Dissertation

Automated Histopathological Image Analysis of Human Bone Marrow Tissue using Supervised Machine Learning

submitted by

Philipp KAINZ, MSc BSc

for the Academic Degree of

Doctor of Medical Science

Dr.scient.med.

at the

Medical University of Graz

Institute of Biophysics

under the supervision of

Ao.Univ.-Prof. Mag. Dr.rer.nat. Helmut AHAMMER

Institute of Biophysics, Medical University of Graz

Graz, July 2016

Eidesstattliche Erklärung

Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig angefertigt und abgefasst, und jene Personen und Institutionen, die am Zustandekommen der Forschungsdaten beteiligt waren, namentlich genannt habe. Andere als die angegebenen Quellen habe ich nicht verwendet und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen habe ich als solche kenntlich gemacht. Die Arbeit an der Dissertation und daraus entstandener Publikationen wurde gemäß den Regeln der „Good Scientific Practice“ durchgeführt.

Statutory Declaration

I hereby declare that this thesis is my own original work and that I have fully acknowledged by name all of those individuals and organisations that have contributed to the research for this thesis. Due acknowledgement has been made in the text to all other material used. Throughout this thesis and in all related publications I followed the guidelines of “Good Scientific Practice”.

Graz, _____ _____

Date Signature

Abstract

Classification of cell types in context of the architecture in tissue specimen is the basis of diagnostic pathology. Decisions for comprehensive investigations rely on a valid interpretation of tissue morphology. Especially visual examination of bone marrow cells in gigapixel histopathological images of biopsy sections consumes a considerable amount of time, where both intra- and inter-observer variability can be remarkable.

This thesis proposes and evaluates approaches based on supervised machine learning for the automated localization and classification of bone marrow cells and their maturity. Their main advantage is that they work on raw images without relying on segmentation or manual feature extraction. A new method termed *proximity score regression* is introduced, where employing the Random Forest (RF) algorithm enables an easy implementation. For each image location, a non-linear monotonous function of the distance to the closest cell nucleus center is predicted. Cell centers can then be identified by revealing locally maximal locations in the proximity score map. On five challenging datasets, the proposed approach outperforms current state-of-the-art methods in terms of detection reliability, spatial localization accuracy, and speed.

To classify maturation stages, a rotation-invariant classification scheme for multi-class Echo State Networks (ESNs) is proposed. Based on representing 2D single-cell image patches as temporal sequence of rotations, ESNs robustly recognize cells of arbitrary orientations by taking advantage of their short-term memory capacity. A comparison to a standard RF classifier is provided and discussed.

This thesis provides evidence that the application of supervised machine learning facilitates reliable image analysis systems, characterized by a predictable error. Driven by the key requirement of having reliable ground truth data, a web-application is presented that enables the conduction of controlled inter-observer reliability studies. While the presented results look promising for computer-aided diagnosis, an assessment of the agreement among algorithms and human observers must be studied in future work.

Kurzfassung

Zelltypenklassifikation im Kontext der Gewebsarchitektur ist Grundlage der diagnostischen Pathologie, Entscheidungen hinsichtlich umfassender Untersuchungen beruhen auf einer validen Interpretation der Gewebsmorphologie. Speziell die visuelle Untersuchung von Knochenmarkszellen in histologischen Bildern ist zeitaufwendig und kann zu einer markanten Intra- und Interobservervariabilität führen.
In der vorliegenden Dissertation werden Methoden basierend auf überwachtem maschinellen Lernen zur automatischen Lokalisation und Klassifikation von Knochenmarkszellen in verschiedenen Reifegraden vorgestellt und evaluiert. Diese Methoden können direkt auf Bilddaten ohne vorangehende Segmentierung oder manuelle Merkmalsextraktion angewendet werden. Eine neue Methode namens *Proximity Score Regression* wird gezeigt, deren Implementierung basierend auf dem Random Forest (RF) Algorithmus sehr einfach möglich ist. Für jede Bildposition wird eine nichtlineare, monotone Funktion der Distanz zum nächstgelegenen Zellkernmittelpunkt vorhergesagt, wobei die Positionen lokaler Maxima den Lokalisationen der Zellkerne entsprechen. Untersuchungen auf fünf anspruchsvollen Bilddatenbanken zeigen, dass dieser Ansatz hinsichtlich räumlicher Lokalisierung und Detektionssgenauigkeit verlässlicher und schneller funktioniert als bisherige State-of-the-Art Methoden. Zur Klassifikation der Reifegrade wird ein neues rotationsinvariantes Klassifikationsschema für Echo State Networks (ESNs) vorgestellt. Bilder einzelner Zellen werden durch Rotation in Zeitsignale transformiert. Es wird gezeigt, wie ESNs aufgrund ihres inherenten Kurzzeitgedächtnisses Einzelzellen robust, und unabhängig von deren Orientierung, erkennen können. Ein Vergleich mit einem RF Klassifikator wird durchgeführt und diskutiert.
Diese Arbeit zeigt, dass überwachtes maschinelles Lernen den Einsatz verlässlicher, präziser Bildanalysesysteme ermöglicht, deren Fehler abschätzbar ist. Motiviert durch die Anforderung dieser Methoden an valide Beispieldaten wurde eine Web-Applikation entwickelt, die eine Durchführung kontrollierter Interobserverstudien ermöglicht. Die

vielversprechenden Resultate zeigen Möglichkeiten für computergestützte Diagnoseprozesse auf. In zukünftiger Forschungsarbeit muss die Kongruenz zwischen mehreren Pathologen und den automatischen Analysemethoden studiert werden.

Acknowledgements

I would first like to thank all people who unconditionally supported me over the last few years, above all my family, significant other, lab mates, fellow PhD student inmates and all other mates I am very happy to call my friends. This entire PhD project and this thesis would have not been possible without their motivational speeches, driven by their faith in me and my work. In particular, I am deeply grateful to Nici and Mike for their support and efforts especially in final phases of this undertaking, when things got tough unexpectedly.

Special thank goes to my supervisors Helmut Ahammer, Harald Burgsteiner, and Martin Asslaber for their active participation in the discussions of novel concepts and for contributing valuable ideas how to approach new scientific and technological challenges. Unfortunately, I cannot mention the entire multitude of people explicitly here, but several people supported the progression of this work in many different ways. First, I would like to acknowledge the work of our collaborating pathologists in this project. Martin Asslaber and Ariane Aigelsreiter from the Institute of Pathology at the Medical University of Graz, and Carlos Abrantes from the University Hospital of Coimbra, Portugal, contributed to a very important first step towards a validated hematopoietic cell dataset for machine learning purposes with controlled inter-observer variability. I further like to thank Patrick Wiedner and Michaela Janschitz from the Institute of Biophysics at the Medical University of Graz, for digitialization of most of the histopathological glass slides and providing figure drafts, respectively.

During the last two years, I have closely collaborated with several people from the Institute of Computer Graphics and Vision (ICG) at Graz University of Technology, to whom I attribute a significant leap in my progress. Hence, I would like to thank Vincent Lepetit, Paul Wohlhart, Samuel Schulter, and Martin Urschler for their collaboration

on the MICCAI 2015 project on cell detection in terms of invaluable discussions, coding, conduction of experiments, and the provided funding to attend the conference.

In 2015, during a six month research visit at the Institute of Neuroinformatics (INI) at the University of Zurich and ETH Zurich, Switzerland, I started working on Deep Learning under the supervision of Michael Pfeiffer. I would like to thank Michael for a highly efficient and productive time, very informative discussions and important input during the EANN 2015 project on cell classification. In addition, we successfully worked together with Martin Urschler on a MICCAI grand challenge project. In relation to the challenge project, I would like to further acknowledge the support of Daniel Neil and Julien Martel (INI), who were exceptionally patient in explaining concepts of deep Neural Networks, particularly in early phases of the challenge.

The third, very important and close collaboration has been established in 2014 with the team of the Cytomine project at GIGA-Research and the Systems and Modeling research unit of the Department of Electrical Engineering and Computer Science at the University of Liège, Belgium. Besides the entire team, I would specifically like to thank Raphaël Marèe, Loïc Rollus, and Renaud Hoyoux for their support in both engineering the Cytomine-IRIS platform, as well as the scientific collaboration when disseminating the results. Thanks to Raphaël for being great host and tourist guide during my visit, and to Natacha Rocks, for introducing me to the blithe and iridescent life of Outremeuse during August 15 celebrations in Liège.

Last but not least, I am very grateful to all the magnificent musicians and their bands, who carried my mind through countless hours of planning, programming, video preparations, thesis writing and self-denial, while many of my friends got married, built houses, got dogs and were blessed with healthy offspring.

Without friends his delicate psyche snapped like snappyest snappersons, the snappyest kid in Snappadelvia, snap snappersons since snaps photography specialists!

Kowalski

The Penguins of Madagascar

Contents

Nomenclature

Abbreviations

ACC	Classification Accuracy
API	Application Programming Interface
AUC	Area under the (precision-recall) curve
BM	Bone Marrow
BM-HE	H&E stained bone marrow dataset (cell detection)
BM-MGG	MGG stained bone marrow dataset (cell detection)
BM-HE-MK	H&E stained bone marrow dataset (cell detection, megakaryocytes)
CART	Classification and Regression Trees
CIE	Commission internationale de l'éclairage (International Commission on Illumination)
CNN	Convolutional Neural Network
cRF	classification Random Forest (cell detection method)
CV	Cross-validation
ICPR-BC	H&E stained breast cancer dataset (cell detection)
ESN	Echo State Network
FN	False Negative
FP	False Positive
FF-NN	Feed-forward Neural Network
F1	F1-score
IHC	Immunohistochemistry
H&E	Hematoxylin&Eosin
HSC	Hematopoietic Stem Cell
HoG	Histogram of Oriented Gradients
LBP	Local Binary Patterns
LoG	Laplacian of Gaussian
LSM	Liquid State Machine
MGG	May-Grünwald-Giemsa
MT-HE	H&E stained dataset of multiple tissues (cell detection)

MRF	Markov Random Field
MSER	Maximally Stable Extremal Regions
N:C ratio	Nucleus-to-Cytoplasm Ratio
NMS	Non-Maximum Suppression
NRMSE	Normalized Root Mean Squared Error
PAS	Periodic Acid Schiff
PRC	Precision
RBC	Red Blood Cells
rRF	*single-target* regression Random Forest (cell detection method)
SD	Standard Deviation
srRF	*spatial-averaging* regression Random Forest (cell detection method)
SPC	Specificity
SVM	Support Vector Machine
SSVM	Structured SVM (classifier)
SSVM	MSER-SSVM (cell detection method)
TN	True Negative
TP	True Positive
REC	Recall
RF	Random Forest
RNN	Recurrent Neural Network
WBC	White Blood Cells
WSI	Whole Slide Image, Whole Slide Imaging (from context)

Symbols and Functions

Part II - Cell Localization

f	classification model
Ω	image domain
I	image
$\mathbf{u}$	image location, 'Pixel'
u_x	x-coordinate of an image location
u_y	y-coordinate of an image location
$\mathcal{C}$	set of annotated ground truth cell centers
$\{\cdot\}$	a set of elements
$\mathbf{c}^{(i)}$	i-th cell center in $\mathcal{C}$
$c_x^{(i)}$	x-coordinate of the i-th cell center
$c_y^{(i)}$	y-coordinate of the i-th cell center
$\mathbf{y}$	proximity score map
$y(\mathbf{u})$	proximity score at an image location $\mathbf{u}$
$\mathfrak{D}_{\mathcal{C}}$	Euclidean distance transform of the set $\mathcal{C}$

g	regression model
d_M	maximum peak width in the proximity score map
α	shape of the peaks in the proximity score map
r_{obj}	(average) object radius
s	scaling factor for α
T	total number of decision trees in a Random Forest
T_t	the t-th decision tree in a Random Forest
g_t	regression model learned by the t-th decision tree
γ	aggregation function over all g_t
$\mathcal{X}$	input space
$\mathcal{Y}$	output space
$I(\mathbf{u})$	local image patch, centered on an image location $\mathbf{u}$
Φ	the set of visual feature channels
$\lvert \cdot \rvert$	cardinality of a set, length of a vector
$\mathcal{P}$	supervised learning dataset
$N_{(\cdot)}$	general symbol for denoting the number of elements in a particular domain
$\mathbf{x}^{(i)}$	i-th input image patch in a learning dataset
$y^{(i)}$	ground truth proximity score (label) of the i-th image patch
p_{in}	size in pixels of a local input image patch
$\ominus$	morphological erosion operator
$\oplus$	morphological dilation operator
$\mathcal{K}$	convolution kernel
∇_d	first order image gradient operator in the d-th dimension
Δ_d	second order image gradient operator (Laplace) in the d-th dimension
μ_G	orientation parameters of Gabor filters
ν_G	scale parameters of Gabor filters
τ_{bg}	'foreground-background' threshold for dataset sampling
N_{τ^+}	total number of foreground patches
N_{τ^-}	total number of foreground patches
ϕ_j	split function at non-terminal node j of a decision tree
θ_j	selection function at non-terminal node j of a decision tree
τ_ϕ	threshold of a split function
N_θ	total number of tested selection functions
N_j	total number of samples at node j
M_j	number of randomly drawn samples to evaluate a split function at node j
$Var(\cdot)$	variance of a numeric set
$RV(\phi)$	reduction in variance according to a split function ϕ
$\mathcal{P}_f(\phi)$	the s-th subset of $\mathcal{P}$ according to a split using ϕ
ω_s	weight of the s-th subset
ϕ^*	optimal split function
T_{md}	maximum depth of a tree
$\hat{y}_t$	proximity score predicted by the t-th tree
$\hat{y}$	final proximity score predicted by the Random Forest

$\hat{\mathbf{y}}$	predicted proximity score map (for a local patch, or an entire image)
W_S	pixel stride of the sliding window
p_{out}	output patch size (spatial-averaging regression)
O	output dimension (spatial-averaging regression)
$\mathbf{y}^{(i)}$	ground truth label (vector) of the i-th training sample (spatial-averaging regression)
V	total number of output patches containing a particular image location
$\Delta\mathbf{u}$	offset vector relative to a center location $\mathbf{u}$
$\bar{\mathbf{y}}_{\mathcal{P}}$	average proximity score vector of a dataset $\mathcal{P}$
$\hat{\mathbf{y}}_t$	proximity score vector predicted by the t-th tree
$\hat{\mathbf{y}}^{(v)}(\mathbf{u})$	the v-th output patch containing location $\mathbf{u}$
$\hat{y}^{(v)}(\mathbf{u})$	proximity score at location $\mathbf{u}$ in the v-th output patch
$\hat{y}(\mathbf{u})$	final proximity score at location $\mathbf{u}$
$\mathcal{G}$	Gaussian filter
κ	proximity score threshold (regression), foreground probability threshold (classification)
W_{NMS}	size of the non-maximum suppression window
r_{NMS}	radius of the non-maximum suppression window
$\mathbf{q}^{(i)}$	nearest neighbor of a ground truth cell center
$\mathrm{d_E}$	Euclidean distance function
$\mu_{\mathrm{d_E}}$	mean of the Euclidean distance function
$\sigma_{\mathrm{d_E}}$	standard deviation of the Euclidean distance function
$n_{r_{obj}}$	total number of cells used to estimate r_{obj}
μ	arithmetic mean (from context)
σ	standard deviation from the arithmetic mean (from context)
μ_d	mean of the distances d between a true positive detection and its associated ground truth
σ_d	standard deviation of the distances d between a true positive detection and its associated ground truth
μ_n	mean of the absolute difference between the number of all true positive detections and the number of all ground truth cell centers
σ_n	standard deviation of the absolute difference between the number of all true positive detections and the number of all ground truth cell centers
ξ	evaluation parameter, maximum distance between ground truth and hypothesized center to be counted as true positive
$k^{(i)}$	ground truth class label of the i-th training sample
$IG(\phi)$	information gain according to a split function ϕ
$\mathcal{H}(\mathcal{P})$	Shannon entropy of a set $\mathcal{P}$
$p(c=k\|\mathcal{P})$	class label probability for class k. given a set $\mathcal{P}$
$Gini(\mathcal{P})$	Gini impurity of a set $\mathcal{P}$
$GG(\phi)$	Gini gain according to a split function ϕ
$p_t(c=k\|I(\mathbf{u}))$	class label probability for class k predicted by the t-th tree, given an image patch

$\bar{p}(c = k\|I(\mathbf{u}))$	class label probability for class k predicted by the Random Forest, given an image patch
$\hat{\mathbf{p}}$	probability map returned by a classification Random Forest
$\mathbf{f}$	feature vector for the *SSVM* method
$\mathbf{w}_b$	weight vector of the binary SVM
$\mathbf{w}_s$	weight vector of the SSVM
b	inference bias of the SSVM
Λ	hyper-parameter space
$\lambda^{(l)}$	the l-th vector of hyper-parameters
$\lambda_k^{(l)}$	the k-th entry in the l-th hyper-parameter vector
$\overline{\mathrm{AUC}}$	average AUC computed from pooled TP, FP, and FN detections
$\overline{\mathrm{F1}}$	average F1-score computed as mean over all N_I images in a dataset
$\mathcal{O}\left(\lambda^{(l)}\right)$	optimality criterion for stage III hyper-parameter search
$\lfloor\cdot\rceil$	rounding operator to the nearest integer
p_{in}^*	optimal input patch size
p_{out}^*	optimal output patch size (spatial-averaging regression)
κ^*	optimal proximity score threshold (regression), optimal foreground probability threshold (classification), resulting in the highest F1-score
$\sigma_{\mathcal{G}}$	standard deviation of the zero-mean Gaussian filter $\mathcal{G}$

Part III - Cell Classification

K	total number of readout units (classes)
L	total number of input units to the Echo State Network
N	total number of reservoir units
$\mathbf{W}$	recurrent reservoir weights
$\mathbf{W}^{in}$	input weights
$\rho(\mathbf{W})$	spectral radius of reservoir weight matrix
$\mathbf{W}^{out}$	readout weights
$\mathbf{u}(t)$	input to the ESN at time t
$\mathbf{x}(t)$	reservoir state at time t
α	reservoir leaking rate
$\mathbf{X}$	collection of temporal reservoir states
$y_k(t)$	regression target value for class k at time t
$\hat{\mathbf{Y}}_k$	predicted temporal output for class k
Υ	inference period
$\bar{y}_k$	average output for class k in the inference period
$\hat{k}$	predicted class label for a cell
$I(\mathbf{c})$	local image patch, centered on location $\mathbf{c}$
φ_0	starting angle, original orientation of a cell
φ	(counter-clockwise) rotation angle
$\Delta\varphi$	number of skipped rotations
$I(\mathbf{c}, \varphi)$	local image patch, rotated centrically by an angle φ
$I(\mathbf{c}, \varphi + \Delta\varphi)$	local image patch, rotated centrically by an offset $\Delta\varphi$ to the current angle φ

List of Figures

List of Tables

Part I.

Background and Material

1. Introduction

Imaging tissue structures has become an integral part of state-of-the-art diagnostics in medicine and biology. Modern biomedical diagnostics has benefited tremendously from technological advances in both anatomical and functional imaging in the past decade. Image resolutions are continuously increasing for volumetric imaging modalities such as magnetic resonance imaging (MRI) or computed tomography (CT), and microscopic whole slide imaging (WSI), allowing important visual examinations of organ and tissue structures at a highly detailed level. In addition, several advanced imaging modalities for both *in-vivo* and *in-vitro* imaging were developed, enabling new insights, but also posing new challenges to data analysis.

Digital pathology has emerged as an active field of research [1]. Telepathology, virtual microscopy and other sub-disciplines heavily rely on robust digital image processing methods and information technology. Every day, hundreds of glass slides frequently have to be processed manually [2], which is a tedious and error-prone activity. As WSI has been well accepted as an alternative to the conventional glass slides, lots of new means in the slide analysis process were revealed [3, 4]. Digital slides are used in a variety of applications like digital diagnostics, research, education, or digital archiving [3–5]. The information content with respect to cytological and histological tissue properties captured in whole slide images is enormous. Due to their quantity and complexity, analysis of entire whole slide images by human experts is cumbersome and calls for automated quantitative solutions.

1.1. Motivation and Problem Definition

Hematopoiesis is the physiological process of blood cell maturation in the bone marrow. The examination of human bone marrow is essential in hematopathology, and is performed on clinical indication such as suspected blood disorders. On average, a pathologist spends considerable time visually, hence qualitatively, analyzing tissue samples, where two main tasks need to be performed. First, the identification of individual cells in a tissue specimen, i.e. cell localization and counting, and secondly the recognition of its type. Since cell maturation is a continuous process, determining distinct stages is sometimes infeasible, even with existing guidelines and significant experience of the observers. A well known, and one of the most prominent problems of the qualitative analysis is the lack of reproducibility and comparability among different observers (intra- and inter-observer reliability). Despite the human brain is capable of very impressive performance on a variety of cognitive tasks where machines are simply not yet intelligent enough, counting hundreds of individual cells of multiple types is still challenging in many ways. In current routine diagnosis, histomorphometric analysis of the cellular bone marrow components is increasingly done semi-quantitatively in digitized glass slides (virtual slides), using basic image processing algorithms. In fact, such repetitive and monotonic tasks hold great potential for successful automatization. Advanced computer-assisted methods for automated tissue analysis have been subject of research over the past years, but very few concepts have reached clinical practice yet.

Automated tissue analysis methods based on intelligent and adaptive algorithms can help to increase the quality of the diagnostic process by speeding up image analysis while reducing the observer variability. Research in image processing, pattern recognition, computer vision and machine learning has resulted in a rich set of tools that can be assembled to robust systems capable of learning specific tasks such as recognizing objects from example images. Once trained, such systems can be applied to previously unseen images to infer information. Further, they can be expected to exhibit a lower and more predictable variability than humans. The ability to scale with increasing demands can truly be seen as a significant advantage, also with respect to the tremendous mass of imaging data being produced every day in clinical practice.

1.2. Hypotheses and Goals

Given the motivation and problem statement in the previous sections, the following hypotheses are examined in this work:

1. Machine learning based computer vision algorithms are able to robustly localize cells in histopathology images of human bone marrow tissue.
2. The discrimination of a variety of bone marrow cell lineages in different stages of maturation can be performed satisfactorily by machine learning algorithms directly from digital image data.
3. In both cell localization and recognition, quantitative learning-based algorithms yield highly accurate and reproducible results, and have both an acceptable and a predictable error.

The overall goal of this work is to examine the applicability of object localization and recognition algorithms that learn from example images to support the automated analysis of bone marrow tissue samples at the cellular level. Here, a machine learning approach is proposed that learns to estimate the distribution of bone marrow cell classes in a virtual slide. Since each slide may contain thousands of instances of different cell classes, the primary goal in automated analysis is the estimation of the cell class distribution. A secondary goal is finding the relative location of different cell types in context of the tissue architecture. These problems can be solved by either pursuing an integrated solution that directly copes with both localization and classification at once, or splitting the problem into two distinct parts. Both strategies have benefits and disadvantages, but a multi-stage approach is considered sufficient here. Hence, cell-like objects are first localized as candidates in stained histopathological images of healthy human bone marrow. Subsequently, these candidates are subject to classification into different stages of maturation in erythropoiesis, granulopoiesis, and megakaryopoiesis. Pursuing this approach facilitates omitting the very challenging and probably error-prone task of accurately *segmenting* multiple cell nuclei within the image.

1.3. Scientific Challenges and Potential

In digital pathology, the development of standardized methods recognizing morphological tissue properties would support the selection of the appropriate immunohistochem-

ical markers in the diagnostic process. Furthermore, it would be of highest value in diagnostic decision making, where particular markers are not established, for instance, reliably detecting atypical characteristics in early myelodysplasia. The universal approach of the present study is qualified to be adapted to many other diagnostic questions, especially when reactive changes must be discriminated from early neoplasia. The close cooperation between pathology and quantitative image analysis significantly enhances the emerging field of computer-aided pathology. Using the findings of this thesis, we will be able to show that the current manual standard method of cell quantification in complex tissue such as bone marrow can be improved with respect to a statistically predictable variance.

The problem at hand is scientifically challenging due to its interdisciplinary character. It requires expertise in computer science, especially machine learning-based computer vision, software engineering, basic processes in biology as well as image-based histopathological diagnostics. Each of these fields is rapidly progressing, and especially – but not exclusively – in computer science and medicine, the number of scientific publications per year is increasing exponentially, which poses a significant challenge to keep pace with all novel developments and consider state-of-the-art research findings. Hence, not all developments that happened in parallel to the compilation of this thesis could be fully considered. However, the connection between the proposed methods and the recent advances will be discussed adequately.

1.4. Original Contributions

In this thesis, several original contributions will be made. Firstly, a novel bone marrow cell localization algorithm is developed and implemented based on randomized decision trees. Its superiority over a current state-of-the-art and baseline method is demonstrated on multiple challenging histopathological image datasets. In this context, a novel benchmark dataset of eleven 1200×1200 pixel images depicting healthy, Hematoxylin-Eosin (H&E) stained, bone marrow tissue was publicly released [6]. This dataset is fully annotated and contains $4,205$ labels covering *foreground* or *unknown* objects.

Secondly, an alternative approach to conventional cell classification is presented to quantify the hematopoietic cell maturation. A Recurrent Neural Network (RNN) is trained to distinguish multiple blood cell precursors with very high accuracy. Due to a

novel training scheme, the network is learning rotationally covariant features directly from raw image data. The temporal features computed by the RNN coherently encode appearance information of the cell, which can be learned in a globally optimal fashion by minimizing a least squares objective and yield a rotation-invariant classifier.

Finally, following the urgent need for reliable ground truth data to apply supervised machine learning in biomedical imaging, an intuitive tool for conducting inter-observer reliability studies has been developed as part of the Cytomine [7] project. It is based on state-of-the-art web-application frameworks and technologies, has been released under a permissive open-source license and is publicly available[1]. Results of a first application in collaborative ground truth data labeling in whole slide images will be briefly reported in the final part of this thesis.

1.5. Organization of this Thesis

In this chapter, the requirements for automated image analysis of bone marrow tissue have been stated. Potentials of machine learning methods have been introduced alongside the research questions. The goals to be achieved and a brief overview of the pursued approach were outlined. The remainder of this thesis is organized in three major parts that reflect the proposed approach. In Chapter 2, the cytological and histological properties of healthy bone marrow tissue are described. Part II covers the localization of cell nuclei in histopathological images and the evaluation of a novel learning-based cell detection algorithm on different datasets, including publicly available benchmark datasets[2]. Part III is focused on hematopoietic cell classification using RNN and empirical evaluations on carefully selected available datasets[3]. Finally, Part IV covers attempts towards reliable ground truth data[4], a discussion of both localization and classification of bone marrow cells, conclusions drawn from the results of this thesis as well as links to previous work, and potential directions for future research.

[1] Available from https://github.com/Cytomine/Cytomine-IRIS.
[2] An excerpt of Part II has been published in [6, 8].
[3] Excerpts of Part III have been published in [9–11], or were submitted for publication [12].
[4] Parts of Chapter 11 have been published in [7].

2. Bone Marrow Tissue

This chapter briefly introduces the biological background of bone marrow tissue. The basic anatomical structure and functions of hematopoietic tissue are described first, laying out the basics to understand the hematopoiesis, i.e. the process of blood cell maturation, in bone marrow. Then, different procedures to harvest tissue and related tasks in histopathological analysis are characterized. A vast amount of research has been attributed to the hematopoietic system, driven by interests in both its physiology and its pathologies. Further, due to their unique characteristics, the origin and versatility of hematopoietic stem cells (HSCs) are subject of current research for new treatments of blood disorders.

With respect to the aims of this thesis, i.e. the analysis of bone marrow histopathology images, the focus in this chapter is set on the characteristics and maturation of precursor cells of erythrocytes, granulocytes, and megakaryocytes, because these cells can be observed in images of stained bone marrow obtained via light microscopy [13]. Further, as opposed to HSCs, cells in these stages of maturation can be discriminated by their morphology, an essential cue for computer-assisted automated visual tissue analysis.

2.1. The Hematopoietic System

Anatomically, bones consist of periosteum, hard outer cortical bone, endosteum, and medulla (bone marrow) in the hollow interior [14], cf. Fig. 2.1 (a). Bone marrow occupies the interstitial space provided by the medullary cavity and is categorized as either red or yellow marrow, both of which are highly vascular [15]. All types of blood cells can be produced in red marrow, which is also referred to as the cellular bone marrow, or hematopoietic tissue. Yellow marrow mainly contains (inactive) adipose tissue.

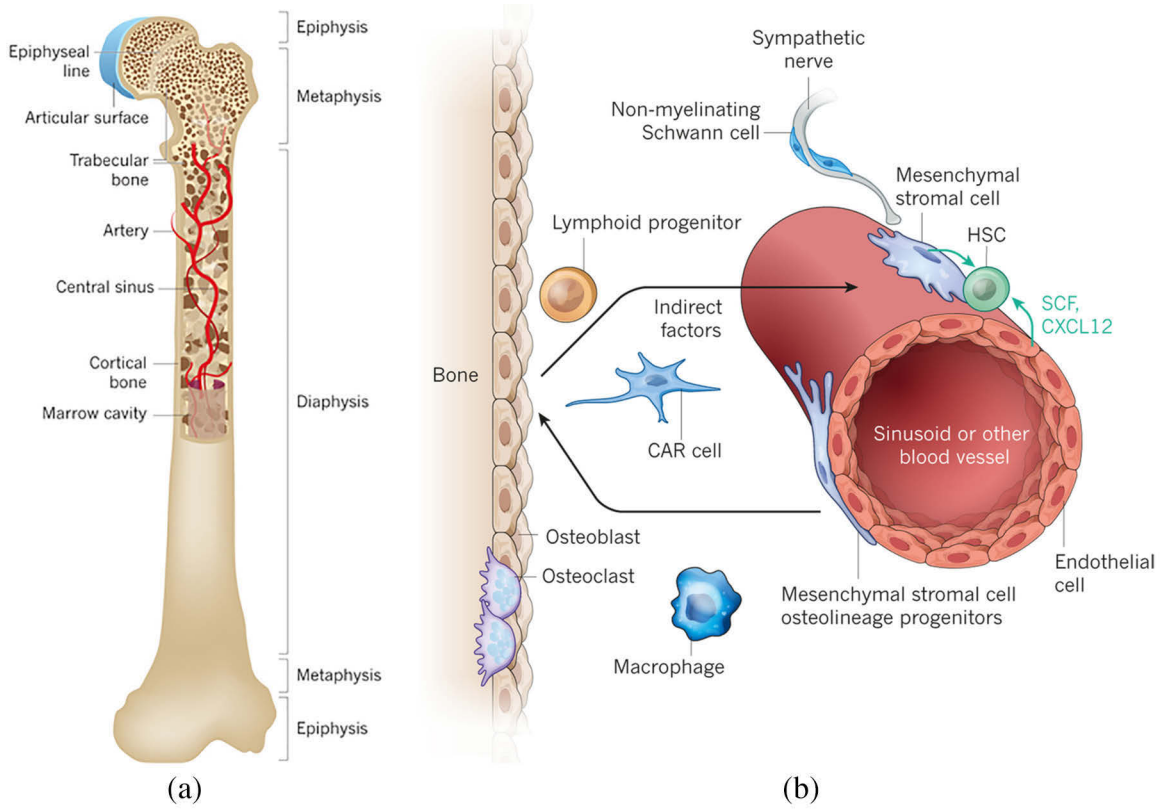


Figure 2.1.: Anatomy of bone marrow. (a) In adults, HSCs mainly reside within marrow of long bones that occupies the medullary cavity. Oxygen and nutrients are transported to the bone cells via arteries and arterioles. (b) Specialized venules (sinusoids) serve as gateway for HSCs and mature cells to the circulatory system.

Adapted by permission from Macmillan Publishers Ltd: Nature. Morrison SJ, Scadden DT. The bone marrow niche for haematopoietic stem cells. Nature. 2014 Jan;505(7483):327–334, copyright 2014.

Hematopoiesis is the physiological process of blood cell formation from HSCs that takes place in and is controlled by special micro-environments (niches) [16, 18–20]. HSCs are multipotent stem cells with the unique ability of highly frequent self-renewal [21, 22] that can differentiate into any blood cell type. Located in the bone marrow niches, they are sustained and regulated by a multitude of different cell types with very specific roles [23]. The concept of niches has been accepted as a model [24]. However, their concrete function to maintain HSC pools is still under active investigation [25]. Conducting this research is non-trivial, since retaining the spatial structure is difficult during histological preparation and applying proper staining methods [16]. A categorization into endosteal and vascular (sinusoidal) niches has been proposed [24, 26, 27], but the accurate spatial locations of HSCs in the bone marrow was not entirely determined yet. Nevertheless, results of previous studies coincided that the HSCs are perivascular and

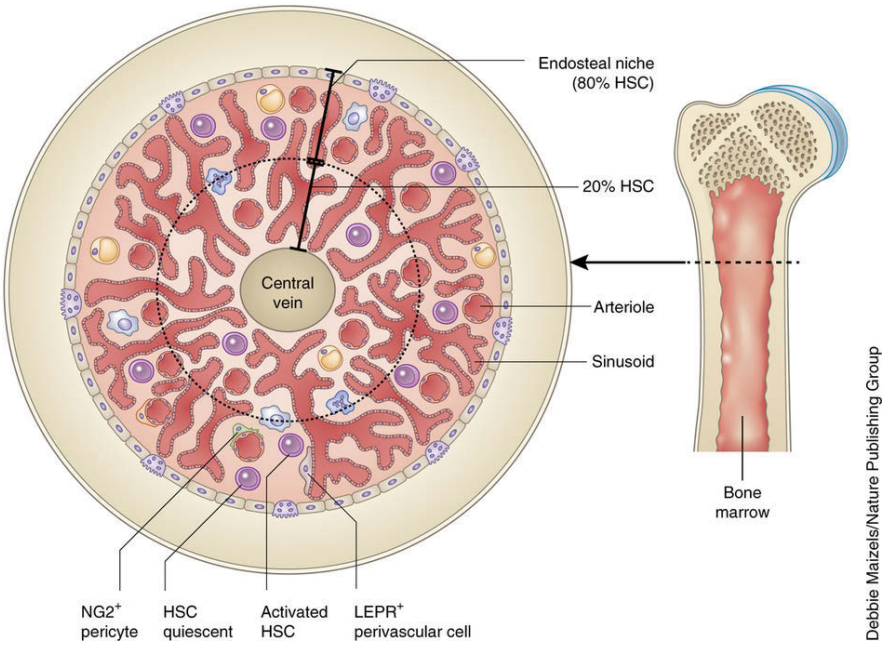


Figure 2.2.: HSCs seem to prefer endosteal parts closer to the bone surface (80%). Only around 20% of all HSCs are located within half distance from endosteum to the central vein [17].

Reprinted by permission from Macmillan Publishers Ltd: Nature Medicine. Mendelson A, Frenette PS. Hematopoietic stem cell niche maintenance during homeostasis and regeneration. Nature Medicine. 2014 Aug;20(8):833–846, copyright 2014.

prefer the highly vascular endosteal regions [17, 28], cf. Fig. 2.2. Being exposed to appropriate external stimuli, they can also differentiate into other specialized cells such as cardiomyocytes, or endothelial cells [29].

During embryonic and fetal development, HSCs are developed at multiple anatomical sites [30] and migrate from there to the skeletal bones, which postnatally become the primary hematopoietic organ in the body [13, 16, 31]. Until puberty, bone marrow in skeletal bones (scull, vertebrae, pelvis, femur, etc.) predominantly consists of red marrow, while red and yellow marrow becomes of almost equal quantity in adulthood. Due to age-dependent resorption of cancellous[1] bone, more space is available in the medulla that gets populated by yellow marrow. In response to physiological stress, red marrow may quickly expand locally to satisfy the demands for increased blood cell formation [27]. In addition, peripheral yellow marrow can also be reactivated to form blood cells [13]. However, hematopoiesis in adults is usually confined to the bone marrow, whereas extramedullary hematopoiesis with some exceptions [14, 32] is likely an indicator for malignant disorders [13, 33].

[1] Cancellous bone is a sponge-like bone structure, as opposed to dense cortical bone.

The hematopoietic cells are embedded in stroma, an essential part in the medulla that indirectly influences hematopoiesis [25] and consists of a dense network of blood vessels, nerve fibers, fat cells, macrophages, lymphocytes, and fibroblasts [28]. The stroma further contains multipotent mesenchymal stem cells (MSCs, skeletal stem cells) [28] that are able to differentiate into a diverse set of cell types, such as endothelial cells [13]. Hematopoietic cells are contained in special fibers (cords) between thin-walled vessels (sinusoids) [13, 14, 34]. Sinusoid walls are composed of a single layer of endothelial cells, a basal membrane and reticular adventitial cells. An important property of adventitial cells is that they support anchoring hematopoietic cells and stroma [13]. At the end of their maturation process, differentiated cells egress bone marrow into the circulatory system by surpassing the endothelial layer of the sinusoids, cf. Fig. 2.1 (b).

Despite the fact that a common model of the human hematopoietic system is actively discussed in recent literature [13, 35, 36], for our purpose it is sufficient to adhere to the classical model. The hematopoietic system consists of a lymphoid and myeloid component, which generate red blood cells (RBCs, erythrocytes), white blood cells (WBCs, leukocytes), and platelets [13]. Fig. 2.3 illustrates a coarse overview of the classical hematopoietic hierarchy. The reader is further referred to *Rubin's Pathology* [13, p. 1082] for a detailed schematic illustration of the widely accepted hierarchical model of the hematopoietic system.

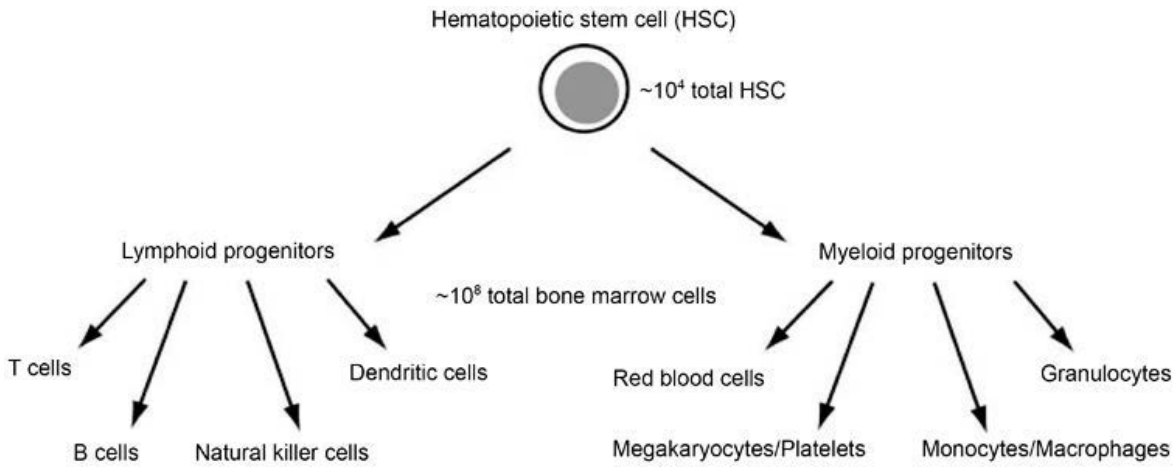


Figure 2.3.: Hierarchy of the hematopoietic system according to the widely accepted classical model. The two main components of the hematopoietic system are lymphoid and myeloid branch, leading to a different set of differentiated red and white blood cells as well as platelets.

Reprinted by permission from Macmillan Publishers Ltd: Cell Research. Nemeth MJ, Bodine DM. Regulation of hematopoiesis and the hematopoietic stem cell niche by Wnt signaling pathways. Cell Research. 2007 Sep;17(9):746–758, copyright 2007.

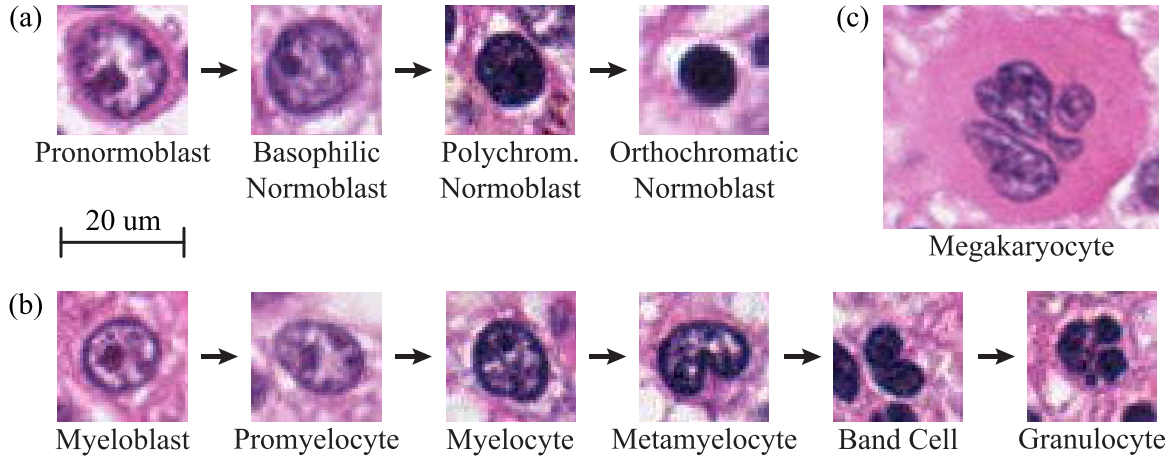


Figure 2.4.: Examples of maturation stages in erythroid and myeloid cell lineages, Hematoxylin-Eosin stained tissue obtained by trephine biopsy, scanned at 40× magnification. (a) Erythropoietic, (b) (neutrophilic) granulopoietic maturation stages, and (c) a megakaryocyte. Mature red blood cells (reticulocytes and erythocytes) do not have nuclei, hence cannot be identified well in this nuclei-specific staining. Moreover they are rarely present in healthy bone marrow.

The maturation of blood cells is a hierarchical, continuous process. Any blood cell origins a (pluripotent) mono-nuclear HSC and differentiates into a multipotent HSC, before becoming a progenitor cell. Despite they lose the ability of self-renewal when differentiating into a progenitor cell, they give rise to common lymphoid and myeloid progenitor cells [29]. Myeloid progenitors then advance to erythrocyte/megakaryocyte, granulocyte/macrophage, and dendritic cell progenitors [27]. On the other hand, lymphoid progenitors eventually mature to lymphocytes outside the bone marrow [13]. Subsequently, the progenitors differentiate into lineage-committed precursors (blasts) of a mature cell. Different lineages of blood cells can be defined, depending on the level in the hierarchy [38]. Here, we distinguish between the erythroid, myeloid, and megakaryocytic lineage. Please note that the myeloid lineage produces myeloid leukocytes, i.e. granulocytes and monocytes, where the latter can further differentiate into macrophages in case of tissue damage or infection [39], or dendritic[2] cells [40]. Nevertheless, we will briefly characterize cytology and histology of erythropoiesis, granulopoiesis, and megakaryopoiesis in the following sections, as they are most relevant in the current context.

[2] For instance, Langerhans cells.

Erythropoiesis Erythropoiesis describes the differentiation of an early erythroid precursor cell into a mature erythrocyte (RBC). The first stage of maturation that can be recognized by morphological features is termed pronormoblasts[3]. Pronormoblasts are around 12-20 µm in diameter and are characterized by a large nucleus, i.e. a high nucleus-to-cytoplasm (N:C) ratio, and a pale perinuclear area. They advance to basophilic normoblasts and polychromatophilic normoblasts and finally to orthochromatic normoblasts. With progressing cell maturity, both N:C ratio and size decreases [40]. Orthochromatic normoblasts lose their nuclei, hence their proliferation ability, and advance to reticulocytes. Reticulocytes are immature RBCs and represent the last stage in this cell lineage that reside within the bone marrow [13]. They are immediate predecessors of functional erythrocytes that leave the bone marrow via the sinusoids. Normoblasts can be found as erythroblastic islands, i.e. clusters of concentric circles of cells around a macrophage, which are located closely to the sinusoids [13]. Cells of later maturation stages are more frequent in healthy bone marrow than earlier stages, cf. Figs. 2.4 (a) and 2.5 (a). With the exception of newborns, only enucleated reticulocytes and erythrocytes enter the blood stream. Hence, these two last maturation stages are infrequently discovered in bone marrow.

Granulopoiesis Granulopoiesis is the physiological process of maturation from granulocyte/monocyte precursors to granulocytes. The N:C ratio increases, while the shape of the nucleus changes from circular/oval to kidney-like, and segmented forms [40], cf. Fig. 2.4 (b). The myeloblast is the first recognizable form and is similar in size to the proerythroblast (12-20 µm), but rather irregular in shape [14]. Myeloblasts are able to divide and advance to promyelocytes, which are slightly larger in size. Promyleocytes can further differentiate into three sub-lineages of granulocytes: neutrophilic, eosinophilic, and basophilic, cf. Fig. 2.5. Each of these series continues maturation in three stages as myelocytes, metamyelocytes and band cells, which share many morphological features. However, they are smaller than promyelocytes, but vary in size (10-20 µm) and their nuclei do not have nucleoli. The three sub-lineages now become distinguishable, because they develop specific granules that are visible as different colors in staining. Myelocytes differentiate into metamyelocytes, that can be recognized by their characteristically U-shaped nuclei. At this stage the cells are not able to divide any more, but mature to the end-stage forms in granulocytic series: band and segmented cells. Band cells do not develop filaments among their bilobed nuclei, while segmented

[3] Other commonly used terms for the maturation stages in erythropoiesis can be found in the work of Naeim *et al.* [40].

cells are charcterized by multiple nuclei segments connected by filaments [40]. Granulocytic precursors are located closer to the trabeculae farther away from sinusoids, but once cells reach metamyelocyte stage, they move towards sinusoids [14]. At the stage of band cells, they usually pass the endothelial layer of the sinusoids and enter the circulation. In healthy adults, the majority of cells (21-40%) in bone marrow are neutrophilic band cells, cf. Fig. 2.5. The number of cells in the eosinophilic and basophilic lineage is quite low in comparison.

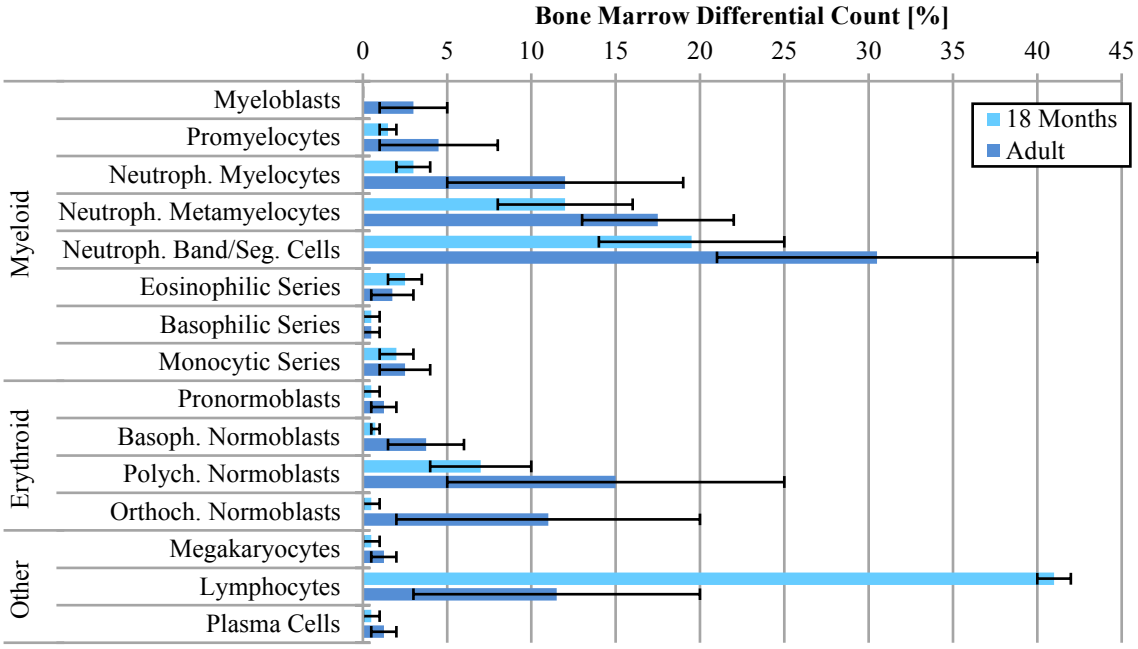
Megakaryopoiesis and Thrombopoiesis Originating from the common erythrocyte/megakaryocyte progenitor cell, megakaryoblasts are the first form that can be identified based on morphological features [40] in bone marrow[4]. It is the only precursor of a megakaryocyte. During maturation, megakaryocytes increase in size and become large multilobed cells, cf. Fig. 2.4 (c). Having reached their final form in bone marrow (group III, or granular megakaryocytes [40]), their cytosol is being released into the blood stream as ribbons containing platelets [13]. Megakaryocytes are the largest cells in normal bone marrow (30-160 µm). Similar to erythroblastic islands, megakaryocytes are located closely to the sinusoids [13], but usually do not form clusters [14]. The differential count of megakaryocytes is between 0.5 and 2% in healthy adults, cf. Fig. 2.5 (b).

The distribution of hematopoietic cells in healthy bone marrow is illustrated as differential blood count (DBC) at the age of 18 months and in adulthood in Fig. 2.5. In adults, myeloid cells constitute the majority (around two thirds) of all cells, with a myeloid-to-erythroid ratio of 3-3.5:1 [40]. Bone marrow serves as an important repository for mature neutrophilic granulocytes.

2.2. Histopathological Analysis

Bone marrow is assessed in terms of cellularity (the overall area of a specimen covered by cellular marrow components), the tissue architecture, and the distribution of hematopoietic maturation stages. Usually, bone marrow is examined in case of suspected blood-related disorders [41–43], and obtaining the marrow is connected to an invasive procedure. Common malignant blood disorders that can be diagnosed directly

[4] Other researchers [14] argue, that the first form is the megakaryocyte itself, and that megakaryoblasts can be recognized in the presence of abnormal hematopoiesis.



(a)

Cell Type	18 Months	Adult	Cell Type	18 Months	Adult
Myeloid			**Erythroid**		
Myeloblasts	–	1-5	Pronormoblasts	<1	0.5-2
Promyelocytes	1-2	1-8	Basoph. Normoblasts	0.5-1	1.5-6
Neutrophilic Series			Polych. Normoblast	4-10	5-25
Myelocytes	2-4	5-19	Orthoch. Normoblasts	<1	2-20
Metamyelocytes	8-16	13-22	**Other**		
Band/Segmented Cells	14-25	21-40	Megakaryocytes	<1	0.5-2
Eosinophilic Series	1.5-3.5	0.5-3	Lymphocytes	40-42	3-20
Basophilic Series	<1	<1	Plasma Cells	<1	0.5-2
Monocytic Series	1-3	1-4	**Myel.-Ery. Ratio**	4-5:1	3-3.5:1

(b)

Figure 2.5.: Bone marrow differential count in percent in healthy persons [40]. (a) Compared to a person at the age of 18 months, the distribution of hematopoietic cells is shifting in adults: Lymphopoiesis decreases, while myeloid and erythroid lineages are more present. Error bars denote the minimum and maximum of the differential counts in (b). The number of myleoid cells is higher than erythroid cells, both at the age of 18 months and in adults, denoted by *Myel.-Ery. Ratio* in (b). Bone marrow serves as reservoir for mature neutrophilic granulocytes.

from bone marrow specimen comprise myelodysplastic syndromes [44], and myeloid leukemia [45]. On the other hand, e.g. in case of anemia (a low number of RBC) bone marrow examination is additionally performed to confirm diagnosis. Marrow specimens

for comprehensive cytological and histological examinations are usually obtained via core and aspiration biopsies, frequently performed in conjunction to report a complete status [14, 46–49]. Other methods include open biopsy (under general anesthesia) and autopsy. For the core biopsy procedure, a drill, or a trephine needle is used to obtain the samples from the anterior or posterior iliac crest, whereas for aspiration biopsies, the sternum can alternatively be used as marrow source[5].

From aspirated material, basically three preparations are possible [40]: particle clot, smear, and touch preparation. Touch preparation is not well suited for morphological assessment due to significant procedure-conditioned artifacts. Smear preparations are used to examine the DBC, cellular details, stages of maturation as well as the myeloid-to-erythroid ratio. Finally, particle clot sections are preparations from aspirated material, but are processed similar to trephine biopsies without the decalcification procedure. They are used to assess the cellularity, megakaryocyte morphology and tumor infiltration and frequently complement trephine biopsies [50].

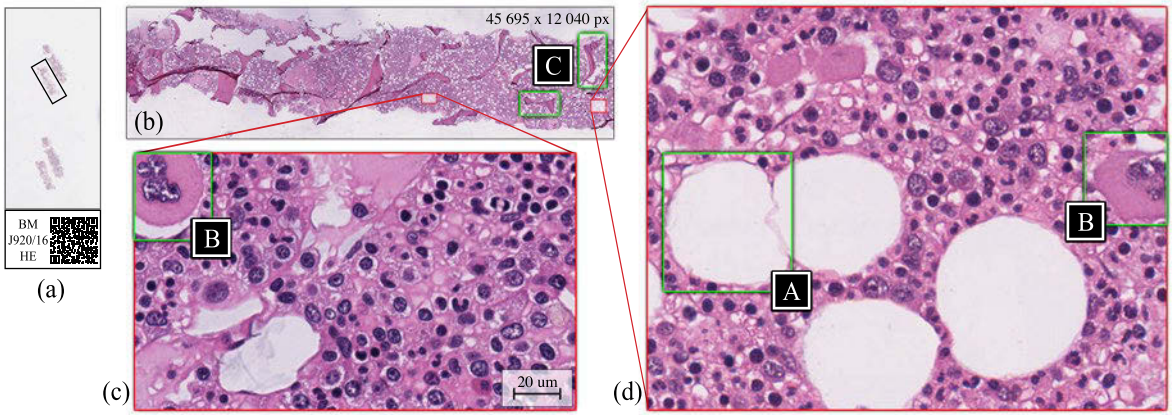


Figure 2.6.: Healthy bone marrow trephine biopsy sample stained with Hematoxylin-Eosin, mounted on a glass slide (a). (b) Digital image obtained from (a) by whole slide imaging at 40× magnification. (c,d) Red highlighted areas from (b) showing tissue regions at full resolution. Green highlighted areas show (A) fat cells, (B) megakaryocytes, and (C) trabecular bones.

Samples of a histopathological image used for bone marrow analysis are shown in Fig. 2.6. Despite cellularity can be assessed in all preparations from aspirated material, only trephine biopsies preserve the tissue architecture and hence include rigid anatomical parts such as trabecular bones as well, cf. Fig. 2.6 (b). The biological specimen shown on the glass slide in Fig. 2.6 (a) was obtained by trephine biopsy, embedded in acrylate and stained with H&E before it was digitized at 40× magnification, cf. Fig. 2.6

[5] In [14, 46, 47], authors describe advantages and disadvantages of both methods in more detail.

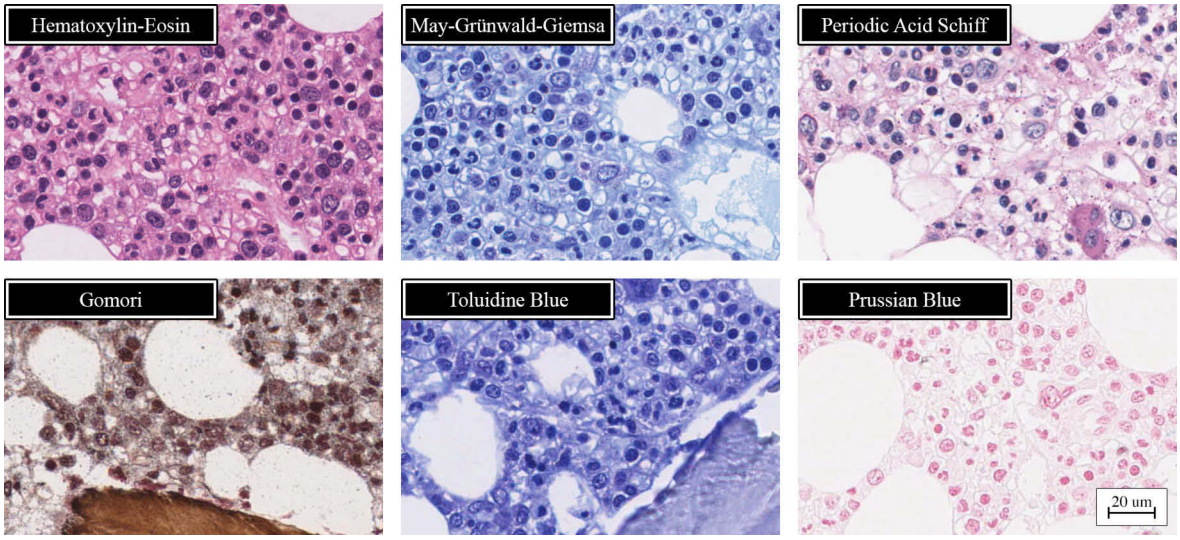


Figure 2.7.: Healthy bone marrow trephine biopsy specimen stained using different staining protocols. The images show the tissue at 40× magnification. Hematoxylin-Eosin, and May-Grünwald-Giemsa are two standard stainings used in routine histopathology of bone marrow. Prussian Blue is usually not used in routine diagnostics of trephine bone marrow specimen [51], but to specifically stain iron storage [52] blue. Gomori is used to stain reticulin fibers [53], Toluidine Blue for staining mast cells (closely related to basophilic granulocytes), and Periodic Acid Schiff for assessing the cellularity as an alternative to H&E [40].

(b). Due to the histological preparation procedure, which we will not elaborate on here, fat cells appear as white blob-like structures in these images.

The standard stainings in routine diagnosis of core biopsies are Hematoxylin-Eosin, May-Grünwald-Giemsa (MGG), or alternatively Periodic Acid Schiff (PAS), cf. Fig. 2.7. Aspirated material is frequently stained with Wright's stain or Giemsa [40]. Depending on the ultimate goal of cytological or histological analysis, general tissue stainings such as the ones mentioned earlier are used. Alternatively, e.g. to confirm suspected diagnoses, specific immunohistochemical (IHC, antibody-antigen-based) stains highlight particular tissue components, or even individual hematopoietic cell lineages. Morphological features of individual cell nuclei and cytoplasm are well-depicted when using standard stainings such as H&E, which can be performed rather quickly compared to the more time-consuming IHC procedures.

Part II.

Localization of Bone Marrow Cells

3. Learning to Detect Cells

3.1. Introduction

Histopathological analysis of biopsy specimen is very common in modern cell biology and medicine. It frequently relies on histomorphometry of individual cells, or larger structures formed by them, to assess the tissue and detect possible pathological alterations. The biological specimen is usually stained to visualize the tissue structure (nuclei, cytoplasm, mitotic cells, etc.) or metabolism. The examination of such biopsy sections is still done manually by experts using light microscopy, or virtual slides, which are obtained by digitizing glass slides using whole slide imaging.

One of the basic questions, though essential, in tissue analysis is determining the cellularity of a particular specimen, i.e. to count the number of cells that are present. Counting and recognizing individual cells usually works better at higher magnification, where morphological features can be observed in greater detail. However, a major disadvantage when examining large specimen is the limited field of view at higher resolutions. Depending on the tissue type, not every part of the specimen is equally important for estimating the cellularity. For instance, some regions in bone marrow specimen do not contain any cells but in fact show trabecular bone, or adipose cells. On the other hand, epithelial tissue may be composed of many layers of cells, where every part is relevant. In current practice, the field of view is changed to the regions of a specimen that seem particularly relevant to the (human) observer, hence counting is very subjective. Despite the very accurate number of cells in a whole slide image may sometimes be seen uncritical, manually counting cells in larger specimen quickly becomes cumbersome and is thus prone to errors and omissions. Given the typically huge number of cells ($\approx$ 3500-4000/mm^2) contained in histological specimen, the qualitative

visual analysis can lead to considerable inter-observer variability and irreproducible results due to intra-observer variability [54, 55].

Computer-assisted methods for automatic cell detection do have the potential to overcome these limitations and are therefore highly desirable. In light of the aim of assessing the distribution of different cell types in a digital histopathological image, the detection of individual cells, more precisely, the reliable localization of individual cell nuclei is a key requirement. Hence, this problem can be formulated as a general object localization problem, which we approach using supervised machine learning techniques. Despite a lot of effort has been invested to solve this challenging computer vision problem, novel methods are still required to cope with the huge variability of histopathological images [56].

