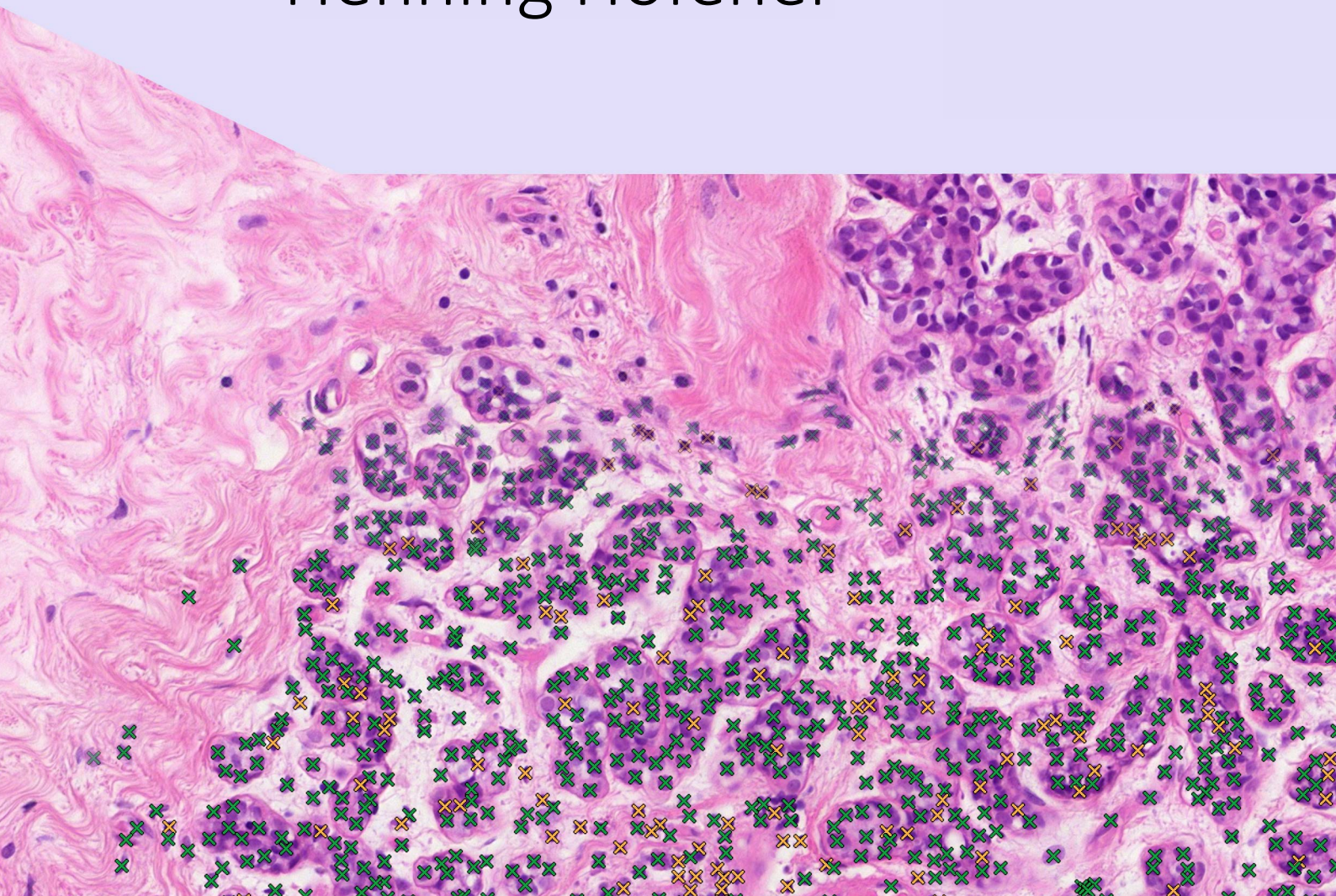
Automated Quantification of Cellular Structures in Histological Images

Henning Höfener



Automated Quantification of Cellular Structures in Histological Images

von

Henning Höfener

(geb. Kost)

Dissertation

zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften
- Dr.-Ing. -

vorgelegt im Fachbereich 3
Mathematik und Informatik
der Universität Bremen

Mai 2019

Gutachter:
Prof. Dr.-Ing. Horst Karl Hahn, Jacobs University Bremen
Prof. Dr. Ron Kikinis, Universität Bremen
Prof. Dr.-Ing. Tim Wilhelm Nattkemper, Universität Bielefeld

Datum des Promotionskolloquiums: 02. September 2019

This work was conducted at
Fraunhofer Institute for Digital Medicine MEVIS

Henning Höfener
Fraunhofer MEVIS
Am Fallturm 1, 28359 Bremen, Germany
Email: henning.hoefener@mevis.fraunhofer.de

Cover image by the Author, histological image kindly provided by Dr. Jeroen van der Laak, Radboudumc, Nijmegen, The Netherlands.

Abstract

Examination of tissue in pathology plays a central role in many diseases, including most cancers. They are essential for producing therapy decisions. Both qualitative and quantitative aspects have to be considered. Pathologists are remarkably good at conducting qualitative investigations, including finding and understanding different tissue patterns and textures. However, difficulties arise with quantitative examinations. These are mostly required for the assessment of cellular structures. Quantitative measures contain large inter- and intra-observer variability. This is caused in particular by the large visual variability of the histological tissue sections. Also, due to lacking time, pathologists are often forced to examine only small portions of the tissue sections and estimate quantitative measures rather than counting them.

Automated quantification of cellular structures using digitized histological tissue sections promises to improve accuracy, reproducibility and efficiency of quantitative assessments. However, these also face challenges that impede its wide use in practice. The large variability is a problem for automated analysis as well. Artifacts in the images represent a further obstacle to automated diagnosis. Additionally, cellular structures often form very dense clusters, making their detection as individual structures difficult.

This cumulative dissertation is based on four scientific publications. The aim is to bring the automated quantification of cellular structures closer to practical applicability. To this end, analyses will be developed that are optimized with regard to the challenges mentioned above and efficiency.

Training of machine learning algorithms will be optimized to be able to better deal with the large variability and artifacts in the images. Generation of training data sets is improved by extracting such training examples from reference data that contain the most valuable information for the algorithm. Additionally, for deep neural networks, the most important training parameters are identified and optimized. For quantification tasks in research settings, interactive training methods are proposed that allow quick adaption of an analysis to new data.

The examination of clusters of structures is improved by using probability or proximity map approaches. These enable detection and quantification within clusters, even if the individual structures cannot be identified unambiguously. For clusters in which the structures lie too closely together for this method, a density-based approach is proposed to enable quantification nonetheless.

The efficiency of automated analysis is another focus of this work, as it is a critical factor for the practical applicability of the analyses. The approaches described in this thesis achieve high efficiency by keeping the complexity of machine learning models low or even actively reducing them through training set optimization.

Zusammenfassung

Die Untersuchung von Gewebe in der Pathologie spielt bei vielen Krankheiten, darunter den meisten Krebserkrankungen, eine zentrale Rolle. Sie ist essentiell, um Therapieentscheidungen zu treffen. Dabei müssen sowohl qualitative als auch quantitative Aspekte berücksichtigt werden. Pathologen sind bemerkenswert gut darin sind, qualitative Untersuchungen durchzuführen, unter anderem im Auffinden und Verstehen verschiedener Gewebemuster und Texturen. Schwierigkeiten entstehen jedoch bei quantitativen Untersuchungen. Diese sind zumeist für die Untersuchung zellulärer Strukturen erforderlich. Quantitative Maße weisen eine große Inter- und Intra-Observer-Variabilität auf. Dies wird insbesondere durch die große visuelle Variabilität der histologischen Gewebeschnitte verursacht. Außerdem sind Pathologen häufig aus Zeitmangel gezwungen, lediglich kleine Teile der Gewebeschnitte zu untersuchen und quantitative Maße zu schätzen anstatt durch Zählen zu bestimmen.

Die automatisierte Quantifizierung zellulärer Strukturen auf Grundlage digitalisierter histologischer Gewebeschnitte verspricht eine Verbesserung von Genauigkeit, Reproduzierbarkeit und Effizienz quantitativer Untersuchungen. Allerdings steht diese auch vor Herausforderungen, die ihre Verbreitung in der Praxis verhindern. Die große Variabilität ist auch für die automatisierte Analyse ein Problem. Artefakte, die sich in den Bildern befinden, stellen eine weiteres Hindernis für die automatisierte Befundung dar. Darüber hinaus bilden zelluläre Strukturen oft sehr dichte Cluster, sodass die Erkennung der einzelnen Strukturen erschwert wird.

Die vorliegende kumulative Dissertation basiert auf vier wissenschaftlichen Publikationen. Ziel ist es, die automatisierte Quantifizierung zellulärer Strukturen der praktischen Anwendbarkeit näher zu bringen. Zu diesem Zweck werden Analysen entwickelt, die im Hinblick auf die genannten Herausforderungen und die Effizienz optimiert sind.

Um besser mit der großen Variabilität und den Artefakten in den Bildern umgehen zu können, wird das Training von Algorithmen des maschinellen Lernens optimiert. Die Erzeugung von Trainingsdatensätzen wird verbessert, indem solche Trainingsbeispiele aus Referenzdaten extrahiert werden, die für den Algorithmus die wertvollsten Informationen beinhalten. Zusätzlich werden für tiefe neuronale Netze die wichtigsten Trainingsparameter identifiziert und optimiert. Für Quantifizierungsaufgaben aus dem Forschungsumfeld werden interaktive Trainingsverfahren vorgeschlagen, die es ermöglichen, eine Analyse mithilfe neuer Daten schnell anzupassen.

Die Untersuchung von Clustern von Strukturen wird durch den Einsatz von Probability- oder Proximity-Map-Ansätzen verbessert. Diese ermöglichen die

Erkennung und Quantifizierung innerhalb von Clustern, auch wenn die einzelnen Strukturen nicht eindeutig voneinander abgegrenzt werden können. Für Cluster, in denen die Strukturen auch für dieses Verfahren zu eng zusammenliegen sind, wird ein dichtebasierter Ansatz vorgeschlagen, um eine Quantifizierung dennoch zu ermöglichen.

Die Effizienz automatisierter Analysen ist ein weiterer Schwerpunkt dieser Arbeit, da diese ein entscheidender Faktor für die praktische Anwendbarkeit der Analysen ist. In den Ansätzen, die in der vorliegenden Arbeit beschrieben werden, wird hohe Effizienz erreicht, indem die Komplexität von Modellen des maschinellen Lernens gering gehalten oder durch eine Optimierung der Trainingssets sogar aktiv reduziert wird.

Acknowledgements

At this point I would like to thank those people who accompanied and supported me during the past years.

First of all, I would like to thank my advisors Horst Hahn and Ron Kikinis, and also my mentor Andrea Schenk for their continuous support and encouragement. Thank you for letting me be part of Fraunhofer MEVIS and for providing me the opportunity to find my own research path. Also, I would like to express my gratitude to Tim Nattkemper for being an examiner of the thesis.

Special thanks go to André Homeyer, an expert and enthusiast in the field of computational pathology. I am very grateful for all the help, the technical and scientific advice I received and for all the useful comments and thoughts on my paper manuscripts. Thank you for being a mentor, colleague and friend to me.

I am grateful for all colleagues at Fraunhofer MEVIS, making this an awesome place to work. The friendly and respectful atmosphere, the dedication for the topics and curiosity for new paths are an invaluable inspiration for me. Thank you also for those senseful and senseless discussions during lunch and coffee breaks, which I really enjoy. Especially, I would like to thank the Histo-Team Nick Weiss, Johannes Lotz, Lars Ole Schwen, Ruben Stein and Tim-Rasmus Kiehl. It is a pleasure to work with you on all the different projects to advance computational pathology.

I would like to thank my family, especially my parents, who have been there for me with support and advice whenever I needed it. And I would like to express my love and gratitude to you, Nadia, my dear wife, for your endless loving support.

Cumulative Thesis

This thesis is based on the following papers:

Paper 1. Henning Kost, André Homeyer, Peter Bult, Maschenka C.A. Balkenhol, Jeroen A.W.M. van der Laak, Horst K. Hahn. **A generic nuclei detection method for histopathological breast images**. Medical Imaging 2016: Digital Pathology. Presented at the SPIE Medical Imaging, p. 97911E. 2016.

Paper 2. Henning Kost, André Homeyer, Jesper Molin, Claes F. Lundström, Horst K. Hahn. **Training nuclei detection algorithms with simple annotations**. Journal of Pathology Informatics 8, 21. 2017.

Paper 3. Henning Höfener, André Homeyer, Nick Weiss, Jepser Molin, Claes F. Lundström, Horst K. Hahn. **Deep learning nuclei detection: A simple approach can deliver state-of-the-art results**. Computerized Medical Imaging and Graphics 70, 43–52. 2018.

Paper 4. Henning Höfener, André Homeyer, Mareike Förster, Hans-Ulrich Schildhaus, Horst K. Hahn. **Automated density-based counting of FISH amplification signals for HER2 status assessment**. Computer Methods and Programs in Biomedicine 173, 77–85. 2019.

Author Contributions

For each of the above publications, Henning Höfener was the main contributor to the development of the evaluated algorithms, the design and execution of the study, and the writing of the manuscript.

Table of Contents

1 Introduction

Under the microscope, the fascinating micro-cosmos of the human body becomes visible. Pathologists can examine the complex structures, interactions and also abnormalities in cells and tissue. This world of small structures is also the world of large numbers. Our body consists of an estimated 37.2 trillion (37,200,000,000,000) cells, as reported by Bianconi et al. (2013). They also assume a volume of the human body of $65.22l$, which results in an average of about half a million cells per mm^3. A single histological slide obtained from a typical core biopsy with a needle diameter of about $2mm$ produces approximately $13mm^2$ of examinable tissue on histology slides (Helbich et al., 1998), containing tens of thousands of cells. And each one of them may carry important information for the pathologist's assessment.

1.1 Pathology

The word pathology consists of its components *pathos* and *logos*, and translates to the *study of suffering* (Kumar et al., 2007). Pathology as a medical field involves the investigation of the causes and mechanisms of disease. In clinical practice, pathology assessments are carried out to identify changes of cells, tissue and body fluids (Kumar et al., 2007). Histopathology describes pathology assessments that involve the microscopic examination of tissue (Collins Dictionaries, 2014).

To perform a histological examination, a tissue sample is prepared and a thin slice of it is placed on a glass slide. The tissue can then be stained in different ways to highlight different aspects of it, ranging from high-level structures like tissue types and tissue components up to very fine structures like certain protein expressions or single genes in cell nuclei. The tissue preparation process is explained in more detail in Section 2.1. Pathologists examine the stained tissue sections by searching for abnormal changes or assessing certain tissue features.

To render a diagnosis and to guide therapy, pathology is of utmost importance. The root causes of many diseases are only revealed under the microscope. Pathology is vital for clinical diagnosis as it allows precise examination of tissue and cells, their behavior and characteristics and can often provide more thorough insights

than other medical disciplines. This can be seen by the fact that it is considered to be the gold standard for the diagnosis of numerous diseases, including almost the entire range of cancers (Rubin et al., 2011). Pathology also plays an important role in research as it helps to better understand the mechanics of the body and of diseases, which in turn is essential for drug development.

1.2 Quantification of cellular structures

Histological features can be divided into qualitative and quantitative features. Pathologists usually assess features of both types during their examinations. Qualitative assessments have categorical results, often only binary. Examples in tumor diagnosis are the presence of isolated tumor cells or micrometastases in regional lymph nodes (de Boer et al., 2009). Also, when a tumor is surgically resected, pathologists need to assess whether the resection margin is free of tumor cells, indicating that the tumor has been completely removed from the patient (Emmadi and Wiley, 2012).

The assessment of the type of a breast tumor is another example for qualitative measures. Pathologists need to determine whether a lesion is an atypical but benign growth of the epithelial tissue, like a flat epithelial atypia or an atypical ductal / lobular hyperplasia, or whether it is a malignant carcinoma (Tot, 2010). In that case, they need to further discover whether the tumor cells only reside inside ductal or lobular areas of the tissue (ductal / lobular carcinoma in situ) or infiltrate the surrounding tissue areas (invasive ductal / lobular carcinoma).

Grading of prostate cancer is based on qualitative assessments as well. For grading according to the Gleason grading system (Epstein et al., 2016; Gleason, 1966), the glandular architecture of the tissue is examined. The distinction between different histological patterns is made based on textual descriptions and schematic drawings.

In contrast to qualitative assessments, quantitative assessments result in numeric values. These values are typically sizes, counts, or densities.

Quantification of cellular structures, especially nuclei, is the most prominent example of quantitative assessments. Numerous biomarkers are based on the quantification of nuclei of certain type or such that exhibit certain characteristics. Using the most frequently applied staining hematoxylin and eosin (H&E), pathologists can differentiate between nuclei and other structures and some cell types can be distinguished. To distinguish more types and subtypes of cells and to reveal further characteristics of nuclei, specific stainings can be applied. A widely used technique for this is immunohistochemistry (IHC), which allows highlighting certain proteins. Nuclei characteristics of interest are often associated with the expression of certain proteins. Thus, IHC can highlight these nuclei in one color, while others are stained with another color.

An example for nuclei quantification using H&E staining is the assessment of epithelial cells and lymphocytes. By doing so, lymphocytic infiltrations of tumors can be detected and quantified. Such infiltrations have been shown to correlate with the disease-free survival and outcome for some tumor types (Alexe et al., 2007). Also, those nuclei can be quantified that are currently in the mitosis stage of the cell cycle. Cells in this stage are recognizable as they appear as characteristic

mitotic figures in H&E stained tissue. Quantifying them is one way to determine the rate of tissue reproduction, which is an important feature for most tumors (Yigit et al., 2013). It is commonly used when staging breast cancer (Elston and Ellis, 1991).

Examples of nuclei quantification in IHC stainings are Ki-67 or progesterone and estrogen receptors (PR and ER). The Ki-67 protein is a cellular marker for cell proliferation. It is expressed during all phases but the resting phase of the cell cycle (Scholzen and Gerdes, 2000). For breast cancer patients it has been shown that Ki-67 is a prognostic parameter for disease-free survival and overall survival (Inwald et al., 2013) The Ki-67 index is measured as the fraction of Ki-67 positive nuclei in a tissue sample and indicates its reproduction rate, similar to mitosis counting.

The progesterone and estrogen receptors are protein groups that are activated by the corresponding hormones progesterone and estrogen. When activated by the hormone, the receptor-hormone complex translocates into the nucleus and regulates different aspects of the cell like the production of certain proteins (Yaşar et al., 2016). For breast cancer, both ER and PR status are predictive for the outcome of hormone treatments (Nardelli et al., 1986). Expression of ER and PR is measured as the ratio of tumor cells that express the receptor to the overall number of tumor cells.

