

Quantification of Nuclei in Synthetic Ki-67 Histology Images of the Breast

Image Analysis in Digital Pathology

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Biomedical Engineering

by

Michaela Weingant, BSc

Registration Number 0571302

to the Faculty of Informatics

at the Vienna University of Technology

Advisor: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Robert Sablatnig

Assistance: Matthew DiFranco, PhD

Vienna, 22nd August, 2016

Michaela Weingant

Robert Sablatnig

Zellkernzählung in Synthetischen Ki-67 Histologiebildern der Brust

Bildanalyse in der Digitalen Pathologie

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Biomedical Engineering

eingereicht von

Michaela Weingant, BSc

Matrikelnummer 0571302

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Robert Sablatnig

Mitwirkung: Matthew DiFranco, PhD

Wien, 22. August 2016

Michaela Weingant

Robert Sablatnig

Erklärung zur Verfassung der Arbeit

Michaela Weingant, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 22. August 2016

Michaela Weingant

Acknowledgements

Thank you to Prof. Sablatnig for supervising my progress on a topic of the Biomedical domain and giving me the necessary hints and background on a "proper" academic approach and its demands.

A huge acknowledgement goes to my co-supervisor Matthew DiFranco for his time, effort and encouragement and his genuine interest in my topic. His belief in me and my skills helped me considerably towards finishing my studies while giving me the chance to experience real, applied research in a multinational team with all its downs and ups.

I want to thank my family for their continuous moral and financial support and constant believe in me and my path.

My gratitude also goes towards all my friends who kept encouraging me while at the same time providing many joyful and diverse opportunities for recreation and diversion.

My conversations with Elke as my "librarian" really helped me to put all my books in their appropriate shelf and thereby allowed me to focus on my work.

Peter deserves a special shout-out for having spent many hours, day and night, listening to me, thinking with me, questioning, proof-reading and correcting my work, for the save room and time he provided by his moral and hands-on support.

Last but not least, I am deeply grateful for the motivation which Wuserl and Wurmi gave me to finish my studies and especially the patience of Wurmi who had to share me with this thesis in her first weeks with us.

Abstract

Breast cancer is a common disease and the diagnosis as well as treatment decisions are among other factors largely based on a number of examinations by the pathologist. They are traditionally conducted on tissue samples sliced into thin slides. In the last decade the field of Digital Pathology has emerged, which after scanning of the tissue slides allows the digital viewing and analysis of tissue slides formerly manually examined under the microscope. One of the exams in breast cancer diagnosis includes the quantification of the proliferative activity of nuclei, which is an indicator of tumor growth. The proliferative activity is expressed as the Labeling Index (LI, rate of dividing vs. non-dividing nuclei) and is made visible on a tissue slide with a stain called Ki-67.

The aim of the work presented is to develop and evaluate an algorithm for the automatic quantification of the Ki-67 LI on a digitized slide. The algorithm is based on the color deconvolution of the images, dividing the image into two channels: one showing the dividing (Ki-67 positive) and one showing the non-dividing (Ki-67 negative) nuclei. Three deconvolution approaches are implemented and tested. Each channel is hereafter post-processed using a pipeline of well-established image processing steps and the result is a segmentation of all nuclei found in each channel. The amount of nuclei in each channel is quantified and yields the LI for each image. No supervised training on labeled data is required prior to the image analysis.

In order to evaluate the performance of the algorithm, a fully labeled dataset of Ki-67 stained images of the breast is required. Because no benchmark dataset of this tissue and stain type is available, a synthetic dataset is built, using nuclei manually extracted from digitized clinical Ki-67 slides and a novel synthesis method, allowing the definition of varying nuclear arrangement, LIs and stain appearances. The images in the datasets generated provide detailed ground truth information.

An in-depth evaluation based on the synthetic images points out that the presented algorithm is able to estimate the LI in an image with an absolute error of 1.5%.

Kurzfassung

Brustkrebs ist eine häufige Erkrankung und sowohl die Diagnose, als auch die Entscheidung für die bestmögliche Behandlung basieren zu einem großen Teil auf einer Reihe an Untersuchungen durch Pathologen und Pathologinnen. Diese Untersuchungen werden traditionellerweise an Gewebeproben durchgeführt, die hierfür in dünne Scheiben geschnitten werden. Die im letzten Jahrzehnt aufgekommene Digitale Pathologie erlaubt es, eingescannte Gewebsschnitte digital zu betrachten und zu analysieren, die zuvor manuell unter dem Mikroskop begutachtet wurden. Eine der Untersuchungen in der Brustkrebsdiagnose berücksichtigt die Messung der Proliferationsaktivität der Zellkerne, welche einen Indikator für Tumorwachstum darstellt. Die Proliferationsaktivität wird als Labeling Index bezeichnet (LI, Verhältnis an sich teilenden vs. sich nicht teilenden Zellkernen) und kann durch eine Färbemethode namens Ki-67 auf dem Gewebsschnitt sichtbar gemacht werden.

Das Ziel der vorliegenden Arbeit ist es, einen Algorithmus zu entwickeln und zu evaluieren, der den Ki-67 LI auf einem digitalen Schnitt automatisch misst. Der Algorithmus fußt auf der Farbzerlegung der Bilder, der sie in zwei Kanäle aufteilt: einen Farbkanal, der die in Teilung befindlichen (Ki-67 positiven) und einen, der die nicht in Teilung befindlichen (Ki-67 negativen) Zellkerne zeigt. Dafür werden drei Ansätze zur Farbzerlegung implementiert und getestet. Jeder Kanal wird hiernach in mehreren Schritten weiterverarbeitet, wobei etablierte Bildverarbeitungsschritte angewandt werden, und in einer Segmentierung aller Kerne in jedem Kanal resultiert. Die Anzahl der Kerne in jedem Kanal wird gezählt und daraus ergibt sich der LI für jedes Bild. Vor der Bildanalyse sind keine überwachten Lernschritte auf annotierten Daten vonnöten.

Um das Ergebnis des Algorithmus zu evaluieren wird ein vollständig annotiertes Datenset von Ki-67 gefärbten Bildern der Brust benötigt. Da kein solches Vergleichs-Datenset dieses Gewebe- und Färbetyps existiert, wird ein synthetisches Datenset erstellt, das manuell extrahierte Zellkerne von digitalisierten, klinischen Ki-67 Schnitten verwendet und anhand einer neuartigen Synthese-Methode die Definition von variablen Kernverteilungen, LIs und Färbungseigenschaften erlaubt. Die so generierten Bilder stellen eine detaillierte Grundwahrheit dar.

Eine ausführliche, auf den synthetischen Bildern basierende Evaluierung ergibt, dass der vorgestellte Algorithmus in der Lage ist, den LI in einem Bild bis auf einen absoluten Fehler von 1.5% zu messen.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Problem Statement	2
1.3	Research Questions	3
1.4	Methodological Approach	3
1.5	Scope of Discussion	3
1.6	Structure of the Work	4
2	Pathology	5
2.1	Processing Steps prior to Analysis by Pathologist	6
2.2	Staining Protocols in Breast Cancer Diagnosis	6
2.3	Breast Cancer Diagnosis with the Bloom-Richardson Method	7
2.3.1	Proliferation Assessment via Ki-67 Scoring	8
2.3.2	Sources of Variability	9
2.3.3	Summary	10
3	Digital Pathology	11
3.1	Motives for the Advancement of Digital Pathology	11
3.2	State-of-the-Art	13
3.3	Challenges	14
3.4	Frequent Digital Image Analysis Challenges in Digital Pathology	15
3.4.1	Artifacts	15
3.4.2	Color Deconvolution	16
3.4.3	Tissue Classification	18
3.4.4	ImmunoHistoChemical (IHC) Quantification	18
3.4.5	Nuclei Detection and Segmentation	19
3.5	Digital Imaging Analysis Approaches in Ki-67 Assessment	22
3.6	Evaluation of Digital Image Analysis Solutions	25
3.6.1	Nuclei Quantification	26
3.6.2	Nuclei Segmentation	26
3.6.3	Ground Truth from Pathologists	28
3.6.4	Ground Truth from Synthetic Datasets	29
3.6.5	Summary	30

4	Suggested Solution	33
4.1	Synthetic DataSet (SDS)	33
4.1.1	Creation of Synthetic Background	35
4.1.2	Nuclei Extraction from Real Data	36
4.1.3	Nuclei Placement	38
4.1.4	Extraction of Staining Characteristic Variability from Real Data	41
4.1.5	Color Deconvolution and Color Normalization	42
4.2	Nuclei Quantification and Segmentation	43
4.2.1	Deconvolution	45
4.2.2	Normalization and Contrast Adjustment	49
4.2.3	Thresholding of the Stain Channels	49
4.2.4	Segmentation and Quantification	50
4.3	Evaluation Methods	59
4.3.1	Evaluation of Nuclei Quantification	60
4.3.2	Evaluation of Nuclei Segmentation	62
5	Results and Discussion	67
5.1	Synthetic Dataset	67
5.1.1	Colorspace Variation	67
5.1.2	Dataset Characteristics	70
5.1.3	Summary and Discussion of Synthetic Datasets Results	73
5.2	Nuclei Quantification and Segmentation Metrics	75
5.2.1	Nuclei Quantification	76
5.2.2	Nuclei Segmentation	83
5.2.3	Summary and Discussion of Nuclei Quantification and Segmentation Results	90
6	Conclusion	95
7	Future Work	97
7.1	Recommendations on the Improvement of the Suggested Solution	97
7.2	Further Opportunities with the Suggested Image Synthesis Method	101
	Appendix	103
	Abbreviations	103
	Nomenclature	104
	Bibliography	107

Introduction

This chapter provides an overview of the motivation, research question and structure of this work. First, an introduction into the motivation for this topic is given. Then, the research question is defined, followed by the scope of discussion. Finally, the chapter concludes with an outline of the thesis.

1.1 Motivation

Pathological exams of tissue specimen are a vital part of routine cancer diagnosis [53]. The tissue is routinely being examined under the microscope, which is time-consuming (around 25 minutes to count 2000 cells) and requires an expert pathologist [55, 83, 88]. The diagnosis is also subject to various sources of variability, it can for example be different from pathologist to pathologist or depend on the processing steps of transferring the tissue onto a glass slide [53]. While the latter factor can be alleviated via rigorous standardization measures, the pathologists variability is a human factor which can never be fully avoided [7, 21, 53, 83]. In the last decade, digitization of slides via scanning has found its way into pathology [32] which has prompted a noticeable response in the research community, as can be seen in Figure 1.1, showing the number of times, relevant keywords have appeared in publications on Pubmed¹. The possibility of slides being available as digital images promises the opportunity to apply digital image analysis methods on the tissue [32]. While it is self-evident that any analysis software is intended to support, rather than replace the pathologist and his wealth of expertise, the possibilities of digital image analysis lie in the facilitation and speed-up of daily clinical routine tasks as well as in the opportunities for large research studies on current or retrospective data [32].

¹Pubmed: <http://www.ncbi.nlm.nih.gov/pubmed>, accessed on January 26th 2016, 09:50



Figure 1.1: Appearance of terms in title or abstract of search results on Pubmed

1.2 Problem Statement

The grading of breast cancer cases required for rendering a diagnosis and concluding the best possible treatment is based on findings from mammography as well as on examinations of extracted breast tissue on glass slides [31]. These histological exams include counting cells in the tissue under the microscope. With breast cancer being the most frequent cancer among women between 40 and 60, the quantification of cells is a frequent task in the daily routine of a pathologist [31]. At the same time, it is of a time-consuming nature because in theory it requires to manually count cells in a selected or given area of the sample [18]. Besides the disadvantageous time factor, the outcome of such a manual assessment has also shown to be highly variable depending on factors such as the pre-processing steps of the slide or the expertise of the pathologist [53].

One of the exams for grading breast cancer studies the proliferation rate of cells, because it is an indicator for the aggressiveness of a breast tumor [18]. The percentage of dividing (i.e. proliferating) cells with respect to the total number of cells is termed Labeling Index (LI) and has a large impact on the treatment decision [19]. Cells which undergo division can be made visible in tissue by being colored with a specific stain called Ki-67 and they will appear as brown nuclei, while non-dividing cells and nuclei are typically colored with an unspecific blue stain [82]. Since the introduction of digital pathology, there have been research efforts to find ways of support the manual visual assessments (which are referred to as LI estimation rather than LI determination) with assessments via digital analysis. Most of them require either training of the pathologist on the software, request manual inputs to adapt to every new case or heavily rely on assumptions based on different aspects of the tissue or nuclei which bears the risk of delivering biased results [32]. Adding to these constraints, one of the key factors of any digital image analysis solution is a thorough evaluation step to ensure that the algorithm actually delivers a

quantifiable answer to the problem [30]. An obvious ambiguity lies in the fact that the ground truth datasets used for verification are typically created by pathologists who label the cells based on their manual counts, despite the fact that the high variability of such manually created assessments was one of the driving forces for creating a non-variable digital solution. This factor favors the use of an artificial dataset, where the undisputable ground truth LI is available from the production process of the dataset [58, 68]. To date and to the best knowledge of the author, no such dataset synthesis method for training and testing Ki-67 LI estimation algorithm exists.

1.3 Research Questions

In line with the current state of the art, this work seeks to tackle two of the current challenges concerning the automated Ki-67 labeling on digitized slides. The research questions are:

Can scoring of the Labeling Index in Ki-67 stained slides (LI estimation) be performed in an automated fashion, requiring neither prior training of the program nor manual inputs from the user? If so, to which point of accuracy?

Is a synthetic, labeled dataset suitable to evaluate the performance of such an algorithm?

1.4 Methodological Approach

A pipeline is assembled to test and accomplish the aims of the work, using Matlab as a prototyping-tool, which provides a large library of readily built-in functions especially for image processing. The pipeline includes sections to adapt to images with varying staining characteristics, to classify, segment and count the cells. After extensive literature research, the most relevant solutions for each section are coarsely implemented to test the functionality for the given use case and the most appropriate solutions are implemented within the final pipeline.

The evaluation of the Ki-67 LI algorithm is based on a custom-built synthetic dataset of Ki-67 breast images. Established criteria for nuclei quantification and segmentation algorithms are employed to rate the performance of the algorithm.

1.5 Scope of Discussion

The presented work focuses on finding answers and insights to the research questions stated in Section 1.3, while aspects like the time-wise performance of any suggested solution, as well as the user-experience (e.g. in terms of a graphical user-interface) are not investigated further. This work does not aim at providing a complete, clinically usable software solution or directly binding into any existing software or systems for handling pathology data.

The synthetic dataset generated within the scope of this work is not intended to mimic all aspects of natural appearance and behavior of human breast tissue, but is specialized in allowing the evaluation of a nuclear quantification and segmentation algorithm.

1.6 Structure of the Work

In the beginning of the work, the state of the art about the topics touched is presented: Chapter 2 first covers the aspects of the medical topic of pathology and paves the way for Chapter 3. There, the motives and challenges for the development of digital pathology solutions and the status thereof are outlined and a focus is laid on the necessities and prerequisites of advancing research in automated Ki-67 labeling.

Chapter 4 then expands on the suggested solutions for the research questions, namely the generation of a synthetic Ki-67 dataset and the implementation of a nuclei quantification and segmentation algorithm, concluding with a section about the conducted evaluation criteria and methods.

Chapter 5 reveals the outcome and characteristics of the synthetic dataset and expands on the performance of the nuclei quantification and segmentation algorithm by illustrating as well as discussing the quantitative and qualitative results given by the evaluation. The chapter concludes with recommendations for further investigations on the suggested solution and highlights a number of potential further use-cases for the synthetic dataset.

The thesis concludes by summarizing the major findings with regards to the stated research questions in Chapter 6.

CHAPTER 2

Pathology

Pathology is the diagnostic practice to assess tissue and give a diagnosis which serves as the basis for further, customized patient treatment. The correctness of the diagnosis is important, as it is a decisive factor for the therapeutic response¹ of this patient to the consequent steps, be it a mere follow-up, or local excision, medical treatment or even surgery and chemotherapy [53].

Pathology becomes relevant during a diagnostic process, when the physician determines that a histological confirmation/assessment is required in order to proceed with the treatment [53]. It includes the examination of solute cells, called "Cytology", and the examination of whole tissue, called "Histology" or "Histopathology". In the context of this work, the focus lies on Histology.

There is a clear distinction between clinical practice and the research domain when it comes to the pathologists goal and timely issues [53]. In clinical practice, pathologists have a clear focus on delivering an accurate diagnosis within a short time-span [53]. They analyze a number of features in the histology slides, such as tissue architecture, nuclear morphometry and quantification or cellular color and texture [53]. For example it ideally only takes about 20 minutes to render a diagnosis based on an inter-operatively extracted tissue sample [42]. In research, as a contrast, the pathologist tries to quantify and describe the differences between histology samples, while looking at similar features such as nuclear morphometry, stroma quantity, tissue classification etc. The time required for research analysis tasks does not play such an essential role as for clinical routine tasks [53].

¹therapeutic response is a term for the success of treatment

2.1 Processing Steps prior to Analysis by Pathologist

The tissue to be examined undergoes certain processing steps, before it can be viewed and analyzed under the microscope by a pathologist [53, 88]. First, tissue is taken from the patient in form of a biopsy or a tumor excision, both of which are usually done in the operating room. In the pathology lab, the tissue is fixated by formalin as soon as possible, which prevents natural biological processes such as apoptotic cell death or growing microorganisms [42, 53, 64, 88]. Specifically in histology, the fixation is conducted in such a way that the biological reactivity of proteins stays intact, which is important for the detection of enzymes or antigen structures [42]. Then the tissue specimen are dehydrated and embedded in paraffin. In order to achieve near-transparent characteristics for bright-field microscopy, the block of tissue embedded in paraffin is then cut into small slices of 3-5 micrometers using a microtome, which is a high precision slicing instrument. The slices are placed on glass slides for convenient manipulation under the microscope and dried in an incubator [64, 88]. Because the tissue is still near-transparent at this stage, the structures of interest have to be highlighted. This is achieved via dying with a stain that targets different structures. The standard staining protocol, which is in use for around a century, uses Hematoxylin and Eosin (H&E) to highlight tissue structures such as nuclei in different colors [42, 88].

2.2 Staining Protocols in Breast Cancer Diagnosis

In line with the focus of this work, this section will only discuss staining protocols common in breast cancer diagnosis. The term staining protocol actually refers to the instructions of applying a stain to a slice of tissue on a glass slide. Many diagnostic histological processes, also besides breast cancer, utilize the so-called "H&E-staining", which is a long established standard-protocol, as mentioned previously. According to this protocol, the tissue is exposed to two substances, Hematoxylin (H) and Eosin (E). Hematoxylin dyes the nuclei blue/purple, because it binds to DNA and Eosin dyes the other structures (such as stroma or cytoplasm) pink, because it binds to proteins [85]. This images treated with this stain have nuclei in a blue tone and the cytoplasmic structures appear clear to red or purple, depending on the constituents [53]. The foremost advantage of H&E is that it visualizes almost all cellular components and provides fair contrast at opposite ends of the spectrum [53] although the latter claim is challenged in recent publications like [34]. For further, more functional investigations, a range of sophisticated staining techniques called Immuno-Histo-Chemistry (IHC) can be applied [85]. IHC utilizes antibodies, which bind to and therefore highlight specific antigens. In breast cancer diagnosis, such antigens are the Human Epidermal Growth factor 2 (HER2), Estrogen (ER), Progesterone (PR) or Ki-67. [85] The presence of the Ki-67 antigen is an indicator of tumor growth and it can be found as a monoclonal antibody at the nucleus of a proliferating cell, thus it allows the assessment of cell proliferation [26, 85]. More details on the Ki-67 stain are given in Section 2.3.1. IHC staining targets only exclusive, functional parts of the otherwise transparent tissue in a slide [69]. Therefore it is combined with a counter-stain,

most commonly Hematoxylin. Hematoxylin dyes both nuclei and the surrounding tissue architecture in a blue shade and enables the visualization of IHC-negative nuclei, the latter being of importance as a contrast to the nuclei stained via IHC [85].

Since the receptor status (i.e. the quantity of specific antigens) has an impact on the prognostic outcome of each patient, it is vital that the staining protocols are quality-controlled and strictly standardized and adhered to in order to enable reproducible, reliable results [85]. In breast cancer research, a frequent task is to conduct standardized cross-patient IHC examination. The best solution for this problem so far is to insert many small biopsy samples into one paraffin block in a strictly grid-structured manner. This block is called Tissue MicroArray (TMA). Its slices contain tissue samples from a larger patient group and the subsequent staining process is automatically standardized across the entire group [85].

2.3 Breast Cancer Diagnosis with the Bloom-Richardson Method

Equal to all diagnosis rendered via the analysis of histological specimen under the microscope, it is of crucial importance that careful attention is paid to the preparation of the breast tissue [18]. This includes e.g. prompt fixation after biopsy or excision [18].

The semi-quantitative method for assessment of what is called the histological grade in breast carcinoma includes several features [18]. They are viewed in Table 2.1 and each variable is assessed separately [18].

The sum of all scores yields the overall tumor grade, where 3-5 is a grade I tumor (well differentiated) and has the best prognostic outcome for the patient, 6-7 points is a grade II tumor (moderately differentiated) and 8-9 points is a grade III tumor (poorly differentiated) and has the worst prognostic outcome [18, 93]. In this context, the term "prognostic outcome" refers to the theoretical outcome if the patient remains untreated [20]. It is necessary to try and minimize the subjectivity of deriving the grade by adhering to strict criteria for each of these evaluation steps in order to ensure reproducibility of

Table 2.1: Assessed features and resulting scores to yield histological grading of breast cancer [18]

Feature	Criterion	Score
Tubule formation	Majority of tumor (>75%)	1
	Moderate degree (10-75%)	2
	Little or none (<10%)	3
Nuclear pleomorphism	Small, regular uniform cells	1
	Moderate increase in size and variability	2
	Marked variation	3
Mitotic counts	Dependent on microscope field area	1-3

these scorings and grades [93]. As an example, for the assessment of the mitotic activity only nuclei in which clear morphological features of mitosis are expressed may be counted, while apoptotic or hyperchromatic nuclei as well as lymphocytes have to be ignored [18]. A certain impact of subjectivity yet can never be ruled out [93]. Basically, all these features can be assessed on tissue stained with the standard H&E protocol [18]. However, one way to alleviate some of this subjectivity is to substitute the H&E-based mitotic count with the assessment of the proliferative activity using an IHC stain targeting a nuclear antigen only present in phases of proliferation, such as Ki-67 [16].

In addition to the cancer grading, indicating the prognostic outcome, there are also histological analysis techniques aiming at the predictive outcome of the patient – how well the patient will respond to different treatment strategies and the theoretical outcome if the patient is treated [20]. To this end, several receptor statuses are tested, using the aforementioned IHC staining protocols for ER, PR, HER-2 and Ki-67 [20]. The examination results from determining these receptor statuses can greatly influence the subsequent treatment plan [85]. They are commonly determined via counting the percentage of IHC-positive nuclei in relation to the IHC-negative nuclei [85]. A certain threshold divides overall positive tissue from overall negative tissue, where the thresholds vary depending on the type of IHC and country (e.g. in Europe it is 10% for the ER and PR examination, in the U.S.A. 1%) [85]. The status of the HER2 receptor, in contrast, is not based on the nuclei quantification, but on the solidity and staining intensity of the cell membranes [85].

2.3.1 Proliferation Assessment via Ki-67 Scoring

A prognostically highly significant and thus important part of the histological grading procedure targets the proliferation activity of the tumor, expressing how fast the tumor grows in terms of cell growth and division [88]. This can be done via a mitotic count² in common H&E staining [79, 88], whereas mitotic cells can easily be confused with other structures or constructs visible in histopathological images, such as apoptotic, necrotic or merely epithelial cells as they are highlighted by the same stain [74]. Alternatively, the proliferative activity can also be assessed via using antibodies against cell phase specific antigens such as Ki-67. The Ki-67 nuclear protein is expressed in all phases of the nuclear cycle but the resting phase (G0-phase). Consequently, it reveals all phases in which the cell is undergoing proliferation [20]. Ki-67 scoring has proven to be an appropriate substitute for the mitotic count and even provides highly significant predictive information about treatment efficacy when used as the only criterion [20, 79].

However, many details about the Ki-67 assessment are subject to open discussions [20]: The definition of the threshold or cut-off value for a "positive" result (e.g. >1%, >10%, >20% etc.) remains a challenging questions yet unsolved, although the significant correlation between Ki-67 index and treatment efficacy has already been established [20, 93]. Data suggests that the cut-off value is helpful in identifying the patients with

²The mitotic count describes the percentage of cells currently undergoing mitosis, i.e. cell division

the highest chance to profit from chemotherapy [70]. It is also not yet fully established whether it is the percentage of positive nuclei which is more relevant or the percentage of stained area [20, 70]. Another open issue concerns the locations or so-called microscopic fields, in which the assessment should be conducted [26, 65]. It is agreed, however, that the evaluation of Ki-67 should be performed in areas with the highest percentage of positive nuclei, e.g. the invasive border of the tumor tissue [70]. As suggested in [26], the hotspot containing the highest percentage of positive nuclei, is determined via visual judgment at low 40x magnification of the tissue. This view shows what is called a low-power field [26]. One recommendation is to consecutively count the nuclei in three to five high-power fields (400x) within this low-power field [16, 37]. The challenge lies in the heterogeneity of the Ki-67 staining intensity, which increases towards the tumor edge and is particularly prevalent in hotspots, so the question remains whether these high-power fields should focus on hotspots, only include them or avoid them altogether [16]. Between 500 and 2000 tumor cells are observed during the assessment [16, 43, 83].

The level of Ki-67 expression, i.e. how strong and frequent the stain attaches to tissue nuclei, is commonly described and referred to as Ki-67 Labeling Index (Ki-67 LI) or proliferation index and the process of deriving this index is referred to as Ki-67 scoring or LI estimation [12, 94]. The scoring should optimally only include cells within the tumor region and exclude other areas such as connective tissue [41].

2.3.2 Sources of Variability

Histology-based diagnosis has yet not been standardized to an extent which eliminates all possible sources of variations [53]. The reproducibility of diagnosis in Ki-67 scoring and histology in general stems from a range of factors: First, there is a natural variability among the characteristic of the human body [53]. Together with the long chain of pathological deviations which can affect a tissue, this leads to an infinite number of biological appearances – the biological variability [53]. Secondly, two pathologists can assess the same slide and conclude different diagnoses from it – this is called the inter-observer variability [21, 53]. Thirdly, the same pathologist can render different results on the same slides when assessing them at two time-instances, which is termed intra-observer variability [7, 83]. Adding to this, there are also considerable inter-laboratory differences in the staining techniques, introducing yet another factor of variability with significant impact on the Ki-67 LI [54]. And last but not least, the non-uniform diagnostic evaluation criteria as mentioned in Section 2.3, also play a role in the low reproducibility of assessments.

All of the mentioned factors are motivators for the development of more standardized solutions of histology-based diagnosis. Among other efforts, this has supported the necessity for and advancement of developing non-subjective digital solutions for pathology-related tasks. The next chapter gives an overview about the field of so-called digital pathology.

2.3.3 Summary

Pathology is a medical discipline for the assessment of tissue in a qualitative and quantitative manner via examination under the microscope. For this purpose, the tissue is excised and treated in order to be placed on a glass slide and display the biological structures. One of the treatment steps is staining of the tissue to highlight distinct structures, such as nuclei or cell membranes. The most common stain is Hematoxylin and Eosin, H&E, which dyes the nuclei blue/purple and the other structures (such as cytoplasm) pink. There are also stains summarized under the term Immuno-Histo-Chemistry (IHC), which target functional characteristics of the tissue. Ki-67 is one of these IHC stains, which dyes the nuclei of proliferating cells in a brown shade. It is counter-stained with Hematoxylin to be able to differentiate between proliferating (Ki-67 positive) and non-proliferating (Ki-67 negative) cells. The quantification of positive versus negative cells is called Labeling Index (LI) scoring and is a diagnostic measure for tumor growth which is routinely examined in breast cancer cases, because it has an impact on the treatment decisions and thus the prognostic outcome of each patient.

Various sources of variability make the reproducibility of pathological diagnosis difficult, such as the natural variability of human body, the variability within pathologically occurring deviations of diseases, the variability of a diagnosis rendered by a single pathologist on two different time instances, the variability of a diagnosis rendered on one slide by different pathologists and even the variability of the tissue preparation steps which can differ within and between laboratories. Adding to this, the diagnostic evaluation criteria are not standardized.

With breast cancer being one of the most frequently occurring cancer types, rigorous evaluation standardization as well as reproducible, non-subjective quantifications are desired. Therefore the task of Ki-67 LI assessment can greatly benefit from digital image analysis solutions.

Digital Pathology

Digital pathology basically describes the digitization of a physical pathology slide into an image, which can be viewed on a computer screen [76]. Several terms exist for describing the field of digital pathology, e.g. computational pathology, [21] virtual microscopy [89], digital histopathology [35] or pathology virtual slide technology [72] and for the denomination of created images, e.g. (digital) whole slide imaging/images [15, 27, 90] or virtual slides [90]. In the last two decades, the field has grown largely (see Figure 1.1) and in 2014 the term also included the entire pathology information system, complete with archiving management, real-time evaluation, tele-viewing and –consultation, and applications in education, clinical routine diagnosis, research and the development of artificial intelligence instruments [76]. While the field of radiology has been digital at least 10 years before pathology, the concepts utilized in radiology cannot simply be reused for pathology, because there are vast differences to be considered [76]. In radiology, live specimen (e.g. patients) are imaged, whereas in pathology there is a constant interchange between wet, hands-on laboratory conditions and digital observations [76].

The following sections give an insight into the motives for the ongoing advancement of the field of digital pathology, the technical state-of-the-art, as well as into the most challenging factors in the development of solutions. Also, an excerpt of frequent digital image analysis problems in digital pathology is highlighted, with a focus on digital solutions for Ki-67 scoring. In the end, evaluation approaches for quantification and segmentation algorithms in digital pathology are investigated.

3.1 Motives for the Advancement of Digital Pathology

The inclusion of pathology images and metadata into information systems is on track to the ultimate goal of creating a large cradle-to-grave electronic patient record and today this goal is technically in foreseeable reach [32]. The digital availability of data also

allows the use of old cases for data mining or re-assessing therapies, offering potential benefits for future patients [32].

The handling of glass slides is an efficient way to make an initial diagnosis but they are overall expensive, time-consuming and largely inefficient in terms of storage, research, education and re-consultation [32]. The conversion of glass slides into digital Whole Slide Images (WSI) offers more cost-effective and efficient means of archiving, presenting and transmitting pathology information [32]. Compared to glass slides, virtual slides offer several opportunities, such as being easily viewed by several people at the same time, cases can be assessed anywhere and anytime, lab-space used for pathology can be in multi-use for other purposes, WSI can be repeatedly annotated or quick switching between slides is possible [32]. Currently, institutions are required by law to keep their glass slides and tissue blocks for at least ten years, but the digitization provides the option of storing the slides using considerably less physical space and avoids the risk of being lost, damaged or fade over time [53].

The use-cases include making tissue diagnosis, educational purposes for all stages of proficiency, consultation (also via telepathology, where a diagnosis is made in another location than the physical glass slide location), quality management, archive management and, last but not least, research [32]. The largest hopes of pathologists concerning digitizing pathology are pinned on factors like improved ergonomics for the pathologist, diagnostic accuracy, measurement accuracy, time saved and speed of slide navigation [49]. Together with foreseen positive impacts on laboratory aspects like slide preparation or slide handling, overall positive effects are expected to be noticed in the overall quality of healthcare, the economy of healthcare pathways, the economy of pathology departments and the development pace of diagnostics [49]. A study has shown that over a 5-year span, the use of digital pathology can be a huge cost saving factor by increasing the productivity, by lab consolidation and by avoiding unnecessary treatment costs thanks to more accurate cancer diagnoses even from non-subspecialty pathologists [29].

The field of digital imaging analysis has therefore gained increasing attention over the past two decades, both in research and in clinical practice [95]. The motivation for digital image analysis to aid in the diagnostic process is founded in the fact that manual diagnosis under the microscope does not only foster various factors of variability, but is also time-consuming and requires high expertise and skill from the examiner [41, 95]. Furthermore, pathology and microscopy images exhibit very complex natures, which makes manual assessment challenging and causes large variability of the examination results [95]. Thus, one of the major advantages, which also drives the motivation for this work, is that Computer-Aided Diagnosis (CAD) based on digitized slides provides the opportunity for quantitative analysis of pathology images with a high throughput rate [95]. The CAD can reduce the bias and deliver reproducible and accurate estimates of diseases, thereby decrease various factors of variability [95].

While computer algorithms on digital slides can automate some of the pathologists routine tasks, such as screening pap smear slides, and thereby reduce the staff expenses and allow more time for challenging cases, the motive for introducing digital image analysis

in research is even more profound [53]. The analysis results rendered by a pathologist are incomplete with regard to the fact that the human visual system is unable to fully identify and classify all biologically or clinically important features [53]. In contrast to manual assessment, CAD also allows the extraction of rigorous quantifiable measures of image features, which does not only facilitate the clinical workflow, but also opens the door for vast comparative studies of older cases in order to deepen insights into potential prognosis and individualized treatment options [95]. For example, it is hard to demonstrate that the average nucleus diameter is bigger in one specimen than the other or to quantify the chromatin distribution, which can form very complex patterns [53]. Here, Digital Image Analysis (DIA) can prompt answers to these tasks on a comparative, reproducible and reliable basis [53].