3.1.1. Related Work

Automatic cell detection and segmentation methods have seen much research effort in the previous decades, and are still an active field of research in medical image analysis [56–58]. Detection of individual objects in histopathological slides has been an essential ingredient for several applications such as mitosis detection [59–66], cancer diagnosis and grading [67–71], high-throughput brightfield microscopy [72–74], or detection of myelodysplastic syndromes in bone marrow [75]. Detection and segmentation of blood cells was researched in peripheral blood [76–78], and in bone marrow [79, 80], mostly with a focus on identifying individual white blood cells (WBC) to diagnose leukemia [81], as well as prostate, breast, and lung cancer [82–87]. Others focused on localizing individual malignant centroblasts in images of follicular lymphoma [88].

In many studies, the detection problem was formulated in the context of cell segmentation [77, 89–95], or cell tracking [96], where the detection hypotheses served as initial seed points for subsequent segmentation algorithms. Popular conventional image processing algorithms such as thresholding, seeded region growing, marker-controlled watershed, hough and distance transforms, multiscale filtering, or morphological operations were then employed to segment the objects [70, 84, 91, 96–102]. Reasonable success in adressing the cell detection problem was achieved using carefully selected low- and high-level image processing techniques, for instance based on local symmetry features [103, 104]. Others used normalized cuts and spectral graph theory to segment cells [105]. A large body of publications has focused on tackling the challenging problem of separating close and overlapping objects in fluorescent images [106, 107], and images

using immunohistochemical [108–110] and histochemical [84, 111, 112] staining. Many studies concluded that the problem of segmenting individual cells is non-trivial and that prior knowledge is often an enabling component. Moreover, authors worked on quite diverse datasets and proposed well-working solutions to their own research problems, which were seldomly evaluated across different datasets [101, 113] or microscopic imaging modalities [114–116]. Despite individual successes, Irshad *et al.* [56] concluded in their recent review that cell detection and segmentation is still not fully resolved yet, and that a general-purpose detection and segmentation solution would require more advanced methods. Rather than relying on heuristics and rigid image processing, modern methods in histopathological image analysis employ machine learning and optimization techniques to create more robust solutions [66, 67, 71, 82–84, 108, 114, 116–123]. For instance, Petushi *et al.* [67] observed a considerable accuracy improvement in detecting different types of nuclei when using decision tree learning compared to a non-learning based version of their algorithm.

In order to detect lung cancer cells, Zhou *et al.* [82] employed feed-forward Neural Network (FF-NN) ensembles to learn a classifier from color and morphological cell features that were extracted from segmented cells. Wählby *et al.* [97] proposed a region-based approach involving a distance transform, a combination of texture and shape information, and the watershed transform to segment cell nuclei in 2D and 3D images. Yang *et al.* [77] introduced an unsupervised segmentation approach based on color gradient vector flow active contour models that achieved an accuracy comparable to supervised approaches when leveraging prior knowledge of the target application. In [100], Dalle *et al.* proposed a scheme to score nuclear pleomorphism by selectively segmenting just critical cell nuclei rather than detecting all cells in an image. As part of a pipeline to grade lymphocytic infiltration of breast cancer tissue, Basavanhally *et al.* [83] detected the centers of cell nuclei using a combination of region growing and a Markov Random Field (MRF). Al-Kofahi *et al.* [101] employed graph cuts and multi-scale Laplacian-of-Gaussian (LoG) filtering as general approach to detect and segment cell nuclei. Some other work relied on contour-based cell models [113], active contours [84, 111, 112, 114, 124], or leverage shape and appearance priors [94, 111, 125] in a global optimization strategy. Veillard *et al.* [94] produced a pixel-wise nuclei probability map, to which they applied active contour models with shape prior to extract cells after applying morphological hole-filling. Support Vector Machines (SVM) [126] and AdaBoost [127] were used previously [122, 123, 128, 129] to learn cell object detectors using a diverse set of features extracted from images. Moreover, learning-based

approaches have recently proven to achieve state-of-the-art results on cell detection benchmarks such as [130]. On this benchmark, the work of Arteta *et al.* [116] to date outperformed other approaches. It is based on extracting a large number of cell candidates using the maximally stable extremal region (MSER) detector [131, 132]. This set of candidates is then pruned using several, increasingly complex classifiers based on structured Support Vector Machines (SSVM). A similar approach has been pursued earlier on renal cell carcinoma tissue by Fuchs *et al.* [71]. They first segmented a large cohort of candidate regions revealed by a combination of morphological operations and Canny edges [133] and employed a soft-margin SVM classifier to remove false positives. Other approaches applied a classifier densely over the input images in a sliding window fashion [94, 134] or learned regions revealed by the scale-invariant feature transform (SIFT) [135] keypoint detector [136].

Even more recently, Deep Learning methods, i.e. deep FF-NN, were used by many authors to solve diverse tasks in histopathological image analysis such as mitosis and cell detection [60, 65, 137–146], or semantic segmentation [147, 148]. These methods have been developed in parallel to the cell localization method that is going to be proposed in this work. Further, due to the availability of accelerated training on graphics hardware, large convolutional Neural Networks (CNN) [149, 150] could be trained in reasonable time and demonstrated tremendous success on many image analysis tasks when huge image datasets were available. Hence, such approaches have set the bar very high for other methods to compete [65, 151–154]. On the other hand, one of the main drawbacks of Deep Learning in general is the requirement for huge labeled image datasets, which are usually less frequently available in biomedical imaging. Further, properly training deep models is tedious, and requires a certain extent of experience to avoid common pitfalls such as over-fitting small datasets and designing the network architectures. Nevertheless, many successful applications have been published recently. In [141, 142], authors proposed a cell nuclei detector based on structured regression and a novel voting scheme using CNN. Cao *et al.* [87] employed CNNs to learn semantic high-level tissue features and combined them with lower-level nuclei and intensity features to improve breast cancer grading, while Sirinukunwattana *et al.* [146] used a spatially constrained CNN to detect and classify cell nuclei in colon cancer images.

A related, but different problem in histopathology image analysis is to count cells without explicitly detecting them. This was done by estimating their density in an image [155], for instance using regression Random Forest [156], or Fully Convolutional Neural Networks [140]. However, these approaches did not produce precise locations of

the cells or cell nuclei, which are important to perform cell type recognition.

Despite there are many methods available to detect cells, we pursue a strategy that does not require a precise delineation of the nuclei. As it will be shown in this work, it is in fact sufficient to have a weak ground truth estimate of the cell nuclei centers to learn a robust detection model from labeled training image data in a supervised manner.

3.1.2. Obtaining Ground Truth Cell Locations

Learning-based image analysis increases robustness to changing conditions. A common prerequisite for supervised machine learning algorithms is having a 'gold standard', or ground truth. However, one of the main challenges is obtaining a *reliable* ground truth. Ground truth resembles a set of labeled examples, where a prediction model can be built by employing some learning algorithm. Yet, different issues need to be considered when creating ground truth annotations. Selecting a suitable complexity of annotations actually depends on the task to be solved as well as the amount of required domain-specific knowledge. In domains such as natural images depicting landscapes and animals, urban scenes showing pedestrians, facades, and cars, ground truth for object detection and semantic segmentation tasks can rather easily be created by citizen scientists [157]. Perhaps the most prominent example in computer vision research is the *LabelMe* database [158], which allows users without domain-specific knowledge to mark or outline objects in any kind of images in a web-application. In the biomedical domain, however, creating a ground truth dataset can rarely be accomplished without experts, who underwent years of special training in e.g. cyto-, neuro-, or hematopathology to correctly recognize objects, tissue, or disease patterns. These experts are rare for special applications and hence our inescapable annotations must be created most effectively.

Assume we were given an object detection task such as localizing cells in images, the simplest possible annotation is placing a single-pixel dot on each object we want to detect. These weak annotations are quite effective, if the objects of interest for instance do not consist of many independently moving parts, or are rather small in size and compact in appearance. Another strategy to quickly annotate objects is drawing brush strokes onto objects, which allows capturing more than a single pixel of the object in one sweep. A more complex approach is drawing a minimum bounding box around each object, which first requires a thorough identification of the object boundaries. Although bounding boxes provide a more evident delineation of an object's extent, it

is not always required. Moreover, it can get very tedious for hundreds or thousands of objects. Pixel-level semantic image segmentation tasks require each pixel in an image to be annotated with the target label. Despite segmentation of microscopy images is a common task in histopathological image analysis, fully pixel-wise annotated datasets such as in [159–161] are rarely available due to the complexity of manual ground truth creation.

Previously, segmentation of individual objects has been frequently used as an intermediate step in selective object detection. Statistical and morphological features were used to characterize these obtained regions to be learned by a classifier. To localize the centers of cell nuclei, we are going to pursue a learning-based approach that does not require any delineation of the cell borders. Hence, we are able to rely on simple dot-annotations given by experts in histopathology. Furthermore, no sophisticated tools are required to place dots onto the cell centers, and the experts are able to generate a reasonable amount of valuable annotations in a minimum of time.

3.1.3. Pixel-Wise Binary Classification for Cell Detection

One of the standard learning-based approaches to cell detection is classifying each pixel in an image whether it belongs to a cell center or background. Such a system is trained with a set of local image patches, where the center pixel represents the cell object's center and gets labeled with a binary label, i.e. foreground (1) or background (0). A binary classifier f is learned and given an unseen test image, the probability of each image location belonging to a cell center is predicted by applying f densely over the image. While this approach is relatively straightforward, it contains some drawbacks with respect to the localization accuracy.

Plateau Formation A prediction of a cell center may be inconclusive, since a binary classifier tends to produce similarly high probabilities around an actual ground truth cell center. Fig. 3.1 illustrates this problem. Especially, but not exclusively, for cells with highly homogeneous staining, these responses form blobs (plateaus) in the probability map, because learning a binary discrimination does not encode the location of a given patch with respect to the cell center. All red highlighted dots in the enlarged view denote pixels located at the border of the cell nucleus, but their classifier response does not differ from the actual center pixel prediction, denoted by the green dot.

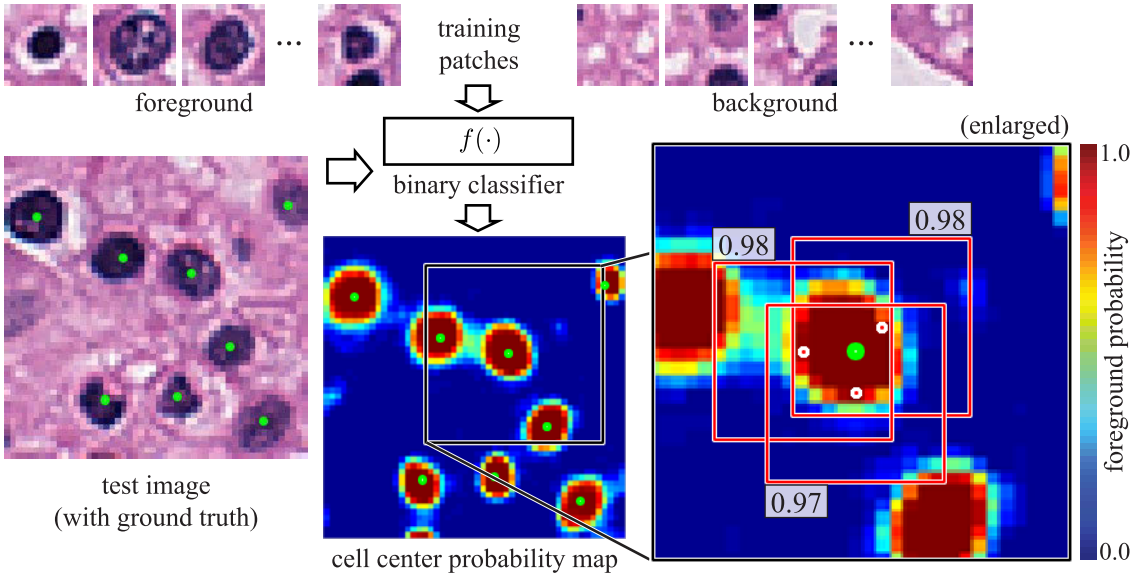
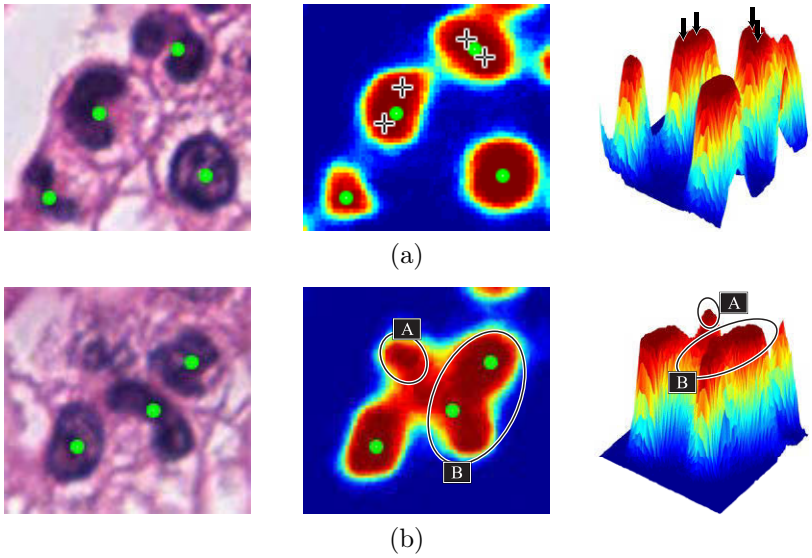


Figure 3.1.: Plateau formation is a drawback of pixel-wise binary classification. Local image patches, labeled as foreground (1) and background (0) are used to train a binary classifier f that predicts a cell center probability map from a given test image. Colors towards red indicate higher probability for a cell center, green dots denote ground truth cell centers. In the enlarged view, similar classifier responses computed from local patches (highlighted red dots) foster plateau formation in the adjacency of the actual cell center without any clear local maximum. The red squares delineate the patches centered on the red dots, which were used as visual context for prediction.

Multiple Cell Center Hypotheses Especially for cells with anisotropic shape, the binary classifier frequently predicts multiple clusters of high probabilities on a single object. Each cluster may theoretically correspond to a cell center. In Fig. 3.2 (a), this unwanted behaviour is demonstrated for a sample image, where multiple responses per object are delineated with crosses in the probability map, and arrows in the 3D model. Furthermore, the 3D model of the prediction surface also shows plateau formation.

Cell Merging Binary classification may merge adjacent objects, since the foreground probabilities between two close objects is still too high to separate the objects. In Fig. 3.2 (b), the ellipses highlight the regions, where two cells were merged, and also a false positive detection occurred. This frequently happens to very closely located cells, and to adjacent non-cell objects, which easily confuse the classifier when their appearances are too similar.



(a)

(b)

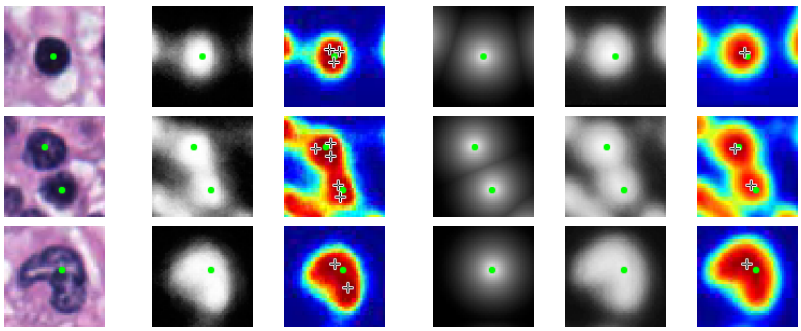
Figure 3.2.: Drawbacks of pixel-wise binary classification used to detect cell centers. The first column contains a source image, the second column the prediction as heatmap and third column a 3D model of the prediction surface. Green dots denote ground truth nuclei centers. (a) The anisotropic shape of the cells receive multiple clusters of similarly high responses (indicated by crosses and arrows), which could be misunderstood as center hypotheses on one object. (b) In some cases, false positive centers are predicted for close non-cell objects with similar appearance (A). Individual cells can be merged if they are located very closely (B).

3.2. Cell Detection via Proximity Score Regression

The drawbacks of using standard pixel-wise binary classification methods for cell detection have been outlined in the previous section. This section describes an alternative, novel approach to cell detection based on the regression of a proximity score, which overcomes the current limitations of binary classification approaches, cf. Fig. 3.3.

Given a local image patch, the algorithm learns to predict a function of the distance to the closest cell center. The construction of the proximity score map is inspired by the work of Sironi *et al.* [162, 163], who recently proposed a scheme to robustly reveal linear structures in an image. Defining a continuous learning target shifts the learning task from pixel-wise classification to proximity score regression and allows us to encode spatial information in the regression target.

Fig. 3.4 contains an overview of the proposed cell localization method. Given a set of dot-annotated cell nuclei, a proximity score map is computed as learning target and a regression model is trained on local image patches. Applying the model in a



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

Figure 3.3.: Comparing classification and regression for cell detection. First row: one fully stained nucleus demonstrating the plateau formation, second row: two closely located nuclei, third row: one nucleus of anisotropic shape and non-uniform staining. The green dots indicate ground truth annotation of the cell centers. (a) Three input patches, centered on one or two cells. (b) Probability maps provided by a classifier applied to these patches, and (c) the local maximums of these maps. They exhibit many local maximums – indicated by crosses – around the actual cell centers, while only one maximum is expected. (d) The expected proximity score map that a regression model should predict and (e) the actual predictions. (f) The local maximums of these predictions correspond much better to the cell centers and do not suffer from multiple responses. The regression method is able to overcome these problems arising from the binary classification approach.

Kainz P, Urschler M, Schulter S, Wohlhart P, Lepetit V. You Should Use Regression to Detect Cells. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. vol. 9351 of Lecture Notes in Computer Science. Springer International Publishing; 2015. p. 276–283. With permission of Springer.

sliding window fashion over an unseen test image, a proximity score map is predicted, where the local maximums correspond to the hypothesized cell centers. Post-processing, comprising, smoothing, score filtering and non-maximum suppression, subsequently reveals the cell centers. In the following sections, we detail the construction of the score map, training and evaluation of the regression model as well as the post-processing steps.

3.2.1. Constructing a Proximity Score Map

Let $\mathbf{u} = (u_x, u_y), \mathbf{u} \in \Omega$ be a location in the domain of image I, and $\mathcal{C} = \left\{\mathbf{c}^{(i)}\right\}$, $\mathbf{c}^{(i)} = \left(c_x^{(i)}, c_y^{(i)}\right)$ be the set of annotated ground truth cell center locations. We now seek a smooth, continuous target function $y(\mathbf{u})$ that expresses the proximity of a pixel $\mathbf{u}$ to the closest cell center. A straightforward way of defining the *proximity score map*

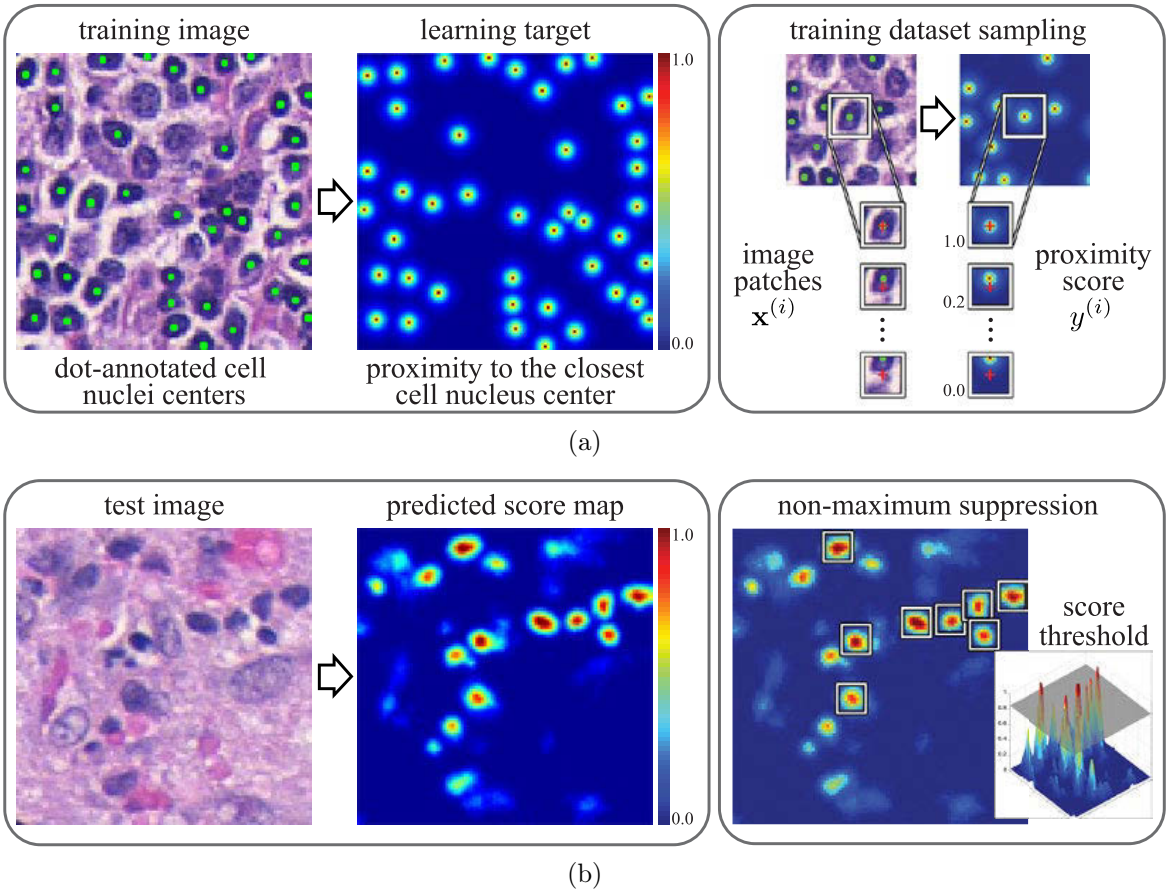


Figure 3.4.: Overview of the cell localization approach using the regression method. (a) Creation of the proximity score map as learning target from dot-annotated cell nuclei centers. Local image patches and their corresponding proximity score comprise the training dataset, where a regression model is trained on. (b) A previously unseen test image and the predicted proximity score map. Applying a score threshold removes 'weak' cell nuclei center hypotheses and non-maximum suppression reveals the locations of the centers.

$\mathbf{y} := \{y(\mathbf{u})\}$ is to define it as a function of the Euclidean distance transform $\mathfrak{D}_\mathcal{C}$ of the set $\mathcal{C}$: $y(\mathbf{u}) = -\mathfrak{D}_\mathcal{C}(\mathbf{u})$. However, while being initially intuitive, this approach suffers from several drawbacks such as producing high proximity scores even in background areas. Moreover, it forces a regression model g to predict varying scores for very similar background regions and finally the individual cell centers are not well-defined. Fig. 3.5 illustrates these issues using an example image of H&E stained bone marrow cells.

Hence, it is better to predict a smooth function of the Euclidean distance transform $\mathfrak{D}_\mathcal{C}$, that is uniform in the background regions and has better localized, distinctive peaks [162, 163]:

$$y(\mathbf{u}) = \begin{cases} \exp\left(\alpha\left(1 - \frac{\mathfrak{D}_{\mathcal{C}}(\mathbf{u})}{d_M}\right)\right) - 1 & \text{if } \mathfrak{D}_{\mathcal{C}}(\mathbf{u}) < d_M \\ 0 & \text{otherwise} \end{cases}, \quad \forall \mathbf{u} \in \Omega, \tag{3.1}$$

where d_M and α control the shape of the exponential 2D function centered on the ground truth cell nuclei annotations. The function is maximal at a ground truth center and decreases exponentially with increasing distance, until it approaches zero at a distance d_M. This method overcomes existing problems by defining a more suitable learning target.

Although there is no strict requirement, it is reasonable to include some prior knowledge into the learning target, i.e. the shape of the peaks in $\mathbf{y}$ such that they are modeled in a reasonable fashion and relate to the objects of interest in the images. In this specific application, we have weakly dot-annotated cell nuclei centers as ground truth and assume idealized cell nuclei, i.e. isotropic shapes. To establish the link between the shape of the peaks and our real-world cell objects, we estimate the average object radius r_{obj} from the image data.

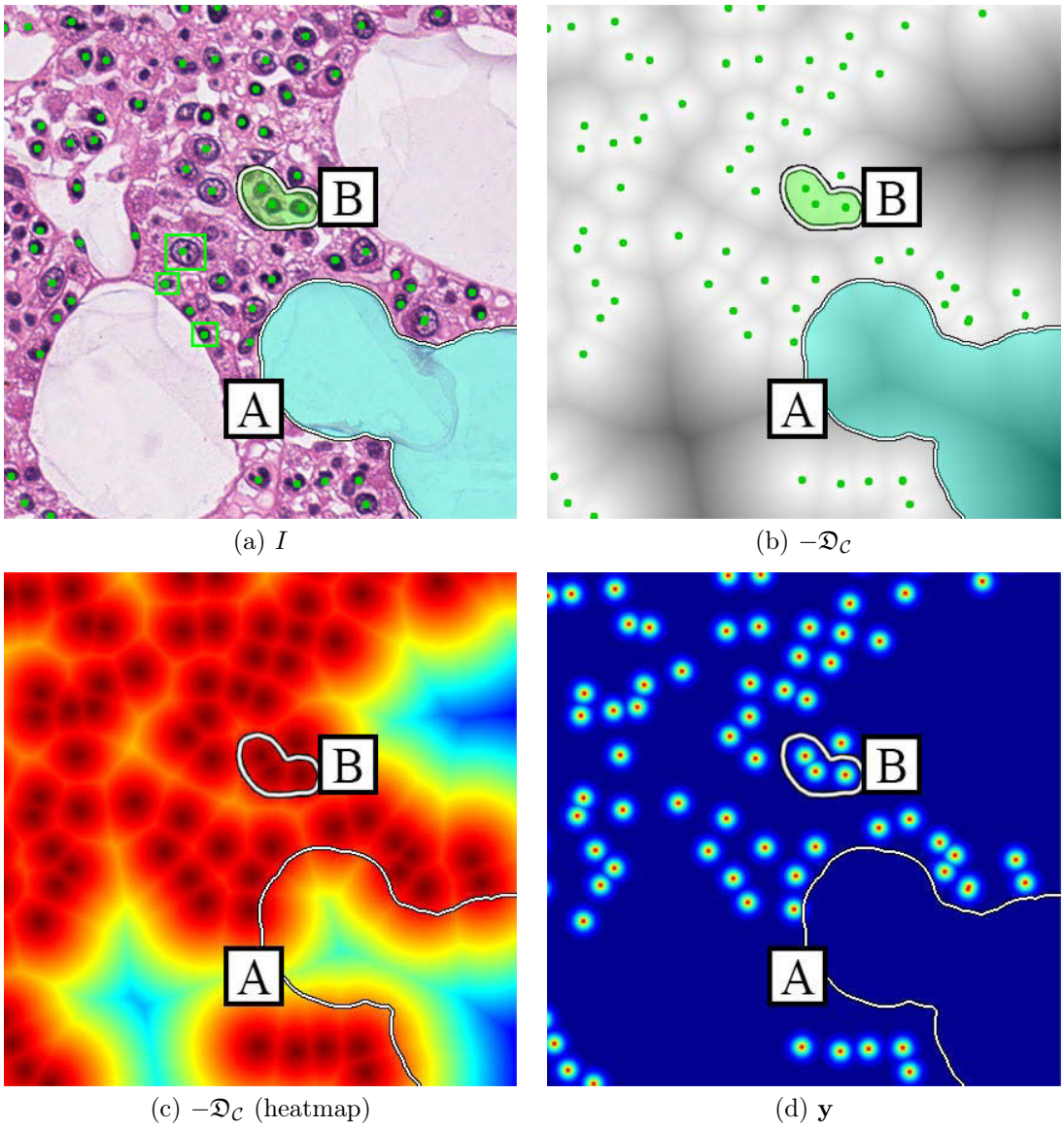
This information comes for free, if minimum bounding-box annotations are available, but as discussed in the previous section, providing a bounding-box for many small cell object is tedious and time-consuming. Moreover, it may not be of significant additional value for this particular application. For dot-annotated datasets, manually drawing a few minimum bounding-boxes is thus usually sufficient to estimate r_{obj} with acceptable accuracy. Nevertheless, it has to be mentioned that a proximity score map based on this prior may not be ideal for large datasets, where the variance of the cell nuclei size is large and the background is very heterogeneous. In that case, bounding-box annotations would be better, since a manual selection is less reproducible and we could rather easily miss representatives of each size, which could harm the localization performance.

To obtain a proximity score map, we first compute the Euclidean distance transform of the cell centers $\mathfrak{D}_{\mathcal{C}}$. Then, in Eq. (3.1) prior information on the datasets is used to restrict the map with

$$d_M = 2 \cdot r_{obj}, \quad \text{and} \tag{3.2}$$

$$\alpha = \frac{s}{\sqrt{d_M}}. \tag{3.3}$$

A scaling factor of $s = 20$, empirically derived from our available datasets, was used to calculate the shape constant α of the peaks in the target maps.



(a) I (b) $-\mathfrak{D}_{\mathcal{C}}$

(c) $-\mathfrak{D}_{\mathcal{C}}$ (heatmap) (d) $\mathbf{y}$

Figure 3.5.: Construction of a ground truth proximity score map $\mathbf{y}$ based on the Euclidean distance transform of the cell centers $\mathfrak{D}_{\mathcal{C}}$. (a) The original image with the ground truth cell centers, (b) the negative Euclidean distance transform $-\mathfrak{D}_{\mathcal{C}}$, (c) heatmap of (b), and (d) the final proximity score map $\mathbf{y}$ (cf. Eq. (3.1)) as heatmap. Green dots denote the ground truth cell center locations, the green boxes were manually drawn to estimate the average object size. In the heatmaps, colors towards red denote high proximity to a cell center, while image locations colored towards blue are farther away. Drawbacks of using $-\mathfrak{D}_{\mathcal{C}}$ as target are clearly visible: in (c), homogeneous background regions are represented by heterogeneous target values (area A), the slow decrease of the score produces high values in background regions, and therefore less pronounced peaks (area B). In (d), the final proximity score map has well-defined peaks on the cell center annotations. Even very close objects can be represented properly, and background is represented by a single score value.

3.2.2. Learning a Regression Model from Image Data

Many options are available to learn a regression model g from a given dataset, such as Neural Networks, decision trees, or Bayesian methods [164]. In this thesis, the localization of cell nuclei is based on Random Decision Forests (RF) [165], because they are fast to evaluate, have been shown to perform well on many image analysis problems involving machine learning [166, 167], and are relatively easy to implement. An RF is an ensemble method of statistical learning and consists of T individual decision trees T_t, $t \in \{1, \ldots, T\}$. Each tree individually learns a mapping function $g_t : \mathcal{X} \rightarrow \mathcal{Y}$, such that $g = \gamma(\{g_t\})$, where $\mathcal{X}$ and $\mathcal{Y}$ denote input and output spaces, and γ denotes an aggregation function over all individual trees (e.g. averaging). During training, the internal nodes of a tree select and store the best parameters to split the dataset into two subsets, while the leaf nodes essentially store the target values of the data samples. Once an RF is trained, previously unseen samples can be propagated through the trees, and a consolidated prediction can be made.

Input Data Representation: Visual Image Features

With respect to the input representation, there are two common ways of training an RF on image data. Firstly, an image I can be seen as an n-D *vector*, where all pixel values of all color channels are stacked in a single vector. However, the vector representation discards all spatial relationships among neighboring pixels and considers them as individual features. The second option considers the image as an n-D *array*, which preserves the spatial relationships for pattern mining in the spatial domain. Here, we use an RGB image I as a 3-D array representation, which allows us to compute additional visual features such as edges to provide a comprehensive representation of the image content. Instead of exclusively relying on raw pixel intensities, we train and evaluate the RF on visual image features [168, 169] extracted from local input image patches $I(\mathbf{u})$ centered on an image location $\mathbf{u}$. Each image patch is represented by a set of visual feature channels $\Phi = \{\Phi_i\}$. The image features for an image patch showing a H&E stained megakaryocyte from bone marrow tissue are illustrated in Fig. 3.6. Assuming an RGB patch as original image, we define $|\Phi| = 54$ different channels that resemble intensity, structure, and texture information, cf. Table 3.2.

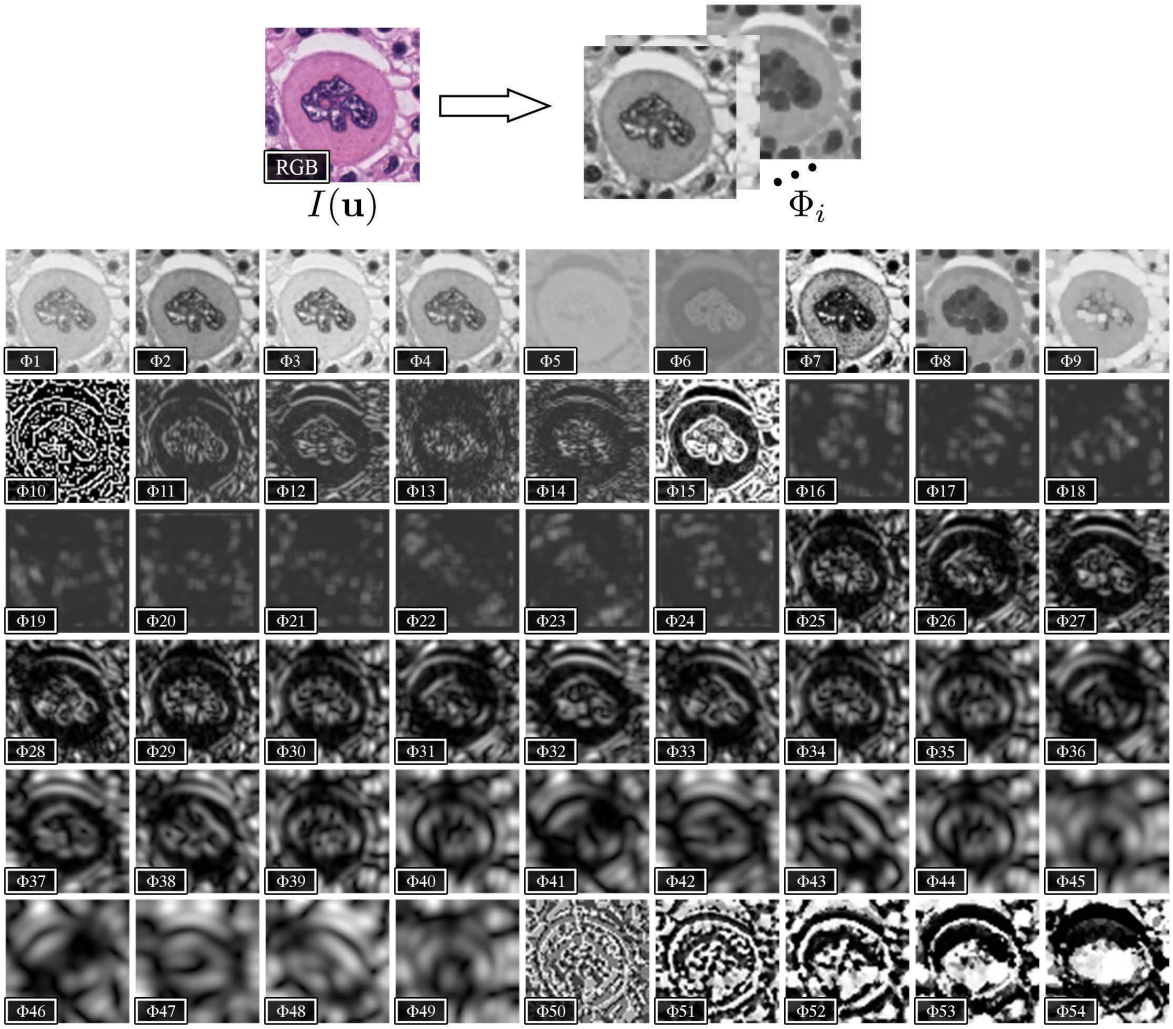


Figure 3.6.: Illustration of the set of 54 visual image features Φ_i for cell localization using a local RGB image patch $I(\mathbf{u})$ depicting an H&E stained bone marrow cell (megakaryocyte) as an example.

Training Dataset Sampling

In this supervised regression problem, each random tree in the forest is provided with a dataset $\mathcal{P} = \left\{\left(\mathbf{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{N_\mathcal{P}}$ of $N_\mathcal{P} = |\mathcal{P}|$ training samples, where $\mathbf{x}^{(i)} := I\left(\mathbf{u}^{(i)}\right)$ is the i-th training image patch centered on $\mathbf{u}^{(i)}$, and $y^{(i)} := y\left(\mathbf{u}^{(i)}\right)$ is the corresponding proximity score normalized in the range $[0, 1]$.

When an RF predictor is trained on local image patches, the input patch size p_{in} essentially defines the spatial context (field of view). Here, a square patch size is used, despite there is no general restriction for that choice. However, since a cell must not necessarily be isotropic and may occur in arbitrary rotations, a square patch size seems

IDs	Name	Source Space	Details/Results
Φ_{1-3}	RGB intensities	CIE RGB	–
Φ_{4-6}	CIE L*a*b intensities	CIE RGB	–
Φ_7	Histogram-equalized intensities	grey-scale	–
Φ_8	Local minimum intensities, i.e. erosion with a uniform kernel $\mathcal{K}$ [170]	grey-scale	$I(\mathbf{u}) \ominus \mathcal{K}_{(3\times3)}$
Φ_9	Local maximum intensities, i.e. dilation with a uniform kernel $\mathcal{K}$ [170]	grey-scale	$I(\mathbf{u}) \oplus \mathcal{K}_{(3\times3)}$
Φ_{10}	Canny edges [133]	grey-scale	–
Φ_{11-12}	First-order image gradients in (x,y)-direction, computed with the proper Sobel-Feldman kernels [171]	grey-scale	$\nabla_x I(\mathbf{u}), \nabla_y I(\mathbf{u})$
Φ_{13-14}	Second-order image gradients in (x,y)-direction, computed with the proper Sobel-Feldman kernels [171]	grey-scale	$\Delta_x I(\mathbf{u}), \Delta_y I(\mathbf{u})$
Φ_{15}	Magnitude of Sobel gradients	$\nabla I(\mathbf{u}) = \sum_d \nabla_d I(\mathbf{u})$	$\lvert\nabla I(\mathbf{u})\rvert$
Φ_{16-24}	HoG-like image features [166, 172]	grey-scale	–
Φ_{25-49}	Gabor features in five orientations μ_G, and five scales each ν_G	grey-scale	$\mu_G \in \{0, 2, 4, 6, 8\}$, $\nu_G \in \{0, 1, 2, 3, 4\}$
Φ_{50-54}	Rotation-invariant Local Binary Patterns (LBP) [173–175] at five scales	grey-scale	2^n, $n = 0, \ldots, 4$

Table 3.2.: Summary of the visual image features and parameters used for representing the input space for learning to localize cells.

to be reasonable. An important aspect to be considered is that it provides enough spatial context such that the method is able to choose from meaningful textural and structural features during training.

A 'foreground-background' threshold τ_{bg} can be used to include prior knowledge into patch sampling for the model. This threshold draws more attention to learning regions with high scores in immediate proximity of the cell centers. An intuitive default choice is to set $\tau_{bg} = 0.5$, i.e. to consider all target values ≥ 0.5, which is supposed to cover the most significant parts around a cell's center, since it is a function of the mean object size.

Sampling the training set proceeds in two steps. Firstly, the set of patches of size N_{τ^+}, where the center pixel is $\geq \tau_{bg}$, is added to the training dataset. Secondly, from the remaining image locations ($< \tau_{bg}$) that cover pure background (i.e. $y(\mathbf{u}) = 0$) and locations farther away from the center, a certain number of patches N_{τ^-} is randomly sampled and added to the training dataset, too.

Since we are considering a regression problem, class label imbalance is not considered a hindering issue as it may be for some classification problems. However, a distinct ratio parameter controls N_{τ^-} with respect to N_{τ^+}. Setting it to 1.0 results in $N_{\tau^-} \equiv N_{\tau^+}$ and ensures that the training set contains patches sampled from the entire value range of the target space.

Training and Inference

In the training phase, starting at the root node of a tree T_t, the objective at each non-terminal (split) node is to assign all available data samples to either a *left* or *right* subset by evaluating a number of split functions. A split function ϕ consists of a selection function θ and a random threshold τ_ϕ. A data sample is assigned to the left subset, if the response of the selection function is smaller than the threshold, or to the right subset otherwise.

The selection function θ essentially defines, which parts of the input sample are evaluated in which way by the optimization algorithm. In particular, we consider for evaluation single pixel values, pixel value differences, Haar-like functions [176], and spatially constrained pixel value differences, where the second location for difference computation was chosen within a distance of 10 pixels, and clamped at the patch borders. Haar-like functions are obtained by taking differences of two randomly selected areas, which can efficiently be computed using integral images [176]. For all selection functions except single pixel values, a 50% chance exists to use the values from locations within the same or from a different feature channel. Exhaustively testing all possible selection functions on the available feature space $(p_{in})^2 \cdot |\Phi|$ is computationally inefficient and essentially not required. If we consider the feature space of a single image location, we would use e.g. $\sqrt{|\Phi|}$ random features to optimize the dataset split at a node instead [164, 165]. As a consequence, we derived the number of randomly drawn selection functions as a function of the patch size and number of available feature channels: $N_\theta = \left\lfloor p_{in} \cdot \sqrt{|\Phi|} \right\rfloor$. Each selection function was compared to 20 random thresholds, resulting in $20 \cdot N_\theta$ split function tests at each node.

Let N_j be the number of individual samples arriving at a non-terminal node j, we randomly select $\min\{N_j, M_j\}$ samples to search for the optimal binary split decision[1]. Since each tree is trained on the entire dataset, selecting random samples for the

[1] Since we address a single node j in the following formal description of the optimization, we will omit the index j for the sake of clarity.

split evaluation is an alternative concept to tree bagging (subsampling with replacement) [177] in order to decrease the variance of the model, without increasing the bias [164]. An optimal split decision minimizes the variance of all target labels $y^{(i)}$ over both subsets. Hence, as in classical CART systems [178], reduction in variance (RV) is used as split quality criterion, which can be formulated as

$$RV(\phi) = Var(\mathcal{P}) - \sum_{s \in \{L,R\}} w_s \, Var\left(\mathcal{P}_s(\phi)\right), \tag{3.4}$$

where

$$w_s = \frac{|\mathcal{P}_s(\phi)|}{|\mathcal{P}|} \tag{3.5}$$

is the weight of the left and right subset $\mathcal{P}_{s \in \{L,R\}}(\phi)$ after applying a candidate split function ϕ, and $|\cdot|$ denotes the cardinality of the set.

With $\bar{y}_{\mathcal{P}} = \frac{1}{|\mathcal{P}|} \sum_{i=1}^{|\mathcal{P}|} y^{(i)}$ being the arithmetic mean of a set, the variance (i.e. the averaged sum of squares) is computed using

$$Var(\mathcal{P}) = \frac{1}{|\mathcal{P}|} \sum_{i=1}^{|\mathcal{P}|} \left(y^{(i)} - \bar{y}_{\mathcal{P}}\right)^2, \tag{3.6}$$

using $\frac{1}{|\mathcal{P}|}$ as normalization factor.

The optimal split function ϕ^* is selected by obtaining a Maximum A-Posteriori (MAP) estimate

$$\phi^* = \arg\max_{\phi} RV(\phi), \tag{3.7}$$

and is stored at each non-terminal node.

Equivalently to maximizing $RV(\phi)$, minimizing the sum of variances over both left and right subset leads to the optimal split function [164] without the requirement to compute the variance of the total set separately:

$$\phi^* = \arg\min_{\phi} \sum_{s \in \{L,R\}} w_s \, Var\left(\mathcal{P}_s(\phi)\right). \tag{3.8}$$

Splitting resumes recursively in a best-first node expansion strategy until a node contains only one sample, i.e. the node is pure. Then, a terminal (leaf) node is created. Early stopping criteria such as the maximum tree depth T_{md} or a minimum number

of samples per node are used to limit the depth and are employed to smooth the predictions of individual trees by aggregating multiple samples in the leaf nodes, e.g. via averaging. A single value for each dimension of the target variable is stored at each leaf, such that when the tree is queried with an unseen sample, predictions can be obtained. Many options are available to produce the final values in a leaf node, ranging from simple averaging to sophisticated non-linear models [179]. While non-linear models may have a better fit to the target function that needs to be learned from the training data (Eq. (3.1)), computing the mean of all predictions results in a piece-wise constant approximation of the target function, which is usually sufficient. Hence, we consider it less likely that the regression model overfits the training data and in this work, predictions are computed using the average of the target values of all samples reaching a particular leaf node.

To infer the proximity score $\hat{y}_t$ of a test image patch $I(\mathbf{u})$, it is passed down each tree T_t and at each split node, the learned split function ϕ^* is applied to decide the subsequent pathway. Once all T trees in the forest have been traversed from the root to the final leaf node, the predictions are consolidated by simply averaging them over all trees to produce a final prediction $\hat{y}$:

$$\hat{y}_t = g_t\left(I(\mathbf{u})\right), \text{hence} \tag{3.9}$$

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t. \tag{3.10}$$

The entire proximity score map $\hat{\mathbf{y}}$ for an unseen image is obtained by applying g in a sliding window fashion with pixel stride $W_S = 1$ over all image locations.

Formally, a regression function g is learned to map from a d-dimensional input space $\mathcal{X} = \mathbb{R}^d$ to a one-dimensional output space $\mathcal{Y} = \mathbb{R}^+$. Henceforth, we refer to this method as *single-target* regression (*rRF*).

Spatial-Averaging Regression

Learning the proximity scores can be formulated as a classical regression problem with a single output variable, or alternatively as regression with multiple output variables. In the latter case, for a given training image patch $\mathbf{x}^{(i)}$, an entire output patch representing the proximity scores of locations around the center pixel is learned during training. Let p_{out} be the (square) size of that output patch, and $\mathbf{y}^{(i)} \in \mathbb{R}^{O\times 1}$ the label in column

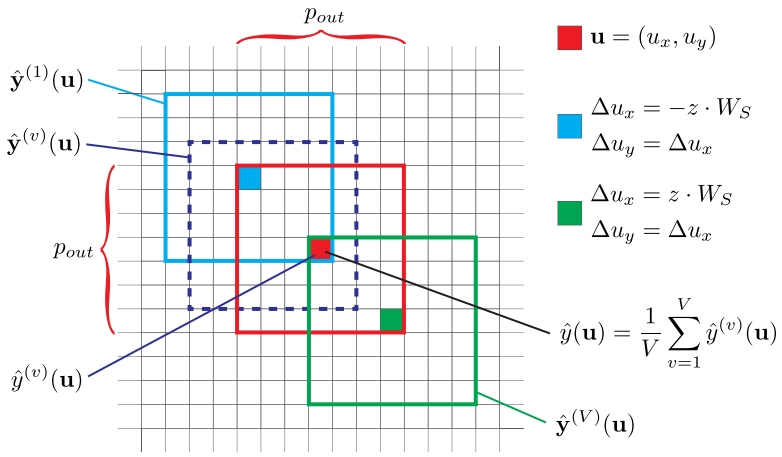


Figure 3.7.: Overlapping output patches of the spatial-averaging regression using a patch size of $p_{out} = 7$ and sliding window pixel stride of $W_S = 1$. Each image location $\mathbf{u}$ receives V individual predictions $\hat{y}^{(v)}(\mathbf{u})$ that are averaged into $\hat{y}(\mathbf{u})$, with V depending on p_{out} and W_S.

vector form, where $O = (p_{out})^2$. Formally, the learned regression function takes the form of $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$.

Training proceeds similar to the *single-target* regression, but optimization and inference generalize to O dimensions. However, now we consider the dataset $\mathcal{P} = \left\{\left(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right)\right\}_{i=1}^{N_\mathcal{P}}$, where $\mathbf{y}^{(i)} = \left\{y_1^{(i)}, \ldots, y_O^{(i)}\right\}$. With $\bar{\mathbf{y}}_\mathcal{P}$ as the average proximity score vector, computing the variance of the set $\mathcal{P}$ in the node optimization is done explicitly in each dimension:

$$Var(\mathcal{P}) = \frac{1}{|\mathcal{P}|} \sum_{i=1}^{|\mathcal{P}|} \left(\mathbf{y}^{(i)} - \bar{\mathbf{y}}_\mathcal{P}\right)^2. \tag{3.11}$$

To infer an output patch for a given input patch $I(\mathbf{u})$, the consolidation of the individual tree predictions $\hat{\mathbf{y}}_t$ into a final one is denoted as

$$\hat{\mathbf{y}}_t = g_t\left(I(\mathbf{u})\right), \text{hence} \tag{3.12}$$

$$\hat{\mathbf{y}} = \frac{1}{T} \sum_{t=1}^{T} \hat{\mathbf{y}}_t. \tag{3.13}$$

The predicted vector output $\hat{\mathbf{y}}$ is mapped back to spatial locations around the patch center location $\mathbf{u}$. Moreover, each image location receives multiple predictions from its neighbors due to the overlapping output patches, cf. Fig. 3.7. In particular, if we consider a pixel stride W_S of a sliding window being applied over an image, each image

location $\mathbf{u}$ is contained in

$$V = (2 \cdot z + 1)^2, \tag{3.14}$$

$$z = \left\lfloor \frac{p_{out}}{2 \cdot W_S} \right\rfloor, \tag{3.15}$$

output patches, which we will index by $v = 1, \ldots, V$. Therefore,

$$\hat{\mathbf{y}}^{(v)}(\mathbf{u}) := \hat{\mathbf{y}}(\mathbf{u} + \Delta\mathbf{u}) \tag{3.16}$$

denotes the v-th output patch containing location $\mathbf{u}$, centered on $\mathbf{u} + \Delta\mathbf{u}$. This patch is formed by the set of individual predictions $\hat{\mathbf{y}}^{(v)}(\mathbf{u}) = \left\{\hat{y}^{(v)}(\mathbf{u})\right\}$. More formally, the relative offset of the output window with respect to $\mathbf{u}$ is denoted as

$$\Delta\mathbf{u} = (\Delta u_x, \Delta u_y), \text{with} \tag{3.17}$$

$$\Delta u_x, \Delta u_y \in \{-z \cdot W_S, \ldots, 0, \ldots, z \cdot W_S\}. \tag{3.18}$$

From each output patch we collect the proximity score for that location $\hat{y}^{(v)}(\mathbf{u})$, to average them into a final prediction:

$$\hat{y}(\mathbf{u}) = \frac{1}{V} \sum_{v=1}^{V} \hat{y}^{(v)}(\mathbf{u}). \tag{3.19}$$

Averaging inherently causes smoothing of the proximity score map and makes it more robust to outliers. Additionally, since $p_{out} > 1$, it is possible to apply g with $W_S > 1$ as long as $W_S < p_{out}$, which results in a speedup during evaluation. We refer to this method as *spatial-averaging* regression (*srRF*).

3.2.3. Localizing Cells in a Proximity Score Map

Given a predicted proximity score map $\hat{\mathbf{y}}$, simple post-processing is applied to reveal the center of a cell nucleus as a local maximum. First, the prediction images are smoothed with a Gaussian filter $\mathcal{G}$ to remove noise and enforce individual peaks as one local maximum per hypothesized object. Then, a 'score' threshold κ is applied that discards center hypotheses with very low proximity scores. Finally, non-maximum suppression

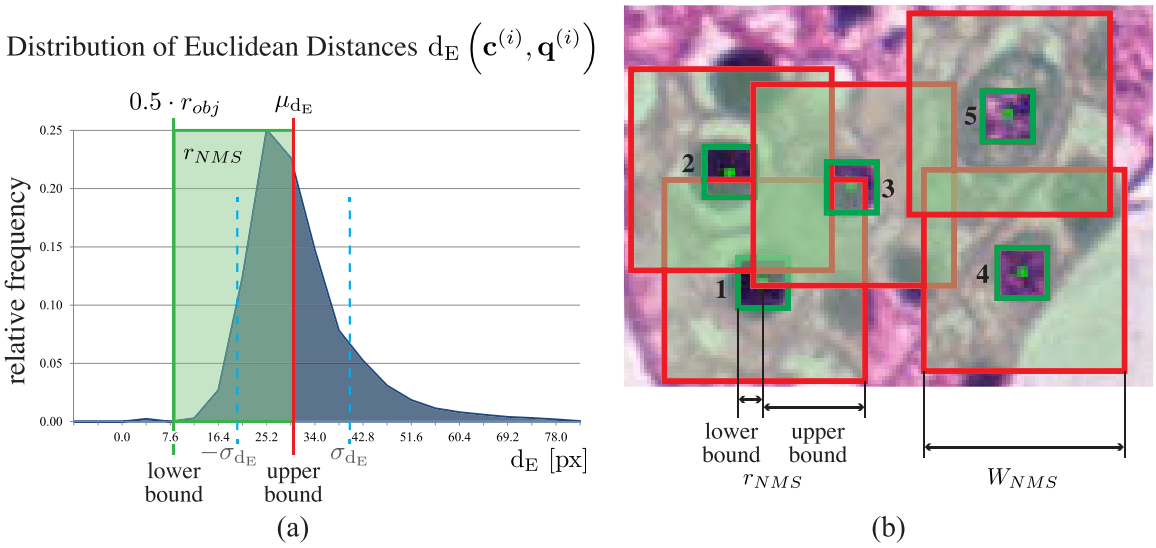


Figure 3.8.: Quantitative and qualitative illustrations of post-processing parameter choices using ground-truth annotations from an H&E stained bone marrow dataset, where $r_{obj} = 16$, $\mu_{\mathrm{d_E}} = 31$, and $\sigma_{\mathrm{d_E}} = 10$. (a) Distribution of Euclidean distances $\mathrm{d_E}$ between a ground-truth cell center and its nearest neighbour. Data for $\mathrm{d_E} > 86$ pixels are not shown in this chart. The distribution mean $\mu_{\mathrm{d_E}}$ is delineated by a red solid line, the standard deviation (SD) $\pm\sigma_{\mathrm{d_E}}$ by blue dashed lines. The proposed suitable range for selecting $r_{NMS} \in [0.5 \cdot r_{obj}, \mu_{\mathrm{d_E}}]$ is illustrated by a green overlay. (b) A qualitative illustration of proposed lower and upper bounds for r_{NMS} (green colored areas). The green dots denote ground truth cell locations of 5 cells (numbered) as an example. Here, cells 1 and 3 would mask each other, if the upper bounds (red squares) would be selected to construct W_{NMS}, while all other cells are just contained within their own window.

(NMS) [180, 181] is applied using a window size of

$$W_{NMS} = 2 \cdot r_{NMS} + 1, \tag{3.20}$$

which reveals the cell centers as local maximums.

Though, the radius r_{NMS} of the window should not be defined arbitrarily but is rather subject to the statistics of a particular dataset. A suitable size satisfies two criteria that influence the localization performance with respect to precision and recall. Firstly, the window must be large enough to cover a sufficient area around a cell center to avoid multiple detection hypotheses on single cell objects. Secondly, it must be small enough to avoid masking adjacent cell centers. This occurs, for instance, when multiple center hypotheses ($> \kappa$ and locally maximal) are located within one NMS window.

Assume we have a representative set of fully center-dot annotated cell nuclei as ground truth. To robustly detect the center of a single nucleus, we can consider the average object radius r_{obj} as cue for a suitable lower bound of r_{NMS}, which has already been estimated as part of the proximity map construction (cf. Section 3.2.1). To find the upper bound of r_{NMS} for a dataset, we define the distance of an annotated cell center $\mathbf{c}^{(i)}$ to it's nearest neighbour $\mathbf{q}^{(i)}$ to be the Euclidean distance $\mathrm{d_E}\left(\mathbf{c}^{(i)}, \mathbf{q}^{(i)}\right)$. Assuming that the distances approximately follow a normal distribution, we are able to obtain a mean and SD: $\mathrm{d_E} \sim \mathcal{N}\left(\mu_{\mathrm{d_E}}, \sigma_{\mathrm{d_E}}\right)$. In Fig. 3.8 (a), the distribution of $\mathrm{d_E}$ for a H&E stained bone marrow dataset is depicted, where relevant statistical parameters are delineated by colored lines. Having derived reasonable bounds, we can now select r_{NMS} for example from the range $[0.5 \cdot r_{obj}, \mu_{\mathrm{d_E}}]$. Nevertheless, the optimal value needs to be determined empirically, since the proposed lower bound may underestimate, and the upper bound may overestimate this range for particular datasets. Fig. 3.8 (b) shows a qualitative illustration of the proposed bounds for r_{NMS}. The green squares depict the lower bounds, while the red squares depict the upper bounds of the range. In this example, the chosen upper bound would probably cause lots of false negatives and decrease recall, since the red squares often contain more than one ground truth dots. For instance, if the goal is to optimize for precision, a larger window size can be selected, taking into account that some very close cells may be missed during localization. On the other hand, when recall is considered to be more important, the window size should be chosen somewhere closer to the lower bound.

3.3. Strategies for Cell Detection at Different Scales

We have considered the fact that all cells of interest are non-overlapping and approximately at the same object scale. The latter of which applies to maturation stages in granulo- and erythropoiesis, but not to megakaryopoiesis, cf. Chapter 2. Regarding size and appearance of bone marrow cells, megakaryocytes are the largest cells. They are up to five times larger than cells in granulo- or erythropoiesis, and differ in appearance in terms of large cell nuclei, surrounded by extensive cytoplasm. Thus, we need to consider them as a distinct cell class, which cannot be covered by simple multi-scale training of granulo- or erythropoietic cells in general. In standard multi-scale object detection, different scales of a particular object class can be learned for instances by training on a (Gaussian) scale-space pyramid. However, in the case of megakaryocytes, the texture, shape, and size of their multilobed cell nuclei is morphologically similar

to the ones of orthochromatic normoblasts at 40× magnification. Given these similarities, fortunately a quite discriminating feature of megakaryocytes is their size and large perinuclear cytoplasm. Therefore, an additional cell detector is required, which focuses on megakaryocytes. This detector can be trained on a different magnification level of the specimen (e.g. 20×, or even 10×). The expected benefits are reduced runtime and reducing the probability of false positive identifications, since the large cytoplasm gains more weight in the correct localization of the cell center. The optimal magnification for megakaryocyte detection in bone marrow slides is explored in Chapter 5.

4. Experimental Setup and Implementation Details

In this chapter, the setup for empirical evaluations is described along the implementation details. The regression method is thoroughly evaluated and compared to two other methods using five challenging dot-annotated cell datasets. We first characterize the datasets, then we define performance evaluation metrics and describe the comparative methods. The remainder of this chapter contains a description of the setups for cross-validation (CV) experiments for hyper-parameter selection.

4.1. Datasets

The cell detection approach was evaluated on five challenging histopathological image datasets. Each of these datasets contains dot-annotated cell nuclei centers. The main focus of this work is to accurately localize cells in images of healthy bone marrow. Since there exist many different histochemical stainings for bone marrow, three novel cell datasets stained with the most popular ones (H&E and MGG) were examined first. Additionally, two publicly available datasets were examined to evaluate the performance and stability of the proposed regression method on other tissue types [113], and on a publicly available benchmark dataset [130].

In the previous chapter we defined some hyper-parameters, such as suitable ranges for the input patch size or the radius for the NMS window, that can be derived from certain dataset characteristics. Hence, we used dedicated training sets to estimate suitable hyper-parameters for each of the cell detection methods in CV, cf. Section 4.4. The characteristics of the training sets are summarized in Table 4.1, Figs. 4.1-4.3 show

statistics and samples. In Section 4.4.3, we define benchmark experiments using dedicated test sets for the two bone marrow datasets *BM-HE* and *BM-MGG*, which were not characterized the same way. All microscopic slides were anonymized prior to processing.