Besides highlighting nuclei, there are also IHC stainings that bind to the cell cytoplasm or membrane. The human epidermal growth factor receptor 2 (HER2) is an example of an antigen expressed at the membrane of cells. HER2 expression is a biomarker that is both prognostic and predictive especially for breast cancer: It is associated with increased relapse and decreased overall survival (Slamon et al., 1987) but is also used for the decision for an antibody therapy (Slamon et al., 2001). When applying HER2 staining usually the nuclei are additionally stained with a different color. This is necessary because, just like ER and PR, HER2 expression is measured as the ratio of tumor cells that express the receptor to the overall number of tumor cells, as can be seen in the ASCO/CAP guidelines (Wolff et al., 2018). However, the guidelines also show that to report HER2 expression, the ratio and the staining intensity are combined to generate a semi-quantitative measure with possible values being 0 (negative), 1+ (negative), 2+ (equivocal) and 3+ (positive).

Although the analysis of cell nuclei is the most common use case, other cellular structures exist that can be assessed quantitatively. Fluorescence *in situ* hybridization (FISH) is a technique to highlight certain parts of the DNA within nuclei. In some tumors, mutations of the cells cause certain gene sequences to be copied several times, which is called amplification. An example is the ERBB2 gene in some breast cancers, which resides on chromosome 17. Amplification of ERBB2 is very closely connected to the expression of HER2 as described earlier. ERBB2 amplification is measured as the ratio of the number of ERBB2 copies to the number of copies of chromosome 17. This ratio is averaged over at least 20 tumor cell nuclei (Wolff et al., 2018).

Another example for quantitative assessments in histopathology is the analysis of macrovesicular steatosis in liver biopsies. Steatosis is quantified by measuring the area of fat droplets in liver tissue or by counting the number of steatotic hepatocytes. Assessing the amount of hepatic steatosis is an important step to

diagnose fatty liver disease or to decide whether a donor liver can be considered suitable for transplantation (Goceri et al., 2016).

For several pathology examinations including breast or colorectal carcinomas (Compton, 2000; Elston and Ellis, 1991), gland (or tubule) formation is taken into account. It is reported as the proportion of tumor that forms glands and thus resembles the behavior of healthy tissue. As there is a gradual transition between a glandular structure and a structure that is no longer a gland, it is important to have an objective cut-off value to define what degree of morphological change is accepted in order to still identify glands as such.

1.3 Automated quantification in pathology

Pathologists are remarkably good at finding relevant patterns and morphologies in tissue. They are very well trained and experienced in matching complex visual structures with diseases and understanding the processes in a given tissue sample. Difficulties arise, however, when it comes to quantitative assessments. Although quantification is objective and reproducible in theory, large inter- and intra-observer variability occurs in practice. They are caused by several challenges that pathologists face. First, many structures have a large variety in which they can occur. For cells, this might be for example staining intensity of their nuclei, membranes or cytoplasm. When counting positively or negatively stained cells, the pathologist has to find a staining threshold, that is, the minimum degree of positive staining that is required to consider a cell to be positive. This threshold needs to be consistent throughout the whole assessment and even over the course of multiple assessments in order to receive comparable results. Secondly, tissue features are often distributed heterogeneously, especially in tumors (Hamilton et al., 2014; Heppner, 1984). When a quantitative feature is assessed, taking the whole slide into account is usually not feasible because of limited time of the pathologist. Consequently, the result of the assessment depends on the examined regions. Although there are often rules which regions are to be considered, a certain amount of subjectivity remains (Gudlaugsson et al., 2012). Thirdly, as another consequence of lacking time, pathologists tend to estimate quantitative measures by eyeballing. Since humans are not good at estimating numbers or ratios, such results are poorly reproducible even for well-trained pathologists (Tang et al., 2012). Finally, quantitative assessments are partially reported in a semi-quantitative manner as seen in Section 1.2: Instead of continuous numbers, discrete measures with few possible values are used to describe tissue properties. This is frequently found for the assessment of IHC stained tissue. Using such semi-quantitative measures impedes both comparability and finding of correlations because a rather large span of tissue characteristics can lead to the same values.

Automated image analysis is one possible answer to these challenges. Assessing quantitative tissue features using image processing promises to increase reproducibility dramatically. Positivity-thresholds do not change over time for a once trained algorithm. Also, more robust biomarkers can be derived by assessing much larger areas of tissue. Automated analysis can also reduce the workload of pathologists, especially for simple routine tasks. This could provide pathologists with more time to concentrate on difficult cases.

Automated image analysis for the quantification of cellular structures is an active field that has been researched for many years. Already in the mid-1950s, first algorithms have been developed to automate nuclei detection in histological images (Meijering, 2012). These algorithms use intensity thresholds and intensity-derived features in conjunction with static decision rules. Due to the variability in the images, those algorithms had only limited success. Later, more sophisticated features for nuclei detection have been used and machine learning algorithms have been applied to create more complex and flexible rule sets (Arteta et al., 2012; Kårsnäs et al., 2011; Vink et al., 2013). With the use of machine learning, decision rules and thresholds do not need to be explicitly coded into the algorithm. Instead, the algorithms are able to be trained with example data to optimize those values themselves. Recent nuclei detection algorithms mainly use convolutional neural networks (CNN) (Litjens et al., 2017). With CNNs, not only the thresholds and decision rules but also the features are automatically learned from training data.

Besides for nuclei, for many other quantitative assessments automated approaches have been developed, including the examples given in Section 1.2. To quantify macrovesicular steatosis in liver biopsies, automated approaches identify and measure fat droplets in histological images, which appear as empty blobs. The main challenge is the discrimination between droplets and other bright structures like vessels and tissue cracks, which can be performed using machine learning (Homeyer et al., 2017).

The automated segmentation of glands or tubules enables the extraction of fully quantitative measures of the tubule architecture. Especially in carcinomas, variability of shape and size of these structures is large, making automated segmentation challenging. Most state-of-the-art algorithms employ CNNs for this task (Sirinukunwattana et al., 2017).

Besides the advantages associated with automated quantification, there are also challenges. These challenges are illustrated in Figure 1.1. Presumable the most important challenge is the plethora of variability in histological images (cf. Figure 1.1, left column). Caused by varying tissue processing and lab conditions as well as by pathological abnormalities, the appearance of tissue can change dramatically (McCann et al., 2015). While pathologists are used to these variations and sometimes do not even consciously perceive them, algorithms need to incorporate measures to cope with them. Additionally, histological images virtually always contain artifacts (cf. Figure 1.1, center column). Folds and cracks in the tissue, stain deposits, foreign bodies or air bubbles are just a few common defects histological slides may contain (Taqi et al., 2018). Such artifacts need to be sorted out by automated quantification methods. Specifically for the quantification of cellular structures, clustering is another important challenge (cf. Figure 1.1, right column). Especially in tumor areas, structures like cell nuclei often form dense clusters, impeding accurate individual delineation of them. Automated analysis needs to incorporate measures to robustly detect and quantify such structures as well.

To be applicable in practice, automated analysis also needs to be runtime efficient. For automated analysis, histological slides are digitized in high resolution using automated microscopes. If the whole glass slide is digitized (Whole slide image, WSI), this results in gigapixel sized images and thus large amounts of data to be processed. Additionally, the amount of cases in pathology departments is huge.

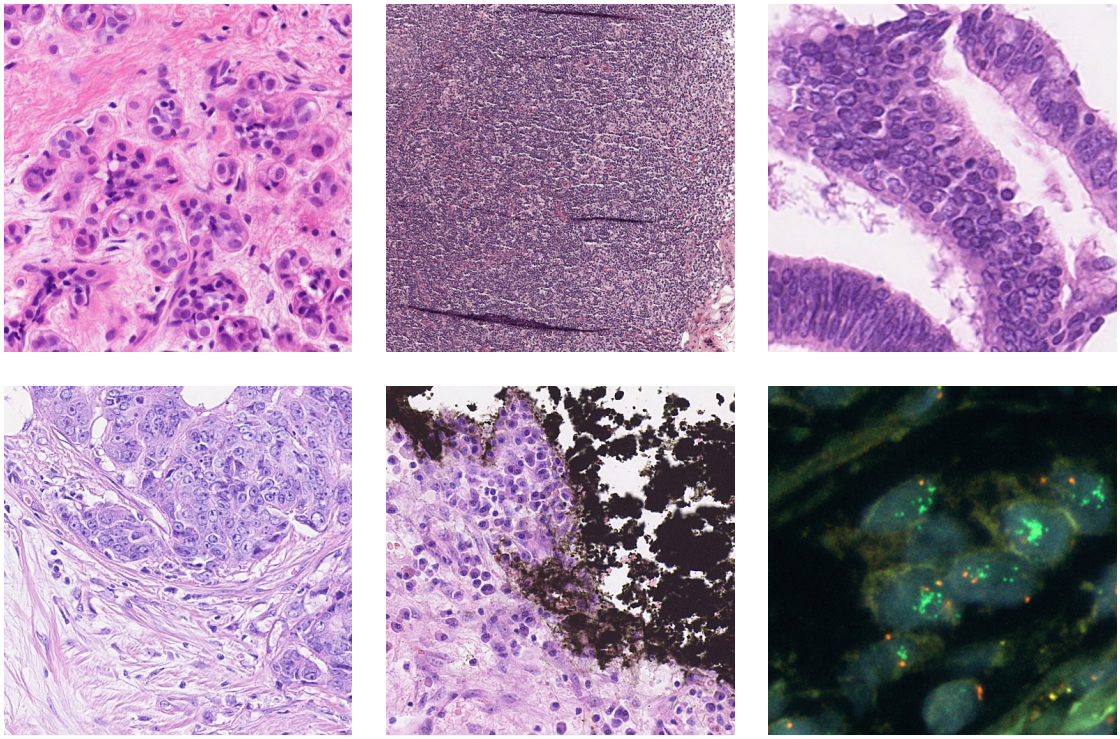


Figure 1.1: Challenges in histological images. Left column illustrates variability. Both images are stained with H&E but have a very different appearance. Center column shows artifacts, being tissue folding (top) and stain deposit (bottom). Right column shows clustering of nuclei (top) and FISH signals (bottom). Top left image kindly provided by Dr. Jeroen van der Laak, Radboudumc, Nijmegen, The Netherlands; Top center image from Litjens et al. (2018); Top right, bottom left and center images from Sirinukunwattana et al. (2016); Bottom right image kindly provided by Dr. Norbert Drieschner, ZytoVision GmbH, Bremerhaven, Germany.

Therefore, algorithms need to be fast in order to reduce pathologists' workload, especially if larger areas of the slides should be assessed to gain robust measures and to capture heterogeneity. Figure 1.2 illustrates the size of histological images and the amount of cellular structures to be quantified.

Automated algorithms are not only relevant in routine pathology, but can also increase quality and efficiency of assessments in pathology research settings. There, quantification tasks change frequently, as for example novel tissue preparation procedures are evaluated. In contrast to human observers that are often able to instantly get accustomed to new appearances of tissue, automated analyses need to be adapted or retrained to be able to be used advantageously.

1.4 Outline of the thesis

Automated quantification of cellular structures in histopathology has great advantages over manual assessment. It has the potential to both increase reproducibility and robustness of the quantification dramatically. However, variability within and between images, artifacts and the formation of clusters of the structures impede automated approaches. Also, for practical usability, automated analysis needs to be highly efficient.

This work is based on a collection of four publications addressing these challenges in the context of the detection and quantification of cell nuclei and fluorescence

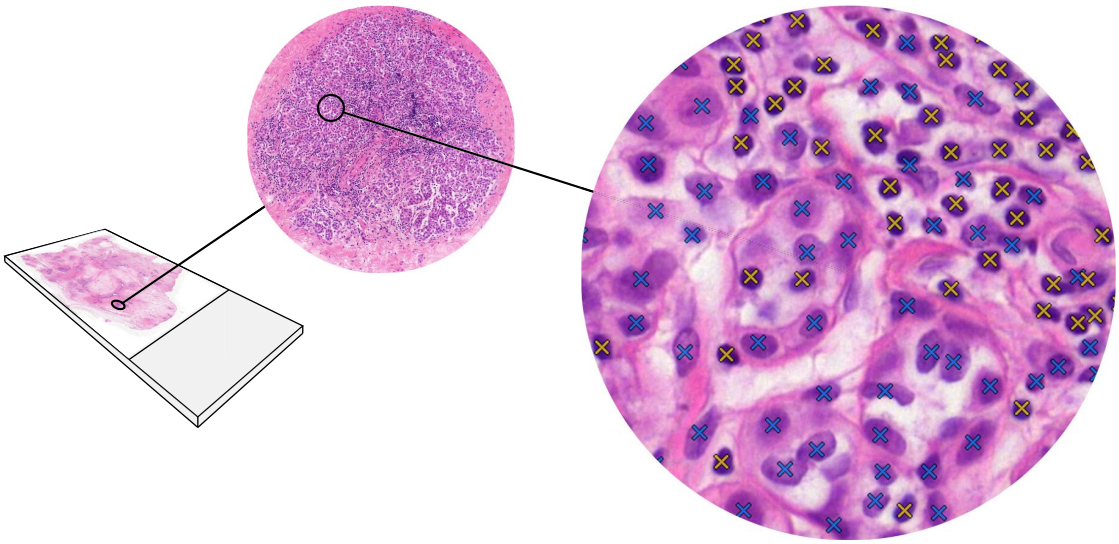


Figure 1.2: Illustration of size and resolution of a digitized histological slides. Left side shows visualization of tissue on the glass slide. Center and right side show magnified versions of circulated regions. Right side additionally visualizes automated analysis result where epithelial and lymphocytic nuclei have been detected. WSI kindly provided by Dr. Jeroen van der Laak, Radboudumc, Nijmegen, The Netherlands.

signals. Different machine learning algorithms are used and training procedures are optimized to manage variability and artifacts in images. So called PMap approaches and density-based quantification are applied that allow quantification of structures that form dense clusters. At the same time, complexity of the machine learning models is reduced to enable fast processing of large amounts of image data.

The remainder of this work is organized as follows. Chapter 2 briefly explains the workflow required to turn a tissue sample into a digitized histological image. Then, machine learning, Random Forests and convolutional neural networks are introduced. In chapter 3 the publications forming this cumulative thesis are summarized and brought into context. The results of the publications are discussed in chapter 4. Chapter 5 presents concluding thoughts and future prospects. Finally, the publications, upon which this thesis is built, are appended.

2
Background

2.1 From tissue to image

The path from a tissue sample to a digitized histological slide is very complex. In the following, the procedure required for light microscopy is briefly described. Figure 2.1 illustrates the most important steps of that process. The aim of this section is not only to give an overview of the steps involved but also to give an impression of the intricacy and the large number of sources for artifacts and variation in image quality. The following explanations are based on the works of McCann et al. (2015), Sucaet and Waelput (2014) and Suvarna et al. (2013).

2.1.1 Collecting

Tissue is mostly collected through biopsies. The size of a biopsy can vary substantially from fine needle aspiration (less than $1mm$ diameter) to the resection of a whole lesion or even of a whole organ.

2.1.2 Processing and embedding

Afterwards, the tissue is fixated to stop all biological activity, including biological decay or cell growth. There are many different fixatives for the different applications that all have their individual advantages and shortcomings. Some, for example, cause distortion of the tissue or alter molecules needed for certain stainings. Thus, care must be taken to choose a fixative and a fixation protocol that is appropriate for the type of tissue as well as the intended staining and application.

In case of larger tissue samples, the next step is the gross inspection and the excision of relevant small parts. Afterwards, the tissue is stabilized. This can also be done in various ways, but is most commonly performed chemically. Three steps are required for stabilization. First, the sample is dehydrated to remove unbound water and aqueous fixatives from the tissue. Then, a clearing agent is applied that replaces the dehydrating agent in the tissue. Third, the clearing agent is in turn

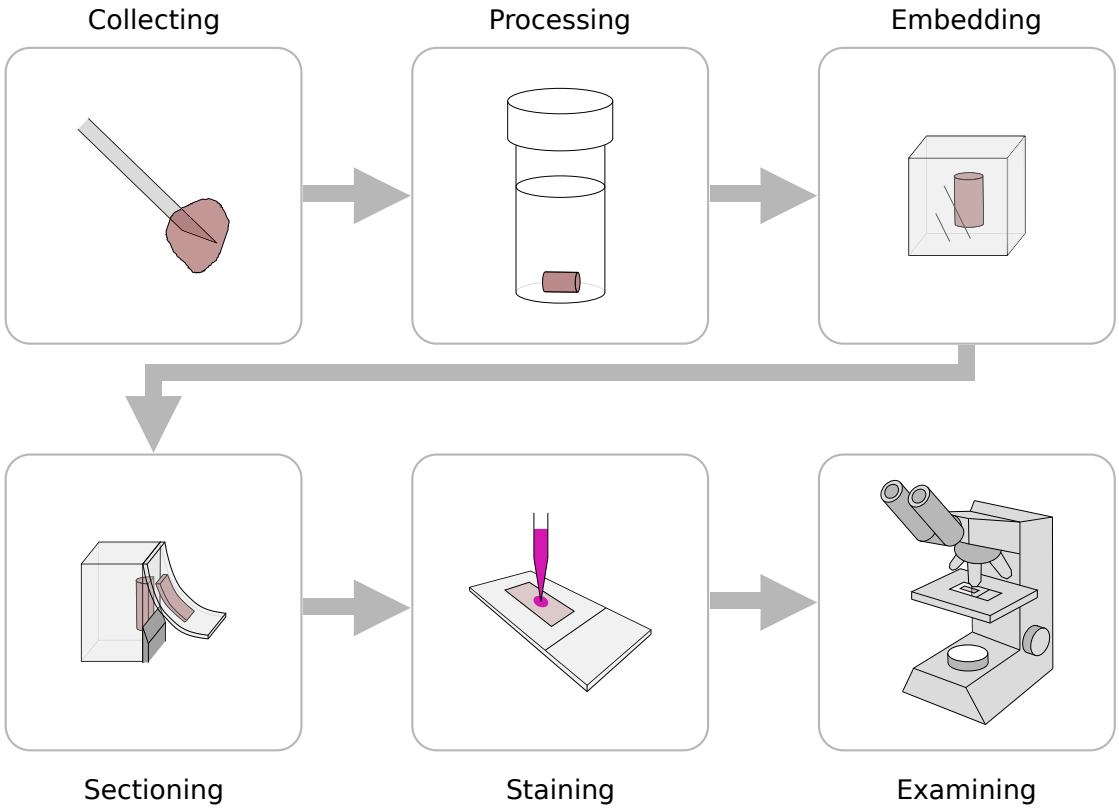


Figure 2.1: Workflow of tissue preparation. Tissue is collected, processed, embedded into paraffin, sectioned into slices, and stained. Afterwards, the resulting glass slides can be and examined under the microscope or digitized using an automated microscope.

replaced by an infiltration agent, usually paraffin wax. By the infiltration with paraffin the tissue is hardened and thus stabilized for further processing.