3.2 State-of-the-Art

In the early stages of digital pathology, the digitization was conducted by capturing still images with a digital camera mounted onto the ocular of the microscope [85]. As of 2016, the scanning procedure is handled automatically by Whole Slide Imaging scanners (WSI-scanners) [85]. They conduct all steps including loading the slides onto the scanning platform, detecting the tissue regions on the slide and selecting a focus point, as well as image acquisition, compression, registering and storing them on a laboratory information systems [85]. Several vendors and file formats can be found on the market: Aperio produces images with an extension called .svs, which is a tiff-based single-file format, the .vms-files by Hamamatsu are jpeg-based multiple images, the .scn-files by Leica are bigTIFF with XML metadata in single-file format and 3DHistech allows the use of various image formats in their multiple-file .mrxs-files [32].

WSI scanners perform rapid slide-scanning at $20\times$ or $40\times$ magnification¹ creating high-resolution images with around 0.5 to 0.25 micrometers per pixel, respectively [53, 85]. To reduce the file size, the images are stored in a pyramid structure with varying magnification levels, which facilitates fast navigation and multi-scale image analysis [85]. Storing a common tissue area of a glass slide of $15\text{mm} \times 15\text{mm}$ yields in the range of 3GB of data, which may be reduced to 200-500MB by compression [69]. The storage space of an entire case (up to 30 different stains applied to different sections of the specimen) may amount to 10s of GB, with extremes going into 100s of GB when multiple focal planes are stored per slide [69]. These numbers exceed data sizes found in radiology studies by at least an order of magnitude [69].

While all of the available digital pathology systems allow the inclusion of relevant metadata, the missing common format and data model hinders straightforward sharing between devices and laboratories [32]. There has been an effort to create DICOM² and

¹Those numbers always refer to the magnification in addition to a $10\times$ objective, resulting in total perceived magnifications of 200 or 400 times the real size

²DICOM = Digital Imaging and Communications in Medicine, the most common medical imaging and communications standard to facilitate the interchange of images and metadata

HL7³ standards as well as IHE⁴ profiles to facilitate interchange of images including their metadata, leading to the publication of an IHE framework in 2010 and two DICOM supplements in 2014, but they have yet to be adopted by all the vendors [32].

In 2015, the Food and Drug Administration had not yet approved the solitary use of digital pathology for primary diagnosis [32]. However, the concordance rates between examination results from light microscopy and WSI is above 90% for different fields of pathology, e.g. dermatological pathology, gastrointestinal tract pathology or breast pathology [32]. The discordance is claimed to be induced mostly by lack of experience in the handling of digital pathology [32].

Concerning the acceptance of digital pathology, a 2014 symposium on digital pathology in Sweden conducted surveys among the participants and showed that digital pathology for various fields of application (such as diagnosis, re-reviewing, secondary consultation or education) is in average still used in only 40% of all cases, but the pathologists and related healthcare personnel have high expectations that this number will increase in the upcoming years, predicting a use in almost 80% of all cases by the end of 2016 [49].

3.3 Challenges

While improvements of image-viewing solutions, image quality or scan times are happening, the complete conversion to digital pathology like in radiology is not yet reality [32]. The WSI technology has successfully found its way into niche applications of clinical, research and educational purpose but some challenges still require more attention to allow full integration [32]. These include, among other aspects, the lack of standard development for practice and validation guidelines, work-flow adaptations, regulatory issues or the huge amounts of data generated [32]. The lack of standardization concerns e.g. the sample processing and viewing process: Tissue sample, staining, optical properties of microscope and scanner, storage format, display calibration can vary [34, 85].

As stated in Section 2.3.2, the three sources of variability, especially the technical variability, can significantly pose an obstacle for digital image processing systems [53]. The human visual system of the pathologist who evaluates the respective sample is also a criterion which benefits from more physiological prerequisites – the contrast on the slides introduced by the applied stains may not be optimal for human visual perception [34].

At the current stage of viewers for digitized whole slide images, studies have shown differences in diagnoses based on assessment of digitized images compared to physical slides under the microscopes [12, 56]. Yet, pathologists still prefer the microscope, which to date offers faster panning and focusing as well as an hardly reproducible optical impression of being nearer to the tissue [53]. Also, the point at which pathologists trust and routinely use such systems to an extent that justifies the large costs put into purchase and development of these tools has not been reached yet [76]. Pathologists

³HL7 = Health Level 7, another major medical communication standard [32]

⁴IHE = Integrating the Health-care Enterprise, an international organization [32]

are non-subject matter experts, which means that they are not trained to use and/or understand image processing software. Thus any program for the pathologist community should be designed in a way that allows to be fully utilized by a non-technical subject matter expert [28].

With the large data volume created in digital pathology and the necessity to have large amounts available online at any time, cost-efficient storage solutions are still an issue [49].

In 2015, despite increasing numbers of publications in the last decade, the research on automated histology analysis is still scattered in comparison to automated radiology analysis [71]. This leads to methods being tailored to limited, private datasets and a standard on quantitative criteria to be reported is still lacking [53]. There is an evident need for public benchmark datasets and standardized evaluation criteria in order for the field of digital pathology to advance to the next critical step [53].

3.4 Frequent Digital Image Analysis Challenges in Digital Pathology

In the scope of this work, the focus lies on digital analysis for images acquired by brightfield-microscopy (as opposed to fluorescence or multispectral microscopy) since the majority of diagnostic steps in breast cancer is performed for this kind of imaging [85]. Most to-date officially approved DIA solutions in digital pathology concern the features of grading systems, like the quantification of bio-marker expression such as ER, PR and HER2⁵, or the automatic calculation of the mitotic rate [41].

3.4.1 Artifacts

Histopathology images exhibit artifacts, which can pose a challenge to any DIA design [95]. They stem from different steps of the slide processing or digitization pipeline and include tissue deformations, background clutter, noise, blurred regions, poor contrast and more [95]. Fixation errors for example introduce changes in tissue morphology and imperfections during mounting or variation of staining may lead to missing parts or out-of-focus regions and over- or under-staining [85]. Bad fixation can also lead to shrinkage artifacts, which can result in clefts being mistaken as lumina⁶[18].

Artefacts and problems during the staining process include poor or excessive contrast and saturation [64]. Tissue slices are by their nature very thin (3-5 micrometers) and can happen to form tissue deformations such as folds when being mounted onto the glass slide [53]. These folds typically appear as areas of high color saturation and are hard to address and avoid in DIA, yet some efforts have been published [53]. Furthermore, variations in the water content of different tissue areas can result in tearing when the tissue is drying during the processing steps [53]. These tears appear as white cracks,

⁵Estrogen (ER), Progesterone (PR) and Human Epidermal Growth factor 2 (HER2), see Section 2.2

⁶Biological term describing a tissue opening such as the inside of a tubular structure, artery or ducts

which have no biological implications. As of 2015, no work has yet addressed this artifact specifically [53]. Also, dull blades of the microtome can lead to alternating light and dark regions, known as chatter artifacts [53].

Another class of common artifacts stems from the stitching process [53]: since WSIs are not scanned all on one, but in strips and tiles, the whole image is created by stitching these parts together [53]. This happens in the scanner software [53]. The stitching becomes challenging when a large specimen has to be sectioned and scanned in several steps [53]. So far only manual solutions have been proposed [53].

3.4.2 Color Deconvolution

Depending on the task, e.g. the aforementioned IHC quantification or nuclei detection, DIA methods can profit from separating the stains which were applied to dye different tissue structures [85]. Several approaches have been proposed [85]. One category of solutions suggest to cluster or classify the RGB (red, green and blue) pixel values to gain binary images or probability maps for the applied stains [85]. The clusters which correspond to the respective stains have to be identified or the data has to be labeled. A different set of approaches is based on the physical background of the staining process [67, 85]. In brightfield microscopy, the images are formed according to the Lambert-Beer law about light absorption. Using this model, the concentration of a stain is proportional to the optical density (the logarithm of the intensity of this stain) in the tissue [67, 85]. The concentration of each of up to three stains can be found by linear decomposition, based on the fact that the image was acquired using three detection channels in the image. The identified stain concentrations allow the derivation of single-stain images by reversing the previously applied approach [67, 85]. Ruifrok et al. [67] was the first to report this approach for quantification of IHC by color deconvolution, which came to be a fundamental work in this field.

However, for the solutions based on these technique, the specific absorption spectrum for each stain has to be provided by the user, as e.g. in [85], or vectors for deconvolution can also be extracted or estimated using manual selection of representative regions, as in [51]. As laid out in [34], many methods resort to fixed vectors for conducting the color deconvolution, as for example provided by Ruifrok [67]. Deconvolution steps relying on fixed vectors are for example utilized in [47, 63, 89].

Some proposed DIA solutions try to overcome this limitation either by being robust to staining variabilities [85] or by standardizing/normalizing the appearance prior to further DIA steps, as in [50]. The normalization process includes the identification of the inherent stain concentrations in each pixel before all channels are normalized and re-mixed to obtain a standardized appearance [85]. Most methods rely solely on the image-inherent color information and do not consider spatial dependencies of different structures in the tissue, which potentially limits their robustness [4].

The mentioned normalization approach by Macenko et al. [50] bases the deconvolution step on the assumption that the reference stain vectors are implicitly given in the non-

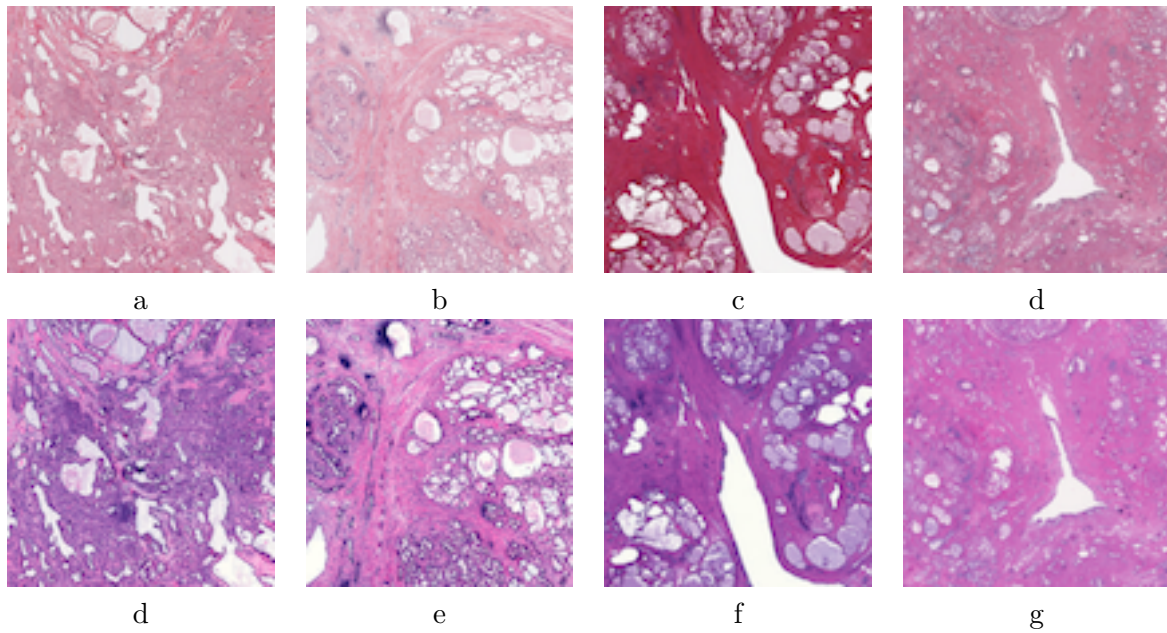


Figure 3.1: Examples for the heterogeneous appearances of images produced in different laboratories, **(a)** and **(b)** from one laboratory and **(c)** and **(d)** from another. Images **(e)**-**(h)** show their more homogeneous normalized equivalents after normalization using the approach by [50] (images taken from [91])

linear "optical density" representation of the pixel distributions (Equation 4.1). The fringes of the pixel cloud are said to constitute the stain vectors and can be identified with the help of a singular vector decomposition [50]. However, [4] accredits the method by [50] only limited applicability as it would not sufficiently adapt to strong staining variations and yield poor stain vector estimates. Examples of the normalizing effect of applying the method in [50] on prostate whole slide images from different laboratories can be viewed in Figure 3.1. The upper row shows the appearance of images as they were originally digitized in two different institutions and the lower row shows their respective appearance after normalization.

Another approach for color deconvolution utilized as a component in a method for grading nuclear pleomorphism (i.e. roughly speaking the shape variations) is published in [11] and based on the assumption that the stains can be linearly separated. It analyses the Cyan Magenta Yellow and Key (CMYK) representation of the original image for dominantly purple and non-purple values, computes the main axes thereof and then projects the original RGB pixel values onto axes orthogonal to the main axes. This approach is used in a method for detecting regions of interest in H&E stained images by Bahlmann et al. [3].

3.4.3 Tissue Classification

The typical tissue area on a histology slides measures 15x15mm, with resolutions reaching into the gigapixel range [85]. For this reason, it is common to first identify the regions of interest within the given tissue [85]. As a first step, most WSI scanners already exclude large portions containing white background from the scanning process [85]. To further reduce the regions for DIA, relevant areas for the respective task are identified [3, 10, 33, 59, 62]. When classifying breast cancer into benign or malign classes for example, only the epithelial areas of the tissue are relevant [85]. Quantification of IHC or the performing of histological grading requires only the tumor tissue, hence all non-tumor regions should be excluded from the DIA [85]. Generally, epithelial and stromal regions of tumor contribute differently to the prognostic outcome [85]. Hence, it is common to segment the tumor into the classes "epithelium" and "stroma" before proceeding with other DIA or grading steps [85]. For tissue segmentation/classification, the division can be done by supervised pixel-wise classification of small sub-image blocks, based on features such as color and/or texture [85, 3]. Unsupervised methods have also been proposed [36].

3.4.4 ImmunoHistoChemical (IHC) Quantification

In contrast to ubiquitous H&E stained slides, where the features of interest are rather complex – the size of the nuclei, their texture and shape, their spatial arrangement and tubule formation, the stroma interaction etc. – the features of interest in IHC stained slides are foremost contained in the staining intensity and color [85]. This factor makes IHC samples applicable for DIA. One readily available feature is the percentage of pixels which are positively stained with a specific target antigen (e.g. "brown" pixels, denoting the presence of the Ki-67 antigen), in relation to the negatively stained pixels (e.g. "violet" pixels, denoting the absence of the Ki-67 antigen) [85]. Owing to these circumstances and the fact that visual IHC examinations are prone to variability among pathologists even when following strict guidelines, the American Society of Clinical Oncology together with the College of American Pathologists encourage the use of DIA techniques in order to improve the consistency of interpretation of IHC slides [85].

Most commercially available platforms for DIA on pathology images supply algorithms for nuclei and membrane staining quantification [85]. Automatic scoring solutions have shown to highly agree with expert scoring results [85, 6].

It is advised not to draw a direct conclusion from the stain intensity derived via color deconvolution to the quantity of IHC stain in a sample, as not all stains are stoichiometric, i.e. the amount of stain visible does not necessarily reflect the amount of histochemical reaction products [21, 78]. Adding to it, some stains such as DAB⁷ do not follow the Lambert-Beer Law, thus they are not true absorbers - a scattering effect can broaden the observed spectrum [21, 81].

⁷Diaminobenzidine, an anti-Ki-67 antibody

3.4.5 Nuclei Detection and Segmentation

Even though a diagnostic factor can be concluded via the percentage of positively stained nuclear area without the need for segmenting single nuclei, DIA approaches customarily include nucleus detection or segmentation steps [85, 80]. The ER, PR and Ki-67 receptor status for example are commonly examined via the rate of positively stained nuclei [85]. For this task, it is sufficient to detect the cells, whereas detection refers to obtaining the rough object location rather than delineating its boundaries [95]. It results in one marker/seed per nucleus, which can be one pixel or a small Connected Component (CC) within the object of interest [95]. A large variety of solutions for nuclei identification has been proposed, including different approaches for all pre- and post-processing steps [95, 85]. It is commonly combined with segmentation of nuclei, either as a preceding step, generating seeds for the segmentation, or following segmentation of nuclear areas as an adjacent step, where the seeds serve as markers for the separation of nuclei clumps into individual nuclei [85, 95].

Methods for nuclei detection rely on common algorithms also utilized in other image processing areas, such as distance transforms, morphology operations, H-maxima and H-minima transform, Laplacian of Gaussian filters, Hough transforms, radial symmetry-based voting procedures or supervised learning. It is common to combine several of these algorithms to achieve the identification [95, 30, 92].

Nuclei segmentation remains one of the most challenging problems in pathology DIA, especially for slides stained with H&E, due to the varying tissue appearance and imperfections during the staining process [85]. Additionally to the challenges introduced by artifacts and challenges before or during the digitization process, the tissue has an inherent heterogeneity, which any DIA solution has to tackle [95]. This heterogeneity includes factors such as varying nucleus sizes, shapes and even intracellular intensity variations. Nuclei/cells can also overlap or touch when they build clumps [95]. Adding to this difficulty, the appearance of epithelial cancer nuclei can differ to a great extent – from almost perfectly round to highly enlarged or irregularly shaped nuclei, containing marginalized or coarse chromatic and prominent sub-nuclear particles (nucleoli) [85]. Furthermore, non-tumor nuclei types such as fibroblasts and lymphocyte nuclei can appear at the same sites as epithelial nuclei, which can hamper the specificity when only epithelial nuclei are sought to be segmented [85]. Also, even when the exclusive identification of epithelial nuclei was successful, they may still be difficult to segment into individual nuclei due to overlapping, clustering or clumping [85, 30]. Last but not least, even small "junk" particles, which also absorbed the Hematoxylin stain, may appear in high grade tumors and complicate the segmentation [85].

In the following sections, an insight into some basic techniques used in nuclei detection and segmentation methods relevant for this work is given.

Binarization

A step ultimately required for object segmentation is the binarization of an image into foreground and background at some stage [40]. At some point, all binarization procedures require a local or global threshold, which can be determined in several ways, e.g. via a method by Otsu [60] as used in [92] or via applying a clustering step like k-means [46] on the image histogram, as used in [72], [40] or [13]. The histogram clustering via k-means by Lloyd [46] partitions the data via iteratively assigning the n data-points to a predefined number k of clusters, each represented by a centroid. The procedure is as follows:

1. Initially, k cluster centers i.e. centroids are chosen
2. The distances between each data-point and each centroid are computed
3. All data-points are assigned to the cluster with the closest centroid
4. In the next iteration, each centroid is updated to the average of all data-points assigned to this cluster
5. Steps 2-4 are repeated until the specified maximum of iterations is reached or until the assignments of data-points to clusters do not change any more

In the original publication by Lloyd [46], the cluster centers are initially seeded randomly. Arthur et al. [2] propose a heuristic for accelerated initial centroid seeding. Regardless of the cluster seed initialization, the entire k-means clustering procedure can be repeated several times to avoid falling into local maxima. The repetition returning the lowest sum-of-distances for all clusters is regarded as the fittest clustering.

The histogram shown in Figure 3.2 displays an example of the result k-means delivers when applied on the histogram of a stain intensity image after deconvolution. The three colors represent the assignment of the each intensity value to one of the three classes.

Morphology Operation

Performing a binary morphological operation describes the filtering of a binary image using a certain structural element – typically a circle, square, cross or other basic geocriterional shapes [75]. It examines the geocriterional and/or topological structures of inherent objects with the defined shape [75]. Some of the basic operators are: dilation, erosion, closing and opening, which in combination extend to operations like boundary extraction, skeletonizing, hole-filling, ultimate erosion and more [95]. Ultimate erosion repeatedly applies erosion to an image until the remaining CC would be removed by another iteration [95]. The resulting CC serve as a marker for each nuclei and can serve as seed-points for consequent separation of touching objects [95]. The difficulty of morphological operations is based on the fact that it relies on perfect binarization, which is hard to achieve on histopathology images [95].

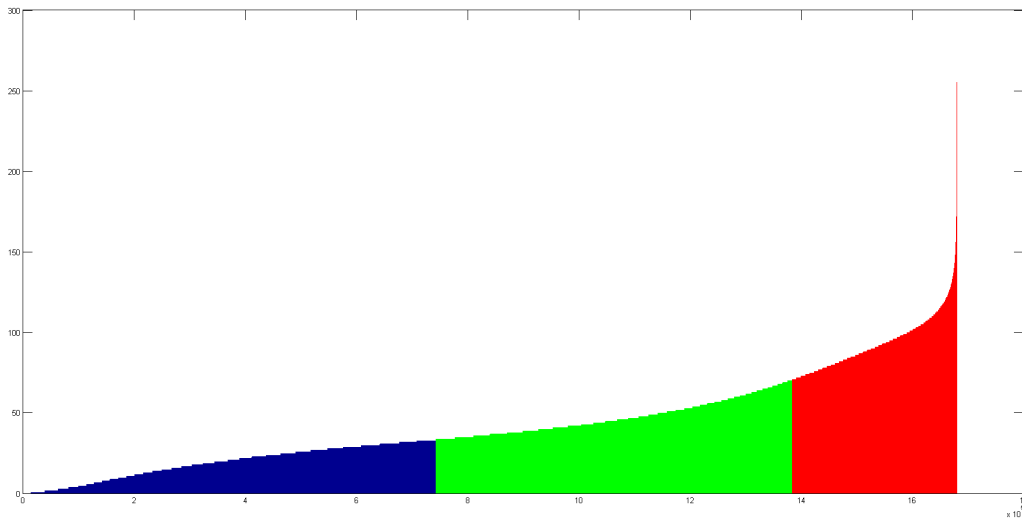


Figure 3.2: A histogram clustered into three classes via k-means

Distance Transform

Generally, the Distance Transform (DT) is a representation of an image, in which the value of distance (in pixels) to the nearest feature point is assigned to each pixel. In nuclei identification, the feature points are made up of edge pixels in a binary image. Most distances used for this use-case are the Euclidean distance [95], less common alternatives are e.g. the Manhattan, Mahalanobis [92] or the Dijkstra distance [41]. The local maxima of the Distance Transform make up the candidates for nuclei centers and require a step for rejecting unfitting candidates [95].

The applicability of DT is limited to regular shapes in binary images, because it results in false local maxima when the edge pixels exhibit even small variations. Thus, it fails in detecting overlapping or touching nuclei. This drawback can be partly alleviated when adding the original intensity to the distance map and /or using a Gaussian filter to eliminate noise. However, these improvements are still insufficient for handling large variances in the complex histological appearance of tissue and can lead to over-detection [95]. DT is usually combined with subsequent watershed segmentation, where the remaining maxima serve as seed points for the flooding [95].

Watershed Transform

The input for a watershed transform is a grayscale image typically made up of the DT of a binary image, which is inverted such that the foreground pixels farthest away from the object borders appear as local minima. The functioning principle is to regard the input image as a topographical surface which is flooded starting from each local minimum, called *basin*. The water level increases constantly across the image, lower minima are flooded previous to higher minima. At each point where the floods stemming from two

or more basins would touch, a dam is built. It represents the watershed line between those two local minima and is built higher with the rising flood to prevent the basins from merging. When all basins are flooded and all floods would have merged, the process is finished.

The result is a binary image in which the dams, i.e. the watershed lines make up the foreground (1) and all so-called *catchment-basins* between watershed lines form the background (0), where each basin contains precisely one previously flooded minimum [5]. To avoid oversegmentation during watershed, the inverted DT can be processed prior to the watershed transform. Such processing may include weighting the DT with a Gaussian kernel [94] or combining it with the intensity gradient information [57]. A common method also incorporates a priori knowledge into the watershed algorithm by inducing markers into the image [30], which can be generated via preprocessing steps such as spatial filtering or morphological operations [24]. The seeds of previously detected nuclei can serve as markers [85]. When markers are then defined as the only allowed regional minima during watershed, noise or over-segmentation can be greatly reduced [24].

Blurring/Gaussian Filtering

Blurring belongs to a range of image filtering methods to enhance or reduce certain image features, such as edges or noise [41]. In a spatial filtering operation a filter mask, typically much smaller than the image, is moved from pixel to pixel across the entire image and conducts a so-called convolution, where the values of the filter mask are multiplied with the current image pixel values at each position [41]. Despite having high resolution images available (see Section 3.2), it is necessary or helpful to remove details such as noise by applying a blur in the form of a Gaussian smoothing filter to the image [30]. In a step for mitosis segmentation, Sirinukunwattana et al. [74] employ blurring to remove noise in the red channel before applying a threshold to binarize it. According to [95], two cell segmentation algorithms are based on shortest path searching within the previously blurred and regularized image.

3.5 Digital Imaging Analysis Approaches in Ki-67 Assessment

While there are numerous publications about nuclei detection and segmentation approaches for H&E and IHC stained images, solutions for other specific stains such as Ki-67 have been more seldom tackled by the research community: in a review by Xing et al. [95] about cell detection and segmentation in digital pathology, algorithms for Ki-67 are not even mentioned. Alike many other breast cancer diagnosis steps, in the assessment of Ki-67 stained slides the detection of nuclei and the identification of nuclei features such as size, shape, quantity or chromatin texture are vital factors [85]. There are only a few publications on the automated assessment of the LI in breast cancer and most of them do not simultaneously tackle nuclei detection and segmentation. In a 2014

review about breast cancer histopathology image analysis ([88]), only a single publication dealing with the segmentation of Ki-67 stained breast images is mentioned ([80]). This attests how underrepresented the field of DIA solutions for Ki-67 LI scoring is. However, it is also evident that a robust method for digital assessment would help reduce the time necessary to render a Ki-67 scoring, reduce intra- and inter-observer variabilities and enhance the accuracy and reproducibility [12]. Following, the approaches and results of some works evolving around the Ki-67 assessment, also in other domains than breast cancer, will be highlighted.

One publication dealing with the automatic scoring of the Ki-67 proliferation index is [94], it is specialized on Neuro-Endocrine Tumor (NET) cases. In order to obtain a Ki-67 score, they combine seed detection followed by segmentation and cell feature extraction with texture feature extraction to derive a tumor/non-tumor classification and subsequently classify into positive and negative nuclei [94]. In a supervised manner, they train their algorithm on 20 and test it on 109 images of the same annotated dataset [94]. They reach a Precision, Recall and F1-score of 0.89, 0.91 and 0.90, respectively [94].

A work on Ki-67 expression in prostate cancer by Desmeules in 2015 compares the visual estimate of the LI found by pathologists against the estimation obtained with an existing DIA method in TMAs [12]. Readily available software packages for tissue recognition and nuclei segmentation by Calopix and Agfa Healthcare were used. Both steps included manual re-adjustment and verification. The comparison is done via comparing the distributions of both estimation methods described by mean, standard deviation, median and IQR (Inter-Quartile Range) of the LI over the whole dataset of 225 patients. The means of visual estimate and DIA are 2.23 (Standard Deviation $SD = 1.98$) and 2.05 ($SD = 1.74$), while the medians are 1.61 ($IQR = [0.71, 3.23]$) and 1.48 ($IQR = [0.86, 2.83]$), respectively. Means as well as medians lie close together. The comparison of these distributions alone would not necessarily prove any correlation if regarded over the entire dataset, thus an additional statistical test was conducted on the mean distributions, proving that visual estimates and DIA give similar results [12]. Two drawbacks of the presented DIA solution are the necessity for manual input during the tissue recognition and segmentation phase and the manual setting of thresholds based on direct visual judgement.

Some DIA solutions are based on plugins of the freely available open source image analysis software ImageJ⁸. The available methods can e.g. serve to work out the potential correlation between DIA-derived and manually derived LI as in Kim et al. ([37], assessing the Ki-67 LI in meningiomas). They report a 0.98 correlation coefficient between the medians of DIA and expert-derived LI, while taking around 11s for the DIA output on each High Power Field⁹ (HPF) at 40x magnification [37]. The publicly available, online DIA solution presented in [80] uses one user input, non-adaptive color deconvolution and subsequent adaptive thresholding to achieve the percentage of positively (Ki-67) stained nuclear area per total nuclear area. With the help of a non-linear correction function

⁸Image Processing and Analysis in Java, <https://imagej.nih.gov/ij/>

⁹Portion of a WSI visible when applying the highest available zoom

on the relation between visual LI (ground truth) and DIA LI, the method yields a final LI correlation of 0.98 [80]. They also point out the limitations of their software when confronted with badly stained samples and allow a background correction via integration of a blankfield image¹⁰ [80].

A widespread approach for nuclei detection, not only in Ki-67, is to fit each identified nuclei segmentation mask to an ellipse with the same second moments as the detected nucleus mask [22, 38, 89]. Ellipses are also used as shape priors for classifying/detecting nuclei, where un-elliptical candidates are rejected [4] or as a mean to separate clumped nuclei [73]. In [92] they use a minimum-model approach to detect nuclei in H&E images, independent of their shape. They point out that a shape prior such as the "roundness" would introduce a bias of excluding relevant, but highly pleomorphic, i.e. irregularly shaped, nuclei. While the practice of ellipse fitting or using it as a shape prior might yield a visually more attractive output than potentially ragged or dented outlines of nuclei segmentations, it carries the risk of occluding physiologically essential information from the viewer.

As mentioned before, there is no consensus on the site or sequence of the Ki-67 scoring. It has been suggested that the Ki-67 LI can coarsely be assessed using visual judgement in 10%-steps on hot spots at only 10x or 20x magnification, which yields a fair correlation to the ground truth LI of 0.94, however leaving a grey zone between 10% and 30% where more rigorous assessment is required to derive a final and reliable Ki-67 assessment [26]. Another assessment method utilizing a step-wise counting strategy was suggested in [65]. Despite the undisputable importance of Ki-67 assessment, the large variety of publications trying to conclude a final and universal assessment method and LI definition underlines the need for a reliable and reproducible assessment method which can adapt to yet varying definitions of a hotspot and LI-thresholds and still save time for the pathologist.

In general it can be said that the more a method relies on assumptions about the characteristics of nuclei, the more likely it is that it wrongfully misses or rejects relevant candidate nuclei, because these dependencies can induce instability [30]. Examples include the dependency of concavity point detections on correct curvature segmentation, the dependency of region growing approaches to shape and size of nuclei, the dependency of marker-controlled watersheds to correct nuclei seeds and the inability of ellipse fitting to include arbitrary nuclei shapes [30]. Adding to this, such assumptions also require prior knowledge [30]. All of these factors make the exact segmentation of nuclei, especially when presenting large overlapping or touching portions, an ongoing challenge in the research domain [30]. This difficulty can be seen for example in a work by Laurinavicius et al. [43], which tested the accuracy of Ki-67 DIA estimation as delivered by a software commercially available since 2011. They had to use several manual tuning and calibration iterations, including the knowledge of ground truth reference values, to reach a misclassification rate (patient considered "positive" or "negative" with respect to a given LI threshold) of 5-7%.

¹⁰A slide scanned empty, without specimen

3.6 Evaluation of Digital Image Analysis Solutions

The metrics reported as results of DIA solutions differ widely from publication to publication, making it hard to define a set of analysis method for comparing performances of different solutions [7]. While publications aiming at classifying tissue in a binary manner (such as malign or benign) tend to report the accuracy or error, these ratings are futile for object detection tasks [7]. There, the declaration of sensitivity and specificity or precision and recall, respectively are more descriptive [7]. Regardless of the metrics reported, all evaluations have to be based on the comparison with some form of ground truth.

The more complex the method to be evaluated, the more difficult or tedious it is to obtain reliable ground truth data for comparison [68, 96]. For instance the evaluation of a method not only limited to the correctness of the object count, but also including measures about the equality of the individual objects in segmentation requires higher quality of the test data than for mere object count correctness [68, 96].

In terms of evaluating segmentation solutions, the image to be segmented is referred to as *Test Image*, the outcome of the segmentation is called *Segmented Image* [96]. The reference image, to which the segmented image is compared is referred to as *Gold Standard* or *Ground Truth* [22, 96].

In cases where the Test Images consist of real, clinical images the Gold Standard is derived by manual creation of annotated/segmented images via human visual inspection of the test images [96]. In cases where the Test Images are generated synthetically, the Gold Standard is typically derived from the initial image generation procedure [9, 96], as demonstrated in Figure 3.3.

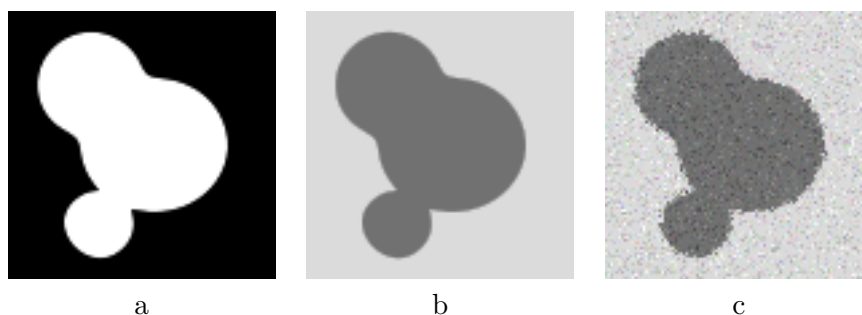


Figure 3.3: Example for Gold Standard extraction from image generation procedure: **(a)** Ground Truth object mask **(b)** Synthetic test image **(c)** Another version of the synthetic test image

In the following sections, possible criteria for the evaluation of nuclei quantification and nuclei segmentation are given. Afterwards, details on evaluations using real annotated images and synthetically derived images as Ground Truth respectively are given.