Parameters	***BM-HE***	***BM-MGG***	***BM-HE-MK***		***ICPR-BC***	***MT-HE***
Magnification	40×	40×	20×	10×	20×	40×
Staining	H&E	MGG	H&E	H&E	H&E	H&E
Training Set (Hyper-Parameter Selection/Cross-Validation)						
Slides	8	10	24	24	n/a	n/a
Images	11	16	269	269	20	36
Size	$1,200 \times 1,200$	800×800	192×192	96×96	100×100	600×600
Annotations	4,205	3,811	335	335	665	7,931
$(\mu \pm \sigma)$	(382 ± 51)	(238 ± 107)	(1.2 ± 0.5)	(1.2 ± 0.5)	(33 ± 13)	(220 ± 112)
$r_{obj} \pm \sigma$	16.5 ± 3.2	16.5 ± 3.5	24.9 ± 4.6	12.4 ± 2.3	5.7 ± 0.8	14.7 ± 2.4
$n_{r_{obj}}$	220	220†	100	20	20‡	180
d_M	33	33	50	25	11	29
α	3.48	3.48	2.82	4.00	6.03	3.71
$\mu_{\mathrm{d_E}} \pm \sigma_{\mathrm{d_E}}$	31 ± 10	28 ± 10	69 ± 21	34 ± 10	11 ± 4	23 ± 9
Test Set						
Slides	4	4	-	-	-	-
Images	4	4	-	-	-	-
Size	$1,200 \times 1,200$	800×800	-	-	-	-
Annotations	1,382	944	-	-	-	-
$(\mu \pm \sigma)$	(346 ± 56)	(236 ± 122)	-	-	-	-

Table 4.1.: Characteristics of the dot-annotated cell datasets (in columns) used to train and evaluate the cell detection approaches. Both 20× and 10× magnifications of the megakaryocyte dataset (*BM-HE-MK*) were examined, which we list as two separate datasets here. $\mu \pm \sigma$ denotes the mean±SD total number of annotations per image, and $r_{obj} \pm \sigma$ the mean±SD object radius in pixels, estimated from $n_{r_{obj}}$ cells. The mean object size d_M (in pixels) and α control the width and shape of the peaks in the proximity score map, cf. Eq. (3.3). $\mu_{\mathrm{d_E}} \pm \sigma_{\mathrm{d_E}}$ denote mean±SD of the Euclidean distances between a ground truth cell center and its nearest neighbor in pixels. The number of digital slides was not available for *ICPR-BC* and *MT-HE*. Only the *BM-HE* and *BM-MGG* were evaluated in the benchmarks, other datasets did not have dedicated test sets.
† The estimates from *BM-HE* were used here.
‡ The estimates of the 20× magnification were divided by two.

BM-HE Bone marrow tissue samples were obtained from the iliac crest of healthy humans via trephine biopsy during routine examinations, performed at the University Hospital Graz. At the Institute of Pathology, Medical University of Graz, samples were cut in sections of ≈ 1-2 µm thickness, stained with H&E, and digitized at 40× magnifi-

cation (0.245 µm per pixel) using an Aperio ScanScope (Leica Biosystems GmbH) whole slide scanner. The training set contains eleven $1,200 \times 1,200$ pixel images, cropped from eight WSIs, depicting healthy human bone marrow from eight different patients, cf. Fig. 4.1 (b-e). All cell nuclei were labeled as *foreground* by providing the location of the center pixel as dot-annotation. Debris and staining artifacts were labeled as *background*. Ambiguous blob-structures, e.g. when a cell nucleus could not be clearly determined as such, were labeled as *unknown*. However, we treated all ambiguous objects as foreground, since the detection method proposed in this work is supposed to identify these objects as candidates for a subsequent cell type classification, which can be used to filter out these objects. Hence, a total of $4,205$ annotations covers *foreground* and *unknown* labels. The focus of evaluating on this dataset was to detect bone marrow cells of similar size, i.e. all cell types except megakaryocytes, which were therefore not labeled as foreground. This dataset has been used in recent work [8] and has been made publicly available [6]. An additional test set consists of four images from four slides, comprising $1,382$ annotated cells. This test set will be used in the benchmark experiments.

BM-MGG The training subset of this novel dataset contains 16 images of May-Grünwald-Giemsa (MGG) stained healthy human bone marrow tissue from ten patients. Obtaining the biological samples, digitalization of the stained specimen and creation of the ground truth annotations (without megakaryocytes) was performed according to the procedure described for the *BM-HE* dataset. Each image was cropped at a size of 800×800 pixels from ten WSIs, and a total of $3,811$ locations was labeled, cf. Fig. 4.1 (g-j). The goal here was to assess the transferability of the proposed detection method to another common staining besides H&E. An additional test set consists of four images from four slides, comprising 944 annotated cells. This test set will be used in the benchmark experiments.

BM-HE-MK The ability of the method to detect megakaryocytes was evaluated in this separate, also novel, H&E stained cell dataset. Obtaining the biological samples and digitalization of the stained specimen was performed according to the procedure described for the *BM-HE* dataset. While the original magnification was $40\times$, we prepared this dataset at $20\times$ and $10\times$ to make use of prior knowledge of the size and cytological properties of megakaryocytes, cf. Section 3.3. Furthermore, we wanted to examine whether the regression method worked better on either magnification. A total

of 269 images of size 192 × 192 pixels (at 20×), containing at least one megakaryocyte, was cropped from WSIs of 24 individual patients, cf. Fig. 4.2 (b-e). The center locations of the cell nuclei were annotated, resulting in a total of 335 cells. Unlike for the other datasets, the distribution of d_E did not comply to a normal distribution, cf. Fig. 4.2 (a). This was most likely caused by the small number of images that contained two or more cells (most of them contained just a single one). At 20× magnifications, the median (Q_2) of the distribution was at 66 pixels, the first and third quartile were at 53 and at 87 pixels, respectively. These characteristics are closely related to the mean and SD of the distribution: $\mu_{d_E} = 69$ pixels and $\sigma_{d_E} = 21$ pixels. In order to be consistent with the remaining datasets, mean and SD could therefore be selected to derive method hyper-parameters for this dataset as well. However, we can expect that in larger images containing multiple megakaryocytes, the distribution of d_E eventually becomes approximately normally distributed.

ICPR-BC Gurcan *et al.* [130] conducted the ICPR 2010 Pattern Recognition in Histopathological Images Contest and provided a public benchmark dataset of H&E stained breast cancer tissue. It contains 20 100 × 100 pixel images at 20× magnification, cf. Fig. 4.2 (f-j), and facilitates a comparison with other, already published work on cell detection [103–105, 116, 182–184]. The reader is referred to the original publication [130] for more details on the image data. This dataset contains 665 labeled lymphocytes in neoplastic tissue, and has been included into the method evaluations to show the broad applicability of the proposed regression method. Furthermore this dataset is of relevance, because lymphocytes originate from hematopoietic stem cells in the bone marrow, cf. Chapter 2. In addition to their accurate localization, a major challenge in this dataset is that lymphocytes must be discriminated from cancer cells, which are sometimes highly similar in appearance.

MT-HE Wienert *et al.* [113] provided a cell dataset containing benign and malignant cases of multiple tissue types: breast cancer, bone marrow, normal liver and kidney tissue, and intestinal mucosa. It contains 36 images at 20× magnification, each cropped from WSI and tissue microarrays to measure 600 × 600 pixels , cf. Fig. 4.3 (b-e). The reader is referred to their publication [113] for more details on the acquisition protocol. The 7,931 labeled cell nuclei are distributed over a great variety of tissue samples, which poses a significant challenge to the generalization requirement of a single, generic cell detection method.

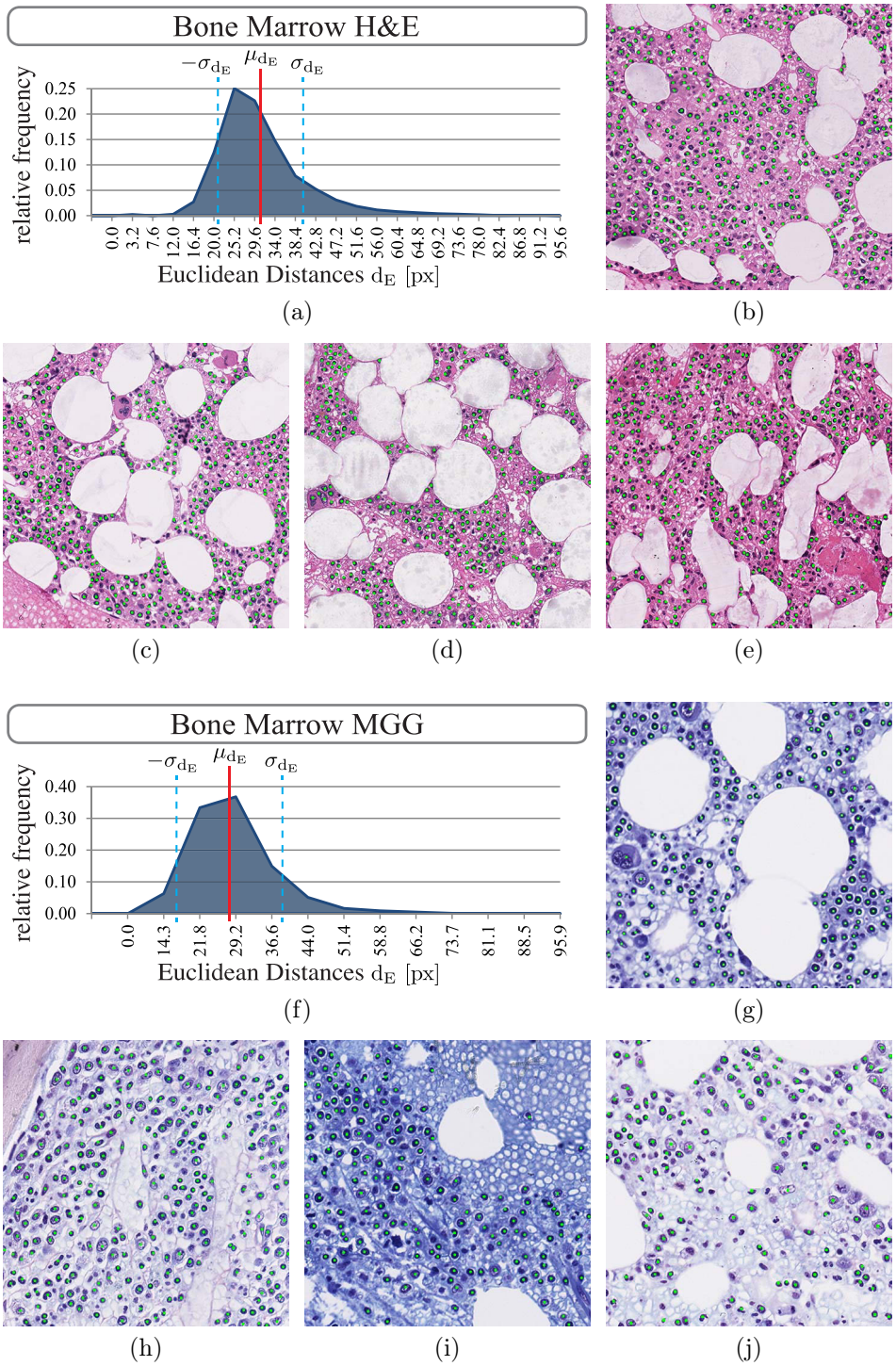


Figure 4.1.: Statistics and samples of the *BM-HE* (a-e) and *BM-MGG* (f-j) datasets. (a,f) Histograms of Euclidean distances d_E between a cell center and its nearest neighbour, $d_E > 100$ pixels are not shown. Mean±SD ($\mu_{d_E} \pm \sigma_{d_E}$) are illustrated as solid red and dashed blue lines. (b-e, g-j) Samples from the datasets at 40× magnification. Green dots denote the ground truth cell center annotations.

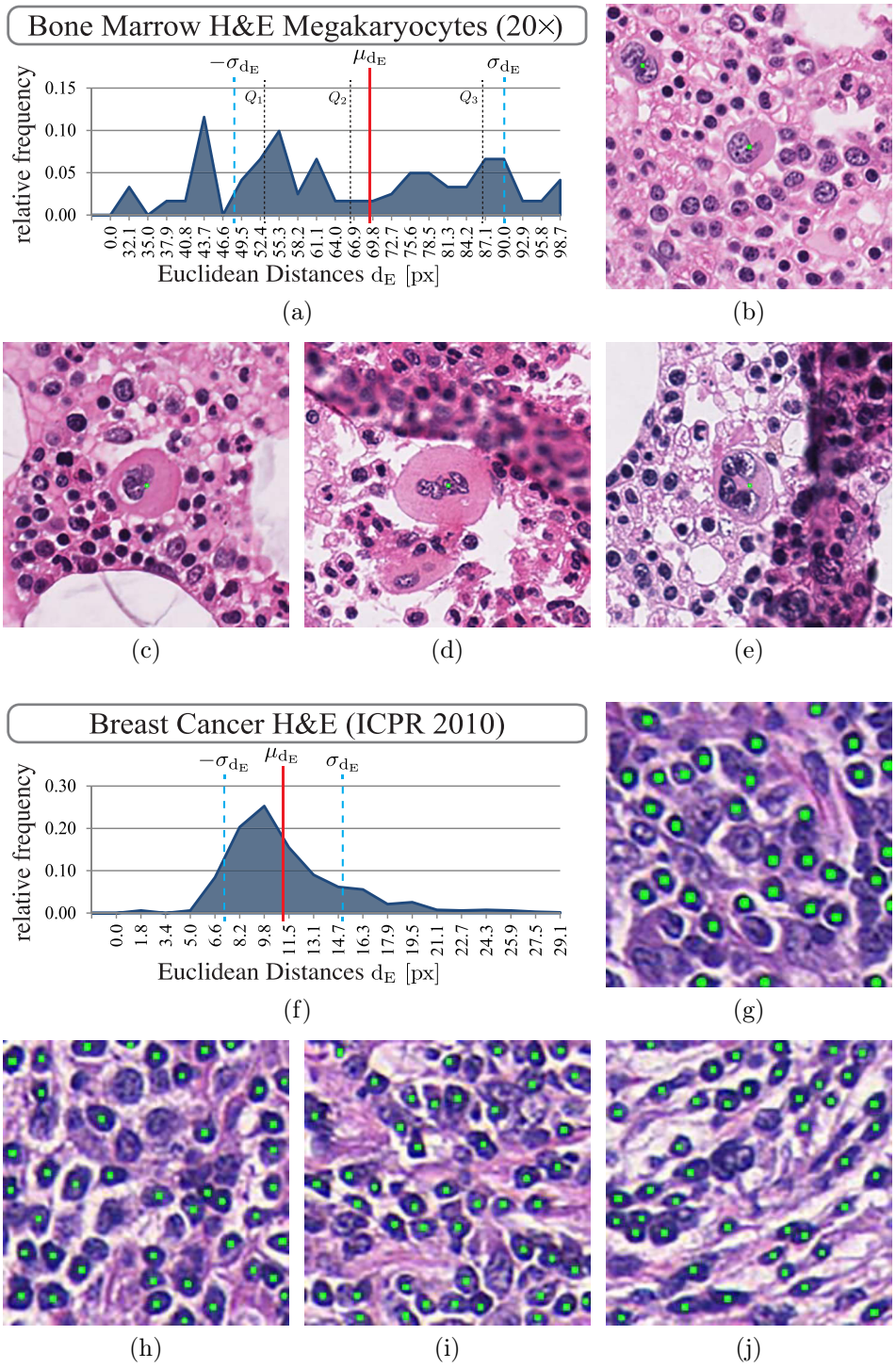


Figure 4.2.: Statistics and samples of the *BM-HE-MK* (a-e) and *ICPR-BC* (f-j) datasets. (a,f) Histogram of Euclidean distances d_E between a cell center and its nearest neighbour, $d_E > 100$ pixels are not shown in (a), $d_E > 30$ are not shown in (b). Mean±SD ($\mu_{d_E} \pm \sigma_{d_E}$) are illustrated as solid red and dashed blue lines. Please note that the shape of the histogram in (a) also applies to the *BM-HE-MK* dataset at 10× magnification, but all values on the x-axis need to be divided by two. Q_1-Q_3 denote the first, second and third quantile of the distribution. (b-e, g-j) Samples from the datasets at 20× magnification. Green dots denote the ground truth cell center annotations.

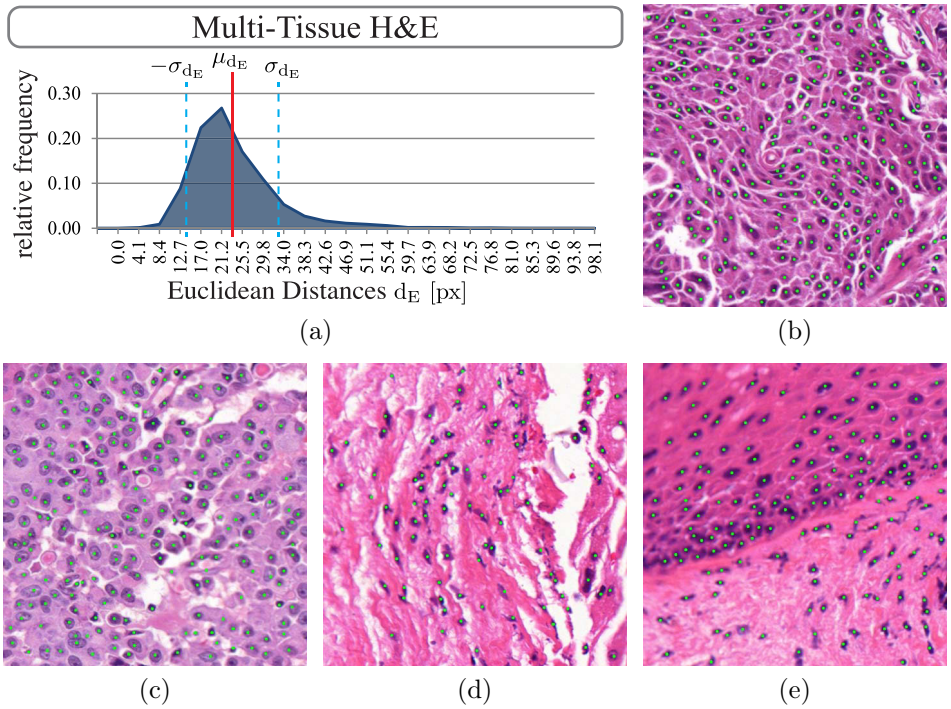


Figure 4.3.: Statistics and samples of the *MT-HE* dataset. (a) Histogram of Euclidean distances d_E between a cell center and its nearest neighbour, $d_E > 100$ are not shown. Mean±SD ($\mu_{d_E} \pm \sigma_{d_E}$) are illustrated as solid red and dashed blue lines. (b-g) Samples from the dataset at 40× magnification. Green dots denote the ground truth cell center annotations.

4.2. Localization Performance Evaluation Metrics

The localization performance is reported quantitatively by different measures. In order to associate a detection hypothesis with a ground truth location, a maximum distance threshold ξ between a ground truth center and a hypothesis was introduced for the evaluations. Let us define TP as true positive, FP as false positive, and FN as false negative detections. If the distance between a detection and a ground truth annotation is $\leq \xi$, we count the detection as TP. If more than one detection hypothesis is present in this area, we assign the most confident one (i.e. the highest proximity score) to the ground truth location and consider the others as FP. A detection hypothesis farther away than ξ from any ground truth location is counted as FP, and all ground truth annotations without any sufficiently close detection hypotheses are FN.

Overall detection performance is evaluated as precision (PRC), recall (REC), and F1-score:

$$\mathrm{PRC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{4.1}$$

$$\mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{4.2}$$

$$\mathrm{F1} = \frac{2 \cdot \mathrm{PRC} \cdot \mathrm{REC}}{\mathrm{PRC} + \mathrm{REC}}. \tag{4.3}$$

Further, the average absolute difference and standard deviation (SD) $\mu_n \pm \sigma_n$ between the number of ground truth annotations ($|\mathcal{C}|$) and correct detection hypotheses (TP) is assessed with $||\mathcal{C}| - \mathrm{TP}|$. The localization accuracy in the spatial domain is reported using the average Euclidean distance and SD $\mu_d \pm \sigma_d$ between a TP detection and its correctly assigned ground truth location.

4.3. Comparative Methods

4.3.1. Classification Random Forest

As first comparison method, a standard classification RF was trained as pixel-wise classifier to predict the posterior distribution of the class labels on the center pixel of a given image patch. To facilitate a direct comparison, input data was represented by the very same 54 image features. Since the goal here is the evaluation of the performance difference that can directly be attributed to classification instead of regression, all hyper-parameters were inherited from the search performed for the regression RF that will be detailed in Section 4.4.1.

The classifier learned a binary classification task using a discrete target variable. Hence, Eq. (3.8) could not be used to evaluate the quality of the split decision. Further, the dataset now took the form of $\mathcal{P} = \left\{\left(\mathbf{x}^{(i)}, k^{(i)}\right)\right\}_{i=1}^{N_\mathcal{P}}$, with $N_\mathcal{P}$ as the number of training samples. Each training patch $\mathbf{x}^{(i)}$ was labeled with a corresponding discrete class label $k^{(i)} \in \{0, 1\}$, where 0 denotes background and 1 denotes the center of a cell (foreground), hence the classification function takes the form of $f : \mathbb{R}^d \rightarrow \{0, 1\}$. A commonly used criterion for optimal split decision at a non-terminal node in decision trees (e.g. ID3 [185] and C4.5 [186]) is the information gain (IG), denoted as

$$IG(\phi) = \mathcal{H}(\mathcal{P}) - \sum_{s \in \{L,R\}} w_s \, \mathcal{H}(\mathcal{P}_s(\phi)), \tag{4.4}$$

where w_s denotes the weights of the subsets determined by Eq. (3.5), and $\mathcal{H}$ denotes the entropy.

The goal in classification trees is to minimize an objective that measures the class impurity in a dataset, and hence the quality of the split decision. This optimization objective is frequently defined by the entropy $\mathcal{H}$, denoted as follows for a binary classification problem:

$$\mathcal{H}(\mathcal{P}) = -\sum_{k=0}^{1} p(c=k|\mathcal{P}) \ \log_b \left[p(c=k|\mathcal{P})\right], \tag{4.5}$$

where $p(c=k|\mathcal{P})$ is the probability that an instance in set $\mathcal{P}$ belongs to class k, and b is a logarithmic base (usually 2, e, or 10).

In this thesis we consider the Gini index (sometimes also termed Gini impurity), known from standard CART systems [178], as alternative to the entropy for class impurity of two subsets:

$$Gini(\mathcal{P}) = \sum_{k=0}^{1} p(c=k|\mathcal{P}) \ p(c\neq k|\mathcal{P}). \tag{4.6}$$

Substituting $\mathcal{H}(\mathcal{P})$ with $Gini(\mathcal{P})$ in Eq. (4.4) results in the Gini gain

$$GG(\phi) = Gini(\mathcal{P}) - \sum_{s\in\{L,R\}} w_s \ Gini(\mathcal{P}_s(\phi)). \tag{4.7}$$

To maximize $GG(\phi)$, we select and store at each non-terminal node the optimal split function

$$\phi^* = \arg\max_{\phi} GG(\phi). \tag{4.8}$$

Using Gini index and entropy as impurity measures may result in different ϕ^*. However, they yield qualitatively comparable results. Furthermore, the Gini index facilitates reduced execution times, since the computation of $\log[\cdot]$ is not required. Once certain criteria apply to stop the growth of the tree (cf. Section 3.2.2), a leaf node is created. The leaf stores the class label histogram over all instances.

To infer a class label for an unseen image patch $I(\mathbf{u})$, it is propagated down each tree that predicts the posterior probability distribution over the labels, i.e. returns the class label histogram. The averaged class label distribution resulting from the classification RF is given by

$$\bar{p}(c=k|I(\mathbf{u})) = \frac{1}{T}\sum_{t=1}^{T} p_t(c=k|I(\mathbf{u})), \tag{4.9}$$

where $p_t(c=k|I(\mathbf{u}))$ is the probability distribution over the labels predicted by the

t-th tree. Taking the most likely label results in the final class prediction

$$\hat{k} = \underset{k \in \{0,1\}}{\arg\max}\, \bar{p}(c = k | I(\mathbf{u})). \tag{4.10}$$

A probability map $\bar{\mathbf{p}}$ for an image I is obtained by applying f on each image location $\mathbf{u}$ in a test image using a sliding window and storing

$$\max_{k \in \{0,1\}} \bar{p}(c = k | I(\mathbf{u})), \forall \mathbf{u} \in \Omega. \tag{4.11}$$

This has several advantages over using Eq. (4.10), where we would obtain a non-smooth, discrete label map, similar to a segmentation map, with a significant amount of (shot) noise that is not well-suited for this object localization task. Detecting the cell centers from the smooth probability map is easier and was performed according to the post-processing procedure described in Section 3.2.3. Henceforth, we refer to this method as *cRF*.

Training Data Augmentation The regression target was defined for multiple locations around an actual cell center, which results in a quite extensive training dataset, if $\tau_{bg} < 1$. To provide a fair comparison between the RF predictors, foreground patches in the classification forest were augmented by three rotations (multiples of 90°) as well as horizontal, vertical, and diagonal flipping. Furthermore, the training dataset was fully balanced using additional, randomly sampled patches from non-foreground locations in the training images.

4.3.2. MSER-SSVM

The second method we compare our approach to is one of the state-of-the-art methods in non-overlapping cell detection, that also learns from dot-annotations on cell objects.

Arteta *et al.* [116] presented an approach based on the classification of maximally stable extremal image regions (MSER [131, 132]) via structured support vector machines (SSVMs), that proceeds in three steps. First, a large set of cell-nuclei-like candidates is extracted by the MSER detector, which reveals image regions that are coherent across different levels of intensity thresholding. Then, the appearance of these candidate regions is characterized by a statistical feature vector $\mathbf{f}$, and subsequently scored by

evaluating a linear SVM classifier ($\mathbf{w}_b \cdot \mathbf{f}$). The training set for this binary classifier consists of image regions that contain a single ground truth dot-annotation as positive class ($+1$), and all other regions as negative class (-1). The features comprise intensity histograms, a shape descriptor, and an area descriptor. The binary SVM produces a new ground truth configuration for the subsequent structured learning approach by scoring all regions. For each ground truth dot, only the region containing that dot and having the highest score is selected as a positive sample. Using a non-overlapping constraint and the new ground truth dataset, the weights $\mathbf{w}_s$ of an SSVM classifier are optimized to discriminate between cell objects and non-cell regions. The learning objective is implemented using a custom loss function that models the deviation from the one-to-one correspondences between the ground truth dots and regions that contain only that dot. Finally, they employ dynamic programming to select the optimal subset of all scored, non-overlapping regions stored in a tree structure, such that the sum of the scores is maximized over the entire training set [116].

Given an unseen image and the learned weight vector, the method first predicts the score of each region revealed by the MSER detector using $\mathbf{w}_s \cdot \mathbf{f} + b$. Finally, it infers the locations of the cell nuclei centers as the centroids of the best non-overlapping subset of regions from the trees. The inference bias b is able to influence the selection of the subsets by steering the performance towards more precision, or recall, respectively.

The authors [116] provided a publicly available implementation of their method[1], which was used here. Henceforth, we refer to this method as *SSVM* and describe the choice of hyper-parameters in Section 4.4.2.

4.4. Experiments

Selection of suitable hyper-parameters is a necessary step to explore the performance range of the cell localization strategy, which consists of a learning part (RF) and a post-processing part (NMS). The hyper-parameter search was performed across reasonable parameter ranges, with the objective to optimize a particular performance measure.

However, to assess the generalization performance on unseen test images, distinct experiments are performed on each available dataset. The following sections detail the

[1] Available from: http://www.robots.ox.ac.uk/~vgg/software/cell_detection.

parameter ranges and experimental setups for the evaluations, while the results are reported in Chapter 5.

4.4.1. Random Forest Hyper-Parameter Selection

Let $\Lambda = \left\{\lambda^{(l)}\right\}_{l=1}^{L}$ be the set of L possible hyper-parameter combinations, where $\lambda^{(l)} = \left\{\lambda_0^{(l)}, \lambda_1^{(l)}, \ldots, \lambda_k^{(l)}, \ldots, \lambda_{n-1}^{(l)}\right\}$ is the l-th n-dimensional hyper-parameter vector. All experiments to search for the optimal hyper-parameters λ^* were carried out in CV. In the case of the *BM-HE*, and *BM-MGG* dataset, we selected a leave-one-out cross-validation (LOOCV) over a 10-fold CV due to the small total number of images. For the *ICPR-BC*, *BM-HE-MK* and *MT-HE* dataset a 10-fold CV was performed. For datasets, where the total number of images N_I was not a multiple of 10, we used $\lfloor N_I/10 \rfloor$ samples in the test set of the first nine iterations, and the remaining $N_I - 9 \cdot \lfloor N_I/10 \rfloor$ test samples in the final iteration.

Exploring the entire high-dimensional space Λ in a dense grid search for all possible combinations of hyper-parameters was not feasible, because training many predictive models in CVs is computationally expensive on big datasets. At each depth level of the trees, the number of nodes increases exponentially such that including the root nodes a forest may consist of up to $\sum_{t=1}^{T}\sum_{d=1}^{T_{md}} 2^{(d-1)}$ split nodes (at a theoretical maximum). Depending on the settings used for randomized node optimization, i.e. N_ϕ and M_j, this could take a considerable amount of time. Therefore, the search for $\lambda^* \in \Lambda'$ was designed incrementally to explore a reduced parameter space $\Lambda' \subset \Lambda$ in three stages. This strategy still facilitated using a grid search, which is a simple method to search for a proper solution in a reproducible way [187]. In the first two stages the post-processing parameters were fixed, and the RF parameters were optimized to produce a proximity score map from an image that is as close as possible to the ground truth. The performance resulting from a parameter vector is measured by the normalized root mean squared error (NRMSE). For a given ground truth image $\mathbf{y}$, prediction image $\hat{\mathbf{y}}$ and parameter vector $\lambda^{(l)}$, it is defined as

$$\mathrm{NRMSE}\left(\mathbf{y}, \hat{\mathbf{y}}, \lambda^{(l)}\right) = \frac{1}{y_{max} - y_{min}} \sqrt{\frac{1}{|\Omega|}\sum_{i=1}^{|\Omega|}(y_i - \hat{y}_i)^2}, \tag{4.12}$$

with y_i and $\hat{y}_i$ as the ground truth and predicted value at pixel i, y_{min} and y_{max} as the minimum and maximum value of the target variable, and $|\Omega|$ as the total number of

pixels in the image. The best hyper-parameters were selected according to the minimal average error over all $m = 1, \ldots, N_I$ images in a dataset:

$$\lambda^* = \underset{\lambda^{(l)} \in \Lambda'}{\arg\min} \left[\frac{1}{N_I} \sum_{m=1}^{N_I} \mathrm{NRMSE}\left(\mathbf{y}_m, \hat{\mathbf{y}}_m, \lambda^{(l)}\right) \right]. \tag{4.13}$$

In the final stage, post-processing hyper-parameters were explored and the RF hyper-parameters were fixed to the optimal values resulting from the previous stage. Over all N_I images in a dataset, the mean of the area under the average precision-recall curve (AUC)[2] and average F1-score[3] was considered as the optimality criterion:

$$\mathcal{O}\left(\lambda^{(l)}\right) = \frac{\overline{\mathrm{AUC}}\left(\lambda^{(l)}\right) + \overline{\mathrm{F1}}\left(\lambda^{(l)}\right)}{2}. \tag{4.14}$$

The best hyper-parameters maximized this criterion:

$$\lambda^* = \underset{\lambda^{(l)} \in \Lambda'}{\arg\max}\, \mathcal{O}(\lambda^{(l)}). \tag{4.15}$$

Opting for this joint optimality criterion has the advantage to include both the entire possible performance range (AUC) and the maximum F1-score that can be achieved by a method.

Stage I

The width of the peaks d_M in the proximity score map is modeled as a function of the average object size to cover areas in the adjacency of an annotated cell center. For each dataset, three input patch sizes $p_{in} \in \{0.5 \cdot d_M, d_M, 1.5 \cdot d_M\}$ were examined empirically. These definitions covered cases, where an area (i) less than the average object size, (ii) equal to the average object size, and (iii) larger than the average object size is depicted on a single patch[4]. The training patches were sampled using $\tau_{bg} = 0.5$ for all datasets but *ICPR-BC*, where we used $\tau_{bg} = 0.1$ due to the smaller average object size.

[2] To construct the average precision-recall curve over all N_I images, detection results (TP, FP, FN) in all images were pooled and the mean curve and its AUC were computed. Please note that this value therefore deviates from the one obtained by averaging the AUC measures of the individual images.

[3] Computed as the mean F1-score over all N_I images.

[4] Initial experiments on the *BM-HE* dataset showed that very small and very large input patch sizes result in higher NRMSE, see Fig. A.1 (a) for details on experiments with $p_{in} \in \{6, 128\}$.

			BM-HE-MK			
λ_k	*BM-HE*	*BM-MGG*	20×	10×	*ICPR-BC*	*MT-HE*
d_M	33	33	50	25	11	29
τ_{bg}	0.5	0.5	0.5	0.5	0.1	0.5
p_{in}	$\{17, 33, 50\}$	$\{17, 33, 50\}$	$\{25, 50, 75\}$	$\{13, 25, 38\}$	$\{6, 11, 17\}$	$\{15, 29, 44\}$
p_{out}	$\{3, 7, 11\}$	$\{3, 7, 11\}$	$\{3, 7, 11\}$	$\{3, 7, 11\}$	$\{3, 7, 11\}$	$\{3, 7, 11\}$
r_{NMS}	8	8	13	6	3	7
ξ	21	18	48	24	7	14

Table 4.2.: Choices of hyper-parameters for all evaluated datasets in search stage I. The first rows correspond to hyper-parameters λ_k common to both the single-target and spatial-averaging regression, while the output patch size p_{out} was evaluated for each p_{in} in spatial-averaging regression only. Across all datasets, some parameters were fixed: $T = 16$, $T_{md} = 16$, $\sigma_{\mathcal{G}} = 2$.

All remaining hyper-parameters for the RF and post-processing were fixed across all datasets at this stage, using existing values from earlier experiments performed in Kainz *et al.* [8]: $T = 16$, $T_{md} = 16$, $\sigma_{\mathcal{G}} = 2$ (i.e. a 5×5 pixel kernel[5]), $r_{NMS} = \lfloor 0.5 \cdot r_{obj} \rceil$, where $\lfloor \cdot \rceil$ denotes the rounding operator to the nearest integer, see Table 4.2 for details. For each selected split function 20 random thresholds were tested. Tree growing stopped, if either T_{md} or a minimum number of remaining samples $N_{\mathcal{P}} \leq 20$ was reached.

In each experiment, a proper threshold for the 'confidence' of a hypothesized center is screened in the possible output value range $[0, 1]$ using a step size of 0.02. By varying κ, a precision-recall curve was obtained and the AUC was computed. The optimal input patch size p^*_{in} was picked by using Eq. (4.13), while the best threshold κ^* globally maximizes the F1-score over all predicted images. For all evaluations in stage I, the maximum accepted distance between a detection hypothesis and a ground truth location was set rather loosely to $\xi = \mu_{\mathrm{d_E}} - \sigma_{\mathrm{d_E}}$.

The *spatial-averaging* regression approach required the additional search for a suitable square output patch size $p_{out} \in \{3, 7, 11\}$, while satisfying $p_{out} \leq p_{in}$ (full-patch regression). The maximum output patch size was limited to 11×11 pixels, since computation times and memory consumption increased significantly in initial proof-of-principle experiments for larger sizes. The rest of the experimental setup was identical to the *single-target* regression.

[5] The Gaussian kernel size depends on the selected SD: $(2 \cdot \sigma_{\mathcal{G}} + 1) \times (2 \cdot \sigma_{\mathcal{G}} + 1)$ pixels [188].

			BM-HE-MK			
λ_k	*BM-HE*	*BM-MGG*	20×	10×	*ICPR-BC*	*MT-HE*
r_{obj}	17	17	25	12	6	15
$\mu_{\mathrm{d_E}}$	31	28	69	34	11	23
r_{NMS}	$\{8, 14, 20, 26, 31\}$	$\{8, 13, 18, 23, 28\}$	$\{12, 18, 24, 30, 35\}$	$\{6, 13, 20, 27, 34\}$	$\{3, 5, 7, 9, 11\}$	$\{7, 11, 15, 20, 23\}$
ξ	$\{21, 11, 5\}$	$\{18, 8, 5\}$	$\{48, 27, 5\}$	$\{24, 14, 5\}$	$\{7, 3\}$	$\{14, 5\}$

Table 4.3.: Choices of hyper-parameters for all evaluated datasets in search stage III. For all datasets, Gaussian kernels defined by $\sigma_{\mathcal{G}} \in \{2, 4, 6, 8\}$ were examined and the minimum distance between a ground truth and hypothesized center was set to different ξ.

Stage II

This stage evaluated, whether the initially chosen model complexity was sufficient, needed be increased, or could even be decreased. Therefore, the best input patch size p^*_{in} and output patch size p^*_{out}, respectively, from stage I were selected, and additional RF parameter combinations were evaluated in a grid search: $T \in \{4, 16, 64\}$, and $T_{md} \in \{8, 16, 24\}$. The post-processing parameters remained as defined for stage I, and again evaluating Eq. (4.13) determined a suitable model complexity.

Stage III

In this final stage of the hyper-parameter search, post-processing settings were evaluated for the best performing RF model resulting from stage II. Since the post-processing hyper-parameters rely on heuristics, they are considered more sensitive than the RF hyper-parameters. However, compared to training the RF models, this search was computationally inexpensive and allowed us to perform a denser grid search, cf. Table 4.3.

Given a predicted proximity score map $\hat{\mathbf{y}}$, we first searched for proper zero-mean Gaussian kernels with $\sigma_{\mathcal{G}} \in \{2, 4, 6, 8\}$ to smooth $\hat{\mathbf{y}}$. Secondly, the NMS window radius r_{NMS} was adjusted for each dataset according to the bounds derived in Section 3.2.3.

The maximum accepted distance between a ground truth and hypothesized center was varied to obtain different levels of confidence for the spatial localization accuracy: $\xi \in \{\mu_{\mathrm{d_E}} - \sigma_{\mathrm{d_E}}, \mu_{\mathrm{d_E}} - 2 \cdot \sigma_{\mathrm{d_E}}\}$. The predictions of the center locations were considered more accurate, when in the absence of significant performance loss, ξ could be set to a lower value. Hence, for some datasets, an additional $\xi = 5$ was examined that enabled

a more rigorous definition of TP detections. The optimal set of hyper-parameters λ^* for the entire cell localization method was determined by evaluating Eq. (4.15), and subsequently used in the experiments for benchmarking the methods on the test sets.

4.4.2. MSER-SSVM Hyper-Parameter Selection

In their original work [116], the authors used a feature vector $\mathbf{f}$ consisting of 92 statistical parameters to describe each region. We extended that number to a total of 151 features: 15 for area description, four intensity histograms of 15 bins each for both the mean RGB image and the three RGB color channels. Moreover, 16 features described the difference between the region and its context (eight bins, two scales), and 60 features described the shape using a rotationally invariant contour points distribution histogram [189] (12 angles, five radii).

Despite the method has more hyper-parameters such as particular SVM settings, we did not perform a grid search to explore the parameter space, but relied on the originally provided implementation that earlier achieved state-of-the-art [116] performance on the *ICPR-BC* dataset, and fairly reasonable performance on the *BM-HE* dataset in more recent work [8]. However, for each dataset the MSER detector[6] was tuned towards higher performance by manually exploring hyper-parameter settings. It detects a set of candidate regions, specified by a minimum and maximum area, from a Gaussian-smoothed version of a given grey-scale image. The *dark-on-bright* detector setting was used, since the cell nuclei in both H&E and MGG histopathology images are usually stained darker than their environment and background. The individual results of manual tuning are not reported here.

SSVM parameters were chosen such that false positives get maximally penalized during structured learning. Nevertheless, varying the inference bias b of the model $(\mathbf{w}_s \cdot \mathbf{f} + b)$ allowed us to influence the precision and recall after learning. Hence, a precision-recall curve was obtained by varying b in the range $[-3, 6]$ using a step size of 0.25. The AUC was computed over all images in CV.

[6] The VLFeat [190] implementation of the MSER detector is used in the *SSVM* method [116].

4.4.3. Method Benchmarks

The experimental setup for comparing the four methods to detect cells (*single-target* regression *rRF*, *spatial-averaging* regression *srRF*, (binary) classification *cRF*, and *SSVM*) are defined in this section. The *BM-HE*, *BM-MGG* and *ICPR-BC* datasets consist of a dedicated training and test set, which were used for the benchmark experiments. Using the best settings resulting from the hyper-parameter search, each method is trained on all available training images, and evaluated on the held-out test images. Results are reported as precision-recall curves and AUCs as well as the performance metrics for object detection defined in Section 4.2.

In the bone marrow datasets, some training images have been cropped from the same WSI. However, it was ensured that the test set contained images from previously unseen variations of staining, texture and tissue architecture. To test the generalization ability of the methods to such variations, each individual image in the test set originated a separate histopathological slide, i.e. patient. For the *BM-HE* dataset, this strategy resulted in a training set of eleven images from eight WSIs, and a test set of four images. Sixteen images from ten slides were used for training in the *BM-MGG* dataset, while four images comprised the test set.

An estimation of the generalization ability using a dedicated test set was only performed for the *BM-HE* and *BM-MGG* datasets. However, the dataset split defined in the ICPR 2010 Contest [130] has been considered for the *ICPR-BC* dataset, and results from earlier work [8] are presented and discussed. No benchmark tests were performed for the *BM-HE-MK* and *MT-HE* dataset, because there was not enough data available for the *BM-HE-MK* dataset. On the other hand, the original work of Wienert *et al.* [113] used the *MT-HE* dataset, but did not define a dataset split that could be used in a comparison.

All methods except the *SSVM* extensively rely on random processes during dataset sampling and tree node optimization. Hence, the stability of each RF-based method is evaluated in ten additional independent runs on the available benchmark datasets to acquire a robust estimate of the expected performance.

5. Cell Detection Results

In this chapter, the results of four cell detection methods across five challenging histopathology datasets are presented. For each dataset, the hyper-parameter selection results for the novel proximity score regression methods *rRF* and *srRF* are described first. A grid search across a multitude of combinations was performed to select proper hyper-parameters on the training part of each dataset in cross-validation, cf. Section 4.4.1. Then, direct comparisons of the proposed proximity score regression, a standard binary pixel-wise classification (*cRF*), and the *SSVM* method of Arteta *et al.* [116] are provided according to the experiment definitions in Section 4.4. Significant results are presented here, whereas supplementary information, especially on intermediate results of the hyper-parameter search, can be obtained from Appendix A.

5.1. Model Evaluations

5.1.1. Bone Marrow (H&E)

Hyper-Parameter Selection

In stage I, an input patch size of $p^*_{in} = 33$ pixels resulted in the lowest NRMSE (0.0959) for *rRF*. The *srRF* achieved the lowest NRMSE (0.0769) for an input and output patch size of $p^*_{in} = 33$ pixels and $p^*_{out} = 11$ pixels, respectively. Fig. A.1 illustrates the results of the other stage I hyper-parameter configurations. F1-score is rather stable, regardless of the chosen patch sizes. Though, the patch sizes seemed to determine the trade-off between precision and recall.

Table 5.1 contains the performance measures obtained from stage II and stage III search. Both *rRF* and *srRF* achieved the lowest NRMSE using a forest of $T = 64$

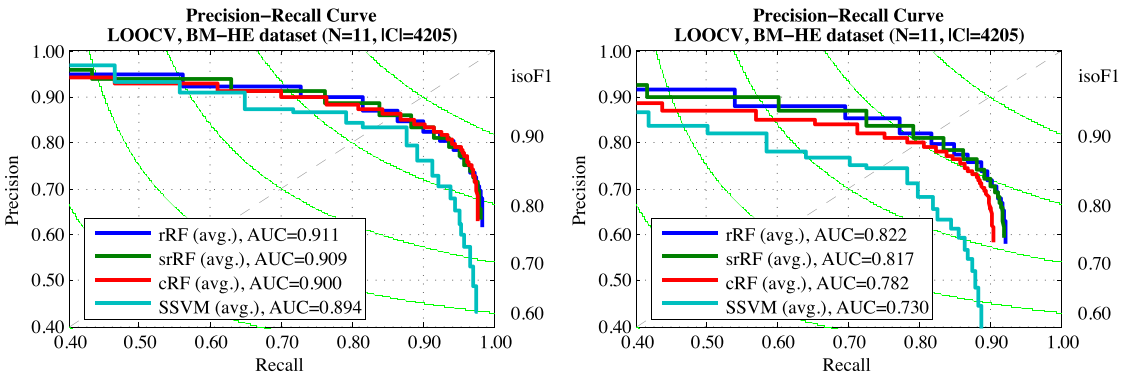
Method	NRMSE	AUC	κ^*	PRC	REC	F1	$\mu_d \pm \sigma_d$	$\mu_n \pm \sigma_n$
				Stage II				
rRF	**0.0905**	0.8879	0.5686	0.7899	0.9270	0.8530	2.54 ± 1.97	27.91 ± 6.79
srRF	**0.0753**	0.8971	0.4706	0.7971	0.9275	0.8573	2.54 ± 1.91	27.73 ± 7.72
Method	$\mathcal{O}$	AUC	κ^*	PRC	REC	F1	$\mu_d \pm \sigma_d$	$\mu_n \pm \sigma_n$
				Stage III				
rRF	**0.8919**	0.9110	0.4706	0.8463	0.9008	0.8727	2.44 ± 1.49	37.91 ± 12.78
srRF	**0.8903**	0.9087	0.3726	0.8324	0.9153	0.8719	2.45 ± 1.51	32.36 ± 11.71
cRF	–	0.9002	0.7451†	0.8239	0.9265	0.8722	2.59 ± 0.28	28.09 ± 13.03
SSVM	–	0.8943	0.0000‡	0.8333	0.8779	0.8550	2.48 ± 1.67	70.27 ± 18.06

Table 5.1.: *BM-HE* dataset: stage II and III performance summary of the detection methods. NRMSE, precision, recall and F1-score are reported as mean over all images in LOOCV. In stage III, results are reported for $\xi = 11$.

† Confidence (probability) threshold for a cell center (*cRF*).

‡ Best prediction bias (*SSVM*).

trees, each with a maximum depth of $T_{md} = 24$. Although we could observe a significant difference in the proximity score prediction errors (-0.0152) between the two methods, their detection-specific performance measures were virtually identical. This indicates that the *srRF* method was able to approximate the proximity map more accurately than *rRF*. The AUC was slightly higher for the *srRF*, due to higher precision at lower recall, cf. Fig. A.2 (d).



(a) $\xi = 11$ (b) $\xi = 5$

Figure 5.1.: *BM-HE* dataset: stage III precision-recall curves for the best hyper-parameter settings of all detection methods. Curves and AUCs represent averages over the LOOCV runs, evaluated for various ξ. We could not observe significant differences for $\xi \geq 11$, all methods worked reasonably well. However, if only hypotheses within five pixels around a ground truth location were considered as true positives, the regression methods performed more reliable than *cRF* and *SSVM*.

The optimal performance on this dataset in stage III was achieved using a Gaussian

kernel width of $\sigma_{\mathcal{G}} = 6$, and a non-maximum suppression radius of $r_{NMS} = 8$. In Fig. 5.1, the localization accuracy in terms of tolerances ξ is illustrated. All methods were evaluated in LOOCV on eleven images in the training set. Results for $\xi \geq 11$ were highly similar. However, it had been shown in earlier work [8] that setting $\xi = 8$ for this dataset resulted in a lower AUC measure of the *SSVM* (0.869) when compared to the RF methods, which stayed similarly high (0.905). In this work, an even stricter tolerance of $\xi = 5$ was examined, cf. Fig. 5.1 (b), and it could be observed that both regression methods performed more reliable than binary classification and the *SSVM* method. Despite at $\xi = 11$ all methods resulted in high AUC measures, the regression methods resulted in superior performance when setting $\xi = 5$.

Method Benchmark and Stability Analysis

The direct comparison of all methods on the *BM-HE* test set showed that all cell detection methods achieved acceptable performance, cf. Table 5.2 and Fig. 5.2. For all evaluations $\xi = 11$ was used as the required distance between a ground truth and a true positive detection. All three RF-based methods expressed very stable F1-scores over ten independent runs, indicating that the influence of the stochastic optimization procedure in the split nodes of the trees is negligible with respect to the ultimate detection performance. The highest variability of precision and recall was observed for *srRF*, the lowest for *cRF*.

Method	PRC	REC	F1	$\mu_d \pm \sigma_d$	$\mu_n \pm \sigma_n$
			Benchmark Results		
rRF	0.8500	0.9059	**0.8771**	2.22 ± 1.51	32.50 ± 19.07
srRF	**0.8835**	0.8669	0.8751	$\mathbf{2.18 \pm 1.48}$	46.00 ± 27.45
cRF	0.8237	**0.9298**	0.8736	2.34 ± 1.61	$\mathbf{24.25 \pm 15.13}$
SSVM	0.8479	0.8859	0.8665	2.21 ± 1.61	38.75 ± 14.57
			Stability Analysis (10 Independent Runs)		
rRF	0.8510 ± 0.0097	0.9055 ± 0.0120	0.8773 ± 0.0012	2.21 ± 1.52	32.65 ± 20.83
srRF	0.8543 ± 0.0209	0.8999 ± 0.0225	0.8760 ± 0.0007	2.23 ± 1.51	34.60 ± 20.55
cRF	0.8223 ± 0.0054	0.9285 ± 0.0074	0.8721 ± 0.0017	2.37 ± 1.60	24.70 ± 15.41

Table 5.2.: *BM-HE* dataset: benchmark and stability analysis results on the held-out test set ($N = 4$, $|C| = 1,382$). All models were evaluated using $\xi = 11$. Stability analysis is reported for RF based methods in the lower panel as mean±SD.

In the benchmark experiment, all methods resulted in similarly high F1-scores. Nevertheless, the RF methods outperformed the *SSVM* on all metrics. The highest recall,

and therefore the lowest number of FN hypotheses was obtained by *cRF*. The highest precision was achieved by *srRF*, while *rRF* resulted in the best F1-score. In terms of spatial localization accuracy, the *srRF* could achieve the lowest average distance of a true positive detection to its corresponding ground truth center. Moreover, *srRF* also achieved the lowest recall, and *cRF* the lowest precision. The AUC measures, however, show that on this test set, the regression methods perform best, and *SSVM* ranks third, being superior over the binary classification approach, cf. Fig. 5.2.

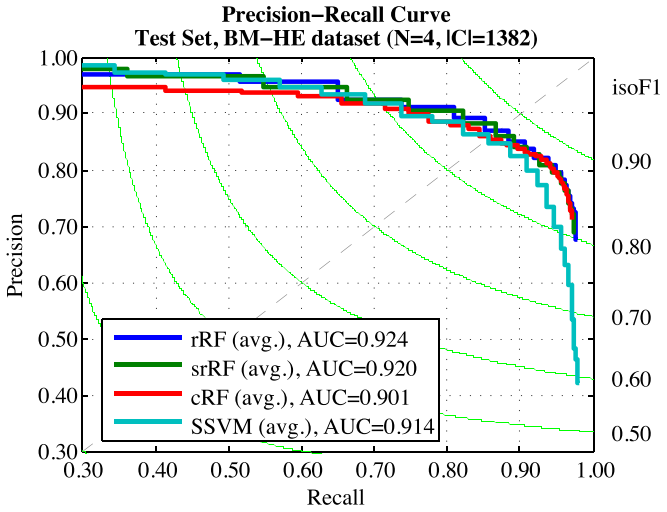


Figure 5.2.: Precision-recall curves for all methods on the *BM-HE* test set, using $\xi = 11$.

Fig. 5.3 depicts qualitative results of all four detection methods in columns for three test samples (A, B, C). Results were obtained using $\xi = 11$. Odd rows show a cropped part of a test image, with ground truth cell nuclei centers illustrated as green dots. True positive detections are denoted by blue x's, false positives by red x's, and false negatives by magenta boxes enclosing the green dots. Even rows show the corresponding model predictions, and detection hypotheses (denoted by white crosses). The heatmaps in the first two columns (regression methods) encode proximity scores, in the third column they encode cell-center probabilities, where blue corresponds to 0 and red to 1. For *SSVM*, the detected MSERs are shown as mask (dark red spots). The qualitative results support the quantitative findings: *srRF* revealed less false positives than the others, *cRF* was able to detect most of the nuclei, but at the cost of lower precision (Sample B). It was not always possible for the RF methods to detect all very close objects as individual ones. This also sometimes applied to the *SSVM* method (Sample A). For the RF methods, it was also not always possible to get just a single response from larger nuclei. On the other hand, this could be observed as a strength of the *SSVM* (Sample C).

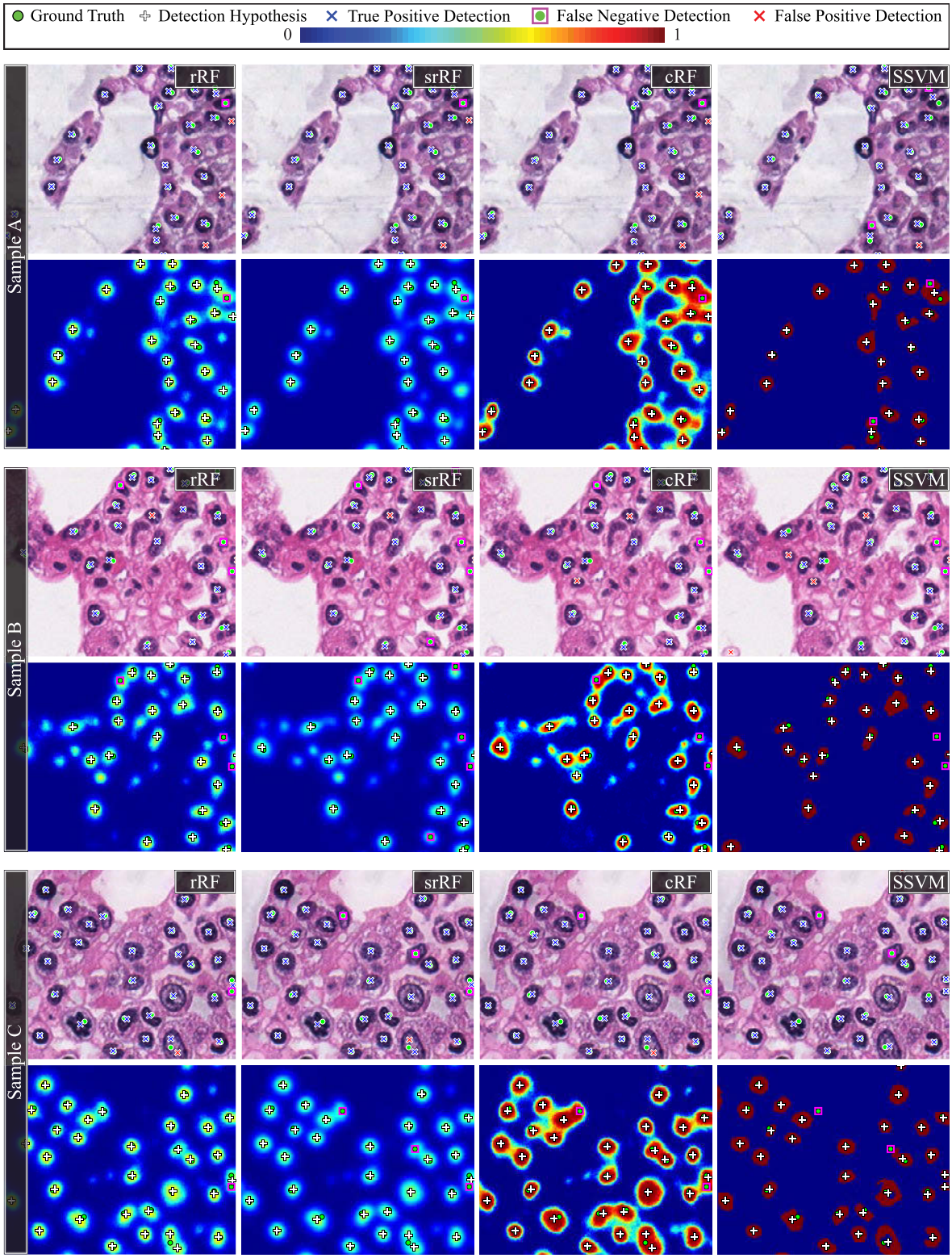


Figure 5.3.: Qualitative cell detection results on the *BM-HE* test set (three samples A, B, C) using $\xi = 11$. Columns correspond to the methods: *rRF*, *srRF*, *cRF*, *SSVM*. Heatmaps encode proximity scores, cell center probabilities, and detected MSERs, where blue corresponds to 0 and red to 1.

5.1.2. Bone Marrow (MGG)

Hyper-Parameter Selection

In stage I, the optimal input patch size was determined to be $p^*_{in} = 33$ pixels for both regression methods, resulting in the minimum NRMSE values of 0.0917 and 0.0758 for *rRF* and *srRF*, respectively. Larger output patch sizes led to lower NRMSE in the spatial-averaging regression approach: the optimal size was $p^*_{out} = 11$ pixels. Over the examined hyper-parameter combinations, only minor variations were observed in the detection-specific measures precision, recall, and F1-score, cf. Fig. A.3.

When comparing stage II to stage I, the average performance in terms of AUC could only be increased by 0.01 using deeper and larger regression forests (Fig. A.4), resulting in the best performance at $T_{md} = 24$ and $T = 64$. However, the NRMSE did improve to 0.0861 and 0.0743 for *rRF* and *srRF*, cf. Table 5.3. The proximity score map could be approximated best by the *srRF* approach, despite the NRMSE did not improve as much when compared to the *rRF* method.

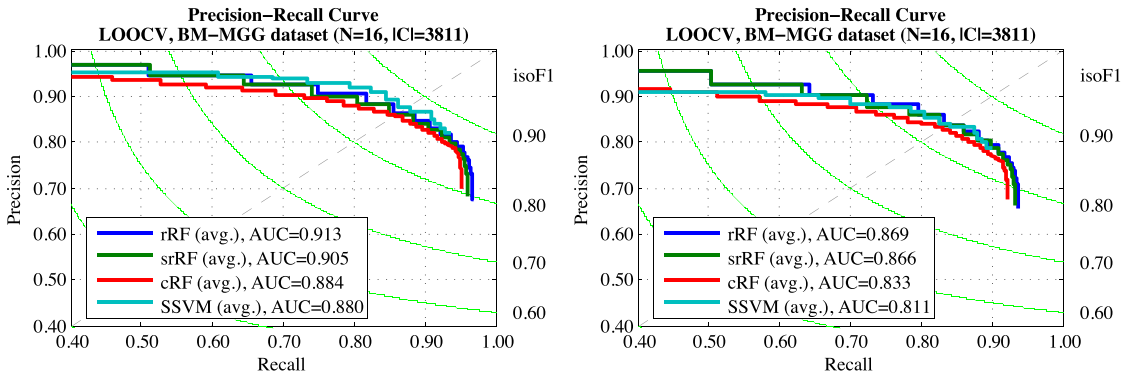
				Stage II				
Method	NRMSE	AUC	κ^*	PRC	REC	F1	$\mu_d \pm \sigma_d$	$\mu_n \pm \sigma_n$
rRF	**0.0861**	0.9103	0.5686	0.8325	0.9040	0.8668	1.81 ± 1.57	22.88 ± 15.69
srRF	**0.0743**	0.9133	0.4510	0.8267	0.9213	0.8714	1.86 ± 1.71	18.75 ± 14.51
				Stage III				
Method	$\mathcal{O}$	AUC	κ^*	PRC	REC	F1	$\mu_d \pm \sigma_d$	$\mu_n \pm \sigma_n$
rRF	**0.8951**	0.9132	0.4118	0.8480	0.9079	0.8770	1.76 ± 1.29	21.93 ± 15.80
srRF	**0.8889**	0.9055	0.3137	0.8258	0.9244	0.8724	1.75 ± 1.27	18.00 ± 13.72
cRF	–	0.8843	0.5294†	0.8203	0.9142	0.8647	1.96 ± 1.32	20.44 ± 15.69
SSVM	–	0.8804	0.7500‡	0.8661	0.9102	0.8876	1.75 ± 1.29	23.31 ± 15.40

Table 5.3.: *BM-MGG* dataset: stage II and III performance summary of the detection methods. NRMSE, precision, recall and F1-score are reported as mean over all images in LOOCV. In stage III, results are reported for $\xi = 8$.

† Confidence (probability) threshold for a cell center (*cRF*).

‡ Best prediction bias (*SSVM*).