In the end, the stabilized sample is put into a cassette which is filled up with paraffin. This results in the sample being embedded in a solid paraffin block.

2.1.3 Sectioning

From the paraffin block containing the tissue very thin slices are cut. The thickness of the slices is usually 3–4 μm. Sectioning is performed with a special device called microtome, which has a similar principle of operation as a carpenter's plane. Manual, semi-automated and fully automated versions of microtomes exist. After cutting, the slices are put into a warm water bath where small wrinkles smooth out. The slices can then be placed on a microscope glass slide.

2.1.4 Staining

The inherent color of tissue is very faint, such that the thin tissue sections on the glass slide are almost transparent. To visualize its structures, the tissue is stained as a next step. The general mechanism of staining is that certain substances bind or have affinity to certain components of the tissue. Lots of stainings exist for various tissue components with different dyes, including fluorescent ones. Each staining has its unique procedure, which is in general a sequence of instructions that include

applying certain chemicals or exposing the tissue to a certain temperature for a given duration.

The most commonly used staining in histopathology is the Hematoxylin and Eosin staining (H&E). Hematoxylin is blue and binds to nucleic acids, while eosin is pink and binds to proteins. For most tissue types, H&E stains nuclei in blue and cytoplasm as well as the collagen fibers of connective tissue in different shades of pink.

immunohistochemistry (IHC) exploits the concept of antibodies binding to certain antigens (Ramos-Vara and Miller, 2014). In a specific staining procedure, the antibodies are applied to the tissue, binding to the antigens (proteins) of interest. Then, in a direct or indirect manner, dye is bound to the antibody. IHC stainings highlight certain proteins and are often used to tag nuclei that express those proteins. Chromogenes of different colors exist for IHC, such that it is possible to highlight a few antigens at the same time. Usually, the tissue is then additionally stained with hematoxylin, coloring the remaining nuclei in blue.

It is also possible to bind fluorescent dyes to the antibodies, making them visible under a fluorescence microscope. Immunofluorescence has the advantage of being able to highlight more antigens at the same time than IHC, using different fluorescent dyes. They can then be distinguished by using different fluorescent filters in the microscope. However, fluorescence microscopes are much less common in clinical practice.

Another staining technique is *in situ* hybridization (ISH). ISH probes bind to specific gene sequences in the tissue. Similar to IHC, dye is then bound to the probes, staining them for example for fluorescence microscopy (fluorescence ISH / FISH) (Langer-Safer et al., 1982).

After staining, a cover slip glass is put on the slide. This is done to protect the tissue and to improve the optical properties of the slide.

2.1.5 Examining

After staining and cover slipping, the glass slides are prepared to be examined by the pathologist. Examination is mostly performed using a manual light microscope, either brightfield or fluorescence. With the microscope, the pathologist can view the tissue with different objectives. For light microscopy, usually a magnification of up to 400× is used. Additionally, the pathologist is able to move the glass slide in X-Y-direction and also in Z-direction to adjust the focal plane.

Instead of examining the glass slide under the microscope directly, digital images of the tissue can be made. There are two common solutions to create digital images. The basic solution is a digital camera mounted to the microscope. Such a device allows the pathologist to quickly take images of the current field of view in order to archive it, append it to the pathology report or to run automated analysis on the image. The other solution are the whole slide scanners. These are fully automated microscopes that can scan and digitize entire slides. They capture parts of the slide tile-wise or line-wise and stitch the individual images together afterwards. With fully digitized glass slides, the pathologist can examine the tissue on the computer screen and directly annotate or measure structures, write the pathology report and also invoke automated analysis.

2.2 Machine learning

In this section, the term machine learning and some of its core components are described. Furthermore, the machine learning methods used in the papers of this theses are explained. This section is based on Bishop (2006), Nielsen (2015), Russell and Norvig (2009).

Machine learning is a part of artificial intelligence and describes the automated creation of knowledge from data. It usually consists of two separate phases, being a training or learning phase and a prediction or inference phase. The training phase comprises the recognition of patterns in data or the discovery of the rules underlying the data. The knowledge gained in that phase can be formalized as a machine learning model. In the prediction phase, the model can then be used to derive information about new data.

2.2.1 Forms of machine learning

Machine learning methods are commonly divided into supervised, unsupervised and reinforcement learning (Bishop, 2006; Russell and Norvig, 2009). The machine learning algorithms used in the papers of this thesis are all supervised. However, all three forms are briefly described in the following.

In supervised learning, a model is learned from pairs of input and output data. An example in the histology domain is to learn whether an image region contains tissue or not, given some basic features of that region, like the mean and standard deviation of the brightness. To train the machine learning algorithm, a set of training samples is required. Each sample consists of an input, called attribute vector or feature vector, as well as an output, called label or target value. In the example above, the feature vector has two values, being the mean and the standard deviation of the brightness. The output is a binary value, whether or not the image region contains tissue. During the training phase, the algorithm learns the mapping between feature vectors given in the training data and their target values. That means that it generates a model or function that describes the relationship between feature vectors and target values as accurately as possible. Afterwards, in the prediction phase, this model can be applied to feature vectors of new image regions to predict whether those contain tissue. Both input and output can comprise categorical and numerical values. In case of categorical output (categories or classes) the machine learning task is a classification task, whereas for numerical output it is a regression task.

In unsupervised learning, only the feature vectors are given to the algorithm during training. Instead of learning to reproduce given target values, the task is to discover relationships of the given data and patterns inside it. An example of unsupervised learning is clustering, where the aim is to discover groups within the data.

Reinforcement learning is a technique used for software agents that interact with a virtual environment. In supervised learning, training samples would consist of the current state of the agent and the environment as input and the optimal next step as target value. In reinforcement learning, the agent instead interacts with the environment for several steps and a reward is given to the algorithm depending on

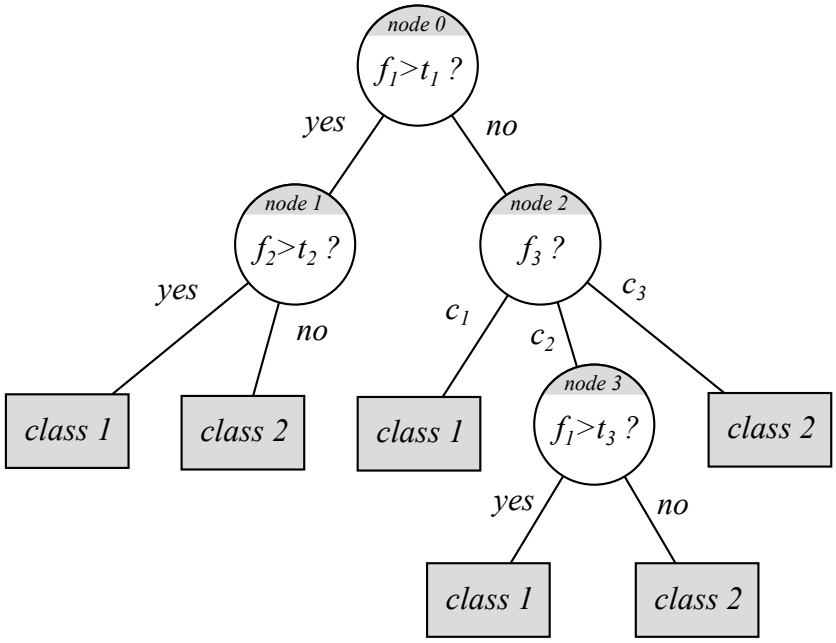


Figure 2.2: Decision tree for classification into *class 1* and *class 2*. Nodes 0, 1 and 3 distinguish whether the feature is above or below a certain threshold, node 2 distinguishes between three possible categorical values that the feature can hold. Leaf nodes are rectangular and show the class prediction associated with them.

the overall achievement of the interaction. Examples are agents that play games and are rewarded for scores or success.

2.2.2 Random Forests

A well-known type of supervised machine learning is decision tree learning. A decision tree is a tree graph with each non-leaf-node being associated with a certain decision criterion. That criterion is one of the training set's features. If that feature is numerical, the decision criterion also includes a threshold value for that feature. The number of possible decisions equals the number of possible values for categorical features, or two (below or above the threshold) for numerical features. For every non-leaf-node in a decision tree, there is one child node for each possible decision. The leaf nodes of the tree are associated with an output value.

To perform a prediction for a given feature vector, the tree is traversed starting from its root. At each node, the decision criterion associated with that node is applied to the given feature vector and the corresponding child node is visited. When a leaf node is reached, the output value associated with that leaf is returned as the prediction value of the decision tree. During traversal the same numerical feature can be assessed multiple times with different thresholds each time. Decision trees can be used both for classification and regression. Figure 2.2 shows an example decision tree for classification into two classes with three features and both categorical and numerical decision criteria.

During training, the decision tree is generated recursively using the training set. When creating a node, the associated decision criterion is generated such that it partitions the training set in the best way according to a given measure. The training set is then split according to the decision criterion. Child nodes are

afterwards generated recursively using the respective subsets of the training set. Whenever all samples in the training subset have the same output value, the building procedure of the respective sub-tree is stopped and the resulting leaf is associated with that output value. There are several metrics to assess the quality of a decision criterion with respect to a given training set. While information gain or Gini impurity are often used in classification tasks, typical options for regression are the mean absolute error or mean squared error. During training, nodes are added to the decision trees until the samples in a node all have the same output value. Therefore, having more training samples usually results in larger decision trees and results in increased run time in the prediction phase.

It has been shown that the use of multiple learners in a so called *ensemble* can improve the quality of the machine learning considerably (Opitz and Maclin, 1999). Breiman (2001) proposed to use multiple decision trees in an ensemble. The prediction of the ensemble is then generated by determining the majority class or by calculating the average of the output values for classification or regression, respectively. For classification, class probabilities can additionally be determined as the fraction of trees predicting that class. In his paper, Breiman specified several parameters settings for the decision trees in the ensemble. Amongst them are that only a random subset of the feature space is searched for the best decision criterion and the usage of bagging. Bagging creates a training set of arbitrary size from an underlying training set by selecting random samples with replacement. The use of an ensemble of decision trees and randomness for multiple aspects of their construction lead to the term Random Forest.

In some of the papers of this thesis, Random Forests are used to perform image analysis. For such tasks, it is hardly ever sufficient to use the red, green and blue color values of each pixel as features for a prediction. Instead, further features can be derived from the original image data. These can be produced for instance by using convolution filters, local binary patterns (Ojala et al., 2002) or histograms. As the choice of the features is a manual task, they are called hand-crafted.

2.2.3 Convolutional neural networks

Neural networks are another machine learning method and are inspired by biological neural networks (McCulloch and Pitts, 1943). They can be used for all of the aforementioned machine learning methods (supervised, unsupervised and reinforcement learning). However, the following section will be limited to their application in supervised learning, as all the approaches described in the papers of this work employ that learning method.

The base element of a neural network is the neuron, which has an inherent activation value. The neurons of a network are connected to each other such that every neuron has a set of incoming connections and outgoing connections. Additionally, a weight is associated with every incoming connection. To determine the current activation of a neuron, the weighted sum of the activations of the neurons at all incoming connections is calculated. The result is then fed into a non-linear activation function. For classical neural networks, a sigmoid function is used, usually the logistic function. To be able to shift the activation function to the left or right, an additional term is added to the weighted sum of input activations. This term is called bias and can also be understood as a further input to the neuron with a

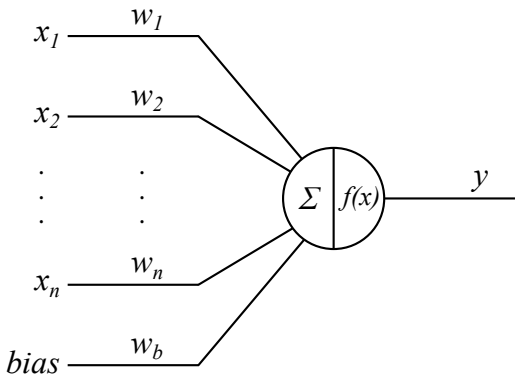


Figure 2.3: Schematic artificial neuron with inputs x and $bias$ and weights w. The neuron computes the weighted sum of the inputs, which is then fed to an activation function $f(x)$. The result is the neuron's own activation y.

constant value of 1 and an associated weight. A neuron is shown schematically in Figure 2.3.

Acyclic neural networks are called feed-forward networks, because activation, and thus information, is always propagated forward towards the end of the network. Networks with cycles are called recurrent networks. As those are not used in this work, they will not be discussed further here. Feed-forward networks are usually organized in layers (Minsky and Papert, 1969), as depicted in Figure 2.4. Neurons of one layer only have incoming connections from neurons of the previous and outgoing connections to neurons of the next layer. In classical neural networks, each neuron of one layer is connected to each neuron in the following layer, resulting in fully connected layers. The first layer of a network is called input layer and has no incoming connections from other neurons. Instead, the feature vector is fed to the input layer, with one neuron for each feature. The last layer is called output layer and has as many neurons as there are output values in the machine learning task. Neural networks can be applied to both classification and regression problems. For regression problems, the activation function used for the output neurons is typically the identity function. For classification problems, there is one output neuron for each class, and a softmax activation function is used. The output can then be interpreted as the class likelihoods. All intermediate layers are called hidden layers. Prediction in a feed-forward neural network is performed by feeding a feature vector to the input layer and calculate the activations of the neurons layer by layer. This is called forward propagation. For regression, the activation of the neurons of the output layer constitute the prediction. For classification, the neuron of the output layer that exhibits the highest activation is determined. The prediction is then the class associated with that neuron.

The behavior of a neural network is defined by its weights, including the biases. To train a network in a supervised manner with a given training set, the weights are adjusted in order to minimize the error (also called loss) between the network predictions and the actual output values of the training samples. Depending on the machine learning task at hand, different loss functions are used to evaluate that error. In classical feed-forward neural networks, the adjustment of the weights is mostly done by backpropagation (Rumelhart et al., 1986; Russell and Norvig, 2009). In essence, backpropagation consists of two steps. First, the gradient is evaluated as the derivative of the loss function with respect to the network weights. Secondly, the weights are adjusted using the derivative in order to decrease the

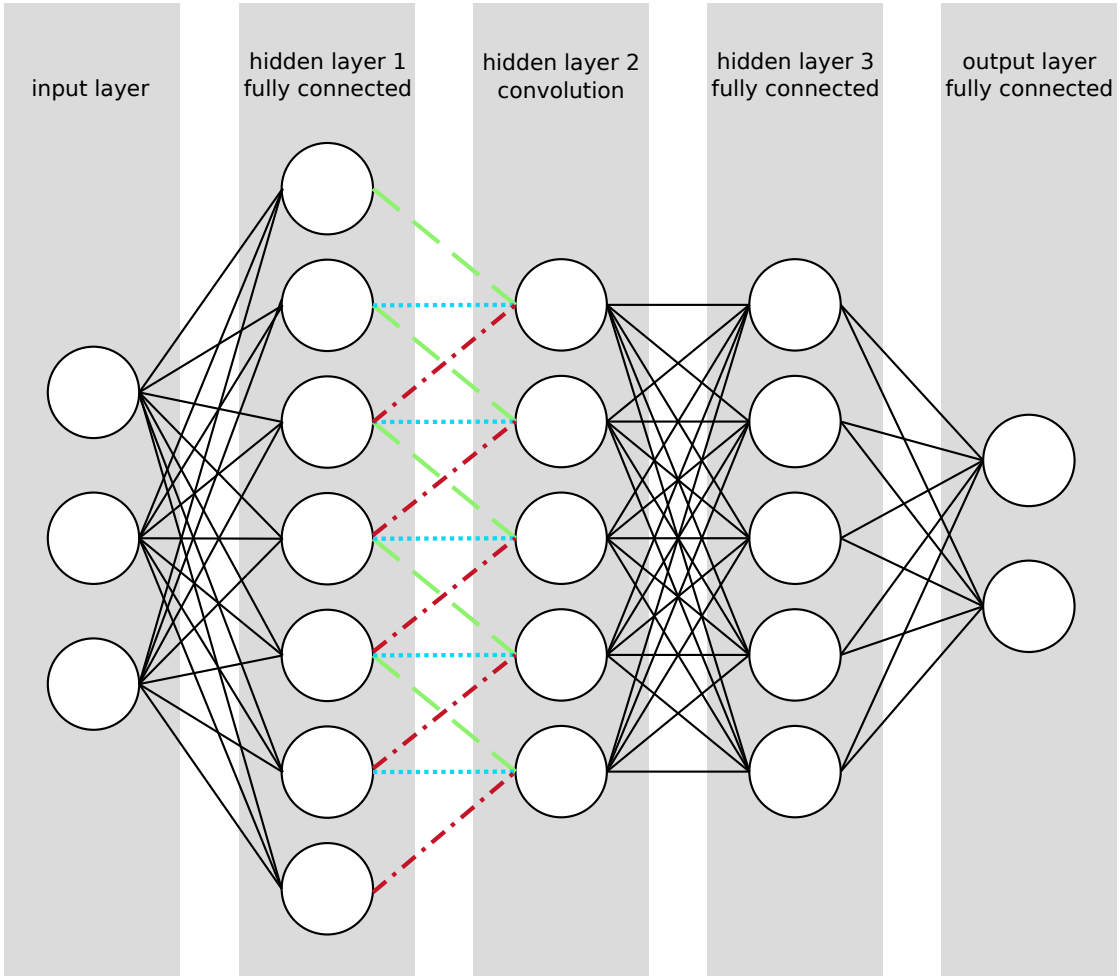


Figure 2.4: Example of a feed-forward neural network with three input and two output neurons and three hidden layers. Input data is propagated from left to right. All layers but the hidden layer 2 are fully connected to their predecessors. The hidden layer 2 is a convolutional layer where each neuron is only connected to the respective neuron of the previous layer and its two neighbors. Weights are shared between the green dashed connections, the blue dotted connections and the red dash-dotted connections.

loss. The simplest technique is gradient descent. The term backpropagation stems from the first step, where the loss is propagated backwards through the network to efficiently determine the derivative.