3.6.1 Nuclei Quantification

The estimation of the Ki-67 LI is based on the correct identification of nuclei. Detection algorithms like [88] base their validation on the Euclidean distance between the located nucleus center and the ground truth nuclei center. In this case, the Ground Truth only contains a single pixel location per nucleus, thus the method requires the definition of a hard global distance threshold between located nucleus center and ground truth nucleus center. In case the located nucleus center is below this threshold, it is counted as a True Positive and otherwise as a False Positive [88]. Furthermore, if multiple nuclei are detected within the threshold distance of a ground truth location, they are only counted as one True Positive [88]. The global threshold is set according to statistical information about the average size of the nuclei to be detected [88]. The criteria of *Precision* (also often referred to as "Positive Predictive Value" [87]), *Recall* (also often referred to as "Sensitivity" [95]) and sometimes *F1*-measure are commonly utilized to quantify the nuclei estimation [95].

3.6.2 Nuclei Segmentation

As stated in [96], there are different approaches to evaluate a segmentation algorithm. The outcome can be evaluated analytically by directly examining and assessing the principles and properties of the algorithms, or empirically by either assessing goodness properties or discrepancy values [96]. Empirical evaluation via goodness includes implicit object characteristics such as intra-region uniformity or inter-region contrast and, alike the analytical evaluation, does not require a priori information about the correct segmentation [96]. On the other hand empirical evaluation via discrepancy measures requires ground truth segmentation [96]. Generally speaking, the value of the discrepancy measure implies the error between the segmented and the ground truth image [96]. Evaluation based on discrepancy measures is the most commonly used method in nucleus segmentation solutions, where the ground truth provides detailed information about the outlines of every single nucleus [95]. In these cases, criteria such as the common Dice similarity Coefficient (DC) published in 1945 [14] and used for example in [35, 87, 89, 88, 97] or the less frequently used Jaccard Index [22] are applied and serve in two ways: they evaluate the segmentation quality and they can also be used as the basis for labeling a detected nucleus as either True Positive, False Positive or False Negative [92]. The equation and portrayal of the Dice Coefficient can be seen in Figure 3.4.

It has to be added, that in DIA algorithm evaluation, the criterion of True Negative is not utilized as the image-nature of the data does not allow a useful definition of objects which are correctly identified as negative. Examples for publications which use True Positive, False Positive and False Negative but not True Negative as criteria for evaluation of nuclei segmentation, are [88, 91, 92].

Some segmentation algorithms also report the accuracy of their segmentation results based on non-standardized criteria such as manually labeling of segmentation outputs as correct or erroneous [86] or defining a certain percentage of contour points lying on the

$$DC = \frac{|A \cap B|}{\frac{1}{2}(|A| + |B|)} \quad \frac{1}{2} \left(\text{Diagram (a)} + \text{Diagram (b)} \right)$$

(a)
(b)

Figure 3.4: The Dice Similarity Coefficient DC as a measure of overlap between two objects, weighted by their average area **(a)** Formula **(b)** Graphical portrayal

true contours as "good segmentation" [66]. [77] report their segmentation accuracy in terms of R_0 , R_1 , R_2 and R_3 , which are expressed as follows:

$$R_0 = \frac{\text{number of pixels well classified}}{\text{number of pixels of the image}} \quad (3.1)$$

$$R_1 = \frac{\text{number of nuclei pixels well classified}}{\text{number of nuclei pixels of the image}} \quad (3.2)$$

$$R_2 = \frac{\text{number of background pixels well classified}}{\text{number of background pixels of the image}} \quad (3.3)$$

$$R_3 = \frac{R_1 + R_2}{2} \quad (3.4)$$

While these rates answer three different questions regarding the classification accuracy, a single rate could be more easily interpreted [45]. Furthermore, it lacks explanation about the quality of the segmentation, meaning whether the segmentation yields correct regions or merely an accurate rate of well classified pixels [45].

The work of [43] uses the Pearson product-moment correlation coefficient, in short *Pearson's Coefficient*, ρ , to report the accuracy of the Ki-67 LI estimation done by DIA in comparison to visual estimates by pathologists (Equation 3.5):

$$\rho(X, Y) = \frac{cov(X, Y)}{var(X) \cdot var(Y)^{\frac{1}{2}}} \quad (3.5)$$

The coefficient ρ gives the degree of linear relationship between two variables, and can lie between -1 (a perfect negative, linear correlation) and $+1$ (a perfect positive, linear correlation), where 0 would mean that no linear correlation is given, however it does not rule out a possible non-linear correlation [23]. The value of ρ is more descriptive as a measure of dependence than covariance alone because ρ is not affected by the scale of X and Y . In the case of [43], the two variables whose relation is given by ρ are the Ki-67 LI estimation of a DIA solution and the Ki-67 LI estimation of the visual estimates, respectively.

In comparison to other measures of association, e.g. the covariance alone, it is invariant to changes of location and scale of the variables. The correlation coefficient ρ never exceeds 1 and its sign can be positive or negative, depending on the covariance. If $\rho = +1$, it states that there is a perfectly linear, positive relationship between the variables (if X increases, then Y increases), while $\rho = -1$ also describes a perfectly linear relationship, but it is negative (if X increases, then Y decreases). In cases where X and Y are entirely uncorrelated from each other $\rho = 0$ is true [23]. For practical purposes however, a correlation is usually only assumed if ρ exceeds a certain threshold, which has to be defined for each individual situation [17]. An obvious feature of interest for the pathologist is the area of the nuclei [68]. Other features include descriptors to divide between typical and atypical cells [76] to assess the nuclear pleomorphism [18], such as solidity and eccentricity. These are both criteria which describe morphological features of a nucleus at an object level [57, 61, 71, 88] and often they are already used in the post-processing steps of nuclei segmentation to discriminate between likely and unlikely candidates [7, 39, 68, 86, 89, 88]:

Solidity describes the ratio between the area of an object and the area inside the objects convex hull [7, 68, 86, 89] and is used mainly in post-segmentation classification to discriminate between likely and unlikely object candidates, based on the assumption that nuclei seldomly exhibit concave shapes [68, 86, 89]

$$\text{Solidity} = \frac{\text{Area}_{\text{Object}}}{\text{Area}_{\text{Convex Hull of Object}}} \quad (3.6)$$

Eccentricity is defined as the ratio of the lengths between the foci and the major axis length of the ellipse that best fits the object [71], i.e. an ellipse with the same second-moments as the object. An eccentricity of 0 describes a circle while an eccentricity of 1 describes a line [8]¹¹. According to [39], properties of elliptical shape models, such as Eccentricity, are among the most prevalent in digital pathology solutions because they are very indicative of cancer predicates.

$$\text{Eccentricity} = \frac{\text{Length}_{\text{Major Axis}}}{\text{Length}_{\text{Minor Axis}}} \quad (3.7)$$

Sections 3.6.3 and 3.6.4 lay out possible applications and benefits of using manually labeled or synthetically derived images as ground truth, respectively.

3.6.3 Ground Truth from Pathologists

Generally, cell image analysis algorithms are validated using a set of representative images which are labeled by one or more subject matter experts, also referred to as observers or investigators [37, 68]. The results from the analysis method are compared to these labeled images [68]. This practice is laborious, and the inter- and intra-observer

¹¹p. 226f

variability in the judgements puts a limit on the validity of the labeled data [52, 68]. In [21], the inter-observer variability for plain nuclei detection exhibits a precision and recall of only 0.92 and 0.91, respectively, when comparing the results of one observer with the *ground truth* labeled by another observer. It is concluded that the apparently straight-forward job of nuclei detection on a histology slide is not an unambiguous task [21]. The intra-observer variability on classification of nuclear atypicalness also averages to 21.2% in this study, the inter-observer variability to even 42%. Detection of cell nuclei, especially if it includes the precise segmentation/delineation of the nuclei, is the most tedious labeling task [21]. Furthermore, to ensure statistical soundness, especially when dealing with biological data in its infinite forms of appearance, nearly impossible quantities of image data would be required to be manually labeled [52].

The academic community greatly benefits from publicly available datasets [21]. However, due to the extensive workload required for their creation, still only a small number of labeled datasets is available [21]. As of the make-span of this work, no such dataset included labeled, let alone segmented Ki-67 images of breast tissue.

3.6.4 Ground Truth from Synthetic Datasets

While using real images as ground truth may be more suitable for the respective use-case, they may not prompt evaluation results appropriate for other domains of application [96]. Therefore, one of the advantages of synthetic images as gold standard is that their generation can be well controlled and they are easily reproducible. This also allows for transferring evaluation results from one application domain to another [96]. Benchmark datasets made of synthetic data bear several benefits: results are comparable among several works, the parameters can be fitted to the respective usecase, observer bias and variability can be ruled out entirely and, foremost, the labeled ground truth can be retrieved [68]. As a consequence, several pathology DIA algorithms validate their solutions on synthetic data. There have been efforts to create software for synthesizing/simulating microscopic images, such as [44], for fluorescence microscope images [58], for multi-parameter cell body images via creating populations with tunable variations among cellular object-level features such as area, length and solidity [68], for fluorescence microscopy images, [52] for pap-smear microscopy images, [48] for Ki-67 hotspot clusters or [1] for synthesizing whole slide images. However, except for [1], all of the present-day solutions are designed for the simulation of cytology rather than histology images, which makes them inapplicable for the creation of Ground Truth datasets for testing histology DIA solutions. The framework presented in [1] applies texture synthesis and texture placement in predefined regions, using textures (cells, fiber pieces, ducts, ...) extracted from real images stained with CD8 or H&E.

3.6.5 Summary

Digital pathology basically describes the digitization of a physical pathology slide into an image, which can be viewed and annotated on a computer screen. This field has grown rapidly in the last two decades. It aids towards creating a thorough digital patient record and allows several examiners to work on the same slide as well as conduct large retrospective studies on existent cases. Conducting diagnosis based on digitized slides, so-called Whole Slide Images (WSI), offers a cost-effective and efficient alternative to glass slides viewed via the microscope. With resolutions of up to 0.25 micrometers per pixel the resulting file sizes range up to 3GB of data per glass slide. These large quantities of data pose a challenge for both their storage and handling. The largest hopes of pathologists from the introduction of digital pathology are pinned on factors like improved ergonomics, diagnostic accuracy, measurement accuracy, time saved and speed of slide navigation. Furthermore, digital image analysis methods allow the application of computer-aided diagnosis, enabling a high throughput rate while reducing bias introduced by a human examiner and instead delivering reproducible and accurate estimates of diseases. Digital pathology can thereby greatly reduce variability in the diagnosis, which is one of the greatest challenges in traditional, analogue pathology.

Any digital image analysis solution in digital pathology needs to be robust to several artifacts as they commonly occur in slides, such as tissue deformations, fixation errors, background clutter, noise or poor contrast. Methods on slides with color- and contrast-related artifacts can profit from color deconvolution, which describes the process of digitally separating the stains which were originally applied to dye and highlight different tissue structures. Several methods have been proposed to solve this task. Some assume a linear relation between the stain appearance on the digital slide and the intensity of the stain applied. They conclude on the original stain colors by information implicitly present in the image. Other methods require manual input or information about the stains to be separated.

A common task tackled with image analysis tools is the detection and segmentation of tissue structures such as nuclei. The major challenges in segmentation can be found in the heterogeneity of nucleus sizes, shapes and even intra-cellular intensity variations, as well as overlapping or touching cells. Methods published in this field employ operations such as thresholding, morphological operations, distance and watershed transforms or gaussian filtering. There are many solutions suggested for nucleus detection and segmentation with the majority focusing on H&E-stained slides. Solutions for Ki-67 stained slides are less frequently proposed and require manual inputs or use a large set of assumptions, which both limit the respective method robustness.

The evaluation of image analysis solutions for digital pathology is not standardized. It can be conducted with the use of an annotated dataset of real digitized slides, where the annotation can either describe mere nucleus locations or to actually delineate segmented nuclei or cells. The annotations are manually drawn by pathologists, which is a time-consuming task, thus there are few publicly available benchmark datasets to compare the

performance of different digital image analysis solutions. As an alternative, an algorithm can be evaluated on synthetically generated images, where the ground truth is implicitly generated during the synthesis process. The advantage lies in the exact, pixel-wise ground truth available, however, synthetic images cannot indistinguishably recreate a realistic image.

Independent of the ground truth used for the evaluation, the research community largely agrees on the criteria to be tested. Concerning the object quantification commonly applied measures are precision, recall and F1-score. For the segmentation performance, the dice coefficient is used as a measure of overlap between objects found by the algorithm and ground truth objects.

Suggested Solution

This chapter deals with the implementation of the work to answer the research question stated in Section 1.3. To begin with, the motivation, design and realization of the synthetic Ki-67 dataset is explained in Section 4.1. The subsequent Section 4.2 deals with the deconvolution of the digital RGB image into relevant channels as well as the first thresholding step and explains the proposed solution for nuclei segmentation and quantification required to obtain the sought Labeling Index. Finally, the steps taken to ensure a thorough algorithm evaluation are covered in Section 4.3. If not mentioned otherwise, all processes are implemented in Matlab.

4.1 Synthetic DataSet (SDS)

As outlined in Section 3.6.4, the standards for an SDS in digital pathology are yet to be defined. In this work, the SDS is designed and built without using existing software solutions for cell body synthesis such as [68] or H&E whole slide [1].as their solutions deliver an inappropriate type of images for developing and testing an algorithm on Ki-67 analysis.

An overview about the entire process of creating synthetic Ki-67 images is illustrated in Figure 4.1. The upper sector *Initial Synthesis*, which is explained in greater detail in Sections 4.1.1, 4.1.2 and 4.1.3, portraits how the two components, namely background and nuclei placement, blend to form an initial synthetic image, SI_{init} . The lower sector *Restaining*, explained in greater detail in Sections 4.1.4 and 4.1.5, illustrates how staining characteristics extracted from real Ki-67 images serve as a basis for creating synthetic images $SI_{restained}$ with varying stain appearance. Agreeing with the appearance of cells in real Ki-67 images, which do not recognizably show the existing cytoplasm or cell membranes around the nuclei (see Figure 4.2 for an example), the synthetic images are created with specifically placing nuclei without cytoplasm, membrane etc. rather than entire cellular structures.

4. SUGGESTED SOLUTION

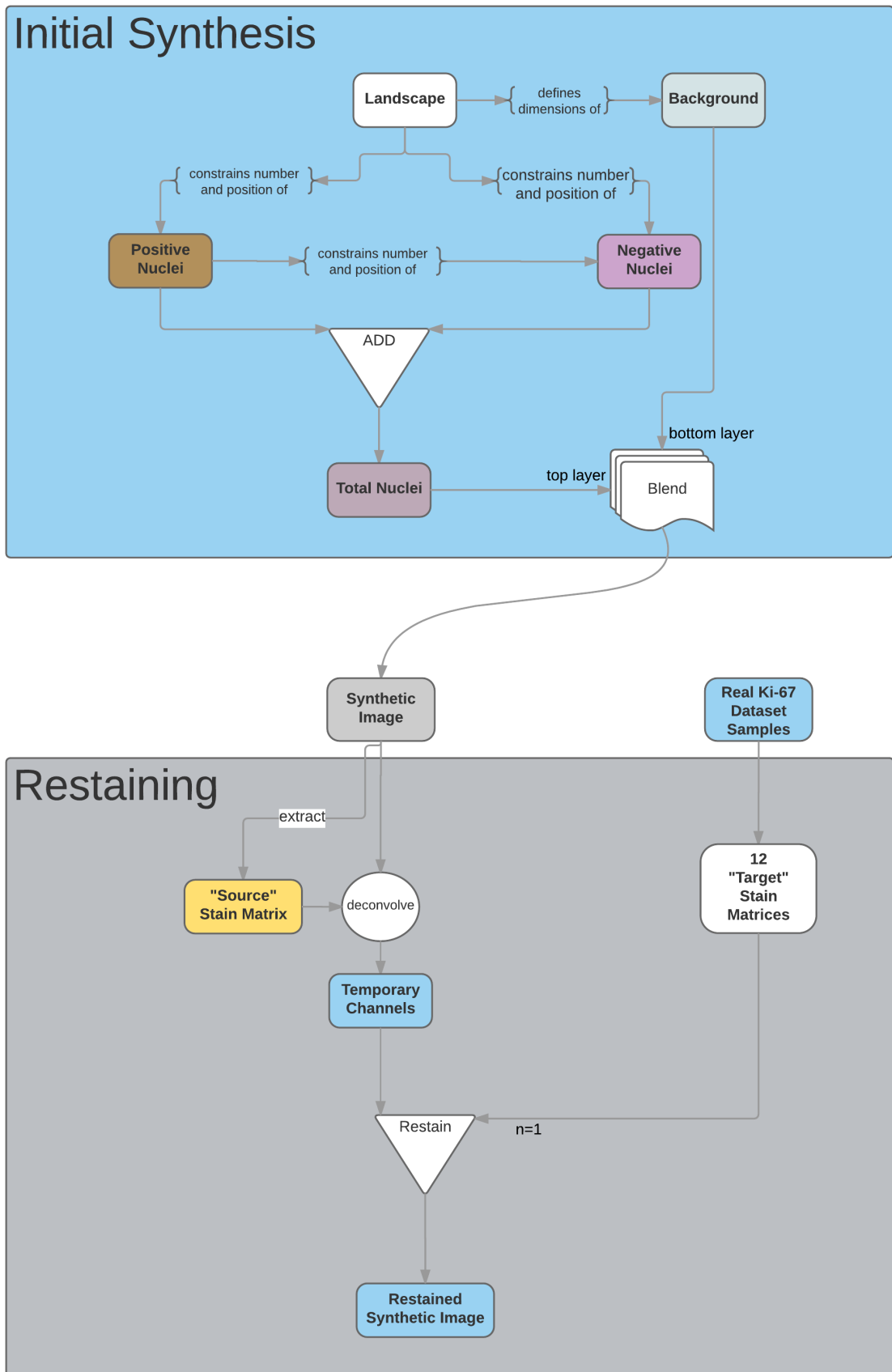


Figure 4.1: Overview over the image synthesis process

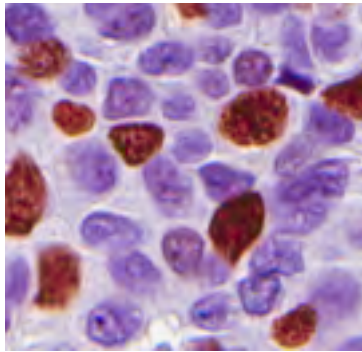


Figure 4.2: Exemplary region of real Ki-67 image, showing Ki-67 positive (brown) and negative (purple) nuclei and barely any cytoplasmic structures or membranes.

The synthetic images are generated with four different resolution. These dimensions are motivated by the following factors:

1680x1050 pixels resolution: The resolution of commonly available computer screens is included to mimic realistic viewing conditions on full-screen for both clinical and research personnel at the highest available zoom level (20x). This resolution is in line with the order of magnitude shown in a microscopic high power field, as also utilized in [88].

1712x980 and 3214x1803 pixels resolution: Two regions of interest of realistic tissue scenarios were also recreated during the synthesis procedure, prompting those specific resolutions.

1200x1200 pixels resolution: Another resolution within the same order of magnitude as the other three was also included to enrich the dataset diversity.

All of the listed pixel resolutions are used in the creation of the SDS. Ultimately, using the presented synthetic image generation method, any image resolution can be generated as long as the synthetic background is generated accordingly and the dimension of the nuclei placement probability map (see Section 4.1.3 for details) agrees.

4.1.1 Creation of Synthetic Background

Even though the nuclei in the synthetic images shall be represented without their cytoplasm or membrane, physiologically they are always embedded in a non-uniform tissue structure, which is perceived as the background of the nuclei in the image by the human eye. The structural and chromatic appearance of this tissue background can severely facilitate or complicate the automated segmentation of nuclei in an image, for example because the structures between nuclei and background seem to interweave or because the background has a similar stain appearance as the nucleus, making the contrast too weak to easily define the border. To mimic a realistic background appearance, a small texture element (see Figure 4.3) from a nuclei-vacant area of a real Ki-67 image was extracted and propagated to fill images with defined resolution.



Figure 4.3: The structural basis element for the creation of the embedding tissue of the synthetic images.

The synthetic background used in the SDSs was generated using a method published in [25] and made available as a plug-in within the GNU Image Manipulation Program (GIMP) called Resynthesize ¹. Given a sample of a texture, it can recreate more of this texture in a randomized manner. The method requires the input of the texture to be resynthesized and the definition of the dimensions for the synthetic image output. Furthermore, three parameters have to be tuned and set for the resynthesis process:

- The neighbourhood size: Adjustable between 1 and 100. It sets how many nearby pixels are to be taken into consideration in the output image. The value chosen empirically for this work is 8.
- Search thoroughness: Adjustable between 1 and 500. It sets how many locations in the input image are examined to find the best match. The value chosen empirically for this work is 200.
- Sensitivity to outliers: Adjustable between 0.00 and 1.00. It sets the allowed error, where 0.00 allows and 1.00 disqualifies a very bad match on a single pixel. The value chosen empirically for this work is 1.00 to ensure a resulting texture without any discontinuities.

4.1.2 Nuclei Extraction from Real Data

The nuclei were extracted from a region within a real Ki-67 image deemed to be a good representation of Ki-67 by a collaborating pathologist. In order to extract the nuclei, the coordinates of this region within the .svs-file were read out using a WSI viewer ² and the desired region itself was stored in a .tif format using the OpenSlide Library in Matlab, which is able to access .svs-files. The region was opened in the Gimp for the actual nuclei extraction. The utilized region of the .svs-file was available in a 20x magnification.

The following procedure to outline the nuclei took place each for the positive and negative nuclei:

On a transparent secondary layer, the outlines of the nuclei were traced in red, giving good contrast to the underlying image. Close attention was paid so that the tracing line lies on the outside border of the nuclei in order to guarantee that the nucleus border

¹<http://registry.gimp.org/node/27986>

²Aperio ImageScope v12.1.0.5029

information would stay intact and part of the resulting mask. Not only solitary nuclei were traced, but also nuclei in adjacent or overlapping arrangement or obviously being in the process of proliferation. The background of the resulting mask as well as the traced borders were filled with black, resulting in a binary mask revealing only the segmented nuclei as foreground. This mask was stored as the Cut Mask, $Mask_{Cut}$, which serves as the evaluation basis for the segmentation performance later described in Section 4.3.2. A duplicate of $Mask_{Cut}$ was dilated and subsequently eroded using a small structuring element (3x3) to merge adjacent nuclei. The 3x3 structuring element was chosen because only objects with a maximum distance of one pixel to each other should be merged by this process. The merging resulted in a "Connected Mask", $Mask_{Conn}$ which serves as the basis for the actual nuclei extraction described in the next paragraph.

Again, the following procedure to produce individual nuclei images was conducted each for the positive and negative nuclei:

The $Mask_{Conn}$ was opened in Matlab and for each "Snippet" (i.e. a CC describing a single nucleus or a clump of overlapping/adjacent/proliferating nuclei), three image objects were stored as .tif-files:

- S_{pos} or S_{neg} : the corresponding region in the RGB Ki-67 .tif-file (I_{RGB} , containing the RGB image of the nucleus with all non-nucleus pixels in black)
- $SnipMask_{Cut}$: containing the corresponding region in the $Mask_{Cut}$
- $SnipMask_{Conn}$: containing the corresponding region in the $Mask_{Conn}$

A naming scheme was applied which assigned a unique identifier to each nucleus (or group of nuclei) and specified the number of nuclei within each Snippet S_{pos} or S_{neg} . The outlining step as well as the results of an S_{pos} , $SnipMask_{Conn}$ and $SnipMask_{Cut}$ are shown in Figure 4.4.

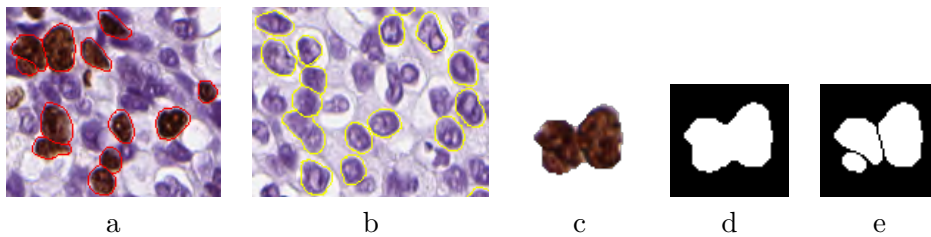


Figure 4.4: (a) Outlined positive nuclei in Gimp (b) Outlined negative nuclei in Gimp (c) Sample of S_{pos} (d) Sample of $Mask_{Conn}$ of this S_{pos} (e) Sample of $Mask_{Cut}$ of this S_{pos}

A total of 129 positive Snippets S_{pos} were created via this manual delineation procedure, whereof 81% (104 S_{pos}) are individual nuclei, 16% (20 S_{pos}) hold two, 2% (3 S_{pos}) hold three and <1% (1 S_{pos}) holds four touching or overlapping nuclei N_{pos} .

A total of 203 negative Snippets S_{neg} were created, whereof 82% (167 S_{neg}) are individual nuclei, 13% (27 S_{neg}) hold two, 4% (8 S_{neg}) hold three and <1% (1 S_{neg}) holds five touching or overlapping nuclei N_{neg} .

4.1.3 Nuclei Placement

The Snippets are randomly placed on the previously generated synthetic background (see Section 4.1.1). The randomness is introduced to avoid a grid-like and artificial appearance of the image. Yet, this randomness is constrained by two decisive factors:

- (1) the probability of nuclei placement for each location in the image is defined using probability Maps $Map_{NucProbability}$ and
- (2) the nuclei snippets S_{pos} and S_{neg} may not overlap or touch unless explicitly specified. These factors will be explained in more detail in the following paragraphs.

(1) The probability of nuclei placement for each location in the image is defined using probability Maps $Map_{NucProbability}$

Using probability maps $Map_{NucProbability}$, the placement of the nuclei snippets S_{pos} and S_{neg} can be arbitrarily constrained to varying probabilities in different areas if desired. This is intended to result in varying densities of the nuclei in different areas of the image and the more nuclei are placed in total, the more easily noticeable this effect is. A $Map_{NucProbability}$ can be designed either randomly (*random likelihood*) or manually (*manual likelihood*), and can differ for positive and negative nuclei, as demonstrated in examples in Figure 4.5. The constraint for snippet placement can also be left out entirely, which leads to a uniform placement likelihood across the entire image area. Thus, this case is referred to as *uniform likelihood* and accordingly the $Map_{NucProbability}$ is uniformly white, which explains the seemingly empty rightmost column in Figure 4.5.

Each $Map_{NucProbability}$ can have values between 0 (black) and 1 (white), where 0 defines areas where nuclei placement is strictly prohibited and 1 defines areas where nuclei placement is unconditionally allowed. To give an example: A value of 0.5 defines that the likelihood for nuclei placement in this area is at 50%. For both positive and negative nuclei placement, one $Map_{NucProbability}$ needs to be specified. If only one $Map_{NucProbability}$ is specified in total, it is used for placing both nuclei types.

(2) Nuclei Snippets may not overlap or touch unless explicitly specified

As can be seen in Figure 4.6 (Nuclei Extraction from Real Data), during the extraction of snippets S not only individual nuclei were obtained, but also touching and/or overlapping groups of nuclei. They serve as a valid ground truth for physiologically appearing nuclei behavior during proliferation phase. During the placement of S on synthetic images, any artificial overlap or merging is thus prohibited to avoid the introduction of possibly non-physiological cellular behavior. It is possible, however, to allow two S to touch, where touching is defined as two $SnipMask_{Conn}$ lying so close that they merge into one 8-CC (Connected Component), but do not yet overlap, i.e. $A \cup B == |A| + |B|$, as shown in Figure 4.6 (b). Overlapping, i.e. $A \cup B < |A| + |B|$, as illustrated in Figure 4.6 (c), is not allowed.

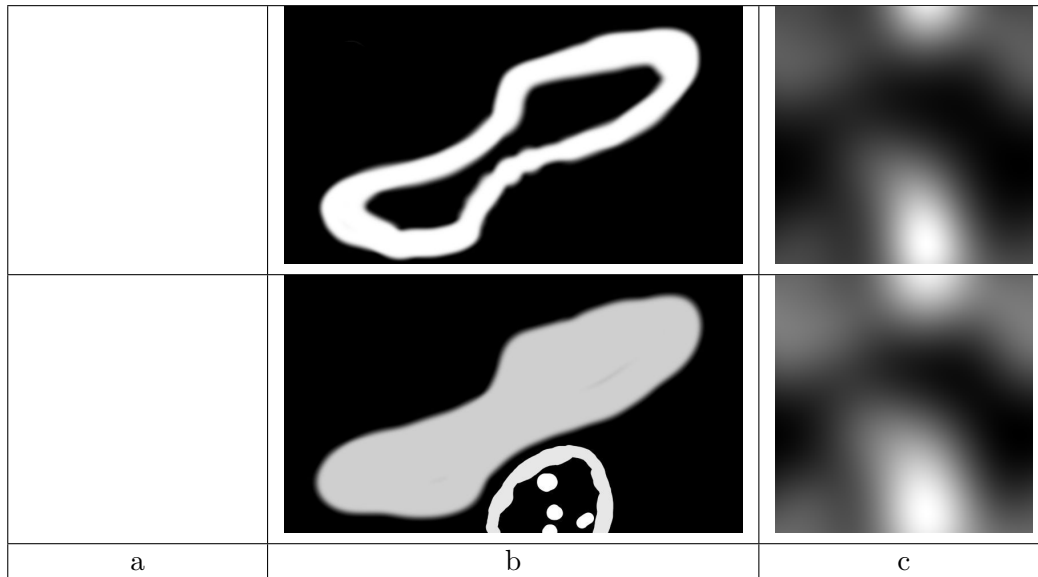


Figure 4.5: Example results of different design methods for $MapNucProbability$, showing areas from 100% placement likelihood (white) to 0% placement likelihood (black)

Upper Map: for the placement of positive Snippets, S_{pos}

Lower Map: for the placement of negative Snippets, S_{neg}

The $MapNucProbability$ can either be (a) uniform (b) manually generated or (c) randomly generated (examples for synthetic images generated with (a) and (b) can be seen in Table 5.1)

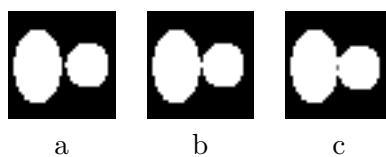


Figure 4.6: Definition of touching and overlapping (a) Nuclei do not touch, two separate CCs (b) Nuclei touch but do not overlap, become one CC (c) Nuclei overlap (grey pixels)

The restraint that nuclei snippets may not overlap is true among S_{pos} and among S_{neg} as well as between S_{pos} and S_{neg} . As will be explained shortly, the S_{pos} are placed first and their placement restrains the possible locations for the subsequent placement of S_{neg} .

Nuclei Placement Procedure

Considering these two constraints, an SI_{init} is generated according to the following procedure, which starts with Steps 1 to 5 defining the basic settings of *where* and *how many* nuclei are to be placed:

1. A $MapNucProbability$ is chosen, either randomly or deliberately
2. The maximum possible total amount of nuclei (positive and negative) to be placed, called $Amount_{possible}$ is restricted by and calculated on the basis of the provided $MapNucProbability$. Each $MapNucProbability$ implicitly has a defined $Amount_{possible}$
3. The actual total amount of nuclei to be placed in SI_{init} , denoted $Amount_{placed}$, is randomly chosen between $[1, Amount_{possible}]$
4. The LI to be generated (i.e. the ratio between N_{pos} and N_{neg} in the image) is randomly chosen between $[0,100\%]$, in steps of 5%.
5. The amount of N_{pos} and N_{neg} are set accordingly, such that the two conditions given by $Amount_{placed}$ and LI are met:

$$\begin{aligned}
 & - N_{pos} + N_{neg} = Amount_{placed} \\
 & - \frac{N_{pos}}{N_{pos}+N_{neg}} \cdot 100
 \end{aligned}$$

In the subsequent steps, the precise placement for every nucleus is defined. These steps are conducted twice: First for placing the positive Nuclei N_{pos} and then for the negative Nuclei N_{neg} . In the following, it is exemplary explained for S_{pos} and N_{pos} .