In stage III, it was found that the post-processing parameters $\sigma_\mathcal{G} = 6$ and $r_{NMS} = 8$ maximize the objective $\mathcal{O}$ for both regression methods, cf. Table 5.3. However, the highest F1-score (0.8876) in the CV experiments could be achieved by the *SSVM* method, followed by *rRF* and *srRF*. This was mainly caused by higher precision of *SSVM*. In terms of AUC, the regression methods were superior, while *SSVM* ranked fourth, cf. Fig. 5.4.



(a) $\xi = 8$ (b) $\xi = 5$

Figure 5.4.: *BM-MGG* dataset: stage III precision-recall curves for the best hyper-parameter settings of all detection methods. Curves and AUCs represent averages over all images (LOOCV), evaluated for various ξ. We could not observe significant differences for $\xi \geq 8$, all methods worked similarly well. However, if just hypotheses within five pixels around a ground truth were considered as true positives, the regression methods performed more reliable than the others.

No noteworthy differences were observed when evaluating the spatial localization accuracy with $\xi > 8$ on this dataset. All methods were able to achieve a similarly high detection performance. The average Euclidean distance of a true positive detection to its associated ground truth was less than two pixels. Compared to $\xi = 8$, the regression methods showed the lowest drop in AUC measures (-0.04) at $\xi = 5$, demonstrating a more stable precision and more reliable spatial localization accuracy than the other methods.

Method Benchmark and Stability Analysis

Quantitative results of the benchmark and stability experiments are reported in Table 5.4 and Fig. 5.5. It could be observed that all RF methods outperformed the baseline results provided by *SSVM*. Both, the highest F1-score and AUC could be achieved by the *rRF* method, followed by the spatial-averaging regression. Despite the classification forest was able to achieve a quite convincing recall, its precision and spatial localization accuracy (μ_d) was the lowest among all tested methods. All RFs resulted in a reliable stability over ten independent runs of training and testing, but in contrast to the *BM-HE* dataset, the *srRF* was the most stable method here. Although the classification forest showed the largest variance of precision and recall, the ranking among the three forest methods did not change in any evaluation metrics.

Method	PRC	REC	F1	$\mu_d \pm \sigma_d$	$\mu_n \pm \sigma_n$
Benchmark Results					
rRF	0.8684	0.9227	**0.8947**	1.94 ± 1.20	18.25 ± 12.92
srRF	**0.8897**	0.8973	0.8935	$\mathbf{1.92 \pm 1.19}$	24.25 ± 16.98
cRF	0.8127	**0.9333**	0.8688	2.17 ± 1.33	$\mathbf{15.75 \pm 11.87}$
SSVM	0.8367	0.8902	0.8626	2.02 ± 1.36	28.75 ± 22.23
Stability Analysis (10 Independent Runs)					
rRF	0.8767 ± 0.0116	0.9162 ± 0.0118	0.8959 ± 0.0013	1.93 ± 1.22	19.78 ± 13.60
srRF	0.8664 ± 0.0091	0.9236 ± 0.0094	0.8940 ± 0.0011	1.92 ± 1.19	18.03 ± 12.02
cRF	0.8358 ± 0.0209	0.9067 ± 0.0244	0.8693 ± 0.0019	2.17 ± 1.31	22.03 ± 16.64

Table 5.4.: *BM-MGG* dataset: method comparison and stability analysis results on a held-out test set. All models were evaluated using $\xi = 8$. Stability analysis is reported for random forest based methods in the lower panel as mean±SD.

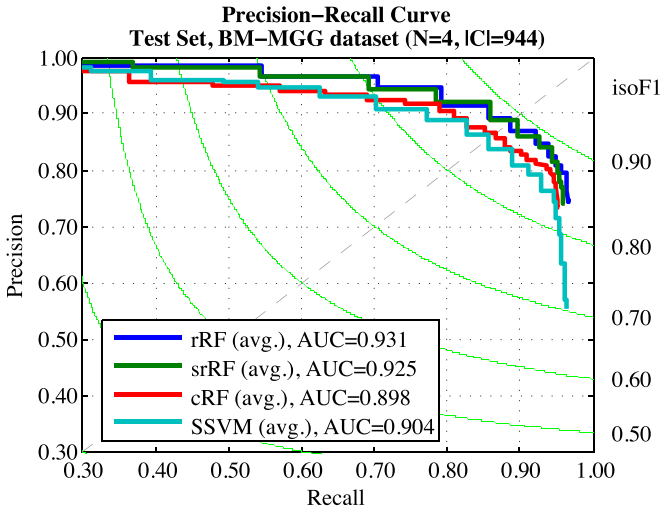


Figure 5.5.: Precision-recall curves for all methods on the *BM-MGG* test set, using $\xi = 8$.

In Fig. 5.6, qualitative results of all four detection methods (columns) for three test samples (A,B,C) are illustrated[1]. Results were obtained using $\xi = 8$. Due to merging indistinctive peaks, very close cell nuclei could sometimes not be identified as individual ones by any method (Sample A). On the other hand, where the baseline method did not detect a cell nucleus, the RF methods actually detected two peaks on a single object (top left corner of Sample B). The classification forest tended to detect more false positives, while the *SSVM* missed a larger number of actual cells (Sample C).

[1] The reader is referred to Section 5.1.1 at page 61 for a more detailed explanation of the figure legend.

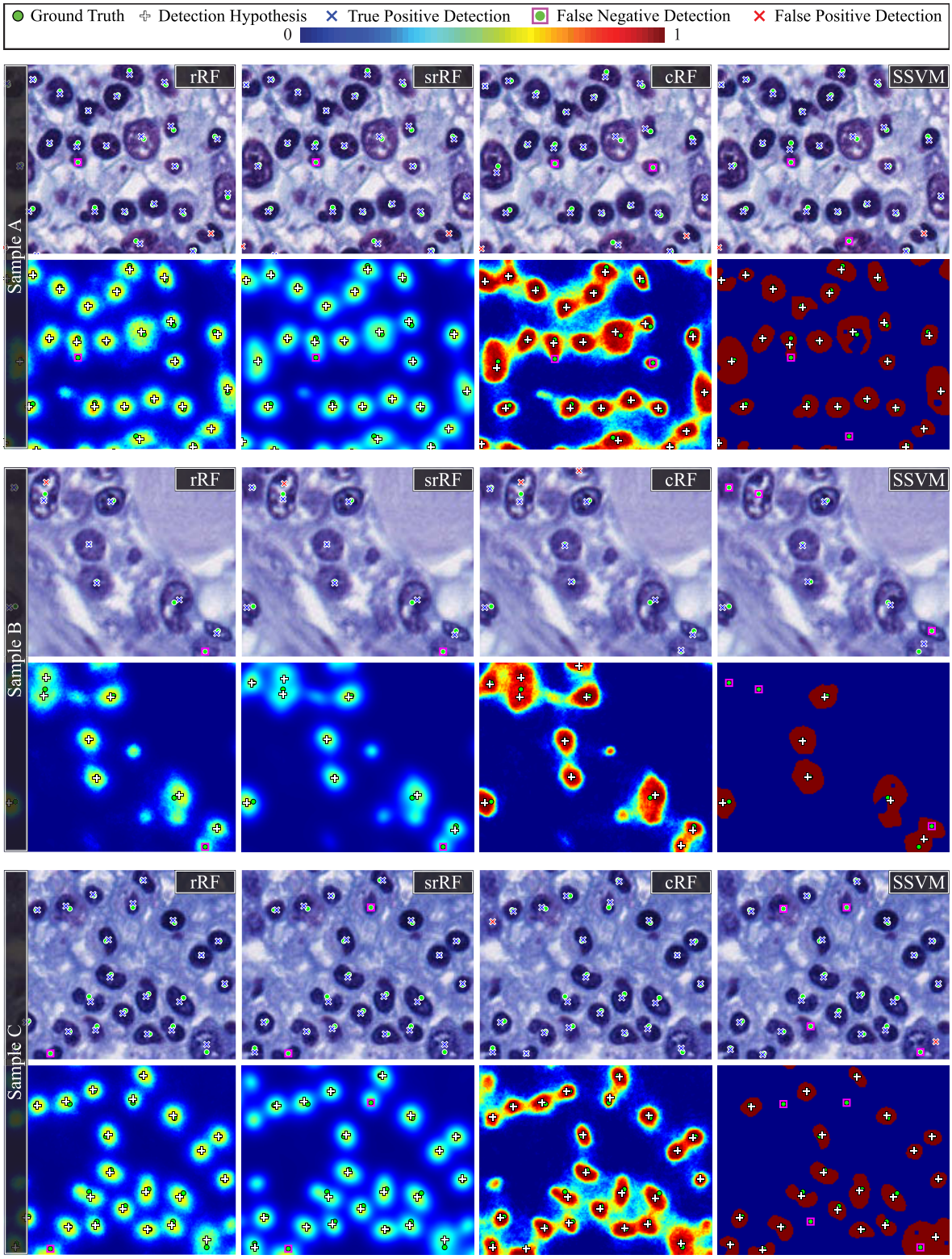


Figure 5.6.: Qualitative cell detection results on the *BM-MGG* test set (three samples A, B, C) using $\xi = 8$. Columns correspond to the methods: *rRF*, *srRF*, *cRF*, *SSVM*. Heatmaps encode proximity scores, cell center probabilities, and detected MSERs, where blue corresponds to 0 and red to 1.

5.1.3. Bone Marrow (H&E) Megakaryocytes

For megakaryocyte localization, prior knowledge on the appearance of this specific hematopoietic cell lineage was leveraged. Megakaryocytes usually have larger, multi-lobed nuclei, and more extensive cytoplasm surrounding the nucleus than other bone marrow cells, cf. Fig. 4.2. Therefore, they can usually be identified rather easily among adjacent granulo- and erythropoietic cells at optical magnifications $< 40\times$.

The first evaluation of the detection methods on this dataset considered megakaryocytes at $20\times$ magnification. An optimal input patch size of $p^*_{in} = 75$ pixels and model complexity of $T_{md} = 24$ and $T = 64$ were determined in stage I and II of the hyper-parameter search, cf. Fig. A.5 and Fig. A.6. Spatial-averaging regression worked best when using $p^*_{out} = 11$ pixels. The optimal post-processing parameters were found to be $\sigma_{\mathcal{G}} = 8$ and $r_{NMS} = 18$. The precision-recall curves illustrated in Fig. 5.8 (a) show that at a distance threshold of $\xi = 27$ all RF-based detection methods provided very high detection accuracy. Nevertheless, for $\xi = 5$, AUC measures dropped considerably for all methods, cf. Fig. 5.8 (b). The quantitative results of the stage III evaluations (Table 5.5) may provide an explanation: the mean distance of a ground truth point to its corresponding true positive detection (μ_d) for the RF methods was almost equal to the selected evaluation threshold, i.e. $4.9 \approx 5$. Additionally, considering the large σ_d, we could not capture a lot of actual true positive detections when setting such a strict evaluation tolerance, hence the sudden drop in detection performance measures and lower AUC measures. Once we selected some $\xi > 5$, the measures became similar to the ones reported in Fig. 5.8 (a).

The second set of evaluations was run using this dataset at $10\times$ magnification. There, we found that the best parameters were $p^*_{in} = 38$ for *rRF*, whereas *srRF* required $p^*_{in} = 25$ and $p^*_{out} = 11$ to achieve the best performance, cf. Fig. A.7. All models minimized the NRMSE using an RF complexity of $T_{md} = 24$ and $T = 64$, cf. Fig. A.8. The optimal post-processing parameters for *rRF* were found to be $\sigma_{\mathcal{G}} = 4$ and $r_{NMS} = 6$, and $\sigma_{\mathcal{G}} = 2$ and $r_{NMS} = 13$ for *srRF*, respectively. Fig. 5.8 (c,d) illustrates the precision-recall curves. Compared to $20\times$ magnification, the performance measures were similarly high for all three RF methods. The spatial localization accuracy in terms of μ_d at $10\times$ magnification was seemingly higher than at $20\times$, but a quantitative examination suggested that they in fact depend on the object scale. For instance, if we relate μ_d of the *rRF* method at both magnifications, we can observe that $4.57 \approx (2 \cdot 2.33)$. Similar observations could be made for the other RF methods, indicating that their cell center

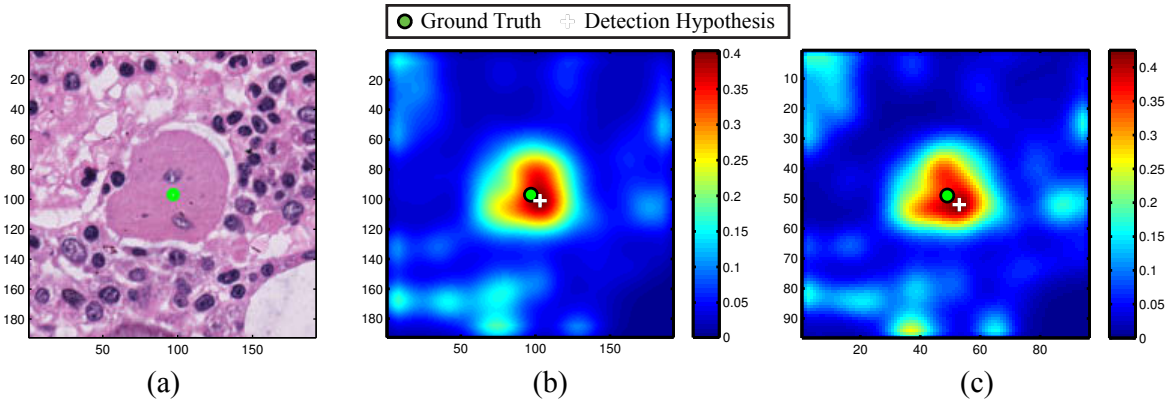
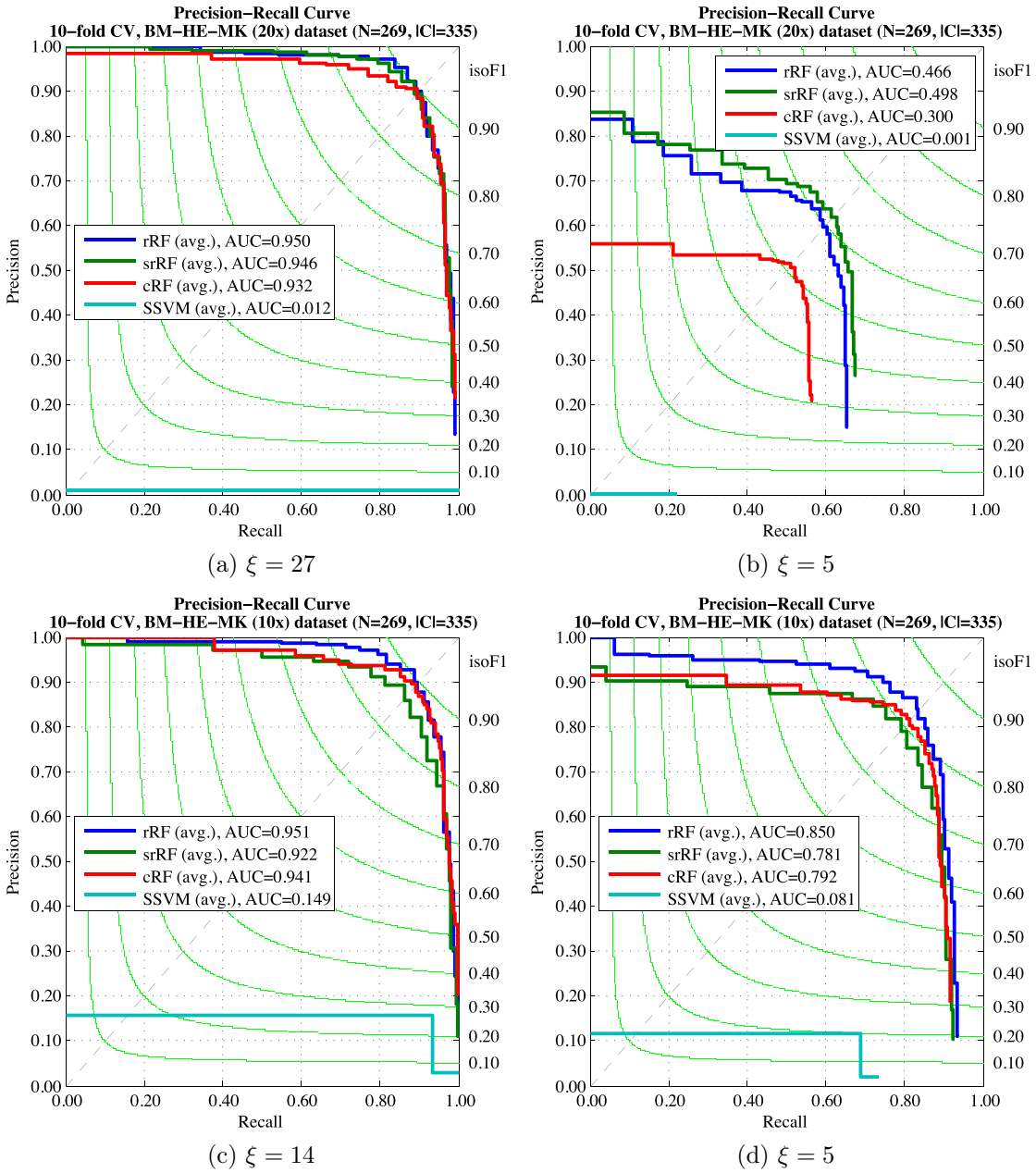


Figure 5.7.: Qualitative comparison at different magnifications. Source image with ground truth (a), proximity score maps at 20× (b), and 10× (c). The location of the predicted center is stable with respect to the ground truth center (green dot). The illustrated proximity score maps were predicted by the *srRF* method, the detections produced with the best post-processing settings for this method.

prediction at both magnifications was rather stable with respect to the relative location of the true center. A qualitative illustration of this effect is depicted in Fig. 5.7 using the *srRF* method as an example.

Interestingly, the *SSVM* method failed detecting the megakaryocytes at both object scales, although it worked well on the myeloid and erythroid cell lineages. This was mainly expressed in very low precision, while recall could reach an acceptable range, cf. Fig. 5.8. Considering the qualitative results in Fig. 5.9, almost all candidates that looked similar to a cell nucleus were predicted as cell centers. On the other hand, multilobed nuclei were frequently not recognized as coherent candidate regions (Sample A), but rather the small cytoplasmic regions between the lobes. Furthermore, many locations in obvious background, or fat cells, were detected and hypothesized as (false positive) cell nuclei centers. It seemed as if the detector learned to detect locations on all nuclei and in the intercellular space rather than on the actual megakaryocytes. Hence, the perfect recall of this method at 20× magnification is just a side effect, cf. Table 5.5.

At both magnifications, the *rRF* method worked best, exhibiting the highest detection performance measures. Compared to results obtained at 20×, the *srRF* method performed not as good as the other RF methods at 10× magnification. In this comparison, most values of the hyper-parameters were scaled by the same factor as the input images, i.e. by 0.5. It is therefore little surprising that the detection-specific performance values look quite similar. However, this also suggests that at each magnification of the



(a) $\xi = 27$ (b) $\xi = 5$

(c) $\xi = 14$ (d) $\xi = 5$

Figure 5.8.: *BM-HE-MK* dataset: stage III precision-recall curves for the best hyper-parameter settings of all detection methods. Curves and AUCs represent averages over the LOOCV runs, evaluated for various ξ. (a,b) 20× magnification: at higher tolerances (a), all RF methods show very high performance, while at lower tolerances (b), their AUC measures drop significantly. (c,d) 10× magnification: in (c) similar performance compared to (a) can be observed. In all cases, also for evaluations using $\xi > 27$ and $\xi > 14$ that are not reported here, *SSVM* fails in this task.

Method	NRMSE	AUC	κ^*	PRC	REC	F1	$\mu_d \pm \sigma_d$	$\mu_n \pm \sigma_n$
			Stage II - 20×					
rRF	**0.0609**	0.9448	0.4314	0.9039	0.8985	0.9012	5.32 ± 4.98	0.12 ± 0.43
srRF	**0.0584**	0.9420	0.4118	0.9195	0.8866	0.9027	4.83 ± 4.16	0.14 ± 0.41
			Stage II - 10×					
rRF	**0.0752**	0.9558	0.4706	0.9446	0.8657	0.9034	2.46 ± 2.14	0.17 ± 0.47
srRF	**0.0603**	0.9310	0.3333	0.8849	0.8716	0.8782	2.66 ± 2.14	0.16 ± 0.48
Method	$\mathcal{O}$	AUC	κ^*	PRC	REC	F1	$\mu_d \pm \sigma_d$	$\mu_n \pm \sigma_n$
			Stage III - 20×					
rRF	**0.9299**	0.9501	0.3726	0.9511	0.8716	0.9097	4.57 ± 3.50	0.16 ± 0.47
srRF	**0.9254**	0.9465	0.3333	0.9198	0.8896	0.9044	4.47 ± 3.23	0.14 ± 0.41
cRF	–	0.9324	0.7451†	0.9036	0.8955	0.8996	5.57 ± 3.47	0.13 ± 0.43
SSVM	–	0.0117	1.2500‡	0.0117	1.0000	0.0232	7.67 ± 3.84	0.00 ± 0.00
			Stage III - 10×					
rRF	**0.9291**	0.9511	0.3922	0.9255	0.8896	0.9072	2.33 ± 1.58	0.14 ± 0.41
srRF	**0.9028**	0.9251	0.3333	0.8926	0.8687	0.8805	2.59 ± 1.77	0.16 ± 0.48
cRF	–	0.9413	0.7255†	0.8952	0.8925	0.8938	2.72 ± 1.85	0.13 ± 0.44
SSVM	–	0.1487	1.0000‡	0.1573	0.9333	0.2692	4.11 ± 2.35	0.01 ± 0.11

Table 5.5.: *BM-HE-MK* dataset: stage II and III performance summary of the detection methods on 20× and 10× magnification. In stage III, results are reported for $\xi = 27$, and $\xi = 14$, respectively. NRMSE, precision, recall and F1-score are reported as mean over all images in 10-fold CV. The perfect recall of the *SSVM* at 20× was just a side effect from a failing detector.

† Confidence (probability) threshold for a cell center (*cRF*).

‡ Best prediction bias (*SSVM*).

dataset the RFs learn a similar predictive model, because they practically have the same contextual information in the input patches.

In Fig. 5.9, qualitative results of all four detection methods in columns for three samples (A,B,C) at 20× magnification are illustrated[2]. The results were obtained by 10-fold CV on the entire dataset using $\xi = 27$. In Samples A and B, all RF methods perfectly detected the megakaryocytes without revealing any false positives, while in Sample C the *cRF* missed one cell. In a minority of cases, multilobed nuclei with minimal surrounding cytoplasm such as shown in Sample A could not be detected by the RF methods. However, *SSVM* always detected a huge number of false positives, where some of them were assigned as true positives as a consequence of the defined performance evaluations.

[2] The reader is referred to Section 5.1.1 at page 61 for a more detailed explanation of the figure legend.

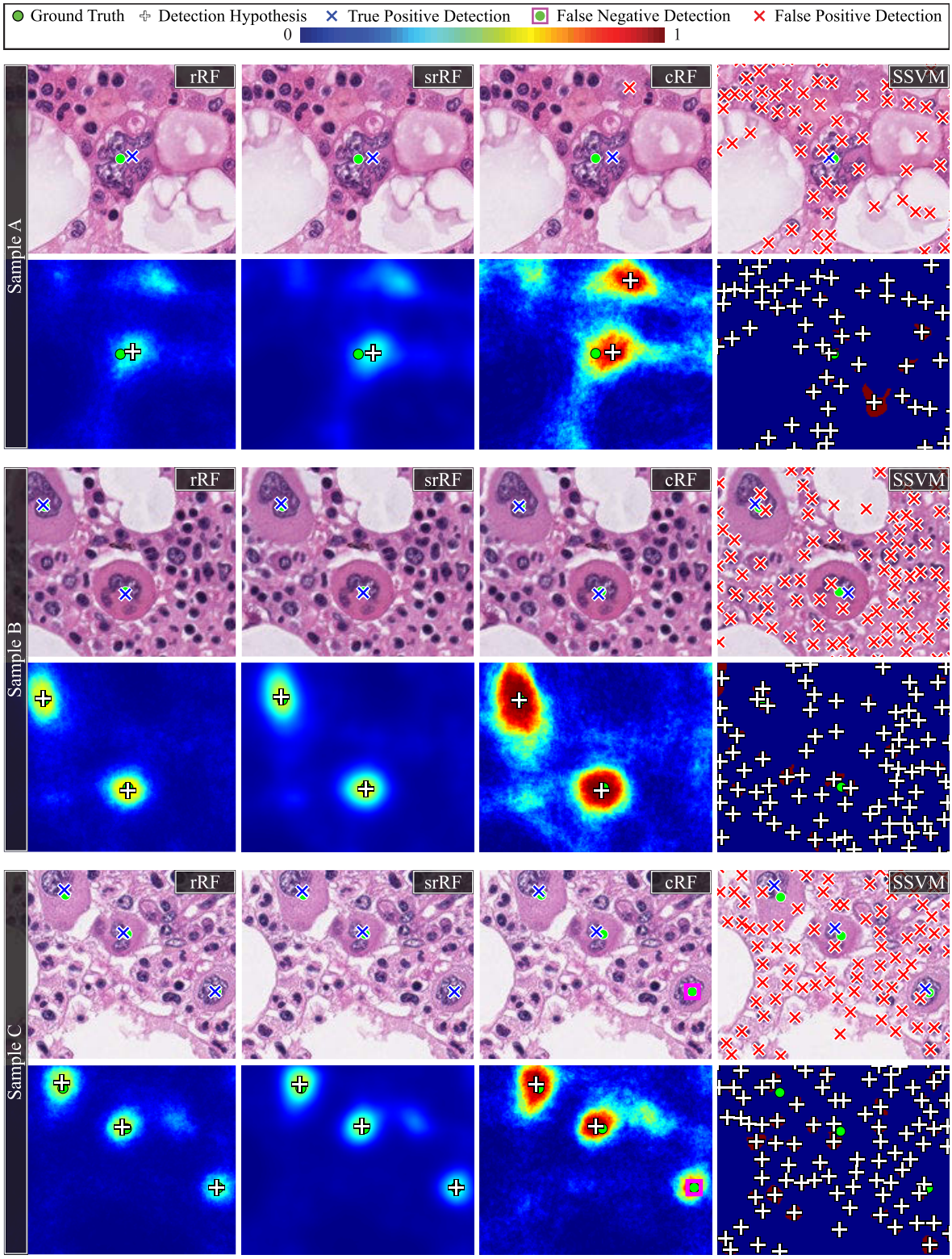


Figure 5.9.: Qualitative cell detection results on the *BM-HE-MK* dataset (three samples A, B, C, from 10-fold CV) using $\xi = 27$ at 20× magnification. Columns correspond to the methods: *rRF*, *srRF*, *cRF*, *SSVM*. Heatmaps encode proximity scores, cell center probabilities, and detected MSERs, where blue corresponds to 0 and red to 1.

5.1.4. Breast Cancer (H&E)

Hyper-Parameter Selection

On this dataset, the optimal input patch size was found to be $p^*_{in} = 11$ for both regression methods. Furthermore, an output patch size of $p^*_{out} = 7$ was optimal for *srRF*. Both models achieved the highest performance when using a random forest with $T_{md} = 24$ and $T = 64$, resulting in a NRMSE of 0.0700 and 0.0634 for *rRF* and *srRF*, cf. Table 5.6. F1-scores are quite stable across all examined parameter configurations, despite precision and recall vary much more compared to the bone marrow datasets. Detailed stage I and II results are illustrated in Figs. A.9 and A.10. The optimal post-processing settings that optimize the stage III objective $\mathcal{O}$ were found to be $\sigma_{\mathcal{G}} = 2$ and $r_{NMS} = 3$.

				Stage II				
Method	NRMSE	AUC	κ^*	PRC	REC	F1	$\mu_d \pm \sigma_d$	$\mu_n \pm \sigma_n$
rRF	**0.0700**	0.9185	0.1961	0.8437	0.9173	0.8790	0.80 ± 0.67	2.75 ± 2.65
srRF	**0.0634**	0.8904	0.1373	0.8264	0.9308	0.8755	0.78 ± 0.66	2.30 ± 2.47
				Stage III				
Method	$\mathcal{O}$	AUC	κ^*	PRC	REC	F1	$\mu_d \pm \sigma_d$	$\mu_n \pm \sigma_n$
rRF	**0.8920**	0.9094	0.1765	0.8189	0.9384	0.8746	0.73 ± 0.64	2.05 ± 2.01
srRF	**0.8812**	0.8926	0.1373	0.8211	0.9248	0.8699	0.77 ± 0.63	2.50 ± 2.59
cRF	–	0.8860	0.4902†	0.8431	0.9053	0.8731	0.78 ± 0.62	3.15 ± 2.81
SSVM	–	0.7895	1.5000‡	0.8142	0.8070	0.8106	0.88 ± 0.67	7.40 ± 5.04

Table 5.6.: *ICPR-BC* dataset: stage II and III performance summary of the detection methods. NRMSE, precision, recall and F1-score are reported as mean over all images in 10-fold CV. In stage III, results are reported for $\xi = 3$.

† Confidence (probability) threshold for a cell center (*cRF*).

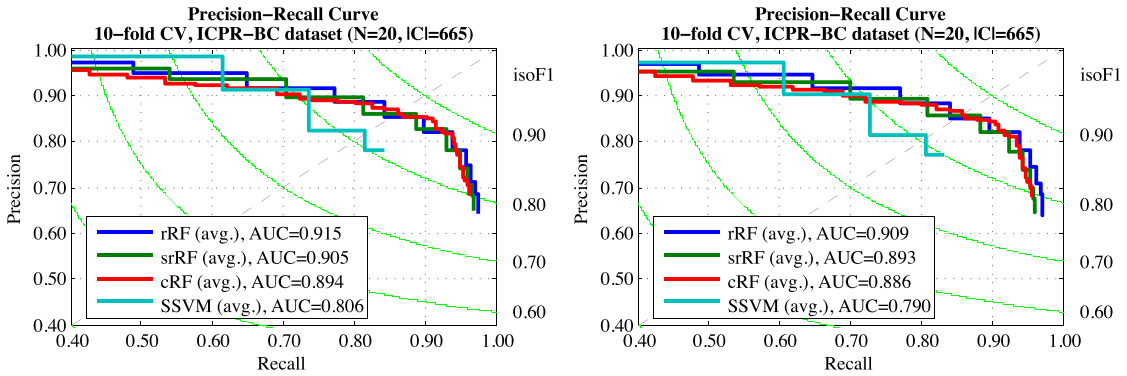
‡ Best prediction bias (*SSVM*).

In the CV experiments, it was not possible to achieve an *SSVM* localization performance comparable to the one reported in the original work of Arteta *et al.* [116], where the detector was evaluated on the dedicated test set provided during the challenge[3]. When using $\xi = 3$, their *SSVM* method could only be optimized to achieve an average F1-score of 0.8106 in the 10-fold CV on the entire dataset. This was mainly caused by a lower recall (0.8142) compared to all other methods (> 0.90). Despite all methods were able to reach sub-pixel spatial localization accuracy (on average), the RF methods

[3] It was also not mentioned in the work of Arteta *et al.* [116], which distance ξ was considered to the report the performance measures.

showed much better detection performance than the *SSVM* baseline, cf. Table 5.6. We could not observe that the regression methods always outperformed the classification approach in the CV experiments on this dataset, because the *cRF* showed the highest precision, and competitive recall.

The AUC measure for all RF methods was quite similar. Also, employing a tighter ξ did not influence these measures much, cf. Fig. 5.10. It has to be noted that the *SSVM* method exhibited higher precision at lower recalls than the other methods. However, a remarkable decrease in precision was found around a recall of 0.73 for both evaluated ξ. This behavior was also observed for $\xi \geq 7$.



(a) $\xi = 7$ (b) $\xi = 3$

Figure 5.10.: *ICPR-BC* dataset: stage III precision-recall curves for the best hyper-parameter settings of all detection methods. Curves and AUCs represent averages over the 10-fold CV runs, evaluated for various ξ. Since all methods exhibit sub-pixel localization accuracy on this dataset, we could not observe significant differences at lower ξ.

In Fig. 5.11, qualitative detection results from the 10-fold cross validation are illustrated for three samples (A,B,C)[4]. The results were obtained on the entire dataset using a strict setting of $\xi = 3$. Several annotations were placed quite close to the image border (e.g. Sample A). This could explain the lower recall of the *SSVM* method, since these cells would probably have been detected when some kind of border extension was used. The RF methods, however, were able to correctly detect these cells most of the time. In Sample B, the number of false negatives was quite low for the RF methods, but *SSVM* missed several cells. *SSVM* also detected many more false positives than the RF methods (Sample C), since stable regions were detected by the MSER detector and the classifier obviously could not learn to properly reject these candidates.

[4] The reader is referred to Section 5.1.1 at page 61 for a more detailed explanation of the figure legend.

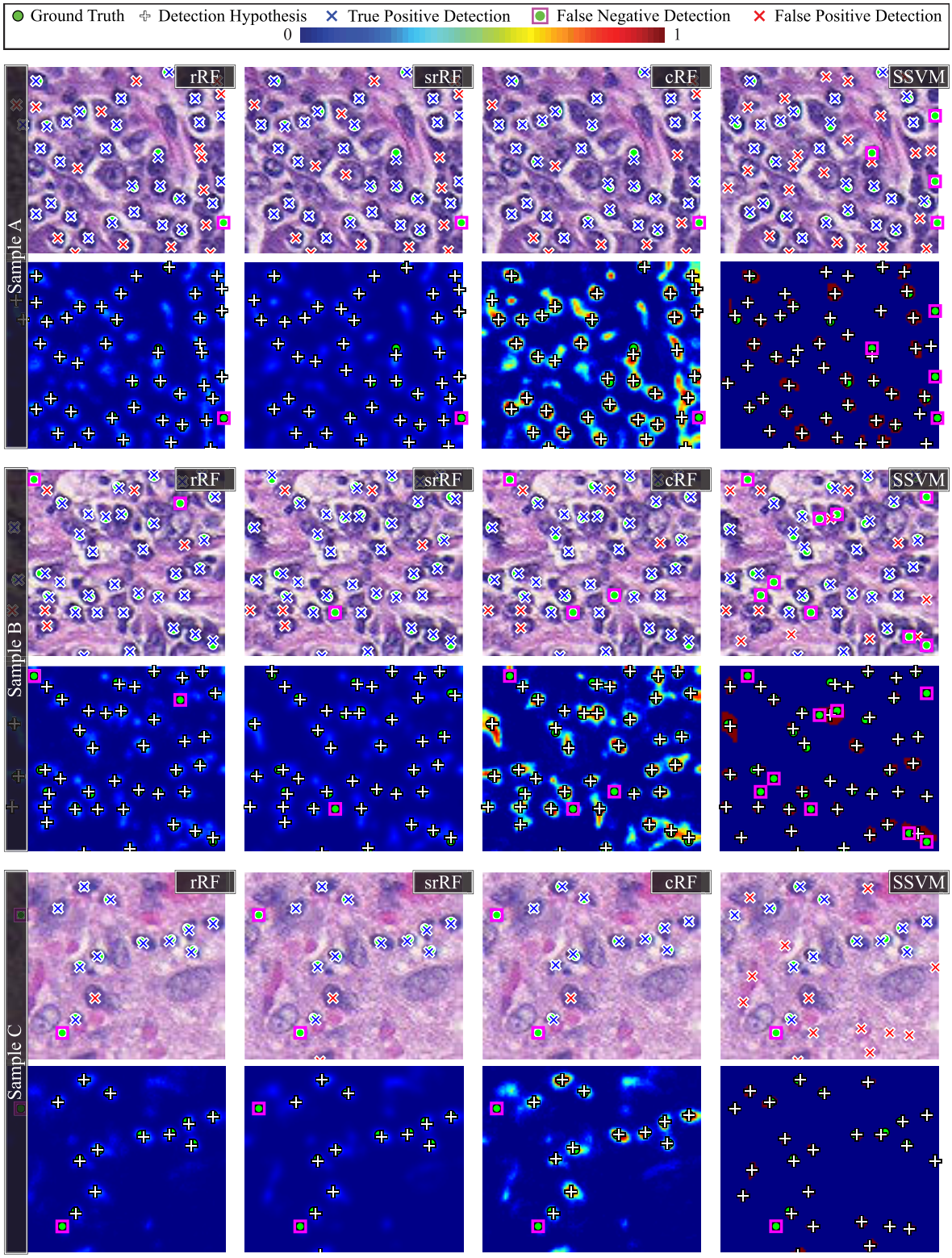


Figure 5.11.: Qualitative cell detection results on the *ICPR-BC* dataset (three samples from A, B, C, from 10-fold CV) using $\xi = 3$. Columns correspond to the methods: *rRF*, *srRF*, *cRF*, *SSVM*. Heatmaps encode proximity scores, cell center probabilities, and detected MSERs, where blue corresponds to 0 and red to 1.

Benchmark Results

The *ICPR-BC* dataset was released as part of the Pattern Recognition in Histopathological Images Contest [130]. In earlier work, regression-based cell localization [8], in particular *rRF*, was shown to outperform standard classification-based, and the method of Arteta *et al.* [116], which was considered state-of-the-art until then. The RF hyperparameters were used from an estimation on a more complex, bone marrow dataset [6], and subsequently the models were trained on ten training images, and evaluated on ten test images.

Method	PRC	REC	F1	$\mu_d \pm \sigma_d$	$\mu_n \pm \sigma_n$
Kainz *et al.*, *rRF* [8]	**0.9133**	**0.9170**	**0.9150**	$\mathbf{0.80 \pm 0.68}$	3.26 ± 2.32
Kainz *et al.*, *cRF* [8]	0.9066	0.8972	0.9018	0.86 ± 0.68	4.04 ± 2.26
Arteta *et al.*, *SSVM* [116]	0.8699	0.9003	0.8848	1.68 ± 2.55	$\mathbf{2.90 \pm 2.13}$
Kuse *et al.*, LIPSyM [104]	0.7021	0.7008	0.7014	3.14 ± 0.93	4.30 ± 3.08
Kuse *et al.* [103]	0.6523	0.6999	0.6729	3.04 ± 3.40	14.01 ± 4.40
Bernardis & Yu [105]	–	–	–	2.84 ± 2.89	8.20 ± 4.75
Cheng *et al.* [183]	–	–	–	8.10 ± 6.98	6.98 ± 12.50
Graf *et al.* [182]	–	–	–	7.60 ± 6.30	24.50 ± 16.20
Panagiotakis *et al.* [184]	–	–	–	2.87 ± 3.80	14.23 ± 6.30

Table 5.7.: *ICPR-BC* dataset: comparison of different cell detection methods on the held-out benchmark test set used in the ICPR 2010 contest [130]. The results of Kainz *et al.* [8] were obtained using $\xi = 4$, whereas all others used less strict values. Superior results are printed in bold.

In Table 5.7, the quantitative results of all participating teams are listed along the results reported in [8]. The *rRF* and *cRF* models were evaluated using $\xi = 4$ [8]. All other methods used less strict values of $\xi > 4$. It was not possible to exactly reproduce the performance originally reported by Arteta *et al.* [116] using the provided implementation of their *SSVM* method, hence the values from their publication are reported here. Both *rRF* and *cRF* method outperformed all previous methods on all standard metrics used in this benchmark dataset. Further, the results of these two methods are reported as the mean performance measures across ten independent iterations of training and testing, which indicates convincing stability and underlines the superiority in spatial localization accuracy. Considering the given challenging task of detecting leukocytes in breast cancer histology images, the proposed proximity score regression method was able to more reliably discriminate cell types than previous methods.

5.1.5. Multi-Tissue (H&E)

Hyper-Parameter Selection

The hyper-parameter search in stage I revealed an optimal input patch size of $p^*_{in} = 29$ pixels for both regression methods. Additionally, the optimal output patch size was found to be $p^*_{out} = 11$ pixels for *srRF*. In stage II, the lowest NRMSE of 0.1103 and 0.0870 could be achieved for *rRF* and *srRF*, using a random forest with $T_{md} = 24$ and $T = 64$. The spatial-averaging regression was able to predict the proximity scores much more precisely than the pixel-wise regression, but this had in fact very little influence on the detection performance in terms of F1-score, cf. Table 5.8. The reader is referred to Section A.5 for more detailed illustrations of performance results.

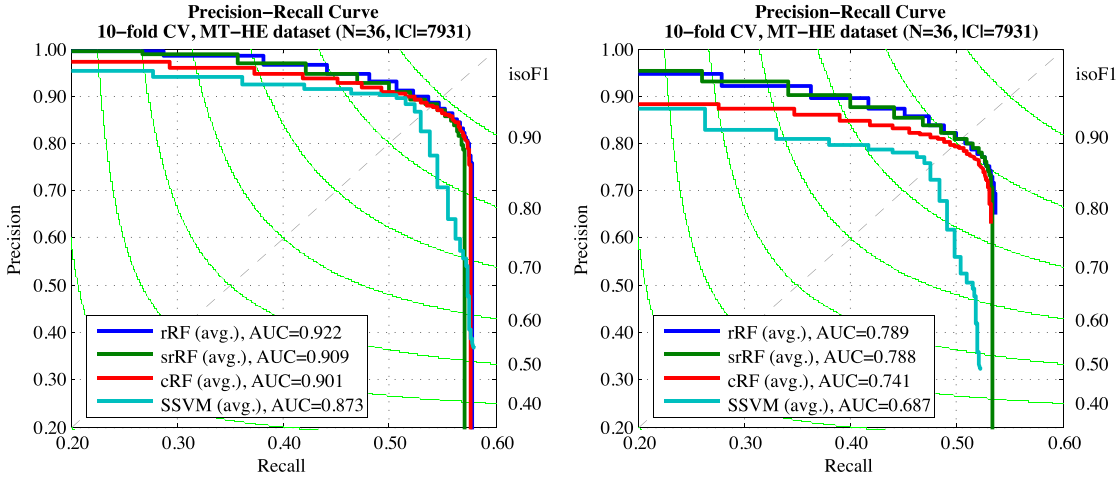
				Stage II				
Method	NRMSE	AUC	κ^*	PRC	REC	F1	$\mu_d \pm \sigma_d$	$\mu_n \pm \sigma_n$
rRF	**0.1103**	0.8955	0.5098	0.8479	0.9143	0.8798	2.65 ± 1.93	18.89 ± 23.91
srRF	**0.0870**	0.8981	0.3529	0.8526	0.9150	0.8827	2.61 ± 1.87	18.72 ± 20.92
				Stage III				
Method	$\mathcal{O}$	AUC	κ^*	PRC	REC	F1	$\mu_d \pm \sigma_d$	$\mu_n \pm \sigma_n$
rRF	**0.9057**	0.9215	0.4510	0.8748	0.9056	0.8899	2.59 ± 1.85	20.81 ± 24.61
srRF	**0.8979**	0.9085	0.3137	0.8714	0.9037	0.8872	2.57 ± 1.79	21.22 ± 23.73
cRF	–	0.9006	0.6471[†]	0.8569	0.9179	0.8864	2.70 ± 1.94	18.08 ± 19.47
SSVM	–	0.8729	0.5000[‡]	0.8830	0.8482	0.8653	2.67 ± 1.87	24.14 ± 20.95

Table 5.8.: *MT-HE* dataset: stage II and III performance summary of the detection methods. NRMSE, precision, recall and F1-score are reported as mean over all images in 10-fold CV. In stage III, results are reported for $\xi = 14$.

[†] Confidence (probability) threshold for a cell center (*cRF*).

[‡] Best prediction bias (*SSVM*).

The two post-processing parameters $\sigma_{\mathcal{G}} = 4$ and $r_{NMS} = 7$ resulted in the optimal stage III objectives for *rRF* ($\mathcal{O} = 0.9057$) and *srRF* ($\mathcal{O} = 0.8979$), cf. Table 5.8. All results were obtained in 10-fold CV on the entire dataset using $\xi = 14$. The precision-recall curves for all methods are illustrated in Fig. 5.12. On this complex dataset, which contains multiple tissue types, the *rRF* method achieved the remarkable AUC of 0.9215, followed by *srRF* and *cRF*. Both regression methods had higher precision than the classification and the *SSVM* approach. The latter could not achieve the same spatial localization accuracy as the regression methods, but still slightly better than *cRF*. The maximum achievable recall was similar for all methods, though.



(a) $\xi = 14$ (b) $\xi = 5$

Figure 5.12.: *MT-HE* dataset: stage III precision-recall curves for the best hyper-parameter settings of all detection methods. Curves and AUCs represent averages over the 10-fold CV runs, evaluated for various ξ. Despite all methods resulted in satisfactory performance at $\xi = 14$, the *SSVM* method showed a considerably lower AUC at $\xi = 5$.

In Fig. 5.13, qualitative detection results from the 10-fold cross validation are illustrated for three samples (A,B,C)[5], obtained using $\xi = 14$. Due to multiple types of tissue, this dataset was characterized by a large variability of cells in terms of nuclei staining, shape, and size. Sample A shows some false positive detections close to the right image border. Considering similar cells that were detected as true positives in this dataset, it is unclear why these cells have not been annotated in the ground truth. The reported quantitative detection performance, especially precision, could therefore have been even higher, cf. Table 5.8. Further, the classification approach detected two peaks on a single object more frequently than the proposed regression. From Sample B it could be observed that tiny, and elongated cells in blurry image regions could not be identified properly by the *SSVM*. Spatial-averaging regression and classification are able to detect most of these cells. The number of false positive detections was very low in this sample. The hypothesized cell centers were very similar for all methods in Sample C. Nevertheless, the *cRF* again produced multiple peaks on single objects more frequently than the regression methods, accompanied by more false positives. None of the four methods was able to detect the two very close individual objects at the lower image border as such. The RF methods further seemed to have issues detecting very tiny cells, but this was likely caused by some questionable ground truth annotations.

[5] The reader is referred to Section 5.1.1 at page 61 for a more detailed explanation of the figure legend.

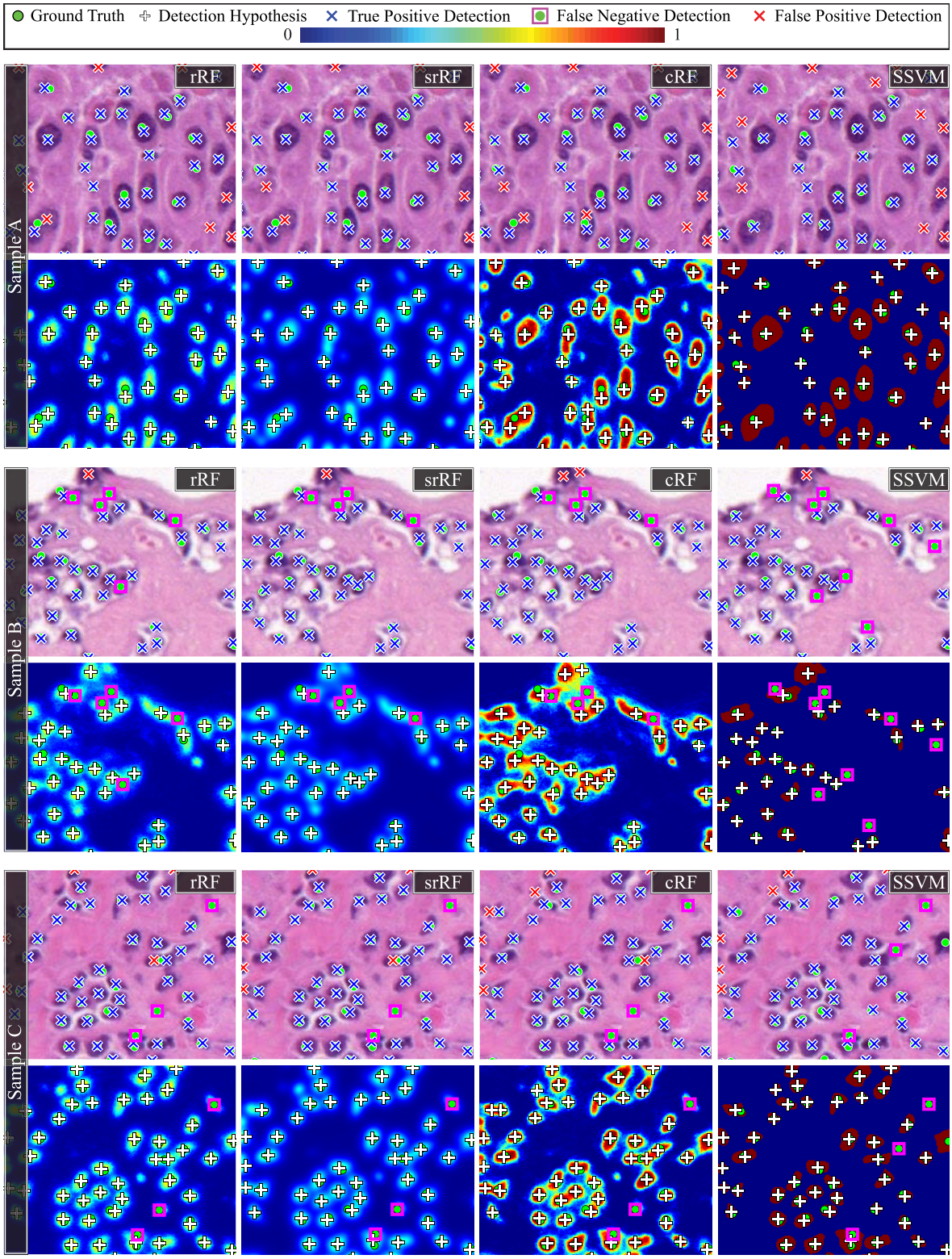


Figure 5.13.: Qualitative cell detection results on the *MT-HE* dataset (three samples A, B, C, from 10-fold CV) using $\xi = 14$. Columns correspond to the methods: *rRF*, *srRF*, *cRF*, *SSVM*. Heatmaps encode proximity scores, cell center probabilities, and detected MSERs, where blue corresponds to 0 and red to 1.

5.2. Transferability of Learned Detection Models

In the previous sections, the performance of the cell detectors was evaluated on datasets, where the tissue had been stained with a dedicated dye only. As it has been shown, the cell detectors may be re-trained from scratch on new datasets, but in bone marrow histopathology, several histochemical staining protocols exist that allow a similar visualization of the cell morphology, e.g. H&E and MGG. Re-using learned predictors for similar histopathological images could therefore remove time-consuming training procedures and provide a more general method for automated tissue analysis.

In this section, we compared the performance of the cell detectors when they were learned on the training set of the *BM-HE* dataset and evaluated on the test set of the *BM-MGG* dataset, and vice versa. To enable learning from multiple stainings with the same hyper-parameters, the datasets were pre-processed: we just used features that could be computed from grey-scale (RGB mean) versions of all images. Color-related features were thus removed, leaving a set of 48 visual image features for the RF methods, cf. Section 3.2.2. The features used for *SSVM* were also reduced to be computed from grey-scale images. All four methods were evaluated using the best hyper-parameters estimated for each of the two bone marrow datasets. The training and test subset were defined previously in Section 4.4.3.

		Test Set					
		H&E			**MGG**		
Method	**Training Set**	PRC	REC	F1	PRC	REC	F1
rRF	H&E	0.8377	0.9298	0.8813	0.8628	0.8994	0.8813
	MGG	0.8462	0.9313	0.8867	0.8657	0.9078	0.8863
srRF	H&E	0.8480	0.9204	0.8827	0.8378	0.9248	0.8792
	MGG	0.8531	0.9248	0.8875	0.8560	0.9195	0.8866
cRF	H&E	0.8529	0.9059	0.8786	0.8332	0.9100	0.8699
	MGG	0.8416	0.9190	0.8786	0.8129	0.9248	0.8652
SSVM	H&E	0.8263	0.9048	0.8638	0.8263	0.9048	0.8638
	MGG	0.7737	0.8922	0.8288	0.8449	0.8841	0.8641

Table 5.9.: Transferability of learned detection models for two different histochemical stainings in bone marrow histopathology. In terms of F1-score, transferring from H&E to MGG and vice-versa is very stable for all RF methods. The *SSVM* method has some difficulties to be transferred from MGG to H&E, where the precision is considerably lower compared to the other experiments.

The performance was compared quantitatively in terms of precision, recall, and F1-

score in Table 5.9. Transferring the learned models worked well both ways for all RF methods. In terms of F1-score, the results on the two test sets were practically identical for the two regression-based methods, which also outperformed the classification-based and *SSVM* method. However, it could be observed that transferring the *SSVM* classifier from MGG to H&E staining did not work as well as vice versa, or compared to the other methods. This was expressed by a higher number of false positive detections (lower precision: 0.7737), when the learned classifier was transferred from MGG to H&E staining. One of the reasons was that the *SSVM* approach relied on intensity histogram features to learn cells from the candidate regions revealed by the MSER detector. The intensity distribution is different for H&E and MGG stained nuclei, also when only grey-value images are considered, hence limiting the transferability. In the *BM-MGG* dataset, we found more homogeneously stained nuclei than in the *BM-HE* dataset, cf. samples in Fig. 4.1, which likely led to problems in the correct classification of the candidates.

5.3. Observations from Empirical Evaluations

5.3.1. Examination of Optimal Split Function Components

The RFs were trained on a set of $|\Phi| = 54$ pre-computed visual features, cf. Fig. 3.6. All feature channels were available for each patch at each randomized internal node optimization in order to find a binary split of the dataset using an optimal split function[6] ϕ^*. Fig. 5.14 visualizes the components of the optimal split functions at different levels of the random forests with $T = 64$ trees, each with a maximum depth of $T_{md} = 24$. The first two columns correspond to the proposed regression methods *rRF* and *srRF*, while the last column shows the results for the binary classification forest (*cRF*). The large panels illustrate the selection frequency of a feature channel Φ_i, where pixels in red to black colors encode lower and yellow to white colors encode higher frequencies[7]. The horizontal axis in these panels corresponds to the feature channel identifiers defined in Table 3.2. Pixels in the smaller panels depict the frequencies of the evaluation operations in the optimal selection functions θ^*, where lower to higher frequencies are

[6] The optimal *split* function ϕ^* was defined to comprise two parts. First, a *selection* function θ chooses the feature channel(s) Φ and an evaluation operation, e.g. a single pixel value, in a particular channel. Secondly, the response of θ is compared to a random threshold τ_ϕ, cf. Section 3.2.2 for more details.

[7] We only considered the first selected feature channel for this visualization, ignoring whether a selection function θ required two channels.

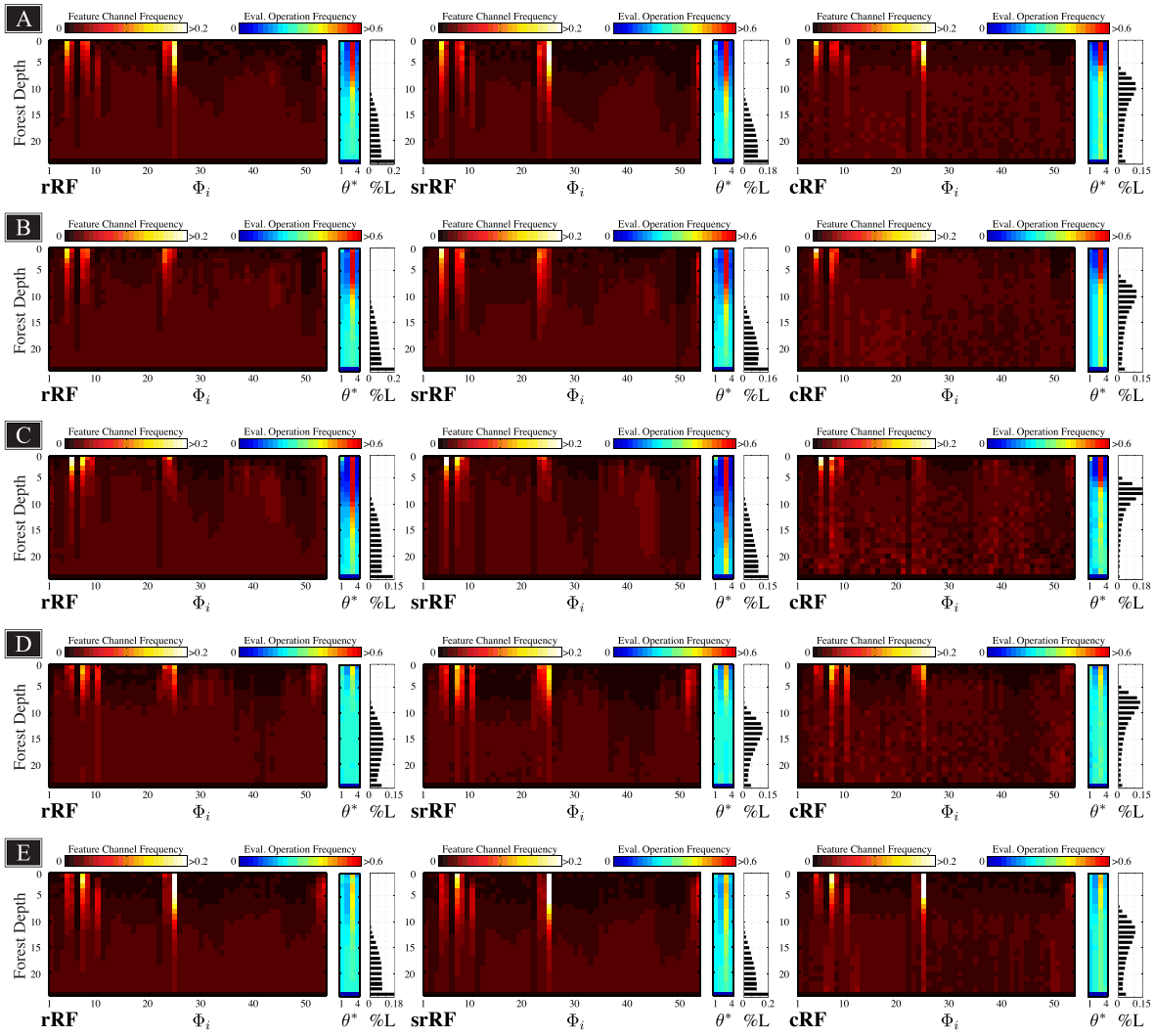


Figure 5.14.: Visualization of split decisions in the RFs ($T_{md} = 24$, $T = 64$), for all datasets as average over all trees and CVs: (A) *BM-HE*, (B) *BM-MGG*, (C) *BM-HE-MK* (20×), (D) *ICPR-BC*, and (E) *MT-HE*. Panels: Horizontal axes represent individual feature channels (Φ_i), evaluation operations in the optimal selection functions (θ^*), and the fraction of leaf nodes (%L) at each depth level of the forest (vertical axes).

NB: Use the electronic version for improved visibility. Blurring may occur due to the document viewer.

encoded by a continuum from blue to red colors[8]. In particular, four evaluation operations were defined: (1) single pixel values, (2) pixel value differences, (3) Haar-like functions, and (4) constrained pixel value differences (both image locations within an Euclidean distance of 10 pixels). For the purpose of comparison among the methods,

[8] Please note that here we borrow the symbol θ to denote *only* the evaluation operations uncorrelated to the feature channels.

the color coding for the selection frequencies was clipped at 0.2, and 0.6, respectively. In addition, the fraction of leaf nodes (%L) at each depth level is illustrated by black horizontal bar charts. For each depth level, mean frequencies for all datasets are reported in Fig. 5.14 (A-E), where values were first averaged over all trees, and subsequently over all CV runs to obtain a complete representation of a forest that is interpretable at a higher level. Some common and specific observations could be made for the datasets and methods, which are described in the following paragraphs.

Common Observations The first ten channels, resembling color and Canny edge features, were among the most frequently ($> 5\%$) selected channels across all datasets. RGB channels were generally of little relevance. While feature selection in the classification trees appeared to be more random, they focused on high-level Gabor features (Φ_{25-40}) more frequently in early levels than the regression trees. Until a tree depth of approximately 15, clearer preferences towards specific feature channels could be observed. This tendency vanished in deeper levels, where all features were selected with almost equal frequency. The most frequent features, defined as being a component of $> 5\%$ of the optimal split functions, were L* and a* and histogram-equalized grey-scale intensities ($\Phi_{4-5,7}$), local grey-scale intensity minima and maxima (Φ_{8-9}), Canny edges (Φ_{10}), HoG-like gradients and high-level Gabor features (Φ_{23-25}), as well as mid- and low scale LBP features ($\Phi_{52,54}$). Haar-like functions were the most frequent evaluation operations until a tree depth of 10-15, sometimes with a selection rate of $> 60\%$. Interestingly, it seemed that some correlation existed between the selection frequency of Haar-like operations and the tree depth, when first intermediate leaf nodes were created. Leaf node creation started once alternative evaluation operations were considered with equal frequency compared to Haar-like. This qualitative observation has not been further investigated quantitatively in this work, though. All methods constructed rather balanced trees until a depth of eight levels, where first intermediate leaf nodes were constructed. The distribution of the leaf nodes over the depth levels seemed to be specific to the methods, hence these characteristics will be described later in this section.