Neural networks with multiple hidden layers are called deep neural networks (Bengio, 2009; Schmidhuber, 2015). The application of deep neural networks is termed deep learning. Due to their layered structure, those networks can learn more complex functions even with the same amount of neurons. However, problems occur when training them. Most prevalent is the vanishing gradient problem (Hochreiter et al., 2001). Due to the sigmoid activation function of the neurons, gradients tend to become smaller and smaller when propagated backwards through the network. As a result, the weights of early layers get adjusted much slower than those of later layers, leading to a poorly trained network. Instead of vanishing gradients, under some circumstances the gradients can also increase dramatically (exploding gradient) or just be unstable. Another important problem of deep neural networks is the large amount of iterations required for training, making it computationally expensive.

In the recent years, important breakthroughs have been made to address the prob-

lems of deep neural networks. The vanishing gradient problem is reduced by using rectified linear functions (Glorot et al., 2011) as activation instead of sigmoid functions. The rectified linear function is defined as the positive part of its argument, which means that it equals the identity function for positive input values. For negative input values, it outputs zero. This has the advantage that on the positive side of the function the gradients are not saturated. Additionally, the gradients are very easy to compute. On the other hand, the rectified linear function is not zero-centered and the hard saturation of the gradient for negative values can lead to neurons being stuck in an inactive state. However, in the context of deep neural networks the advantages outweigh these problems and the rectified linear function became the most popular activation function in 2015 (LeCun et al., 2015).

The problem of high computational expense when training deep neural networks is addressed in multiple ways. Using convolutional layers (LeCun et al., 1989) instead of fully connected layers reduces the amount of adjustable weights dramatically. In convolutional layers, each neuron has an input connection from only a few neighbored neurons of the previous layer, for instance the 10th neuron being connected only to neuron 8th-12th of the previous layer. The weights of the neurons in convolutional layers are additionally shared, which decreases the amount of weights to be adjusted and increases the stability of the gradients. Since such layers perform mathematical convolution, they are called convolutional layers and networks using such layers are called convolutional neural networks (CNN). Although CNNs had been known for a long time, they only became established when their training and prediction could be highly parallelized and thus accelerated by using GPUs instead of CPUs (Oh and Jung, 2004).

Another approach to reduce the computational expense of deep neural networks is to decrease the amount of data between two layers. This can be implemented by striding, where only every n-th neuron of a given layer is connected to neurons of the next layer. This reduces the amount of data available for that next layer. Instead of completely disregarding the information of the neurons that are not connected further, pooling layers are often used as an intermediate layer. In pooling layers (Cireşan et al., 2011), the activation of the neurons is recomputed from the activations of the neighboring neurons, most commonly by determining the maximum or the average of them. Especially these improvements of deep neural networks led to their breakthrough in image analysis (Krizhevsky et al., 2012) and other machine learning applications (Schmidhuber, 2015).

When considering image analysis, a main difference between classical machine learning methods and CNNs is the way how features are extracted from the underlying data. Classical methods like Random Forests require hand-crafted features to be conceived and extracted. Those features are often derived from the images by applying convolutional filters. 2D-convolutional layers in CNNs are equivalent to applying such a filter. Instead of designing suitable filters manually, with CNNs, the convolutional filters are learned automatically. Thus, instead of hand-crafted features, raw image data can be fed to CNNs. To enable that, the layers of the CNN are 2-dimensional, performing for example 2D-Convolutions or 2D-Pooling.

The layered architecture of CNNs leads to the network learning increasingly abstract representations of the input data (LeCun et al., 2015). For image analysis, the features learned in the first layer typically represent small, local details, like points and edges. Building upon these, successive layers learn to represent more

abstract image features that take into account more context, like lines and simple shapes. This continues to layers learning to recognize complex image structures like faces, animals, or specific tissue characteristics in histological images. The last layers of the network can then use these abstract representations to perform the actual classification or regression task at hand.

The fact that features are extracted from the raw image data and the structure of propagation of information are the main advantages of CNNs over classical machine learning methods. The ability to learn hierarchical features with increasing abstraction is key for solving complex image analysis tasks. It was shown that for many applications, including histological image analysis, specifically designed CNNs achieve better accuracies than classical machine learning methods (cf. Section 1.3).

2.2.4 Overfitting

Real-world regression or classification problems are often very complex. In order to still generate a capable machine learning predictor, two paths can be taken. The first is to increase the complexity of the machine learning model, that is, the function mapping from input to output data is given more degrees of freedom. The second path is to increase the amount of information available for each training sample, by adding more features. For CNNs, where features are not explicitly extracted, this is comparable to increasing the number of output channels of the last convolutional layer. This way, the last layers of the network that perform the classification or regression are provided with more features. Both paths, however, can result in a model that too closely learned the specific characteristics of the training dataset instead of learning the general rules of the problem. Such a missing ability to generalize is called overfitting and can mostly be observed as low errors when predicting the training set but high errors when predicting unseen data.

The reason for overfitting can be explained by interpreting the training samples as points in the feature space. The number of features the samples have, is the number of dimensions of the feature space. Considering a binary classification task, in which positive and negative samples are to be distinguished, a model with many degrees of freedom can define a very complex shape in feature space separating positive from negative samples perfectly. Even in the presence of noise in the data, the model will learn to separate the samples and thus also learn the specific noise of the training set. Figure 2.5 visualizes overfitting of a complex model to noisy data.

If the training samples have many features, then the feature space becomes very high-dimensional. The volume of the feature space grows exponentially with the number of dimensions and quickly becomes very sparsely populated with training samples. This effect is called *curse of dimensionality* (Bellman, 1957) and makes it very difficult to create a model that accurately reflects the general rules of the machine learning problem. In order to compensate for this, the number of training data must increase dramatically as the number of features grows. Another issue that arises with the sparsity is that it becomes easy to find even a simple hyperplane that can separate positive from negative samples perfectly. This is called the *Hughes' phenomenon* (Hughes, 1968) and is illustrated in Figure 2.6.

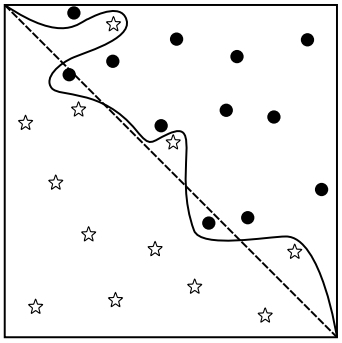


Figure 2.5: Noisy data set of samples with two features and two classes. A complex machine learning model resulted in the solid line, a simple one in the dashed line. While the former perfectly splits the samples in this dataset, the latter will produce better results when applied to yet unseen data. The complex model overfitted the training samples.

As stated above, overfitting is the lack of generalizability and results in low errors on the training set but high errors on unseen data. To assess the generalizability of a trained model, a separate test set is required. Just like the training set, the test set consists of samples with both input data and desired output values. The only difference is that the test set may not be used during training. This way, the quality of the predictions of the model on the yet unseen test set shows how well the model generalizes. It is important that no information of the test set may be used to improve the model. This also means that the results obtained with the test set may not lead to decisions about model or training parameters. Otherwise, the test set is not independent anymore and cannot be used to give valid information about generalizability.

To avoid or reduce overfitting and thus improve generalizability, regularization is applied. A large number of regularization techniques exist. In the following, those of them are described that are used in this thesis.

First of all, the amount of training data is closely related to overfitting. A sufficient number of training samples need to be seen by the machine learning algorithm in order to extract the underlying rules for two reasons: First, the probability to learn noise from the training data is reduced by an increasing number of samples. Secondly, it is important that the training data describes the underlying rules of the machine learning problem well. To do so, the decision boundaries in the feature space must be sampled densely. As stated above, this becomes more and more

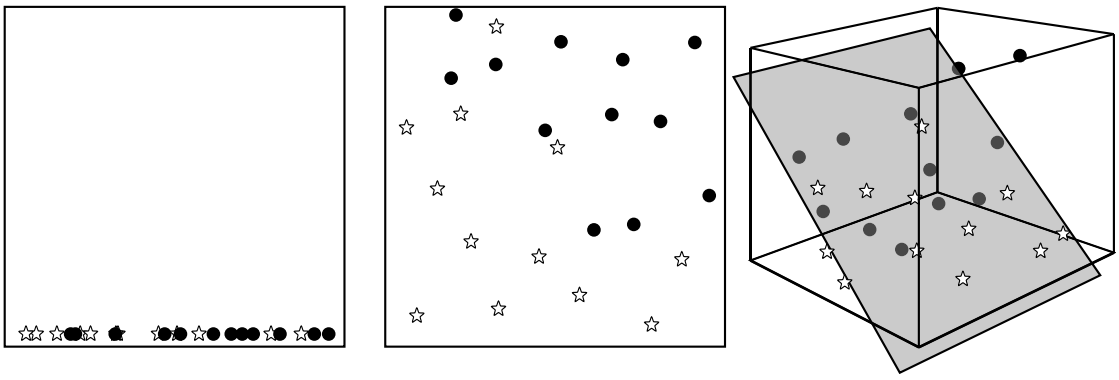


Figure 2.6: The images show the same dataset with 1, 2 and 3 features, respectively. Increasing the number of features eventually allows linear separation of the classes, the separating hyperplane is however usually only valid for the training data.

difficult as the number of features grows. Acquiring a sufficient amount of data is a major obstacle especially for deep learning methods. Data augmentation is a regularization technique to artificially increase the amount of data (Baird, 1992). For image analysis, this can be done by applying transformations to the data like rotating or flipping input images. Adding pixel-wise noise or slightly changing the color values are further augmentation techniques. When modifying the training data that way, care must be taken that the target values remain valid.

Another regularization technique is early stopping, which can be used for iteratively trained models like CNNs (Prechelt, 2012). For such models, overfitting often starts at some point in the training process. With an independent set of samples not used in the training process itself, the generalizability of the model can be assessed after each iteration. The training process can be stopped as soon as overfitting is detected. However, the sample set used to make that stopping decision is afterwards not independent any more. Therefore, instead of the test set a separate validation set is used for this procedure.

Preventing the model from accessing all available information during the training process is a further regularization method. For CNNs, dropout is a technique to set the activation of randomly selected neurons to zero (Hinton et al., 2012). This reduces the risk of overfitting, as the network is forced to learn more robust representations of the features and not to rely on specific pieces of information. For Random Forests, the training procedure also contains regularizations in the form of reducing the amount of information available. First of all, the training sets used for the individual decision trees do not contain all samples available. Although this increases the risk of overfitting for the individual trees, overfitting is prevented by merging the outputs for prediction. Secondly, when training the decision trees, for each decision only a subset of features is considered. Similarly to dropout, these procedures reduce the dependence of the Random Forest on specific pieces of information.

3

Overview of papers

As presented in Section 1.2, quantification of cellular structures is an important assessment in histopathology and has lots of applications. The papers of this thesis involve quantification of two different cellular structures. Papers 1, 2 and 3 focus on the detection of nuclei. Paper 4 presents an approach to quantification of FISH signals.

3.1 Nuclei quantification

Nuclei of various types and characteristics can be assessed by applying different stainings. With automated detection of these nuclei, many quantitative measures can be derived by assessing densities or ratios of different nuclei.

Papers 1, 2 and 3 make use of probability or proximity maps. The approaches of the three papers have in common that for each pixel position of an input image, a value is predicted using a machine learning algorithm. If the machine learning algorithm is a classifier, that value is the probability of the position belonging to the *nucleus center* class. For a regressor, the value represents the proximity to the nearest nucleus center. While these values are semantically different, they share the same core feature. High values (towards 1.0) are given to positions close to a nucleus center, lower values to positions further away from a center and lowest values (towards 0.0) to positions in the background, far away from any center. Replacing each pixel value in an input image by its predicted nucleus probability or proximity results in a map which will be called PMap in the following. This term was coined in Paper 3 to conveniently refer to both probability and proximity maps. Nuclei positions can be derived by detecting the local maxima of the PMap (cf. Paper 3) or by applying watershed transform (cf. Paper 1 and 2). By calculating densities or ratios of the detected nuclei, quantitative measures of the tissue can be determined. Because the PMap plays a central role in the detection approaches, we refer to them as PMap approaches.

3.1.1 Generic nuclei detection

Paper 1 presents a generic approach to nuclei detection. The approach uses a Random Forest classifier (cf. Section 2.2.2) to predict a PMap as explained above. The two possible classes for the Random Forest are called *nuclear* and *non-nuclear* and are slightly different to the classes described above and in the other papers. Here, PMap values are not only high at nuclei centers, but within the whole nuclear regions.

The first step of the proposed approach is a color deconvolution that separates the individual dyes of the staining. Afterwards, only the dye that highlights the nuclei of interest needs to be considered, which for example is hematoxylin in H&E stained images. Color deconvolution is parameterized by giving examples of all dyes involved at regions where they are predominant.

The next step of the approach is the classification of the *nuclear* and *non-nuclear* regions. The features used for the Random Forest classifier are hand-crafted and are all based on the nucleus highlighting dye as described above. Training of the classifier is performed interactively. Regions that belong to nuclei and such that belong to background can be marked by the user by painting over them in the original image. At any time, the user can invoke retraining of the Random Forest with the current training set. Immediately, the current field of view of the image is processed by the newly trained classifier and the classification result is shown as an overlay on the original image. The user can then mark further samples of either class to improve the classifier. By iterating this procedure, training samples are collected in an active learning fashion (Settles, 2009). After the training is finished, the PMap is processed by a watershed transform (Beucher and Meyer, 1993). The watershed transform is modified in a way that too small segments are merged in order to avoid oversegmentation. Nucleus positions are then extracted from the resulting segments by calculating their center of gravity.

The major evaluation goal of the paper is to demonstrate that the approach is generic, that is, that it can be quickly adapted to new nuclei detection tasks. The generalizability of the approach in terms of a trained model capable of dealing well with yet unseen data is of less interest. Therefore, no separate test set is used. Instead, the model is trained on the images interactively and the quality of the current model is inspected visually until no further improvements can be observed. Then, the model is evaluated with respect to reference annotations.

The tasks used for evaluation are the detection of either general nuclei or specifically lymphocytic nuclei in H&E stained images as well as the detection of positively stained nuclei in PR stained images. To detect lymphocytic nuclei and PR-positively stained nuclei, 800 training samples are required to achieve f1-measures of 0.830 and 0.937, respectively. When detecting nuclei without restriction to any particular cell type, more training is required. In two different datasets, $4,800$ and $3,900$ samples are needed to achieve f1-measures of 0.922 and 0.851, respectively. Although the number of training samples might seem large, they are very quickly produced. Generation of each training set in the interactive fashion requires only few minutes. The runtime of the algorithm is 0.27 megapixels per second (3.17 seconds per megapixel) on an Intel Core i7-3740QM using a single core.

3.1.2 Automated optimization of training data

Paper 2 presents an optimized training procedure for the approach in Paper 1. The features used by the Random Forest are optimized and the focus of the approach is changed from quick adaptability to good generalizability. The interactive training is replaced by an offline training. For training, only center marker annotations are available for the histological images. That means, for each nucleus in a defined region of each training image, the position of the center of the nucleus has been marked by a human observer.

The aim of the paper is to turn those simple annotations into training sets. To do so, different methods are proposed and evaluated by assessing the nuclei detection quality of the approach, when trained with the resulting training set. Training set generation is performed by sample extraction methods and sample reduction methods. Two different methods for the sample extraction are introduced. The first method is called distance-based and extracts pixels near the center marker annotations as *nucleus center* class and pixels far away from those as *background* class. The second method is called Voronoi-based and additionally generates a Voronoi tessellation of the annotation markers. Afterwards, further *background* samples are extracted along the edges of the Voronoi tessellation.

Sample reduction is performed to improve the model by removing redundant information from the training and to reduce the complexity of the decision trees to improve efficiency. To reduce the amount of samples, three different methods are presented. The first method is a simple stratified random subsampling. In the second method, two Kd-trees (Bentley, 1975) with limited depth are constructed, one from the *nucleus center* samples and one from the *background* samples. Afterwards, a single sample is randomly selected from each leaf or the Kd-tree. The third method employs an automated active learning procedure. To do so, a small random subset of training samples is used to train the classifier. Then, iteratively, the following three steps are performed. For all remaining samples, their class is predicted and the uncertainty of the prediction (probability of the *nucleus center* class near 0.5) is assessed. The samples being classified most uncertainly are then added to the training set and the classifier is retrained.

Evaluation is performed with both extraction methods each in combination with zero, one or two sample reduction methods. Voronoi-based sample extraction appears to outperform distance-based extraction. Applying one or more reduction methods improves the results considerably. Most important is the application of a reduction method that performs class balancing, which is the case for stratified random subsampling and Kd-tree subsampling. The highest quality measures for nuclei detection in the paper is achieved by the combination of stratified random subsampling followed by an automated active learning.

Evaluation is conducted using images of Ki-67 stained breast cancer tissue. Using the best-performing training set generation approach leads to an f1-measure of 0.826 on the test set.

3.1.3 Nuclei detection with convolutional neural networks

In Paper 3, the Random Forest is replaced by two variants of a CNN, being a classification network and a regression network. For the latter, the PMap is realized as a proximity map and the loss function is the mean squared error. For classification, class balancing is applied and cross-entropy is used as loss function.

In the literature, many other approaches for nuclei detection with CNNs use a workflow similar to the PMap approach described earlier. However, the results of the approaches differ substantially, and so do different aspects of the approaches. Thus, in Paper 3, we identify five of these aspects, or parameters, and collect them alongside the different settings of them found in the literature or proposed ourselves. The first two parameters are regularization options, namely whether or not to use dropout and data augmentation. The remaining parameters affect the appearance of the resulting PMaps and thereby its suitability for extracting nuclei positions. The first of them is which function is used to generate target values. Different options each for classification and regression networks are compared. The second parameter controls the way downsampling is compensated that occurs in the network layers. This compensation is required to receive PMaps of the same size as the input images. This is done by either replacing the strides in the layers with dilation or by adding an upsampling layer to the end of the network. And finally, the last of the five parameters is the post-processing of the PMaps. These parameters are systematically evaluated to assess both their general impact on the detection results and the best setting of each of them.

Evaluation shows that the parameters have a great impact on the quality of nuclei detection. Most important is the post-processing of the PMap, where the proposed combination of median and small Gaussian filter achieves best results. Although most state-of-the-art nuclei detection approaches are much more complex, we are able to show that with proper parameter settings, the rather simple PMap approach leads to detection results of the same high quality.

Evaluation of the parameter setting is performed using H&E stained colorectal adenocarcinoma tissue sections, where an f1-measure of 0.816 is achieved. Additionally, the best combinations of parameter settings are evaluated using the Ki-67 dataset already used in Paper 2, resulting in an f1-measure of 0.819. Unfortunately, these result cannot be directly compared to the results in Paper 2, as the distance threshold between reference and predicted nucleus position is more strict in Paper 3. On a machine with an Intel(R) Core(TM) i7-7700 CPU and an Nvidia GeForce GTX 1080 graphics card the proposed approach is able to process 4.15 megapixels per second, which, due to the usage of a GPU, is considerably faster than the approach in Paper 1.