6. A S_{pos} , holding between 1 and 4 nuclei, is randomly chosen from the available set of S_{pos}
7. The random row and column, Loc_{Row} and Loc_{Col} , for the placement of this S_{pos} are generated
8. A random number between 0 and 1, Rnd_S , is generated
9. Rnd_S is compared against the likelihood P_{Loc} of $MapNucProbability$ at Loc_{Row} and Loc_{Col} . If $Rnd_S \geq P_{Loc}$, then Loc_{Row} , Loc_{Col} as well as Rnd_S are generated anew until $Rnd_S < P_{Loc}$.
10. It is checked whether the chosen S_{pos} overlaps with any previously placed S_{pos} on this SI_{init} via the logical operation $S_{pos} \& Mask_{Conn}$ (at the considered placement area). If $S_{pos} \& Mask_{Conn} > 0$, then steps 5-8 are repeated until a Loc_{Row} , Loc_{Col} are found where $S_{pos} \& Mask_{Conn} == 0$

11. The S_{pos} is placed at Loc_{Row} , Loc_{Col} (left upper corner of S_{pos} is located at Loc_{Row} , Loc_{Col})
12. After each placement of S_{pos} , the $SnipMask_{Conn}$ and the $SnipMask_{Cut}$ is saved to the $Mask_{Conn}$ and $Mask_{Cut}$ for this very SI_{init} .

Like other image synthesis methods proposed, the natural appearance of nuclei can be enhanced via applying a blurring step in order to reduce sharp, unnatural edges between nucleus and other tissue [44, 52]. This happens during Step 11. Hereby, the entire S_{pos} is filtered with a 2×2 averaging filter kernel, but only those pixels within a dilated 5-pixel-band of the snippet boundary are adopted into the synthetic image, thus the inside of the nucleus remains non-blurred. The size of the averaging filter was chosen to be as small as possible in order to retain as much structural information of the nuclei as possible.

4.1.4 Extraction of Staining Characteristic Variability from Real Data

For the purpose of creating a representative variety of stain appearances, as it would occur in a clinical laboratory (see section 2.2 and 2.3.2) it is necessary to create a variability of the staining characteristics among the synthetic images produced. A dataset of eleven real, clinical Ki-67 breast WSI from one Pathology Laboratory at the Linköping University Hospital, Sweden, implicitly offers the necessary information on staining variability. Using information gained via the aforementioned method of adaptive color deconvolution (see Section 3.4.2), the individually specific stain vector of each image within this dataset can be accessed and utilized. The total of all stain vectors then form the basis for introducing a credible and realistic variation into the stain appearances within the SDS.

From a set of 11 available WSI, one representable region was identified in each WSI (one particularly diverse WSI prompted 2 regions) using the software ImageScope, which allows fast pan and zoom operations. The representative regions were chosen to contain both Ki-67 positive and negative cells, i.e. both Ki-67 and H stains. In the case of one WSI, the observed intra-slide stain appearances indicated that the selection of two regions instead of one can substantially add to the diversity of the extracted stain appearances, thus two regions were chosen. The chosen regions were captured as .tif-files at the displayed resolution and the highest available magnification (20x) using the built-in function "Save Snapshot". Thanks to the specifically chromatic purpose of this step, it was not necessary to use a library such as OpenSlide via Matlab to access the native, unaltered pixel-wise data of the region. The resulting snapshots are depicted in Chapter 5, Section 5.1 on page 68).

A stain normalization method by Macenko ([50] implemented by Mitko Veta ³ in Matlab is used to extract the image-specific stain matrices from the 12 representative regions.

³<https://github.com/mitkovetta/staining-normalization> (c) 2013, Mitko Veta, Image Sciences Institute, University Medical Center, Utrecht, The Netherlands

Contrary to the originally intended purpose of this software, which is bringing images with differing chromatic appearance to a uniform, "normalized" appearance, it was used here in a reversed manner:

The generation of synthetic images initially yields images with uniform, normalized appearance. Hence they are deconvolved into two stain channels using a fixed stain matrix and re-stained back to RGB channels using varying stain matrices. These varying stain matrices are harvested from the 12 representative regions (see bottom of Figure 4.1 for clarification on the process summary).

On these representative regions, the identification of the image-specific optimal stain vectors as published in [50] and implemented by Mitko Veta (2013) was realized. The main mathematical operation is carried out on the Optical Density (OD) representation of the image.

$$OD = -\log_{10}(I) \tag{4.1}$$

Equation 4.1 converts the RGB image data (I) into OD vectors. After removal of OD values below a certain threshold (to exclude transparent pixels from further analysis), the eigenvectors of the remaining tuples are calculated. All tuples are projected onto the plane spanned by the two eigenvectors corresponding to the two largest eigenvectors and brought to unit length. Via examining the angle between each tuple and the first eigenvector, the two robust extremes are identified (α^{th} percentile and $(1 - \alpha^{th})$ percentile). The two corresponding initial OD-tuples represent the two prominent stain vectors in this image. They are stored as the image-specific optimal 2×3 stain matrix, where each of the two rows describes the three OD-values for the two stains, respectively. This procedure yields 12 stain matrices M_{target} .

Even though this stain vector identification is intended to identify H&E- and not Ki-67-specific stain vectors, it is of equally essential usefulness when later applied as the basis for restaining of images deconvolved with a fixed H&E stain matrix (M_{source}).

4.1.5 Color Deconvolution and Color Normalization

As pointed out in the previous Section 4.1.4, an essential step in the meaningfulness of the SDS involves the variation of the staining appearance among the generated synthetic images. In order to serve this purpose, the mean of the 12 stain matrices is defined as the *fixed stain matrix* M_{source} designated for image deconvolution. Each of the images in the SDS is deconvolved into the same two channels defined the by M_{source} and then re-stained using one of the 12 available, varying stain matrices M_{target} identified in Section 4.1.4. Both steps are realized using the normalization method published in Macenko et al. [50].

In the following, the most important formulae to understand this process are described. First, the image values are converted to OD-space, equal to the previous section 4.1.4, where (I) again denotes the RGB image data:

$$OD = -\log_{10}(I) \quad (4.2)$$

Then, the image is deconvolved using Equation 4.3, where M_{source} is the fixed stain matrix previously defined:

$$C = M_{source}/OD \quad (4.3)$$

Now, the original three-channel RGB image has been converted into a two-channel image, C , where the two channels correspond to the intensity of each stain described in the fixed stain matrix M_{source} for every pixel location. The intermediate channel images are neither intended nor capable of representing the optimal deconvolution of any given image, but for allowing an insight into the functionality of this process, Figure 4.7 displays a sample image I and the two corresponding channel images, C_{pos} and C_{neg} .



Original RGB Image, I Deconvoluted Channel C_{pos} Deconvoluted Channel C_{neg}

Figure 4.7: Example image of the deconvolution of an RGB image into a positive and negative channel

In the next vital operation, the actual restaining takes place: Multiplication of the two-channel image C with one of the 12 matrices, M_{target} , in Equation 4.4 inserts the new stain appearance. The exponentiation of the entire expression brings an image I from OD-space back into RGB-space, creating a restained image $SI_{restained}$.

$$I_{restained} = e^{-M_{target}*C} \quad (4.4)$$

4.2 Nuclei Quantification and Segmentation

In this section, the algorithm developed within the course of this thesis is presented. First, an overview about the methods used is given and then each subsection expands on the details of one of the methods.

The algorithm aims at analyzing the nuclei in a Ki-67 stained image in order to derive the LI, thus the present proliferative activity. The LI calculation bases on the segmentation outputs of the nuclei in two channels: Firstly, a channel containing proliferating nuclei,

C_{pos} , and secondly a channel containing nuclei which are in non-proliferative phases of the cell cycle, C_{neg} . The nuclei in each channel are segmented and the number of resulting connected components CC in each segmentation mask provides the information for calculating the LI. Figure 4.8 gives an overview of the steps necessary to reach this aim. Each operation is depicted as a blue ellipse and the in- and outputs to each operation are denoted by arrows and images:

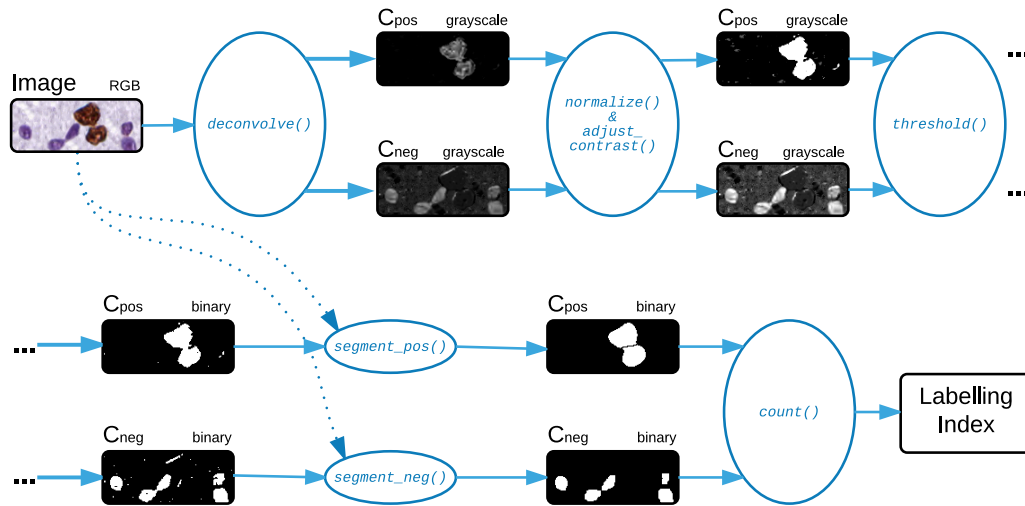


Figure 4.8: Overview of the steps involved in the nuclei segmentation and LI derivation

In the following, each of the operations depicted in Figure 4.8 is coarsely explained, before they are covered in more detail in the subsequent Sections 4.2.1 to 4.2.4:

deconvolve()

First, the RGB image to be analyzed undergoes a deconvolution. In essence, the term deconvolution in this context connotes splitting the information present within the three RGB channels into two channels, which correspond to the two stains of the image: Ki-67 (positive for the proliferation activity stain) and Hematoxylin (negative for the proliferation activity stain). These channels are referred to as positive Channel C_{pos} and negative Channel C_{neg} and initially contain grayscale-values.

normalize() & adjust_contrast()

In this step, the channels are each normalized and their contrast is adjusted, after which their intensities are scaled to $[0, 1]$.

threshold()

A clustering-based thresholding step converts each channel to a binary image.

segment_pos() & segment_neg()

The subsequent segmentations are conducted differently for C_{pos} and C_{neg} and both employ information won from the original RGB image.

count ()

The objects in the resulting segmentation masks for each channel are counted and the Labeling Index is derived via Equation 4.5.

$$\frac{CC_{pos}}{CC_{pos} + CC_{neg}} \quad (4.5)$$

4.2.1 Deconvolution

The deconvolution of the image to be analyzed embodies the first step in the entire processing pipeline (see step `deconvolve()` in Figure 4.8) and the quality of the results hinges on the quality of the deconvolution. To test this statement, three existing deconvolution approaches are considered in this work [11, 50, 67]. They all have in common that they aim at splitting a three-channel RGB image I_{RGB} into two channels, where each contains the intensity values for one of the stains in each pixel. The characteristic of each target channel is typically defined as a three-dimensional vector in a defined three-dimensional colorspace (e.g. RGB, CMY, OD) and the intensity values for one of the stains in each pixel is derived via orthogonally projecting the original pixel value onto this vector. The difference between the methods lies in the principles utilized to arrive at these vectors and once the vectors are identified, I_{RGB} is deconvolved according to Equation 4.6 for all three methods, where M denotes the matrix containing the two stain vectors which characterize the channels:

$$C = M/OD \quad (4.6)$$

Method by Cosatto

The deconvolution approach by Cosatto [11] is based partly on a set of assumptions about the colors of the stains and partly on the histograms of the current image. Thus, it can be described as a semi-adaptive deconvolution. Originally, the method is part of a larger pipeline used to grade the pleomorphism (in this case the area characteristics of nuclei) in H&E stained breast cancer histopathology images. Thus, it targets the deconvolution of H&E stained slides. The deconvolution vector identification process by Cosatto is initiated by converting the image from RGB to CMY. The two color vectors, henceforth called H and E, are retrieved using the following set of Equations:

$$C = \frac{\sum w_i P_i}{\sum w_i} \text{ with } w_i = (P_i^{Cyan})^4 \quad (4.7)$$

$$C = \frac{\sum w_i P_i}{\sum w_i} \text{ with } w_i = \left| P_i - \frac{(P_i \cdot C)C}{C^2} \right|^4 \quad (4.8)$$

$$H = C - \frac{(C \cdot M)M}{|M|^2} \quad (4.9)$$

$$E = M - \frac{(C \cdot M)C}{|C|^2} \quad (4.10)$$

P_i is the CMY color vector for a pixel i . In Equation 4.7, a vector C accumulates colors with a large cyan component, where w_i is the weight assigned to the cyan component P_i^{Cyan} . Next, a vector M , using a different weight w_i , sums up all colors unexplained by the C vector (Equation 4.8). Finally, the vectors H and E are derived by calculating the vectors orthogonal to M and C , respectively (Equations 4.9 and 4.10). These calculations are motivated by assumptions about the properties of the stains, namely that the eosin intensity is correlated to the intensities of the cyan component and the Hematoxylin intensity is correlated to the intensities of the other components. After deriving the vectors H and E , the CMY-intensities of each pixel is projected onto both vectors. The distance between a projected pixel and the point of origin describes the intensity of this pixel in the respective stain. Pixels where the Hematoxylin stain is more intense than the Eosin stain ideally are further along H than along E and vice versa. Despite the fact that this method is intended for use on H&E-slides and the profound assumptions supporting this usage, the usability of the same method for the deconvolution of Ki-67 stained images is tested in this work.

Method by Macenko

The main assumption in the deconvolution approach by Macenko [50] is, that the stains are statistically separable. No expectations about the hue of these two stains are incorporated, as will emerge from the following description of the method, which is summarized as Pseudocode in Algorithm 4.1 on page 47. Each statement of the algorithm is described in more by detail, using the keyword **Step 1, 2, 3...**, respectively.

Step 1: The RGB values of the image in question are first converted to OD, where a linear combination of stains yields a linear combination of OD values [67]. To facilitate this step, the RGB-values are uniformly normalized to [0,1] and reshaped into an $m \times n$ array of 3-element-vectors (m being the rows, n the columns of the RGB image). The transformation is conducted for every tuple (i.e. the RGB vector of each pixel), using Equation 4.11, which is the same as the previously quoted Equation 4.1. Again, I is a vector containing the red, green and blue intensity value of each pixel, normalized to [0, 1].

$$OD = -\log_{10}(I) \quad (4.11)$$

After this conversion, all OD values are still normalized and range between [0, 1]. Samples of the normalized RGB and converted OD values are shown in Figure 4.9. It can be

Algorithm 4.1: Pseudocode for Calculation of Optimal Stain Vectors (as originally published in [50])

Input : RGB Slide

Output : Optimal Stain Vectors

- 1 Convert RGB to OD
 - 2 Remove data with OD intensity less than β
 - 3 Calculate SVD on the OD tuples
 - 4 Create plane from the SVD directions corresponding to the two largest singular values
 - 5 Project data onto the plane, and normalize to unit length
 - 6 Calculate angle of each point wrt the first SVD direction
 - 7 Find robust extremes (α^{th} and $(100 - \alpha)^{th}$ percentiles) of the angle
 - 8 Convert extreme values back to OD space
-

seen that the purple and pink pixels are hardly linearly separable (from the origin of the coordinate system) in the RGB space, while such a linear separation is more easily achieved in OD space.

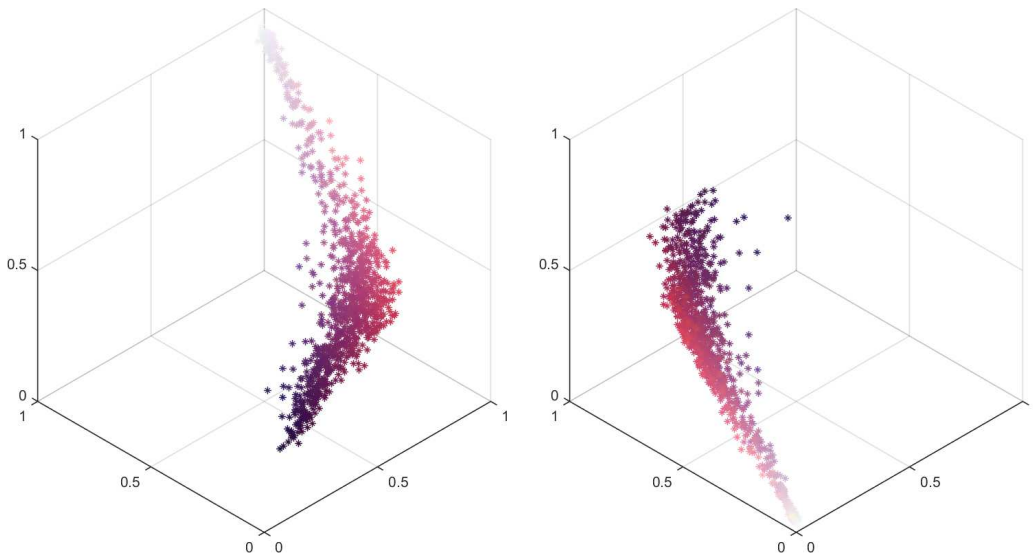


Figure 4.9: Colors in an H&E stained slide: **(Left)** Samples of normalized RGB pixels depicted in their original RGB color, **(Right)** Same samples in OD space depicted in their original RGB color, also normalized

Step 2: In order to eliminate the influence of background areas from the subsequent calculations, all data with OD values less than a threshold beta, β are eliminated, resulting in a smaller vector array called \hat{OD} . In this work $\beta = 0.15$.

Step 3: Singular Value Decomposition (SVD) is calculated based on $\hat{O}D$. This is equivalent to the calculation of the eigenvectors of the covariance-matrix of these tuples (Equation 4.12).

$$V = \text{eig}(\text{cov}(\hat{O}D)) \quad (4.12)$$

Step 4: A plane is created from the two largest eigenvectors, shown in Figure 4.10 left.

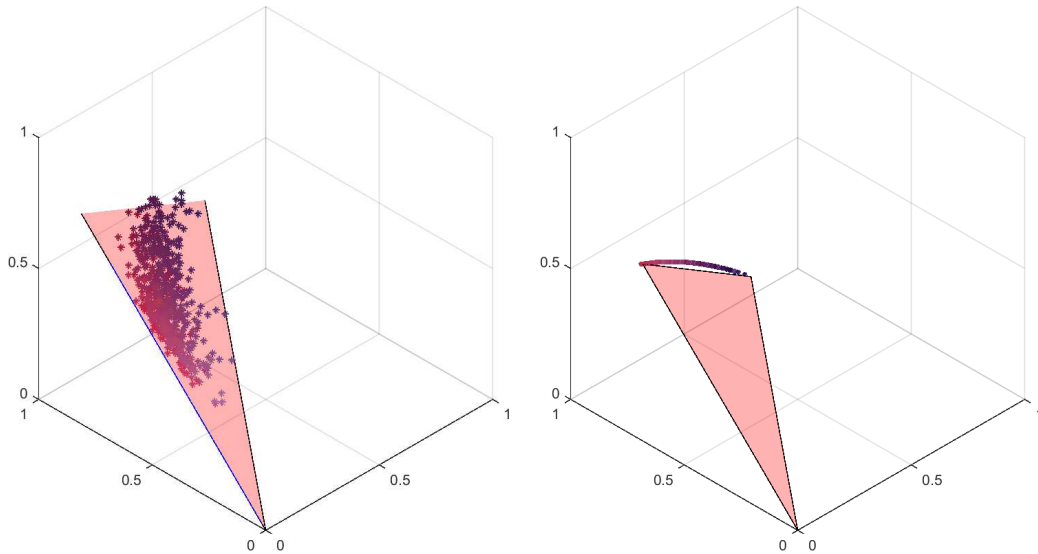


Figure 4.10: **(Left)** Representative depiction of the plane spanned by the two largest eigenvectors, **(Right)** Samples projected onto plane, at uniform length.

Step 5: All $\hat{O}D$ -tuples are projected onto this plane and brought to uniform length (1), which is shown in 4.10.

Step 6 and 7: In order to exclude outliers from the delicate step of identifying the inherent stain vectors, the angle of each tuple with respect to the first eigenvector direction is calculated. The alpha, α of the extremes in Algorithm 4.1 is set to $\alpha = 1$, hence the 1st and 99th percentile of these angles are defined to be robust extremes.

Step 8: The tuples representing the previously found two extremes are the OD values representing the inherent stain vectors. They do not need to be converted back to RGB value as the subsequent deconvolution requires the stain matrix to be in OD-space.

In the course of this work, the recreation of the method as described in [50] and its implementation led to in-depth understanding about all steps. However, a readily implemented version of this method with a copyright notice granting free permission to use was discovered after the custom implementation [84]. Since the readily implemented

version was faster than the custom implementation by an order of magnitude it was decided to use the former for this pipeline.

Fixed Values Method

Alike in other publications (see Section 3.4.2), the computationally cheapest deconvolution method involves the usage of fixed deconvolution vectors, as they do not have to be computed per given image or dataset. In this work stain vectors are utilized which are commonly deployed in existing DIA solutions (e.g. [63]), namely vectors defined by Gabriel Landini and Ruifrok [67] in the Color Deconvolution Plugin of the software "FIJI". These values are taken from a software by Gabriel Landini⁴ and stated in Table 4.1, where DAB stands for Diaminobenzidine, an anti-Ki-67 antibody.

Hematoxylin	0.650	0.704	0.286
DAB	0.268	0.570	0.776

Table 4.1: Fixed stain matrices from the deconvolution plugin of FIJI used in this work

4.2.2 Normalization and Contrast Adjustment

As the different deconvolution techniques yield channels with varying values and contrasts, a necessary prerequisite for further processing of each channel image is to bring it to a uniform contrast. To this end, each image is first normalized. In this context, the term *normalization* signifies the linear operation of transferring the histogram of an image from any arbitrary scale to the limited, normalized scale of $[0, 1]$. Subsequently, the contrast of each channel C is adjusted in such a way that 1% of the data is saturated at high and low intensities, respectively, of C which increases the contrast.

4.2.3 Thresholding of the Stain Channels

At this stage RGB image has already been deconvolved into the two grayscale channels, C_{pos} and C_{neg} , and the values of each channel are normalized and their contrasts adjusted. Following, each of the channels is converted into a binary image with two classes, foreground and background. This is realized via a k-means clustering step on the grayscale image histogram for all pixel values > 0 . The settings for the k-means clustering are the following:

- Number of clusters to be built:
2 (C_{pos}), 3 (C_{neg})
- Maximum number of iterations:
200

⁴Gabriel Landini: http://imagej.net/Colour_Deconvolution, accessed on 19.05.2016, 11:15

- Action when losing all members of a cluster:
create new cluster from point furthest away
- Number of times the clustering is repeated with new initial cluster centers:
5

The outputs of the k-means clustering are the following:

- The cluster centers of the repetition resulting in the lowest sum of distances from all points to their centers
- The assignment of each datapoint to a cluster center

The different numbers of clusters to be built for C_{pos} and C_{neg} are based on the physical background of the staining steps:

Hematoxylin (the ‘negative’ stain visible in C_{neg}) dyes all parts of the tissue – cytoplasm, inter-cellular space and nuclei. While the nuclei are stained more intensely, the cytoplasm and inter-cellular space appear less intense and the tissue-void space of the slide remains unstained. Thus, the clustering step for C_{neg} is initiated with 3 cluster centers.

Ki-67, on the other hand, unexceptionally stains proliferating nuclei and all other parts of the slide remain unstained. Therefore, the clustering step for C_{pos} is initiated with only 2 cluster centers.

Figure 4.11 demonstrates the meaningfulness of choosing different numbers of clusters. The positive channel C_{pos} clearly shows two different classes (high staining intensity of the nuclei and low intensity elsewhere) while the negative channel C_{neg} can be differentiated into three classes (high staining intensity of the nuclei, medium intensity of the surrounding tissue and low intensity). In both channels, only those pixels assigned to the brightest class are retained as foreground-pixels in the binary image BW_{pos} and BW_{neg} , respectively.

4.2.4 Segmentation and Quantification

At this stage, both channels have been converted to binary images BW_{pos} and BW_{neg} . However, as can be seen in Figure 4.11 (c) and (f), these are still very coarse, contain a lot of noise, holes and unwanted structures.

The following sections describe how these coarse binary images are gradually cleaned from noise, holes and other unwanted structures. At the end of these sequences, the result for each channel is the segmentation mask as a binary image, where each CC represents a nucleus.

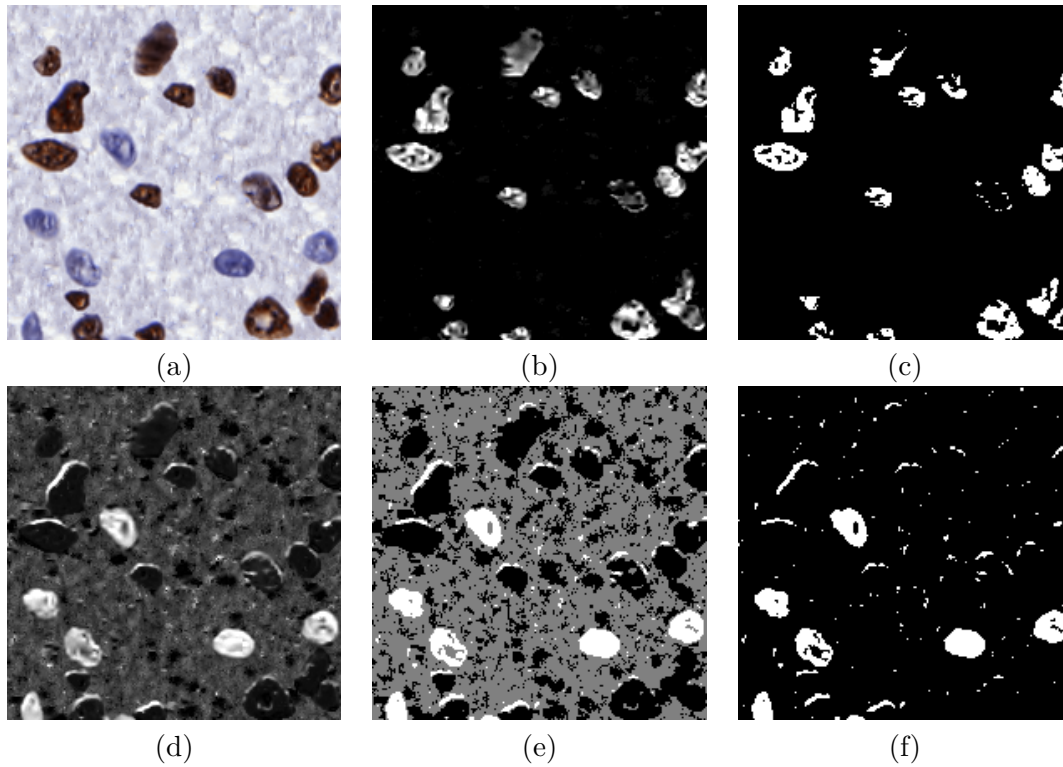


Figure 4.11: Example of clustering the deconvolved channels with k-means **(a)** Original image **(b)** Channel C_{pos} **(c)** Pixels in C_{pos} classified as foreground (white) and background (black) after k-means clustering, yielding binary positive channel BW_{pos} **(d)** Channel C_{neg} **(e)** Pixels in C_{neg} classified as foreground (white), and background (gray and black) **(f)** Binary negative channel BW_{neg}

Positive Channel

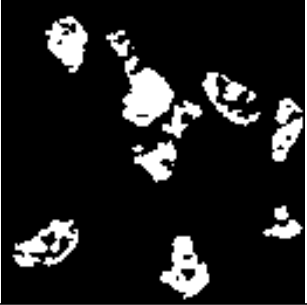


This section describes the operation `segment_pos()` of Figure 4.8, which is the segmentation of the positive channel. First, the settings for variables used during the segmentation are listed in Table 4.2.



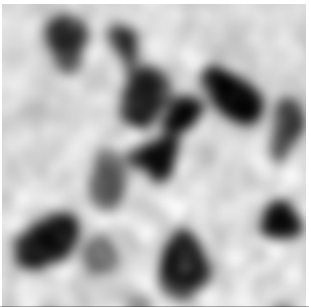

In the following, the single steps to convert the coarse binary image after k-means clustering into a segmentation mask are listed, accompanied by images to aid in understanding the effect of each measure. BW_X denotes a Black-White image (binary), GS_X denotes a Gray-Scale image, RGB_{grayX} denotes the grayscale-version of the original RGB image and WS denotes a Water-Shed transform.

4. SUGGESTED SOLUTION

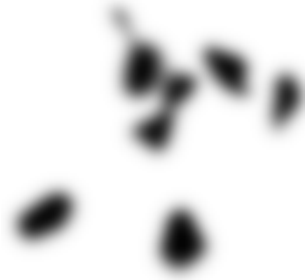



Variable Name	Size	Description
d	$5\mu m$	Globally defined minimum nucleus diameter
SE	$r = \frac{1}{3}d$	Structuring Element for morphological operations, disk-shaped with radius r
t_{sol}	0.96	Solidity threshold
t_{area_max}	$(d \cdot 5)^2 \cdot \pi$	Maximum area threshold
k_1	$SD = \frac{d}{3}$	Gaussian kernel 1 with standard deviation SD
k_2	$SD = \frac{d}{4}$	Gaussian kernel 1 with standard deviation SD


Table 4.2: Settings for the segmentation of the positive Channel C_{pos}

Step #	Input	Action	Output	Output Image
1	-	Binary input image obtained after thresholding C_{pos}	BW_1	
2	BW_1	Filter with k_1	GS_1	
3	GS_1	Cluster histogram into 2 classes (foreground, background) using a custom, fast k-means implementation based on the histogram of every 10^{th} pixel	BW_2	

Step #	Input	Action	Output	Output Image
4	BW_2	Eliminate CC with minor axis length smaller than d , as they are considered noise (none in this example)	BW_3	
5	BW_3	The area and solidity of all CC in BW_3 are calculated and all CC with a solidity below t_{sol} and area above t_{area} are labeled as "irregular" (brighter CC in image) in contrast to "regular" (darker CC in image)	GS_2	
6	RGB_{gray0}	Filter grayscale version of native RGB image, RGB_{gray0} , with Gaussian kernel k_2	RGB_{gray1}	
7	RGB_{gray1}, GS_2	Investigate the intensities in RGB_{gray1} within the bounding box of every irregular CC and use k-means to divide them into three classes. This creates a locally adaptive classification. The brightest class is merged with the background, only the two darker classes are retained	GS_3	

4. SUGGESTED SOLUTION

Step #	Input	Action	Output	Output Image
8	GS_3	Filter GS_3 with Gaussian kernel k_1	GS_4	
9	GS_4	Separate possibly connected nuclei in GS_4 using watershed	WS, BW_4	 
10	BW_4	Eliminate nuclei candidates which are above t_{Area_max} as they are probably not nuclei but stromal areas or the like (no candidates in this example)	BW_5	

Step #	Input	Action	Output	Output Image
11	BW_5, GS_2	Result of Steps 9-12 is binary mask containing both irregular CC (now separated, if necessary) and regular CC	BW_{final}	

Negative Channel

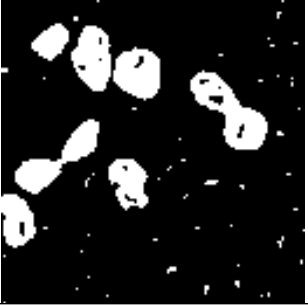
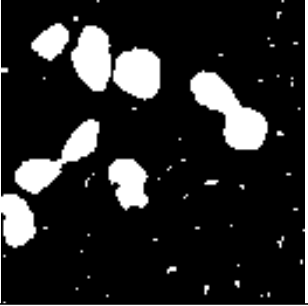


This section describes the operation `segment_neg()` of Figure 4.8, which is the segmentation of the negative channel. First, the settings for variables used during the segmentation are listed in Table 4.4.



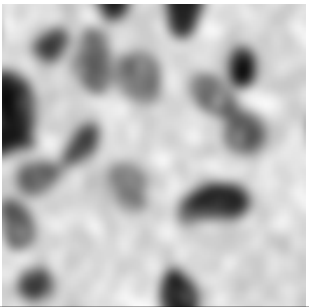

Variable Name	Size	Description
d	$5\mu m$	Globally defined minimum nucleus diameter
SE	$r = \frac{1}{3}d$	Structuring Element for morphological operations, disk-shaped with radius r
t_{sol}	0.96	Solidity threshold
t_{area_min}	$(d \cdot 5)^2 \cdot \pi$	Maximum area threshold
t_{area_max}	$d^2 \cdot \pi$	Maximum area threshold
k	$SD = \frac{d}{4}$	Gaussian kernel with standard deviation SD

Table 4.4: Settings for the segmentation of the negative Channel C_{pos}


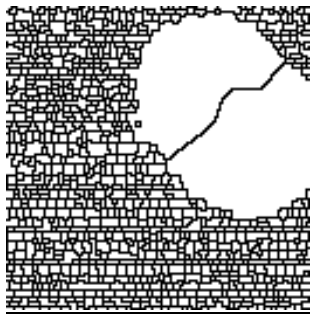


In the following, the single steps to convert the coarse binary image after k-means clustering into a segmentation mask are listed, accompanied by images to aid in understanding the effect of each measure.