Dataset-Centered View Compared to other datasets, the three bone marrow datasets required the trees to utilize mid- to low-scale Gabor features more frequently at early depth levels, cf. Fig. 5.14 (A-C). Especially in the *MT-HE* dataset, the majority of optimal splits until a depth of approximately 20 involved a particular Gabor feature

channel (Φ_{25}), while low-level gradient orientation and magnitudes (Φ_{11-22}) were rather infrequently selected, cf. Fig. 5.14 (E). While at the root nodes in the *ICPR-BC* and *MT-HE* datasets Haar-like functions were used most frequently ($\approx 40\%$), alternative evaluation operations gained more importance after a depth of five to seven. When larger objects such as megakaryocytes needed to be localized, the focus of all methods was set on coarser features such as low-level Gabor and color features that were evaluated using Haar-like selection functions, cf. Fig. 5.14 (C). The mode of the leaf node distribution in the *ICPR-BC* dataset shifted noticeably to shallower levels, suggesting that a maximum tree depth of 24 was probably not required to learn the variability contained in this dataset. For all other datasets, deeper (regression) trees seemed possible.

Method-Centered View Apart from the commonly selected features described above, no clear tendency towards specific channels could be observed for the classification forests, but LBP features (Φ_{50-54}) were among the least frequently used. While for the regression methods the feature frequency patterns looked quite smooth, the results of the node optimization for the classification method indeed looked quite random. A less noisy frequency histogram could be observed solely in the *MT-HE* dataset, for which also less intermediate leaves were constructed. It seemed that these random-looking patterns in the feature selection started at depth 10-15, where the number of intermediate leaf nodes started to decrease again. As a result, a hypothesis that would require further examination could be that the distribution of the leaf histogram is somehow related to these patterns and that the trees already learned most of the information from the dataset until a particular depth. In early levels of the regression forests, the single pixel value evaluation did not contribute much to the selection function histograms. Haar-like functions on the other hand were more frequent only in shallow levels, whereas in deeper levels all operations were selected with almost equal average frequency. The classification forest tended to construct many intermediate leaf nodes at rather early levels ($\approx 8-10$), and very little leaves towards the deeper levels (> 15). Hence, these trees could perhaps be pruned earlier. Conversely, deeper regression trees seemed to be possible as the mode of the leaf nodes distribution was at the maximum depth, where tree growing stopped but still enough samples were available for further splitting. Despite these trees seemed quite balanced until a depth of ten, growing deeper trees might result in slightly more imbalance.

5.3.2. Runtime and Efficiency Profiling

The runtime of all detection algorithms was compared for training and prediction on the benchmark data split defined by the *BM-HE* dataset. All algorithms were executed on a workstation hosting two Intel Xeon E5-2630v2 2.60 GHz CPUs and 64 GB main memory. The RFs were profiled in several configurations and were subsequently compared to the runtime of the *SSVM* baseline cell detector. First, we examined them using the *BM-HE* benchmark settings from Section 5.1.1 ($T = 64$, $T_{md} = 24$, $|\Phi| = 54$), then using less complex models that achieved the same detection performance (F1-score) during the hyper-parameter search, i.e. $T = 16$ and $T_{md} = 16$. Finally, the RF configuration presented in Kainz *et al.* [8] ($T = 64$, $T_{md} = 16$, $|\Phi| = 21$, denoted by the * symbol) was profiled. The number of parallel execution threads was limited to 12 for all algorithms[9]. All timings are reported as wall time, final detection performance is reported as F1-score, cf. Table 5.10.

Training

The number of patches used for RF training differs due to the different dataset sampling strategies of the regression and classification methods. A training set of $479,156$ patches was sampled from 11 images ($\tau_{bg} = 0.5$) for the regression methods in around 103 seconds using all 54 feature channels. Since the classification forest used data augmentation of the original images, sampling lasted around 11 minutes for $58,878$ patches. The *SSVM* baseline method took around 9 minutes to generate $44,442$ training samples, and learned the model in around 5 hours. For the largest RFs, fitting the *rRF* model took around 1.5 hours, while training the *srRF* consumed 5.7 hours. Net learning time of the *cRF* was significantly faster than regression (by factor 14-52×) and took only 6.6 minutes. However, the data sampling took even more time than learning the model. In the reference RF configuration (*), we saw a remarkable speedup in data sampling for all RF methods: *rRF* and *srRF* approximately by factor 3, and for *cRF* by factor 25. This could be attributed to omitting the Gabor and LBP feature channels (30 in total), because these implementations have not yet been optimized. Pre-computing $|\Phi| = 21$ simple intensity and gradient feature channels was much faster. RF configurations using $T_{md} = 16$ and $T = 16$ resulted in a training speedup of approximately 6.5-7× for the regression methods.

[9] The RF implementations were parallelized using OpenMP, while the *SSVM* method was using parallelization in MATLAB and optimized C/C++ code using the MEX interface.

***BM-HE* Dataset**				Training: 11 Images			Test: 1 Image				
Method	T_{md}	T	$\lvert\Phi\rvert$	**Sampling** ($N_{\mathcal{P}}$)		**Fitting**	W_S	**Prediction**		AUC	F1
rRF	24	64	54	479,156	102 s	5,428 s	1	78.5 s	25.6 s	0.9243	0.8759
	16	16	54	479,156	100 s	836 s	1	78.6 s	6.4 s	0.9190	0.8742
	24	64	21	479,156	39 s	3,459 s	1	9.2 s	24.1 s	0.9185	0.8749
⋆	16	64	21	479,156	43 s	1,485 s	1	9.5 s	16.2 s	0.9103	0.8713
	16	16	21	479,156	24 s	482 s	1	9.0 s	4.6 s	0.9084	0.8702
srRF	24	64	54	479,156	105 s	20,620 s	1	84.1 s	38.4 s	0.9201	0.8794
	16	16	54	479,156	106 s	2,884 s	1	80.5 s	8.5 s	0.9122	0.8757
	24	64	21	479,156	39 s	13,115 s	1	10.0 s	29.3 s	0.9152	0.8720
							4	3.1 s	2.4 s	0.9139	0.8703
⋆	16	64	21	479,156	30 s	6,442 s	1	9.9 s	19.1 s	0.9092	0.8684
							4	2.8 s	1.7 s	0.9084	0.8672
	16	16	21	479,156	38 s	1,913 s	1	10.0 s	7.1 s	0.9052	0.8658
							4	2.9 s	0.6 s	0.9079	0.8672
cRF	24	64	54	58,878	633 s	400 s	1	78.4 s	24.7 s	0.8963	0.8733
	16	16	54	58,878	633 s	365 s	1	78.0 s	4.6 s	0.8764	0.8682
	24	64	21	58,878	24 s	22 s	1	9.1 s	18.3 s	0.8780	0.8648
⋆	16	64	21	58,878	25 s	16 s	1	8.7 s	13.6 s	0.8760	0.8654
	16	16	21	58,878	24 s	4 s	1	8.9 s	4.5 s	0.8737	0.8657
SSVM	–	–	151	44,442	541 s	17,748 s	–	26.1 s	5.2 s	0.9144	0.8665

Table 5.10.: Runtime and efficiency profiling results of the cell detection algorithms on the *BM-HE* benchmark dataset ($1,200 \times 1,200$ pixel images). All timings are reported as wall time. Training was profiled regarding duration of training dataset sampling as well as the net model parameter fitting time. Inferring cell center locations on a single test image is reported as average over all four test images. The fastest detector is *srRF* (highlighted in light gray), being 9× faster than the *SSVM* baseline. The first timing refers to feature computation, the second to the net model inference and applying post-processing. Detection performance is given as F1-score for all models. Variations in runtimes are due to background activities of the test machine.

⋆ Reference RF configuration from Kainz *et al.* [8].

Inference

To report the inference times on a single test image (1.4 million 33×33 pixel test image patches), we averaged data sampling, and net model inference time over all four test images in the *BM-HE* dataset. The baseline method [116] predicted a single test image in 31 seconds, while net inference time of the structured SVM was only 5 seconds. All RF methods were considerably slower in their largest configurations, but their F1-score was slightly higher than the baseline. Post-processing the proximity score and probability maps took about 0.9 seconds, which was included as constant

in the prediction times. The required runtime for feature channel computation for a single image could be reduced to 9-10 seconds, when using less channels, in the absence of detection performance loss. Interestingly, the detection performance did also not decrease much as less and shallower trees were employed. However, the biggest benefit of using simpler RF models was a decreased prediction time, which could be reduced by a factor of ≈ 5.25 down to 5-7 seconds. Compared to the other methods, predictions using *srRF* took a little longer, since the vector output for each input patch was mapped back to the pixel coordinates. The *rRF* with the lowest evaluated complexity was able to achieve higher performance in terms of runtime ($\approx 2\times$ faster) and cell detection accuracy than the *SSVM* approach [116]. The other RF methods showed similar runtime and comparable detection performance. These findings are in accordance with the results of earlier work using the reference configuration (*) [8]. While in [8] the original *SSVM* implementation as provided on-line by the authors was used, it was further optimized for the experiments in this thesis. Hence, the baseline speedup between the reference configuration and the optimized *SSVM* method was only $\approx 1.23\times$.

In order to obtain predictions for an unseen image, a sliding window is moved over each image location to predict the center pixel. The *rRF* and *cRF* methods were able to achieve a speedup of $\approx 2\times$, but applying the *srRF* detector using a window stride of $W_S = 4$ pixels (output patch size $p_{out} = 11$ pixels) facilitated a major speedup of $9\times$ compared to the *SSVM* baseline. We could reduce the complexity of the RF to a minimum, which led to a runtime for detecting cell center locations in a single image of only 3.5 seconds. A speedup of approximately 7-12$\times$, depending on the RF size, was observed compared to densely predicting each image location. Please note that this speedup over the baseline could be achieved in the absence of detection performance loss. However, dense predictions using more complex models resulted in higher performance. Further optimizations in terms of a lower number of feature channels were not explored so far.

6. Discussion and Conclusions - Cell Detection

6.1. Proximity Score Regression using Random Forests

In the previous chapters, a novel cell localization method was proposed and evaluated on multiple challenging histopathology datasets. A regression Random Forest algorithm was employed to learn, for each image location, a smooth, non-linear function of the distance to the closest cell nucleus center, we termed *proximity score*. Given a set of local image patches, each patch is labeled with the proximity score. For unseen images, the trained forest predicts the proximity score map using a sliding window, where local maximums revealed in a post-processing step by non-maximum suppression correspond to cell nuclei center detection hypotheses. The proposed detection methods via regression relies on a set of hyper-parameters for both the learning algorithm and post-processing. In Section 3.2, reasonable ranges based on prior knowledge of the data were proposed. The determination of the average cell nucleus size and Euclidean distance between a nucleus center and its nearest neighbor is required, and can easily be done manually as an initial step on a tiny subset of ground truth cells, if center-dot annotations are available instead of bounding boxes or segmentation masks.

The local image patches were represented by a set of pre-computed visual features that comprised intensity and texture features based on different color spaces and image gradient information, cf. Section 3.2.2. The non-trivial task of accurately segmenting single objects has been circumvented, no shape-related features were extracted. However, features computed from object shape are quite discriminative and contribute to

detection methods that work at multiple object scales. Despite for instance Gabor and LBP features were used to describe the image content at multiple scales, the localization method cannot yet be considered as scale-invariant with respect to the object scale. Nevertheless, this representation seemed to be sufficient, since we defined separate problems here for the localization of granulopoietic and erythropoietic cells, and megakaryocytes, which are up to five times the size of other hematopoietic cells, cf. Section 2.1.

A research goal of future work could be to perform the detection of all hematopoietic cells in a single step. To endow a proximity score predictor that simultaneously detects different cell lineages characterized by different object sizes would require further experiments, which probably are going to result in learning different, and more complex models. More importantly, it requires at least bounding boxes [166], or segmentation masks as ground truth annotations. In Section 3.1.2, it was argued why simple dot-annotations enable a fast ground truth generation, thus avoids object segmentation. While it is a strength of the proposed method to learn from dot-annotated objects, it is somehow limited to a certain object scale and size. This limitation becomes visible as soon as multilobed megakaryocytes are present at 40× magnification. The appearance of a single lobe is locally very similar to a nucleus of early erythroid and myeloid precursor cells and get therefore detected as individual objects, see the example in Fig. 6.1. It seems that at this scale, the RF does not consider perinuclear cytoplasmic areas, but only nucleus appearance as visual cue to learn the nuclei centers. This is actually reasonable given that the proximity score maps are heuristically computed from annotations using the average object size of non-megakaryopoietic cells only. However, since megakaryocyte detection works convincingly well at lower magnification, cf. results in Section 5.1.3, linking these detection hypotheses to locations in the same images at full resolution (40×) should be trivial. Further, it should be interesting to see in prospective empirical evaluations, whether the high precision and recall of the regression methods for megakaryocyte detections can be maintained in an integrated approach.

The optimal hyper-parameters were determined in grid search using CV for each combination. The search proceeded in three stages, where stage I and II determined suitable hyper-parameters to minimize the NRMSE of the single-target and spatial-averaging regression models, and stage III for common post-processing steps to maximize the detection-specific optimality criterion $\mathcal{O}$, cf. Eq. (4.15). The set of visual image features was fixed for both methods in all stages. While this strategy drastically reduced the search space, we now cannot except that a more optimal result for each dataset could

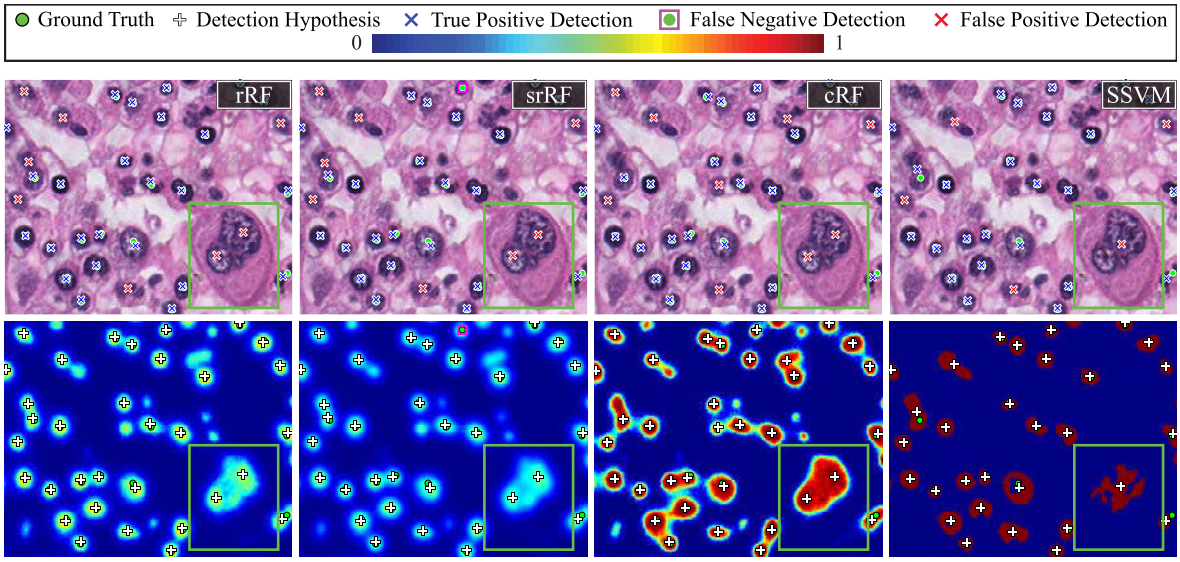


Figure 6.1.: Results of detectors trained to recognize erythropoiesis and myleopoiesis in H&E images at 40× magnification. Due to the similar appearance, each lobe of the megakaryocyte nucleus is detected as individual (false positive) object by the RF methods (green square). The *SSVM* method detected just one nucleus.

have been determined by including the selection of suitable feature channels (from all available) in the hyper-parameter search as well. Nevertheless, the search would have been much more time consuming accompanied by an increased risk to create an *ad hoc* solution. An alternative approach to search for optimal hyper-parameters in a high-dimensional space is to use random search as proposed by Bergstra and Bengio [187]. However, it is questionable whether the random search would have resulted in superior results in our case, since they argue that in low-dimensional spaces, grid search is reliable. Moreover, we selected a reasonable grid covering the value ranges.

A certain extent of robustness of the proposed regression methods against variations of hyper-parameters was further confirmed in stage I results, especially on the *BM-MGG*, *ICPR-BC*, and *MT-HE* datasets, cf. Figs. A.3 (a), A.9 (a), and A.11 (a) in Appendix A. Considering the input patch size p_{in}, the NRMSE as well as the F1-score is rather stable across all tested values, suggesting that the actual value within this range is uncritical for this method to work. Possible values for input patch sizes were stressed in initial experiments on the *BM-HE* dataset using $p_{in} \in \{6, 128\}$ before deriving reasonable ranges, cf. Fig. A.1 (a). It could be observed that when p_{in} was chosen very small, the field of view was too limited and the image content could not be represented properly by the 54 feature channels. More precisely, patches sampled from locations on the cell nuclei and preparation artifacts such as cell debris were too similar,

leading to ambiguous training data. Hence, the RF learned a model that predicted high proximity scores for insignificant patches that had 'nuclei-like' appearance. While this was beneficial for recall, it led to significantly lower precision, i.e. more preparation artifacts were detected as false positive cells, cf. Fig. A.1 (a). Nevertheless, we did not yet consider the *background* labels that are available in the *BM-HE* and *BM-MGG* datasets. By including these samples as negative class, the detectors could likely be trained to be more robust with respect to such artifacts.

Deeper and larger forests generally produced a lower NRMSE. One major reason is the fact that the smooth non-linear target is approximated by piece-wise constant functions, stored as mean proximity scores in the leaves of the regression trees. Naturally, deeper trees are equipped with more leaves and are therefore able to learn a more detailed approximation. We could not observe any tendency to over-fit the datasets in stage II of the hyper-parameter search. However, once the optimal input and output patch sizes had been determined, the quantitative detection results with respect to the F1-score were surprisingly constant. In fact, despite more complex models predicted more accurate proximity score maps, the F1-score did not differ much between forests consisting of four shallow trees or 64 trees with a maximum depth of 24. This behavior was observed in all but the megakaryocyte (*BM-HE-MK*) datasets and suggests that the post-processing is capable of compensating imprecise proximity score maps. This further questions the time-consuming training of more complex RF models, if such simple heuristics-based post-processing results in similarly high detection performance. F1-score, however, is just a condensed representation of a method's selectivity, hence precision and recall must be inspected separately. Yet, we did not observe any considerable variations of the latter measures across different model complexities either (except for the *BM-HE-MK* dataset).

The sampling strategy for the regression methods did not consider the entire range of the proximity scores due to the foreground-background threshold τ_{bg}. This constraint was set mainly due to the host memory limits, since training was implemented in an off-line manner. However, setting $\tau_{bg} = 0.1$ and sampling background locations accordingly did not yield improved detection performance, suggesting that sampling small areas around a true cell center using $\tau_{bg} = 0.5$ was in fact sufficient.

6.2. Additional Value of Spatial-Averaging Regression

Pursuing a patch-wise prediction using the spatial-averaging regression showed some advantages, and unveiled potentials for improvement. Considering the predictive efficiency of the two regression-based methods, it could be observed that the NRMSE is usually lower for spatial-averaging regression (*srRF*) compared to single-target regression (*rRF*). In stage II of the hyper-parameter search (optimal RF complexity) it was found that *srRF* significantly relaxes the requirement for complex trees to result in acceptable predictive accuracy, see Fig. A.2 (a,b) for an example on the *BM-HE* dataset. Using the *lowest* possible complexity of $T_{md} = 8$ and $T = 4$, spatial-averaging regression was able to produce NRMSE values that were equal or even lower than the ones produced by single-target regression using the *highest* tested complexity of $T_{md} = 24$ and $T = 64$. This gap further increased when the complexity of the RF for spatial-averaging regression was increased. The same observation could be made on all other datasets as well.

Naturally, one would expect the lower NRMSE of *srRF* to be pandering for improved detection performance. Surprisingly, in the majority of experiments it was observed that the *srRF* yielded results comparable to *rRF*, but scarcely ever surpassed the performance of *rRF* in terms of precision. One possible reason for that behavior is that the *srRF* method inherently produced a smoother proximity map by merging multiple predictions for each image location, cf. Eq. (3.19). The output patch size indirectly controlled the smoothness of the map by the number of predictions per image location. In other words, larger patches produced smoother maps. Hence, the resulting local maximums were usually of lower magnitude as for the *rRF* method, where individual locations were predicted only once and the maps were a little noisier, cf. Fig. 6.1. While smoothing the proximity score map of *rRF* with a Gaussian kernel $\mathcal{G}$ sometimes suppressed weak peaks, it did not have the same effect in the *srRF* approach. Applying $\mathcal{G}$ caused over-smoothed maps and hence reduced the difference between lower and higher proximity scores. At the examined resolution of κ (each 2% of the proximity score range) it was not possible to remove as much false positives as for *rRF* during post-processing. Though, in all datasets but the *ICPR-BC*, the largest tested output patch size was optimal ($p^*_{out} = 11$) and resulted in the smallest NRMSE. The optimal proximity score threshold κ^* was found to be usually higher for *rRF*, favoring less false positive detections after diminishing local maximums. Recall, on the other hand,

was approximately the same for both methods across all datasets. Yet, we did not perform evaluations with increased resolution of κ and must leave this subject to future research.

Spatial-averaging regression provided very low prediction errors and can therefore be expected to overcome current limitations of the single-target regression method. We have shown that another noteworthy advantage of using *srRF* over other evaluated methods lies in using a sliding prediction window stride of $W_S > 1$. In the absence of detection performance loss, the proposed method is able to achieve a 9× speedup over the baseline method (*SSVM*), and 12× over dense prediction. However, to take full advantage of the mostly considerably lower prediction error over single-target regression, future work should evaluate ways disclosing the full potential of spatial-averaging regression.

6.3. Discussion of Qualitative Localization Results

Our initial hypothesis that common problems of the classification method can be solved by switching to regression could only be confirmed to some extent. Sometimes multiple detections on single objects, or merging of very close cells could not be avoided by any approach, but their occurrence was reduced and the spatial localization accuracy was increased when using regression.

Cells in early stages of granulo- and erythropoiesis are characterized by their larger size, more granulated chromatin staining in their nuclei, and high intensity gradients at their nuclei borders, cf. Fig. 2.4 (a,b). Some false positive detections were observed on such cells, since the RF predicted multiple high responses on their nuclei that could not suppressed by the available NMS window size. False negative detections were most frequently observed by all methods when the cell nuclei were very tiny, or too close to another cell. In the first case, the predicted proximity score and prediction confidence was too low, or the SSVM rejected the candidate, whereas the latter case the non-maximum suppression window only considered the stronger peak, and the MSER detector merged the candidate regions, respectively. Less frequently, myeloblasts and proerythroblasts were missed by the regression and *SSVM* methods, but were recognized by the classification approach. However, obtaining their locations via classification suffered from multiple responses on single nuclei.

With very few exceptions, recall was higher than precision in all methods. Nuclei of more differentiated cells in later stages of maturation are smaller and usually exhibit a more homogeneous staining that in extreme cases became just an easily detectable dark blob. Hence, the training dataset contained a large number of such cells that could not be discriminated properly from cell debris and favored a higher number of false positives.

Despite their unique appearance among the hematopoietic cell lineages, localizing megakaryocytes in an image is subject to other factors. Due to their size, they can be cut at many more different positions in 3D space when the histopathological sections are processed, which leads to a larger variability in their appearance. Frequently, the characteristic multilobed nuclei and extensive surrounding cytoplasm are visible (Fig. 6.2 (a)), but sometimes only a tiny part of the nucleus is depicted (Fig. 6.2 (b)). In the latter case, all evaluated methods had difficulties detecting the cell properly in the images, despite the response of the *rRF* detector is well positioned on the cell center and locally maximal. However, this local maximum was lower than the optimal proximity score threshold κ^*, which consequently resulted in a FN detection. In our representative *BM-HE-MK* dataset, only a small fraction of cells was cut similar to the one depicted in Fig. 6.2 (b), but there could of course be many more like that in a whole slide image. Megakaryocytes only constitute roughly 0.5-2% of the DBC (cf. Section 2.1), hence missing many of these cells may distort the results. Tackling this problem requires a more robust inference scheme than the current simple proximity score thresholding and must be considered in future work.

When considering the different magnifications of the datasets, we could observe that the spatial localization accuracy among the 40× datasets *BM-HE*, *BM-MGG*, and *MT-HE* are quite similar. However, μ_d was quite large for the megakaryocyte dataset at 20×. One possible reason becomes apparent once we consider the cytology of megakaryocytes with respect to the other cells. For instance, other bone marrow cell lineages, or lymphocytes in breast cancer tissue do not have such extensive perinuclear cytoplasm. Hence, the center is frequently well-defined as approximately the center of gravity of the minimum bounding box around the nucleus. Cell objects in the *BM-HE-MK* dataset have been annotated similarly, but depending on their cutting plane, megakaryocytes are subject to much more variability of their observable N:C ratio. Further, the cytoplasm may not appear concentrically around the nucleus. Hence, this resulted in the ground truth center-dots being (randomly) placed on either a (perhaps only partly depicted) nucleus, or into the perinuclear cytoplasm. Apparently, the RF

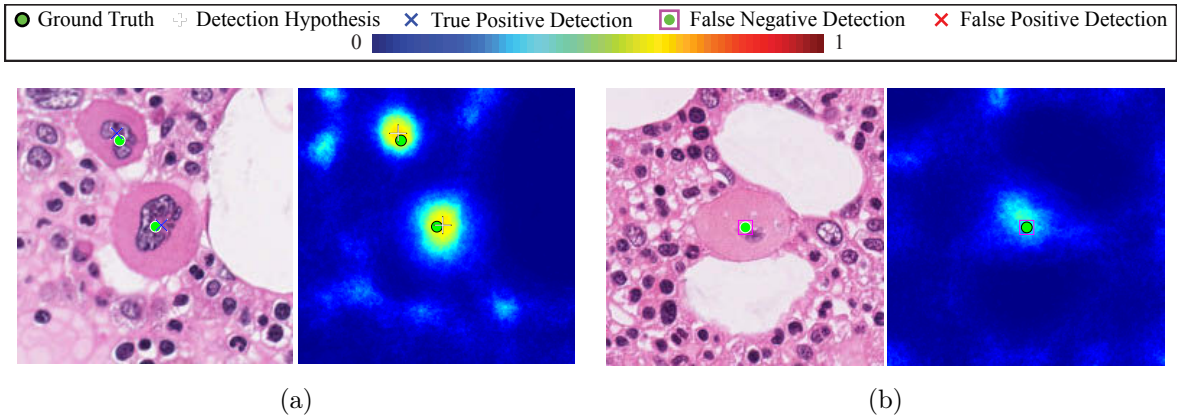


Figure 6.2.: Detection results of the *rRF* method on megakaryocytes, cut at different positions in 3D space. (a) The cell nucleus is well depicted and surrounded by a large area of cytoplasm, suggesting a beneficial cutting position, perhaps close to the equatorial plane. (b) This cell was cut at a location farther away from the equatorial plane, resulting in only a partially depicted nucleus surrounded by cytoplasm. The cell detection worked as expected in (a), but the detector generally had difficulties detecting cells such as the one shown in (b). However, the prediction of the RF is well located on the cell center, but the local maximum was lower than the optimal proximity score threshold κ^*. A more robust scheme to reveal local maximums is required to deal with these shortcomings.

learned to prefer the nuclear parts of the megakaryocytes as centers, placing the hypotheses frequently farther from the true center, which was located within cytoplasm. For applications with the primary focus on sole detection and counting, larger deviations from the true megakaryocyte nuclei centers should not be considered an issue, as long as a megakaryocyte gets detected as a true positive. However, as we are aiming at an maximally accurate estimate of the cell centers, using a lower magnification (10×) has been shown to be more expedient and less distracting for the RF.

The *SSVM* method generally performed well at detecting mononuclear cells in our datasets, especially when the nucleus staining was quite uniform, which was favorable to the candidate detector (MSER). Particularly, this could be observed on the *BM-MGG* dataset, where this method achieved the highest detection performance (F1-score) in the CV experiments. However, it failed at detecting megakaryocytes, expressing similar symptoms at both object scales under examination. For megakaryocytes, the appearance of the cell nuclei does not solely determine the cell type, but also perinuclear cytoplasm represents a significant cue. However, as argued above, many of the ground truth center annotations in the *BM-HE-MK* dataset were actually located in the cytoplasm. The unsatisfactory performance of the *SSVM* method can be pinpointed to

the way candidate regions, revealed by the MSER detector and pruned by a binary classifier, are presented to the final SSVM classifier, cf. Section 4.3.2. The classifier apparently was not able to discriminate cytoplasm and the intercellular matrix, which were represented similarly in the feature space (at least in our H&E stained dataset). From this we can conclude that the *SSVM* method is not able to learn a megakaryocyte detector from simple dot-annotations. Many false positive locations were hypothesized in areas that are in fact background, cf. Fig. 5.9. Using the proposed RF methods, we pursued a different approach that does not suffer from these drawbacks, rendering them more robust and versatile for such tasks.

6.4. Potential for Improvements and Future Work

The detection performance of the proposed proximity score regression methods was satisfactory across all evaluated datasets. Since good generalization properties was observed in our experiments, we currently did not consider any additional pre-processing steps such as denoising or staining normalization [191, 192]. The cell detectors were explicitly trained on the H&E and MGG staining, respectively. Transferring learned models to predict the untrained staining also worked well, but it was not examined whether training on a mixed staining dataset is feasible. This would facilitate building a general-purpose cell detector that works simultaneously on multiple stainings, which at this point is subject to future research. However, the current object scale limitation needs to be addressed separately.

In all RF experiments recall was higher than precision and F1-score seemed to saturate. Ground truth annotations are usually created manually, and as a consequence contain a certain level of label noise (object mislabeling, inaccurate locations, etc.). While inaccurate labeling can be neglected for dot-annotations, a rigorous identification of an object being a part of a cell, staining artifact, or debris is sometimes not feasible. In our bone marrow datasets, such questionable objects were labeled as *unknown* and learned to detect them along typical cell nuclei. Hence, the detectors were a bit biased towards deciding in favor of recall and tolerated more false positives than false negatives, which in the context of the current application is actually beneficial. This leads to more candidates that ultimately need to be classified into one of many cell maturation stages, background, or unknowns. Further potential improvements regard adding more power to the detection method by forcing it to reduce the number of false positives in order to

filter the dataset for a subsequent classifier. Since for bone marrow negative labels are already available, it can be expected that their use in training increases the selectivity of the classifier. For bone marrow cell detection it was perhaps not ideal to use the highest available magnification, since more false positives were detected than false negatives. It must be examined in future research, if the detection of small artifacts decreases at magnifications lower than 40×, while maintaining the current recall. Touching and overlapping cells were rare in all datasets. Therefore it was not of high priority to our problem to explicitly consider separation strategies. However, we have seen that our algorithm has difficulties to separate some close cells. Options to separate close objects have been previously proposed [193–196], which could be applied to our problem as well. Though, it would be more beneficial, and even more elegant, if the localization algorithm was able to perform the detection and the separation of close objects in a single step.

Ensemble methods (meta-algorithms) are a well-known strategy to improve and stabilize the performance of learning algorithms. Hence, the proposed cell detection method relies on the RF algorithm [165] to learn the proximity score regression model. In contrast to a single regression tree, an ensemble of many randomized decision trees is used in RFs, whose predictions are consolidated via averaging. De-correlation of the tree ensemble was achieved using randomized node optimization. Averaging produces smoother final predictions and reduces the overall prediction variance [164], i.e. 'random' errors are minimized. A closely related ensemble method for decision tree training is bootstrap aggregating (bagging) [177], where standard CART trees are learned on a dataset that is randomly subsampled with replacement from all available training samples. In principle, bagging could also be used in RFs by training the randomized trees on bootstrap subsets. However, we did not observe any performance gain when this was tested in initial experiments and therefore opted for RF without monitoring the out-of-bag error rate. While RF and bagging train a set of (either weak or strong) base learners in parallel, boosting[1] [127] combines a sequential cascade of weak learners to produce the final prediction as weighted average. In each 'round' of boosting[2], the samples in the dataset are reweighted such that falsely classified samples get higher weights in the subsequent round, where the weak learner is forced to predict these samples correctly. Each weak learner is weighted according to the prediction confidence (weighted error) in the final averaging step. A closely related an alternative method to learn is

[1] In particular, we refer to the AdaBoost algorithm proposed by Freund and Schapire [127]. Other variants of boosting have also been described in literature, e.g. boosting by filtering [197].

[2] The term 'round' actually refers to a particular stage of a sequential cascade of weak learners.

gradient boosting [164], which was originally employed by Sironi *et al.* [162, 163] in their center line detection algorithm that inspired our proximity score regression for cell detection.

While it first seems appealing, applying the idea of sample reweighting is of little value here, since we used a different random subset of M_j samples to evaluate the split quality in the randomized node optimization. Using all samples was computationally expensive and did not result in a performance gain that justified these increased runtimes. Nevertheless, inspired by concepts of boosting, we could perform iterative RF training to enrich the training set with difficult samples in order to increase the robustness of the predictors [169, 198]. Instead of pursuing the proposed sampling strategy using all available training data at once (cf. Section 3.2.2), we could initially train an RF on a bootstrap subset. We would then predict the remaining dataset and retain difficult samples with high prediction errors. In each subsequent round, a new bootstrap dataset is added to the retained samples and a new predictor is learned. These iterations are repeated until a defined stopping criterion applies. Future research must hence evaluate proper stopping criteria and whether the models become more robust after a few rounds of bootstrapping.

Further, since we relied on randomized node optimization, an important parameter that influenced the performance of RFs is the number of randomly selected split functions in the internal nodes. In a separate experiment, which was not part of the extensive evaluations, we examined whether an increased number of randomly drawn split functions influences the detection performance. Depending on the three choices of $p_{in} \in \{17, 33, 50\}$, the total number of tested selection functions was $N_\theta \in \{124, 242, 367\}$. Compared to these automatically determined numbers, we tested $1,000$ selection functions with 20 random thresholds each in a LOOCV on the *BM-HE* dataset during stage I of the hyper-parameter optimization, i.e. $N_\phi = 20,000$. We could not observe noticeable differences in the performance measures when using a higher number of split tests N_ϕ, cf. Table 6.1. Training the models required $2.7 - 8\times$ more time, though. The lack of difference in overall performance and the elevated training times in fact did not justify choosing a larger number of random node tests. With respect to the total number of available features (i.e. pixels in the feature channels), the number of tested split functions attributing to higher performance is saturating at some point. Hence, it becomes more likely for smaller patch sizes that highly similar split functions were tested redundantly.

Moreover, in Section 5.3.2 the runtime has been profiled for RF models that achieved

p_{in}	N_ϕ	PRC	REC	F1
17	2,498	0.8001	0.8854	0.8406
	20,000	0.7576	0.9408	0.8393
33	4,850	0.7961	0.9211	0.8540
	20,000	0.7895	0.9313	0.8546
50	7,348	0.7816	0.9277	0.8484
	20,000	0.8007	0.9037	0.8491

Table 6.1.: Detection performance of the *rRF* on the *BM-HE* dataset in LOOCV for different total numbers of split tests $N_\phi = 20 \cdot N_\theta$.

a very similar qualitative detection output (0.8706 ± 0.0043). From these results we can conclude that in fact we did not require all 54 visual features and that forests of lower complexity can be used as well. A set of color and gradient features [8] has been shown to be sufficient and resulted in a considerable speedup during testing. It can be expected that there is further potential for runtime improvement, for instance when further reducing the size of the forest. However, this needs to be examined in further experiments. On the other hand, the feature channels were selected manually, which also raises potential for improvement. For instance, as an alternative, required features could be learned in an end-to-end approach, e.g. using CNNs [149]. Others [199] proposed a method based on decision trees to learn required features from random subwindows in conjunction with a generic image classification problem. It has to be evaluated how their approach can be applied to a localization problem. Schulter *et al.* [168] proposed alternating decision forests for object detection, which performed better than standard RF and boosting while producing more compact trees. Since reduced runtime is an important property for cell detection in whole slide images, this method could be considered in future research.

Kirby: "You know, buddy, we've got a lot in common! I'm Matt Kirby."
Walsh: "... and I am not."

Matt Kirby alias Terence Hill
Wilbur Walsh alias Bud Spencer (1929–2016)

Part III.

Quantifying Hematopoietic Cell Maturation

7. Classifying Bone Marrow Cells

In the previous chapters, a general cell localization method was proposed that provides the required locations of objects to perform cell type classification. The following chapters concern the classification of granulopoietic and erythropoietic cells using an alternative approach to image classification.

7.1. Introduction

The initial step of diagnostic work in histopathology is the assessment of cellularity in context of tissue architecture. Especially the diagnosis of bone marrow specimen requires a valid interpretation of different cell types with respect to their local distribution. The maturation of blood cells (hematopoiesis) is categorized into granulopoiesis, erythropoiesis, and megakaryopoiesis, which refer to maturation of white blood cells, red blood cells, and megakaryocytes, cf. Chapter 2.

In healthy individuals, hematopoiesis mainly occurs in bone marrow, whereas extramedullary hematopoiesis is observed during fetal development, or may indicate pathological alterations [33]. In bone marrow specimen, several thousand cells of multiple classes in various stages of maturation have to be interpreted by the hematopathologist, and the class distributions need to be reported. This qualitative and sometimes semi-quantitative classification is usually performed on H&E stained tissue sections. The correct classification based on cell morphology and spatial cell distribution heavily dependents on the observer's experience, since in hematopoiesis disparities between subsequent development stages are frequently indistinct. Even equal maturation levels of different myeloid precursor cells share morphological characteristics, cf. Fig. 7.1. As a consequence, both inter- and intra-observer variability can be considerable, affecting the accurate diagnosis of reactive or even premalignant, and early malignant alterations.

Thus, automated cell recognition systems exhibiting low variance, high classification accuracy and predictable error are desirable to facilitate valid repeatable quantitative biomedical diagnostics [57]. Since virtual microscopy using high-resolution whole slide images has been emerging to a standard in pathology departments [5], proper computer-aided pathology using computer-assisted image analysis systems could easily be implemented in routine diagnostic processes.

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

Figure 7.1.: Samples of hematopoietic cell nuclei in the human bone marrow at 40× magnification and stained with H&E. (a) myelocyte, (b) metamyelocyte, and (c) band cells are subsequent maturation stages of a granulopoietic cell. Especially in early stages, where the cells are not fully differentiated, different cell lineages share morphological characteristics, e.g. (d) myelocytes (granulopoiesis) and (e) orthochromatic normoblasts (erythropoiesis).

7.1.1. Related Work

A remarkable amount of research has been conducted on blood cell counting, segmentation and classification in histopathological images for various applications in recent years. Motivated by the aggressiveness of blood cancer and the requirement for early diagnosis, most works were related to leukemia research, in particular identifying different types of leukemia by classifying WBC from histopathology images of peripheral blood smear [76, 81, 137, 200–207] or bone marrow, obtained by aspiration [80, 208–216] and trephine biopsy [79]. Particularly, some work focused on classification of WBC in healthy tissue [201, 206, 217], while others dealt with detecting pathological alterations from morphological characteristics of entire cells and cell nuclei [80, 81, 213–215].

In analogy to the related work reported for cell detection in Section 3.1.1, a vast majority followed a conventional pattern recognition approach and used distinct steps for cell detection, segmentation, extraction of rotation and translation invariant features, and classification. Some studies mainly addressed detection and segmentation and relied on standard image processing techniques such as Hough space analysis [207], watershed transform, Gabor filters and adaptive thresholding [217], or intensity clustering [213, 214]. Others pursued supervised learning-based cell detection using FF-

NN [218], fuzzy cellular neural networks [203], and Random Forests [8]. Several approaches used statistical pattern recognition and classification techniques [219] such as SVM [126, 137, 210, 216], FF-NN [137, 197, 200, 204, 208, 211, 212, 220], or Bayesian classifiers [81, 202, 221] to learn feature vectors representing individual cell objects that were obtained by segmentation. Decision tree based methods such as regression trees [79], hierarchical trees using genetic algorithms for node optimization [222], or RFs [80, 165, 215] were used as well as k-nearest neighbour [80, 81, 209, 223], or heterogeneous classifier ensembles [215, 223]. Employing hierarchical models has been shown to be more powerful than using single-stage classifiers [205, 216].

Shape and texture features of cell nuclei were most frequently used and seemed to provide more discriminative power than statistical features from intensity histograms. This is reasonable, since – despite proper histopathological staining protocols – nuclei of different cell classes share very similar intensity patterns after staining [76], cf. Fig. 7.1. However, the choice and importance of features depends on the application. For instance, it was shown that features computed from cytoplasm could even be omitted for WBC classification and that the problem could be downscaled to using features from cell nuclei only [212]. In the context of another application, using features from both nuclei and cytoplasm conjointly resulted in higher classification performance [224]. Recent work of Reta *et al.* [80] on bone marrow cells concluded that features extracted from nuclei and cytoplasm separately are more discriminative than features from entire cells. In previous studies, the total number of features varied from a small set of four to over 190, comprising object-level features as well as global image features such as wavelet coefficients [200]. Nevertheless, handcrafting features from images requires prior knowledge and experience, they are not easily transferable to other problems and may as well remove significant information, or introduce non-discriminative information. Thus, authors of previous papers frequently extracted a huge set of feature candidates and applied automatic selection procedures to extract the most significant subset, hence compress the available information to achieve a better generalization performance [137, 202, 206, 223]. It has been shown that this strategy generally improved the classification results compared to using all available features [202, 210, 216] on specific problems. On the other hand, working directly on image intensity data provides a directly observable object representation that is not influenced by errors of preceding segmentation steps that are frequently inevitably when extracting object-level features. Nevertheless, only a minority of previous work focused on learning a classifier from raw cell images [137, 203, 225], but reported promising results.

Very little work has been reported on quantitative analysis of bone marrow trephine biopsy images [79], or quantification of blood cell maturation [217]. Tissue micro-architecture is usually well preserved after histological preparation in bone marrow trephine biopsy samples. At the proper magnification, and using suitable histological staining protocols, this enables the inspection of the morphological differences between subsequent maturation stages, but also introduces and emphasizes background structures irrelevant to cell classification, cf. Fig. 2.7 in Section 2.2. The most common staining used for tissue specimen are H&E and MGG (for bone marrow) and Wright's stain (for peripheral blood), since morphological characteristics of the cells of interest can be well represented. Feature-based discrimination of cells is usually less complicated in peripheral blood smear images depicting differentiated cells than in trephine biopsies, where immature forms of blood cells are frequent. Further, segmentation methods can more easily be applied to the cell objects without getting distracted by heterogeneous background. Despite the efforts of previous work, several issues have not yet been addressed, and the quantification of blood cell maturation in the bone marrow has not been sufficiently studied yet.

7.1.2. Goals and Organization of this Part

In this part, an alternative approach to bone marrow cell classification is proposed. It is based on the direct application of a Recurrent Neural Network (RNN) to raw images of H&E stained bone marrow trephine biopsy tissue. Under the conceptual framework of *reservoir computing*, two related effective training methods for RNNs have been developed independently: Echo State Networks (ESN) [226] and Liquid State Machines (LSM) [227]. Both approaches use a randomly and recurrently connected pool of hidden units, and learn to classify the observed temporal activities by adapting the readout weights only. While LSMs are considered as a biologically more realistic model, applying ESNs is usually easier due to a reduced number of hyper-parameters. Face recognition using a combination of an ESN and a FF-NN has been presented by Woodward and Ikegami [228]: the ESN extracted features, and inference was performed by the FF-NN. However, since not of paramount importance for their application, their approach did not consider any rotational invariant aspect.

The main motivation for this work is based on the fact that cells can be observed under arbitrary in-plane rotations in a histopathological section. In a conventional rotation-invariant supervised learning setting, one could train exhaustively on additional sam-

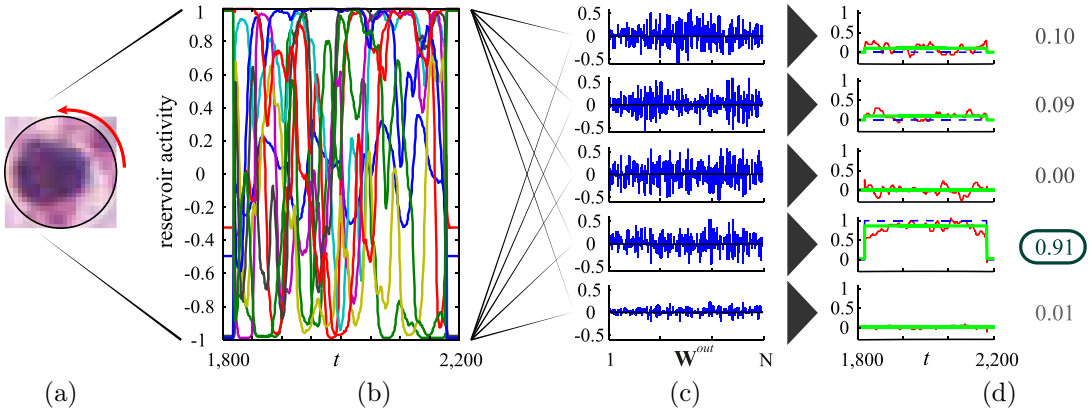


Figure 7.2.: The proposed rotation-invariant multi-class ESN training scheme. Counter-clockwise rotations of a cell patch (a) cause reservoir activity (i.e. feature computation) over time (b). (c) For each class, a set of readout weights is learned. (d) The readout unit with the highest mean output (green curve) over the image presentation time finally determines the class. The blue dashed curve is the binary target function, which is set to one for the correct class and zero everywhere else. The red curve is the actual network output.

Kainz P, Burgsteiner H, Asslaber M, Ahammer H. Robust Bone Barrow Cell Discrimination by Rotation-Invariant Training of Multi-Class Echo State Networks. In: Iliadis L, Jayne C, editors. Engineering Applications of Neural Networks - EANN 2015. vol. 517 of Communications in Computer and Information Science. Rhodes, Greece: Springer International Publishing; 2015. p. 390–400. With permission of Springer.

ples representing independent rotations of the cell without taking into account the relations among consecutive rotations. Motivated by how RNNs can capture appearance information in temporal features, we propose a rotation-invariant learning scheme for cell classification using pure ESNs, cf. Fig. 7.2.

In the following chapters it will be shown that it is possible to train an ESN with standard ridge regression directly on raw image data in a way such that its classification accuracy is independent of the rotation of the cell. While previous work heavily relied on explicit feature extraction from segmented cells, nuclei or cytoplasm, this approach does not include such steps and can be applied to single-cell image patches directly. Hence, a dedicated segmentation step of cell nuclei and cytoplasm, can be omitted, which in fact is not always possible in our cell samples, cf. Fig. 7.1. This work is based upon results from earlier work [9, 10], where a similar, but only binary cell recognition problem was considered. We already explored the extension of this approach to a multi-class problem in Kainz *et al.* [11]. In addition to [11], this contribution is extended by providing a direct comparison with an RF [165] image classifier that was trained the conventional way to achieve rotation-invariance. An RF is an ensemble of de-correlated, binary

decision trees, that are individually trained and produce a consolidated prediction, cf. Section 4.3.1. In machine learning and computer vision, forests are well known for their efficiency and good generalization ability without the tendency to overfit the data [164, 179]. Further, they perform well on unbalanced and small datasets, have become a key ingredient in patch-based medical image analysis [167, 229–231], and have recently shown state-of-the-art performance on similar bone marrow data [8]. Further, we will discuss strengths and weaknesses of both approaches.

In the subsequent sections we are going to elaborate on the proposed novel rotation-invariant training scheme for ESNs, provide the outline of the experimental setups and present the results. The findings are discussed and an outlook to future work will be given in the final chapter of this part.

7.2. Echo State Networks for Cell Recognition

An ESN-approach to rotation-invariant blood cell maturation recognition in the human bone marrow is proposed here. An overview of the classification scheme starting

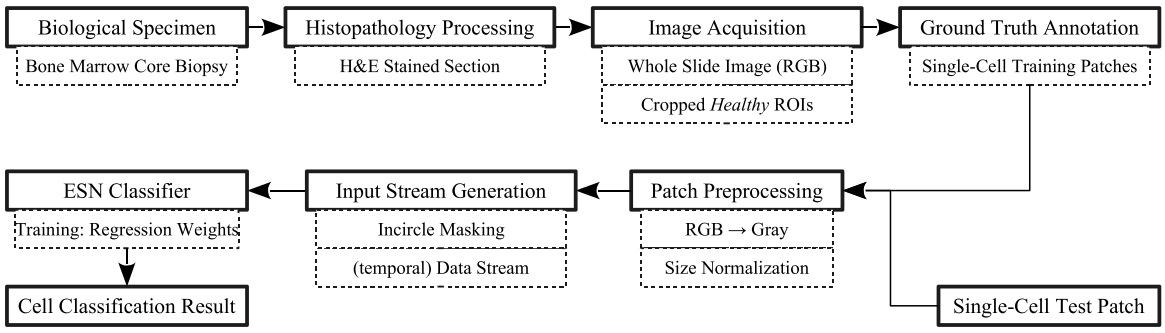


Figure 7.3.: Overview of the proposed cell recognition pipeline using ESNs. Biological tissue specimen is prepared in the histopathology laboratory according to standard protocols for H&E staining. The sections are digitized to RGB whole slide images, and regions of interest (ROIs) containing healthy tissue are cropped. Typical bone marrow cells of four classes are labeled as ground truth by an expert pathologist. Single-cell RGB patches are manually extracted as part of the ground truth labeling, converted to gray-scale (RGB mean), and rescaled to a predefined size. A temporal input data stream is generated by rotating the incircle-masked static images, which is then fed into the ESN classifier, cf. Fig. 7.2. During training, ridge regression determines a set of weights that can be used to predict the class label of a test patch using a threshold-based inference scheme.

with the biological sample is illustrated in Fig. 7.3. Our supervised learning system is trained using local image patches, which are first labeled by an experienced pathologist as one of multiple foreground classes, or background. Please note that automatic cell detection is omitted here, and we will focus on classifying manually cropped image patches in this proof-of-concept study. Nevertheless, we eventually discuss the promising cell localization method proposed in Chapter 3 that is going to be applied when considering this classification approach in an integrated system.

7.2.1. Multi-Class Echo State Networks

Echo State Networks are a way to train RNNs for temporal prediction tasks [226]. Many different ESN architectures have been proposed [232], but in this work we focus on the classical architecture proposed by Jäger [226], and adapt it to solve a multi-class classification task. The reservoir, a randomly connected RNN composed of N units, models short-term memory and non-linear input expansion. Reservoirs in ESNs have to ensure the 'echo state' property to be an universal function approximator. Hence, the recurrent reservoir weights $\mathbf{W} \in \mathbb{R}^{N \times N}$ must be scaled, such that the spectral radius $\rho(\mathbf{W}) < 1$ [226]. In practice, the spectral radius is a global control parameter that defines how fast the reservoir activity vanishes [233]. Hence, a larger spectral radius results in slower decay and longer self-interaction of reservoir activity.

Fig. 7.4 illustrates the architecture of a multi-class ESN. The L-dimensional input at a particular point in time t, given by $\mathbf{u}(t) = [u_1(t), u_2(t), \ldots, u_L(t)]^{\mathrm{T}}$, and a bias unit are connected to the reservoir units via the input weights $\mathbf{W}^{in} \in \mathbb{R}^{N \times (1+L)}$. When an input is presented to the input layer, it causes non-linear activity in the reservoir. This activity represents the (temporal) features, which the recurrent reservoir units compute from the input stimulus at each observable time step. The weights $\mathbf{W}^{in}$ and $\mathbf{W}$ may be sparse and remain fixed after random initialization and meeting task-specific scaling criteria [234].

A binary classification task can already be performed by a single readout unit. Given a training set of one positive class and one negative class, the unit is trained to recognize samples belonging to the positive class and ignore the negative samples. This simple scheme can easily be extended to solve multi-class problems: each class $C_k, k \in \{1, \ldots, K\}$ is represented by a single readout unit, which is trained in the *one-versus-all* scheme. Hence, the ESN learns a classifier $f : \mathcal{X} \to \mathcal{Y}$, with $\mathcal{X} = \mathbb{R}^L$ and $\mathcal{Y} = \{C_k\}$.

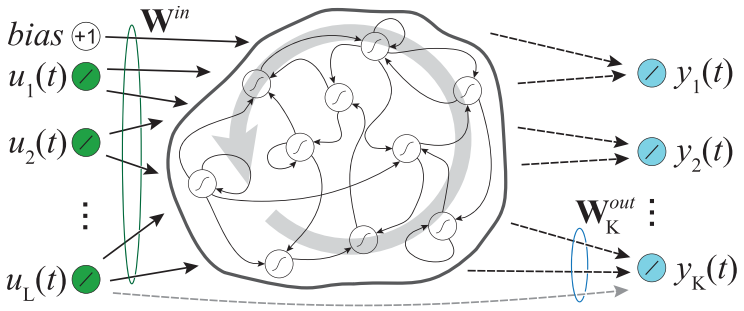


Figure 7.4.: Multi-class ESN architecture. At each time step t, L linear input units (green) feed input $\mathbf{u}(t)$ into the reservoir via input weights $\mathbf{W}^{in}$. Each of the K linear readout units (blue) corresponds to a specific class. The reservoir consists of N (internal) units with hyperbolic tangent (*tanh*) activation function. Readout-to-reservoir feedback connections are omitted in our architecture. The input layer is fully connected to each readout unit, symbolically illustrated for one unit by the grey dashed arrow at the bottom. This provides contextual information on the original input in parallel to the temporal features. After learning readout weights $\mathbf{W}^{out}_k$, the output $y_k(t)$ is determined for the readout units.
Kainz P, Burgsteiner H, Asslaber M, Ahammer H. Robust Bone Barrow Cell Discrimination by Rotation-Invariant Training of Multi-Class Echo State Networks. In: Iliadis L, Jayne C, editors. Engineering Applications of Neural Networks - EANN 2015. vol. 517 of Communications in Computer and Information Science. Rhodes, Greece: Springer International Publishing; 2015. p. 390–400. With permission of Springer.

When stimulating the reservoir, the state update equation at time step $t+1$ is given as

$$\mathbf{x}(t+1) = (1-\alpha)\mathbf{x}(t) + \alpha \tanh(\mathbf{W}^{in}[1;\mathbf{u}(t)] + \mathbf{W}\mathbf{x}(t)), \tag{7.1}$$

where $\mathbf{x}(t)$ denotes the state vector at the previous time step t. The leaking rate α defines the short-term memory capacity, i.e. how strong the reservoir activity at time step t influences activity at $t+1$. The ESN can be set to a generative mode, where the input $\mathbf{u}(t)$ is switched off. We use the term *generative mode* here to avoid confusion with the very similar *pattern generator* principle [233]. The difference is that we do not use output-to-reservoir feedback or apply the output at t as input at $t+1$, but compute the state updates purely from the remaining activity within the reservoir using

$$\mathbf{x}(t+1) = (1-\alpha)\mathbf{x}(t) + \alpha \tanh(\mathbf{W}\mathbf{x}(t)). \tag{7.2}$$

During a recording time Ψ, input and state vectors are concatenated in a large state matrix $\mathbf{X} \in \mathbb{R}^{(1+L+N)\times\Psi}$. Using a single reservoir facilitates automatically capturing the activities caused by multiple classes. The individual readout weights $\mathbf{W}^{out}_k \in \mathbb{R}^{(1+L+N)}$

for C_k are learned via ridge regression with Tikhonov regularization:

$$\mathbf{W}_k^{out} = \mathbf{Y}_k \mathbf{X}^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}} \mathbf{X} + \beta \mathbf{I})^{-1}, \tag{7.3}$$

where $\mathbf{Y}_k \in \mathbb{R}^{\Psi}$ denotes the desired target function of class C_k, $\mathbf{X}^{\mathrm{T}}$ the transpose of the state matrix, and $\mathbf{I}$ the identity matrix. The Tikhonov regularization coefficient is fixed to $\beta = 10^{-2}$ [233].

A piece-wise constant target function is regressed for each class C_k at a recorded time step t, which is given by

$$y_k(t) = \begin{cases} 1 & \text{if } \mathbf{u}(t) \in C_k \\ 0 & \text{otherwise} \end{cases}. \tag{7.4}$$

Each readout unit produces an output $\hat{\mathbf{Y}}_k = \mathbf{W}_k^{out} \mathbf{X}$, and a score over a predefined inference period Υ, computed as the mean output:

$$\bar{y}_k = \frac{1}{|\Upsilon|} \sum_{t=-\Upsilon/2}^{\Upsilon/2} \hat{y}_k(t). \tag{7.5}$$

The unit with the maximal mean score subsequently determines the winner class:

$$\hat{k} = \arg\max_k \{\bar{y}_k\}. \tag{7.6}$$

7.2.2. Rotation-Invariant Cell Classification

Given an image patch $I(\mathbf{c})$ centered on an individual cell at an image location $\mathbf{c} = (c_x, c_y)$, we need to transform it into time-dependent input for the ESN classifier. In order to generate temporal input from $I(\mathbf{c})$, we take advantage of the fact that cells can occur in arbitrary rotations within tissue. Fig. 7.5 illustrates the generation of a temporal input stream Θ_i for a cell by concatenating subsequent rotations of the patch $I(\mathbf{c}, \varphi)$. While rotating by an angle φ, we ignore the patch corners and just consider the pixels within the incircle radius r. For this purpose, a receptive field $\mathbf{V}$ of radius r is defined as the input layer and forwards the pixel intensities into the reservoir. All patches are required to be normalized to a fixed size of $2r \times 2r$ beforehand.

Rotation-invariance of the classifier is achieved by letting the reservoir generate covariant features for each $I(\mathbf{c}, \varphi)$, $\varphi = 0, \ldots, 359°$, starting at an arbitrary angle φ_0 that

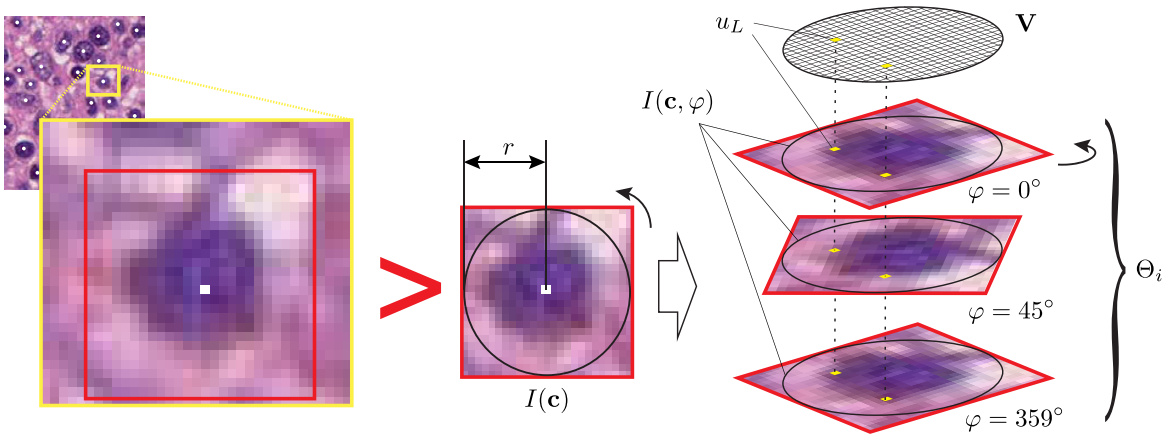


Figure 7.5.: Illustration of the temporal input stream generation for the rotation-invariant learning scheme using ESN. A patch containing a single cell is extracted from a virtual slide. It is then normalized to a predefined size $2r \times 2r$ to fit a receptive field $\mathbf{V}$. A static image patch $I(\mathbf{c})$ is transformed into a stream Θ_i by concatenating subsequent rotations $I(\mathbf{c}, \varphi)$. For each rotation, $\mathbf{V}$ forwards the pixel intensity within the incircle of $I(\mathbf{c}, \varphi)$ into the reservoir.