3.2 Fluorescence signal quantification

As described in Section 1.2, FISH is a technique which can highlight certain DNA sequences in tissue. These highlights are called fluorescence signals and appear as small, bright dots. If these DNA sequences are copied multiple times as a result of a pathology, they are called amplified. Especially when amplified, the signals tend to form clusters so dense that even human observers often cannot identify

the individual signals any more. Therefore, in Paper 4, a density-based signal quantification is proposed.

Technically, the workflow of density-based quantification is similar to the PMap approaches described above. Instead of a probability or proximity map, a density map is generated by the CNN. Each pixel value in that map represents the local density of signals at the position of that pixel. The number of signals in a particular region of the image can then be determined by forming the integral of the density map over that region. Density-based counting is a technique mostly applied to the task of crowd counting, where high density and large occlusion can occur as well. To train the CNN, images with fluorescence signals have been annotated with center markers as accurately as possible. In the case of dense clusters, the overall number of signals in the cluster has been estimated by the human observer and the same number of annotations has been placed into the image. Target density maps are generated by placing Gaussian curves into the map at each annotation's position.

The density-based approach is compared to three other approaches. Two of them are derived from other publications. The first uses a difference-of-Gaussians filter to enhance the signals and extracts connected components after performing thresholding. Each connected component is then counted as a signal. The second one uses a CNN to generate signal probability maps and also extracts connected components after thresholding. The third compared approach is a modification of the second one. Instead of counting each connected component as one signal, the area of the components is measured. The number of signals is then determined by multiplying the overall signal area with a factor determined using a validation set.

All approaches are evaluated using breast cancer tissue stained with ERBB2 / CEN17 dual color fluorescent probes. In both the ERBB2 and the CEN17 fluorescence channels, signals are quantified. The clinically relevant distinction is, to put it simply, whether the ratio of ERBB2 signals to CEN17 signal exceeds 2.0. In this case, ERBB2 is considered amplified. The density-based approach outperforms the other approaches in all evaluated measures. On a per-case level, concordance with a pathologist of the distinction between amplified and not amplified is 0.971. Using the same machine as Paper 2 (Intel(R) Core(TM) i7-7700 CPU, Nvidia GeForce GTX 1080), the runtime of the approach is 30.95 megapixels per second for a single fluorescence channel.

Additionally, the impact of uncertainty in the training annotations is assessed for the density-based quantification approach. Marker positions are repositioned randomly inside a circular area of increasing radius before generating the target density maps. The approach appears to be robust against such alterations.

4 Discussion

The quantification of cellular structures plays an important role in modern histopathology. Counting of nuclei or other structures is required for an increasing amount of applications. Manual assessment of such values is labor- and time-intensive, and has been shown to have poor reproducibility. Automated quantification approaches have the potential to overcome these issues. But there are also challenges for the automated quantification. Most importantly, they must be able to handle both artifacts in and variability in an between histological images. Additionally, to hold the promise of decreasing the time required for assessment, they have to be efficient. The following sections discuss these issues and how they are addressed by the papers in this thesis.

4.1 Variability and artifacts

Histological images are virtually always subject to variability and artifacts. Variability occurs between both different images and tasks. Machine learning can be employed in order to cope with that problem. Instead of relying on given fixed rules, machine learning algorithms are trained using many examples of actual images. Thus, the algorithms can derive own rules that work well on all the different appearances of tissue included in the training set. Additionally, the algorithms can be confronted with different types of artifacts in order to learn to ignore them.

The requirements for a training set are high. Not only a sufficient amount of training data, but also the right data has to be collected.

Both Paper 1 and 2 use Random Forests to perform nuclei detection. However, their approaches to data collection and training are completely different.

Paper 1 focuses on training the algorithm interactively. By giving instant feedback about the classification results obtained using the current training set, an active learning procedure is achieved. This way, when improving the classifier by giving more training samples in falsely classified regions, the training set is augmented with the most informative data available. For multiple applications, we are able

to show that only a relatively small amount of training data is needed to achieve high nuclei detection quality.

The proposed approach enables the end user to quickly adapt the nucleus detector for new tasks. To do so, no knowledge about the inner mechanics of the algorithm is required. The training process requires the user only to point out areas of certain characteristics to the computer. It is thus completely visually driven, a realm especially pathologists are very familiar with.

In research settings, tissue preparation methods, stainings or the cellular structures of interest and thus the quantification tasks change frequently. Therefore, if automated quantification is desired in such settings, quick adaptability is essential. The approach can thus be an enabler for the use of automated analysis in research settings and has the potential to increase efficiency and reproducibility there.

The described approach is, however, not feasible for clinical routine applications, where there is a rather fixed set of quantification tasks. Instead, for those tasks, robust trained machine learning models should exist prior to application. Due to the variability in and between the images, such models need to be trained using large amounts of training data to cover all or at least most of the variability expected. Training the model once with that amount of data in the described interactive fashion is not feasible. It does not only require much time and effort, but the benefits of the active learning approach also diminish with larger amounts of training images. After each augmentation of the training set, the user has to review the quality of the new model, which is hardly manageable with lots of training images.

Instead, ideally, there would be training samples for each pixel in the training images in order to automatically train and evaluate the model. However, generating such data is very time consuming and labor-intensive. A more feasible way to produce training data is to only mark all nuclei centers in the training images. Therefore, Paper 2 presents an approach to extract training sets from such simple center annotations. By using such annotations, reference data can be generated in more and larger regions of multiple histological images. Therefore, the data covers a larger portion of variability and artifacts that may be expected during routine use. The extracted training data is afterwards condensed by an automated active learning procedure.

Paper 2 shows the importance of training set balancing. Although many clustered nuclei are present in the training images, the vast majority of training samples generated by the extraction methods belong to the *background* class. The decision trees of the Random Forests are constructed using Gini impurity or information gain, which both are sensitive to imbalanced training data and result in a bias towards the majority class (Flach, 2003). In Paper 2, class imbalance was very high, resulting in a severe underestimation of the probability of a sample being classified as *nucleus center*. This in turn makes the distinction between noise and real *nucleus center* regions more difficult and leads to detection errors.

Both Paper 1 and 2 show that active learning is a good way to generate small yet powerful training sets. Regardless of being used in a human-in-the-loop fashion (Paper 1) or automated (Paper 2), with active learning, such samples are added to the training set that contain most information for the classifier to the training set. The success of the training set reduction methods in Paper 2 implies that removing

training data can be beneficial for both the quality of the training set and the efficiency of the machine learning algorithm. At least for the Random Forest, this demonstrates that not only large quantities but also high-quality training data is required.

From a machine learning standpoint, this success is counterintuitive, as reducing the amount of training data increases the risk of overfitting. Nevertheless, it is carried out in this paper because we assume that large training sets for the classification task at hand contain much redundant information, especially when using small sets of hand-crafted features. Having multiple samples with redundant information leads to the classifier focusing on that information more than on other information and thus deteriorate classification quality.

In Paper 3, CNNs are used to detect and quantify nuclei. In contrast to Random Forests, CNNs have the advantage of not requiring hand-crafted features. When using hand-crafted features, the machine learning algorithm requires that the feature set contains all information needed to produce a prediction. The criteria humans use when performing nucleus detection manually, are implicit and often not used deliberately. Therefore, it is challenging to turn those, often complex, criteria into formalized features. It is especially difficult to cover variability and artifacts in such feature sets. Instead of relying on good hand-crafted features, CNNs omit explicit feature extraction and generate the features themselves implicitly during training.

PMap generation with CNNs can be formulated either as a classification or a regression problem. In the case of a classification problem, it turns out that, like in Paper 2, class balancing is very important. Without balancing, the optimizer is not able to reliably find an acceptable local minimum. The reason for that is that the trivial solution of always predicting the majority class already leads to a relatively low loss value, because the loss function is the cross-entropy averaged over all samples. It is very unlikely that the model can break out of this local minimum and find a better minimum during the training process. When formulating nuclei detection as a regression problem, a proximity map is used instead of a probability map. This formulation better fits to the fact that only center marker annotations are available. Proximity maps can directly be derived from such annotations without any further assumptions like an average nucleus radius. Also, better detection quality is achieved when using regression.

4.2 Dense structure clusters

In histological tissue sections, cellular structures are often densely packed. This constitutes a major challenge for automated quantification. Without clusters, classification of *structure* and *background* regions in conjunction with a connected component analysis would suffice to detect and quantify such structures. In the presence of clusters, however, this procedure leads to connected components often spanning multiple neighboring instances of the structures. In clusters, the structures often occlude or compress each other. Therefore, the number of structures contained in a connected component cannot be derived solely by its area. Furthermore, some pathological conditions also cause heterogeneity of the structures' size, like different nuclei sizes in many tumors. This additionally prevents count estimation by area.

In the nuclei quantification approaches of Paper 1, 2 and 3, PMaps generated by machine learning algorithms are employed to produce robust structure detection. The PMap approaches are thoroughly evaluated in the papers.

In Paper 2, generation of training sets is optimized for the classification-based PMap approach. The proposed training set generation consists of two distinct steps, being the extraction and the reduction of training samples. For the extraction, it appears that adding training samples of the *background* class along the edges of the Voronoi tessellation improved detection quality. As a machine learning classifier is used, those samples are important. In the frequent case of clustered nuclei where there is no background in between, those samples ensure that the PMap values decrease between such nuclei. Even if the border between clustered nuclei is not classified as background, the reduction of the *nucleus* class probability and thus the PMap values is sufficient for the subsequent processing step to split the cluster into separate nuclei.

In Paper 1, 2 and 3, it appears that post-processing of the PMap is crucial. When extracting local maxima or performing watershed transform on the PMaps, these need to be sufficiently smooth. Noisy PMaps lead to too many local maxima and oversegmentation in the watershed transform. However, in practice, PMaps do exhibit noise. Each value in the PMap is determined rather independently by the machine learning algorithm. Due to the nonlinear mapping between input and output of the machine learning algorithms, output values can differ significantly even for neighboring pixels with very similar inputs. Therefore, proper post-processing of the PMaps is of great importance. In Paper 1 and 2, smoothing is performed by an average filter. Paper 3 evaluates the impact of different post-processing methods. There, the proposed combination of a median filter and a Gaussian kernel results in the best quality of nucleus position extraction. The additional median filter removes outlier values from the PMap and thus reduces the detection of false positive nuclei.

PMaps are a simple and straight-forward way to perform quantification of cellular structures. Nevertheless, when parameterized properly, state-of-the-art results can be achieved, as shown in this thesis.

In Paper 4, the structures to be quantified are fluorescence signals. Here, clusters can be extremely dense, such that individual signals are not identifiable any more even for the human eye. Therefore, it seems that even machine learning algorithms are not able to obtain accurate information about the positions of the signals. For that reason, instead of detection-based quantification, we employ density-based quantification.

Density-based quantification has two major advantages to detection-based approaches. First of all, it is not reliant on detection. As there is no detection involved whatsoever, the algorithm does not try to find signal, background or border regions. Instead, only the amount of signal present at any position is predicted. The second advantage is the ability of the approach to deal with fractions of signals. Thus, small errors only affect the resulting measure by a small amount. For detection, if a detection error occurs, the number of signals is wrong by at least 1.

But density-based quantification also has drawbacks. The advantage of not having signals detected with their positions is a major drawback at the same time. For the user, it is rather opaque why the approach outputs a certain signal count. In

case of detection, each detected signal can be visualized and reviewed by the user. Also, corrections can easily be made by the user. For density-based quantification, only a number can be shown for an image region. While such a number can be corrected as well, doing so is less intuitive.

Nevertheless, in case of very dense clusters, density-based quantification provides a good alternative to other approaches. It is also suitable for several other applications, including nuclei quantification, where it already has been used (Lempitsky and Zisserman, 2010).

4.3 Efficiency

For the automated analysis of histological tissue sections, efficiency of the algorithms is crucial. The main reason is the size of the images to be analyzed. Images of 100,000 pixels in one dimension are encountered frequently. The more efficient an automated quantification is, the larger is the part of the images for which it is feasible to perform automated quantification. Especially in tissue that exhibits pathologies, biomarkers are often distributed heterogeneously. Having more efficient algorithms thus helps increasing robustness and reproducibility.

The training procedure of Random Forests demands that all decision trees are trained until the leafs only contain samples of the same class. Therefore, an increase of training samples usually leads to deeper trees. The time required to predict a single sample is directly dependent on the depth of the trees. In Paper 1 and 2, active learning is used to decrease the amount of samples required for the Random Forest to perform high-quality nuclei detection. Paper 2 compares different approaches to craft small training sets from the annotations available.

CNNs have been shown to outperform classical machine learning methods for several image analysis problems. However, state-of-the-art CNN methods for nuclei detection often have complex architectures with large amounts of parameters. In Paper 3, we use a CNN approach that is very simple. The basic idea of the approach is commonly seen in the literature, but the detection quality is often below state-of-the-art. We identify several important parameters of the approach and are able to adjust them such that the detection quality of the approach is equally good as the quality of state-of-the-art methods. Having a much simpler architecture, our approach is much more efficient both during training and prediction.

5 Conclusions

Individual conclusions of the works of this thesis are discussed in the respective sections of the papers. In the following, some more general conclusions are drawn and implications of the performed work are described.

5.1 Impact of quantitative analysis

This thesis aims at improving automated quantification of cellular structures in histological images. It addresses some of the major challenges that impede the wide use of automated analysis. Optimization of the training of machine learning methods is performed in order to be able to cope with the large variability and artifacts being present in histological images. Analysis of clustered structures is optimized by using PMap approaches and by performing density-based quantification. Also, the efficiency of the algorithms is improved by reducing the complexity of machine learning models.

Nowadays, pathology is still very little quantitative. Although more and more quantitative biomarkers exist, they are usually estimated by eyeballing and are often captured using semi-quantitative scores. Additionally, they are only examined in small regions of interest, which may be prominent regions like hot-spots or just randomly selected regions. Automated analysis helps making pathology more quantitative. With efficient automated analysis, much larger areas can be examined, resulting in statistically more robust and reproducible assessments.

Larger areas of examination not only improve the robustness of the assessments but also allow for heterogeneity analysis. For several measures it has been shown that the average value over a large or multiple smaller regions is not sufficient if the values differ substantially in different regions (Hamilton et al., 2014; Heppner, 1984). With quantification across large parts of the slides, heterogeneity can be assessed and quantified very easily. Combining histological measures with measures about their heterogeneity could further improve diagnosis and therapy decision.

When searching for prognostic or predictive measures in pathology, biomarkers are often correlated with patient survival or therapy response. To obtain meaningful

correlation results, those biomarkers need to have a good signal-to-noise ratio. Thus, first of all, they should be as correct as possible, since incorrect values increase the noise. Secondly, they should be robust in terms the underlying tissue regions being as large as possible to reduce sampling errors. And thirdly, they should be as precise as possible. This means that continuous valued measures are to prefer in contrast to semi-quantitative or categorical measures which result from qualitative assessments, as those reduce signal. Automated quantification can improve the results in all three aspects and can thus lead to better signal-to-noise ratio compared to manual determined measures. As a consequence, automated quantification can improve finding patient groups and sub-groups for which more tailor-made prognoses can be given and, most importantly, more targeted and precise therapies can be carried out.

5.2 Automated quantification in practice

Automated quantification in histological images still has to make the transition into practice. To accomplish that, automated analysis has to be of sufficient quality and efficiency. Both factors are covered in this thesis.

The quality of quantification algorithms is the most obvious aspect for practicality. Only if algorithms, and thereby their results, are of sufficient quality, they can take over parts of pathologists' tasks and thus reduce their workload. Considering practical applicability, there are two stages of algorithm quality.

The first stage are algorithms delivering results that are good enough as a first suggestion. This means that they do not have to be sufficiently good to directly paste them into the pathology report. Instead, the pathologist is required to validate the results or, if necessary, to correct them. Sufficient quality for the first stage is accomplished when, on average, correcting and validating the results requires the pathologist less effort than performing the analysis manually.

For the second stage the quality of the algorithm's results has to be of such a high level that it can be trusted blindly. This does not mean that the algorithm is "perfect", but rather that the deviation from a gold standard lies within the inter-observer variance of human pathologists. If this is the case, such an algorithm can be considered to be as good as a human pathologist.

To reach practical applicability and entering clinical routine, the first stage is already sufficient. Obviously, transition from the first to the second stage is fluent. In this way, automated analysis can be used at an early stage and relieve pathologists more and more, while they are further improved until the second stage is reached. In research settings, the second stage will often not be reached due to quickly changing tasks with limited training data and time. However, for such non-routine tasks, users can profit from automated analyses of the first stage as well.

The second aspect regarding practical applicability is the efficiency of the automated analysis. Efficiency is of great importance, but is often neglected. An algorithm of great quality that requires more time to compute than a pathologist needs to perform the same analysis manually is of limited use. Although it still offers many benefits such as more robust results, inefficient algorithms will not be viable in practice for economic reasons.

It is often argued that long-running analyses can be performed overnight and that the results can be presented to the pathologist the next day. But this argument is not necessarily valid. Given a limited amount of computing capabilities, consequently, only a limited number of analyses can be performed each day. The less efficient an automated algorithm is, the smaller is the number of slides that can be analyzed in a given amount of time. On the other hand, the better the quality of the algorithms becomes, the more tasks they are going to take over from pathologists. Thus, more analyses have to be performed each day. Therefore, in order to be practically applicable, developing highly efficient algorithms is of great importance.

We are convinced that practicality in general and efficiency in particular are central aspects for the establishment of automated analysis in histology. Therefore, such considerations can be found in all papers of this thesis.

For non-routine tasks, efficiency is also to be considered in terms of efficient training of algorithms. High quality needs to be achieved with often very limited amounts of training data and training effort to make the algorithm applicable in practice. Therefore, in this work, optimizations of the training process are proposed.

5.3 Future prospects

While the papers of this work raise their own individual questions that are worth being investigated, this section presents future research directions that result from the overall view of the papers. The proposed directions comprise technical optimizations, workflow improvements as well as practical benefit considerations.

For training of the CNNs, different technical improvements appear to be beneficial. Smoothing of PMaps appears to be crucial for the prediction of nuclei. However, smoothing a PMap as a post-processing step always discards some of the information gained from the network. Therefore, an integration of smoothness constraints into the loss function, for example by penalizing quick changes in the PMap, might improve the results. A built-in smoothness constraint would make an additional smoothing step unnecessary while still reducing noise. Furthermore, omitting this extra step reduces runtime.