4. SUGGESTED SOLUTION

Step #	Input	Action	Output	Output Image
1	-	Binary input image obtained after thresholding C_{neg}	BW_1	
2	BW_1	Holes of binary input image BW_1 are filled	BW_2	
3	BW_2	Marker <i>marker</i> is produced via erosion of binary input image BW_2	<i>marker</i>	
4	BW_2 , <i>marker</i>	Noise is eliminated via reconstruction of BW_2 with <i>marker</i>	BW_3	

Step #	Input	Action	Output	Output Image
5	BW_3	Morphological Opening of each CC with SE	BW_4	
6	BW_4	The area and solidity of all CC in BW_4 are calculated and all CC with a solidity below t_{sol} and area above t_{area_max} are labelled as "irregular" (brighter CC in image) in contrast to "regular" (darker CC in image)	GS_1	
7	RGB_{gray0}	Filter grayscale version of native RGB image RGB_{gray0} with Gaussian kernel k	RGB_{gray1}	
8	RGB_{gray1}, GS_1	Investigate the intensities in RGB_{gray1} within the bounding box of every irregular CC and use k-means to divide them into three classes. This creates a locally adaptive classification. The brightest class is merged with the background, only the two darker classes are retained	GS_2	

4. SUGGESTED SOLUTION

Step #	Input	Action	Output	Output Image
9	GS_2	Filter GS_2 with Gaussian kernel k	GS_3	
10	GS_3	Separate possibly connected nuclei in GS_3 using watershed	WS, BW_4	 
11	BW_4	Eliminate nuclei candidates which are above t_{Area_max} as they are probably not nuclei but stromal areas or the like (no candidates in this example)	BW_5	

Step #	Input	Action	Output	Output Image
12	BW_5, GS_1	Result of Steps 9-12 is binary mask with now separated irregular blobs	BW_6	
13	BW_6	Removal of all CC in result with Area smaller than than $\frac{t_{Area_min}}{4}$ as they are considered noise (no candidates in this example)	BW_{final}	

For each identified nucleus, be it in C_{pos} or C_{neg} , the segmentation mask is stored at the corresponding location in the Result Mask, RM . RM is a binary mask with the first two dimensions m, n equivalent to the True Mask, TM , with m, n and the original image m, n, o , where o is the third dimension describing the three color channels red, green and blue (see Figure 4.12).

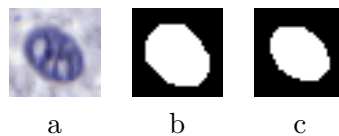


Figure 4.12: Detail from (a) Original RGB nucleus (b) Ground truth segmentation of this nucleus in the TM (c) segmentation of this nucleus in the RM

4.3 Evaluation Methods

In accordance with common practice as outlined in Section 3.6, the evaluation of the quality of both nuclei quantification and segmentation is based on a number of criteria. They are derived by comparing each RM , containing the derived segmentation, with the corresponding TM , containing the ground truth segmentation stemming from the image synthesis process. Thus it is an operation conducted on binary image data.

The subsequent spreections explain which criteria are used and how they are derived for the particular use case.

4.3.1 Evaluation of Nuclei Quantification

The criteria to be examined for the evaluation of the nuclei quantification are:

- Labeling Index Error $Error_{LI}$
- Precision P
- Recall R
- F1-score F_1

Each of these will be discussed in more detailed in the following paragraphs.

Labelling Index Error $Error_{LI}$

One of the main questions of this work is how accurately the presented algorithm is able to estimate the Labeling Index LI which represents the ratio of positive nuclei in respect to the total number of nuclei in the image, i.e. the proliferative activity of the tissue. In order to evaluate this aspect, a criterion called Labeling Index Error, $Error_{LI}$, is introduced. It reports the absolute Error between estimated LI (LI_{est}) and ground truth LI (LI_{true}) as formulated in Equation 4.13.

$$Error_{LI} = LI_{est} - LI_{true} \quad (4.13)$$

LI_{est} and LI_{true} range between $[0,1]$, where 0.0 signifies that 0% of the nuclei in the image are Ki-67 positive (all nuclei are negative) and 1.0 signifies that 100% of the nuclei in the image are Ki-67 positive (no nuclei are negative). $Error_{LI}$ ranges between $[-1,1]$, where $Error_{LI} = -1.0$ represents an extreme example and signifies that LI_{true} was underestimated by 100%, which is the case if $LI_{true} = 1.0$ and $LI_{est} = 0.0$. At the opposing end of the scale, $Error_{LI} = 1.0$ signifies that LI_{true} was overestimated by 100%, which is the case only if $LI_{true} = 0.0$ and $LI_{est} = 1.0$. A value of $Error_{LI} = 0.02$ predicates that LI_{true} was overestimated by 2%. The desired value is $Error_{LI} = 0.0$, which states that the estimated LI matches the true LI exactly, $LI_{est} == LI_{true}$.

LI_{est} is derived by counting the number of CC in the RM of each channel, RM_{pos} and RM_{neg} , conducting the following calculation (Equation 4.14):

$$LI_{est} = \frac{\#CC \text{ in } RM_{pos}}{\#CC \text{ in } RM_{pos} + \#CC \text{ in } RM_{neg}} \quad (4.14)$$

Accordingly, LI_{true} is derived by counting the number of CC in the TM of each channel (Equation 4.15).

$$LI_{est} = \frac{\#CC \text{ in } TM_{pos}}{\#CC \text{ in } TM_{pos} + \#CC \text{ in } TM_{neg}} \quad (4.15)$$

Precision P , Recall R and F1-score F_1

Precision, Recall and F1-score are described by the Equations 4.16, 5.2 and 5.3, where TP stands for "True Positive", FP for "False Positive" and FN for "False Negative".

$$P = \frac{TP}{TP + FP} \quad (4.16)$$

$$R = \frac{TP}{TP + FN} \quad (4.17)$$

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (4.18)$$

In order to derive Precision, Recall and F1-score, the counts of TP, False Positive FP and False Negative FN in each image are required. Given the fact that the nuclei are counted via segmentation, the derivation of TP, FP and FN is based on the comparison between the True Mask TM and the Result Mask RM . The basic algorithm for this step is outlined in Algorithm 4.2.

Algorithm 4.2: Pseudocode for Derivation and Definition of TP, FP and FN

```

1 for each  $CC$  in  $TM$  do
2   identify  $CC$  in  $RM$  having overlap with this  $CC$ ;
3   if number of  $CC$  in  $RM$  identified == 0 then
4     copy current  $CC$  in  $TM$  to  $FN$ -mask;
5   else
6     identify  $CC$  in  $RM$  with largest  $DC$ ;
7     store largest  $DC$  in  $DC$ -list;
8     store  $CC$  in  $RM$  with largest  $DC$  in  $TP$ -mask;
9     eliminate  $CC$  in  $RM$  with largest  $DC$  from  $RM$  to prevent double-counting;
10  end
11 end
12  $TP$ -count = number of  $CC$  in  $TP$ -mask;
13  $FP$ -count = number of remaining  $CC$  in  $RM$ ;
14  $FN$ -count = number of  $CC$  in  $FN$ -mask;

```

The definition of "overlapping" in this case, as examined in Line 2 of Algorithm 4.2, is equivalent to the definition of a true positive in [89]: if the Dice Coefficient (DC) between the blob in the TM and the blob in the RM is below a certain threshold, the blob in the RM is not regarded as overlapping and will not be counted as a TP. The threshold is set to 0.2, as in the works of [89]. The explanation of the Dice Coefficient DC and details of its implementation are laid out in the adjacent Section 4.3.2.

Once TP, FP and FN are defined, identified and counted for each image, the other criteria Precision P , Recall R and F1-score F_1 can be determined for each image. For each, the resulting value per image is passed on to evaluate the performance of the algorithm on an entire dataset. These criteria are chosen according to their frequent reporting in similar publications, thus they allow comparing the algorithm performance.

4.3.2 Evaluation of Nuclei Segmentation

The criteria to be examined for the evaluation of the nuclei segmentation are::

- Dice Similarity Coefficient DC
- Area Estimation $Area_{Est}$
- Pearson's Coefficient of the Area ρ_{Area}
- Pearson's Coefficient of the Solidity $\rho_{Solidity}$
- Pearson's Coefficient of the Eccentricity $\rho_{Eccentricity}$

Each of these criteria is calculated for both positive and negative nuclei and is based on the respective TP nuclei. The values reported per image are the median of the entire population of nuclei in each image and the values reported per experiment (Quartiles, Median etc.) are based on all images in the respective dataset. The median is chosen to be the preferable measure of central tendency over the mean, because there are no clear grounds to assume a normal distribution among the reported data.

Each of these will be discussed in more detailed in the following paragraphs.

Dice Coefficient DC

In order to define to which extent a nucleus was correctly segmented, the commonly used Dice Coefficient DC is examined as a measure of weighted overlap between segmentation result and ground truth (Equation 4.19):

$$DC = \frac{|A \cap B|}{\frac{1}{2}(|A| + |B|)} \quad (4.19)$$

Literature research (see Section 3.1) revealed that the motivation for developing nuclei segmentation DIA solutions is among other factors founded in the pathologists' need

for the quantification of morphological nuclei features, such as area or atypicalness (e.g. how elliptical it is, how many indentations it has etc.) to assess the nuclear pleomorphism. However, as a criterion for evaluating the segmentation, the commonly used Dice Coefficient alone does not indicate whether or not the true object area has been over- or underestimated, since it is merely a measure of overlap. In this light, the *DC* by itself is only of limited expressiveness concerning the quality of the segmentation, since the two model cases exemplified in Figure 4.13 reveal the same dice coefficient. As a consequence, the pathologist would make the same assumptions about the area/size of the nuclei for both cases, even though case (a) clearly underestimates and case (b) clearly overestimates the area.

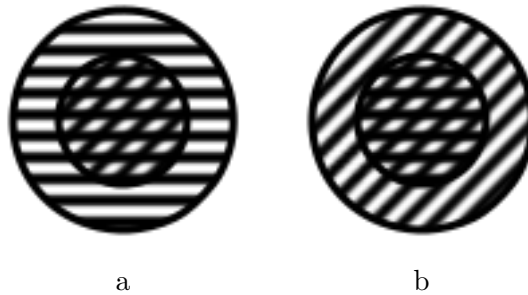


Figure 4.13: Example of two segmentation cases resulting in the same *DC* (Horizontal stripes = *TM*, Stripes at 45° angle = *RM*) (a) *RM* is entirely surrounded by *TM* (b) *TM* is entirely surrounded by *RM*

Thus, this work also specifically examines the over- or underestimation of the area, as laid out in the next paragraphs.

Area Estimation $Area_{Est}$

Via comparing the segmented area RM_{Area} (CC in the *RM*) to the ground truth area TM_{Area} (CC in the *TM*) of each true positive nucleus and returning the ratio for each nucleus, the called Area Estimation Coefficient, $Area_{Est}$, is calculated. It is given in Equation 4.20:

$$Area_{Est} = \frac{RM_{Area}}{TM_{Area}} \quad (4.20)$$

The criteria reported is the median of all nuclei area estimations per image. It casts light onto whether the nuclei areas are generally over- or underestimated.

Pearson's Coefficient of the Area ρ_{Area} , Solidity $\rho_{Solidity}$ and Eccentricity $\rho_{Eccentricity}$

As previously laid out in Section 2.3, morphological nuclei features such as shape and area are also of interest in the context of diagnosing breast cancer in histopathological images.

Thus it is desirable to examine the correlation between the outputs of the segmentation algorithm and the ground truth features of the nuclei.

To this end, the Pearson's Correlation Coefficient ρ is utilized as a measure of correspondence between two variables (Equation 4.21):

$$\rho(X, Y) = \frac{cov(X, Y)}{(var(X)var(y))^{\frac{1}{2}}} \quad (4.21)$$

It is applied on a small representative, but not exhaustive set of nuclear features covering three high-level morphological feature families: area as a representative for size features, solidity as a representative for convex hull features and eccentricity as representative for elliptical features [7]. The features chosen are also used for verification in other publications [39, 68, 89] and demonstrate the possible applicability of the algorithm in the mentioned context. These features, identified via segmentation, are correlated to their ground truth equivalents via the mentioned Pearson's Correlation Coefficient ρ , as laid out in the following Equations, which are conducted for each CC which is a TP.

Equation 4.22 for Area Correlation

$$\rho_{Area} = \rho(Area_{RM}, Area_{TM}) \quad (4.22)$$

Equation 4.23 for Solidity Correlation (Sol = Solidity):

$$\rho_{Sol} = \rho(Sol_{RM}, Sol_{TM}) \quad (4.23)$$

Equation 4.24 for Eccentricity Correlation (Ecc = Eccentricity):

$$\rho_{Ecc} = \rho(Ecc_{RM}, Ecc_{TM}) \quad (4.24)$$

The value of ρ indicates whether there is a meaningful and reproducible correlation between the identified features and ground truth values. If ρ is high, it means that a feature is systematically under- or overestimated and the output value can be corrected by a linear correctional term, thus it is reliable enough as an output to be reported to the pathologist. If ρ is close to 0, however, the identified features are revealed to be close to random guessing and do not add to the diagnostic value of the segmentation results.

Area Estimation and ρ_{Area} are features which are complementary to each other and both necessary in order to fully report the correctness of the segmentation regarding the size of the nuclei. While Area Estimation reveals information about the extent of over- or underestimation across the entire image for N_{pos} and N_{neg} , it could counterbalance and consequently hide systematically or randomly occurring cases of both over- and under-detection, resulting in a median acceptable area estimation, but stemming from severe

segmentation errors. The criteria ρ_{Area} adds essential information to this impairment as it investigates whether there is a consistent correlation between the value reported and the ground truth. The closer ρ_{Area} to +1, the more reliable the value reported by $Area_{Est}$. Vice versa, the $|\rho_{Area}|$ alone does not include any hint on the extent of over- or underestimation of the area.

Results and Discussion

This chapter deals with the resulting output of the dataset synthesis and the performance of the nuclei segmentation and quantification algorithm.

5.1 Synthetic Dataset

In this section, the result of the SDS generation is presented and discussed, aided by sample pictures. First, the effect of the colorspace variation on the SDS is highlighted, then the most important characteristics of the datasets generated for further experiments are illustrated.

5.1.1 Colorspace Variation

Illustrated by Figures 5.1, 5.2, 5.3 and 5.4 the effect of the color deconvolution and normalization, as described in Section 4.2.1 is displayed.

Figure 5.1 shows Ki-67 snapshots taken from the real, clinical Ki-67 breast WSI from the Pathology Laboratory at the Linköping University Hospital, Sweden. In this illustration, the necessity of any nuclei detection algorithm to adapt to locally prevailing staining characteristics becomes obvious, as the shades of Ki-67 (brown) and Hematoxylin (purple) vary from image to image, even though they stem from the same laboratory. The leftmost and middle-left image of the second row in Figure 5.1 even stem from the same original .svs-file.

This hand-selected collection of snapshots was taken as the basis for extracting the set of 12 stain matrices M_{target} to be used for colorspace variation of the synthetic images SI_{init} via restaining. Figure 5.2 depicts the fixed stain matrix M_{source} used for the deconvolution of each SI_{init} (empty rings), as well as the 12 varying stain matrices M_{target} used for restaining them to varying appearances.

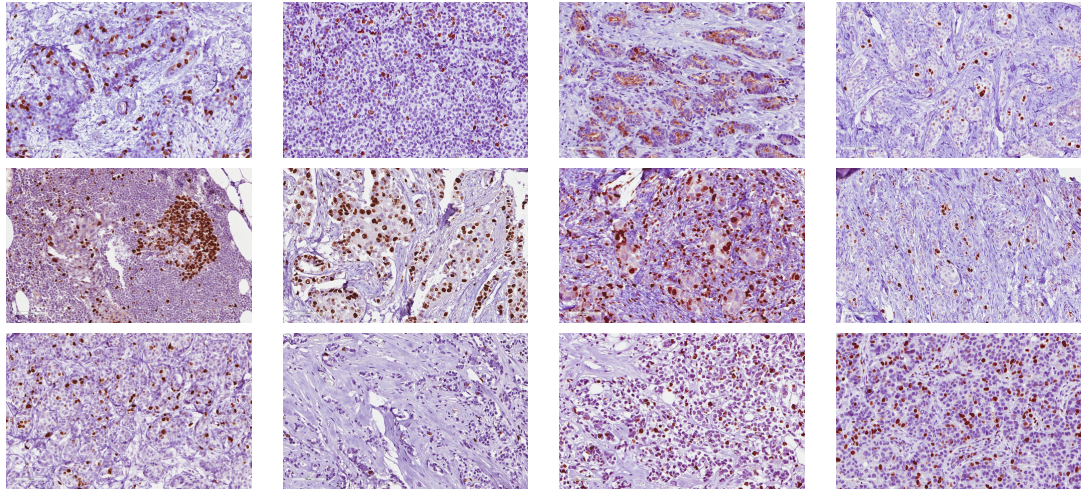


Figure 5.1: Set of Ki-67 snapshots taken from real, clinical images

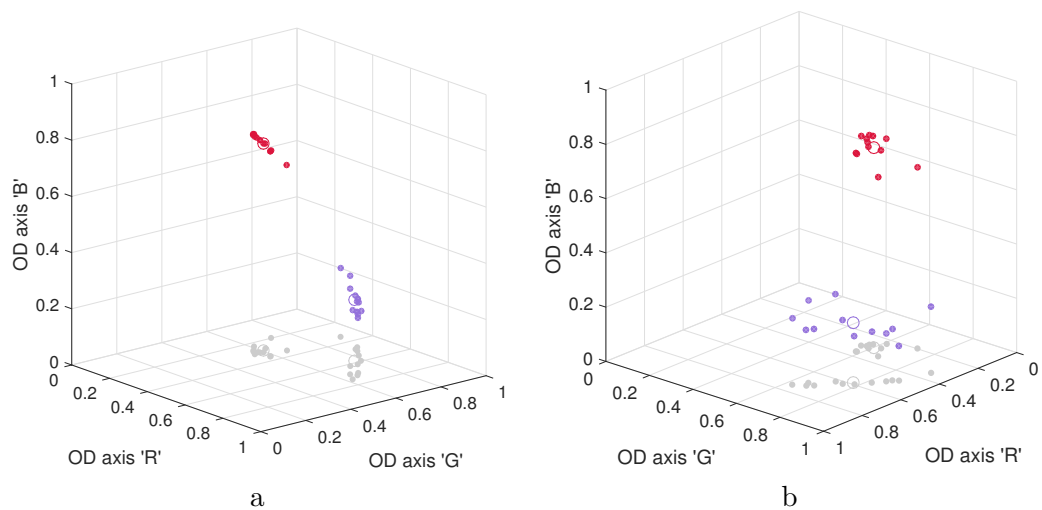


Figure 5.2: Two viewing angles **(a)** and **(b)** on the OD-values of the 12 stain matrices M_{target} represented by filled dots: The purple dots (lower cluster) represent the Hematoxylin values, the red dots (upper cluster) represent the Eosin values. The larger, empty ring within each cluster represents the values of M_{source} . The light-gray dots are the projections of the all tuples cast along the OD axis B, to emphasize their three-dimensional position.

The axes are labelled according to RGB scale corresponding to the respective OD scale (see Equation 4.1).

Following, a set of 12 synthetic images SI_{init} is shown in Figure 5.3. It gives an idea of the stain appearance right after the initial image synthesis. It can be seen that all sample images have the same stain appearance, i.e. all backgrounds, positive and negative nuclei are equally colored within a certain range and the most obvious difference in appearance stems only from the varying densities and ratios and positive and negative nuclei.

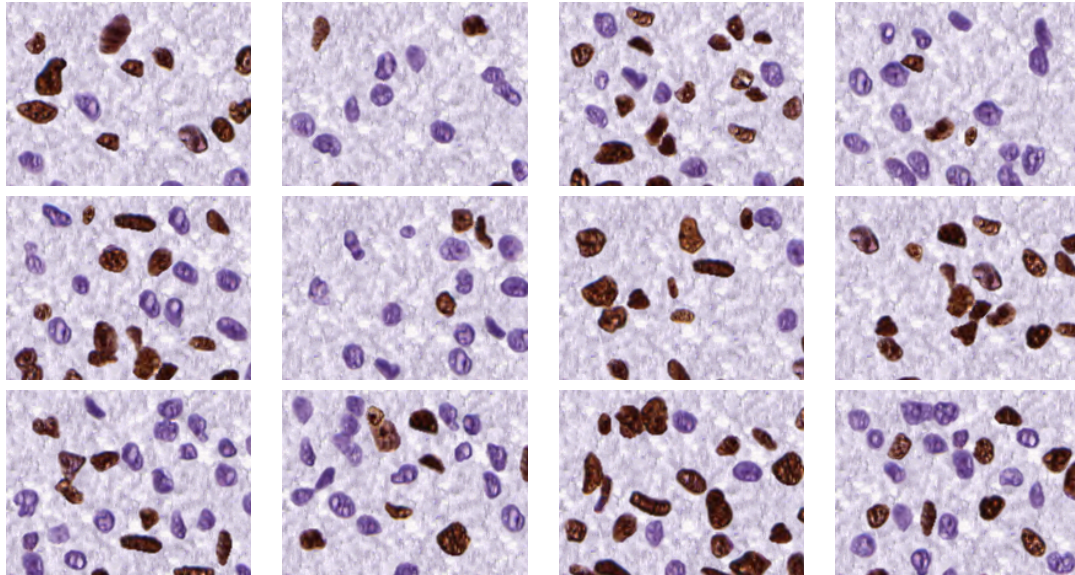


Figure 5.3: Clips from 12 representative synthetic images SI_{init} before restaining

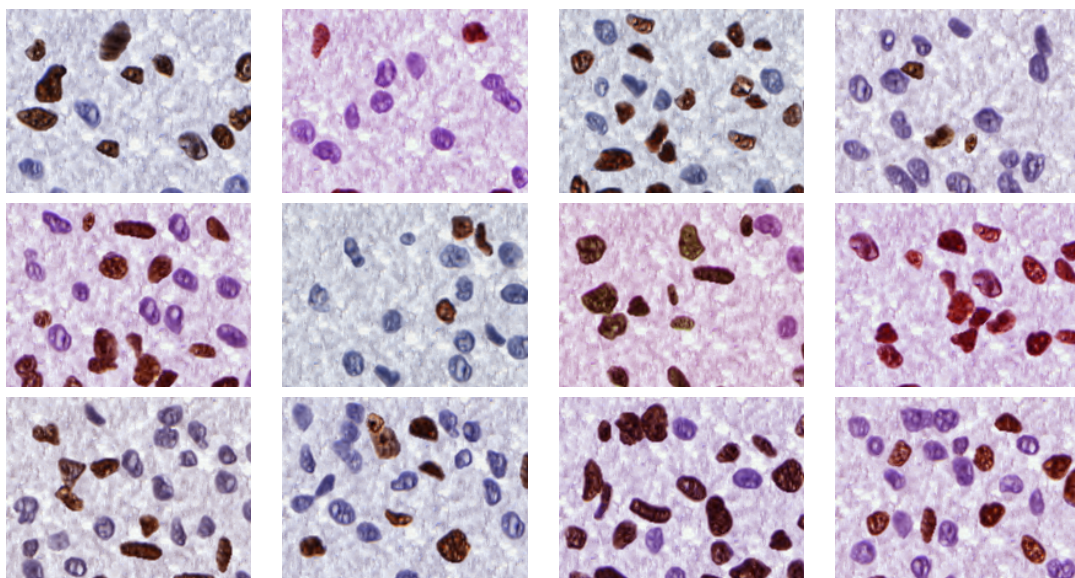




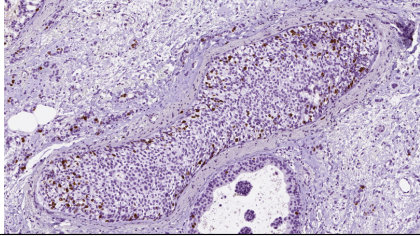
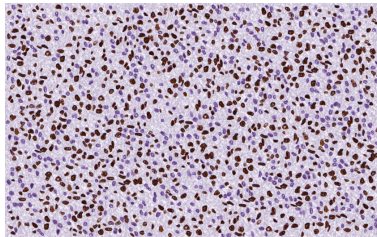
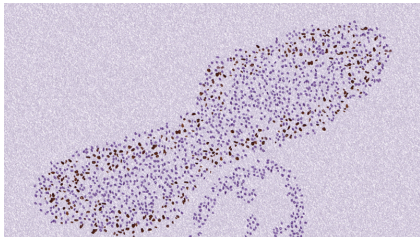
Figure 5.4: Clips from 12 representative synthetic images $SI_{restained}$ after restaining

Figure 5.4 demonstrates the impact of even small staining variations contained in a supposedly coherent dataset stemming from a single laboratory. All images now exhibit a different stain appearance, regarding background, positive and negative nuclei. The overall density or ratio of positive and negative nuclei does not in any way affect the resulting appearance of $SI_{restained}$ since the deconvolution and restaining process is conducted using predefined matrices, M_{source} and M_{target} . This fact needs to be stressed, as – in contrast – the density and ratio immanent in these images strongly affect the performance of the nuclei quantification algorithm presented in Section 4.2. Summarizing, adapting the normalization method published in Macenko et al. [50] to be used in a reversed manner leads to the expected result of successfully introducing realistic stain variations into a given set of synthetic images.

5.1.2 Dataset Characteristics

In this work, 2 different SDSs, namely a "uniform" and a "designed" Dataset, have been created. They are called $D_{uniform}$ and $D_{designed}$ respectively. The following Table 5.1 presents the characteristics of these two Datasets:

	"Uniform distribution" Dataset $D_{uniform}$	"Designed distribution" Dataset $D_{designed}$
Distribution Function for Nuclei	The probability of a nucleus to be placed is uniform across the entire image, thus the definition nuclear placement probability maps, $Map_{NucProbability}$, are obsolete	The probability of a nucleus to be placed, thus the distribution function, can be manually designed via nuclear placement maps, $Map_{NucProbability}$
Example of $Map_{NucProbability}$ (white = 100% probability for nuclei placement, black = 0% probability for nuclei placement)	none (uniform probability would yield only-white $Map_{NucProbability}$)	- for placement of S_{pos}  - for placement of S_{neg} 

	”Uniform distribution” Dataset $D_{uniform}$	”Designed distribution” Dataset $D_{designed}$
Example of real Ki-67 model image	none (uniform cell distribution is no natural behaviour)	
Example of resulting synthetic image SI_{init}		
Image Size(s)	Uniform (1680x1050)	Varying (1200x1200, 1712x980, 3214x1803)
# of Synthetic Images Per Dataset	50	50
# of different Stain-Appearances	12	12
Versions available (see Figures 5.3 and 5.4)	- SI_{init} (Original) - $SI_{restained}$ (Restained)	- SI_{init} (Original) - $SI_{restained}$ (Restained)

	"Uniform distribution" Dataset $D_{uniform}$	"Designed distribution" Dataset $D_{designed}$
Density of Nuclei ($N_{total}/\mu m^2$)	<p>Box plot showing the distribution of the density of nuclei ($N_{total}/\mu m^2$) for the Uniform and Designed datasets. The y-axis is scaled by $\times 10^{-3}$. The Uniform dataset shows a higher median density (around 2.2) compared to the Designed dataset (around 0.7). Both distributions have whiskers extending from 0 to 4.</p>	
Distribution of Labelling Index in Dataset (N_{pos}/N_{total})	<p>Box plot showing the distribution of the labelling index (N_{pos}/N_{total}) for the Uniform and Designed datasets. The y-axis ranges from 0 to 1. The Uniform dataset shows a lower median labelling index (around 0.42) compared to the Designed dataset (around 0.58). Both distributions have whiskers extending from 0 to 1.</p>	

Table 5.1: The characteristics of the two generated dataset Datasets $D_{uniform}$ and $D_{designed}$

The fact that $D_{designed}$ possesses a smaller overall nuclear density, as visible in Table 5.1 is due to the fact that the image synthesis algorithm calculates the number of maximally possible nuclei based on the $Map_{NucProbability}$ specified. The maps utilized in $D_{designed}$ allow nuclei placement only in restricted regions, thus the total number of nuclei, N_{total} , is smaller than in the images found in $D_{uniform}$, where the maps allow uniform placement across the entire image. This factor can be seized to examine whether the algorithms performance varies when confronted with differing nuclear densities.

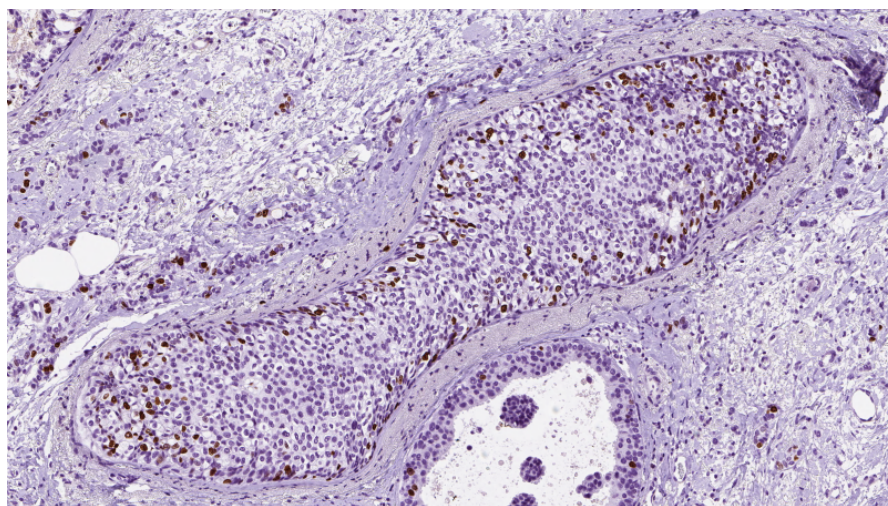
As laid out in Section 4.1.2, the nuclei in the SI_{init} have been extracted from one exemplary sample region of a real Ki-67 svf-file. Even though not explicitly notable at a first glance, the nuclei display a directional bias. This means that their mean major axis is biased along a certain direction. As illustrated in Figure 5.5 (b) and (c), the positive nuclei show a notable bias towards being rotated about -70° against the x-axis, the negative nuclei show a bias at about -30° . This originates from the physiological tendency of nuclei to align with the surrounding tissue scaffolding they are embedded into, which can be observed in Figure 5.5 (a).

While this observation may call for the implementation of a directional bias correction, its correction would not induce any difference for the functionality or performance of the presented nuclei quantification and segmentation method because the latter disregards any rotational nuclei features.

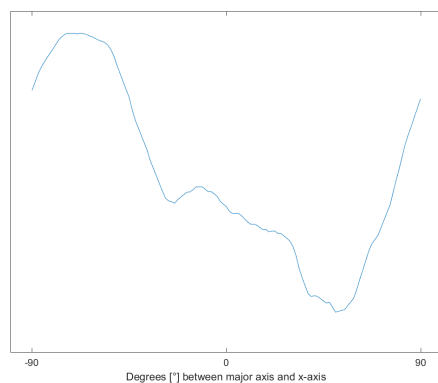
5.1.3 Summary and Discussion of Synthetic Datasets Results

Using the procedure for generating synthetic images described in Section 4.1, two datasets have been created. They mostly differ in the distribution of the nuclei on each image: in the dataset $D_{uniform}$ the nuclei are uniformly distributed in the dataset $D_{uniform}$ and in the other dataset $D_{designed}$ the distribution is variable across the image, confining both the allowed areas and the relative densities of the nuclei. Each dataset contains 50 images available both in a standard stain appearance version (SI_{init}) and in one of 12 different varying stain appearance versions ($SI_{restained}$). In both datasets, some images are extremely sparse and some are very densely populated (measured in $N_{total} / \mu m^2$). In average, $D_{uniform}$ features a higher nuclei density than $D_{designed}$ due to the constraint on the nuclei placement in the generation of the latter. Both datasets exhibit the full possible range of Labeling Indices LI, namely between 0 and 100%.

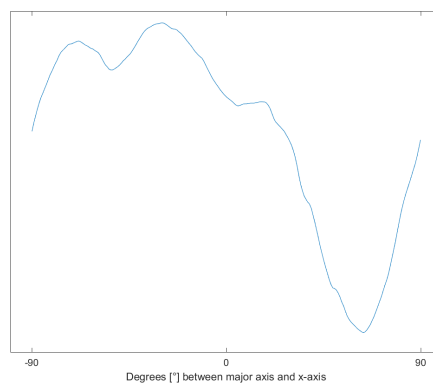
The optical appearance of the synthetic images, showing both isolated and touching or overlapping nuclei on a background structure, provides a realistic testing opportunity for any nuclei detection and/or segmentation algorithm which is either specialized on Ki-67 LI estimation or on general nuclei quantification. The ground truth mask provided alongside every single image, precisely outlining every nucleus, permits the exact evaluation of any DIA solution in both quantitative (how good is the nuclei detection) and qualitative (how good is the segmentation) aspects. As such, the research question whether a SDS generation is suitable for the evaluation of a nuclei quantification and segmentation algorithm can be answered in the affirmative.



a



b



c

Figure 5.5: (a) Nuclei tend to align along their surrounding tissue structure
 Frequency of angles between x-Axis and major axis of ellipse with same second moment
 as positive (a) and negative (b) nuclei used in the generation of the dataset (y-label
 omitted to show relationality, not absolute values)

5.2 Nuclei Quantification and Segmentation Metrics

Twelve experimental setups have been conducted to test the impact and performance of variable settings. In essence, three different deconvolution methods (Cosatto, Macenko and fixed, see Section 4.2.1) were applied on the two versions of each image (original synthetic image SI_{init} and restrained synthetic image $SI_{restrained}$) within each of the two dataset Datasets $D_{uniform}$ and $D_{designed}$. Table 5.2 introduces the naming scheme for the experimental setups and their characteristics.