Kainz P, Burgsteiner H, Asslaber M, Ahammer H. Robust Bone Barrow Cell Discrimination by Rotation-Invariant Training of Multi-Class Echo State Networks. In: Iliadis L, Jayne C, editors. Engineering Applications of Neural Networks - EANN 2015. vol. 517 of Communications in Computer and Information Science. Rhodes, Greece: Springer International Publishing; 2015. p. 390–400. With permission of Springer.

relates to a cell's arbitrary orientation in a slide. These reservoir states are harvested by evaluating Eq. (7.1) and the target function is approximated at each recorded time step. After the network saw all $I(\mathbf{c}, \varphi)$, the final class is determined using Eq. (7.6).

The generative mode of ESNs also enables skipping $\Delta\varphi$ rotations after receiving external input $I(\mathbf{c}, \varphi)$. Due to the memory and decaying reservoir activity we are still able to obtain discriminative features using Eq. (7.2), even without external input driving the reservoir. The reservoir activity usually approaches a resting state without external input and thus a properly selected $\rho(\mathbf{W})$ ensures that there is enough activity left before the next input $I(\mathbf{c}, \varphi + \Delta\varphi)$ is presented.

In general, two kinds of memory can be observed in an ESN. Firstly, memory that can be controlled by the leaking rate parameter α. We name it the 'state' memory in this context. Secondly, the 'immutable' memory, which is inherently modeled by the recurrent weights $\mathbf{W}$ in the reservoir. The state memory models the influence of previously computed reservoir states on the current state and therefore controls smoothing of the temporal features. If $\alpha \ll 1$, the internal states caused by previous rotations may significantly influence subsequent ones and may cause over-smoothed

states. On the other hand, if $\alpha = 1$, the state memory is turned off and features for each observed rotation $I(\mathbf{c}, \varphi)$ are less influenced by previous states. However, setting $\alpha = 1$ does not entirely turn off the memory capacity of the reservoir, since the recurrent connections defined by the internal weights $\mathbf{W}$ are not affected. To find a suitable amount of state smoothing, α needs to be chosen accordingly.

8. Experimental Setup and Implementation Details

8.1. Bone Marrow Cell Dataset

We challenge our approach on a non-neoplastic human bone marrow cell dataset composed of three consecutive maturation stages in granulopoiesis as well as one class from erythropoiesis, cf. Table 8.1. Myelocytes, metamyelocytes and band cells are three consecutive maturation stages of WBC in the bone marrow, and are characterized by a high intra-class variability and a small inter-class distance. Biological samples were taken from the human iliac crest by trephine biopsy, embedded in acrylate, cut into slices of $\approx$ 1-2 µm thickness, and stained with H&E. Cell patches were extracted from virtual slides of two patients (digitized at 40$\times$ magnification using an Aperio whole slide scanner) and labeled by an expert pathologist. All cells appeared at the same object scale. Considering the problem of hematopoietic cell classification without megakaryocytes, the object size within a class does not vary by more than approximately 20%, which still can be compensated by the proposed approach. The total number of original patches was extended by a factor of six using non-linear warping transformations (circular distortion by $\pm 35°$) as well as horizontal flipping, resulting in 744 foreground patches of four classes. Despite this data augmentation strategy, we ensured that both transformed and original images were unique, and that rotation of the extended dataset did not introduce any duplicates. In addition, we used 200 randomly sampled background patches from the same virtual slides as control class. With respect to all positive classes (i.e. foreground classes), the control class served as negative class. Hence, samples of this class did not contain any centered cells. It was used to verify the capability of a classifier to discriminate among different positive classes, as well as between all positive

	Group	Name	n_c	Sample Patches
C_1	G	Band Cell	200	
C_2	G	Metamyelocyte	144	
C_3	G	Myelocyte	200	
C_4	E	Orthochromatic Normoblast	200	
C_{bg}	–	Background	200	
		Total	**944**	

Table 8.1.: The non-neoplastic bone marrow dataset used for method evaluations consists of a total of $n = 944$ patches. Group G denotes cell classes belonging to granulopoietic maturation, while E refers to erythropoietic maturation.

classes and the negative class. All patches were converted to grey scale by averaging the color channels (RGB mean). At 40× magnification, the average single-cell patch size in our dataset was 33×33 pixels. Hence, we could normalize the patches to a fixed size of 20×20 pixels using bilinear interpolation without losing significant appearance information or introducing artifacts. The receptive field of the ESN was connected to all incircle ($r = 10$) pixels of that patch, resulting in $L = 332$ network inputs. In total, the dataset comprised $n = 944$ patches, of which 66% ($n = 623$) were assigned for training and hyper-parameter optimization, and 34% ($n = 321$) were held-out for testing.

8.2. Echo State Network

In order to avoid learning the sequence of cell classes rather than the appearance of each cell, we introduced periods with zero-input of random length between two consecutive cell image patches. They were furthermore required to let the reservoir 'forget' about the previous image and learn each instance separately. Depending on the spectral radius $\rho(\mathbf{W})$, the network approximately required 20 to 50 time steps to reach the resting state after the last input stimulus has been presented. Therefore, we randomly sampled the zero-input length in the range $[50, 100]$. Please note that this is a different concept than setting the network to the generative mode after presenting an image input, because we did not apply another input stimulus as long as it did not completely reach the

resting state.

Starting with ESN hyper-parameters that were determined by simulated annealing [235] for a binary classification task [9], we continued with manual fine-tuning in this multi-class problem. Considering the influence of the hyper-parameters to achieve higher performance [233], they were successively optimized for our cell recognition task using 10-fold cross-validation (CV) on the training set ($n = 623$). Dense connectivity was used in the input weights $\mathbf{W}^{in}$, while only 30% of the reservoir connections $\mathbf{W}$ were non-zero. The normalized pixel intensity was bounded within $[0, 1]$, so we shifted and scaled it to $[-1, 1]$ to avoid using only the linear part of the *tanh* activation function of the reservoir units. For a presented input $I(\mathbf{c})$, all readout units showed rather high activity in the first and last few recording steps. Since these reservoir activities did not contribute to the actual classification, we bounded the inference window Υ to start 5% after the first and to end 5% before the last sample.

8.3. Random Forest

The performance of the ESN was compared to a classification Random Forest [165]. A standard RF[1] was employed to solve the same five-class classification problem that was previously defined for the ESN. In analogy to the ESN setup, the input data for the RF were the grey-value single-cell patches. Training a decision tree T_t in a forest is based on the principle that the dataset arriving at an internal node gets split into a left and right subset based on randomly selected criteria. The Gini index [178] reflects the 'purity' over the five classes in a dataset $\mathcal{P}$:

$$Gini(\mathcal{P}) = \sum_{k \in \{1,\ldots,4,bg\}} p(c = k|\mathcal{P})\, p(c \neq k|\mathcal{P}). \tag{8.1}$$

The optimization problem at each node is given by maximizing the Gini gain objective function to find the best split decision ϕ^*. Accordingly, evaluating Eqs. (4.7) and (4.8) resulted in the optimal parameters. Suitable hyper-parameters such as number of random node tests, number of samples to test a split and split functions, were evaluated in 10-fold CV experiments. A single, yet effective split function that randomly selected two locations in the image patch and compared the intensity difference to a randomly selected threshold (pixel value difference) was employed. As described in Section 3.2.2,

[1] See Section 4.3.1 for details on the formulation of a binary classification RF.

a common rule defines $\lfloor\sqrt{p}\rfloor$ random node tests for each split [164], where p is the number of features per sample. For 20×20 pixels images, each pixel being a feature, this results in only 20 node tests ($p = 400$). However, we found that increasing the number of split function tests per node to 100, and comparing each one to 20 random thresholds (i.e. 2000 tests per node) resulted in superior performance with negligible runtime prolongations. Each split decision was evaluated on a subset of 200 samples that were randomly selected from the data available at a node. A terminal (leaf) node was constructed when either the maximum tree depth T_{md} was reached or the size of the dataset arriving at a node was smaller than a predefined number of 20 samples. Leaf nodes store the class label histogram of the data. Bagging did not result in better performance during forest training. Hence, each tree was trained on all available training data samples. Once the trees were constructed, we could propagate a single-cell image patch $I(\mathbf{c})$ through the forest. First, individual trees were evaluated and their results consolidated:

$$\bar{p}(c = k|I(\mathbf{c})) = \frac{1}{T}\sum_{t=1}^{T} p_t(c = k|I(\mathbf{c})), \tag{8.2}$$

where $p_t(c = k|I(\mathbf{c}))$ denotes the probability for class k predicted by the t-th tree. Consequently, the most likely class label results in the final prediction:

$$\hat{k} = \underset{k\in\{1,\ldots,4,bg\}}{\arg\max}\ \bar{p}(c = k|I(\mathbf{c})). \tag{8.3}$$

8.4. Classification Performance Metrics

Both ESN and RF classification performance were assessed quantitatively. The overall performance is reported as mean accuracy (ACC) weighted by the class distribution:

$$\mathrm{ACC} = \frac{1}{n}\sum_{k} n_k\,\mathrm{TP}_k, \tag{8.4}$$

with n as the total number of samples and n_k as the number of samples in class C_k. TP_k denotes true positive, FP_k false positive, TN_k true negative, and FN_k false negative predictions for class C_k. In order to observe the performance at the class level, more detailed measures than the overall accuracy were required. Class-wise performance is

reported as precision (PRC_k), recall (REC_k), specificity (SPC_k), and F1-score ($\mathrm{F1}_k$):

$$\mathrm{PRC}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k}, \tag{8.5}$$

$$\mathrm{REC}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}, \tag{8.6}$$

$$\mathrm{SPC}_k = \frac{\mathrm{TN}_k}{\mathrm{TN}_k + \mathrm{FP}_k}, \tag{8.7}$$

$$\mathrm{F1}_k = \frac{2 \cdot \mathrm{PRC}_k \cdot \mathrm{REC}_k}{\mathrm{PRC}_k + \mathrm{REC}_k}. \tag{8.8}$$

Actual values of performance measures are reported as mean and standard deviation (SD).

8.5. Experiment Definitions

We were interested in the performance of the proposed rotation-invariant approach in terms of overall and class-wise accuracy under different conditions. Therefore, we defined the following four experimental settings, where for each individual cell image 360 reservoir states were collected. We used bilinear interpolation when rotating the images to reduce artifacts caused by aliasing.

1. *Experiment I*
   In the first experiment we were interested in the general applicability of the proposed approach to multi-class cell classification. The ESN was trained on a sequence of all possible (integer) rotation angles $\varphi = 0, \ldots, 359°$ of an individual image patch. Hence, no generative mode was used ($\Delta\varphi = 0$).

2. *Experiment II*
   In the second experiment the generative mode of the ESN in the context of image classification was evaluated. In particular, we focused on whether or not meaningful features could be computed when we inserted a predefined stimulus-free period ('zero-input') between showing single rotations of the image. Instead of having 360 individual inputs from an image, we skipped five subsequent rotation angles ($\Delta\varphi = 5$) and recorded the decaying reservoir activity. For instance, the input stimulus sequence for the first 12 time steps included three actual image inputs

at the rotation angles 0°, 6°, and 12°, respectively: $[\mathbf{u}(t_0) = I(\mathbf{c}, 0),\ \mathbf{u}(t_{1,\ldots,5}) = 0,\ \mathbf{u}(t_6) = I(\mathbf{c}, 6),\ \mathbf{u}(t_{7,\ldots,11}) = 0,\ \mathbf{u}(t_{12}) = I(\mathbf{c}, 12),\ \ldots]$.

3. *Experiment III*
   Here, we increased the duration of the generative mode and skipped 10 rotation angles ($\Delta\varphi = 10$).

4. *Experiment IV*
   To simulate a concrete real-world application, the classifier is learned from scratch on all available training data ($n = 623$) and tested on the held-out test data ($n = 321$). Further, we used $\Delta\varphi = 5$ (see also experiment II) to examine whether the ESN is able to deal with short periodical, but different input stimuli, and compared it to the RF trained the conventional way to achieve rotation-invariance. A classifier is considered to be invariant to rotations, if the very same object is always labeled with the same class label under arbitrary in-plane rotations. Hence we examined, whether both ESN and RF were able to recognize the very same cell again under a randomly selected, different rotation angle. The number of samples per class in the test set was $n_1 = 69$, $n_2 = 56$, $n_3 = 61$, $n_4 = 65$, and $n_{bg} = 70$, which also approximately reflected the distribution in the training set. We did not perform any additional data augmentation to balance the training set.

Using the settings from experiments I-III, we examined the influence of the generative mode on the classification performance of the ESN in 10-fold CV experiments on the training set. The size of the reservoir was varied ($N = \{200, 500, 1000, 2000, 3000, 4000\}$) to assess the required memory capacity of the ESN for the tasks. For each N, the reported values correspond to the best results of the hyper-parameter fine-tuning, which was stopped once the performance reached saturation on our dataset. A similar approach was carried out for the RF, where we varied the number of individual trees and their depth along other hyper-parameters determined by the preceding search. The samples in the dataset were randomly shuffled at the beginning of the experiments. Despite that the main focus of this work was set on evaluating the rotation-invariant ESN classifier, we employed the previously described RF as baseline classifiers to validate the results.

9. Cell Classification Results

9.1. Model Evaluations

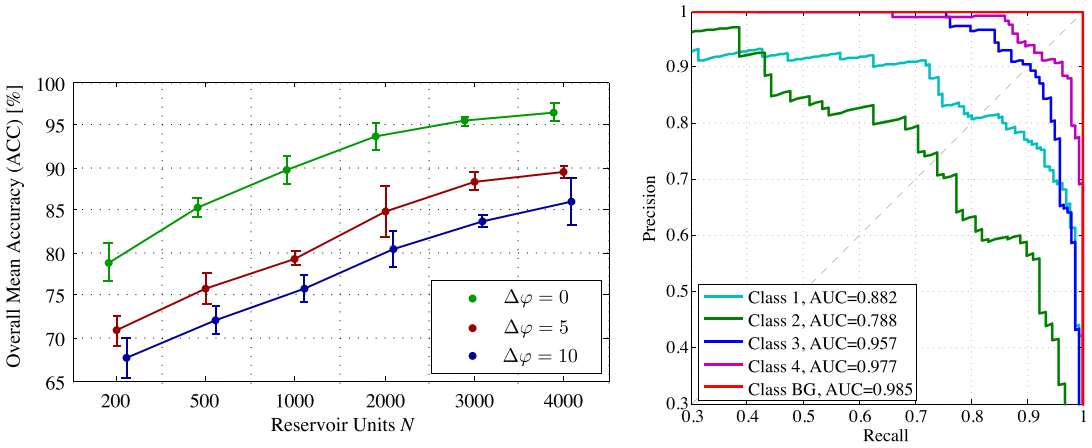
9.1.1. Echo State Network

The classification performance of the proposed rotation-invariant approach was evaluated in terms of weighted mean overall accuracy (ACC). Five independent 10-fold CVs were run in experimental settings I-III, cf. Section 8.5. In order to fix a suitable amount of short-term memory, prevent over-smoothing the state space, and account for proper reservoir decay, the following parameter tuples $(\Delta\varphi, \rho(\mathbf{W}), \alpha)$ were used for the ESN experiments: $(0, 0.6, 0.85)$, $(5, 0.8, 0.85)$, and $(10, 0.95, 0.85)$. We collected 360 reservoir states for each individual cell image patch, starting at rotation $\varphi_0 = 0$.

			Reservoir Size N					
Exp.	$\Delta\varphi$		200	500	1000	2000	3000	4000
I	0	ACC	78.85	85.32	89.76	93.63	95.37	96.43
		SD	(0.75)	(0.37)	(0.55)	(0.51)	(0.17)	(0.35)
II	5	ACC	70.82	75.85	79.37	84.73	88.39	89.49
		SD	(0.59)	(0.63)	(0.27)	(1.01)	(0.35)	(0.22)
III	10	ACC	67.72	72.08	75.82	80.42	83.64	85.96
		SD	(0.77)	(0.53)	(0.54)	(0.70)	(0.22)	(0.92)

Table 9.1.: Results of the ESN model evaluations. Five independent 10-fold cross-validation experiments were run on the training dataset. Values are reported in percent as mean weighted accuracy (ACC) and standard deviation (SD), cf. Fig. 9.1 (a).

Quantitative results of the CVs are reported in Table 9.1 and visualized in Fig. 9.1 (a). Obtaining higher performance generally required larger reservoirs. For instance, to reach approximately 80% accuracy, it took ten times more reservoir units to get similar



(a) (b)

Figure 9.1.: Performance evaluation of the ESN model. (a) Mean weighted accuracy (ACC) over five independent 10-fold CV experiments, error bars refer to three standard deviations (SD). The longer the ESN is in generative mode, the more reservoir units are required to generate sufficient features for classification via linear regression. (b) Precision-recall curve for one ESN cross-validation experiment with $N = 1000$ and $\Delta\varphi = 0$. The area under the curve (AUC) for C_1 and C_2 is lower than for the other classes. With larger reservoirs even these two classes could obtain higher performance.

performance in experiment III compared to experiment I: ($N = 2000$, $\Delta\varphi = 10$) versus ($N = 200$, $\Delta\varphi = 0$), cf. Table 9.1.

Discriminating subsequent maturation stages of granulopoietic cells (i.e. C_1, C_2, and C_3) was challenging for the ESN. Taking a closer look at a single CV run from experiment I with $N = 1000$ and $\Delta\varphi = 0$, we observed that this task required larger reservoirs to capture the subtle differences in the cells' appearance. Considering the area under the curve (AUC) in Fig. 9.1 (b), learning to recognize band cells (C_1) and metamyelocytes (C_2) in this setting seems to be harder than learning the other classes. The confusion among these classes may be caused by their indistinct class borders, because we did not observe this effect among C_3 and C_4. All samples of the background class C_{bg} could be recognized correctly.

9.1.2. Random Forest

Similarly, the dataset for the baseline classifier (RF) experiments was augmented with the same number of rotations and rotation angles that were available as inputs to the ESN (i.e. $360/\max\{\Delta\varphi, 1\}$). In a grid search, 16 combinations of two main RF

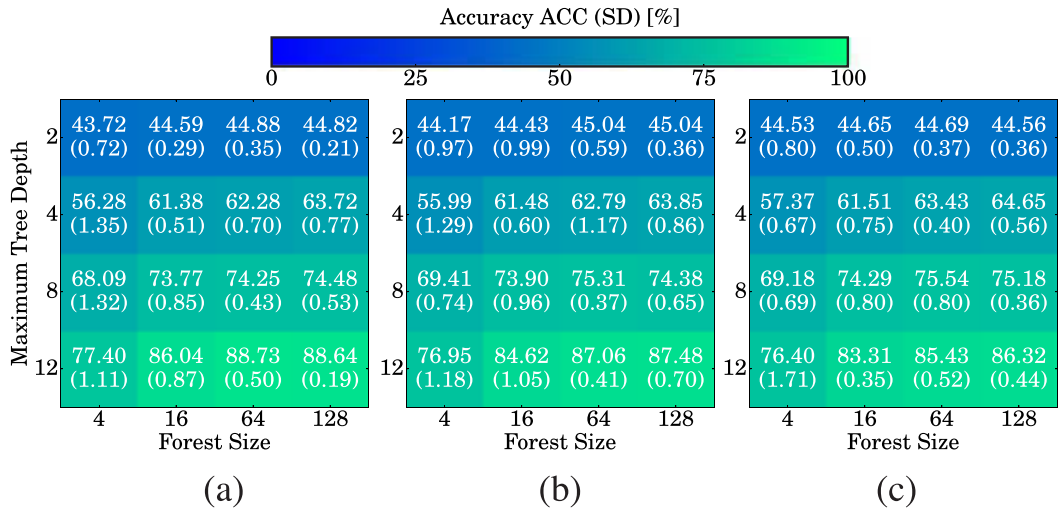


Figure 9.2.: Results of the experiments to select the best RF model. In a grid search, 16 combinations of two main RF parameters forest size (T), and maximum tree depth (T_{md}) were evaluated. Performance values for (a) $\Delta\varphi = 0$, (b) $\Delta\varphi = 5$ and (c) $\Delta\varphi = 10$, are reported in percent as weighted mean accuracy (ACC) and standard deviation (SD, in brackets). Higher overall mean accuracy corresponds to green color, where low accuracy is represented by blue color.

parameters were evaluated: the number of individual trees $T \in \{4, 16, 64, 128\}$, and the maximum tree depth $T_{md} \in \{2, 4, 8, 12\}$.

Like for the ESN, five independent 10-fold CVs were run on the training dataset using the RF. Fig. 9.2 illustrates the results. The classification accuracy could significantly be improved when larger and deeper forests were used. However, with respect to higher performance, the depth of the individual trees was more important than the total number of trees in the forest.

9.1.3. Classifier Comparison

Compared to the ESN, the RF was less sensitive to omitted rotation angles. The difference between maximum and minimum mean accuracy among experiment I-III was in the range of $0.1 - 3.3\%$ for the RF, and in the range of $10.5 - 14.0\%$ for the ESN, cf. Fig. 9.2. This may be explained by the fact that in the RF training, each rotation was considered as an individual sample, while the ESN received a continuous stream formed by the rotations of an image patch and omitting rotations causes unforeseen interruptions. These interruptions could only be compensated by larger reservoirs. Conversely, this also indicates that training a RF on each possible integer rotation ($\Delta\varphi = 0$)

is not necessary, and using much less training data already results in similarly high performance.

9.2. Robustness for Random Cell Orientations

We have shown in experiment I that generalization of an ESN works well when it is trained on all 360 rotations of an image patch. Further, it is also capable of learning a classifier that works with periods of zero-input between two consecutive rotations (experiments II-III). The CV experiments suggested that increasing the reservoir size also increases the classification performance on all classes, cf. Fig. 9.1. Using $\varphi_0 = 0$ and $\Delta\varphi = 5$ as reference scheme, we examined whether the very same cells could be recognized equally well when the rotation started at a random $\varphi_0 \neq 0$. Therefore, the best ESN and RF models were selected from the CVs with respect to experiment II

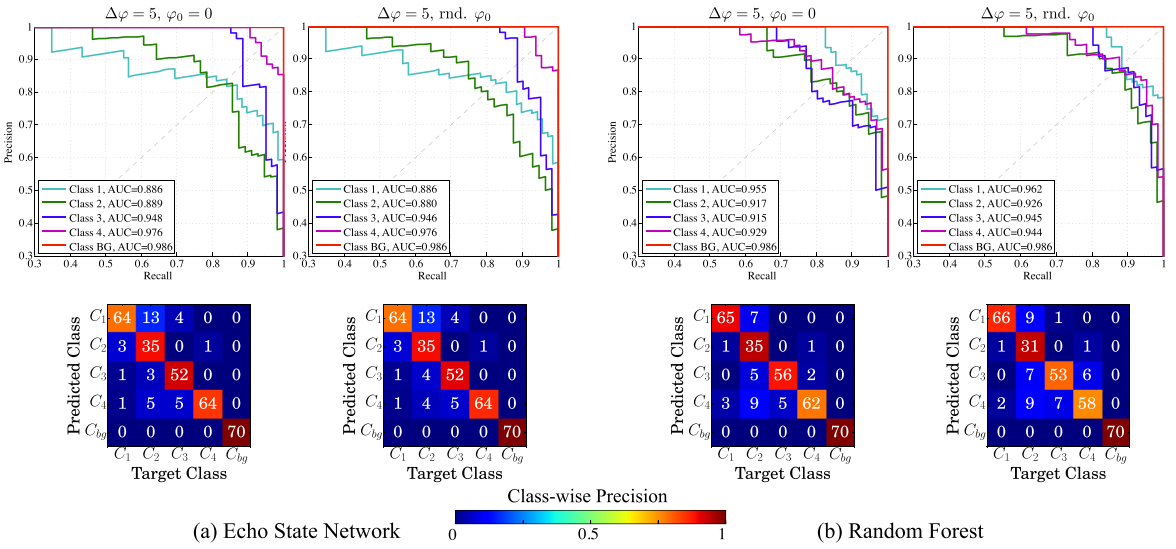


Figure 9.3.: Classifier performance on the test set in experiment IV ($n = 321$, $\Delta\varphi = 5$). The top row shows precision-recall curves and the area under the curve (AUC), the bottom row the confusion matrices. The first column in (a) illustrates the results for the ESN classifier trained on the original initial starting angles $\varphi_0 = 0$, while the second column refers to results obtained using a random φ_0. Similarly, in (b) the results are illustrated for the RF classifier. The color bar encodes the class-wise precision (i.e. the ratio of true positives per predicted class), colors towards red correspond to higher precision, i.e. less false positives. While the ESN frequently confused C_1 and C_2, the predominant misclassifications of the RF were among C_3 and C_4. However, the RF showed better results for C_1 and C_2.

(ESN: $N = 4000$, RF: $T = 128$, $T_{md} = 12$).

In Fig. 9.3 precision-recall curves, AUCs, and the confusion matrices on the fixed test set ($n = 321$) are reported for both classifiers. Generally, the results were of almost equal quality with respect to the weighted mean measures for a patch rotation starting at $\varphi_0 = 0$ and random φ_0. Considering precision, the ESN achieved 89.40% and 89.30%, the RF 90.50% and 87.60%, respectively. The recall of the ESN is constant 88.80%, but marginally decreases for the RF from 89.70% to 86.60%. While only two out of 321 cells (0.62%, Fig. 9.4 (q,r)) were predicted differently by the ESN, the RF predicted 16 cells (4.98%, Fig. 9.4 (a-p)) differently. Classification measures reported in Table 9.2 are more stable for the ESN than for the RF. More specifically, very stable precision and recall values indicate that the recognition accuracy does not necessarily depend on the initial rotation angle and that the proposed approach to train ESNs for image recognition works very well. The absolute value differences between $\varphi_0 = 0$ and random φ_0, denoted as $|\Delta|$, is close to zero for most of the measures. These results suggest that the ESN is able to robustly predict the same class label of a particular cell in 99.38% of all test cases, even if φ_0 randomly falls within $\Delta\varphi$, where the network is in the generative mode. However, the baseline classifier (RF) achieved comparable performance in almost all measures, cf. Table 9.2.

The precision-recall curves and confusion matrices were quite diverse for the foreground classes. While curves for C_3 and C_4 were close to optimal, curves for other classes showed that the network had troubles discriminating among C_1 and C_2. This could also be observed by inspecting the confusion matrices in Fig. 9.3 (a), where the ESN frequently predicted C_1 when the true class was C_2. On the other hand, the RF was better in recognizing C_1 and C_2, but showed a tendency to predict C_4 when the true class was C_2 or C_3, and C_3, when the true class was C_2 or C_4, cf. Fig. 9.3 (b). The area under the curve (AUC) over all classes was also more stable for the RF. The control class C_{bg} has always been perfectly classified by both ESN and RF.

		PRC			REC			SPC			F1		
		$\varphi_0 = 0$	rnd. φ_0	$\lvert\Delta\rvert$	$\varphi_0 = 0$	rnd. φ_0	$\lvert\Delta\rvert$	$\varphi_0 = 0$	rnd. φ_0	$\lvert\Delta\rvert$	$\varphi_0 = 0$	rnd. φ_0	$\lvert\Delta\rvert$
C_1	ESN	0.791	0.791	**0.000**	0.928	0.928	**0.000**	0.934	0.934	**0.000**	0.853	0.853	**0.000**
$n_1 = 69$	RF	**0.903**	**0.868**	0.035	**0.942**	**0.957**	0.015	**0.972**	**0.960**	0.012	**0.922**	**0.910**	0.012
C_2	ESN	0.897	0.897	**0.000**	0.625	**0.625**	**0.000**	0.985	0.985	0.000	0.737	**0.737**	**0.000**
$n_2 = 56$	RF	**0.946**	**0.939**	0.007	0.625	0.554	0.071	0.992	0.992	0.000	**0.753**	0.697	0.056
C_3	ESN	**0.929**	**0.912**	**0.017**	0.853	0.853	**0.000**	**0.985**	**0.985**	**0.000**	0.889	**0.881**	**0.008**
$n_3 = 61$	RF	0.889	0.803	0.086	**0.918**	**0.869**	0.049	0.973	0.950	0.023	**0.903**	0.835	0.068
C_4	ESN	**0.853**	**0.865**	**0.012**	**0.985**	**0.985**	**0.000**	**0.957**	**0.961**	**0.004**	**0.914**	**0.921**	**0.007**
$n_4 = 65$	RF	0.785	0.763	0.022	0.954	0.892	0.062	0.934	0.930	0.004	0.861	0.823	0.038
C_{bg}	ESN	1.000	1.000	0.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	1.000	0.000
$n_{bg} = 69$	RF	1.000	1.000	0.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	1.000	0.000
w.m.	ESN	0.894	**0.893**	**0.001**	0.888	**0.888**	**0.000**	0.971	**0.971**	**0.000**	0.884	**0.884**	**0.000**
$n = 321$	RF	**0.905**	0.876	0.029	**0.897**	0.866	0.031	**0.974**	0.966	0.008	**0.894**	0.860	0.034

Table 9.2.: Class-wise performance of the ESN and RF on the fixed test dataset ($n = 321$) in experiment IV. The last row contains the weighted mean (w.m.) of the measures according to the class distribution in the test set. The differences between the performance at the default ($\varphi = 0$) and the random starting angle (rnd. φ_0) are computed as absolute differences $|\Delta|$. The ESN recognized the same cells under different rotation angles more constantly. Superior results are printed in bold. The best hyper-parameters of the classifiers were chosen according to the best cross-validation results of experiment II (ESN: $N = 4000$, RF: $T = 128$, $T_{md} = 12$.)

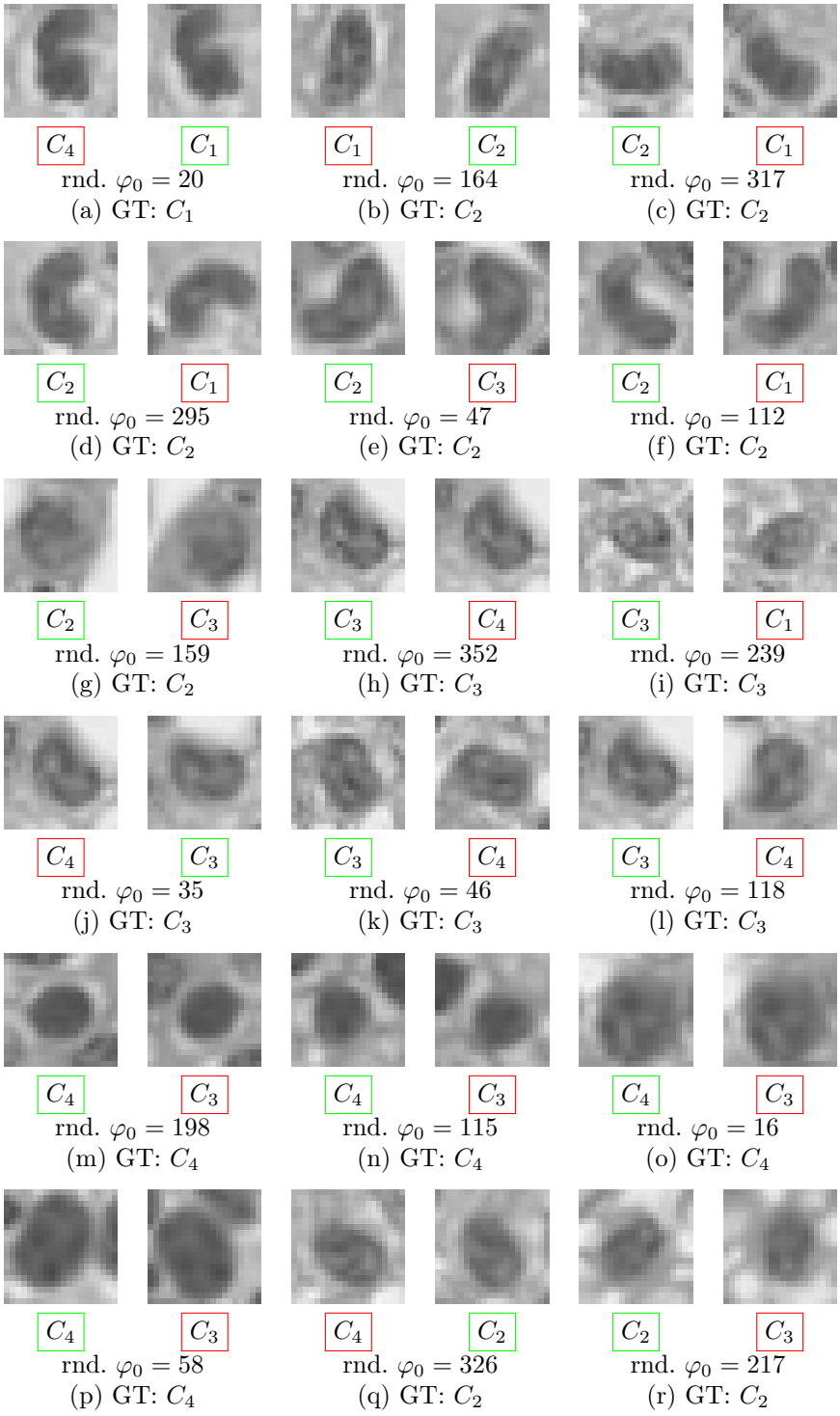


Figure 9.4.: The 16 samples (4.98%) classified differently by the RF classifier in experiment IV (a-p), and the two samples (0.62%) classified differently by the ESN (q,r). The left image shows the original cell patch at $\varphi_0 = 0$, the right image at a random starting angle (rnd. φ_0). Below these images, the predicted classes are enclosed in green boxes, when they were recognized as true positives, or in red boxes otherwise. The used counter-clockwise initial rotation angle and ground truth class (GT) is shown below the predictions.

10. Discussion and Conclusions - Cell Classification

In the previous chapters a novel cell classification approach was proposed and evaluated that considers subsequent rotations of a single-cell image patch as temporal signal to drive an ESN reservoir as feature generator. Finally, linear regression was used to learn a multi-class classification problem and the final class was determined by the *winner-takes-all* principle. The proposed approach was compared to a baseline RF classifier, achieving comparable or higher performance. However, the runtime of both algorithms was not compared in detail, since the proof-of-principle implementation of the proposed ESN approach was not yet optimized, while the already optimized RF implementation from Section 4.3.1 was used. A comparison to other image classifiers that work on raw image data such as SVM, or FF-NN, is considered in future research to validate the rotation-invariant classification scheme. Tuning these methods to be used as a competitive baseline was not feasible within the scope of this thesis, since they are much more sensitive to hyper-parameter tuning than RFs.

While previous work of Kainz *et al.* [9] focused on binary classification of similar cells from raw image patches, this thesis showed that the proposed approach robustly generalizes to multi-class problems as well. Further, performance comparable to that of an RF classifier was observed that was trained the conventional way to achieve rotation invariance, i.e. using multiple rotated versions of an image patch. The model evaluation revealed that learning the temporal features works better for showing all rotations ($\Delta\varphi = 0$) in a continuous stream than for skipping some rotations ($\Delta\varphi > 0$) when ESNs of the same complexity (in terms of reservoir capacity) were used. Yet, the ESNs were able to extract discriminative features in the generative mode, but significantly larger reservoirs were required to achieve similar classification performance.

We have considered short periods of zero-input between subsequent rotations of the same image to explicitly examine the capability of the reservoir to generate meaningful features from just a few external stimuli at specific points in a temporal input stream (experiments II-IV). However, this may not be an optimal setting for the ESN. Results from our model evaluations, especially with $\Delta\varphi = 0$ in experiment I (Section 9.1.1), suggest that applying continuous input, i.e. without interruptions of the stream, may potentially deliver higher performance for this cell recognition task. That could for instance be realized by providing non-zero input only, such as keeping the input between consecutive rotations constant. Future work must therefore evaluate a more economical way of using the proposed rotation-invariant approach, such as subsampling the input sequence at specific rotations, which would additionally decrease the required runtime. However, any change to the ESN input scheme likely requires an adaption of hyper-parameters, and may even lead to different architectures, which could not be examined anymore in this thesis.

Considering the current approach, the runtime for feature computation in the reservoir could potentially be reduced by resetting the reservoir state instead of waiting until the resting state has been reached after the last rotation of an image has been presented. A first obvious advantage would be that we could use much shorter random periods between the inputs, or omit them entirely. Further, this would enable parallel feature generation, since we could collect the temporal features from parallel copies of the reservoir and concatenate them (with, or without random gaps) before learning the regression weights off-line. Similar applies to testing, where inference could be done for multiple images in parallel. However, it needs to be investigated critically, whether omitting the sequences of random length between individual images is feasible, and how this (positively or negatively) influences the performance on this task.

Interestingly, we could only observe a minor improvement in the baseline RF classification results when more training data was used. Even when we skipped 10 rotations, the mean overall accuracy was approximately as high as when we trained on all 360 rotations. From that we can conclude that the performance of the RF can only be slightly improved when we train on ten times the original dataset size. We presume that due to interpolation the images within a range of ten subsequent rotation angles are too similar to increase the diversity in the training dataset and surpass the saturation of the classification performance.

We see advantages in using the proposed ESN approach for cell recognition: multiple classes can be learned from a single, randomly connected RNN, which is driven

by raw image data. When compared to other, gradient descent-based training methods [236, 237], training via ridge regression is guaranteed to result in a global optimum. Besides the proposed regression inference scheme, more sophisticated (also non-linear) schemes may lead to superior performance. Nevertheless, obtaining good hyper-parameters is highly task-specific and remains a tedious duty. CNNs [149] have demonstrated reasonable performance over SVM classifiers in white blood cell classification [137]. Despite the study of Habibzadeh *et al.* [137] used a very limited number of samples for their evaluation, we can consider them as a promising candidate in future research regarding the quantification of maturation stages in bone marrow. While CNNs also operate on raw images, they may be more robust in capturing the high intra-class variance while coping with small inter-class distance of blood cell maturation by learning significant features directly from the cell images. Nevertheless, robustly training supervised deep FF-NN usually requires huge image databases that are rarely publicly available for biomedical imaging problems. Our approach, on the other hand, works well even with a small number of samples, and skewed class distributions [9]. We will later (in Chapter 11) discuss current work towards creating a reliable bone marrow ground truth dataset for supervised learning that should finally enable employing Deep Learning models.

We consider a classifier as *truly* rotation-invariant, if the same object can be recognized as the true class under arbitrary rotations. However, a rectangular image is bounded by definition, and when it gets rotated while keeping the original image dimensions, some parts of the rotated image naturally become undefined. A common strategy to overcome this problem is to use border extension techniques [170], for instance filling these regions with uniform intensities, or mirroring border pixel intensities. Depending on the ratio of the image dimensions, this may introduce significant mirroring artifacts and artificial repetitive patterns. Since we used square patches and the cell nuclei were centered, the introduced border artifacts were just minimal. Moreover, the proposed rotation scheme for the ESN training just considers pixels within the incircle of the square patch. These border artifacts thus became negligible for both ESN and RF classifier. Nevertheless, the information in the original and rotated image is not completely equivalent, and therefore we can only speak of achieving an *approximate* rotation-invariance. It has to be noted that despite the classifiers may have misclassified some samples in either the default (i.e. $\Delta\varphi_0 = 0$) or the random starting angle evaluation (see definition of experiment IV), the other one always resulted in a true positive recognition, cf. Fig. 9.4. Hence, to increase their overall recognition rates and make a step towards more robust

rotation-invariant classifiers, it could be beneficial to classify a given test image several times under varying (e.g. random) angles and predict the final class label using some consolidation procedure.

The classification performance of the ESN on the presented bone marrow dataset has to be interpreted carefully, though. Firstly, due to the minimal inter-class distance that is caused by continuous maturation stages (i.e. C_1-C_3), the cells' appearance is frequently very similar and exacerbates finding a good (linearly) separating hyper-plane. Secondly, even after several years of experience, it is a non-trivial task for expert pathologists to make an exact distinction between consecutive maturation stages. The class distribution in this dataset might also slightly distort the results presented as weighted mean here, since C_2 is the minority – but most difficult – class to be recognized. This under-representation provokes a more optimistic view on our results, since the probability of misclassifying C_2 is lower due to the sample size. However, we provided per-class performance measures in our experimental results, cf. Fig. 9.1 (b). These results showed non-consecutive maturation stages (i.e. C_3 and C_4) being recognized reliably even by less complex ESN models. Using a background class as control enabled assessing the discriminative power of the classifiers with respect to the foreground classes. In comparison to the foreground classes, the background class always shows very high classification accuracy in both classifiers. This behavior increases the overall mean accuracy measures and must be considered when interpreting the results. The collection of hundreds of images for each cell class requires the time-consuming, manual annotation by hematopathology experts. By employing label-preserving data augmentation strategies that mimic morphological variability we were able to generate more samples for this study. Our current research is focused on creating a larger bone marrow dataset to assess the robustness and generalization capability of the ESN and omit artificial data augmentation. Additionally, a thorough evaluation on other, similar datasets is required to evaluate the transferability of the approach. Despite our promising results, an evaluation on a more extensive and fully balanced dataset, obtained from multiple patients is required to derive more precise conclusions.

Authors of previous work employed heterogeneous ensembles of classifiers [80, 223], where each individual instance focused on different aspects of the same feature space, or even different features. They reported superior results of their ensembles over individual classifiers. Our results in experiment IV revealed that the RF classifier has weaknesses, where the ESN actually shows strengths – and vice versa, cf. Fig. 9.3. A combination of the two evaluated classifiers, i.e. ESN and RF, to increase the overall recognition rates in

our experimental settings seems feasible. Deeper trees are expected to further increase the performance, but finding an optimum depth requires further examination. Using multiple ESNs in ensembles could be another opportunity to increase the performance by introducing classifier diversity. One could use different settings for the individual networks, such as the sparsity of input and reservoir weights, different reservoir sizes, input weight scaling, neuron models, etc. Furthermore, training on different levels of a Gaussian scale space pyramid could add robustness against scale variations to a certain extend, where the linear regression would still guarantee globally optimal learning.

Using ESNs to classify bone marrow cells is attractive for applications in biomedical diagnostics due to the reliability of the system. An important measure in medical application settings besides recall is a high specificity, as it is expressed by the proposed recognition system. Since it is a learning-based strategy, it can more easily be transferred to other problems than rigid standard image processing approaches. A big advantage of the proposed approach is that cell segmentation and explicit manual feature extraction is not required, once the locations of cell nuclei are determined. The focus of this thesis part was set on discriminating blood cell maturation stages in bone marrow and thus we omitted including an automatic procedure to localize cell candidates in the histopathology images. Some previously reported approaches treated cell localization as a subproblem of cell counting, but we believe that a separation of matters regarding cell localization and cell classification has more potential here. Results of recent work [8], also described and further refined in Section 5.1 of this thesis, presented state-of-the-art performance in bone marrow cell localization with the ability to tune the detector towards producing a huge set of candidate cells. Such a reliable and accurate cell nuclei detection strategy could easily be employed as a preceding step before applying the rotation-invariant classification scheme proposed in this thesis. Integrating these approaches into a fully automated system to support quantitative bone marrow diagnostics seems feasible, but must be left subject of future research at this point.

... or for short: the F.L.D.S.M.D.F.R.!

Flint Lockwood
Cloudy with a Chance of Meatballs

Part IV.

Data Quality Requirements, Discussion and Conclusions

11. Towards Extensive Reliable Ground Truth Data

Classification experiments involving supervised machine learning approaches require a valid ground truth to maximize the learning success of an algorithm. In the cell classification study described in Part III of this thesis, a rather small dataset was used to create the first proof-of-concept results. The distribution of bone marrow cell lineages could only be partly examined due to the lack of labeled samples for all morphologically distinguishable stages of maturation. Despite this dataset was created by a senior pathologist, the magnitude of neither intra- nor inter-observer variability could not be assessed. In fact, a reliable ground truth would involve multiple experts agreeing on a single class label such that typical representatives of the hematopoietic cells can be learned by the algorithms. Currently, neither labeled nor unlabeled hematopoietic cell datasets are available in the public domain that could be used for training cell recognition systems. The creation of a novel dataset comprising all stages of blood cell maturation would thus be an invaluable contribution to this field.

Several approaches exist for generating ground truth data for various purposes, which mostly depend on the required level of expert knowledge. The more generic a vision problem is, the more laypeople are in fact capable of contributing. In cases where one does not have to rely on any expert knowledge at all, e.g. labeling commonly known instances like pets, or outlining trees, bikes, facades in natural images, crowd-sourcing tools for image annotation like *LabelMe* [158, 238] can be used to efficiently obtain large ground truth datasets from citizen scientists. These tools have been used to generate computer vision research datasets such as scene understanding or facade segmentation [239, 240]. However, generating a ground truth dataset from histopathological images requires people with special training and experience, whereas delineating cars and roads in an urban scene image does not.

Due to the large variability of elements in biological tissue such as individual cells and larger structures formed by them, it is non-trivial to obtain a consensus among experts on a class label, or on an accurate delineation of structure boundaries. Ideally, a *single* label is the correct label and all observers assign this label independent of each other in a blinded fashion. This task is further exacerbated, since criteria for morphological discrimination in healthy tissue usually do not apply to neoplastic tissue. Eventually, all samples with a certain level of inter-observer agreement will be collected as valid and representative instances in the ground truth dataset. As a result of highly specialized research areas, domain experts are usually rare and may as well be located on different continents. Such factors aggravate communication and running inter-observer image analysis studies on a larger scale. Moreover, finding committed experts to contribute to a biomedical image dataset for research purposes is an organizational challenge. Mostly, experts in tissue analysis come from biology, histology and pathology, where their daily business may permit only a small time period for research activities. Thus, all tools for capturing their domain-specific knowledge must be solid and well-performing in order to maximize the likelihood for a successful labeling study. Amazon's mechanical turk (*MTurk*) has been used to create labels for image captioning and classification [241]. However, *MTurk* has not been designed to provide the capabilities for displaying, annotating, or navigating in whole slide images. In order to account for high usability, and since there was no application available to conduct blind inter-observer reliability studies, a novel web-application based on the Cytomine bioimage analysis platform [7] has been developed. In the following sections, this platform and preliminary results of a collaborative labeling study for hematopoietic cells will be described briefly.

11.1. The Cytomine-IRIS Labeling Platform

The Cytomine platform [7, 242] has been created following the demands of collaborative gigapixel image analysis over the internet. An orchestration of state-of-the-art internet technologies and distributed computing frameworks facilitates large-scale studies on extensive microscopic images and allows for integration of external applications via a public API[1]. Intentionally, the image annotation interface of Cytomine (Fig. 11.1) has been created to conduct collaborative studies in a most transparent way including review and proofreading tools. Of course, this contradicts our aim of having independent

[1] An extensive description of the Cytomine platform and its successful utilization in a broad set of projects can be found in the work of Marée *et al.* [7], and the Supplementary Material thereof.

expert opinions on individual annotations in order to create an unbiased ground truth cell dataset and being able to quantitatively measure inter-observer reliability. Hence, a novel labeling application was created leveraging the existing backend functionality: the Cytomine Inter-Observer Reliability Study Module, or Cytomine-IRIS for short. The source code of this application is publicly available under the open-source Apache 2.0 license from Github[2]. A deployed version of the application including tutorials and previews is available as part of the Cytomine demonstration servers[3].

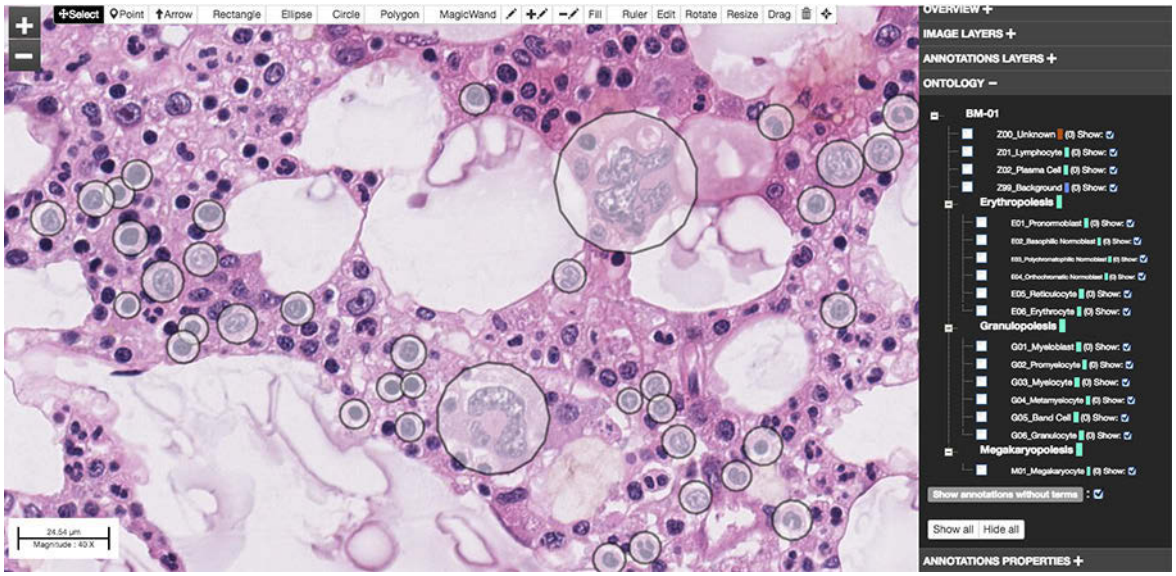


Figure 11.1.: The web interface of the Cytomine platform facilitates a collaborative annotation of gigapixel microscopy images. It comprises many functions such as proofreading and reviewing tools that enable a transparent collaboration, but does not provide options for inter-observer experiments. This interface is being used in an international study to label hematopoietic cells according to a predefined ontology (right panel).

We have launched an international study involving pathologists to assign labels to marked cells in whole slide images of H&E stained healthy bone marrow, cf. Fig. 11.1. We relied on the existing web interface (Fig. 11.1) to create the initial annotations in many images. Subsequently, the labeling interface of Cytomine-IRIS was used by the experts to effectively assign labels to the annotated cells by selecting them from a predefined ontology, cf. Fig. 11.2 (a). To enable the assessment of the cells in the context of high- and low-scale tissue architecture, panning and zooming facilitates proper exploration of the whole slide image.

[2] Available from https://github.com/Cytomine/Cytomine-IRIS.
[3] Available from http://demo-iris.cytomine.be/iris/.

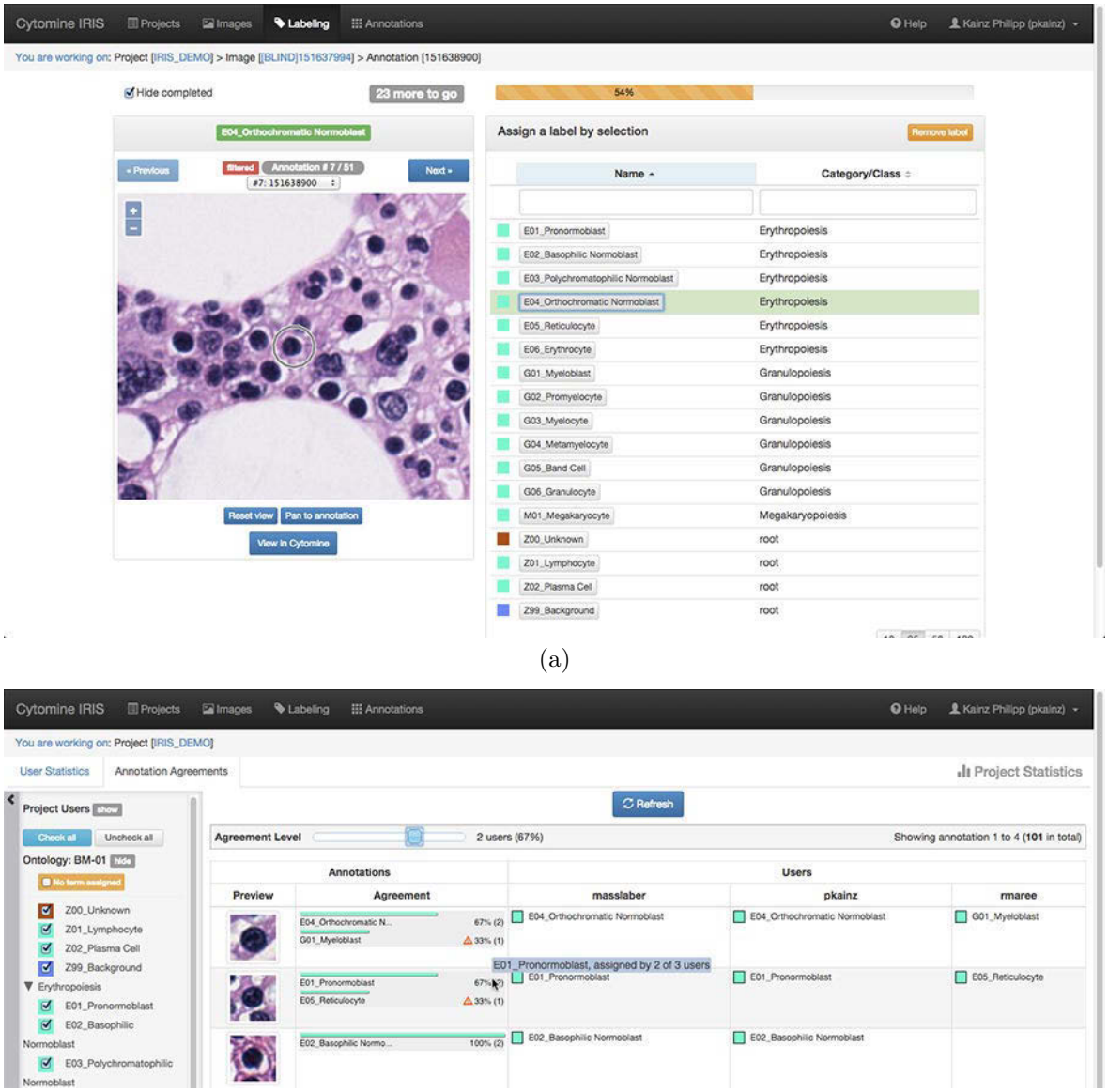


Figure 11.2.: Cytomine-IRIS labeling and annotation agreement statistics interface. (a) Observers can select terms from a predefined ontology and assign them to the pre-annotated object. (b) Designated project coordinators can view the per-annotation observer agreement. Here, the list is filtered to contain only annotations with at least 66% agreement on any assigned label among a set of selected observers.

Marée R, Rollus L, Stévens B, Hoyoux R, Louppe G, Vandaele R, et al. Collaborative analysis of multi-gigapixel imaging data using Cytomine. Bioinformatics. 2016 Jan;32(9):1395–1401, by permission of Oxford University Press.

A convenient navigation in this view enables a quick assessment of many annotations in an efficient way. The labeling progress in the project can be tracked per image by

each observer, cf. Fig. B.1. Since each image may contain several hundred annotations, the application stores the labeling sessions and the user may resume at the last visited annotation from any client. Reviewing the assigned annotations is possible in the gallery view of Cytomine-IRIS, cf. Fig. B.2 (a). Outliers in each category can be rapidly identified by visual inspection and their label can be corrected by dragging the sample onto the correct term from the ontology.

Authorized project coordinators are able to view the progress of individual users in a project, cf. Fig. B.2 (b). This interface allows to identify biased observers, e.g. whether in inconclusive cases a particular observer tends to assign the earlier or later maturation stage. Further, the agreement statistics across all users can be queried and filtered. According to the selected filter, the labeled image data can be exported in a convenient way, cf. Fig. 11.2 (b), utilizing all magnification levels the original whole slide image provided.

11.2. A Novel Bone Marrow Cell Dataset

In a first study using Cytomine-IRIS, approximately $8,100$ cells were annotated in 50 H&E stained whole slide images of 50 individual patients. By the time this thesis was submitted, the bone marrow dataset consisted of $1,324$ representative cells in eleven classes having a full agreement among three independent observers, cf. Fig. 11.3. If we considered two of three agreeing observers, the dataset comprises additional 790 cells. Both distributions for myeloid and erythroid precursors are skewed, but reflect the approximate DBC in bone marrow reported in standard literature, cf. Fig. 2.5.

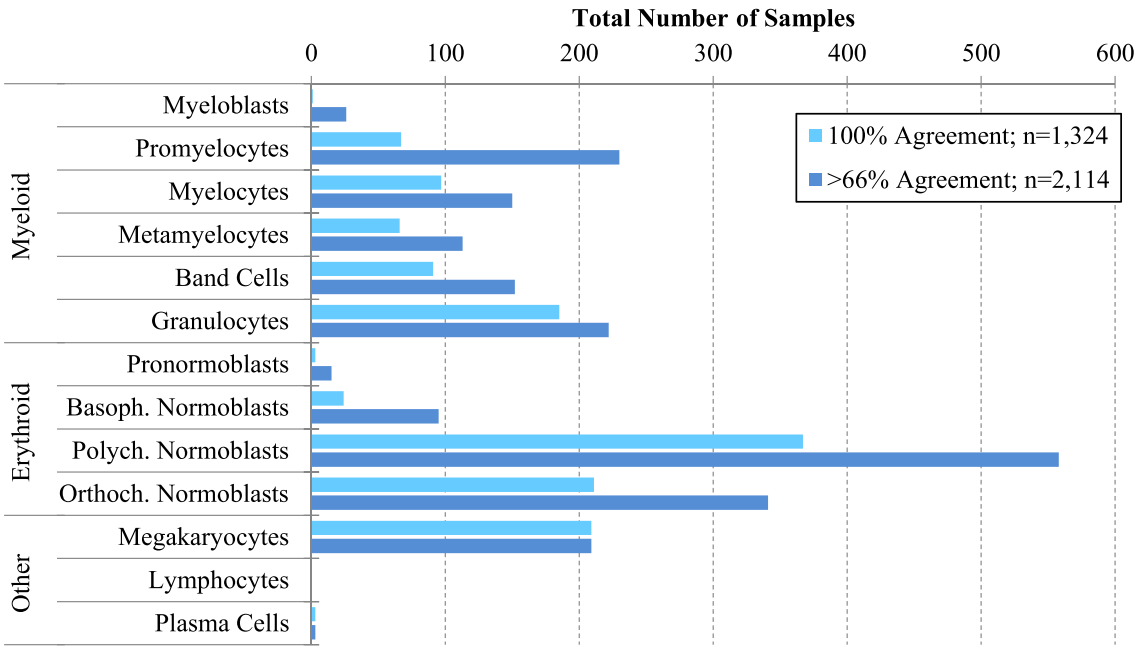


Figure 11.3.: Preliminary results of the data collection using the novel Cytomine-IRIS platform. A total (100%) agreement among all three observers was achieved on $1,324$ cells (cyan data series). The blue data series shows the label distribution for a filter set for lower level of agreement ($> 66\%$).

12. Discussion and Conclusions

12.1. Summary

The goal of the presented work was the development of cell localization and recognition algorithms based on machine learning for the purpose of histopathological analysis of healthy human bone marrow trephine biopsies, stained using a standard dyeing protocol (H&E, MGG) during routine examinations. Hematopathologists need to report, for each specimen, the distribution of several maturation stages in different hematopoietic cell lineages in the context of tissue architecture. Given the huge number of specimen that need to be analyzed in routine diagnostics, automated image analysis methods are demanded to increase reliability and throughput. Fostered by availability and progress in digital microscopy as well as novel methods for computer-aided diagnosis using state-of-the-art machine learning and computer vision methods, the interdisciplinary area of biomedical image analysis has seen much research effort during the past decade. In this work, it was demonstrated that learning-based pattern recognition algorithms achieve quantifiable, reproducible and accurate results for cell detection and classification. Despite a large body of publications on cell detection, segmentation and classification was already available, only very few papers dealt with the automated analysis of bone marrow core biopsies, even less with trephine biopsies. Many studies aimed at improving methods at the level of individual cells, others focused on processing whole slide histopathological images, e.g. to increase throughput in screening programs. In fact, an unexpectedly small number of the previously reported applications that was reviewed in this thesis addressed the quantification of the entire hematopoiesis in routine bone marrow sections, highlighting the urgency and rationale of this work. On the other hand, a vast majority focused on histopathological images from peripheral blood or bone marrow smear, mostly to quantify the number of white blood cells in the con-

text of diagnosing malignant disorders or performing differential blood cell counting. Localization and recognition of blood cell maturation stages were identified as two subsequent, yet challenging, concerns, which have been addressed separately in this thesis.