Another improvement regarding the loss of CNN training for nuclei detection is the usage of the earth mover's distance (Rubner et al., 2000) instead of the mean squared error used in Paper 3. Simply put, the earth mover's distance interprets the PMaps as a landscape and measures how much work would be required to move earth in the prediction PMap such that it equals the target PMap. For nuclei detection, the loss function should be as closely related to the f1-measure as possible. However, the f1-measure itself cannot be used as loss function, as it is not differentiable. Compared to a pixel-wise comparison, the earth mover's distance could better resemble the f1-measure. Using it as the loss function could therefore improve detection quality. Additionally, it could increase robustness of the training to the size and shape of the kernel which the target PMap is generated with.

In paper 2, the Voronoi tessellation is used to select background samples that are located between two nuclei. It helps forcing the PMap to form a valley on

the rims of touching nuclei. This procedure could also be used to train CNNs. Improvements in PMap generation could be achieved by increasing the weight of the loss near the edges of the Voronoi tessellation.

In paper 1, nuclei detection is trained in an interactive fashion using a Random Forest. This allows quick adaption of the analysis for new tasks like nuclei detection in different stainings. Deep learning–based analysis has been shown to outperform classical machine learning methods regarding prediction quality. Obviously, combining both interactive training and deep learning is potentially very beneficial. However, deep learning requires a time-consuming training step, impeding interactive training. Concepts, such as transfer learning (Pan and Yang, 2010) or super-convergence (Smith and Topin, 2018) can reduce the time required for retraining. To further facilitate non-routine tasks that are often present in research settings, future developments should improve the efficiency of training to enable interactive deep learning.

The two stages of algorithm quality are discussed in Section 5.2. In order to convince stakeholders that the time for automated analysis has come, the practical value of first-stage-algorithms needs to be emphasized. To do so, reporting the usual quality measures like accuracy, f1-measure or the area under the receiver-operation-characteristic curve is not sufficient. Instead, the practical value should be measured and reported. This can be done by comparing the time required by the pathologist to perform the assessment, with and without computer support.

When human level quality is required for algorithms of the second stage, future work should pay more attention to the question what exactly human level quality is. In histology, a real ground truth can hardly ever be created, as histological images always leave some room for interpretation. To measure quality of human assessments, inter-observer-variance can be determined. We can consider an automated algorithm to have human level quality, if the algorithm's results lie within the inter-observer-variance of pathologists. Therefore, future evaluations should include determination of inter-observer-variance if not already known and assessment of the algorithm results with respect to that variance. We already conducted such an evaluation for the approach described in Paper 3 with a small number of observers. In that yet unpublished study the we were able to show that the agreement between automated analysis and the observers is higher than the agreement between the observers.

List of own publications

Höfener, H., Homeyer, A., Förster, M., Schildhaus, H.-U., Hahn, H.K., 2018. Automated Density-based Counting of FISH Amplification Signals for HER2 Status Assessment. Submitted to: Computer Methods and Programs in Biomedicine.

Höfener, H., Homeyer, A., Weiss, N., Molin, J., Lundström, C.F., Hahn, H.K., 2018. Deep learning nuclei detection: A simple approach can deliver state-of-the-art results. Computerized Medical Imaging and Graphics 70, 43–52.

Homeyer, A., Hammad, S., Schwen, L.O., Dahmen, U., **Höfener, H.**, Gao, Y., Dooley, S., Schenk, A., 2018. Focused scores enable reliable discrimination of small differences in steatosis. Diagnostic Pathology 13, 76.

Weiss, N., **Kost, H.**, Homeyer, A., 2018. Towards Interactive Breast Tumor Classification Using Transfer Learning, in: Campilho, A., Karray, F., ter Haar Romeny, B. (Eds.), Image Analysis and Recognition. Springer International Publishing, Cham, pp. 727–736.

Homeyer, A., Nasr, P., Engel, C., Kechagias, S., Lundberg, P., Ekstedt, M., **Kost, H.**, Weiss, N., Palmer, T., Hahn, H.K., Treanor, D., Lundström, C., 2017. Automated quantification of steatosis: agreement with stereological point counting. Diagnostic Pathology 12, 80.

Kost, H., Homeyer, A., Molin, J., Lundström, C., Hahn, H.K., 2017. Training nuclei detection algorithms with simple annotations. Journal of Pathology Informatics 8, 21.

Kost, H., Homeyer, A., Bult, P., Balkenhol, M.C.A., van der Laak, J.A.W.M., Hahn, H.K., 2016. A generic nuclei detection method for histopathological breast images, in: Gurcan, M.N., Madabhushi, A. (Eds.), Medical Imaging 2016: Digital Pathology. Presented at the SPIE Medical Imaging 2016, p. 97911E.

References

Alexe, G., Dalgin, G.S., Scanfeld, D., Tamayo, P., Mesirov, J.P., DeLisi, C., Harris, L., Barnard, N., Martel, M., Levine, A.J., Ganesan, S., Bhanot, G., 2007. High expression of lymphocyte-associated genes in node-negative HER2+ breast cancers correlates with lower recurrence rates. Cancer Research 67, 10669–10676.

Arteta, C., Lempitsky, V., Noble, J.A., Zisserman, A., 2012. Learning to detect cells using non-overlapping extremal regions. In: Medical Image Computing and Computer-Assisted Intervention MICCAI 2012. Springer, pp. 348–356.

Baird, H.S., 1992. Document Image Defect Models. In: Baird, H.S., Bunke, H., Yamamoto, K. (Eds.), Structured Document Image Analysis. Springer, Berlin, Heidelberg, pp. 546–556.

Bellman, R.E., 1957. Dynamic programming. Princeton University Press, Princeton, NJ.

Bengio, Y., 2009. Learning Deep Architectures for AI. Foundations and Trends in Machine Learning 2, 1–127.

Bentley, J.L., 1975. Multidimensional binary search trees used for associative searching. Communications of the ACM 18, 509–517.

Beucher, S., Meyer, F., 1993. The morphological Approach to Segmentation: The Watershed Transformation. In: Dougherty, E.R. (Ed.), Mathematical Morphology in Image Processing, Optical Engineering. M. Dekker, New York, pp. 433–481.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M.C., Tassani, S., Piva, F., Perez-Amodio, S., Strippoli, P., Canaider, S., 2013. An estimation of the number of cells in the human body. Annals of Human Biology 40, 463–471.

Bishop, C.M., 2006. Pattern recognition and machine learning, Information science and statistics. Springer, New York.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32.

Cireşan, D.C., Meier, U., Masci, J., Gambardella, L.M., Schmidhuber, J., 2011. Flexible, High Performance Convolutional Neural Networks for Image Classification. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two, IJCAI'11. AAAI Press, pp. 1237–1242.

Collins Dictionaries, 2014. Collins English Dictionary Complete and Unabridged edition: Over 700,000 words and phrases, 12th edition edition. ed. Collins, Glasgow.

Compton, C.C., 2000. Updated protocol for the examination of specimens from patients with carcinomas of the colon and rectum, excluding carcinoid tumors, lym-

phomas, sarcomas, and tumors of the vermiform appendix: A basis for checklists. Cancer Committee. Archives of Pathology & Laboratory Medicine 124, 1016–1025.

de Boer, M., van Deurzen, C.H.M., van Dijck, J.A.A.M., Borm, G.F., van Diest, P.J., Adang, E.M.M., Nortier, J.W.R., Rutgers, E.J.T., Seynaeve, C., Menke-Pluymers, M.B.E., Bult, P., Tjan-Heijnen, V.C.G., 2009. Micrometastases or isolated tumor cells and the outcome of breast cancer. The New England Journal of Medicine 361, 653–663.

Elston, C.W., Ellis, I.O., 1991. 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, 403–410.

Emmadi, R., Wiley, E.L., 2012. Evaluation of Resection Margins in Breast Conservation Therapy: The Pathology Perspective - Past, Present, and Future. International Journal of Surgical Oncology 2012.

Epstein, J.I., Egevad, L., Amin, M.B., Delahunt, B., Srigley, J.R., Humphrey, P.A., Grading Committee, 2016. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. The American Journal of Surgical Pathology 40, 244–252.

Flach, P.A., 2003. The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In: In Proceedings of the Twentieth International Conference on Machine Learning. AAAI Press, pp. 194–201.

Gleason, D.F., 1966. Classification of prostatic carcinomas. Cancer Chemotherapy Reports 50, 125–128.

Glorot, X., Bordes, A., Bengio, Y., 2011. Deep Sparse Rectifier Neural Networks. In: Gordon, G., Dunson, D., Dudík, M. (Eds.), Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, FL, USA.

Goceri, E., Shah, Z.K., Layman, R., Jiang, X., Gurcan, M.N., 2016. Quantification of liver fat: A comprehensive review. Computers in Biology and Medicine 71, 174–189.

Gudlaugsson, E., Skaland, I., Janssen, E.A.M., Smaaland, R., Shao, Z., Malpica, A., Voorhorst, F., Baak, J.P.A., 2012. Comparison of the effect of different techniques for measurement of Ki67 proliferation on reproducibility and prognosis prediction accuracy in breast cancer. Histopathology 61, 1134–1144.

Hamilton, P.W., Bankhead, P., Wang, Y., Hutchinson, R., Kieran, D., McArt, D.G., James, J., Salto-Tellez, M., 2014. Digital pathology and image analysis in tissue biomarker research. Methods, Advancing the boundaries of molecular cellular pathology 70, 59–73.

Helbich, T.H., Rudas, M., Haitel, A., Kohlberger, P.D., Thurnher, M., Gnant, M., Wunderbaldinger, P., Wolf, G., Mostbeck, G.H., 1998. Evaluation of needle size for breast biopsy: Comparison of 14-, 16-, and 18-gauge biopsy needles. American Journal of Roentgenology 171, 59–63.

Heppner, G.H., 1984. Tumor heterogeneity. Cancer Research 44, 2259–2265.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.,

2012. Improving neural networks by preventing co-adaptation of feature detectors. ArXiv preprint arXiv:1207.0580.

Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., 2001. Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. IEEE Press.

Homeyer, A., Nasr, P., Engel, C., Kechagias, S., Lundberg, P., Ekstedt, M., Kost, H., Weiss, N., Palmer, T., Hahn, H.K., Treanor, D., Lundström, C., 2017. Automated quantification of steatosis: Agreement with stereological point counting. Diagnostic Pathology 12, 80.

Hughes, G., 1968. On the mean accuracy of statistical pattern recognizers. IEEE Transactions on Information Theory 14, 55–63.

Inwald, E.C., Klinkhammer-Schalke, M., Hofstädter, F., Zeman, F., Koller, M., Gerstenhauer, M., Ortmann, O., 2013. Ki-67 is a prognostic parameter in breast cancer patients: Results of a large population-based cohort of a cancer registry. Breast Cancer Research and Treatment 139, 539–552.

Kårsnäs, A., Dahl, A.L., Larsen, R., 2011. Learning histopathological patterns. Journal of Pathology Informatics 2, 12.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12. Curran Associates Inc., USA, pp. 1097–1105.

Kumar, V., Abbas, A.K., Fausto, N., Mitchell, R.N. (Eds.), 2007. Robbins Basic Pathology, Eighth Edition, 8th edition. ed. Saunders/Elsevier, Philadelphia.

Langer-Safer, P.R., Levine, M., Ward, D.C., 1982. Immunological method for mapping genes on Drosophila polytene chromosomes. Proceedings of the National Academy of Sciences of the United States of America 79, 4381–4385.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D., 1989. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Comput. 1, 541–551.

Lempitsky, V., Zisserman, A., 2010. Learning To Count Objects in Images. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1, NIPS'10. Curran Associates Inc., USA, pp. 1324–1332.

Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., Manson, Q.F., Stathonikos, N., Baidoshvili, A., van Diest, P., Wauters, C., van Dijk, M., van der Laak, J., 2018. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: The CAMELYON dataset. GigaScience 7.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60–88.

McCann, M.T., Ozolek, J.A., Castro, C.A., Parvin, B., Kovacevic, J., 2015. Automated Histology Analysis: Opportunities for signal processing. IEEE Signal Processing Magazine 32, 78–87.

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in

nervous activity. The Bulletin of Mathematical Biophysics 5, 115–133.

Meijering, E., 2012. Cell Segmentation: 50 Years Down the Road [Life Sciences]. IEEE Signal Processing Magazine 29, 140–145.

Minsky, M., Papert, S., 1969. Perceptrons. M.I.T. Press, Cambridge, MA.

Nardelli, G.B., Lamaina, V., Siliotti, F., 1986. Estrogen and progesterone receptors status in the prediction of response of breast cancer to endocrine therapy (preliminary report). European Journal of Gynaecological Oncology 7, 151–158.

Nielsen, M.A., 2015. Neural Networks and Deep Learning. Determination Press, San Francisco.

Oh, K.-S., Jung, K., 2004. GPU implementation of neural networks. Pattern Recognition 37, 1311–1314.

Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 971–987.

Opitz, D., Maclin, R., 1999. Popular Ensemble Methods: An Empirical Study. Journal of Artificial Intelligence Research 11, 169–198.

Pan, S.J., Yang, Q., 2010. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering 22, 1345–1359.

Prechelt, L., 2012. Early Stopping – But When? In: Montavon, G., Orr, G.B., Müller, K.-R. (Eds.), Neural Networks: Tricks of the Trade: Second Edition, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 53–67.

Ramos-Vara, J.A., Miller, M.A., 2014. When Tissue Antigens and Antibodies Get Along: Revisiting the Technical Aspects of Immunohistochemistry - The Red, Brown, and Blue Technique. Veterinary Pathology 51, 42–87.

Rubin, R., Strayer, D.S., Rubin, E., 2011. Rubin's Pathology: Clinicopathologic Foundations of Medicine, Sixthtion. ed. Lippincott Raven, Philadelphia.

Rubner, Y., Tomasi, C., Guibas, L.J., 2000. The Earth Mover's Distance as a Metric for Image Retrieval. International Journal of Computer Vision 40, 99–121.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536.

Russell, S., Norvig, P., 2009. Artificial Intelligence: A Modern Approach, 3 edition. ed. Pearson, Upper Saddle River.

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural Networks 61, 85–117.

Scholzen, T., Gerdes, J., 2000. The Ki-67 protein: From the known and the unknown. Journal of Cellular Physiology 182, 311–322.

Settles, B., 2009. Active learning literature survey (Computer Sciences Technical Report No. 1648). University of Wisconsin-Madison.

Sirinukunwattana, K., Pluim, J.P.W., Chen, H., Qi, X., Heng, P.-A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., Böhm, A., Ronneberger, O., Cheikh, B.B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead,

D.R.J., Rajpoot, N.M., 2017. Gland segmentation in colon histology images: The glas challenge contest. Medical Image Analysis 35, 489–502.

Sirinukunwattana, K., Raza, S., Tsang, Y.-W., Snead, D., Cree, I., Rajpoot, N., 2016. Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. IEEE transactions on medical imaging 35, 1196–1206.

Slamon, D.J., Clark, G.M., Wong, S.G., Levin, W.J., Ullrich, A., McGuire, W.L., 1987. Human breast cancer: Correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 235, 177–182.

Slamon, D.J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., Baselga, J., Norton, L., 2001. Use of Chemotherapy plus a Monoclonal Antibody against HER2 for Metastatic Breast Cancer That Overexpresses HER2. New England Journal of Medicine 344, 783–792.

Smith, L.N., Topin, N., 2018. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. ArXiv preprint arXiv:1708.07120.

Sucaet, Y., Waelput, W., 2014. Digital Pathology, Springer Briefs in Computer Science. Springer International Publishing, Cham.

Suvarna, S.K., Layton, C., Bancroft, J.D. (Eds.), 2013. Bancroft's theory and practice of histological techniques, 7th ed. ed. Churchill Livingstone Elsevier, Oxford.

Tang, L.H., Gonen, M., Hedvat, C., Modlin, I.M., Klimstra, D.S., 2012. Objective quantification of the Ki67 proliferative index in neuroendocrine tumors of the gastroenteropancreatic system: A comparison of digital image analysis with manual methods. The American Journal of Surgical Pathology 36, 1761–1770.

Taqi, S.A., Sami, S.A., Sami, L.B., Zaki, S.A., 2018. A review of artifacts in histopathology. Journal of Oral and Maxillofacial Pathology 22, 279.

Tot, T., 2010. Breast Cancer: A Lobar Disease. Springer Science & Business Media, London.

Vink, J., Van Leeuwen, M., Van Deurzen, C., De Haan, G., 2013. Efficient nucleus detector in histopathology images. Journal of Microscopy 249, 124–135.

Wolff, A.C., Hammond, M.E.H., Allison, K.H., Harvey, B.E., Mangu, P.B., Bartlett, J.M., Bilous, M., Ellis, I.O., Fitzgibbons, P., Hanna, W., Jenkins, R.B., Press, M.F., Spears, P.A., Vance, G.H., Viale, G., McShane, L.M., Dowsett, M., 2018. Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. Archives of Pathology & Laboratory Medicine 142, 1364–1382.

Yaşar, P., Ayaz, G., User, S.D., Güpür, G., Muyan, M., 2016. Molecular mechanism of estrogenEstrogen receptor signaling. Reproductive Medicine and Biology 16, 4–20.

Yigit, N., Gunal, A., Kucukodaci, Z., Karslioglu, Y., Onguru, O., Ozcan, A., 2013. Are we counting mitoses correctly? Annals of Diagnostic Pathology 17, 536–539.

Paper 1

A generic nuclei detection method for histopathological breast images

Henning Kost, André Homeyer, Peter Bult, Maschenka C.A. Balkenhol, Jeroen A.W.M. van der Laak, Horst K. Hahn.

Proceedings of SPIE 9791, Medical Imaging 2016: Digital Pathology, 97911E. 2016.

A Generic Nuclei Detection Method for Histopathological Breast Images

Henning Kost[a], André Homeyer[a], Peter Bult[b], Maschenka C.A. Balkenhol[b], Jeroen A.W.M. van der Laak[b], and Horst K. Hahn[a]

[a]Fraunhofer MEVIS, Institute for Medical Image Computing, Universitätsallee 29, Bremen, Germany
[b]Department of Pathology, Radboud University Medical Centre, Geert Grooteplein 10, Nijmegen, The Netherlands

ABSTRACT

The detection of cell nuclei plays a key role in various histopathological image analysis problems. Considering the high variability of its applications, we propose a novel generic and trainable detection approach. Adaption to specific nuclei detection tasks is done by providing training samples. A trainable deconvolution and classification algorithm is used to generate a probability map indicating the presence of a nucleus. The map is processed by an extended watershed segmentation step to identify the nuclei positions. We have tested our method on data sets with different stains and target nuclear types. We obtained F1-measures between 0.83 and 0.93.