Table 5.2: Setups for the different experiments conducted on the presented nuclei quantification and segmentation algorithm, using the two SDS created. By testing each of the three deconvolution methods on two versions of the image to be analysed, a total of 6 setup combinations, called A-F, are examined.

Name of experimental setup	A	B	C	D	E	F
Deconvolution method	Cosatto	Cosatto	Macenko	Macenko	Fixed (Ruifrok)	Fixed (Ruifrok)
Image to be analysed	SI_{init}	$SI_{restrained}$	SI_{init}	$SI_{restrained}$	SI_{init}	$SI_{restrained}$
Dataset	$D_{uniform}$ + $D_{designed}$	$D_{uniform}$ + $D_{designed}$	$D_{uniform}$ + $D_{designed}$	$D_{uniform}$ + $D_{designed}$	$D_{uniform}$ + $D_{designed}$	$D_{uniform}$ + $D_{designed}$

The results of the most vital criteria of these experimental setups is illustrated on the succeeding pages using boxplots. The following general information is valid for all plots:

- Each pair of boxplots (e.g. the two left-most columns in a boxplot labelled "A") represents the performance of the algorithm using a specific setup A-F (see Table 5.2) on both the uniform and designed dataset (left and right column of the pair, respectively). The data samples stem from the verification process on each of the 50 images of each dataset.
- The upper boundary of each blue box represents the 25% quantile (q_1) of the data, the red line represents the median and the lower boundary represents the 75% quantile (q_3) of the data. The distance between the quartiles q_1 and q_3 is referred to as Inter-Quartile Range (IQR).
- The plotted lower and upper whiskers (dashed black vertical line above and below each box) include all data values which are not considered outliers.
- The whiskers emerging at the top and bottom of the box denote the extremes, defined as values higher than $q_3 + w \cdot (q_3 - q_1)$ or lower than $q_1 - w \cdot (q_3 - q_1)$, where w is a weighting factor (here: $w = 1.5$). The distance between the whiskers is referred to as Range of Extremes (RoE).

- Any value outside of the RoE is defined as an outlier, represented by red, filled dots.
- The green line in each plot represents the "desired" value for the examined criteria, to facilitate visual judgement of the performance of the different experimental setups.
- The dashed horizontal line on the bottom/top of the plots represent the lowest/largest value of the 20/80 percentile of each distribution, respectively. Any data points lying outside these boundaries are displayed on the dashed horizontal line.

The arrangement of experiments in the boxplots and the division of the results into boxplots for each criteria enables insights into two main factors to be examined in the course of this work:

1. The influence of different deconvolution methods on the algorithm performance (the 4 leftmost columns represent the Cosatto method, the 4 middle columns represent the Macenko method and the 4 rightmost columns represent the fixed deconvolution).
2. The influence of the nuclei density on the algorithm performance (each of the 6 column-pairs has the high-density dataset $D_{uniform}$ on the left and the low-density dataset $D_{designed}$ on the right).

5.2.1 Nuclei Quantification

In the following, the evaluation all the criteria for the nuclei quantification stated in Section 4.3.1 is presented and discussed, namely the LI Error, Precision, Recall and F1-Score.

Labeling Index Error for each Setup and Dataset

In Figure 5.6, the LI error, $Error_{LI}$, for each setup A-F and Dataset $D_{uniform}$ and $D_{designed}$ are shown. More specifically, it depicts how far the LI estimated by the algorithm, LI_{est} , deviates from the true LI, LI_{true} , known from the ground truth of the SDS. As an example: The LI_{true} is 0.90 (i.e. 90% of the nuclei are positive) and the LI_{est} is 0.89 (i.e. 89% of the nuclei are positive), then the $Error_{LI}$ for this model case amounts to: $Error_{LI} = LI_{est} - LI_{true} = 0.89 - 0.90 = -0.01$ Thus in this example LI_{true} was underestimated by -0.01 (i.e. it was underestimated by 1%).

The LI Error is the major criterion for evaluating the performance of the nuclei detection algorithm and one of the principal answers to the research questions of this work. Interpreting the median as a measure of correctness and the RoE as a measure of variability of the estimated LI, the results reveal that setup A on the uniform dataset performs best, with a median of -0.015 and an RoE of only 0.03. This states that the LI

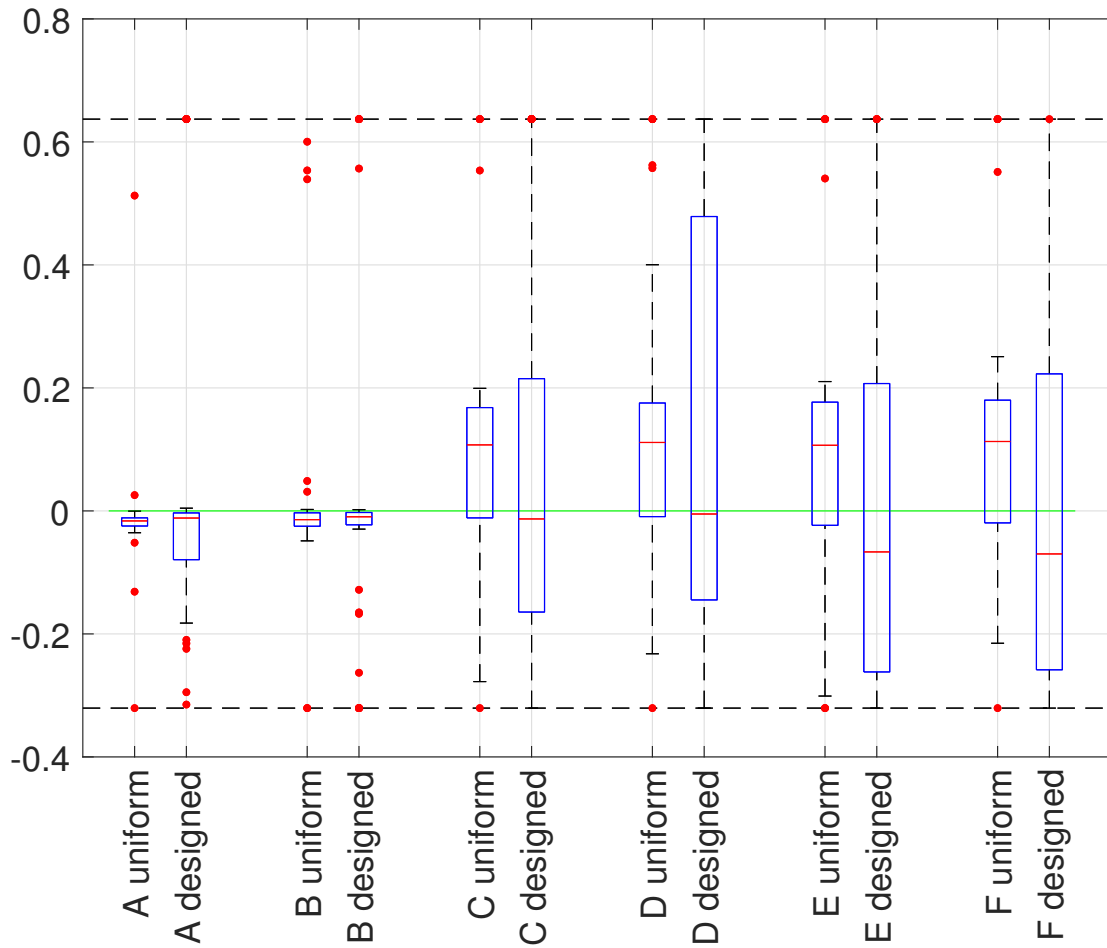


Figure 5.6: Labelling Index (LI) Error for each setup A-F and datasets $D_{uniform}$ and $D_{designed}$

(Note that the plots for positive and negative nuclei have different vertical scalings to allow for the best fitted display of the relevant data.)

is estimated with a median deviation of -1.5% and the RoE indicates that the variability of the estimated LI is very low, thus the LI estimation is a reliable output of the program. Setup D on the designed dataset performs worst, with a RoE of 0.98 and an IQR of 0.62 despite a median of -0.01. Setups E and F are slightly worse than C and D, because the medians for $D_{designed}$ are always lower and the IQR is almost as large as in the worst setup, D on $D_{designed}$.

Regarding the LI error results in the light of the impact of the deconvolution method, it is visible that the algorithm benefits from using the Cosatto deconvolution (setup A and B) as opposed to the Macenko (setup C and D) or fixed (setup E and F) value deconvolution. This implicates that the impact of the deconvolution method on the correctness of the LI estimation is significant.

The fact that Experiments A and B exhibit more outliers according to Figure 5.7 does not infer that they are more likely to fail. The whiskers (bordering the RoE) of the other experiments include all data points possibly lying in the same range, thus have similar values as the so-called outliers of experiments A and B.

A clear tendency of less reliable results can be observed on the $D_{designed}$ in comparison to $D_{uniform}$: for each experimental setup A-F, except B, $Error_{LI}$ is notably larger on $D_{designed}$ than on $D_{uniform}$. This can be explained with the major difference between the number of nuclei in the two datasets: $D_{designed}$ is more sparsely populated with nuclei than $D_{uniform}$, as viewable and discussed in Section 5.1.2. As a consequence, the adaptive deconvolution methods of setups A-D as well as the thresholding steps in the segmentation process, which are both histogram-based operations, are more influenced by the color of the background and less influenced by the color of the nuclei. Concerning setups E and F which use the fixed deconvolution method, the inferior performance on $D_{designed}$ can also be explained with the thresholding step after deconvolution. It fails during the histogram-based clustering when there are too few samples of foreground (nuclei) in comparison to the background.

Precision for each Setup and Dataset

The Precision P is defined as

$$P = \frac{TP}{TP + FP} \quad (5.1)$$

and describes the portion of TP in relation to the amount of detected nuclei (sum of TP and FP). The closer P to 1, the better, because $P = 1$ states that all detected nuclei are TP. $P = 0.5$ states that only 50% of the detected nuclei are TP. The constant $P = 1$ is highlighted by a green line in Figure 5.7.

Based on the values of P for the positive nuclei detection depicted in Figure 5.7 (a), it can be concluded that setups A and B (using the Cosatto deconvolution method) on both Datasets as well as C (using the Macenko deconvolution method) on $D_{uniform}$ perform best. The superiority of the aforementioned setups is clearly visible in the analysis of positive nuclei, with the medians, IQR and RoE of setups A, B and C all located clearly

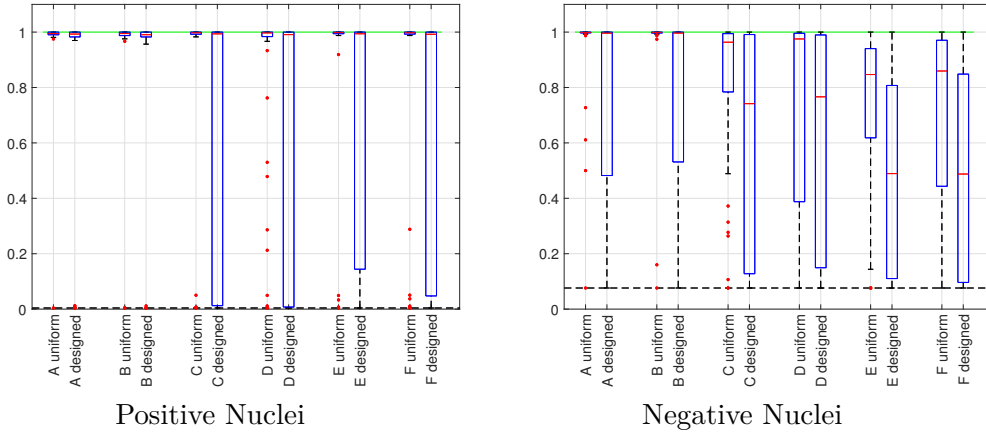


Figure 5.7: Precision of Nuclei Detection for each setup A-F and Datasets $D_{uniform}$ and $D_{designed}$ (Note that the plots for positive and negative nuclei have different vertical scalings to allow for the best fitted display of the relevant data)

above 0.9 and close to 1.0, while the other setups D-F result in massive misperformance. This is observable in a high number of outliers ranging down to $P = 0.0$ for setups D-F on $D_{uniform}$ and large IQRs and RoEs for setups C-F on $D_{designed}$. $P = 0.0$ indicates complete randomness of the assigned "positive" label, inferring that none of the detected nuclei represent the ground truth nuclei, but quite the contrary: all detected nuclei are FP. Taking into consideration the results of Figure 5.7 (b), which depict P for the negative nuclei detection, the previous conclusion regarding setup A and B has to be withdrawn. It is visible, that setups A and B still outperform all the other setups, however this is only true when executed on $D_{uniform}$. The RoE of the algorithm-setup A and B on negative nuclei ranges down to 0.0, making it an unreliable result despite the median P located at almost 1.0. The reason for this is once again to be found in the sparsity of the $D_{designed}$: The essential, histogram-based clustering step necessary for meaningful thresholding between background and foreground fails in the negative channel if the number of negative nuclei is so small that they do not substantially influence the histogram. Summarizing and in line with the LI error results, the choice of the deconvolution method has the largest impact on P , while the second decisive factor is the density of nuclei in the analyzed images, which is higher in $D_{uniform}$ than in $D_{designed}$.

Recall for each Setup and Dataset

The Recall R is defined as

$$R = \frac{TP}{TP + FN} \quad (5.2)$$

and describes the portion of correctly identified nuclei in relation to all true nuclei present in the image. The closer to 1, the better, because $R = 1$ states that all true nuclei have been detected and no true nuclei remained undetected. $R = 0.5$ states that only 50% of

the present true nuclei have been detected. The ideal $R = 1$ is highlighted by a green line in Figure 5.8.

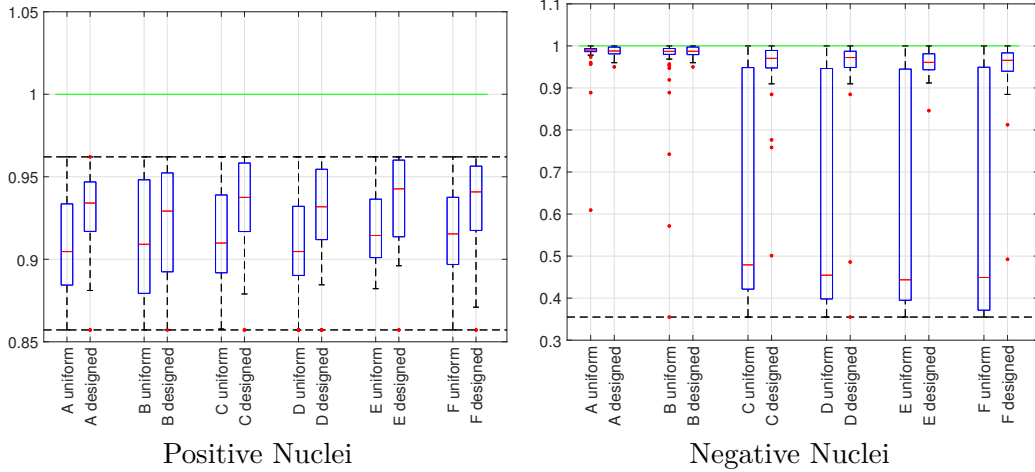


Figure 5.8: Recall of Nuclei Detection for each setup A-F and Datasets $D_{uniform}$ and $D_{designed}$ (Note that the plots for positive and negative nuclei have different vertical scalings to allow for the best fitted display of the relevant data)

The distributions of Recall in Figure 5.8 (a) for the positive nuclei are more similar among all setups than the previously discussed Precision P . A clear tendency towards better results is visible for $D_{designed}$ in comparison to $D_{uniform}$. The medians of R for $D_{uniform}$ all lie between 0.90 and 0.93 while those for $D_{designed}$ lie between 0.92 and 0.94. The same tendency can be observed in the IQR and RoE of all setups: For $D_{uniform}$, the RoE is always larger and ranges further down than for $D_{designed}$. For the positive channel, setup A on the designed Dataset performs best, with a median of about 0.93 and a comparably small IQR and RoE. An entirely different impression is obtained when looking at R for the negative channel in Figure 5.8 (b). Here, the results vary widely: The best result is found for setup A on $D_{uniform}$ with a high median R of 0.99 and IQR as well as RoE of 0.0. The worst result is found for setup F on $D_{uniform}$ with a median of 0.45, an IQR of 0.57 and an RoE of 0.64. Despite the large span of results, the same tendency is visible as in the positive channel, namely that all setups work better on $D_{designed}$ than on $D_{uniform}$. An evident question now is why R suggests that the nuclei detection algorithm generally works better on $D_{designed}$ than on $D_{uniform}$, which is contrary to the conclusions drawn from $Error_{LI}$ and P . This phenomenon can be explained with the definition of Recall per say: A high recall merely states that all relevant items have been identified, but does not express any information on the percentage of irrelevant items. On the sparsely populated $D_{designed}$, the massive misperformance of the thresholding step can lead to segmentation not only of nuclei, but also of other tissue structures visible in the respective channel. Illustrative examples for two contrary cases are shown in Figure 5.9.

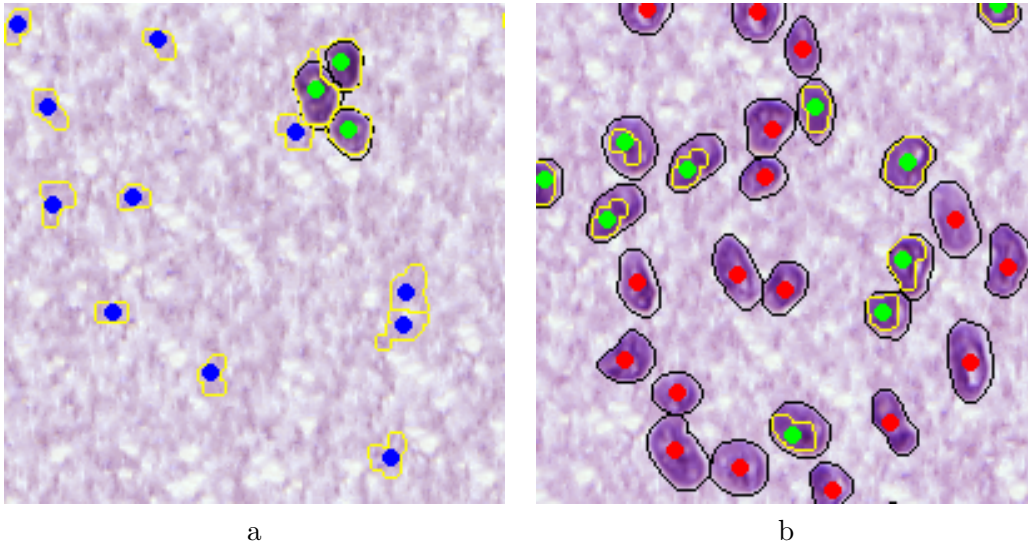


Figure 5.9: Examples for (a) high Recall despite obvious oversegmentation and (b) low recall due to undersegmentation. Both images stem from $D_{designed}$.

This underlines that the Recall alone would not sufficiently describe the performance of a setup and thus it has to be reported and understood in combination with the precision.

F1-score for each Setup/Dataset

The F1-score is defined as

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (5.3)$$

and describes the harmonic mean of Precision P and Recall R . Its value gives a good summary of both criteria, Precision and Recall. The closer F_1 to 1, the better, because an $F_1 = 1$ characterizes a high accuracy. A Recall of 0.5 states that only 50% of the present true nuclei have been detected. The ideal Recall of 1 is highlighted by a green line in Figure 5.10.

In Figures 5.10 (a) and (b), revealing the F1-score for the detection of positive, respectively negative nuclei, it is visible that the algorithm works superior on $D_{uniform}$ and noticeably worse on $D_{designed}$. Thus, the seemingly better Recall of the algorithm on the $D_{designed}$ is more than annihilated when averaged with the Precision. In line with previous findings drawn from interpreting the $Error_{LI}$, P and R of all setups and datasets, the F1-scores suggest the following conclusions:

- The Cosatto deconvolution utilized in setups A and B, clearly outperforms the other deconvolution methods
- The algorithm performs better the more densely populated the image is

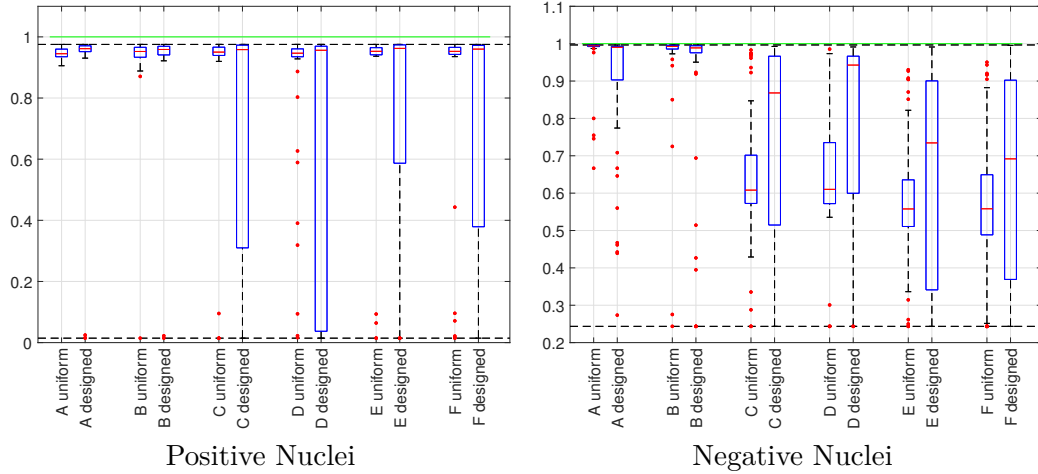


Figure 5.10: F_1 of Nuclei Detection for each setup A-F and Datasets $D_{uniform}$ and $D_{designed}$ (Note that the plots for positive and negative nuclei have different vertical scalings to allow for the best fitted display of the relevant data)

- Using the best setup, the algorithm yields F1-scores of above 0.98 for both positive and negative nuclei, proving it to be a reliable solution

Correlations between Image Features, LI and Labeling Index Error for each Setup and Dataset

To further investigate the claim that there is a correlation between the density of nuclei in an image and the $Error_{LI}$, the Pearson's Product-Moment Similarity Coefficient between the density and the $Error_{LI}$ for each Image was computed. As can be discerned from Figure 5.11, the correlation between the total density of nuclei and the $Error_{LI}$ is not significant, it oscillates around zero for all setups and both Datasets. While this might surprise at first sight, the explanation lies in the correlation between the density of positive nuclei and the $Error_{LI}$: it reveals that for setups C-F, all experiments conducted on $D_{designed}$ show a high negative correlation of less than -0.5 . This indicates that the $Error_{LI}$ tends to increase with decreasing density of positive nuclei in an image.

The $Error_{LI}$ is also investigated in its correlation to LI_{true} and the consistently moderate, but negative correlation is an indication that the $Error_{LI}$ decreases with increasing LI_{true} . This implies that the algorithm is able to more correctly deconvolve and threshold the images if a higher relative density of positive nuclei populates the image, adding weight to the positive nuclei in the histograms. Last but not least, the correlation between LI_{true} and LI_{est} is shown, once again demonstrating that setup A and B on $D_{uniform}$ outperform the other setups, with significant positive correlations of clearly above 0.6 each.

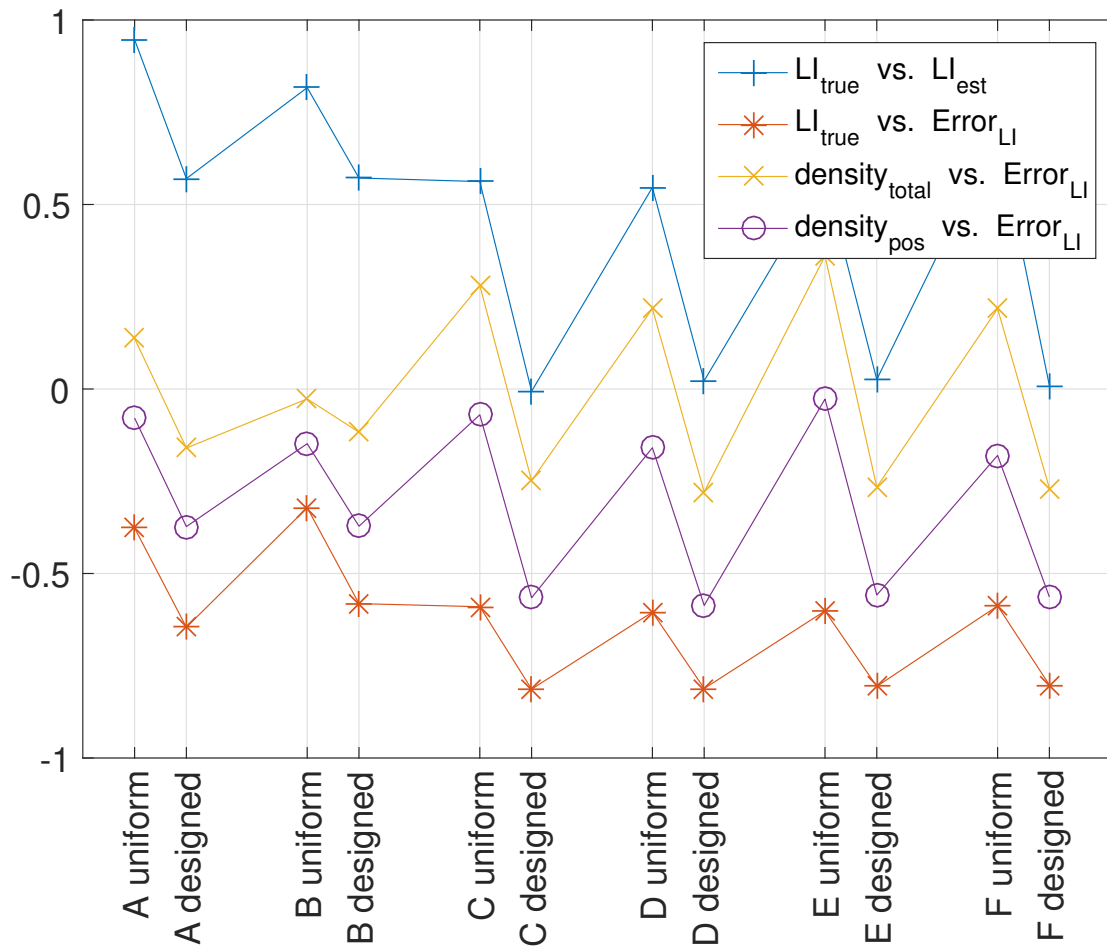


Figure 5.11: Correlations between Image Features, LI and $Error_{LI}$ for each setup and dataset (Lines between data-points added to facilitate identifying each type of correlation, no linear interpolation is expressed by them)

5.2.2 Nuclei Segmentation

This section deals with the performance of the nuclei segmentation. First, the visual output of the segmentation is presented and then a number of criteria demonstrate the performance of the nuclei segmentation and its potential diagnostic applicability for the pathologist.

Examples of the Segmentation Output (Visualization)

Figure 5.12 (a) and (b) illustrate how the segmentation result of both positive and negative nuclei is visualized. The blue dots represent FP. In both cases visible in (a) and (b), a nucleus has been over-segmented and divided into two nuclei while it should only be one. This is visualized by both a green dot (TP) and a blue dot (FP) located on the

same nucleus. In (a), two FN, represented by red dots, can be seen. They mark nuclei which have not been detected by the algorithm. In (b), a FN, represented by a red dot, lies adjacent to a TP (green dot). This is due to the fact that two nuclei were identified as one nucleus. The nucleus having the larger overlap with the RM is considered to be a TP and the green dot is always drawn in the center of the segmentation mask. The second nucleus, having the smaller overlap with the RM, is considered to be a FN and the red dot is always drawn in the center of the FN. While this depiction might not be intuitive at first sight, it proves that the evaluation procedure distinctly considers cases of over- or under-segmentation and labels each connected component within the segmentation mask as either a TP or a FP, while it prohibits a true nucleus from being labelled as TP twice.

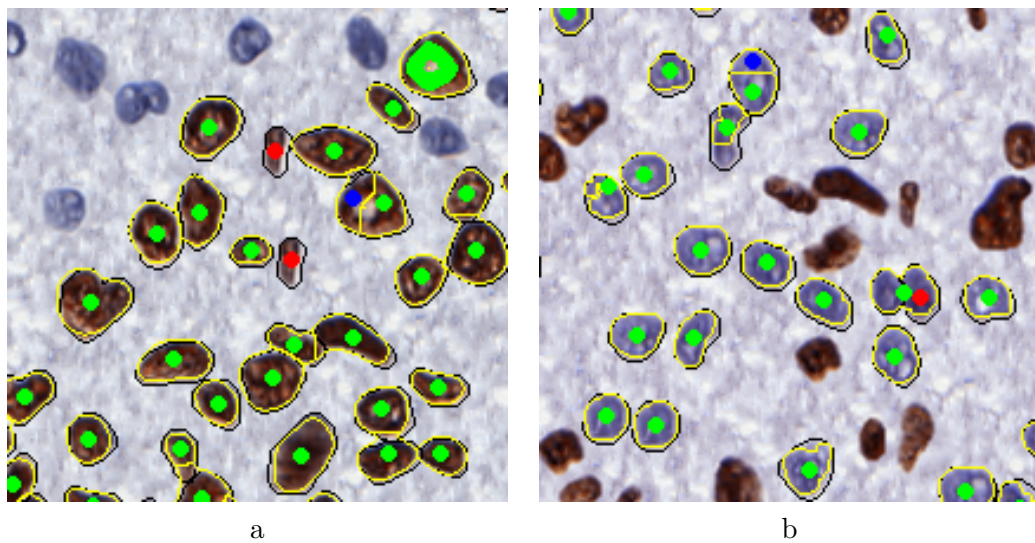


Figure 5.12: Visualization of the segmentation result: The ground truth outlines (True Mask, TM) are held in black, the outlines identified by the algorithm (Result Mask, RM) are held in yellow. TP are represented by green dots, FP by blue dots and FN by red dots. The two images show (a) the segmentation result of the positive nuclei and (b) the segmentation result of the negative nuclei, each in a different image.

In Figure 5.12 (a) there is a green ring visible in the upper right corner. This is due to the process of deriving the illustrative green dot depicting a TP: first, ultimate erosion is applied to the segmentation mask of the nucleus (i.e. shrinking it to one pixel diameter) to define the center of the located nucleus and then it is dilated back to a disk of 7 pixels diameters to be large enough for proper display in the visualization. If the ultimate erosion is applied on a CC with a hole, as in the present case, it is transformed to a ring and the subsequent dilation results in a thick ring. Despite the possibly misleading depiction, this nucleus is nevertheless counted as one TP nucleus.

Dice Similarity Coefficient for each Setup and Dataset

The Dice Similarity Coefficient DC

$$DC = \frac{|A \cap B|}{\frac{1}{2}(|A| + |B|)} \quad (5.4)$$

is a measure of overlap between two regions (in this case, a CC in the TM and a CC in the RM), weighted by their average area. A perfect segmentation yields $DC = 1$ while $DC = 0.5$ states that the object in the TM and the object in the RM only overlap by 50% in relation to their average area. The values reported in the boxplots of Figure 5.13 are based on the median DC of all objects per image.