12.1.1. Cell Localization

Chapter 3 has proposed an innovative approach for robust cell localization in histopathological images that can be applied with minimal prior knowledge to many different tissue types, cf. Chapter 5. For the cells to be detected, a rough estimate of the cell nuclei radius is sufficient. It relies on regression Random Forests [165, 179] that learn to predict, for each image location, a smooth non-linear function of the distance to the closest cell center, which we termed the *proximity score*. The core idea was inspired by the work of Sironi *et al.* [162, 163] on multiscale centerline detection. The predictive models were trained using local image patches sampled from images containing simple dot-annotations that were placed on, or in the immediate adjacency of a cell nucleus center. Local maximums in the predicted proximity score map correspond to the cell center hypotheses, which can be discovered by non-maximum suppression. From a search across an extensive range of the method's hyper-parameters in cross-validation experiments we concluded that the actual values of the hyper-parameters are relatively uncritical for the method to operate properly. In Chapter 5 we have extensively evaluated the detection performance by means of several standard metrics in object localization on five challenging cell datasets containing erythroid and myeloid precursor cells, megakaryocytes, lymphocytes, and a mixed-tissue cell dataset comprising many different cell types. We compared the novel localization method to a state-of-the-art cell detector based on maximally stable extremal regions and structured SVM (*SSVM* [116]) that won the ICPR 2010 Pattern Recognition in Histopathological Images contest on lymphocyte detection by a large margin. As second baseline, a standard binary classification Random Forest was trained similar to the regression methods. We could show that the proposed regression methods outperformed both baseline methods in terms of detection performance and spatial localization accuracy, cf. Chapter 5. In terms of execution speed the RF methods ranked *ex aequo*, being $\approx 2\times$ faster than the *SSVM* method. The *spatial-averaging* extension of the simple center-pixel regressor could even achieve a speedup of $9\times$, while being as reliable as the baseline. It detects cells in a single 1200×1200 pixel image in 3.5 seconds. Both regression and classifica-

tion approaches using Random Forest were able to very reliably detect megakaryocytes. However, the *SSVM* detector was unable to learn a localization from the dot-annotated megakaryocytes, cf. Section 5.1.3. Advantages and disadvantages of the methods across the different datasets were discussed in detail in Chapter 6.

12.1.2. Cell Classification

In Part III of this thesis, the second aspect of the research question regarding the recognition of individual hematpoietic cells was addressed. A novel scheme for cell rotation-invariant classification using Echo State Networks was proposed. Given a set of grey-value image patches depicting single cells at its centers, each static 2D patch is transformed into a temporal signal by rotation. Driven by this signal, the activation patterns of the recurrently connected reservoir units encode the appearance information of the cell in temporal features, which are recorded over time. Finally, ridge regression is employed to learn weights for each cell class to be detected in a globally optimal fashion. In this proof-of-concept study, a five-class classification problem was formulated for the ESN, which is a generalization of earlier work on a similar, but binary, problem [9]. We used a manually extracted bone marrow cell dataset consisting of four classes, and randomly sampled background patches. Similarly to the annotation procedure for the localization datasets, an expert pathologist classified the cells by placing a single dot onto the cell nucleus centers and recording their class label. The bone marrow cells comprised three consecutive stages of myleoid precursors (myelocytes, metamyelocytes, and band cells), and one erythroid precursor stage (orthochromatic normoblasts). The selection of these particular cell types for this study was driven by two main factors. Firstly, a set of typical samples could more easily be generated, because these classes are among the most frequent in bone marrow sections of healthy adults, cf. Fig. 2.5. The second reason regarded the challenging discrimination of subsequent maturation stages within one lineage (myeloid) and the fact that precursor cells share morphological similarities across lineages. Similar to the cell localization approach, we pursue a segmentation-free strategy by learning a discriminative model directly from raw image patches. The proposed ESN approach was evaluated empirically in cross-validation experiments and compared to a standard classification RF (similar to the localization setup) as baseline using a benchmark split of the available dataset. The classification RF was trained the conventional way to achieve robustness against arbitrary object rotations, i.e. by augmenting the dataset via rotation by the same angles as the ESN

used for the continuous input stream generation. The choice for RF as the baseline was inspired by their successful application in cell localization, where they showed confident results on a similar patch-based problem. For both methods satisfactory class-wise and overall classification performance could be reported, and a trend over the evaluated hyper-parameter ranges could be discovered. Due to the internal dynamics of the ESN reservoir, it was possible to generate temporal features from single input stimuli followed by several time steps without any external input. Despite these features still possessed enough discriminative power to complete the classification task, the efficiency of this particular way of feeding input to the ESN is questionable. Hence, this strategy should not be pursued in this current form. In the benchmark experiment we examined what we termed the *true* rotation-invariance, i.e. whether a classifier predicted the same label for a cell in its original rotation and in a rotation defined by a randomly chosen offset angle to this position. While the ESN was able to classify 99.38% of all samples identically in this experiment, the RF achieved 95.02%. In this multi-class study we observed that the proposed ESN training scheme can be applied to rotation-invariant cell recognition and that an extraction of features from a single, untrained, randomly connected RNN was able to achieve higher performance than the RF. Nevertheless, we discussed in Chapter 10 the potentials to improve the current approach of training the ESN in terms of efficiency and runtime, which we could not perform in the scope of this thesis. As soon as more labeled data of different stages of maturation and cell lineages become available it can be tested whether more than five classes can be discriminated as well. First attempts towards this aim have already been undertaken and were briefly described in Chapter 11. Additionally, further comparison experiments with other cell classification methods that work with raw images must be performed.

12.2. Integration of Cell Localization and Recognition for General-Purpose Solutions

Despite this thesis addressed localization (Part II) and recognition (Part III) of bone marrow cells as two separate concerns, integrated solutions are required for practical applications. We have argued that the proposed methods are able to learn predictive models from image data without the requirement for accurate delineation of the objects, i.e. semantic segmentation. The proximity score regression can be used for fast and reliable identification of bone marrow cell nuclei centers in multiple stainings with

an average deviation from the true center of ≈ 2 pixels, which corresponds to ≈ 0.5 µm at $40\times$ magnification of the tested digital whole slide images. Moreover, the object size of the bone marrow cells is known at the scanned magnification and we can therefore obtain the single-cell patches for any subsequent classifier via simple cropping. This assumption represented a special case in our application, which actually allowed us to omit the segmentation step, since the myeloid and erythroid cells do not vary by more than ±15-20% in size (megakaryocytes were addressed separately). While incorporating prior knowledge into systems usually leads to better performance, it bears the risk of creating an *ad hoc* solution. To reduce this risk and working towards a general-purpose solution, additional considerations involving segmentation are inevitable. This applies especially to situations, where the size of cells varies and simple heuristics based on object sizes are not robust enough for automated patch extraction. Fortunately, one can resort to one of the many possible strategies to deal with cell nucleus (and cytoplasm) segmentation as well as separation of touching and overlapping objects in various tissues, which have been proposed in the literature, cf. related work in Section 3.1.1 and 7.1.1. Such an intermediate step to delineate the nuclei borders can easily be added before employing any classifier, e.g. by using the locations revealed by our proposed detector as seed points. However, it has to be ensured that segmentation is reliable, if the subsequent classifier relies on shape information. On the other hand, the classifier must be able to handle minor segmentation errors, which we currently did not need to consider in the presented bone marrow application.

A joint approach towards detection and classification using CNNs has been proposed recently [146] for the analysis of cell nuclei in colorectal cancer histopathology images[1]. Their joint model was trained on an extensive labeled cell dataset, containing about $30,000$ locations of cell nuclei, with about $22,000$ being classified as one of four cell types. The large size of this dataset facilitated studying an end-to-end learning strategy using CNN that predicts the location and the class of a cell using a single model. It shall be interesting to see whether a competitive RF can be trained using this data, but we could not evaluate this anymore within the scope of this thesis. However, it has to be noted that this large dataset has not been created with accompanying inter-observer reliability analysis. Data availability is usually a severe limitation that prohibits Deep Learning, hence an investigation of alternatives that work with less data as well is of high interest. Given the high spatial localization accuracy of the proposed regression method, and the classification accuracy shown by the RF-based cell classifier, such

[1] Their *CRChistoLabeledNucleiHE* dataset is publicly available from http://www.warwick.ac.uk/BIAlab/data/CRChistoLabeledNucleiHE.

a joint model could for instance be learned by decision tree based methods. However, solving both detection and classification using a single model has not yet been examined in this thesis. Others have proposed Hough Forests [166, 243], or joint classification and regression forests [169, 244], which can be taken as a starting point for future work on integrated localization and classification of multiple objects.

In any future cell classification application, *unknown* objects need to be considered as a separate class, especially when a huge set of candidates is provided by a localization algorithm such as the proposed one. Since in that case recall is optimized, more false positive nuclei location will be predicted that need to be filtered in subsequent processing steps. Classifiers need to have the opportunity to classify objects belonging to neither foreground (i.e. any cell class), nor background. This can be of importance in terms of at least two aspects. Firstly, a quantification of the decision confidence for individual instances provides more insight into the actual selectivity of a classifier. If the predicted labels are almost uniformly distributed over all classes, the classifier is likely to become unreliable and of little value for automated analysis. Secondly, provided the selectivity of the classes is reliable, the *unknown* class is used as element of quality control, and aids the estimation of the cellularity in a specimen. If many instances are predicted as this class, the biological sample may contain histological processing artifacts such as cell debris, blurring, and staining artifacts. Further, it could indicate a low count of actually intact cells, which may be caused by underlying disorders. Using fully automated tissue analysis in medical diagnostics requires mechanisms for human experts to validate the results and intervene at suspicious results.

12.3. Conclusions

Despite this work provided promising results for bone marrow cell localization and classification in histopathological images of bone marrow trephine biopsy, it can only be considered as an initial step. We require more robust models to deal with the classification of hematopoiesis, the physiological process of blood cell maturation. Until such computer-aided diagnostic tools can be effectively used in clinical practice, more research and development is required to integrate and optimize the individual components. The successful application of supervised machine learning in biomedical image analysis heavily relies on data availability and quality. By providing novel algorithms, an original cell detection dataset, and a tool to efficiently capture the consensus of many

domain experts, we took a further important step towards advancements in automated tissue analysis in hematopathology.

13. Outlook and Future Work

Many future research directions have already been revealed in the discussions of Part II-IV. This final chapter attempts to give an outlook to future work by raising further aspects worth of being considered.

Leveraging Unlabeled Data by *Humans-in-the-Training-Loop* A considerable number of histopathology sections is produced each day in pathology laboratories during routine tissue examinations. For diagnostic and archiving purposes these sections are scanned at high optical magnification to gigapixel whole slide images. While an in-depth analysis at a cellular level is time consuming, frequently even impossible for human observers, it holds the potential to leverage unlabeled image data to be used in a semi-supervised or interactive learning setting. The detection algorithm proposed in Part II of this thesis fully automatically generates accurate locations of cell nuclei in whole slide images, hence reveals many unlabeled objects for classification. A reliable expert-labeled ground truth cell dataset can be created by using specialized tools such as Cytomine-IRIS (Section 11.1). Semi-supervised techniques such as self-training [245] allow a classifier to learn to refine the decision boundaries between the classes by adding high confidence predictions. However, this may not be optimal for our bone marrow cell problem where expert-knowledge is essential. Ensuing from a cell classifier trained on an initial fully labeled dataset, more unlabeled instances can be predicted and proofread by human experts in an iterative process. Expert knowledge is actively included in the evolution of more robust cell classification system by constantly guiding the learning process. Proofreading predictions gains more importance, since recognition systems must be able to adapt to new requirements [120]. As more reliable data becomes available, i.e. as a result of such controlled studies, predictive models can be entirely retrained or refined on a larger dataset.

You should consider Regression to Classify Cells The nature of the present cell maturity recognition problem in fact allows us to approach it from alternative directions than the proposed multi-class classification. Blood cell maturation is a continuous process and many cells actually represent intermediate stages. A rigorous classification into predefined stages is therefore frequently inconclusive and still dependent on the observer's personal preference, and experience. Creating the ground truth in this work was focused on identifying typical cells for each maturation stage, for which the multi-class formulation and ongoing inter-observer study are suitable. Nevertheless, to capture general tendencies, many more experts are required to participate in such studies. As an alternative formulation, a multi-label classification or regression problem could be considered in future research. This work introduced a novel labeling tool for biomedical images, which enables the decentralized creation of extensive datasets in a convenient way. It can be adapted to facilitate labeling cells to be used in a regression problem as well. The concrete protocol for obtaining continuous labels has to be developed in future work, though. A quantization of the continuous output space could be performed to obtain a discrete class label for being used in a bone marrow differential blood count, which is standard in clinical practice.

There's a scientific word for this: weird.

Kowalski

The Penguins of Madagascar

Bibliography

[1] Cross SS, Dennis T, Start RD. Telepathology: current status and future prospects in diagnostic histopathology. Histopathology. 2002 Aug;41(2):91–109.

[2] Huang CH, Veillard A, Roux L, Loménie N, Racoceanu D. Time-efficient sparse analysis of histopathological whole slide images. Computerized Medical Imaging and Graphics. 2011 Oct-Dec;35(7–8):579–591.

[3] Riber-Hansen R, Vainer B, Steiniche T. Digital image analysis: a review of reproducibility, stability and basic requirements for optimal results. Acta Pathologica, Microbiologica et Immunologica Scandinavica (APMIS). 2012 Apr;120(4):276–289.

[4] Hamilton PW, Wang Y, McCullough SJ. Virtual microscopy and digital pathology in training and education. Acta Pathologica, Microbiologica et Immunologica Scandinavica (APMIS). 2012 Apr;120(4):305–315.

[5] Al-Janabi S, Huisman A, Van Diest PJ. Digital pathology: current status and future perspectives. Histopathology. 2011 Apr 11;61(1):1–9.

[6] Kainz P, Urschler M, Schulter S, Wohlhart P, Lepetit V. Bone Marrow Hematoxylin&Eosin Dataset; 2015. http://dx.doi.org/10.6084/m9.figshare.2056305.v1.

[7] Marée R, Rollus L, Stévens B, Hoyoux R, Louppe G, Vandaele R, et al. Collaborative analysis of multi-gigapixel imaging data using Cytomine. Bioinformatics. 2016 Jan;32(9):1395–1401.

[8] Kainz P, Urschler M, Schulter S, Wohlhart P, Lepetit V. You Should Use Regression to Detect Cells. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. vol. 9351 of Lecture Notes in Computer Science. Springer International Publishing; 2015. p. 276–283.

[9] Kainz P, Mayrhofer-Reinhartshuber M, Burgsteiner H, Asslaber M, Ahammer H. Echo State Networks for Granulopoietic Cell Recognition in Histopathological Images of Human Bone Marrow. Biomedizinische Technik. 2014 Oct;59(S1):S492–S495.

[10] Kainz P, Mayrhofer-Reinhartshuber M, Burgsteiner H, Asslaber M, Ahammer H. The Influence of Image Denoising on Granulopoietic Cell Recognition using Echo State Networks. In: International Biophysics Congress. Brisbane, Qld, Australia; 2014. p. 1.

[11] Kainz P, Burgsteiner H, Asslaber M, Ahammer H. Robust Bone Barrow Cell Discrimination by Rotation-Invariant Training of Multi-Class Echo State Networks. In: Iliadis L, Jayne C, editors. Engineering Applications of Neural Networks - EANN 2015. vol. 517 of Communications in Computer and Information Science. Rhodes, Greece: Springer International Publishing; 2015. p. 390–400.

[12] Kainz P, Burgsteiner H, Asslaber M, Ahammer H. Training Echo State Networks for Rotation-Invariant Bone Marrow Cell Classification. Neural Computing & Applications. In Revision;.

[13] Strayer DS, Rubin E, editors. Rubin's pathology: clinicopathologic foundations of medicine. 7th ed. Wolters Kluwer Health; 2015.

[14] Bain BJ, Clark DM, Wilkins BS. The Normal Bone Marrow. In: Bone Marrow Pathology. Wiley-Blackwell; 2009. p. 1–53.

[15] Nombela-Arrieta C, Pivarnik G, Winkel B, Canty KJ, Harley B, Mahoney JE, et al. Quantitative imaging of haematopoietic stem and progenitor cell localization and hypoxic status in the bone marrow microenvironment. Nature Cell Biology. 2013 26 Jun;15:533–543.

[16] Morrison SJ, Scadden DT. The bone marrow niche for haematopoietic stem cells. Nature. 2014 Jan;505(7483):327–334.

[17] Mendelson A, Frenette PS. Hematopoietic stem cell niche maintenance during homeostasis and regeneration. Nature Medicine. 2014 Aug;20(8):833–846.

[18] Arai F, Hirao A, Suda T. Regulation of Hematopoietic Stem Cells by the Niche. Trends in Cardiovascular Medicine. 2005 Feb;15(2):75–79.

[19] Moore KA, Lemischka IR. Stem Cells and Their Niches. Science. 2006 Mar;311(5769):1880–1885.

[20] Li Z, Li L. Understanding hematopoietic stem-cell microenvironments. Trends in Biochemical Sciences. 2006 Oct;31(10):589–595.

[21] Till JE, McCulloch EA. A Direct Measurement of the Radiation Sensitivity of Normal Mouse Bone Marrow Cells. Radiation Research. 1961;14(2):213–222.

[22] Spangrude G, Heimfeld S, Weissman I. Purification and characterization of mouse hematopoietic stem cells. Science. 1988 Jul;241(4861):58–62.

[23] Oh IH, Kwon KR. Concise Review: Multiple Niches for Hematopoietic Stem Cell Regulations. Stem Cells. 2010;28(7):1243–1249.

[24] Doan PL, Chute JP. The vascular niche: home for normal and malignant hematopoietic stem cells. Leukemia. 2012 Jan;26(1):54–62.

[25] Birbrair A, Frenette PS. Niche heterogeneity in the bone marrow. Annals of the New York Academy of Sciences. 2016;1370(1):82–96.

[26] Huang X, Cho S, Spangrude GJ. Hematopoietic stem cells: generation and self-renewal. Cell Death and Differentiation. 2007 Sep;14(11):1851–1859.

[27] Mosaad YM. Hematopoietic stem cells: An overview. Transfusion and Apheresis Science. 2014 Dec;51(3):68–82.

[28] Anthony BA, Link DC. Regulation of hematopoietic stem cells by bone marrow stromal cells. Trends in Immunology. 2014;35(1):32–37.

[29] Chotinantakul K, Leeanansaksiri W. Hematopoietic Stem Cell Development, Niches, and Signaling Pathways. Bone Marrow Research. 2012;2012:270425.

[30] Mikkola HKA, Orkin SH. The journey of developing hematopoietic stem cells. Development. 2006;133(19):3733–3744.

[31] Boisset JC, Robin C. On the origin of hematopoietic stem cells: Progress and controversy. Stem Cell Research. 2012 Jan;8(1):1–13.

[32] O'Malley DP. Benign extramedullary myeloid proliferations. Modern Pathology. 2007;20(4):405–415.

[33] Kim CH. Homeostatic and pathogenic extramedullary hematopoiesis. Journal of Blood Medicine. 2010;1:13–19.

[34] Nwajei F, Konopleva M. The bone marrow microenvironment as niche retreats for hematopoietic and leukemic stem cells. Advances in Hematology. 2013;2013:953982.

[35] Goodell MA, Nguyen H, Shroyer N. Somatic stem cell heterogeneity: diversity in the blood, skin and intestinal stem cell compartments. Nature Reviews Molecular Cell Biology. 2015 May;16(5):299–309.

[36] Woolthuis CM, Park CY. Hematopoietic stem/progenitor cell commitment to the megakaryocyte lineage. Blood. 2016;127(10):1242–1248.

[37] Nemeth MJ, Bodine DM. Regulation of hematopoiesis and the hematopoietic stem cell niche by Wnt signaling pathways. Cell Research. 2007 Sep;17(9):746–758.

[38] Wolff L, Humeniuk R. Concise Review: Erythroid Versus Myeloid Lineage Commitment: Regulating the Master Regulators. Stem Cells. 2013;31(7):1237–1244.

[39] Yang J, Zhang L, Yu C, Yang XF, Wang H. Monocyte and macrophage differentiation: circulation inflammatory monocyte as biomarker for inflammatory diseases. Biomarker Research. 2014;2(1):1–9.

[40] Naeim F, Rao PN, Song SX, Grody WW. Structure and Function of Hematopoietic Tissues. In: Naeim F, Rao PN, Song SX, Grody WW, editors. Atlas of Hematopathology. Academic Press; 2013. p. 1–24.

[41] Naeim F, Rao PN, Song SX, Grody WW. Granulocytic Disorders. In: Naeim F, Rao PN, Song SX, Grody WW, editors. Atlas of Hematopathology. Academic Press; 2013. p. 663–673.

[42] Naeim F, Rao PN, Song SX, Grody WW. Disorders of Red Blood Cells – Anemias. In: Naeim F, Rao PN, Song SX, Grody WW, editors. Atlas of Hematopathology. Academic Press; 2013. p. 675–704.

[43] Naeim F, Rao PN, Song SX, Grody WW. Disorders of Megakaryocytes and Platelets. In: Naeim F, Rao PN, Song SX, Grody WW, editors. Atlas of Hematopathology. Academic Press; 2013. p. 705–714.

[44] Naeim F, Rao PN, Song SX, Grody WW. Myelodysplastic Syndromes/Neoplasms – Overview. In: Naeim F, Rao PN, Song SX, Grody WW, editors. Atlas of Hematopathology. Academic Press; 2013. p. 111–127.

[45] Naeim F, Rao PN, Song SX, Grody WW. Acute Myeloid Leukemia – Overview. In: Naeim F, Rao PN, Song SX, Grody WW, editors. Atlas of Hematopathology. Academic Press; 2013. p. 219–226.

[46] Bain BJ. Bone marrow trephine biopsy. Journal of Clinical Pathology. 2001 Oct;54(10):737–742.

[47] Bain BJ. Bone marrow aspiration. Journal of Clinical Pathology. 2001;54(9):657–663.

[48] Wilkins BS, Clark DM. Making the most of bone marrow trephine biopsy. Histopathology. 2009;55(6):631–640.

[49] Afkhami M, Vergara-Lluri M, Brynes RK, Siddiqi IN. Peripheral Blood Smears, Bone Marrow Aspiration, Trephine and Clot Biopsies: Methods and Protocols. In: Day EC, editor. Histopathology: Methods and Protocols. New York, NY: Springer New York; 2014. p. 257–269.

[50] Lee SH, Erber WN, Porwit A, Tomonaga M, Peterson LC, International Council for Standardization In Hematology. ICSH guidelines for the standardization of bone marrow specimens and reports. International Journal of Laboratory Hematology. 2008;30(5):349–364.

[51] Stuart-Smith SE, Hughes DA, Bain BJ. Are routine iron stains on bone marrow trephine biopsy specimens necessary? Journal of Clinical Pathology. 2005;58(3):269–272.

[52] Pujara K, Bhalara R, Dhruva G. A study of bone marrow iron storage in hematological disorder. International Journal of Health & Allied Sciences. 2014;3(4):221–224.

[53] Beham-Schmid C, Apfelbeck U, Sill H, Tsybrovsky O, Höfler G, Haas OA, et al. Treatment of chronic myelogenous leukemia with the tyrosine kinase inhibitor STI571 results in marked regression of bone marrow fibrosis. Blood. 2002;99(1):381–383.

[54] Al-Adhadh AN, Cavill I. Assessment of cellularity in bone marrow fragments. Journal of Clinical Pathology. 1983;36:176–179.

[55] Horning SJ. Something Old, Something Few, Something Subjective, Something Déjà Vu. Journal of Clinical Oncology. 2003;21(1):1–2.

[56] Irshad H, Veillard A, Roux L, Racoceanu D. Methods for Nuclei Detection, Segmentation, and Classification in Digital Histopathology: A Review – Current Status and Future Potential. IEEE Reviews in Biomedical Engineering. 2014;7:97–114.

[57] Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B. Histopathological image analysis: a review. IEEE Reviews in Biomedical Engineering. 2009;2:147–171.

[58] Meijering E. Cell Segmentation: 50 Years Down the Road [Life Sciences]. IEEE Signal Processing Magazine. 2012 Sep;29(5):140–145.

[59] Sertel O, Catalyurek UV, Shimada H, Gurcan MN. Computer-aided Prognosis of Neuroblastoma: Detection of mitosis and karyorrhexis cells in digitized histological images. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2009. p. 1433–1436.

[60] Malon C, Brachtel E, Cosatto E, Graf HP, Kurata A, Kuroda M, et al. Mitotic Figure Recognition: Agreement among Pathologists and Computerized Detector. Analytical Cellular Pathology (Amsterdam). 2012;35(2):97–100.

[61] Malon CD, Cosatto E. Classification of mitotic figures with convolutional neural networks and seeded blob features. Journal of Pathology Informatics. 2013 May;4:9.

[62] Khan AM, Eldaly H, Rajpoot NM. A gamma-gaussian mixture model for detection of mitotic cells in breast cancer histopathology images. Journal of Pathology Informatics. 2013 May;4:11.

[63] Irshad H. Automated mitosis detection in histopathology using morphological and multi-channel statistics features. Journal of Pathology Informatics. 2013 May;4:10.

[64] Irshad H, Jalali S, Roux L, Racoceanu D, Hwee L, Naour G, et al. Automated mitosis detection using texture, SIFT features and HMAX biologically inspired approach. Journal of Pathology Informatics. 2013;4(2):12.

[65] Ciresan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. In: Mori K, Sakuma I, Sato Y, Barillot C, Navab N, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013. vol. 8150 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2013. p. 411–418.

[66] Sirinukunwattana K, Khan AM, Rajpoot NM. Cell words: Modelling the visual appearance of cells in histopathology images. Computerized Medical Imaging and Graphics. 2015;42(0):16–24.

[67] Petushi S, Katsinis C, Coward C, Garcia F, Tozeren A. Automated identification of microstructures on histology slides. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2004. p. 424–427 Vol. 1.

[68] Demir C, Yener B. Automated cancer diagnosis based on histopathological images: a systematic survey. Rensselaer Polytechnic Institute; 2005.

[69] Petushi S, Garcia FU, Haber MM, Katsinis C, Tozeren A. Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. BMC Medical Imaging. 2006;6(1):1–11.

[70] Doyle S, Hwang M, Shah K, Madabhushi A, Feldman M, Tomaszeweski J. Automated Grading of Prostate Cancer using Architectural and Textural Image Features. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2007. p. 1284–1287.

[71] Fuchs TJ, Lange T, Wild PJ, Moch H, Buhmann JM. Weakly Supervised Cell Nuclei Detection and Segmentation on Tissue Microarrays of Renal Clear Cell Carcinoma. In: Rigoll G, editor. Pattern Recognition: 30th DAGM Symposium 2008. vol. 5096 of Lecture Notes in Computer Science. Munich, Germany: Springer Berlin Heidelberg; 2008. p. 173–182.

[72] Yang L, Chen W, Meer P, Salaru G, Feldman MD, Foran DJ. High Throughput Analysis of Breast Cancer Specimens on the Grid. In: Ayache N, Ourselin S, Maeder A, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. p. 617–625.

[73] Buggenthin F, Marr C, Schwarzfischer M, Hoppe PS, Hilsenbeck O, Schroeder T, et al. An automatic method for robust and fast cell detection in bright field images from high-throughput microscopy. BMC Bioinformatics. 2013;14(1):1–12.

[74] Arbelle A, Drayman N, Bray M, Alon U, Carpenter A, Raviv TR. Analysis of High-throughput Microscopy Videos: Catching Up with Cell Dynamics. In: Navab N, Hornegger J, Wells MW, Frangi FA, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing; 2015. p. 218–225.

[75] Monaco J, Raess P, Chawla R, Bagg A, Weiss M, Choi J, et al. Image segmentation with implicit color standardization using cascaded EM: Detection of myelodysplastic syndromes. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2012. p. 740–743.

[76] Ramoser H, Laurain V, Bischof H, Ecker R. Leukocyte segmentation and classification in blood-smear images. In: 27th Annual International Conference of the Engineering in Medicine and Biology Society; 2005. p. 3371–3374.

[77] Yang L, Meer P, Foran DJ. Unsupervised segmentation based on robust estimation and color active contour models. IEEE Transactions on Information Technology in Biomedicine. 2005 Sep;9(3):475–486.

[78] Hamghalam M, Ayatollahi A. Automatic Counting of Leukocytes in Giemsa-Stained Images of Peripheral Blood Smear. In: International Conference on Digital Image Processing; 2009. p. 13–16.

[79] Ballarò B, Florena AM, Franco V, Tegolo D, Tripodo C, Valenti C. An automated image analysis methodology for classifying megakaryocytes in chronic myeloproliferative disorders. Medical Image Analysis. 2008 Dec;12(6):703 – 712.

[80] Reta C, Altamirano L, Gonzalez JA, Diaz-Hernandez R, Peregrina H, Olmos I, et al. Segmentation and Classification of Bone Marrow Cells Images Using Contextual Information for Medical Diagnosis of Acute Leukemias. PLoS ONE. 2015;10(6):e0130805.

[81] Scotti F. Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. In: IEEE International Conference on Computational Intelligence for Measurement Systems and Applications – CIMSA; 2005. p. 96–101.

[82] Zhou ZH, Jiang Y, Yang YB, Chen SF. Lung cancer cell identification based on artificial neural network ensembles. Artificial Intelligence in Medicine. 2002;24(1):25–36.

[83] Basavanhally AN, Ganesan S, Agner S, Monaco JP, Feldman MD, Tomaszewski JE, et al. Computerized Image-Based Detection and Grading of Lymphocytic Infiltration in HER2+ Breast Cancer Histopathology. IEEE Transactions on Biomedical Engineering. 2010 Mar;57(3):642–653.

[84] Ali S, Madabhushi A. Active Contour for Overlap Resolution using Watershed based initialization (ACOReW): Applications to histopathology. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2011. p. 614–617.

[85] Yuan Y, Failmezger H, Rueda OM, Ali HR, Gräf S, Chin SF, et al. Quantitative Image Analysis of Cellular Heterogeneity in Breast Tumors Complements Genomic Profiling. Science Translational Medicine. 2012;4(157):157ra143–157ra143.

[86] Jing J, Wan T, Cao J, Qin Z. An improved hybrid active contour model for nuclear segmentation on breast cancer histopathology. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2016. p. 1155–1158.

[87] Cao J, Qin Z, Jing J, Chen J, Wan T. An automatic breast cancer grading method in histopathological images based on pixel-, object-, and semantic-level features. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2016. p. 1151–1154.

[88] Sertel O, Catalyurek UV, Lozanski G, Shanaah A, Gurcan MN. An Image Analysis Approach for Detecting Malignant Cells in Digitized H&E-stained Histology Images of Follicular Lymphoma. In: IEEE International Conference on Pattern Recognition; 2010. p. 273–276.

[89] Lezoray O, Cardot H. Cooperation of color pixel classification schemes and color watershed: a study for microscopic images. IEEE Transactions on Image Processing. 2002 Jul;11(7):783–789.

[90] Yang L, Tuzel O, Meer P, Foran DJ. Automatic Image Analysis of Histopathology Specimens Using Concave Vertex Graph. In: Metaxas D, Axel L, Fichtinger G, Székely G, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008. vol. 5241 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2008. p. 833–841.

[91] Coelho LP, Shariff A, Murphy RF. Nuclear segmentation in microscope cell images: A hand-segmented dataset and comparison of algorithms. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2009. p. 518–521.

[92] Veta M, Huisman A, Viergever MA, van Diest PJ, Pluim JPW. Marker-controlled watershed segmentation of nuclei in H&E stained breast cancer biopsy images. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2011. p. 618–621.

[93] Veta M, van Diest PJ, Kornegoor R, Huisman A, Viergever MA, Pluim JPW. Automatic Nuclei Segmentation in H&E Stained Breast Cancer Histopathology Images. PLoS ONE. 2013 Jul;8(7):e70221.

[94] Veillard A, Kulikova MS, Racoceanu D. Cell nuclei extraction from breast cancer histopathologyimages using colour, texture, scale and shape information. Diagnostic Pathology. 2013;8(Suppl 1):S5–S5.

[95] Chang H, Han J, Borowsky A, Loss L, Gray JW, Spellman PT, et al. Invariant Delineation of Nuclear Architecture in Glioblastoma Multiforme for Clinical and Molecular Association. IEEE Transactions on Medical Imaging. 2013 Apr;32(4):670–682.

[96] Yang X, Li H, Zhou X. Nuclei Segmentation Using Marker-Controlled Watershed, Tracking Using Mean-Shift, and Kalman Filter in Time-Lapse Microscopy. IEEE Transactions on Circuits and Systems I: Regular Papers. 2006 Nov;53(11):2405–2414.

[97] Wählby C, Sintorn IM, Erlandsson F, Borgefors G, Bengtsson E. Combining intensity, edge and shape information for 2D and 3D segmentation of cell nuclei in tissue sections. Journal of Microscopy. 2004;215(1):67–76.

[98] Adawy ME, Shehab Z, Keshk H, Shourbagy ME. A Fast Algorithm for Segmentation of Microscopic Cell Images. In: International Conference on Information Communications Technology; 2006. p. 1–1.

[99] Ko B, Seo M, Nam JY. Microscopic Cell Nuclei Segmentation Based on Adaptive Attention Window. Journal of Digital Imaging. 2008;22(3):259–274.

[100] Dalle JR, Li H, Huang CH, Leow WK, Racoceanu D, Putti TC. Nuclear pleomorphism scoring by selective cell nuclei detection. In: IEEE Workshop on Applications of Computer Vision; 2009. p. 1–6.

[101] Al-Kofahi Y, Lassoued W, Lee W, Roysam B. Improved Automatic Detection and Segmentation of Cell Nuclei in Histopathology Images. IEEE Transactions on Biomedical Engineering. 2010 Apr;57(4):841–852.

[102] Ko BC, Gim JW, Nam JY. Automatic white blood cell segmentation using stepwise merging rules and gradient vector flow snake. Micron. 2011 Oct;42(7):695–705.

[103] Kuse M, Sharma T, Gupta S. A Classification Scheme for Lymphocyte Segmentation in H&E Stained Histology Images. In: Ünay D, Çataltepe Z, Aksoy S, editors. Recognizing Patterns in Signals, Speech, Images and Videos. vol. 6388 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2010. p. 235–243.

[104] Kuse M, Wang YF, Kalasannavar V, Khan M, Rajpoot N. Local isotropic phase symmetry measure for detection of beta cells and lymphocytes. Journal of Pathology Informatics. 2011;2(2).

[105] Bernardis E, Yu SX. Pop out many small structures from a very large microscopic image. Medical Image Analysis. 2011 Oct;15(5):690–707.

[106] Cloppet F, Boucher A. Segmentation of overlapping/aggregating nuclei cells in biological images. In: IEEE International Conference on Pattern Recognition; 2008. p. 1–4.

[107] Cheng J, Rajapakse JC. Segmentation of Clustered Nuclei With Shape Markers and Marking Function. IEEE Transactions on Biomedical Engineering. 2009 Mar;56(3):741–748.

[108] Jung C, Kim C, Chae SW, Oh S. Unsupervised Segmentation of Overlapped Nuclei Using Bayesian Classification. IEEE Transactions on Biomedical Engineering. 2010 Dec;57(12):2825–2832.

[109] Qi X, Xing F, Foran DJ, Yang L. Robust Segmentation of Overlapping Cells in Histopathology Specimens Using Parallel Seed Detection and Repulsive Level Set. IEEE Transactions on Biomedical Engineering. 2012 Mar;59(3):754–765.

[110] Shu J, Fu H, Qiu G, Kaye P, Ilyas M. Segmenting overlapping cell nuclei in digital histopathology images. In: 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2013. p. 5445–5448.

[111] Ali S, Madabhushi A. An Integrated Region-, Boundary-, Shape-Based Active Contour for Multiple Object Overlap Resolution in Histological Imagery. IEEE Transactions on Medical Imaging. 2012 Jul;31(7):1448–1460.

[112] Kulikova M, Veillard A, Roux L, Racoceanu D. Nuclei extraction from histopathological images using a marked point process approach. Proceedings of SPIE. 2012;8314:831428–831428–8.

[113] Wienert S, Heim D, Saeger K, Stenzinger A, Beil M, Hufnagl P, et al. Detection and Segmentation of Cell Nuclei in Virtual Microscopy Images: A Minimum-Model Approach. Scientific Reports. 2012 Jul;2(503).

[114] Mosaliganti K, Cooper L, Sharp R, Machiraju R, Leone G, Huang K, et al. Reconstruction of Cellular Biological Structures from Optical Microscopy Data.

IEEE Transactions on Visualization and Computer Graphics. 2008 Jul;14(4):863–876.

[115] Al-Kofahi Y, Lassoued W, Grama K, Nath SK, Zhu J, Oueslati R, et al. Cell-based quantification of molecular biomarkers in histopathology specimens. Histopathology. 2011 Jul;59(1):40–54.

[116] Arteta C, Lempitsky V, Noble JA, Zisserman A. Learning to Detect Cells Using Non-overlapping Extremal Regions. In: Ayache N, Delingette H, Golland P, Mori K, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012. vol. 7510 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2012. p. 348–356.

[117] Jung C, Kim C. Segmenting Clustered Nuclei Using H-minima Transform-Based Marker Extraction and Contour Parameterization. IEEE Transactions on Biomedical Engineering. 2010 Oct;57(10):2600–2604.

[118] Fuchs TJ. Computational Pathology: A Machine Learning Approach [PhD Thesis]. Swiss Federal Institute of Technology (ETH), Zurich. Zurich, Switzerland; 2010.

[119] Yin Z, Bise R, Chen M, Kanade T. Cell segmentation in microscopy imagery using a bag of local Bayesian classifiers. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2010. p. 125–128.

[120] Fuchs TJ, Buhmann JM. Computational Pathology: Challenges and promises for tissue analysis. Computerized Medical Imaging and Graphics. 2011;35(7–8):515–530.

[121] Monaco J, Hipp J, Lucas D, Smith S, Balis U, Madabhushi A. Image Segmentation with Implicit Color Standardization Using Spatially Constrained Expectation Maximization: Detection of Nuclei. In: Ayache N, Delingette H, Golland P, Mori K, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012. vol. 7510 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2012. p. 365–372.

[122] Veillard A, Bressan S, Racoceanu D. SVM-based Framework for the Robust Extraction of Objects from Histopathological Images Using Color, Texture, Scale and Geometry. In: International Conference on Machine Learning and Applications (ICMLA). vol. 1; 2012. p. 70–75.

[123] Vink JP, Van Leeuwen MB, Van Deurzen CHM, De Haan G. Efficient nucleus detector in histopathology images. Journal of Microscopy. 2013;249(2):124–135.

[124] Sadeghian F, Seman Z, Ramli AR, Kahar BHA, Saripan MI. A Framework for White Blood Cell Segmentation in Microscopic Blood Images Using Digital Image Processing. Biological Procedures Online. 2009 11 Jun;11:196–206.

[125] Nosrati M, Hamarneh G. Segmentation of Cells with Partial Occlusion and Part Configuration Constraint Using Evolutionary Computation. In: Mori K, Sakuma I, Sato Y, Barillot C, Navab N, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013. vol. 8149 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2013. p. 461–468.

[126] Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. 1995 Sep;20(3):273–297.

[127] Freund Y, Schapire RE. Experiments with a New Boosting Algorithm. In: Saitta L, editor. International Conference on Machine Learning. Morgan Kaufmann; 1996. p. 148–156.

[128] Cosatto E, Miller M, Graf HP, Meyer JS. Grading nuclear pleomorphism on histological micrographs. In: IEEE International Conference on Pattern Recognition; 2008. p. 1–4.

[129] Sharma H, Zerbe N, Heim D, Wienert S, Behrens HM, Hellwich O, et al. A Multi-resolution Approach for Combining Visual Information using Nuclei Segmentation and Classification in Histopathological Images. In: Braz J, Battiato S, Imai FH, editors. In Proceedings of the 10th International Conference on Computer Vision Theory and Applications (VISAPP-2015). Berlin, Germany: SciTePress; 2015. p. 37–46.

[130] Gurcan MN, Madabhushi A, Rajpoot N. Pattern Recognition in Histopathological Images: An ICPR 2010 Contest. In: Ünay D, Çataltepe Z, Aksoy S, editors. Recognizing Patterns in Signals, Speech, Images and Videos. vol. 6388 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2010. p. 226–234.

[131] Matas J, Chum O, Urban M, Pajdla T. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In: Proceedings of the British Machine Vision Conference. Cardiff, UK: British Machine Vision Association; 2002. p. 1–10.

[132] Matas J, Chum O, Urban M, Pajdla T. Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing. 2004 Sep;22(10):761 – 767.

[133] Canny J. A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1986 Nov;PAMI-8(6):679–698.

[134] Quelhas P, Marcuzzo M, Mendonça AM, Campilho A. Cell nuclei and cytoplasm joint segmentation using the sliding band filter. IEEE Transactions on Medical Imaging. 2010;29(8):1463–1473.

[135] Lowe DG. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision. 2004;60(2):91–110.

[136] Mualla F, Schöll S, Sommerfeldt B, Maier A, Steidl S, Buchholz R, et al. Unsupervised Unstained Cell Detection by SIFT Keypoint Clustering and Self-labeling Algorithm. In: Golland P, Hata N, Barillot C, Hornegger J, Howe R, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014. Cham: Springer International Publishing; 2014. p. 377–384.

[137] Habibzadeh M, Krzy zak A, Fevens T. White Blood Cell Differential Counts Using Convolutional Neural Networks for Low Resolution Images. In: Rutkowski L, Korytkowski M, Scherer R, Tadeusiewicz R, Zadeh L, Zurada J, editors. Artificial Intelligence and Soft Computing. vol. 7895 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2013. p. 263–274.

[138] Cruz-Roa AA, Arevalo Ovalle J, Madabhushi A, González Osorio FA. A Deep Learning Architecture for Image Representation, Visual Interpretability and Automated Basal-Cell Carcinoma Cancer Detection. In: Mori K, Sakuma I, Sato Y, Barillot C, Navab N, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013. vol. 8150 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2013. p. 403–410.

[139] Wang H, Cruz-Roa A, Basavanhally A, Gilmore H, Shih N, Feldman M, et al. Cascaded ensemble of convolutional neural networks and handcrafted features for mitosis detection. Proceedings of SPIE. 2014;9041:90410B–90410B–10.

[140] Xie W, Noble JA, Zisserman A. Microscopy Cell Counting with Fully Convolutional Regression Networks. In: MICCAI 1st Workshop on Deep Learning in Medical Image Analysis; 2015. p. 25–32.

[141] Xie Y, Kong X, Xing F, Liu F, Su H, Yang L. Deep Voting: A Robust Approach Toward Nucleus Localization in Microscopy Images. In: Navab N, Hornegger J, Wells MW, Frangi FA, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science. Springer International Publishing; 2015. p. 374–382.

[142] Xie Y, Xing F, Kong X, Su H, Yang L. Beyond Classification: Structured Regression for Robust Cell Detection Using Convolutional Neural Network. In: Navab N, Hornegger J, Wells MW, Frangi FA, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science. Springer International Publishing; 2015. p. 358–365.

[143] Liu F, Xing F, Zhang Z, Mcgough M, Yang L. Robust Muscle Cell Quantification Using Structured Edge Detection and Hierarchical Segmentation. In: Navab N, Hornegger J, Wells MW, Frangi FA, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science. Springer International Publishing; 2015. p. 324–331.

[144] Liu F, Yang L. A Novel Cell Detection Method Using Deep Convolutional Neural Network and Maximum-Weight Independent Set. In: Navab N, Hornegger J, Wells MW, Frangi FA, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science. Springer International Publishing; 2015. p. 349–357.

[145] Su H, Xing F, Kong X, Xie Y, Zhang S, Yang L. Robust Cell Detection and Segmentation in Histopathological Images Using Sparse Reconstruction and Stacked Denoising Autoencoders. In: Navab N, Hornegger J, Wells MW, Frangi FA, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science. Springer International Publishing; 2015. p. 383–390.

[146] Sirinukunwattana K, Raza SEA, Tsang YW, Snead DRJ, Cree IA, Rajpoot NM. Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. IEEE Transactions on Medical Imaging. 2016 May;35(5):1196–1206.

[147] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted

Intervention – MICCAI 2015. vol. 9351 of Lecture Notes in Computer Science. Springer; 2015. p. 234–241.

[148] Chen H, Qi X, Yu L, Heng P. DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation. CoRR. 2016;abs/1604.02677.

[149] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998 Nov;86(11):2278–2324.

[150] LeCun Y, Kavukcuoglu K, Farabet C. Convolutional networks and applications in vision. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS); 2010. p. 253–256.

[151] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Bartlett P, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems 25; 2012. p. 1106–1114.

[152] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR. 2014;abs/1409.1556.

[153] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going Deeper with Convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition; 2015. p. 1–9.

[154] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. CoRR. 2015;abs/1502.03167.

[155] Lempitsky V, Zisserman A. Learning To Count Objects in Images. In: Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A, editors. Advances in Neural Information Processing Systems 23. Curran Associates, Inc.; 2010. p. 1324–1332.

[156] Fiaschi L, Nair R, Koethe U, Hamprecht FA. Learning to count with regression forest and structured labels. In: IEEE International Conference on Pattern Recognition; 2012. p. 2685–2688.

[157] Welinder P, Perona P. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In: IEEE Conference on Computer Vision and Pattern Recognition, Workshop on Advancing Computer Vision with Humans in the Loop; 2010. p. 1–8.

[158] Russell BC, Torralba A, Murphy KP, Freeman WT. LabelMe: A Database and Web-Based Tool for Image Annotation. International Journal of Computer Vision. 2008;77(1-3):157–173.

[159] Maška M, Ulman V, Svoboda D, Matula P, Matula P, Ederra C, et al. A benchmark for comparison of cell tracking algorithms. Bioinformatics. 2014 Jun;30(11):1609–1617.

[160] Sirinukunwattana K, Snead DRJ, Rajpoot NM. A Stochastic Polygons Model for Glandular Structures in Colon Histology Images. IEEE Transactions on Medical Imaging. 2015 Nov;34(11):2366–2378.

[161] Arganda-Carreras I, Turaga SC, Berger DR, Ciresan D, Giusti A, Gambardella LM, et al. Crowdsourcing the creation of image segmentation algorithms for connectomics. Frontiers in Neuroanatomy. 2015;9(142).

[162] Sironi A, Lepetit V, Fua P. Multiscale Centerline Detection by Learning a Scale-Space Distance Transform. In: IEEE International Conference on Computer Vision and Pattern Recognition; 2014. p. 1–8.

[163] Sironi A, Türetken E, Lepetit V, Fua P. Multiscale Centerline Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016 Jul;38(7):1327–1341.

[164] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. 2nd ed. Springer Series in Statistics. Springer; 2013.

[165] Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32.

[166] Gall J, Yao A, Razavi N, van Gool L, Lempitsky V. Hough Forests for Object Detection, Tracking, and Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2011 Nov;33(11):2188–2201.

[167] Criminisi A, Shotton J. Decision Forests for Computer Vision and Medical Image Analysis. Springer Publishing Company, Incorporated; 2013.

[168] Schulter S, Wohlhart P, Leistner C, Saffari A, Roth PM, Bischof H. Alternating Decision Forests. In: IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR; 2013. p. 508–515.

[169] Schulter S, Leistner C, Wohlhart P, Roth PM, Bischof H. Accurate Object Detection with Joint Classification-Regression Random Forests. In: IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 923–930.

[170] Gonzalez RC, Woods RE. Digital image processing. Upper Saddle River, N.J.: Prentice Hall International; 2008.

[171] Szeliski R. Computer Vision: Algorithms and Applications. Texts in Computer Science. Springer London; 2011.

[172] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition. vol. 1; 2005. p. 886–893.

[173] Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on featured distributions. Pattern Recognition. 1996;29(1):51–59.

[174] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002 Jul;24(7):971–987.

[175] Ahonen T, Hadid A, Pietikäinen M. Face Recognition with Local Binary Patterns. In: Pajdla T, Matas J, editors. Computer Vision - ECCV 2004. vol. 3021 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2004. p. 469–481.

[176] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: IEEE Conference on Computer Vision and Pattern Recognition. vol. 1; 2001. p. I–511–I–518.

[177] Breiman L. Bagging predictors. Machine Learning. 1996;24(2):123–140.

[178] Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. Monterey, CA: Wadsworth and Brooks; 1984.

[179] Criminisi A, Shotton J, Konukoglu E. Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. In: Foundations and Trends® in Computer Graphics and Vision. vol. 7. Hanover, MA, USA: Now Publishers Inc.; 2011. p. 81–227.

[180] Forsyth DA, Ponce J. Computer Vision: A Modern Approach. 1st ed. Prentice Hall Professional Technical Reference; 2002.

[181] Neubeck A, Gool LJV. Efficient Non-Maximum Suppression. In: IEEE International Conference on Pattern Recognition. Hong Kong, China; 2006. p. 850–855.

[182] Graf F, Grzegorzek M, Paulus D. Counting Lymphocytes in Histopathology Images Using Connected Components. In: Ünay D, Çataltepe Z, Aksoy S, editors. Recognizing Patterns in Signals, Speech, Images and Videos. vol. 6388 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2010. p. 263–269.

[183] Cheng J, Veronika M, Rajapakse J. Identifying Cells in Histopathological Images. In: Ünay D, Çataltepe Z, Aksoy S, editors. Recognizing Patterns in Signals, Speech, Images and Videos. vol. 6388 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2010. p. 244–252.

[184] Panagiotakis C, Ramasso E, Tziritas G. Lymphocyte Segmentation Using the Transferable Belief Model. In: Ünay D, Çataltepe Z, Aksoy S, editors. Recognizing Patterns in Signals, Speech, Images and Videos. vol. 6388 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2010. p. 253–262.

[185] Quinlan JR. Induction of decision trees. Machine Learning. 1986 Mar;1(1):81–106.

[186] Quinlan JR. C4.5: Programs for Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1993.

[187] Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research. 2012 Feb;13:281–305.

[188] Dollár P. Piotr's Computer Vision Matlab Toolbox (PMT, v3.40); 2014. http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html.

[189] Shu X, Wu XJ. A novel contour descriptor for 2D shape matching and its application to image retrieval. Image and Vision Computing. 2011 Mar;29(4):286 – 294.

[190] Vedaldi A, Fulkerson B. VLFeat: An Open and Portable Library of Computer Vision Algorithms; 2008. http://www.vlfeat.org/.

[191] Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, et al. A method for normalizing histology slides for quantitative analysis. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2009. p. 1107–1110.

[192] Khan AM, Rajpoot N, Treanor D, Magee D. A Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using Image-Specific Color Deconvolution. IEEE Transactions on Biomedical Engineering. 2014 Jun;61(6):1729–1738.

[193] Schmitt O, Hasse M. Morphological multiscale decomposition of connected regions with emphasis on cell clusters. Computer Vision and Image Understanding. 2009 Feb;113(2):188–201.

[194] Kong H, Gurcan M, Belkacem-Boussaid K. Partitioning Histopathological Images: An Integrated Framework for Supervised Color-Texture Segmentation and Cell Splitting. IEEE Transactions on Medical Imaging. 2011 Sep;30(9):1661–1677.

[195] Kong H, Gurcan M, Belkacem-Boussaid K. Splitting touching-cell clusters on histopathological images. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2011. p. 208–211.

[196] Sommer C, Fiaschi L, Hamprecht FA, Gerlich DW. Learning-based mitotic cell detection in histopathological images. In: IEEE International Conference on Pattern Recognition; 2012. p. 2306–2309.

[197] Haykin S. Neural Networks - A Comprehensive Foundation. Cambridge, London: Pearson; 1999.

[198] Koestinger M, Wohlhart P, Roth PM, Bischof H. Robust Face Detection by Simple Means. In: Computer Vision in Applications Workshop (DAGM); 2012. p. 1–2.

[199] Marée R, Geurts P, Piater J, Wehenkel L. Random subwindows for robust image classification. In: IEEE Conference on Computer Vision and Pattern Recognition. vol. 1; 2005. p. 34–40 vol. 1.

[200] Micheli-Tzanakou E, Sheikh H, Zhu B. Neural Networks and Blood Cell Identification. Journal of Medical Systems. 1997;21(4):201–210.

[201] Bikhet SF, Darwish AM, Tolba HA, Shaheen SI. Segmentation and classification of white blood cells. In: Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. vol. 6; 2000. p. 2259–2261 vol.4.

[202] Sabino DMU, Costa LF, Rizzatti EG, Zago MA. Toward leukocyte recognition using morphometry, texture and color. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro. vol. 1; 2004. p. 121–124.

[203] Shitong W, Min W. A new detection algorithm (NDA) based on fuzzy cellular neural networks for white blood cell detection. IEEE Transactions on Information Technology in Biomedicine. 2006 Jan;10(1):5–10.

[204] Khashman A. IBCIS: Intelligent blood cell identification system. Progress in Natural Science. 2008 Oct;18(10):1309 – 1314.

[205] Tai WL, Hu RM, Hsiao HCW, Chen RM, Tsai JJP. Blood Cell Image Classification Based on Hierarchical SVM. In: IEEE International Symposium on Multimedia; 2011. p. 129–136.

[206] Ramesh N, Dangott B, Salama ME, Tasdizen T. Isolation and two-step classification of normal white blood cells in peripheral blood smears. Journal of Pathology Informatics. 2012 16 Mar;3:13.

[207] Venkatalakshmi B, Thilagavathi K. Automatic red blood cell counting using hough transform. In: IEEE Conference on Information Communication Technologies; 2013. p. 267–271.

[208] Theera-Umpon N, Gader PD. Training neural networks to count white blood cells via a minimum counting error objective function. In: Proceedings of the 15th International Conference on Pattern Recognition. vol. 2; 2000. p. 299–302 vol.2.

[209] Ongun G, Halici U, Leblebicioglu K, Atalay V, Beksac M, Beksac S. Feature extraction and classification of blood cells for an automated differential blood count system. In: IEEE International Joint Conference on Neural Networks. vol. 4; 2001. p. 2461–2466.

[210] Markiewicz T, Osowski S, Marianska B, Moszczynski L. Automatic recognition of the blood cells of myelogenous leukemia using SVM. In: IEEE International Joint Conference on Neural Networks. vol. 4; 2005. p. 2496–2501.

[211] Theera-Umpon N. White Blood Cell Segmentation and Classification in Microscopic Bone Marrow Images. In: Wang L, Jin Y, editors. Fuzzy Systems and Knowledge Discovery. vol. 3614 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2005. p. 787–796.

[212] Theera-Umpon N, Dhompongsa S. Morphological Granulometric Features of Nucleus in Automatic Bone Marrow White Blood Cell Classification. IEEE Transactions on Information Technology in Biomedicine. 2007 May;11(3):353–359.

[213] Zheng X, Zhang Y, Shi J, Yu Y. Analysis of leukemia development based on marrow cell images. In: IEEE International Congress on Image and Signal Processing. vol. 1; 2011. p. 95–99.

[214] Zheng X, Zhang Y, Shi J, Yu Y. A new method for automatic counting of marrow cells. In: IEEE International Conference on Biomedical Engineering and Informatics. vol. 1; 2011. p. 42–46.

[215] Escalante HJ, Montes-y-Gómez M, González JA, Gómez-Gil P, Robles LA, García CAR, et al. Acute leukemia classification by ensemble particle swarm model selection. Artificial Intelligence in Medicine. 2012;55(3):163–175.

[216] Staroszczyk T, Osowski S, Markiewicz T. Comparative Analysis of Feature Selection Methods for Blood Cell Recognition in Leukemia. In: Perner P, editor. Machine Learning and Data Mining in Pattern Recognition. vol. 7376 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2012. p. 467–481.

[217] Hengen H, Spoor SL, Pandit MC. Analysis of blood and bone marrow smears using digital image processing techniques. Proceedings of SPIE. 2002 May;4684:624–635.

[218] Sjöström PJ, Frydel BR, Wahlberg LU. Artificial Neural Network-Aided Image Analysis System for Cell Counting. Cytometry. 1999 Jan;36(1):18–26.

[219] Jain AK, Duin RPW, Mao J. Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000 Jan;22(1):4–37.

[220] Lin W, Xiao J, Micheli-Tzanakou E. A computational intelligence system for cell classification. In: Proceedings of the 1998 IEEE international conference on information technology applications in biomedicine; 1998. p. 105–109.

[221] Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.

[222] Wu Q, Zeng L, Ke H, Xie W, Zheng H, Zhang Y. Analysis of blood and bone marrow smears using multispectral imaging analysis techniques. Proceedings of SPIE. 2005;5747:1872–1882.

[223] Mohapatra S, Patra D, Satpathy S. An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images. Neural Computing and Applications. 2014 Jun;24(7-8):1887–1904.

[224] Gençtav A, Aksoy S, Önder S. Unsupervised segmentation and classification of cervical cell images. Pattern Recognition. 2012;45(12):4151–4168.

[225] Zheng Q, Milthorpe BK, Jones AS. Direct Neural Network Application for Automated cell recognition. Cytometry Part A. 2004 Dec;57A(1):1–9.

[226] Jaeger H. The "echo state" approach to analysing and training recurrent neural networks - with an Erratum note. German National Research Center for Information Technology; 2001. GMD Report 148.

[227] Maass W, Natschläger T, Markram H. Real-time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations. Neural Computation. 2002;14(11):2531–2560.

[228] Woodward A, Ikegami T. A reservoir computing approach to image classification using coupled echo state and back-propagation neural networks. In: International Conference Image and Vision Computing. Auckland, New Zealand; 2011. p. 543–458.

[229] Lempitsky V, Verhoek M, Noble JA, Blake A. Random Forest Classification for Automatic Delineation of Myocardium in Real-Time 3D Echocardiography. In: Ayache N, Delingette H, Sermesant M, editors. Functional Imaging and Modeling of the Heart: 5th International Conference. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009. p. 447–456.

[230] Giuly RJ, Martone ME, Ellisman MH. Method: automatic segmentation of mitochondria utilizing patch classification, contour pair classification, and automatically seeded level sets. BMC Bioinformatics. 2012;13(1):1–12.

[231] Xie Z, Gillies DF. Patch forest: a hybrid framework of random forest and patch-based segmentation. Proceedings of SPIE. 2016;9784:978428–978428–8.

[232] Lukoševičius M, Jaeger H. Reservoir computing approaches to recurrent neural network training. Computer Science Review. 2009;3(3):127–149.

[233] Lukoševičius M. A Practical Guide to Applying Echo State Networks. In: Montavon G, Orr GB, Müller KR, editors. Neural Networks: Tricks of the Trade. vol. 7700 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2012. p. 659–686.

[234] Verstraeten D, Dambre J, Dutoit X, Schrauwen B. Memory versus non-linearity in reservoirs. In: International Joint Conference on Neural Networks (IJCNN); 2010. p. 1–8.

[235] Bertsimas D, Tsitsiklis J. Simulated Annealing. Statistical Science. 1993 Feb;8(1):10–15.

[236] Steil JJ. Online reservoir adaptation by intrinsic plasticity for backpropagation-decorrelation and echo state learning. Neural Networks. 2007;20(3):353–364.

[237] Werbos PJ. Backpropagation through time: what it does and how to do it. Proceedings of the IEEE. 1990 Oct;78(10):1550–1560.

[238] Torralba A, Russell BC, Yuen J. LabelMe: Online image annotation and applications. Proceedings of the IEEE. 2010;98(8):1467–1484.

[239] Fröhlich B, Rodner E, Denzler J. A Fast Approach for Pixelwise Labeling of Facade Images. In: IEEE International Conference on Pattern Recognition; 2010. p. 3029–3032.

[240] Jampani V, Gadde R, Gehler PV. Efficient Facade Segmentation Using Auto-context. In: IEEE Winter Conference on Applications of Computer Vision (WACV); 2015. p. 1038–1045.

[241] Rashtchian C, Young P, Hodosh M, Hockenmaier J. Collecting Image Annotations Using Amazon's Mechanical Turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. CSLDAMT '10. Stroudsburg, PA, USA: Association for Computational Linguistics; 2010. p. 139–147.

[242] Marée R, Stevens B, Rollus L, Rocks N, Lopez X, Salmon I, et al. A rich internet application for remote visualization and collaborative annotation of digital slides in histology and cytology. Diagnostic Pathology. 2013 30 Sep;8(Suppl 1):S26.

[243] Schulter S, Leistner C, Roth PM, Bischof H, Van Gool LJ. On-line Hough Forests. In: British Machine Vision Conference; 2011. p. 1–11.