Keywords: Digital Pathology, Histology, Nuclei Detection, Biomarker Quantification, Breast

1. INTRODUCTION

The detection of different types of cell nuclei is an essential step in many automatic analysis methods for histopathological breast images. General nuclei detection in digitized Hematoxylin and Eosin (H&E) stained breast cancer tissue sections is an important prerequisite for different quantitative analyses like measuring the local cell density of tissue. Identifying lymphocytes in digitized H&E stained tissue sections enables the detection and quantification of lymphocytic infiltrates, which was shown to correlate with the disease-free survival and outcome for some tumor types.[1] Immunohistochemistry (IHC) offers the possibility to stain cells according to the presence of certain proteins. The detection of positively stained nuclei is a common task in automatic analysis methods for IHC stains.

In this paper we present a generic nuclei detection approach that can be easily adapted to specific nuclei detection tasks. The adaption is done by adjusting a single parameter and providing pixel examples of nuclear and non-nuclear regions.

2. METHODS

The proposed method is divided into three separate steps. First, the input image is deconvolved into its underlying dyes. Then, the input image is transformed into a nucleus probability map. A high probability of a nucleus center being present at a particular pixel leads to a high value at the corresponding pixel in the probability map. In the third step, the nuclei positions are obtained from the probability map by an extended watershed segmentation. Figure 1 illustrates the steps of the method.

Further author information: (Send correspondence to Henning Kost)
Henning Kost.: E-mail: henning.kost@mevis.fraunhofer.de, Telephone: +49 421 218 59244

Medical Imaging 2016: Digital Pathology, edited by Metin N. Gurcan, Anant Madabhushi, Proc. of SPIE Vol. 9791, 97911E · © 2016 SPIE
CCC code: 1605-7422/16/$18 · doi: 10.1117/12.2209613

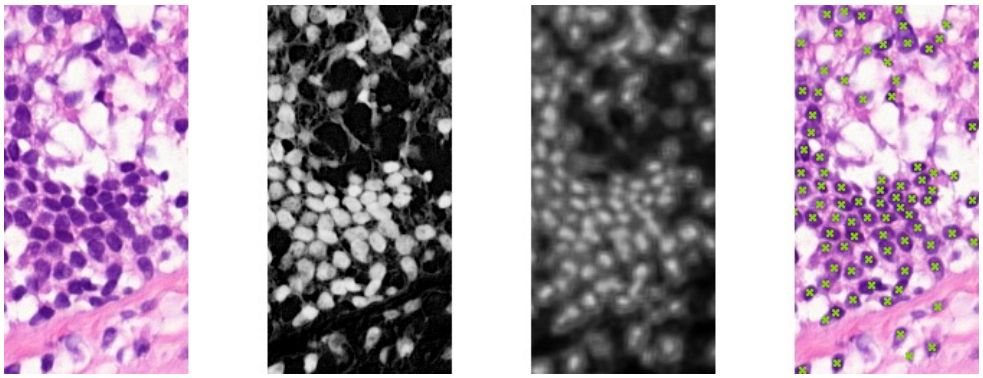


Figure 1. The steps of the proposed method for generic detection of nuclei. From left to right: detail of an image from data set 2; extracted Hematoxylin channel; generated probability map (higher probabilities appear brighter); visualization of detected nuclei.

2.1 Stain Deconvolution

Our approach works on stains that highlight nuclei. Those are present in many stain combinations like the Hematoxylin in H&E. The current step performs the extraction of the nuclear stain from a stain combination described in Ref. 2.

To do so, we perform a color space conversion from RGB to a space where each dye is represented by a separate channel. Instead of working on the intensity RGB values, we work with optical densities, which are computed by $OD = -log_{10}(I)$, where I is the intensity vector of the RGB channels normalized to $[0, 1]$. In this color space combinations of stains result in linear combinations of OD.

Assuming that the stain intensities C for a pixel in the image are known, the optical densities of the RGB channels can be determined by $OD = C \cdot M$, where M is the OD matrix of the stainings:

$$M = \begin{pmatrix} m_{R,0} & m_{G,0} & m_{B,0} \\ & \cdots & \\ m_{R,n} & m_{G,n} & m_{B,n} \end{pmatrix}$$

for n stainings. Consequentially $M^{-1} \cdot OD = C$ can be used to get the stain intensities from the input image.

In our approach, the Matrix M is filled by the user via pixel examples as explained in section 2.4. For each dye i, the values of the respective example pixels are averaged and the optical density vector $\overrightarrow{OD_i}$ of the resulting RGB vector is computed. Each row of M is then populated with the normalized $\overrightarrow{OD_i}$. Generating M from user input this way enables the quick adaptability of the algorithm to different stainings.

From the deconvolved image, only the channel is extracted that represents the nuclear stain. This channel constitutes the input for the next step.

2.2 Nucleus Probability Map

The main step of the proposed algorithm is the generation of the nucleus probability map. This map holds the probability of a nucleus being present for each pixel of the input image. It can be visualized as a gray-scale image ranging from black for lowest to white for highest nucleus probability (cf. image 3 of figure 1).

The main requirement of the nucleus probability map is to yield high values at nuclei centers and lower values at their periphery. These are the borders of the nuclei, which might be tightly connected to borders of adjacent nuclei. In background areas, the probability is required to have the lowest values (cf. figure 2).

The creation of the probability map is based on machine learning. We use a Random Forest classifier[3] for the distinction between the two classes *nuclear* and *non-nuclear*. For each pixel, the Random Forest is applied and

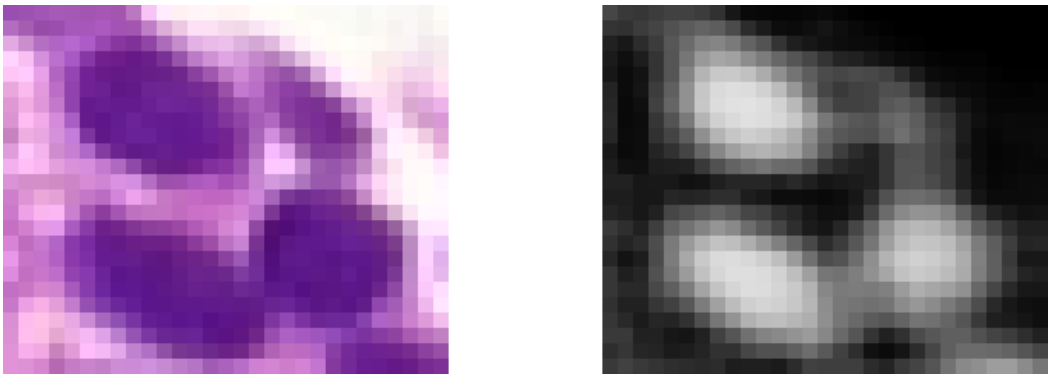


Figure 2. Detail of input image from figure 1 and corresponding probability map. Clustered nuclei appear separable.

the ratio of trees voting for the nuclear class is taken as the probability value. To reduce noise, the probability map is afterwards smoothed using an average filter with a 5x5 square kernel.

Training data for the classifier is given by the user. The training is pixel-based, so the user has to specify pixels for both classes mentioned above, as described in section 2.4.

For the classification, we thoroughly evaluated many different features. We selected a feature set both with respect to the above mentioned requirements and in terms of computational costs. All features in this set are extracted from the single stain channel, which also constitutes the base feature. Further features are the local minima, maxima, sum and dynamic range of the intensity values in a 3x3 pixel neighborhood.

The local minimum and maximum filters are implemented as described in Ref. 4. For the sum filter, an efficient 1D ring buffer implementation is used. The local dynamic range represents the difference of the minimum and maximum pixel of the current position's neighborhood.

In many whole slide image file formats, the image data is available in a pyramidal manner. This means that next to the highest resolution, lower resolutions of the original image are stored. These versions represent higher levels of the image pyramid and result from downsampling the image by a factor 2^l, where l denotes the level. We exploit this way of storing the images and extract our features on multiple levels of the image pyramid (in our case levels 0 to 3). The resulting various scales of the feature set ensure that the surrounding of each pixel is taken into account. This enables the distinction between nuclei center and periphery.

2.3 Nuclei Detection

For the detection of the nuclei, at first a threshold of 0.5 is applied to the probability map. In this manner, regions in which the majority of the decision trees voted against the nuclear class are excluded.

The resulting probability map can be interpreted as a gray scale landscape, where the mountains represent the nuclear regions. We perform a watershed segmentation using hierarchical queues[5] on that map to separate the nuclei.

To reduce oversegmentation, we extend the watershed algorithm with a size constraint, which works as follows: During the watershed, whenever the currently processed pixel touches more than one already labeled region, the sizes of the regions are considered. If there is more than one region of size sz_{min} or larger, the regions remain separated and the current pixel is not labeled. Else, the regions are merged and the current pixel is added to that merged region. Choosing a sz_{min} below the expected minimal size of a nucleus and above the size of noise structures allows for correct splitting of individual nuclei yet avoids oversegmentation.

Figure 3 visualizes the size constraint in a simple 1D example. After processing, the center points of the resulting regions are identified as the nuclei centers.

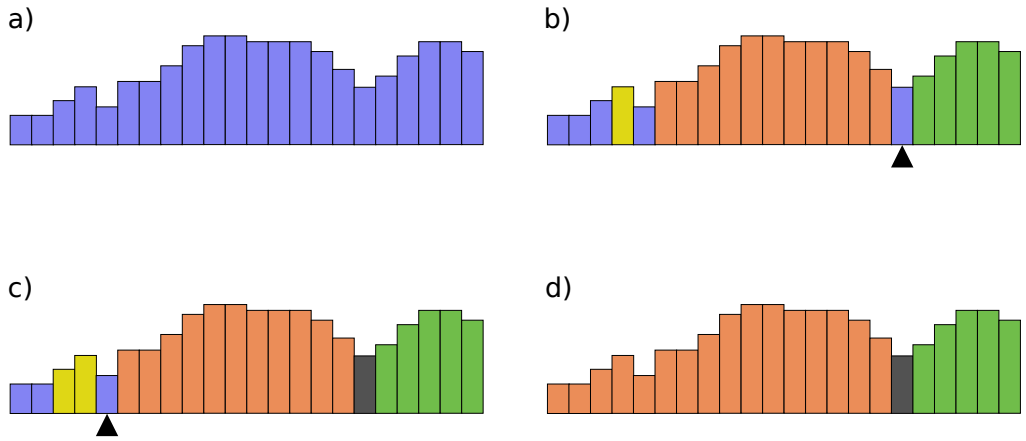


Figure 3. The watershed process for a 1D example. The boxes represent image pixels with heights according to the nucleus probability values. Segmented regions share the same color. The triangles mark the currently processed pixel in b and c. The orange and green regions are larger or equal sz_{min} and thus separated. The yellow region is smaller sz_{min} and merged into the orange region.

2.4 Training

The parametrization of the algorithm consists of the adjustment of the parameter sz_{min}, as stated above, and a set of training examples.

Both the training and the application of the proposed method are implemented within the Fraunhofer MEVIS Histokat software framework. In the graphical user interface histological images can be viewed and pixel examples can be directly marked in the image.

For the color deconvolution the user is requested to mark representative pixels for each dye. Such pixels are in areas of the image, where only a particular dye is visible like nuclei centers for Hematoxylin or stromal areas for Eosin.

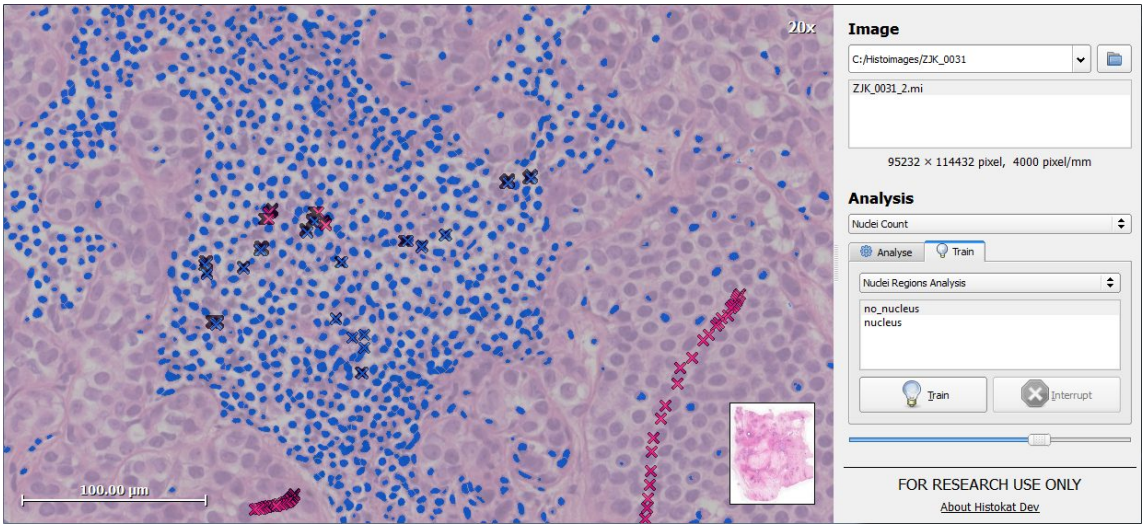


Figure 4. Training in the Histokat user interface. In this example, lymphocytic nuclei should be detected. Blue crosses mark pixel samples of the *nuclear* class, red crosses those of the *non-nuclear* class. The visualization overlay shows detected nuclei regions in blue. During the training process, the algorithm is only applied to the currently viewed image region. This saves computation time and enables a responsive user interaction.

To train the random forest, the user has to select example pixels of the classes *nuclear* and *non-nuclear*. It is possible at any time to invoke the training of the classifier with the selected samples. The trained classifier is then applied to the currently viewed region of the image and the segmented regions are visualized as an overlay (cf. figure 4). In this way, the user receives instant feedback of the training effect and can perform the training in an interactive way.

The user is always able to view and explore any available digitized slide image to check the performance of the current training on other fields of view and add additional training samples if desired. In this way, samples from different images can be selected to form a joint training basis. Whenever the user is satisfied with the training, he or she can invoke the analysis of the entire image or the whole set of images. After that, the quantitative results are output in tabular form.

3. RESULTS

To show the versatility of the proposed method, we have evaluated our method on multiple data sets and different tasks (cf. Table 1). The digitized tissue sections were stained with either H&E or the IHC stain for the progesterone receptor (PR). Tissue preparation and scanning was performed at the Radboud University Medical Centre in Nijmegen, the Netherlands. The data sets comprise field of view images at 20x magnification with image sizes between 400x300 and 900x600 pixels. Each field of view image belongs to a different patient.

For data set 1 and 3, a pathologist has provided comprehensive ground truth annotations of the nuclei. For data set 2 and 4 ground truth nuclei were annotated by a trained non-specialist. The ground truth data consists of labeled mask images with a unique label for each nuclear region.

The evaluation was performed for each dataset independently. First, the training was performed as described in section 2.4 with respect to the images in the dataset. Afterwards, the trained method was applied to all images in the dataset. The output is a set of the center points of the detected nuclei. This set was evaluated against the ground truth annotations.

For each annotated region r in the ground truth, the number of detected nuclei dn_r was counted. A result of $dn_r = 0$ was counted as one false negative (FN), $dn_r = 1$ as one true positive (TP) and $dn_r > 1$ as one TP and $dn_r - 1$ false positive (FP) nuclei. Detected nuclei outside any region were counted as false positives (FP) as well. Based on these value, the precision, sensitivity and F1-measure were computed.

For the different tasks, only the training samples and sz_{min} were adapted. The resulting F1-measures varied between 0.83 and 0.93.

Set	Task	Stain	Images	# Nuclei	# Training Samples	Precision	Sensitivity	F1
1	Lymphocytes	H&E	10	3645	800	0.7901	0.8748	0.8303
2	Nuclei	H&E	7	3505	4800	0.9224	0.9127	0.9224
3	Nuclei	H&E	19	3219	3900	0.8514	0.9217	0.8514
4	Positive Nuclei	PR	7	2019	800	0.9371	0.9009	0.9371

Table 1. Listing of the data sets and the respective results of the proposed method.

The average runtime of the C++ implementation of the proposed method was 3.72 seconds per Megapixel on an Intel Core i7-3740QM utilizing a single core. It became apparent, that the runtime of the algorithm depends on the number of training samples present for that image set and the number of nuclei being present in that image.

Both is expected. Since we use the Random Forest classifier without any pruning, the decision trees will have more nodes when trained with more samples, which leads to more processing time. Also, the more nuclei are present in the image, the more pixels of the probability map will yield values above 0.5. These pixels have to be processed by the watershed algorithm, which also results in a higher processing time. Figure 5 shows the runtimes for all images.

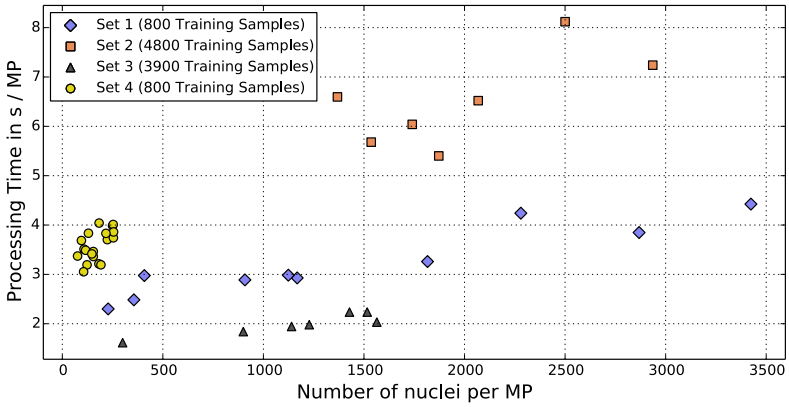


Figure 5. Number of nuclei and processing time in seconds (both per Megapixel) for the images of the 4 test sets. The different markers show the affiliations to the image sets.

4. CONCLUSIONS AND PROSPECTS

Many current nuclei detection algorithms utilize shape information to identify single nuclei or separate multiple clustered nuclei.[6–10] Derived shape information like symmetry or the response to a differential of Gaussian filter are also used.[11–13] However, the variation of nuclear shapes is huge. Especially in breast tumors, nuclei can have very irregular shapes. That is why we do not take shape-based features into account.

Instead, we propose a method for the detection of cell nuclei in digitized histopathological breast cancer tissue sections, which is generic and suited for the quick adaption to many stains or nuclear types. Although nuclei detection algorithms based on probability maps have been published before,[14,15] none of these publications focussed on adaptability and generic applicability.