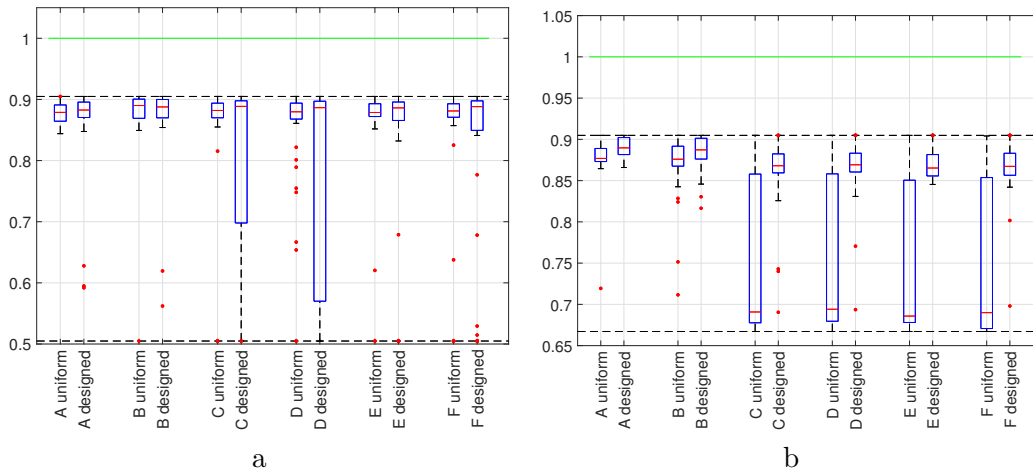


Figure 5.13: DC of Nuclei Detection for each setup A-F and Datasets $D_{uniform}$ and $D_{designed}$ (Note that the plots for positive and negative nuclei have different vertical scalings to allow for the best fitted display of the relevant data)

The DC for the segmentation of both positive (Figure 5.13 (a)) and negative (5.13 (b)) nuclei never exceeds 0.91. However, the DC for positive nuclei is more consistent than for the negative nuclei with values mainly between 0.85 and 0.90, except for setups C and D on $D_{designed}$. For the segmentation of the negative nuclei in setups C-F, the IQR and the mean of the DC systematically range below 0.7 for $D_{designed}$. The fact that the maximum value lies above 0.9 does not infer that all nuclei within even the best performing images have a DC of maximum 0.9, but in contrary – it is the median of the DC of all objects in the image. Thus, half of the nuclei are segmented in a quality higher than $DC = 0.9$. In accordance with the previous results from Section 5.2 these outcomes indicate that the algorithm performs better and more reliable when using the Cosatto deconvolution (setups A and B) and considerably worse on the more sparsely populated $D_{designed}$. In this context, the term "performance" refers to the segmentation performance.

Area Estimation Coefficient for each Setup and Dataset

As stated in Section 3.6.2, the DC alone does not disclose any information on whether the objects to be segmented are over- or underestimated. Hence, the criteria of Area Estimation $Area_{Est}$ is introduced:

$$Area_{Est} = \frac{TM_{Area}}{TM_{Area}} \quad (5.5)$$

It merely puts the segmented area of an object in the result mask, RM, in direct ratio to the true area of the object. This is done for nuclei verified as TP. Ideally $Area_{Est} = 1$, which is graphically stressed by the green line in Figure 5.14.

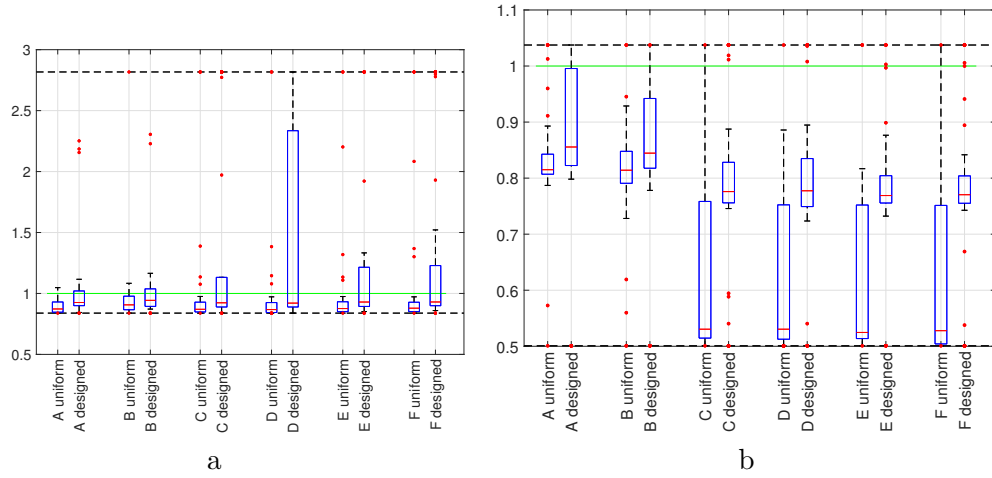


Figure 5.14: Area estimation coefficient $Area_{Est}$ showing the relation between detected and true area of the nuclei for each setup A-F and Datasets $D_{uniform}$ and $D_{designed}$ (Note that the plots for positive and negative nuclei have different vertical scalings to allow for the best fitted display of the relevant data)

The $Area_{Est}$ for both positive and negative nuclei is shown in Figure 5.14 (a) and (b), respectively. The boxplots reveal a tendency of underestimating the area more severely in $D_{uniform}$ than in $D_{designed}$. The medians for $D_{uniform}$ area located between 0.8 and 1 (positive nuclei) and between 0.75 and 0.85 (negative nuclei). The $Area_{Est}$ is generally more reliable and shows less variations for the segmentation of positive nuclei (Figure 5.14 (b)), except for setup D on $D_{designed}$, using the Macenko deconvolution on diversely stained $SI_{retained}$. $Area_{Est}$ of negative nuclei (Figure 5.14 (b)) on setups C-F is generally lower on $D_{uniform}$ than on $D_{designed}$, ranging down to $Area_{Est} = 0.5$.

Mind that the values reported for each setup and dataset are again the medians per image, not values per object within an image. As a consequence, it could happen that extreme over-estimation of some objects balances extreme under-estimation of other objects within the same image and the median value reported is still fairly good. Thus it is advised to

look at the Pearson's Coefficient of the Area Estimates as well for a more coherent picture.

Pearson's Product-Moment Correlation Coefficient of the Area for each Setup and Dataset

To shed a light on the incomplete expressiveness in the Area Estimation Coefficient, the Pearson's Product-Moment Correlation Coefficient between the estimated object areas A_{est} and the true object Areas A_{true} is also examined. The formula for this Coefficient, in short called Pearson's Coefficient or ρ_{Area} , is given in Equation 5.6, where X represents A_{est} and Y represents A_{true} :

$$\rho(X, Y) = \frac{cov(X, Y)}{(var(X)var(y))^{\frac{1}{2}}} \quad (5.6)$$

The Pearson's Coefficient is reported for all TP nuclei.

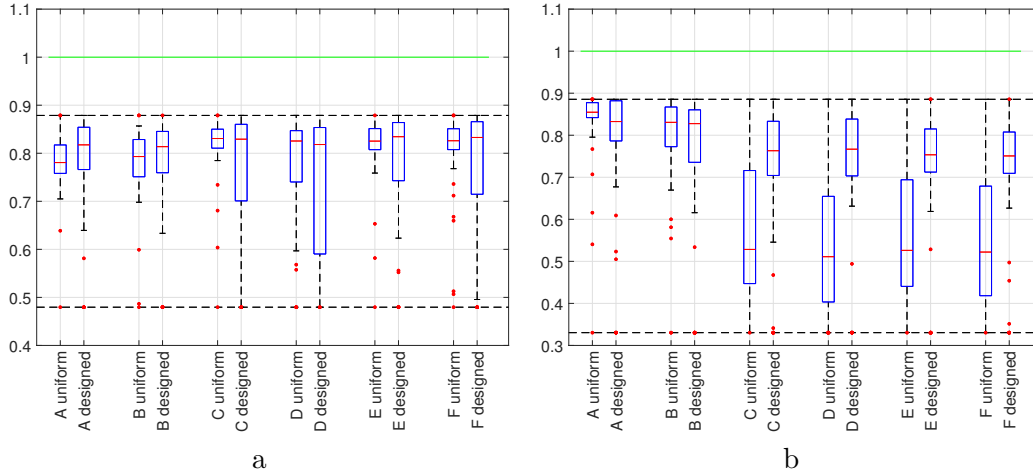


Figure 5.15: The Pearson's Coefficient, indicating the degree of linear relationship between the true area A_{true} and the estimated area A_{est} of the nuclei for each setup A-F and Datasets $D_{uniform}$ and $D_{designed}$.

(Note that the plots for positive and negative nuclei have different vertical scalings to allow for the best fitted display of the relevant data)

Figure 5.15 points out that there is a consistent positive linear correlation between estimated and true area of above 0.73 for both the negative nuclei and the positive nuclei, while the medians are all above 0.85. This indicates a significant correlation between A_{est} and A_{true} . In case of the positive nuclei, the IQR and RoE are always larger on the sparsely populated $D_{designed}$ than on $D_{uniform}$, where the contrary can be observed in case of the negative nuclei. Thus, the algorithm delivers by far more reliable results on the area estimates when using the Cosatto deconvolution (A-B) than with the Macenko or fixed deconvolution method (C-F).

It is very interesting that according to this criteria, the Cosatto deconvolution (setups A-B) performs worse on $D_{uniform}$ than the other two deconvolution methods (setups C-F), such that the former has a slightly lower median and a higher IQR and RoE. This is contrary to previous findings where the Cosatto deconvolution outperforms the other two methods. However, in connection with the previously discussed Area Estimation Coefficient $Area_{Est}$, it can be asserted that the algorithm tends to deliver more accurate measures on the area of the positive nuclei when using the Cosatto deconvolution on the uniform dataset, albeit not as reliably as the Macenko and fixed deconvolution.

Concerning the interpretation of these numbers, it has to be kept in mind, that even with a significantly high ρ , indicating a consistent reliability of the estimates, the areas of the nuclei could still be systematically over- or underestimated on a large scale. If it can be verified that a high ρ is based on a relationship with truly linear nature, any over- or under-estimation of the area could be corrected by a linear correction factor to deliver more accurate results. Figure 5.16 associates each A_{true} with the segmented A_{est} for two exemplary images and gives a visual confirmation that the underlying distribution is generally linear.

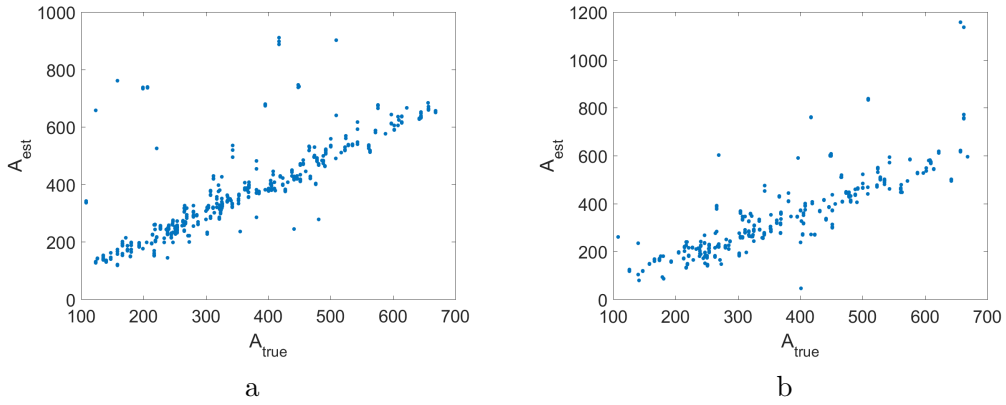


Figure 5.16: Examples for the correlation between A_{true} and A_{est} for an image of setup B on $D_{uniform}$ with reported (a) $Area_{Est} = 1.00$ and $\rho_{Area} = 0.83$ and (b) $Area_{Est} = 0.89$ and $\rho_{Area} = 0.85$

Pearson's Product-Moment Correlation Coefficient of the Solidity for each Setup and Dataset

As for the Area, the Pearson's Coefficient is utilized to examine the correlation between reported solidity, S_{est} , and true solidity, S_{true} as the variables X and Y in this Equation:

$$\rho(X, Y) = \frac{cov(X, Y)}{(var(X)var(y))^{\frac{1}{2}}} \quad (5.7)$$

Contrary to the correlation between estimated and true Area, ρ_{Area} , the correlation between estimated and true Solidity, $\rho_{Solidity}$, is much weaker, as can be seen in Figure

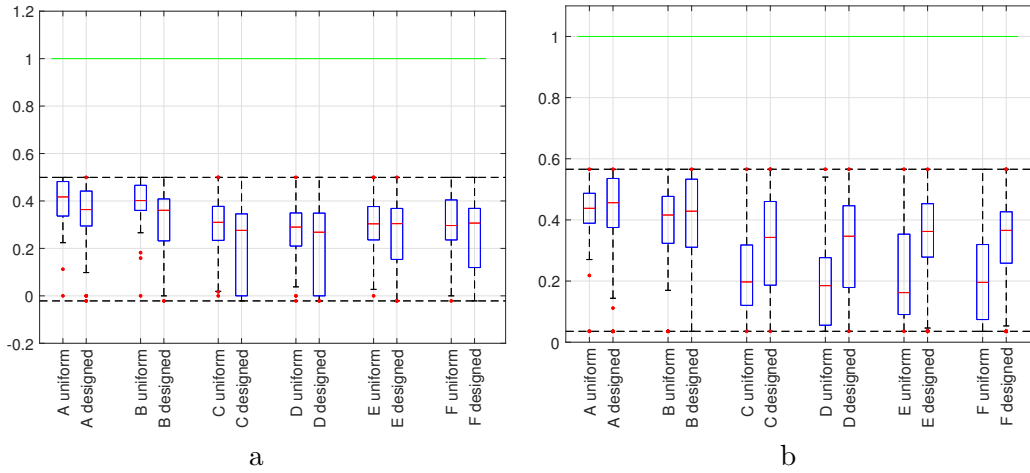


Figure 5.17: The Pearson's Coefficient, indicating the degree of linear relationship between the true solidity S_{true} and the estimated solidity S_{est} of the nuclei for each setup A-F and Datasets $D_{uniform}$ and $D_{designed}$.

(Note that the plots for positive and negative nuclei have different vertical scalings to allow for the best fitted display of the relevant data)

5.17. The medians of the correlation for the both positive and negative nuclei solidity never exceed 0.5. Even though the values for setups A and B are higher than those for the other setups, they are still too low to indicate a significant correlation. Hence, the solidity values of the detected objects cannot be reliably reported to the pathologist for any of the setups.

Pearson's Product-Moment Correlation Coefficient of the Eccentricity for each Setup and Dataset

To examine the estimated eccentricity, i.e. the ratio of the lengths between the foci and the major axis length of the ellipse that best fits the nuclei, the Pearson's Coefficient is utilized again. The estimated eccentricity, E_{est} , and true eccentricity, E_{true} , are represented by the variables X and Y in this Equation:

$$\rho(X, Y) = \frac{cov(X, Y)}{(var(X)var(y))^{\frac{1}{2}}} \quad (5.8)$$

As visible in Figure 5.18, there is a significant positive correlation between E_{est} and E_{true} for both positive and negative nuclei when conducting setups A and B. For the positive nuclei (Figure 5.18 (a)), all setups perform approximately equally well on $D_{uniform}$, with means of 0.80 to 0.85 and IQR of about 0.05. Looking at the performance on negative nuclei (Figure 5.18 (b)), only setups A and B perform better on $D_{uniform}$ than $D_{designed}$, while there is a higher correlation between estimated and true eccentricity for $D_{designed}$ than for $D_{uniform}$ for all other setups. This is an interesting observation, which is in

concordance with the results of the other segmentation criteria, DC , $Area_{Est}$, ρ_{Area} and $\rho_{Solidity}$ which all accredit setups C-F to perform better on the segmentation of negative nuclei when applied on $D_{designed}$ than on $D_{uniform}$.

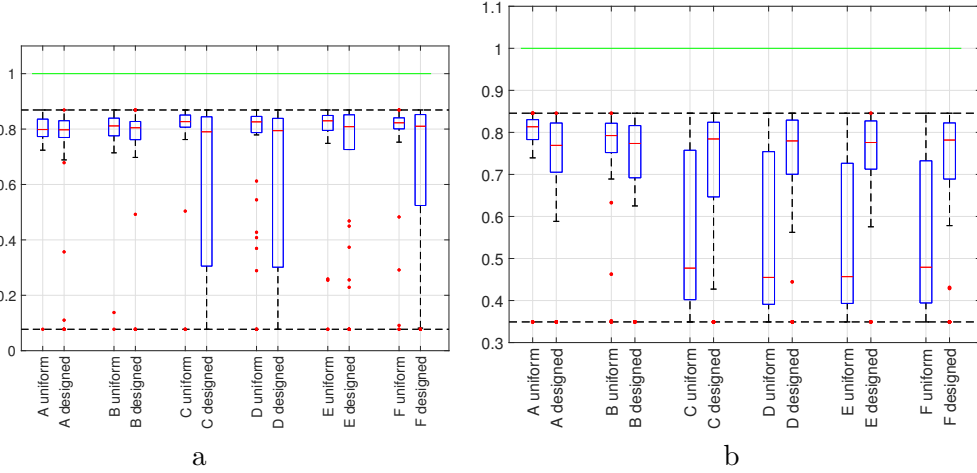


Figure 5.18: The Pearson's Coefficient, indicating the degree of linear relationship between the true eccentricity E_{true} and the estimated eccentricity E_{est} of the nuclei for each setup A-F and Datasets $D_{uniform}$ and $D_{designed}$.

(Note that the plots for positive and negative nuclei have different vertical scalings to allow for the best fitted display of the relevant data)

5.2.3 Summary and Discussion of Nuclei Quantification and Segmentation Results

The major research questions of this work (Section 1.3) was about the feasibility of automating the LI scoring in Ki-67 stained images of the breast and the achievable level of accuracy. This question is foremost answered by the presented Labeling Index Error $Error_{LI}$. It is notable, that to the publication date of this thesis and to the best knowledge of the author, no algorithm specialized on the analysis of the Ki-67 LI in breast images has been published. Hence, the performance of this algorithm can only be evaluated in comparison with DIA solutions for the Ki-67 LI analysis in tissue types other than breast.

Overall, the best achieved $Error_{LI}$ using the most advantageous setup (B) was -1.5% (median value, with IQR of 0.01) on a whole dataset including images with extremely sparse as well as high density, LI_{true} between 0 and 100% and varying staining characteristics. This setup used the deconvolution approach by Cosatto. This lowest $Error_{LI}$ is higher than the error reported for methods compared in Xing et al. [94] for estimation of the Ki-67 LI in NET tumor cases¹, which all range below 1%. A novel method presented

¹Neuro-Endocrine-Tumor

in Xing et al.[94] which is said to outperform all other compared methods however reveals that their method has a Precision, Recall and F1-Score of $P = 0.89$, $R = 0.91$ and $F_1 = 0.90$, respectively, while the values of the here presented work in the best setup (B) are $P = 0.99/0.99$ (for the positive/negative nuclei), $R = 0.90/0.98$ and $F_1 = 0.96/0.99$. Thus, the presented values indicate that the presented method works better than the solution by Xing et al. Yet, it has to be kept in mind that the comparison is flawed as Xing et al. worked on analyzing images which first of all stem from other tissue sites of the human body than breast and secondly which were real, annotated images rather than synthetic images as used here.

Overall, there were no notable differences in both the quantification and segmentation results between running the algorithm on either versions, the SI_{init} and the $SI_{restained}$, of the datasets. This supports the conclusion that for both the Cosatto and the Macenko deconvolution method the sheer diversity of the color appearances did not lead to a deterioration of the results. Nevertheless, as proved by almost all criteria presented in Section 5.2, in total Cosatto provides a more robust deconvolution method than Macenko, which leads to better results. As already stated in the discussion of $Error_{LI}$, the fixed value deconvolution performs the worst. All of these facts implicate that the choice of the deconvolution method as the first step in the image analysis chain has a major impact on the correctness of the automated Ki-67 LI analysis.

As notable in all criteria describing the results of the nuclei quantification, the more sparsely an image is populated with positive nuclei, the worse the results, suggested by the fact that the algorithm tends to perform worse on the overall more sparsely populated dataset $D_{designed}$. The thresholding step at a very early stage of the algorithm strongly depends on the presence of a perceivable contrast between negative/negative nuclei respectively and the tissue background. While this might emerge as a confinement on the SDS, it is highly unlikely that this impairment would also significantly show if the algorithm was applied on real data, as the regions on which such an analysis is usually conducted (so-called hotspots) do not exhibit large nuclei-void regions but in contrary contain a high percentage of positive nuclei. This conclusion hints at an aspect worth investigating further: the dataset characteristics should be limited to physiologically/pathologically occurring densities, which would also have ameliorate the performance of histogram-based deconvolution approaches such as the ones presented here.

Concerning the datasets used in this work it should be pointed out that the superior results of the algorithm on $D_{uniform}$ should not lead to the conclusion that nuclei are more easily detected when distributed uniformly across the image. Since the presented nuclei quantification and segmentation algorithm only conducts global, not local operations, the actual locations of the nuclei on a synthetic image do not in any way alter the result. Thus, if two images of the same size and stain appearance are populated with the same number of positive and negative nuclei, respectively and they only differ in the distribution of the nuclei in each image, the algorithm will output the same LI for both images and the segmentation quality will be equal as well.

Looking at the quantification results it becomes obvious that the low Precision values and F1-Scores in setups D-F for the positive nuclei could have been alleviated by adding an area-discriminative noise elimination step to processing the BW_{pos} , as it is done in the post-processing step for BW_{neg} .

It has to be kept in mind that bad results in the segmentation quality do not necessarily imply bad results of the quantification criteria. As long as a segmentation yields a DC of at least 0.2, it is counted as a TP. Thus, even if the areas of all nuclei are severely under- or overestimated and the segmentation regarding the shape of the objects is in the largest part incorrect, the quantification might still yield a correct result. In conclusion, it can be said that as long as the algorithm is able to differentiate between objects of interest (nuclei) and non-interest (noise and other unwanted structures) and these objects are described with a minimum quality ($DC > 0.2$), the output of the LI estimation can still be correct.

The criteria describing the segmentation quality point to the same conclusions as the criteria describing the quantification outcomes: The Cosatto method is the preferable choice for the deconvolution because it yields the segmentations with the highest Dice Coefficients. Although this does not show equally prominent in the other criteria (Area Estimation Coefficient and Pearson's Coefficient for Area, Solidity and Eccentricity), it is still observable they still support the same conclusion. As with the quality of the quantification results, the segmentation quality suffers from the sparsity of images, which can be seen in generally worse results on $D_{designed}$ than $D_{uniform}$.

Concluding the findings about the color deconvolution methods, it can be said that while the presupposition of stain characteristics, as when using a fixed vector for deconvolution, leads to large errors, the use of an entirely unbiased approach for stain vector identification as in the Macenko method is also not ideal if the contrast of the stains in the image is too weak. Cosatto et al. presented a stable method which is a fair compromise between a small set of assumptions and a sufficient amount of adaptability to varying stain appearances.

Of course, the level of accuracy (in this regard, the correctness of the LI score) can be increased when using methods more complex than the here applied methods, such as supervised training, sophisticated shape-, color- and texture based feature classification steps or different unsupervised learning techniques known from other fields of digital image analysis applications. The presented algorithm is able to demonstrate that a high accuracy can already be achieved by seizing the advantages of basic image processing methods, while making as few assumptions about color-related or physiological aspects of the nuclei and surrounding tissue.

While the evaluation of this algorithm on the SDSs delivers both quantitative and qualitative metrics about its performance, it would be highly desirable and insightful to test it on a real, clinical dataset as well. Since all research in the digital pathology domain aims at either contributing to disease knowledge in large research studies or facilitating and enriching the daily diagnostic tasks of the pathologist, the relevant images of interest

are always WSI. To reach these aims, a SDS can always only substitute real images to a certain degree, but never replace them.

Conclusion

In the following paragraphs, the findings of the work conducted in the course of this thesis are summarized by recapitulating the stated research questions and their respective answers as well as mentioning other noteworthy discovered aspects.

Can scoring of the Labeling Index in Ki-67 stained slides (LI estimation) be performed in an automated fashion, requiring neither prior training of the program nor manual inputs from the user? If so, to which point of accuracy?

In the course of this work, a pipeline of steps has been developed, which deconvolves a given image into two channels, then binarizes and segments each channel via several post-processing steps, using solely unsupervised image processing operations and making barely any assumptions about the shape or size of the nuclei. At the end of the pipeline the nuclei in the synthetic images are segmented.

It turns out that the initial step of color deconvolution is the most decisive factor in the quality of the results. It is concluded that, for Ki-67 images, a partly biased deconvolution method such as the one by Cosatto et al. [11] is more stable to stain variations than an entirely unbiased approach such as the one by Macenko et al. [50] and both are preferable to fixed assumptions about the stain appearance. Concludingly, the deconvolution method of choice for this algorithm is the Cosatto deconvolution.

The algorithm does not need to be trained on a labeled or annotated set of images and it does not require any user-input prior or during its execution. It is able to adapt to stain appearance variations as they occur in data from a single laboratory.

Most room for improvement of the presented algorithm lies in making it more stable towards the lower range of the Labeling Index, such that even images with sparse density and a low rate of Ki-67 stained positive nuclei are correctly segmented.

As proved in the conducted experiments, using different dataset characteristics and algorithm settings, the best results are achieved when the test images are well populated

and the deconvolution is only partly biased. Here, a Labeling Index Error of only 1.5%, together with a Precision of 0.99, a Recall of 0.90 and an F1-score of 0.96 can be achieved. These values are within the range of accuracy found in other publications about Ki-67 Labeling Index estimation.

Is a synthetic, labeled dataset suitable to evaluate the performance of such an algorithm?

While a fully labeled and annotated dataset of real clinical/research WSI of Ki-67 stained slides of the breast, including the entire physiological range of LI and various kinds of tricky tissue scenarios, would certainly promote the development and valid comparison of accurate nuclei detection and segmentation algorithms, to date no such dataset exists. Nevertheless, a synthetic dataset constitutes a profound basis for the task of evaluating an algorithm which aims at nuclei quantification and segmentation in Ki-67 stained breast images. The image generation procedure presented in its current state allows the exact evaluation of the given algorithm in a qualitative (LI-related) and quantitative (pixel-wise segmentation-related) manner.

The potential of these image generation procedure is not yet fully exploited. Images generated this way can not only be used to test and evaluate unsupervised methods, but could also serve for the training, testing and evaluation of supervised DIA solutions. Furthermore, by fully utilizing the possibility of predefining the nuclear placement during image generation a dataset produced can also be applied for training and testing tissue classification methods e.g. to differentiate between suspicious and normal tissue areas. To this end, the inclusion of additional types of nuclei (e.g. epithelial nuclei) or the incorporation of directional nuclei bias features, as well as placing entire annotated cells (including nuclei and cytoplasm) would enrich the applicability for more sophisticated detection and/or classification tasks.

Future Work

This chapter deals with potential directions of improvements of this work and proposes ideas and possibilities on how to tackle them. Furthermore it covers potential directions this algorithm and dataset synthesis could take when investigated further.

7.1 Recommendations on the Improvement of the Suggested Solution

The computational time of the automated Ki-67 plays a subsidiary role in this work, but may aid in giving an insight into the computational complexity of this solution. To compute the score of an image within the range of 1200x1200 to 3214x1803 pixels at 20x magnification, the presented work takes in the range between 26s and 52s, depending on the size of the image and the number of nuclei in it. Compared to a work by [94], which takes about 100s to detect cells on a digitized 2310x2150 pixels image at 20x magnification, the presented segmentation and scoring algorithm is not considerably faster. If a pathologist opens an image and requests its Ki-67 score, these durations would be perceived as too slow, especially since pathologists would not only look at one field of view, but want to be able to freely pan and zoom through entire .svs-files in ranges of 70,000x40,000 pixels or more and still receive the numbers for each studied region. However, the presented CIA solution still has the potential to be a huge time saver for them, because they would otherwise spend time in the order of 30 minutes ([94]) to yield a comparable manual count of all nuclei in a field of view of equal size as the images presented in this work. Furthermore, the automatic Ki-67 scoring can be conducted without user-interaction as soon as an image is stored on the server, thus may present a preliminary result already on opening by the pathologist.

As proved in the results and discussion, when using adaptive deconvolution methods the correct functioning of the algorithm is severely dependent on the density of nuclei,

especially positive nuclei. This can represent a large impairment for analyzing images and regions like the one depicted in Figure 7.1. It shows a region in a real, clinical .svs-file, which is not only generally sparsely populated but (as can be visually judged by the ratio of Ki-67 positive to Ki-67 negative cells) also exhibits a very low LI. The nuclei detection and segmentation algorithm would most likely fail to deliver an accurate result of the LI because the adaptive deconvolution step is not able to handle such sparsity. A possible solution would be to deploy a fixed-vector based deconvolution as a first guidance to coarsely identify Regions Of Interest (ROI) which contain Ki-67 positive cells. In these regions it would be advisable to conduct in-depth Ki-67 scorings, using adaptive deconvolution algorithms. Given one of the definitions of a hotspot (see Section 2.3.1), Ki-67 scorings in locations other than possible hotspots do not render diagnostic consequences, thus a limitation of any analysis to ROI is highly advisable.

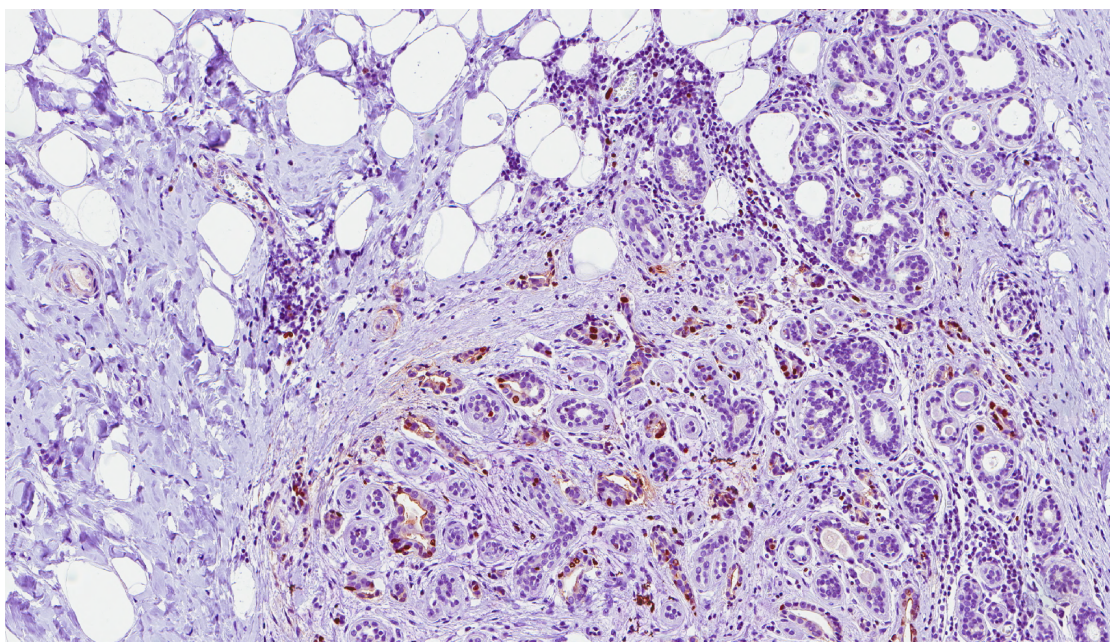


Figure 7.1: Region of a Ki-67 image in which left side is generally sparsely populated and the entire image contains only few Ki-67 positive cells

The topic of ROI identification was briefly investigated in the course of this work and an ROI identification method proposed by Bahlmann [3] was implemented coarsely. It is based on the cumulative histograms of H, respectively E channels for sub-image tiles (Bahlmann suggests the Cosatto deconvolution also discussed in this thesis to derive the channels). The 0, 10, 20, . . . , 100% percentile values of the cumulative histograms of both channels of each tile are combined to yield a feature vector with 22 features per tile. A support vector machine is suggested to classify the feature vectors into two classes, ROI and non-ROI. For this work, Cosatto is also used as the deconvolution method and instead of using a support vector machine for classification, k-means serves to divide the

feature vectors and concomitant tiles into the clusters ROI and non-ROI. In the example image in Figure 7.2, the clustering was initiated with three clusters instead of two as proposed in [3] and the cluster containing the tiles with the highest cumulative histogram of the positive channel was considered to be the ROI regions. The result can be seen in Figure 7.2. If only two clusters are used, the method is too sensitive and labels too many regions, containing the slightest faint of Ki-67 positive stain, as ROI regions.

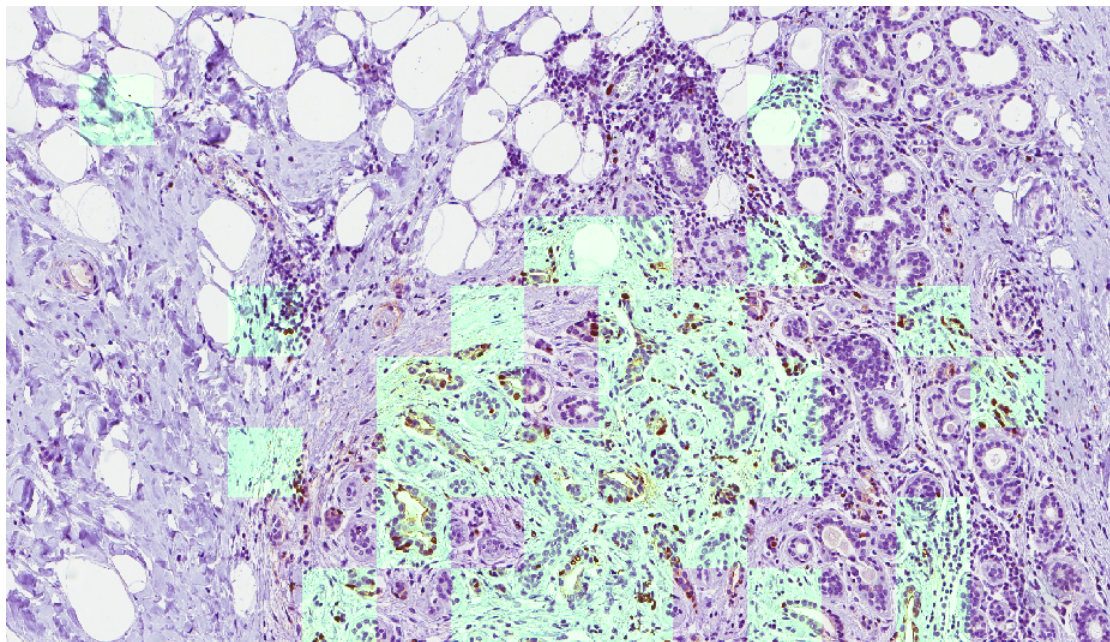


Figure 7.2: Test implementation on ROI identification, based on [3], using three clusters. The regions assigned to the most positive cluster are highlighted in a light-green shade.