[244] Pauly O. Random Forests for Medical Applications [PhD Thesis]. Technische Universität München. Munich, Germany; 2012.

[245] Zhu X. Semi-Supervised Learning Literature Survey. University of Wisconsin – Madison; 2006. [Last modified December 14, 2007].

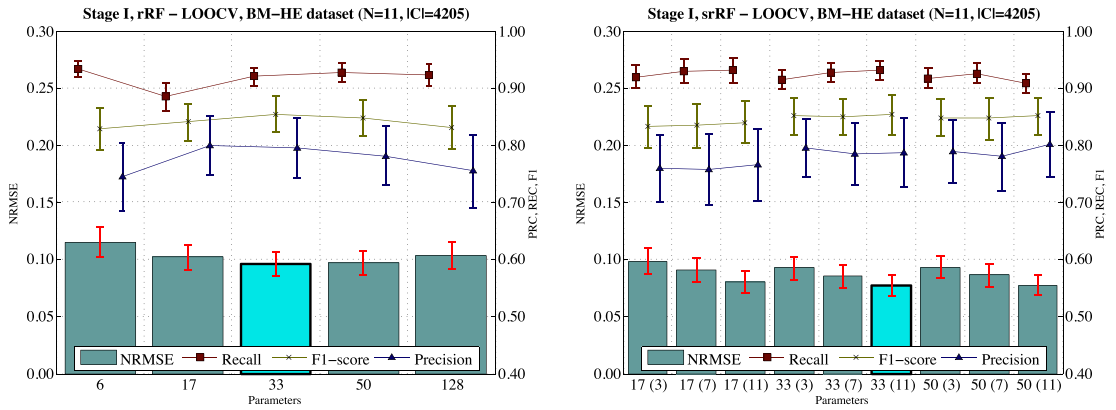
Appendix

Appendix A.

Cell Localization Results

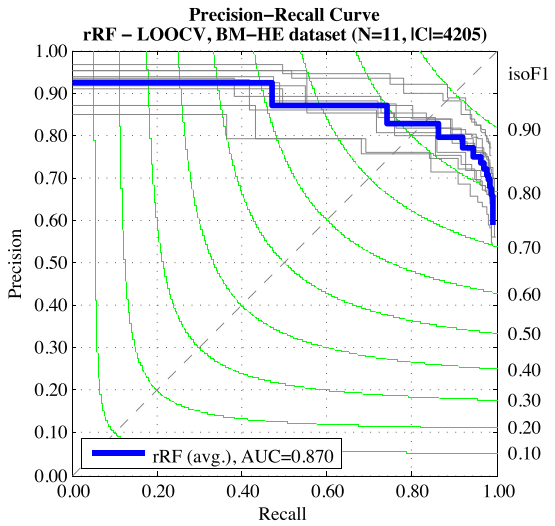
This appendix chapter contains supplementary tables and figures of the cell localization method evaluation. Figures and tables are sorted by dataset.

A.1. Bone Marrow (H&E)

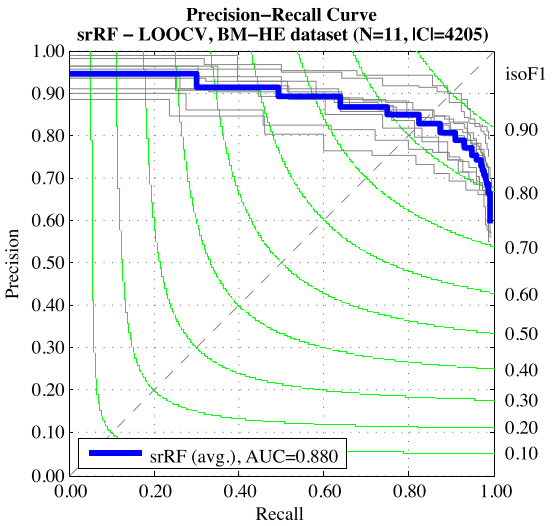


(a) *rRF*: $\kappa^* = 0.5882$

(b) *srRF*: $\kappa^* = 0.4706$

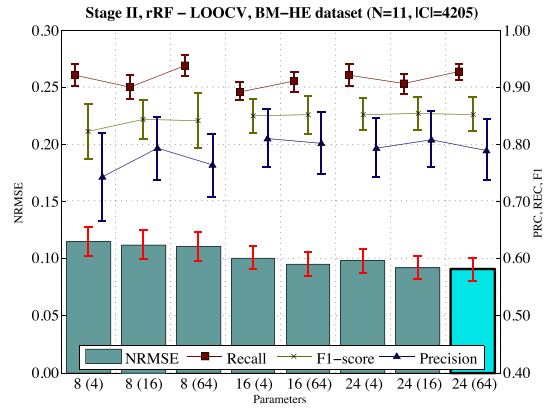


(c) *rRF*: $p^*_{in} = 33$, $p^*_{out} = 1$

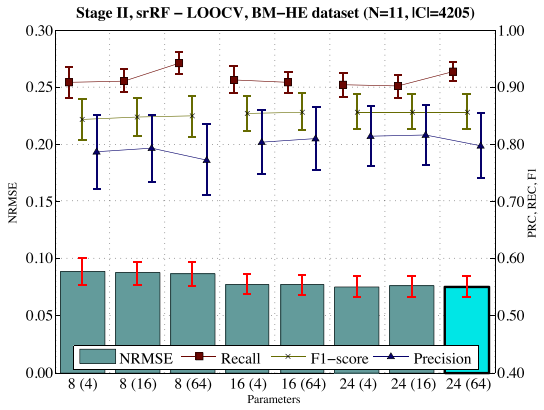


(d) *srRF*: $p^*_{in} = 33$, $p^*_{out} = 11$

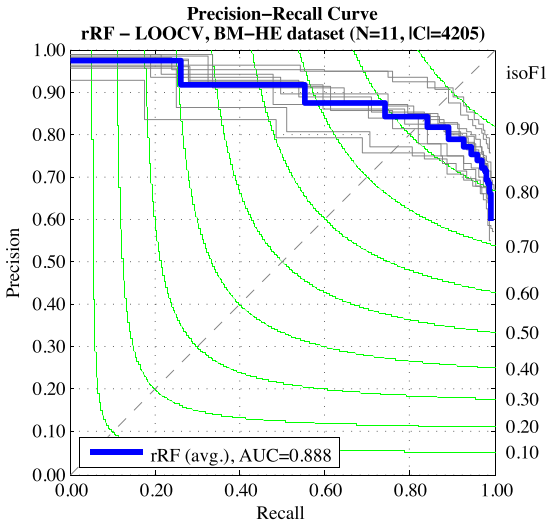
Figure A.1.: *BM-HE* dataset: stage I hyper-parameter selection results for the *rRF* and *srRF* method (a,b). Parameter values on the horizontal axis denote input patch sizes p_{in}. Additionally in (b), numbers in parentheses denote the output patch size p_{out}. The minimum NRMSE, here determining p^*_{in} and p^*_{out}, respectively, is highlighted using cyan color, and a bold border. Error bars denote the SD, κ^* is given for the best configuration. (c,d) Precision-recall curves for the best hyper-parameters. The blue line denotes the average curve over all CV runs, grey solid lines the individual CVs (i.e. images), and green solid lines the iso-contours of F1-score. (a,c) *Single-target*, (b,d) *spatial-averaging* regression forest. See Table 4.2 (p. 54) for details regarding hyper-parameter configuration for this dataset.



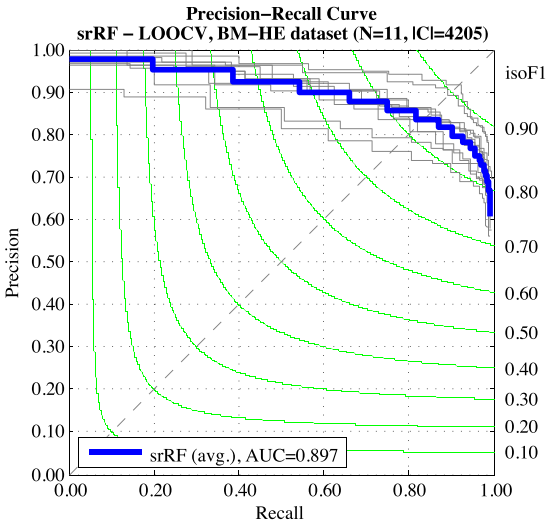
(a) rRF: $\kappa^* = 0.5686$



(b) $srRF$: $\kappa^* = 0.4706$



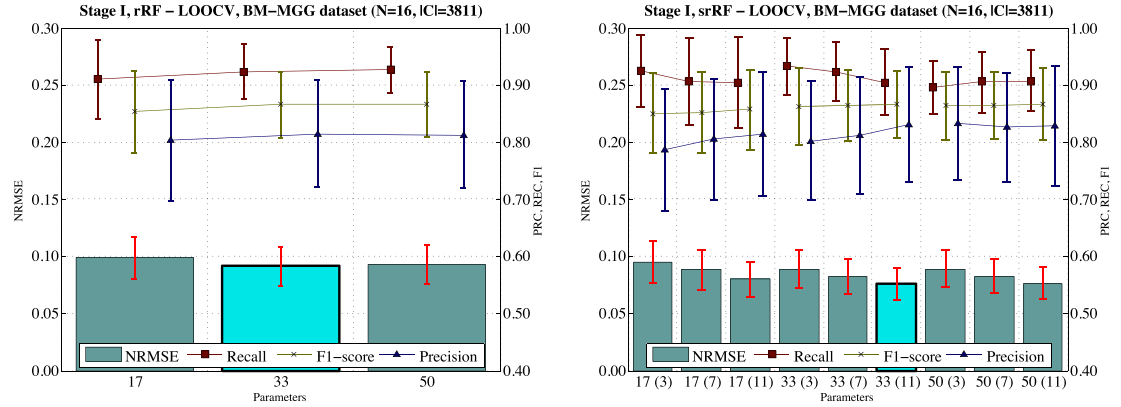
(c) rRF: $T = 64$, $T_{md} = 24$



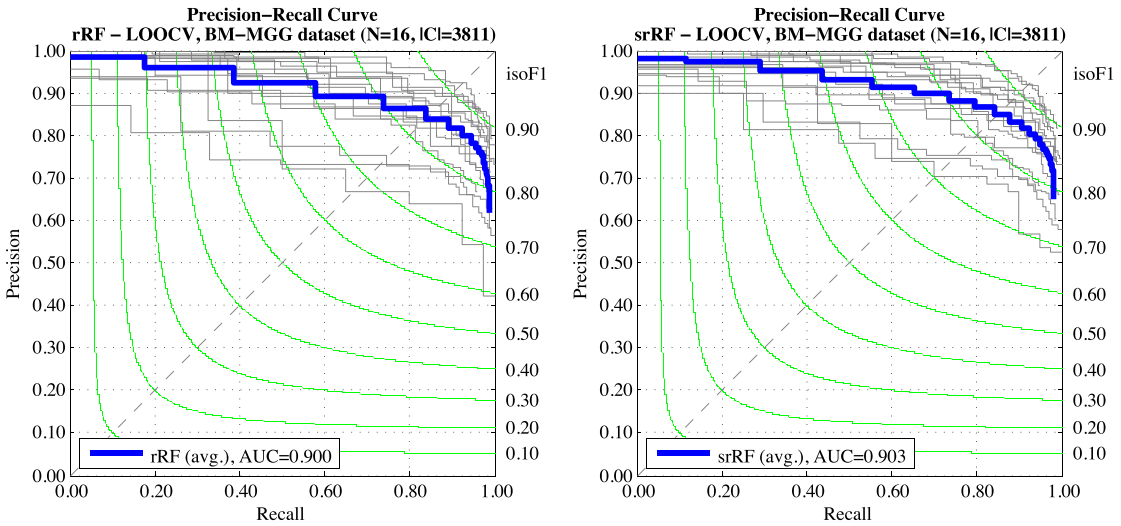
(d) $srRF$: $T = 64$, $T_{md} = 24$

Figure A.2.: *BM-HE* dataset: stage II results for the best hyper-parameters. (a,b) Parameter values on the horizontal axis denote T_{md} (T). The minimum NRMSE, here determining optimal model complexity, is highlighted using cyan color, and a bold border. Error bars denote the SD. (c,d) The blue line denotes the average curve over all CV runs, grey solid lines the individual CVs (i.e. images), and green solid lines the iso-contours of F1-score.

A.2. Bone Marrow (MGG)

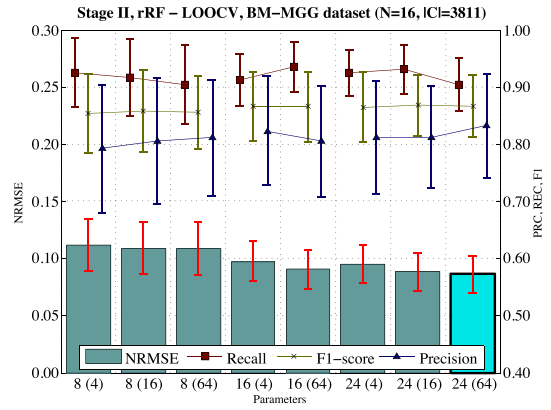


(a) rRF: $\kappa^* = 0.5686$ (b) $srRF$: $\kappa^* = 0.4706$

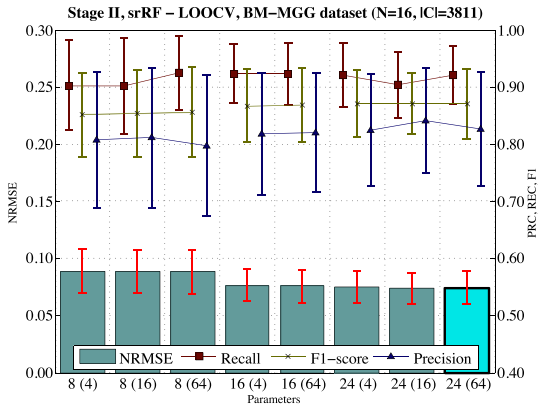


(c) rRF: $p^*_{in} = 33$, $p^*_{out} = 1$ (d) $srRF$: $p^*_{in} = 33$, $p^*_{out} = 11$

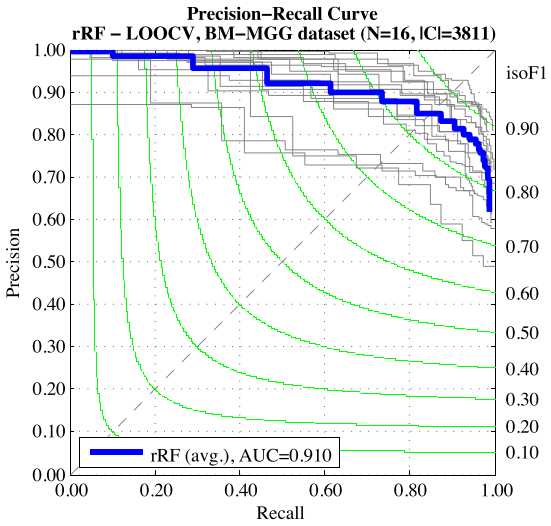
Figure A.3.: *BM-MGG* dataset: stage I hyper-parameter selection results for the *rRF* and *srRF* method (a,b). Parameter values on the horizontal axis denote input patch sizes p_{in}. Additionally in (b), numbers in parentheses denote the output patch size p_{out}. The minimum NRMSE, here determining p^*_{in} and p^*_{out}, respectively, is highlighted using cyan color, and a bold border. Error bars denote the SD, κ^* is given for the best configuration. (c,d) Precision-recall curves for the best hyper-parameters. The blue line denotes the average curve over all CV runs, grey solid lines the individual CVs (i.e. images), and green solid lines the iso-contours of F1-score. (a,c) *Single-target*, (b,d) *spatial-averaging* regression forest. See Table 4.2 (p. 54) for details regarding hyper-parameter configuration for this dataset.



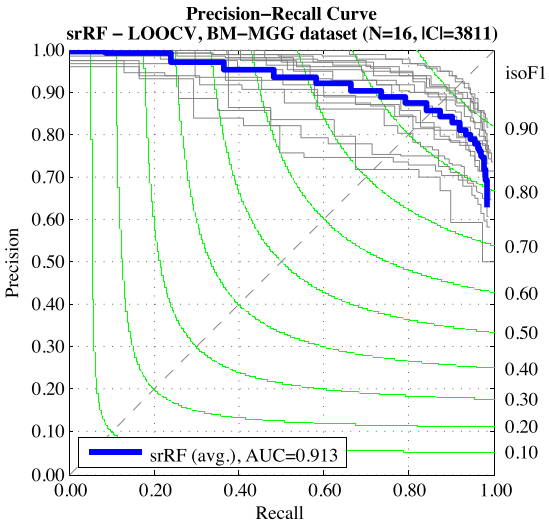
(a) rRF: $\kappa^* = 0.5686$



(b) $srRF$: $\kappa^* = 0.4510$



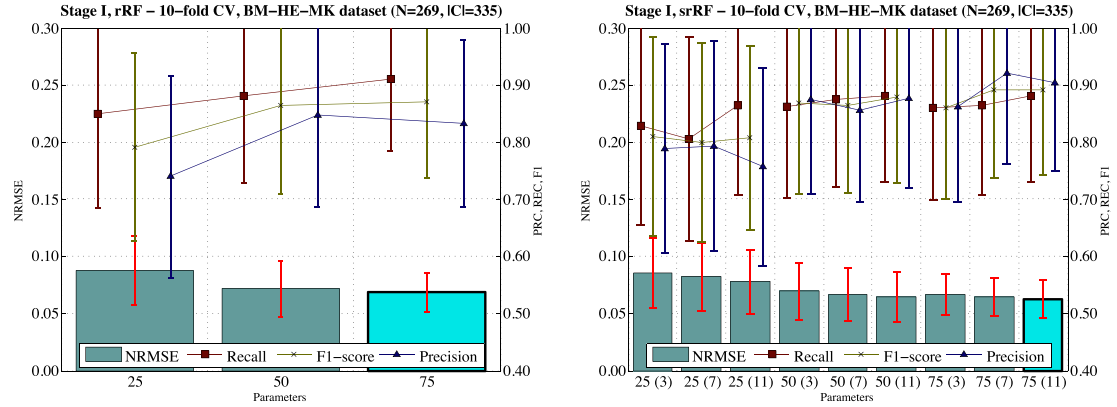
(c) rRF: $T = 64$, $T_{md} = 24$



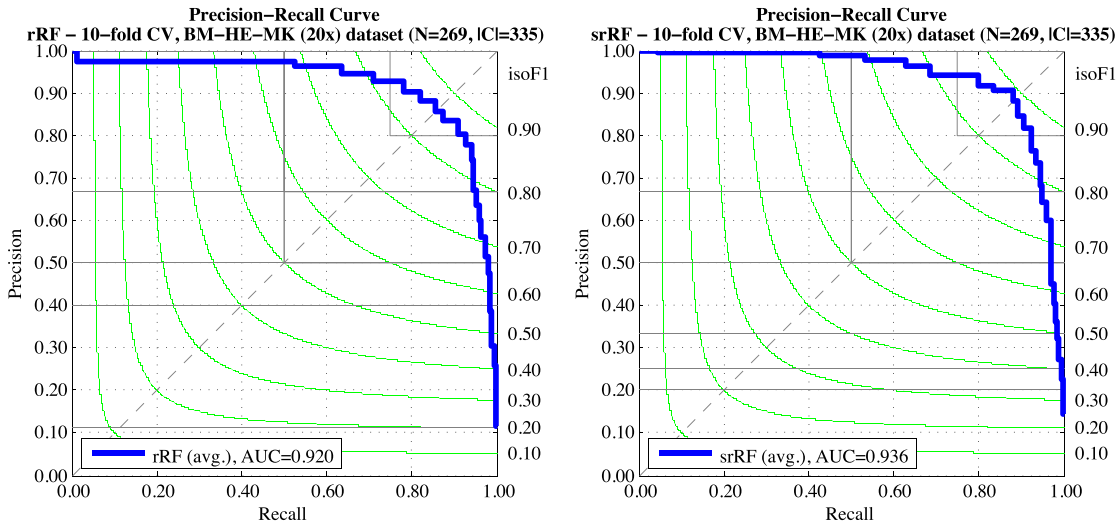
(d) $srRF$: $T = 64$, $T_{md} = 24$

Figure A.4.: *BM-MGG* dataset: stage II results for the best hyper-parameters. (a,b) Parameter values on the horizontal axis denote T_{md} (T). The minimum NRMSE, here determining optimal model complexity, is highlighted using cyan color, and a bold border. Error bars denote the SD. (c,d) The blue line denotes the average curve over all CV runs, grey solid lines the individual CVs (i.e. images), and green solid lines the iso-contours of F1-score.

A.3. Bone Marrow (H&E) Megakaryocytes

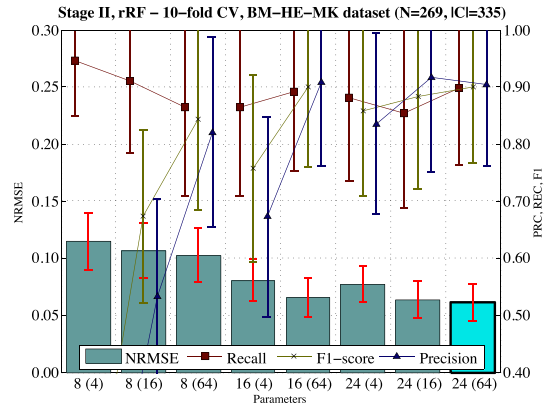


(a) rRF: $\kappa^* = 0.4706$ (b) $srRF$: $\kappa^* = 0.4510$

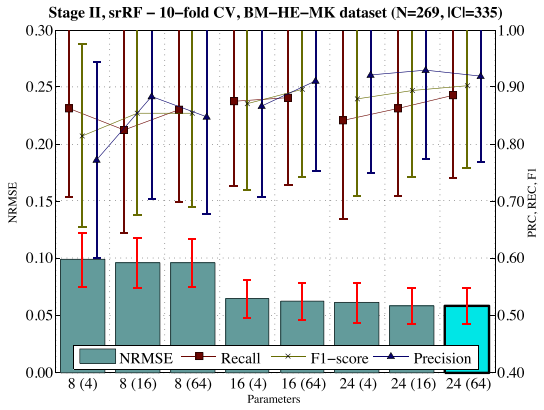


(c) rRF: $p^*_{in} = 75$, $p^*_{out} = 1$ (d) $srRF$: $p^*_{in} = 75$, $p^*_{out} = 11$

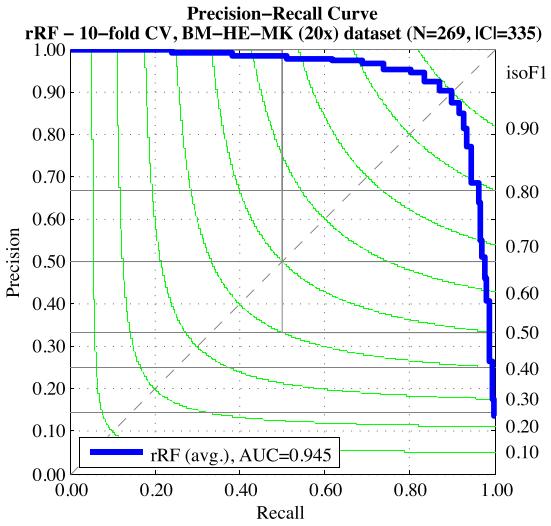
Figure A.5.: *BM-HE-MK* dataset (20× magnification): stage I hyper-parameter selection results for the *rRF* and *srRF* method (a,b). Parameter values on the horizontal axis denote input patch sizes p_{in}. Additionally in (b), numbers in parentheses denote the output patch size p_{out}. The minimum NRMSE, here determining p^*_{in} and p^*_{out}, respectively, is highlighted using cyan color, and a bold border. Error bars for PRC, REC, and F1 denote 0.5·SD (for illustration purpose only), while NRMSE error bars denote SD. κ^* is given for the best configuration. The large SD is explained by the fact that most of the images contained only a single megakaryocyte. (c,d) Precision-recall curves for the best hyper-parameters. The blue line denotes the average curve over all CV runs, grey solid lines the individual images, and green solid lines the iso-contours of F1-score. (a,c) *Single-target*, (b,d) *spatial-averaging* regression forest. See Table 4.2 (p. 54) for details regarding hyper-parameter configuration for this dataset.



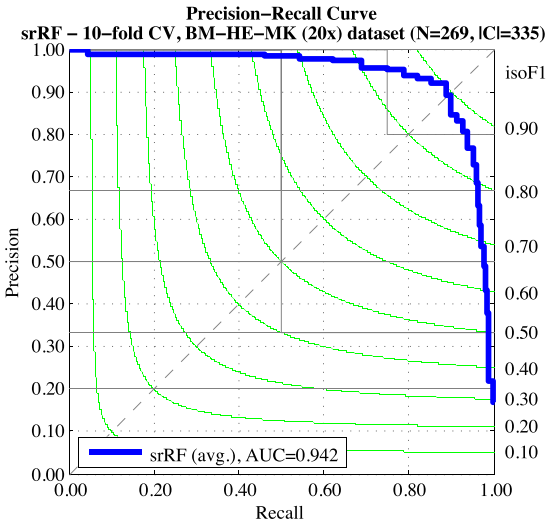
(a) rRF: $\kappa^* = 0.4314$



(b) $srRF$: $\kappa^* = 0.4118$

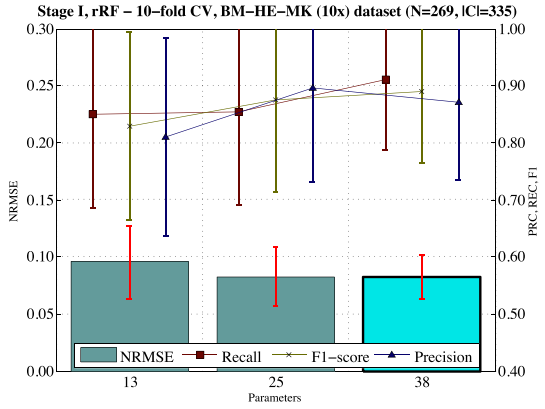


(c) rRF: $T = 64$, $T_{md} = 24$

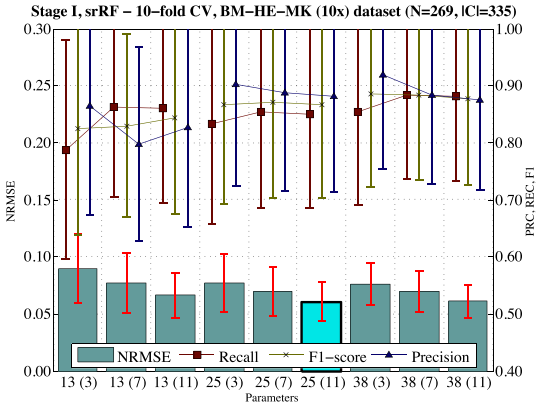


(d) $srRF$: $T = 64$, $T_{md} = 24$

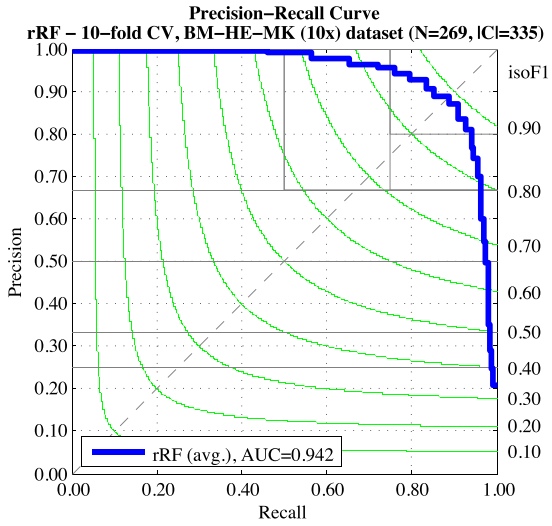
Figure A.6.: *BM-HE-MK* dataset (20× magnification): stage II results for the best hyper-parameters. (a,b) Parameter values on the horizontal axis denote T_{md} (T). The minimum NRMSE, here determining optimal model complexity, is highlighted using cyan color, and a bold border. Error bars for PRC, REC, and F1 denote 0.5·SD (for illustration purpose only), while NRMSE error bars denote SD. The large SD is explained by the fact that most of the images contained only a single megakaryocyte. (c,d) The blue line denotes the average curve over all CV runs, grey solid lines the individual images, and green solid lines the iso-contours of F1-score.



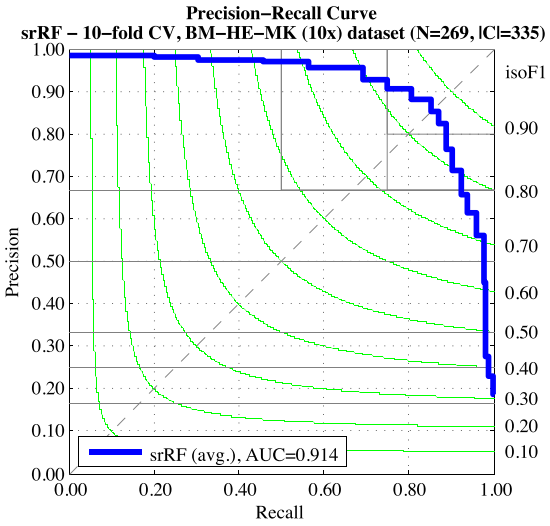
(a) *rRF*: $\kappa^* = 0.4510$



(b) *srRF*: $\kappa^* = 0.3529$

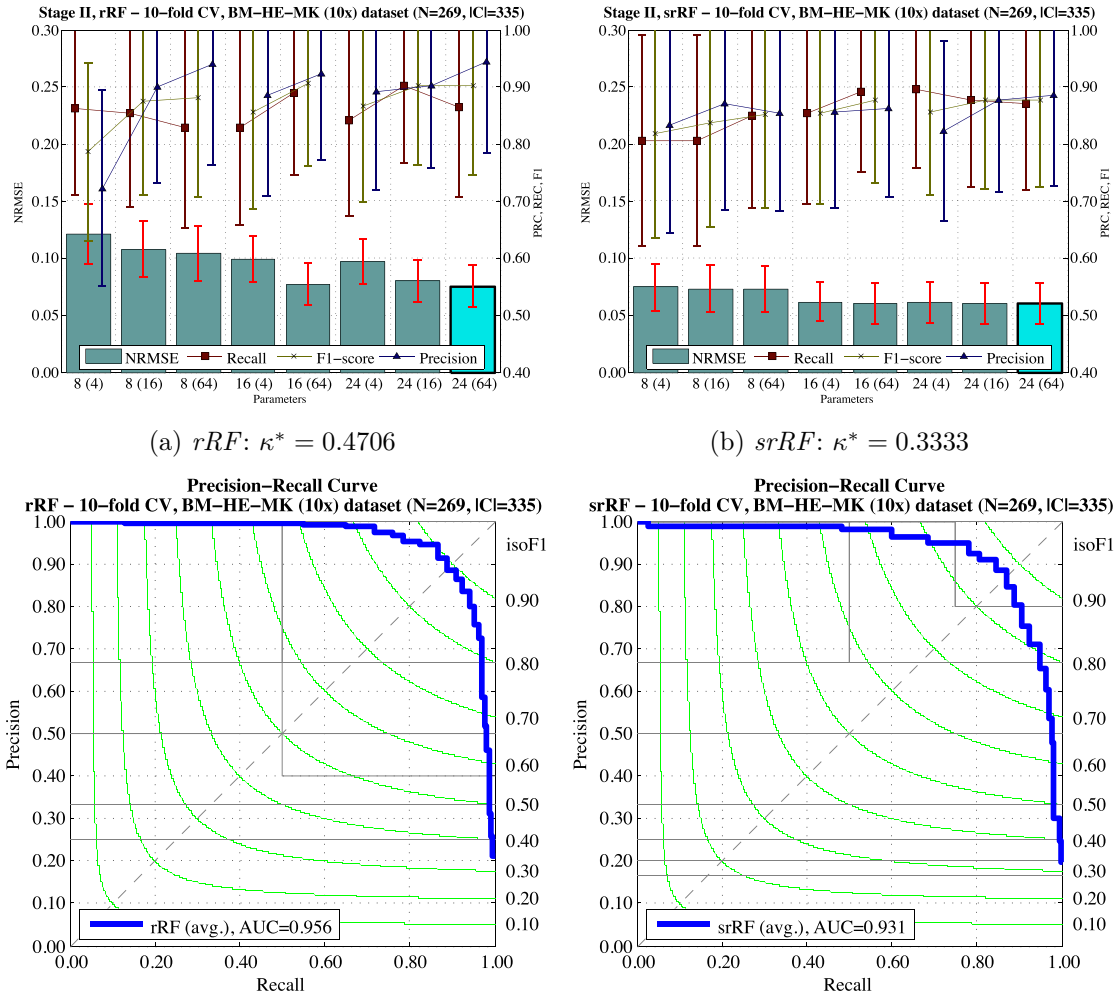


(c) *rRF*: $p^*_{in} = 38$, $p^*_{out} = 1$



(d) *srRF*: $p^*_{in} = 25$, $p^*_{out} = 11$

Figure A.7.: *BM-HE-MK* dataset (10× magnification): stage I hyper-parameter selection results for the *rRF* and *srRF* method (a,b). Parameter values on the horizontal axis denote input patch sizes p_{in}. Additionally in (b), numbers in parentheses denote the output patch size p_{out}. The minimum NRMSE, here determining p^*_{in} and p^*_{out}, respectively, is highlighted using cyan color, and a bold border. Error bars for PRC, REC, and F1 denote 0.5·SD (for illustration purpose only), while NRMSE error bars denote SD. κ^* is given for the best configuration. The large SD is explained by the fact that most of the images contained only a single megakaryocyte. (c,d) Precision-recall curves for the best hyper-parameters. The blue line denotes the average curve over all CV runs, grey solid lines the individual images, and green solid lines the iso-contours of F1-score. (a,c) *Single-target*, (b,d) *spatial-averaging* regression forest. See Table 4.2 (p. 54) for details regarding hyper-parameter configuration for this dataset.



(a) rRF: $\kappa^* = 0.4706$

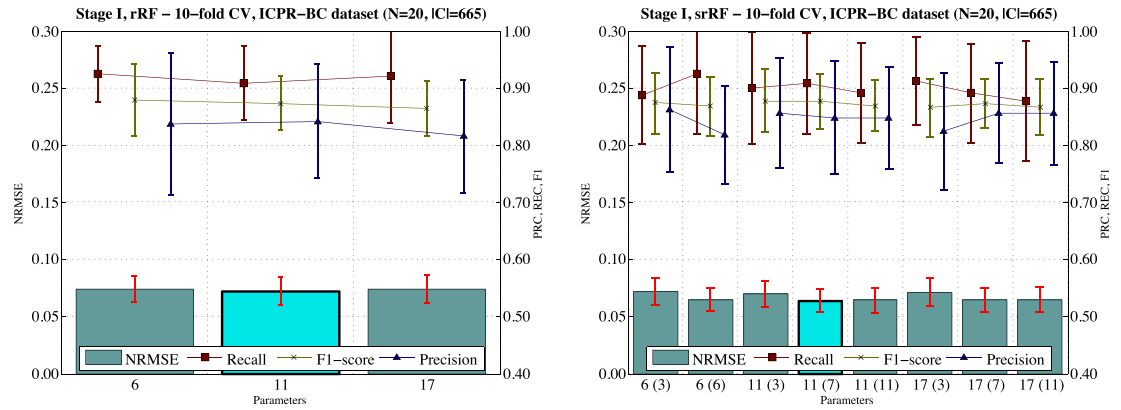
(b) $srRF$: $\kappa^* = 0.3333$

(c) rRF: $T = 64$, $T_{md} = 24$

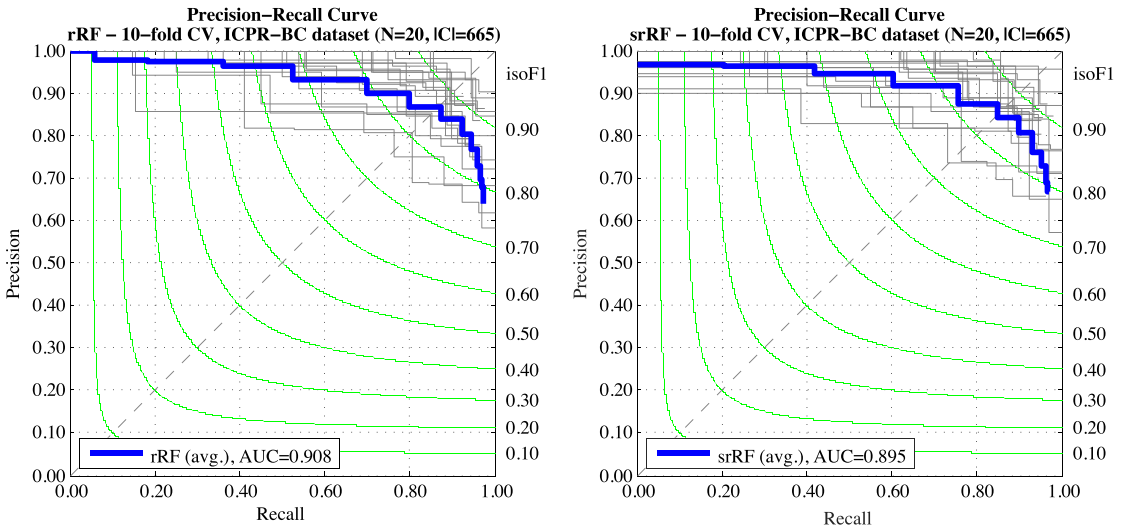
(d) $srRF$: $T = 64$, $T_{md} = 24$

Figure A.8.: *BM-HE-MK* dataset (10× magnification): stage II results for the best hyper-parameters. (a,b) Parameter values on the horizontal axis denote T_{md} (T). The minimum NRMSE, here determining optimal model complexity, is highlighted using cyan color, and a bold border. Error bars for PRC, REC, and F1 denote 0.5·SD (for illustration purpose only), while NRMSE error bars denote SD. The large SD is explained by the fact that most of the images contained only a single megakaryocyte. (c,d) The blue line denotes the average curve over all CV runs, grey solid lines the individual images, and green solid lines the iso-contours of F1-score.

A.4. Breast Cancer (H&E)

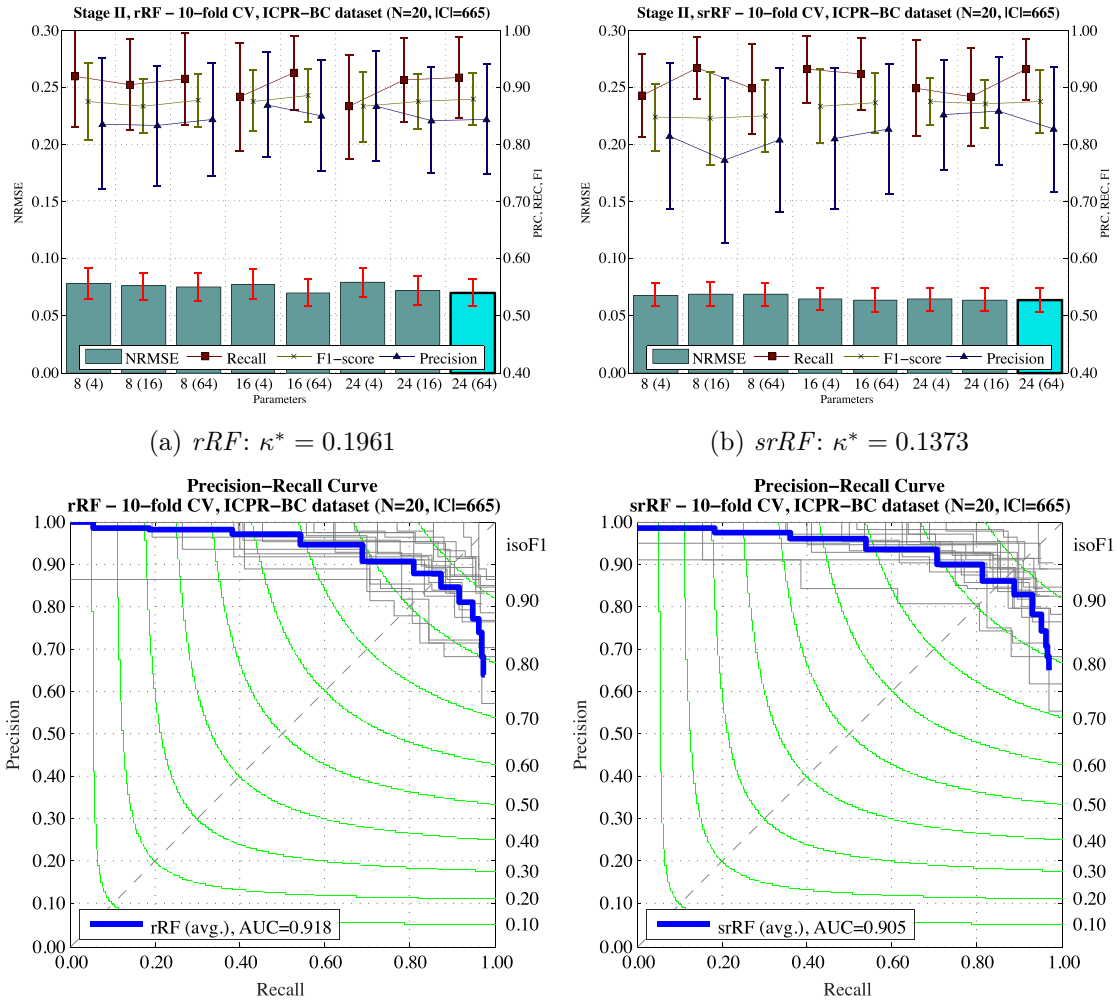


(a) *rRF*: $\kappa^* = 0.1961$ (b) *srRF*: $\kappa^* = 0.1569$



(c) *rRF*: $p^*_{in} = 11$, $p^*_{out} = 1$ (d) *srRF*: $p^*_{in} = 11$, $p^*_{out} = 7$

Figure A.9.: *ICPR-BC* dataset: stage I hyper-parameter selection results for the *rRF* and *srRF* method (a,b). Parameter values on the horizontal axis denote input patch sizes p_{in}. Additionally in (b), numbers in parentheses denote the output patch size p_{out}. The minimum NRMSE, here determining p^*_{in} and p^*_{out}, respectively, is highlighted using cyan color, and a bold border. Error bars denote the SD. (c,d) Precision-recall curves for the best hyper-parameters. The blue line denotes the average curve over all CV runs, grey solid lines the individual images, and green solid lines the iso-contours of F1-score. (a,c) *Single-target*, (b,d) *spatial-averaging* regression forest. See Table 4.2 (p. 54) for details regarding hyper-parameter configuration for this dataset.



(a) rRF: $\kappa^* = 0.1961$

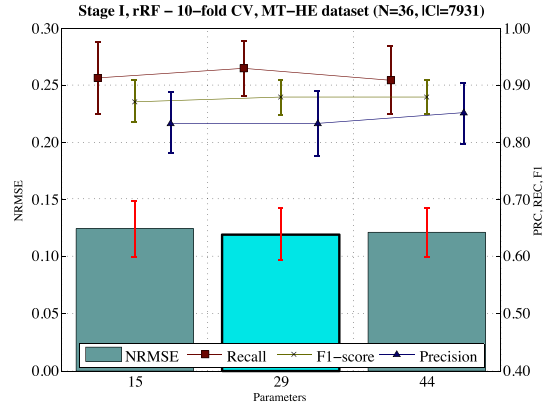
(b) $srRF$: $\kappa^* = 0.1373$

(c) rRF: $T = 64$, $T_{md} = 24$

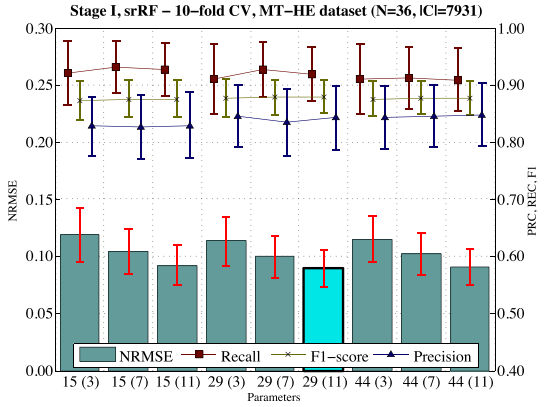
(d) $srRF$: $T = 64$, $T_{md} = 24$

Figure A.10.: *ICPR-BC* dataset: stage II results for the best hyper-parameters. (a,b) Parameter values on the horizontal axis denote T_{md} (T). The minimum NRMSE, here determining optimal model complexity, is highlighted using cyan color, and a bold border. Error bars denote the SD. (c,d) The blue line denotes the average curve over all CV runs, grey solid lines the individual images, and green solid lines the iso-contours of F1-score.

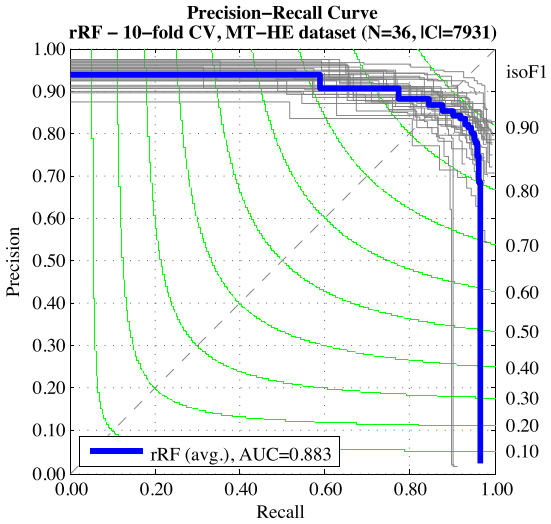
A.5. Multi-Tissue (H&E)



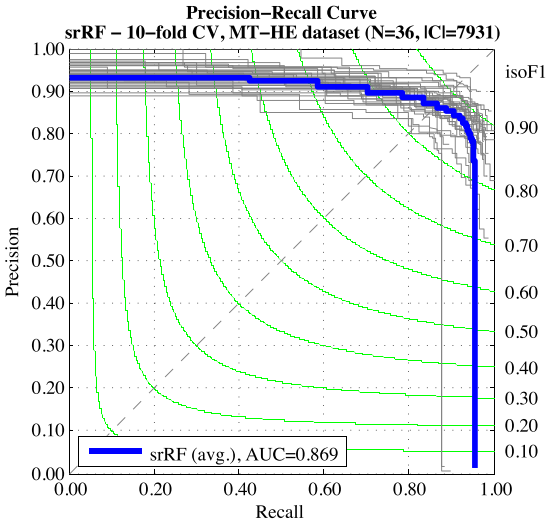
(a) *rRF*: $\kappa^* = 0.5098$



(b) *srRF*: $\kappa^* = 0.3529$

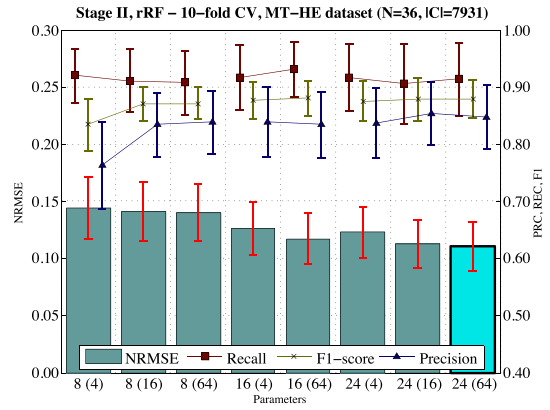


(c) *rRF*: $p^*_{in} = 29$, $p^*_{out} = 1$

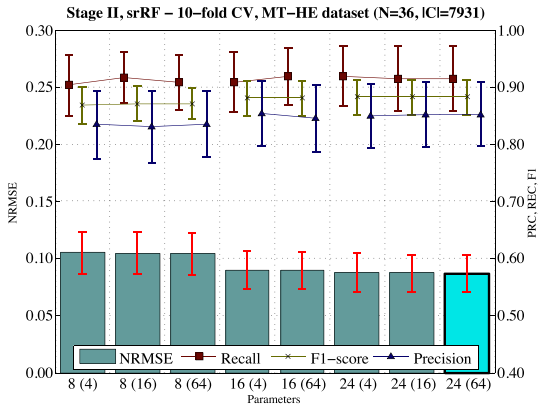


(d) *srRF*: $p^*_{in} = 29$, $p^*_{out} = 11$

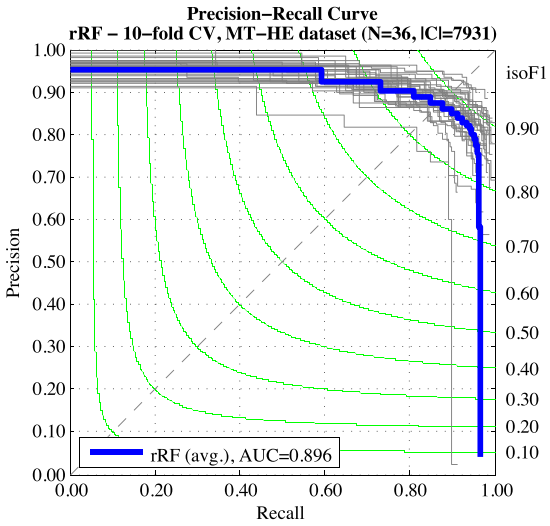
Figure A.11.: *MT-HE* dataset: stage I hyper-parameter selection results for the *rRF* and *srRF* method (a,b). Parameter values on the horizontal axis denote input patch sizes p_{in}. Additionally in (b), numbers in parentheses denote the output patch size p_{out}. The minimum NRMSE, here determining p^*_{in} and p^*_{out}, respectively, is highlighted using cyan color, and a bold border. Error bars denote the SD. (c,d) Precision-recall curves for the best hyper-parameters. The blue line denotes the average curve over all CV runs, grey solid lines the individual images, and green solid lines the iso-contours of F1-score. (a,c) *Single-target*, (b,d) *spatial-averaging* regression forest. See Table 4.2 (p. 54) for details regarding hyper-parameter configuration for this dataset.



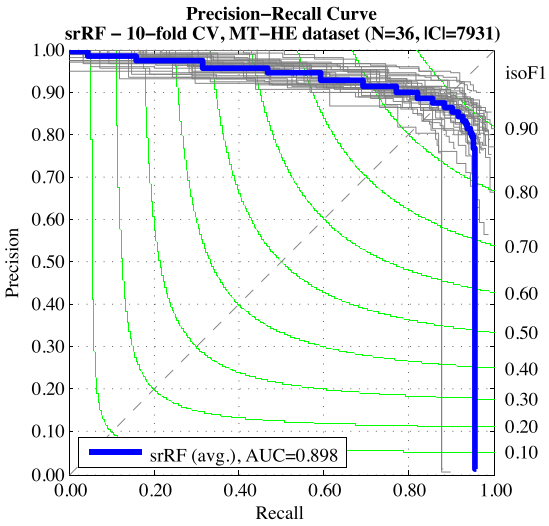
(a) rRF: $\kappa^* = 0.5098$



(b) $srRF$: $\kappa^* = 0.3529$



(c) rRF: $T = 64$, $T_{md} = 24$



(d) $srRF$: $T = 64$, $T_{md} = 24$

Figure A.12.: *MT-HE* dataset: stage II results for the best hyper-parameters. (a,b) Parameter values on the horizontal axis denote T_{md} (T). The minimum NRMSE, here determining optimal model complexity, is highlighted using cyan color, and a bold border. Error bars denote the SD. (c,d) The blue line denotes the average curve over all CV runs, grey solid lines the individual images, and green solid lines the iso-contours of F1-score.

Appendix B.

Cytomine IRIS Labeling Platform

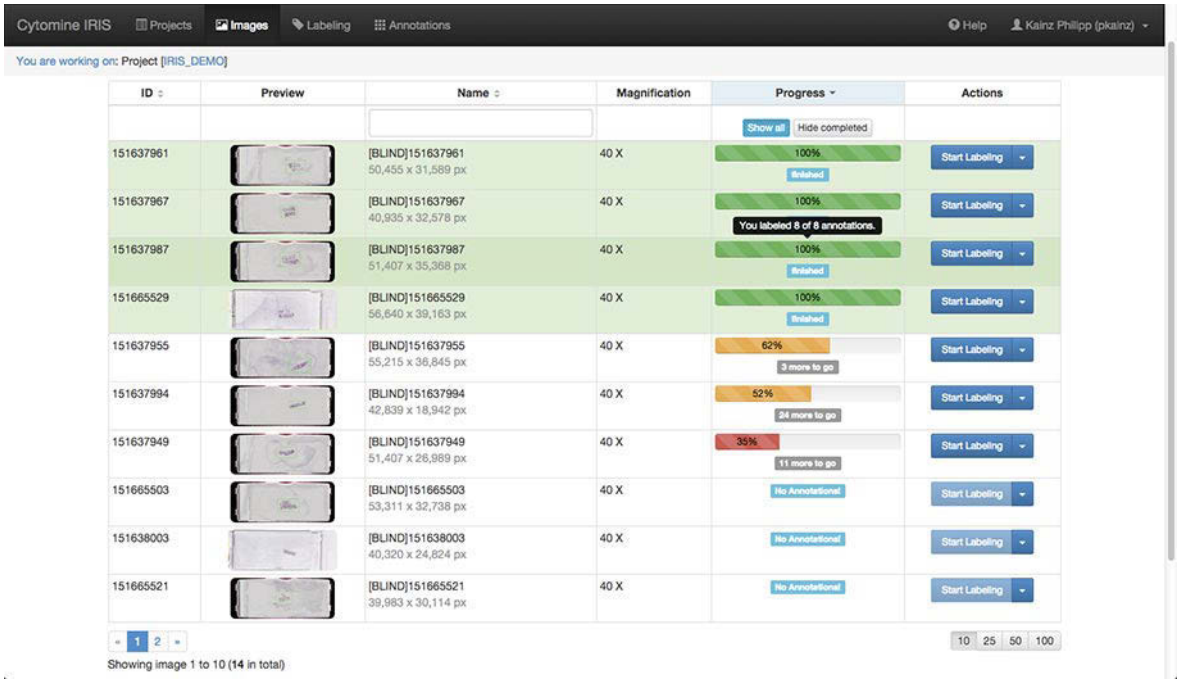
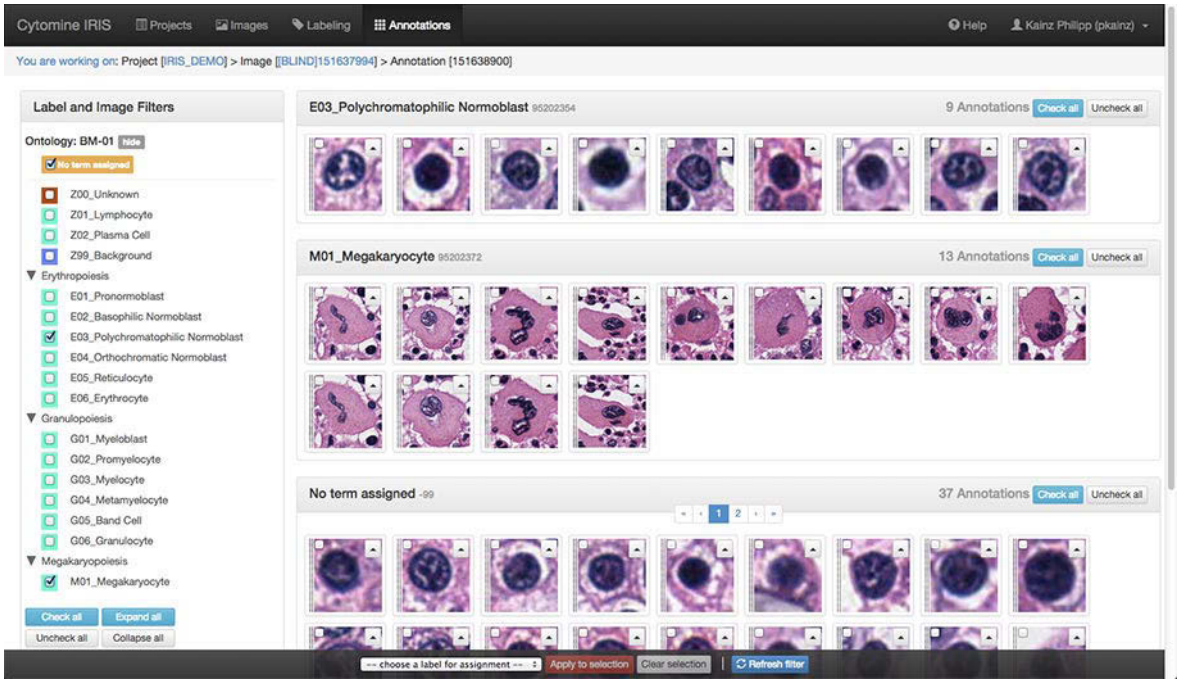
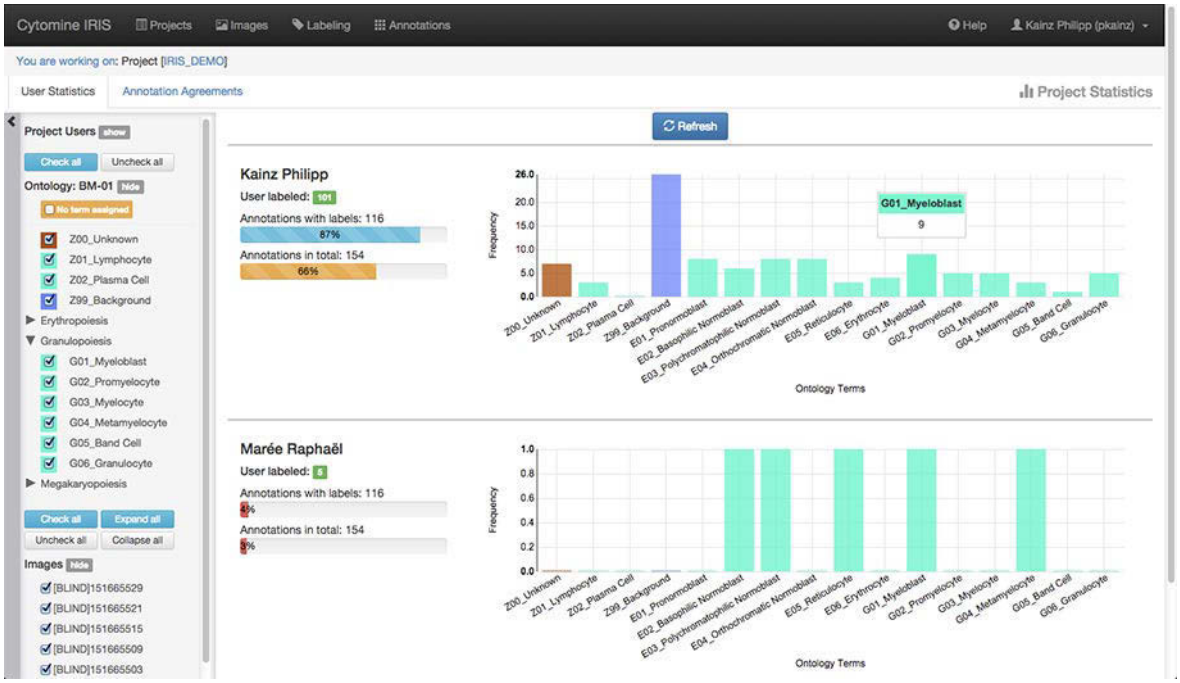


Figure B.1.: Interface of the labeling progress in individual whole slide images. Tracking the progress is essential for labeling a huge number of annotations.

Marée R, Rollus L, Stévens B, Hoyoux R, Louppe G, Vandaele R, et al. Collaborative analysis of multi-gigapixel imaging data using Cytomine. Bioinformatics. 2016 Jan;32(9):1395–1401, by permission of Oxford University Press.



(a)



(b)

Figure B.2.: All labeled (and yet unlabeled) annotations can be clearly reviewed in the Cytomine-IRIS gallery. (a) Corrections of label assignments can easily be performed for samples that were identified as outliers. (b) User statistics in a Cytomine-IRIS project can be viewed by authorized project coordinators. Marée R, Rollus L, Stévens B, Hoyoux R, Louppe G, Vandaele R, et al. Collaborative analysis of multi-gigapixel imaging data using Cytomine. Bioinformatics. 2016 Jan;32(9):1395–1401, by permission of Oxford University Press.