC Cancer 2009, 9:217

14. Floros X., Fuchs T.J., Rechsteiner M., Spinas G., Moch H., Buhmann J.M., Graph-Based Pancreatic Islet Segmentation for Early Type 2 Diabetes Mellitus on Histopathological Tissue, Medical Image Computing and Computer-Assisted Intervention, MICCAI 2009

15. Edgar Dahl, Abdelaziz En-Nia, Frank Wiesmann, Renate Krings, Sonja Djudjaj, Elisabeth Breuer, Thomas J Fuchs, Peter J Wild, Arndt Hartmann, Sandra E Dunn and Peter R Mertens, Nuclear detection of Y-box protein-1 (YB-1) closely associates with progesterone receptor negativity and is a

strong adverse survival factor in human breast cancer, BMC Cancer 2009, 9:410

16. Stefan C. Saur, Philippe C. Cattin, Lotus Desbiolles, Thomas J. Fuchs, Gbor Szkely, Hatem Alkadhi, Prediction Rules for the Detection of Coronary Artery Plaques: Evidence From Cardiac CT, Journal of Investigative Radiology, 2009, vol. 44, no8, pp. 483-490

17. Sudhir Raman, Thomas J. Fuchs, Peter J. Wild, Edgar Dahl, Volker Roth, The Bayesian Group-Lasso for Analyzing Contingency Tables, Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, ICML 2009.

18. Stefan C. Saur, Hatem Alkadhi, Lotus Desbiolles, Thomas J. Fuchs, Gbor Szkely and Philippe C. Cattin, Guided review by frequent itemset mining: additional evidence for plaque detection, International Journal of Computer Assisted Radiology and Surgery, 2009

19. Fuchs TJ, Wild P, Moch H, Buhmannm J, Computational Pathology Analysis of Tissue Microarrays Predicts Survival of Renal Clear Cell Carcinoma Patients", Medical Image Computing and Computer-Assisted Intervention, MICCAI 2008.

20. Fuchs TJ, Lange T, Wild P, Moch H, Buhmann J, Weakly Supervised Cell Nuclei Detection and Segmentation on Tissue Microarrays of Renal Cell Carcinoma, DAGM 2008.

21. Fuürnstahl P, Fuchs TJ, Schweizer A, Nagy L, Szkely G, Harders M, Automatic and Robust Forearm Segmentation using Graph Cuts, Biomedical Imaging: From Nano to Macro, ISBI 2008.

22. Gluz O, Wild P, Meiler R, Diallo-Danebrock R, Ting E, Mohrmann S, Schuett G, Dahl E, Fuchs TJ, Herr A, Gaumann A, Frick M, Poremba C, Nitz UA, Hartmann A., Nuclear Karyopherin 2 expression predicts poor survival in patients with advanced breast cancer irrespective of treatment intensity., International Journal of Cancer 2008 Sep 15;123(6):1433-8.

Curriculum Vitae
Thomas J. Fuchs

Personal Details

Name	Thomas Josef Fuchs
Current Address	Nelkenstrasse 20 8006 Zurich Switzerland
Telefone	+41 44 632 8292
Mobile	+41 76 200 1492
E-Mail	thomas.fuchs@inf.ethz.ch
Web	www.inf.ethz.ch/personal/thomas.fuchs
Date of Birth	April 13^{th}, 1979
Place of Birth	Graz, Austria
Nationality	Austria

Education

12/2005 to date	PhD Candidate at ETH Zurich Pattern Analysis and Machine Learning Group Department of Computer Science
10/27/2005	Diploma Examination
2004–2005	Master thesis at Siemens Corporate Research in Princeton, NJ, USA, "Bayesian Networks Classification in Bioinformatics and Medical Decision Support",
1998-2005	Student of Technical Mathematics/Information Processing at Technical University Graz (Erzherzog-Johann-Universität)
1997–1998	Military Service
1997	High School graduation
1989–1997	High School, Bundesgymnasium Carnerigasse Graz
1985–1989	Elementary School, Volksschule Graz-Schönau