Successfully handling common image-inherent artifacts like smeared cells, uneven illumination conditions resulting in inconsistent intra-image brightness or tissue folds with excessive staining intensities remain an open issue in the presented nuclei quantification algorithm. As an example: the clinical dataset at hand contains an unaccountable optical artifact which is manifested in a purple shade along the upper edge of all nuclei, shown in Figure 7.3. In the lower row it can be seen how prominently this edge lights up in the negative channel, where there should be no signal because it is physiologically impossible that a small portion of the nucleus is in another phase than the principal nucleus thus does not take up the Ki-67 stain, or that this portion is always located at the upper edge of the nucleus. An artefact like this requires shape- and size-dependent distinction between candidate nuclei, a step which is linked to preferably avoidable assumptions.

As discussed in Section 4.1.3, the implementation of the SDS allows the definitions of $Map_{NucProbability}$, which constraints the nuclear placement for positive and negative nuclei, separately. However, only the upper left corner of the bounding box of a Snippet S_{pos} or S_{neg} , containing one or several nuclei, is compared with this map, thus it is

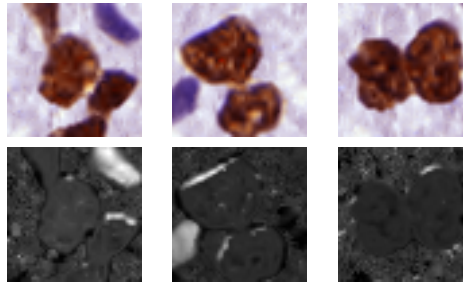


Figure 7.3: A purple shade along the upper edge of Ki-67 positive nuclei in an SDS-image. The lower row depicts the appearance of those shades in the negative channel of each image, deconvolved using Cosatto [11].

possible that nuclei encroach into supposedly prohibited regions. This can be seen in Figure 7.4, where the zones defined for placement of S_{pos} and S_{neg} are also visualized. Some nuclei obviously protrude the allowed regions to the lower right. This limitation can easily be avoided through the incorporation of a comparison between a candidate S_{pos} and the $Map_{NucProbability}$ which considers not only the upper left corner of the snippets bounding box, but the entire $SnipMask_{Conn}$.



Figure 7.4: An example of how the snippets S_{pos} and S_{neg} , containing one or several nuclei, are placed within the constraints of a nuclear probability map, $Map_{NucProbability}$. The dashed-brown line defines the allowed region for S_{pos} , the faint purple line for S_{neg} . Some nuclei appear in the respective forbidden areas.

Another chance for improvement lies in the thresholding step of the nuclei segmentation

and quantification algorithm, after the deconvolution and contrast adjustment. It is advisable to test local instead of global thresholds for thresholding the channels to insure the inclusion of more faintly stained nuclei. An example is shown in Figure 4.11 on page 51, where the upmost positive nucleus partly vanishes from (a) to (c).

Last but not least, the major aspect placing a limitation on the presented work is the use of a single dataset, which is based on images from a single laboratory and is thus constrained in its range of breast cancer cases, as well as in the variation of its staining characteristics, possible artifacts and other aspects. Despite being common practice, the usage of a single dataset for both developing and evaluating an algorithm is always questionable, not only in image analysis in the field of digital pathology. The risk of this methodology is that the method may be over-fitted to the limited training data, which by itself depreciates the expressiveness of the presented results and the applicability on other data. However, this is a limitation which most publications using a specific, homogeneous dataset suffer from [11]. As pointed out before, a publicly available, large benchmark dataset for the quantification and segmentation of Ki-67 images of the breast would clearly augment the validity of any solution proposed in this métier.

7.2 Further Opportunities with the Suggested Image Synthesis Method

The presented image synthesis method offers some additional potential applications besides serving as a Gold Standard for the evaluation of this algorithm. The inclusion of constraint maps allows the definition of ground-truth models of nuclei arrangement as they appear around ducts or in stromal areas. One example can be seen in Figure 7.5, which depicts a scene from a real image with a large, central stromal area and a duct on the bottom, and the corresponding synthetic remake of these relevant structures. The latter is an image of the dataset $D_{designed}$ used in this work. This opportunity can be utilized to generate images for region classification training, where algorithms learn e.g. to differentiate between stromal and ductal regions based on the nuclear arrangement.

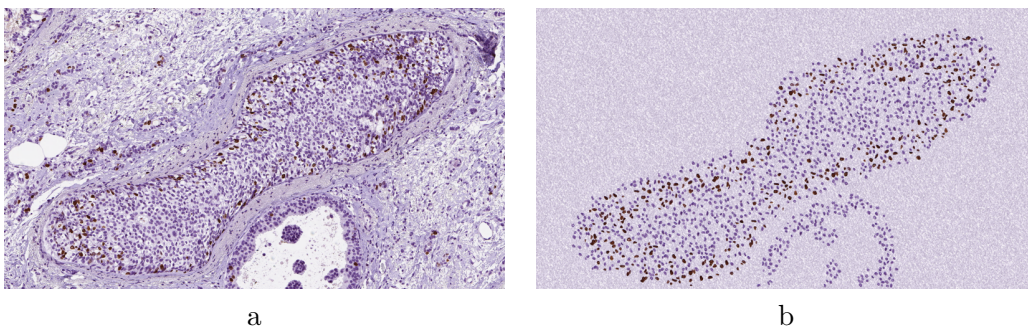


Figure 7.5: Example for (a) a real-world image and (b) its synthetic equivalent for potential training and testing of tissue classification

Referring to the directional bias of the nuclei in the dataset (see Section 5.1.2) the current implementation of the SDS generation limits the usability of the SDS to the verification of methods which also disregard the directionality. If an algorithm includes tissue type analysis (e.g. for the detection of ducts) it might want to seize the directionality feature. In this case it would be beneficial to include a deliberate step in which the nuclei are rotated to a desired angle. This could simply be achieved via providing a map of desired angles in specific areas, much like the $Map_{NucProbability}$, and then rotating each nucleus/snippet accordingly before its placement on the SI_{init} .

To be fully utilized as a dataset for region classification tasks, the image synthesis process would definitely benefit from incorporating a number of additional features beside the directional bias of the nuclei. Two examples are the inclusion of cells (i.e. nucleus and cytoplasm) rather than nuclei, with both nucleus and cytoplasm annotated in detail and/or including nuclei from other tissue types of the breast, such as the more dense and elongated nuclei found in epithelial tissue. Again, since one of the limiting factors in digital pathology research is the time-wise availability of pathologist for annotating data and both suggested measures of improvement require expert annotations, these improvements would come at a high cost. However, considering the limited availability of pathologists it is debatable which of the two options would turn out more beneficiary for the research community: The first option is to set up one fully labeled, annotated, diverse dataset of Ki-67 stained images of the breast and to make it publicly available to allow for a comparable evaluation of DIA solutions. The second option is to enrich and enhance the possibilities of an existing image synthesis method such as the here presented, which would come at an equal or even lower time-wise expense of pathologists and allows the generation of a nearly unlimited amount of images with different characteristics.

The possibility of creating synthetic images does not only bring up the question of how the community can benefit from it, but also whether there is a potential danger in this direction. Further down the road, when synthetic images become more and more realistic, it might become increasingly hard for a pathologist to differentiate between a real and a synthetic image. While the indistinguishability is not ethically questionable in and of itself, the fact that it opens the door to fraudulent misuse in research and clinical practice should be kept an eye on.

Appendix

Abbreviations

CAD	Computer Aided Diagnosis
CC	Connected Component
CMYK	Cyan Magenta Yellow Key
DAB	Di-Amino-Benzidine
DIA	Digital Image Analysis
DT	Distance Transform
ER	Estrogen Receptor
FN	False Negative
FOV	Field of View
FP	False Positive
HER	Human Epidermal Growth-factor 2
HPF	High Power Field
IHC	Immuno-Histo-Chemistry
IQR	Inter Quartile Range
LI	Labeling Index
OD	Optical Density
PR	Progesterone
RoE	Range of Extremes
ROI	Region of Interest
SD	Standard Deviation
SDS	Synthetic Data-Set
SE	Structuring Element
SVD	Singular Value Decomposition
TMA	Tissue Microarray
TP	True Positive
WSI	Whole Slide Images

Nomenclature

$Area_{Est}$	The Area Estimation of each CC
BW_{neg}	Initial binary mask (Black-White) of the negative Channel after Quantization with k-means
BW_{pos}	Initial binary mask (Black-White) of the positive Channel after Quantization with k-means
BW_X	Binary Image (Black-White)
CC_{neg}	Connected Component in the negative Channel
CC_{pos}	Connected Component in the positive Channel
DC	Dice Similarity Coefficient
$Error_{LI}$	The Labeling Index Error
F_1	F1-score
GS_X	Gray-Scale Image
I_{RGB}	The original RGB Image
LI_{est}	The Labeling Index as estimated by the algorithm
LI_{true}	The Labeling Index as in the Ground Truth
Loc_{Col}	Column of the Location chosen for the placement of a N_{pos} or N_{neg}
Loc_{Row}	Row of the Location chosen for the placement of a N_{pos} or N_{neg}
M_{source}	Matrix with two vectors describing the original stain appearance of an image
M_{target}	Matrix with two vectors describing the desired stain appearance of an image
$Map_{NucProbability}$	Map describing probability for placement of a Snippet (0%= forbidden, black; 100% = allowed, white)
$Mask_{Conn}$	Binary Image of a Channel containing the CCs of every placed Snippet (S_{pos} or S_{neg})
$Mask_{Cut}$	Binary Image of a Channel containing the CCs of every placed Nucleus (N_{pos} or N_{neg})
N_{neg}	Ki-67 negative Nucleus
N_{pos}	Ki-67 positive Nucleus
P	Precision
P_{Loc}	The randomly generated Probability of placing a Nucleus at a certain Location (to be compared against $Map_{NucProbability}$ at this location)
R	Recall
RM	Result Mask
RM_{Area}	Area of a CC in the RM
RM_{neg}	Result Mask of the negative Channel
RM_{pos}	Result Mask of the positive Channel
Rnd_S	Randomly chosen Snippet
S_{neg}	Snippet containing one or several connected N_{neg}
S_{pos}	Snippet containing one or several connected N_{pos}
SE	Structuring Element

SI_{init}	Synthetic Image in Initial state (before restraining)
$SI_{restained}$	Synthetic Image in Initial state (after restraining)
$SnipMask_{Conn}$	Image of same size as Snippet S_{pos} or S_{neg} , respectively, containing the the corresponding region in the $Mask_{Conn}$
$SnipMask_{Cut}$	Image of same size as Snippet S_{pos} or S_{neg} , respectively, containing the corresponding region in the $Mask_{Cut}$
TM	True Mask
TM_{Area}	Area of a CC in the TM
TM_{neg}	True Mask of the Negative Channel
TM_{pos}	True Mask of the Positive Channel
WS	Watershed-Transform

Bibliography

- [1] G. Apou, F. Feuerhake, G. Forestier, B. Naegel, and C. Wemmert. Synthesizing whole slide images. In *2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 154–159. IEEE, 2015.
- [2] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.
- [3] C. Bahlmann, A. Patel, J. Johnson, J. Ni, A. Chekkoury, P. Khurd, A. Kamen, L. Grady, E. Krupinski, A. Graham, et al. Automated detection of diagnostically relevant regions in H&E stained digital pathology slides. In *SPIE Medical Imaging*, pages 831504–831504. International Society for Optics and Photonics, 2012.
- [4] B. E. Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Höller, A. Homeyer, N. Karssemeijer, and J. A. van der Laak. Stain specific standardization of whole-slide histopathological images. *IEEE transactions on medical imaging*, 35(2):404–415, 2016.
- [5] S. Beucher and F. Meyer. The morphological approach to segmentation: the watershed transformation. In E. Dougherty, editor, *Mathematical morphology in image processing*, volume 34, pages 433–481. Marcel Dekker AG, 1992.
- [6] K. L. Bolton, M. Garcia-Closas, R. M. Pfeiffer, M. A. Duggan, W. J. Howat, S. M. Hewitt, X. R. Yang, R. Cornelison, S. L. Anzick, P. Meltzer, S. Davis, P. Lenz, J. D. Figueroa, P. D. Pharoah, and M. E. Sherman. Assessment of automated image analysis of breast cancer tissue microarrays for epidemiologic studies. *Cancer Epidemiology Biomarkers & Prevention*, 19(4):992–999, 2010.
- [7] L. E. Boucheron. *Object-and spatial-level quantitative analysis of multispectral histopathology images for detection and characterization of cancer*. PhD thesis, 2008.
- [8] W. Burger and M. J. Burge. *Digitale Bildverarbeitung: Eine algorithmische Einführung mit Java*. Springer-Verlag, 2006.
- [9] T. F. Chan and L. A. Vese. Active contours without edges. *IEEE transactions on Image processing*, 10(2):266–277, 2001.

- [10] Q. Chaudry, S. H. Raza, Y. Sharma, A. N. Young, and M. D. Wang. Improving renal cell carcinoma classification by automatic region of interest selection. In *8th IEEE International Conference on BioInformatics and BioEngineering*, pages 1–6. IEEE, 2008.
- [11] E. Cosatto, M. Mille, H. Grad, and J. Meyer. Grading nuclear pleomorphism on histological micrographs. In *19th International Conference on Pattern Recognition*, pages 1–4, 2008.
- [12] P. Desmeules, H. Hovington, M. Nguilé-Makao, C. Léger, A. Caron, L. Lacombe, Y. Fradet, B. Têtu, and V. Fradet. Comparison of digital image analysis and visual scoring of KI-67 in prostate cancer prognosis after prostatectomy. *Diagnostic pathology*, 10(1):67, 2015.
- [13] S. Di Cataldo, E. Ficarra, A. Acquaviva, and E. Macii. Automated segmentation of tissue images for computerized IHC analysis. *Computer methods and programs in biomedicine*, 100(1):1–15, 2010.
- [14] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.
- [15] M. DiFranco, G. O’Hurley, E. Kay, W. Watson, and P. Cunningham. Ensemble based system for whole-slide prostate cancer probability mapping using color texture features. *Computerized medical imaging and graphics*, 35:629–645, 2011.
- [16] M. Dowsett, T. O. Nielsen, R. A’Hern, J. Bartlett, R. C. Coombes, J. Cuzick, M. Ellis, N. L. Henry, J. C. Hugh, T. Lively, et al. Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. *Journal of the National Cancer Institute*, 103(22):1656–1664, 2011.
- [17] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification and Scene Analysis Part 1: Pattern Classification. *John Wiley & Sons, New York*, 1998.
- [18] C. W. Elston and I. O. Ellis. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410, 1991.
- [19] S. Fasanella, E. Leonardi, C. Cantaloni, C. Eccher, I. Bazzanella, D. Aldovini, E. Bragantini, L. Morelli, L. Cuorvo, A. Ferro, et al. Proliferative activity in human breast cancer: Ki-67 automated evaluation and the influence of different ki-67 equivalent antibodies. *Diagnostic pathology*, 6(1):1, 2011.
- [20] P. L. Fitzgibbons, D. L. Page, D. Weaver, A. D. Thor, D. C. Allred, G. M. Clark, S. G. Ruby, F. O’Malley, J. F. Simpson, J. L. Connolly, et al. Prognostic factors in breast cancer: College of American Pathologists consensus statement 1999. *Archives of pathology & laboratory medicine*, 124(7):966–978, 2000.

- [21] T. J. Fuchs and J. M. Buhmann. Computational pathology: Challenges and promises for tissue analysis. *Computerized Medical Imaging and Graphics*, 35(7):515–530, 2011.
- [22] A. Gertych, A. O. Joseph, A. E. Walts, and S. Bose. Automated detection of dual p16/Ki67 nuclear immunoreactivity in liquid-based Pap tests for improved cervical cancer risk stratification. *Annals of biomedical engineering*, 40(5):1192–1204, 2012.
- [23] J. Gibbons and S. Chakraborti. *Nonparametric Statistic Inference*. New York: Dekker, 4th edition, 2003.
- [24] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Prentice Hall, 2007.
- [25] P. Harrison. *Image Texture Tools*. PhD thesis, Clayton School of Information Technology, Monash University.
- [26] A. I. Hida, Y. Oshiro, H. Inoue, H. Kawaguchi, N. Yamashita, and T. Moriya. Visual assessment of Ki67 at a glance is an easy method to exclude many luminal-type breast cancers from counting 1000 cells. *Breast Cancer*, 22(2):129–134, 2015.
- [27] J. Hipp, S. C. Smith, J. Cheng, S. A. Tomlins, J. Monaco, A. Madabhushi, L. P. Kunju, and U. J. Balis. Optimization of complex cancer morphology detection using the SIVQ pattern recognition algorithm. *Analytical cellular pathology*, 35(1):41–50, 2012.
- [28] J. D. Hipp, J. Y. Cheng, M. Toner, R. G. Tompkins, U. J. Balis, et al. Spatially Invariant Vector Quantization: A pattern matching algorithm for multiple classes of image subject matter including pathology. *Journal of pathology informatics*, 2(1):13, 2011.
- [29] J. Ho, S. M. Ahlers, C. Stratman, O. Aridor, L. Pantanowitz, J. L. Fine, J. A. Kuzmishin, M. C. Montalto, A. V. Parwani, et al. Can digital pathology result in cost savings? A financial projection for digital pathology implementation at a large integrated health care organization. *Journal of pathology informatics*, 5(1):33, 2014.
- [30] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential. *IEEE reviews in biomedical engineering*, 7:97–114, 2014.
- [31] Ł. Jeleń, T. Fevens, and A. Krzyżak. Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies. *International Journal of Applied Mathematics and Computer Science*, 18(1):75–83, 2008.
- [32] K. J. Kaplan and L. K. Rao, editors. *Digital Pathology*. Springer International Publishing, 2016.

- [33] B. Karaçali and A. Tözeren. Automated detection of regions of interest for tissue microarray experiments: an image texture analysis. *BMC Medical Imaging*, 7(1):1, 2007.
- [34] J. N. Kather, C.-A. Weis, A. Marx, A. K. Schuster, L. R. Schad, and F. G. Zöllner. New Colors for Histology: Optimized Bivariate Color Maps Increase Perceptual Contrast in Histological Images. *PloS one*, 10(12), 2015.
- [35] A. Khan, N. Rajpoot, D. Treanor, and D. Magee. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering*, 61(6):1729–1738, 2014.
- [36] A. M. Khan, H. El-Daly, E. Simmons, N. M. Rajpoot, et al. HyMaP: A hybrid magnitude-phase approach to unsupervised segmentation of tumor areas in breast cancer histology images. *Journal of pathology informatics*, 4(2):1, 2013.
- [37] Y.-J. Kim, B. Romeike, J. Uszkoreit, and W. Feiden. Automated nuclear segmentation in the determination of the Ki-67 labeling index in meningiomas. *Clinical neuropathology*, 25(2):67–73, 2006.
- [38] S. Kothari, Q. Chaudry, and M. D. Wang. Automated cell counting and cluster segmentation using concavity detection and ellipse fitting techniques. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 795–798. IEEE, 2009.
- [39] S. Kothari, J. H. Phan, T. H. Stokes, and M. D. Wang. Pathology imaging informatics for quantitative analysis of whole-slide images. *Journal of the American Medical Informatics Association*, 20(6):1099–1108, 2013.
- [40] M. Kowal, P. Filipczuk, A. Obuchowicz, J. Korbicz, and R. Monczak. Computer-aided diagnosis of breast cancer based on fine needle biopsy microscopic images. *Computers in biology and medicine*, 43(10):1563–1572, 2013.
- [41] A. Kårsnäs. *Image Analysis Methods and Tools for Digital Histopathology Applications Relevant to Breast Cancer Diagnosis*. PhD thesis, Uppsala University.
- [42] G. Lang. *Histotechnik: Praxislehrbuch für die biomedizinische Analytik*. Springer-Verlag, 2nd edition, 2013.
- [43] A. Laurinavicius, B. Plancoulaine, A. Laurinaviciene, P. Herlin, R. Meskauskas, I. Baltrusaityte, J. Besusparis, D. Dasevicius, N. Elie, Y. Iqbal, et al. A methodology to ensure and improve accuracy of Ki67 labelling index estimation by automated digital image analysis in breast cancer tissue. *Breast Cancer Res*, 16(2):R35, 2014.
- [44] A. Lehmussola, P. Ruusuvoori, J. Selinummi, H. Huttunen, and O. Yli-Harja. Computational framework for simulating fluorescence microscope images with cell populations. *IEEE Transactions on Medical Imaging*, 26(7):1010–1016, 2007.

- [45] O. Lezoray and H. Cardot. Cooperation of color pixel classification schemes and color watershed: a study for microscopic images. *IEEE Transactions on Image Processing*, 11(7):783–789, 2002.
- [46] S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.
- [47] X. M. Lopez, O. Debeir, C. Maris, I. Roland, I. Salmon, and C. Decaestecker. Ki-67 hot-spots detection on glioblastoma tissue sections. In *ISBI*, pages 149–152, 2010.
- [48] X. M. Lopez, O. Debeir, C. Maris, S. Rorive, I. Roland, M. Saerens, I. Salmon, and C. Decaestecker. Clustering methods applied in the detection of Ki67 hot-spots in whole tumor slide images: An efficient way to characterize heterogeneous tissue-based biomarkers. *Cytometry Part A*, 81(9):765–775, 2012.
- [49] C. Lundström, S. Thorstenson, M. Waltersson, A. Persson, and D. Treanor. Summary of 2nd Nordic symposium on digital pathology. *Journal of pathology informatics*, 6(5):22–28, 2015.
- [50] M. Macenko, M. Niethammer, J. Marron, D. Borland, J. Woosley, X. Guan, C. Schmitt, and C. Thomas. A Method for Normalizing Histology Slides for Quantitative Analysis. In *Proceedings of the 2009 IEEE International Symposium on Biomedical Imaging*, pages 1107–1110, 2009.
- [51] D. Magee, D. Treanor, D. Crellin, M. Shires, K. Smith, K. Mohee, and P. Quirke. Colour normalisation in digital histopathology images. In *Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*, volume 100, pages 100–111, 2009.
- [52] P. Malm, A. Brun, and E. Bengtsson. Simulation of bright-field microscopy images depicting pap-smear specimen. *Cytometry Part A*, 87(3):212–226, 2015.
- [53] M. McCann, J. Ozolek, C. C.A., B. Parvin, and J. Kovacevic. Automated Histology Analysis: Opportunities for signal processing. *IEEE Signal Processing Magazine*, 32, No. 1:78–87, 2015.
- [54] M. Mengel, R. von Wasielewski, B. Wiese, T. Rüdiger, H. K. Müller-Hermelink, and H. Kreipe. Inter-laboratory and inter-observer reproducibility of immunohistochemical assessment of the Ki-67 labelling index in a large multi-centre trial. *The Journal of pathology*, 198(3):292–299, 2002.
- [55] J. S. Meyer, C. Alvarez, C. Milikowski, N. Olson, I. Russo, J. Russo, A. Glass, B. A. Zehnbauer, K. Lister, and R. Parwaresch. Breast carcinoma malignancy grading by bloom–richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. *Modern pathology*, 18(8):1067–1078, 2005.

- [56] Z. Mohammed, D. McMillan, B. Elsberger, J. Going, C. Orange, E. Mallon, J. Doughty, and J. Edwards. Comparison of visual and automated assessment of ki-67 proliferative activity and their impact on outcome in primary operable invasive ductal breast cancer. *British journal of cancer*, 106(2):383–388, 2012.
- [57] A. Mouelhi, M. Sayadi, F. Fnaiech, K. Mrad, and K. B. Romdhane. A new automatic image analysis method for assessing estrogen receptors’ status in breast tissue specimens. *Computers in biology and medicine*, 43(12):2263–2277, 2013.
- [58] T. W. Nattkemper, A. Saalbach, and T. Twellmann. Evaluation of multiparameter micrograph analysis with synthetical benchmark images. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 1, pages 667–670. IEEE, 2003.
- [59] M. Oger, P. Belhomme, J. Klossa, J.-J. Michels, and A. Elmoataz. Automated region of interest retrieval and classification using spectral analysis. *Diagnostic Pathology*, 3(1):1, 2008.
- [60] N. Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, January 1979.
- [61] S. Petushi, F. U. Garcia, M. M. Haber, C. Katsinis, and A. Tozeren. Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. *BMC medical imaging*, 6(1):1, 2006.
- [62] S. Petushi, C. Katsinis, C. Coward, F. Garcia, and A. Tozeren. Automated identification of microstructures on histology slides. In *IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2004*, pages 424–427. IEEE, 2004.
- [63] H. M. Reynolds, S. Williams, A. M. Zhang, C. S. Ong, D. Rawlinson, R. Chakravorty, C. Mitchell, and A. Haworth. Cell density in prostate histopathology images as a measure of tumor distribution. In *SPIE Medical Imaging*, pages 90410S–90410S. International Society for Optics and Photonics, 2014.
- [64] B. Riedelsheimer and S. Büchl-Zimmermann. *Romeis - Mikroskopische Technik*, chapter Färbungen, pages 171–282. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.
- [65] Q. Romero, P.-O. Bendahl, M. Fernö, D. Grabau, and S. Borgquist. A novel model for Ki67 assessment in breast cancer. *Diagnostic Pathology*, 9(1):1–8, 2014.
- [66] M. A. Roula, A. Bouridane, and F. Kurugollu. An evolutionary snake algorithm for the segmentation of nuclei in histopathological images. In *IEEE International Conference on Image Processing. 2004*, volume 1, pages 127–130. IEEE, 2004.
- [67] A. Ruifrok and D. Johnston. Quantification of Histochemical Staining by Color Deconvolution. *Analyt Quant Cytol Histol*, 23:291–299, 2001.

- [68] P. Ruusuvuori, A. Lehmussola, J. Selinummi, T. Rajala, H. Huttunen, and O. Yli-Harja. Benchmark set of synthetic images for validating cell image analysis algorithms. In *16th European Signal Processing Conference*, pages 1–5. IEEE, 2008.
- [69] B. Sabata. Digital pathology imaging—the next frontier in medical imaging. In *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 1–6. IEEE, 2012.
- [70] V. S. Sacchini and A. M. Pluchinotta. Staging and Workup of Invasive Breast Cancer. In *The Outpatient Breast Clinic*, pages 293–314. Springer, 2015.
- [71] O. Sertel. *Image analysis for computer-aided histopathology*. PhD thesis, The Ohio State University, 2010.
- [72] O. Sertel, U. V. Catalyurek, H. Shimada, and M. N. Gurcan. Computer-aided prognosis of neuroblastoma: Detection of mitosis and karyorrhexis cells in digitized histological images. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1433–1436. IEEE, 2009.
- [73] J. Shu, H. Fu, G. Qiu, P. Kaye, and M. Ilyas. Segmenting overlapping cell nuclei in digital histopathology images. In *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5445–5448. IEEE, 2013.
- [74] K. Sirinukunwattana, A. M. Khan, and N. M. Rajpoot. Cell words: Modelling the visual appearance of cells in histopathology images. *Computerized Medical Imaging and Graphics*, 42:16–24, 2015.
- [75] P. Soille. *Morphological Image Analysis: Principles and Applications*. Springer-Verlag, 1999.
- [76] Y. Sucaet and W. Waelput. *Digital Pathology*. Springer, 2014.
- [77] V.-T. Ta, O. L  zoray, A. Elmoataz, and S. Sch  upp. Graph-based tools for microscopic cellular image segmentation. *Pattern Recognition*, 42(6):1113–1125, 2009.
- [78] P. Tadrous. Digital stain separation for histological images. *Journal of microscopy*, 240(2):164–172, 2010.
- [79] H. Trihia, S. Murray, K. Price, R. D. Gelber, R. Golouh, A. Goldhirsch, A. S. Coates, J. Collins, M. Castiglione-Gertsch, and B. A. Gusterson. Ki-67 expression in breast carcinoma. *Cancer*, 97(5):1321–1331, 2003.
- [80] V. J. Tuominen, S. Ruotoistenmaki, A. Viitanen, M. Jumppanen, and J. Isola. ImmunoRatio: a publicly available web application for quantitative image analysis of estrogen receptor (ER), progesterone receptor (PR), and Ki-67. *Breast Cancer Res*, 12(4):R56, 2010.

- [81] C. M. van der Loos. Multiple immunoenzyme staining: methods and visualizations for the observation with spectral imaging. *Journal of Histochemistry & Cytochemistry*, 56(4):313–328, 2008.
- [82] P. Van Diest, E. Van Der Wall, and J. Baak. Prognostic value of proliferation in invasive breast cancer: a review. *Journal of clinical pathology*, 57(7):675–681, 2004.
- [83] Z. Varga, J. Diebold, C. Dommann-Scherrer, H. Frick, D. Kaup, A. Noske, E. Obermann, C. Ohlschlegel, B. Padberg, C. Rakozy, et al. How reliable is Ki-67 immunohistochemistry in grade 2 breast carcinomas? A QA study of the Swiss Working Group of Breast-and Gynecopathologists. *PLoS one*, 7(5):e37379, 2012.
- [84] M. Veta. Staining Nnmixing and Normalization. Matlab-Code.
- [85] M. Veta. *Breast Cancer Histopathology Image Analysis*. PhD thesis, Utrecht University, The Netherlands, 2014.
- [86] M. Veta, A. Huisman, M. A. Viergever, P. Van Diest, and J. P. W. Pluim. Marker-Controlled Watershed Segmentation of Nuclei in H&E Stain Breast Cancer Biopsy Images. pages 618–621, 2011.
- [87] M. Veta, R. Kornegoor, A. Huisman, A. H. Verschuur-Maes, M. A. Viergever, J. P. Pluim, and P. J. van Diest. Prognostic value of automatically extracted nuclear morphometric features in whole slide images of male breast cancer. *Modern Pathology*, 25(12):1559–1565, 2012.
- [88] M. Veta, J. P. W. Pluim, P. J. van Diest, and M. A. Viergever. Breast Cancer Histopathology Image Analysis: A Review. *IEEE Transactions on Biomedical Engineering*, 61(5):1400–1411, May 2014.
- [89] M. Veta, P. J. van Diest, R. Kornegoor, A. Huisman, M. A. Viergever, and J. P. Pluim. Automatic nuclei segmentation in H&E stained breast cancer histopathology images. *PLoS one*, 8(7):e70221, 2013.
- [90] C.-W. Wang. A Bayesian learning application to automated tumour segmentation for tissue microarray analysis. In *Machine Learning in Medical Imaging*, pages 100–107. Springer, 2010.
- [91] M. Weingant, H. M. Reynolds, A. Haworth, C. Mitchell, S. Williams, and M. D. DiFranco. Ensemble Prostate Tumor Classification in H&E Whole Slide Imaging via Stain Normalization and Cell Density Estimation. In L. Zhou, L. Wang, Q. Wang, and Y. Shi, editors, *Machine Learning in Medical Imaging*, volume 9352 of *Lecture Notes in Computer Science*, pages 280–287. Springer International Publishing, 2015.
- [92] S. Wienert, D. Heim, K. Saeger, A. Stenzinger, M. Beil, P. Hufnagl, M. Dietel, C. Denkert, and F. Klauschen. Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach. *Scientific reports*, 2(503):1–7, 2012.

- [93] Y. Wu and A. A. Sahin. Prognostic and Predictive Factors of Invasive Breast Cancer. In *Breast Disease*, pages 187–206. Springer, 2016.
- [94] F. Xing, H. Su, J. Neltner, and L. Yang. Automatic Ki-67 counting using robust cell detection and online dictionary learning. *IEEE Transactions on Biomedical Engineering*, 61(3):859–870, 2014.
- [95] F. Xing and L. Yang. Robust Nucleus/Cell Detection and Segmentation in Digital Pathology and Microscopy Images: A Comprehensive Review. *IEEE Reviews in Biomedical Engineering*, 2016.
- [96] Y. J. Zhang. A survey on evaluation methods for image segmentation. *Pattern recognition*, 29(8):1335–1346, 1996.
- [97] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. Tempany, M. R. Kaus, S. J. Haker, W. M. Wells, F. A. Jolesz, and R. Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index 1: Scientific reports. *Academic radiology*, 11(2):178–189, 2004.