We have shown that our method features high processing speed and good evaluation results for some nuclei detection tasks. These include both small and uniform lymphocytic nuclei as well as nuclei with heterogenious sizes, shapes and staining intensities. Also, the tasks include both conventional and immunohistochemical staining. The good performance on these tasks is an indicator for the generalizability of the proposed method. Applying it to further nuclear types and stainings will be subject of future work.

The adaption of our method requires the user to set one parameter (sz_{min}). That parameter can be chosen according to section 2.3. Apart from that, user can customize the algorithm by only providing representative pixel examples for the applied stain and for nuclear and non-nuclear regions. In contrast to other approaches, a thorough understanding of the algorithm is not necessary to adapt the method to novel tasks. Our method automatically learns from examples.

Given the small size of the available datasets, it was impossible to perform a thorough evaluation of the accuracy and robustness of the method with respect to the variability of images encountered in practice. Instead, our motivation was to show that the method is versatile, that is, that it can perform well in different applications, when properly trained. For this reason we have performed the training on the same images as the evaluation. A more thorough evaluation of the accuracy and robustness of the method is left for future work.

Further future work may be to improve the generation of the nucleus probability map, since it is the crucial part of the method for both detection quality and speed. Approaches could be the optimization of the feature set used by the Random Forest classifier. Alternatively, the classifier could be replaced by another machine learning approach like a deep convolutional neural network, which have been shown to perform well in other histological detection tasks.[16]

ACKNOWLEDGMENTS

The authors wish to acknowledge the financial support by the European Union 7th Framework Program funded research project VPH-PRISM (Virtual Physiological Human: Personalised Predictive Breast Cancer Therapy Through Integrated Tissue Micro-Structure Modeling, FP7-ICT-2011-9, 601040) and the research project AMI (Automation In Medical Imaging) which is funded by the Fraunhofer ICON initiative.

REFERENCES

[1] Alexe, G., Dalgin, G. S., Scanfeld, D., Tamayo, P., Mesirov, J. P., DeLisi, C., Harris, L., Barnard, N., Martel, M., Levine, A. J., Ganesan, S., and Bhanot, G., "High expression of lymphocyte-associated genes in node-negative her2+ breast cancers correlates with lower recurrence rates," *Cancer Res.* **67**, 10669–10676 (Nov. 2007).

[2] Ruifrok, A. C. and Johnston, D. A., "Quantification of histochemical staining by color deconvolution," *Anal. Quant. Cytol. Histol.* **23**, 291–299 (Aug. 2001).

[3] Breiman, L., "Random Forests," *Machine Learning* **45**, 5–32 (Oct. 2001).

[4] van Herk, M., "A fast algorithm for local minimum and maximum filters on rectangular and octagonal kernels," *Pattern Recognit. Lett.* **13**, 517–521 (July 1992).

[5] Beucher, S. and Meyer, F., "The morphological Approach to Segmentation: The Watershed Transformation," in [*Mathematical morphology in image processing*], Dougherty, E. R., ed., *Optical engineering* **34**, 433–481, M. Dekker, New York (1993).

[6] Wienert, S., Heim, D., Saeger, K., Stenzinger, A., Beil, M., Hufnagl, P., Dietel, M., Denkert, C., and Klauschen, F., "Detection and Segmentation of Cell Nuclei in Virtual Microscopy Images: A Minimum-Model Approach," *Sci. Rep.* **2** (July 2012).

[7] Ali, S. and Madabhushi, A., "An Integrated Region-, Boundary-, Shape-Based Active Contour for Multiple Object Overlap Resolution in Histological Imagery," *IEEE Trans. Med. Imaging* **31**, 1448–1460 (July 2012).

[8] Qi, X., Xing, F., Foran, D. J., and Yang, L., "Robust segmentation of overlapping cells in histopathology specimens using parallel seed detection and repulsive level set," *Biomed. Eng. IEEE Trans. On* **59**(3), 754–765 (2012).

[9] Kong, H., Gurcan, M., and Belkacem-Boussaid, K., "Partitioning Histopathological Images: An Integrated Framework for Supervised Color-Texture Segmentation and Cell Splitting," *IEEE Trans. Med. Imaging* **30**, 1661–1677 (Sept. 2011).

[10] Jung, C., Kim, C., Chae, S. W., and Oh, S., "Unsupervised Segmentation of Overlapped Nuclei Using Bayesian Classification," *IEEE Trans. Biomed. Eng.* **57**, 2825–2832 (Dec. 2010).

[11] Veta, M., van Diest, P. J., Kornegoor, R., Huisman, A., Viergever, M. A., and Pluim, J. P. W., "Automatic Nuclei Segmentation in H&E Stained Breast Cancer Histopathology Images," *PLoS ONE* **8**, e70221 (July 2013).

[12] Al-Kofahi, Y., Lassoued, W., Lee, W., and Roysam, B., "Improved Automatic Detection and Segmentation of Cell Nuclei in Histopathology Images," *IEEE Trans. Biomed. Eng.* **57**, 841–852 (Apr. 2010).

[13] Cosatto, E., Miller, M., Graf, H. P., and Meyer, J. S., "Grading nuclear pleomorphism on histological micrographs," in [*Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*], 1–4, IEEE (2008).

[14] Arteta, C., Lempitsky, V., Noble, J. A., and Zisserman, A., "Learning to detect cells using non-overlapping extremal regions," in [*Medical image computing and computer-assisted intervention–MICCAI 2012*], 348–356, Springer (2012).

[15] Kårsnäs, A., Larsen, R., and Dahl, A. L., "Learning histopathological patterns," *J. Pathol. Inform.* **2**(2), 12 (2011).

[16] Cireşan, D. C., Giusti, A., Gambardella, L. M., and Schmidhuber, J., "Mitosis detection in breast cancer histology images with deep neural networks," in [*Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*], 411–418, Springer (2013).

Paper 2

Training nuclei detection algorithms with simple annotations

Henning Kost, André Homeyer, Jesper Molin, Claes F. Lundström, Horst K. Hahn.

Journal of Pathology Informatics 8, 21. 2017.

The original publication is available at:
https://doi.org/10.4103/jpi.jpi_3_17

Research Article

Training Nuclei Detection Algorithms with Simple Annotations

Henning Kost[1], André Homeyer[1], Jesper Molin[2,3,4], Claes Lundström[3,4], Horst Karl Hahn[1]

[1]Fraunhofer Institute for Medical Image Computing MEVIS, 28359 Bremen, Germany, [2]Department of Applied Information Technology, Chalmers University of Technology, 41258 Gothenburg, [3]Sectra AB, 58330 Linköping, Sweden, [4]Center for Medical Image Science and Visualization, Linköping University, 58183 Linköping, Sweden

Received: 10 January 2017 Accepted: 17 March 2017 Published: 15 May 2017

Abstract

Background: Generating good training datasets is essential for machine learning-based nuclei detection methods. However, creating exhaustive nuclei contour annotations, to derive optimal training data from, is often infeasible. **Methods:** We compared different approaches for training nuclei detection methods solely based on nucleus center markers. Such markers contain less accurate information, especially with regard to nuclear boundaries, but can be produced much easier and in greater quantities. The approaches use different automated sample extraction methods to derive image positions and class labels from nucleus center markers. In addition, the approaches use different automated sample selection methods to improve the detection quality of the classification algorithm and reduce the run time of the training process. We evaluated the approaches based on a previously published generic nuclei detection algorithm and a set of Ki-67-stained breast cancer images. **Results:** A Voronoi tessellation-based sample extraction method produced the best performing training sets. However, subsampling of the extracted training samples was crucial. Even simple class balancing improved the detection quality considerably. The incorporation of active learning led to a further increase in detection quality. **Conclusions:** With appropriate sample extraction and selection methods, nuclei detection algorithms trained on the basis of simple center marker annotations can produce comparable quality to algorithms trained on conventionally created training sets.

Keywords: Active learning, machine learning, nuclei detection, training set generation

Introduction

Many pathological assessments depend on the quantification of cell nuclei. In cancer diagnosis, for instance, the quantification of nuclei expressing the Ki-67 protein is a widely used method to determine the proliferation rate of a tumor. Furthermore, the quantification of lymphocytic infiltrates has been shown to be of strong prognostic importance.[1] Another important application is the determination of the progesterone and estrogen receptor status. The latter is arguably the most important predictive biomarker that exists today.[2] In clinical routine, such evaluations are usually done manually by estimating or counting a small number of nuclei, which is highly subjective and often not reproducible.[3] Consequently, the ability to automatically detect different types of nuclei on larger regions becomes increasingly important.

Varying staining and tissue preprocessing conditions, as well as different nuclear types and pathologies, lead to a huge variability in the appearance of nuclei, making their automatic detection very challenging. Recent approaches employ trainable algorithms to address this issue, including traditional machine learning[4-6] as well as deep learning methods.[7-9] Trainable detection methods come with the advantage of being adaptable and refinable by just using different training datasets.

Generating a good training dataset is essential for such methods. Most of these methods learn some kind of pixel-wise distinction between nuclear and nonnuclear regions,[4,6-9] to either create an intermediate segmentation or a probability map. Hence, the optimal training data would consist of exhaustive manual segmentations of all nuclei in several histological images. Unfortunately, creating such annotations requires an expert to accurately draw contour lines around each nucleus, making it a very tedious and time-consuming task.

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_3_17

Address for correspondence: Mr. Henning Kost, Fraunhofer Institute for Medical Image Computing MEVIS, Am Fallturm 1, 28359 Bremen, Germany. E-mail: henning.kost@mevis.fraunhofer.de

How to cite this article: Kost H, Homeyer A, Molin J, Lundström C, Hahn HK. Training nuclei detection algorithms with simple annotations. J Pathol Inform 2017;8:21.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2017/8/1/21/206227

Annotation marks at the nuclear centers constitute an alternative kind of reference data. Center annotations can be created with much less effort because they only require the expert to mark nuclei with a single click. This makes the marking process much faster and, therefore, also allows larger amounts of images to be annotated. Obviously, such annotations comprise much less information than full segmentations.

Center marker annotations have already been employed in the past. The different approaches address their insufficiency by augmenting them in various ways. An iterative thresholding approach was used by Gul-Mohammed *et al.*[10] to distinguish nuclear and nonnuclear areas around the center markers. In a study by Janowczyk and Madabhushi,[9] this distinction is performed by a naive Bayesian classifier with center positions as nuclear training data and randomly selected noncenter positions as nonnuclear training data. However, these approaches tie the capability of the machine-learning algorithm to the capability of the previous step. In both the studies by Sirinukunwattana *et al.*[11] and Xu *et al.*,[8] an assumption regarding the size of the nuclei is incorporated to supplement the annotation data: In a study by Sirinukunwattana *et al.*,[11] a regression is trained using the distance to the next center marker to compute the target value. In a study by Xu *et al.*,[8] nonnuclear training samples were drawn from positions that are further away from any center marker than a given threshold.

The quality of the mentioned approaches is hard to compare as the authors usually use different data sets with different nuclear types and often also different quality measures.

In a study by Vink *et al.*,[4] a nucleus detection method for Her2-stained breast tissue is proposed. The authors report a detection rate, which equals recall, of 0.95. Breast tissue nuclei are also detected in the studies conducted by Xing *et al.*[7] and Xu *et al.*[8] The approaches work on H&E-stained images and yield f1-measures of 0.78 and 0.84, respectively. In the study conducted by Arteta *et al.*[5] and Janowczyk and Madabhushi,[9] lymphocytic nuclei are detected in H&E-stained breast images. They state f1-measures of 0.88 and 0.90, respectively. Kårsnäs *et al.*[6] reported that a detection method for Ki-67-positive nuclei in breast tissue is proposed. The authors announce 1.0% missing objects, 2.6% missing annotations, and 4.1% multiple annotations. A nuclei detection method for H&E-stained colorectal tissue is described by Sirinukunwattana *et al.*[11] and an f1-measure of 0.80 is reported.

In this paper, we perform a systematic comparison of different methods for generating training sets solely from center marker annotations. In addition, we evaluate how the proposed center marker-based sample extraction methods compare with manual segmentations.

Methods

Training set generation

A training set consists of a set of training samples, which in turn consist of a feature vector and a class label. The content of the feature vector depends on the classification method that is to be trained. It might comprise hand-crafted features or in case of feature learning methods such as deep convolutional neural networks, small image patches. In both cases, a training sample is produced with respect to a given position in the image.

All of the examined training set generation approaches consist of two main steps, which are the extraction and the selection of training samples.

Given a set of training images with labeled center markers, the extraction step needs to identify image positions that can be labeled as nuclear or nonnuclear regions and derive a training sample from it. The main difficulty here is that center markers obviously provide far less information about the nuclear and nonnuclear regions in the image, especially with regard to their boundaries.

The output of the extraction step already forms a valid training set. However, the abundance of training samples often deteriorates the analysis quality and the runtime performance for the training process of the classifier. Depending on the type of the classifier, also, the runtime of the nuclei detection can be increased considerably. Thus, the second step of the considered training set generation approaches is the selection of optimal subsets of training samples.

Training sample extraction

We compare two different methods for extracting training samples from a given set of images.

Distance-based

We assume that positions close to the annotated center markers can be considered to represent nuclear regions whereas positions far away from any center marker are very likely to represent nonnuclear regions. Training samples are extracted as follows:

For each position x, y in an image, we compute the distance to the closest center marker. That distance and the index of that closest marker are stored in two maps d(x, y) and m(x, y). To be designated as nuclear region, a position must not be further away from the closest marker than a threshold called t_{nuc}. Thus, all positions x, y where $d(x, y) < t_{nuc}$ can be labeled as nuclear region. To be designated as nonnuclear region, a position may not be closer to any marker than a threshold called t_{bg}. Consequently, all positions x, y where $d(x, y) > t_{bg}$ are labeled as nonnuclear region. For our experiments, we set t_{bg} to 15 pixels and t_{nuc} to 3 pixels each at 20× resolution.

Voronoi-based

The distance-based approach has the drawback that the boundary positions of the nuclei are not considered at all. Boundary positions, however, are very informative because they shape the decision boundaries of the classifier. In our case, nuclear boundaries should be labeled as nonnuclear region so that clustered nuclei can be separated by the classifier. The Voronoi-based extraction method augments the distance-based method with such boundary samples.

The marker map m(x, y) is equivalent to the Voronoi diagram of the center markers. Assuming that neighboring nuclei are similarly sized, the Voronoi boundary between nontouching nuclei only crosses nonnuclear regions. As soon as two nuclei are touching, the Voronoi boundary crosses exactly that touching point. Consequently, for overlapping nuclei, the region of overlap is crossed by the Voronoi boundary. The assumption above may not always be valid, leading to Voronoi boundaries crossing nuclear regions, but we found that being a rare case in our experiments. Thus, the Voronoi boundaries are suited to extract nonnuclear samples along them.

Figure 1 illustrates the sample extraction methods.

Training sample selection

Selecting a subset of training samples from those extracted in the previous step can be beneficial. Reducing the amount of samples leads to a decrease of the runtime of the training process. For some classifiers, such as the random forest, the runtime of the classification is reduced as well.

Moreover, subsets of training samples often result in a higher quality of the nuclei detection if the samples show class imbalance. The extraction methods generally produce more samples of nonnuclear regions than nuclear regions because of the relative area fractions in the image. A small t_{nuc} further increases that imbalance. A classifier confronted with substantial class imbalance is deluged by instances of the majority class, leading it to ignore the instances of the minority class. Such imbalance is a well-known issue in the field of machine learning.[12]

During the training of a machine-learning classifier, the most interesting regions of the feature space are those close to the decision boundary of the classifier. Here, the classifier is most uncertain. That is why samples near the decision boundary are much more informative than samples far away. The ratio between samples of high and low informativeness in a training set can have a strong influence on the resulting detection quality. The samples extracted in the previous step are, in addition to the class imbalance stated above, likely to contain a large amount of uninformative instances.

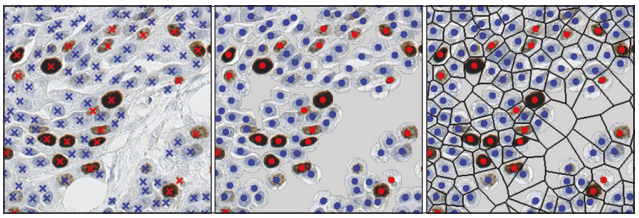


Figure 1: Visualization of the sample extraction methods. The left image shows the original image with overlayed center marker annotations. The center image shows the positions, where nonnuclear samples are extracted in gray and those where nuclear samples are extracted in red and blue for positively and negatively stained nuclei, respectively. The right image additionally shows the Voronoi boundaries in black, where also nonnuclear samples are extracted in the Voronoi-based extraction method

We investigated three different sample selection methods addressing the described issues.

Stratified random subsampling

The most straightforward way to reduce the amount of samples and to achieve a balance of the class labels is stratified random subsampling. From each class, samples are randomly drawn until a target number is reached or until there are no more samples of one class available. This method has the advantage that it can be integrated into the sample extraction methods. The subsampling can be already applied to the image positions before the features are calculated. This leads to a much better runtime performance than subsampling the samples in a separate step afterward.

Kd-tree subsampling

In the study by Pechenizkiy *et al*.,[13] the Kd-tree subsampling is suggested as an alternative or supplementary method for stratified random subsampling. It also reduces the number of samples while retaining their distribution in feature space. The general concept of the Kd-tree is explained by Bentley.[14] For the sample selection task, a Kd-tree with limited depth is constructed on the extracted samples using their features as dimensions. In each node, the splitting feature is chosen as that with maximum variance across the samples of the node and the median is used as pivot, as suggested by Omohundro.[15] Then, a single sample can be drawn randomly from each leaf of the tree. The granularity and the amount of resulting samples can be controlled by adjusting the depth limit of the tree. To also address class imbalance, we apply the Kd-tree subsampling independently for both classes and join the sample sets afterward.

Active learning

Active learning[16] selects samples with respect to their informativeness to the classifier. A classifier is trained using a subset S of the available samples. Then, iteratively, the remainder of the samples is classified and the classification confidence for each sample is considered. The samples with the least confident classifications are added to S for the next iteration. The iterations are terminated as soon as the size of S reaches a target number. By following this uncertainty sampling approach, the most informative samples are chosen from the training set. In our implementation, to produce a training set of n samples, the first subset is generated by randomly choosing $n/10$ samples from the available samples and $n/100$ samples are added in each iteration. In contrast to the previous methods, active learning does not address any class imbalance of the sample set.

The training sample selection methods are applied to either the samples extracted from a single image or the whole set of extracted samples. They can also be combined to utilize their different strengths.

Experimental setup

The different sample extraction and selection methods were compared using image data from a study described by Molin *et al*.[17] In that study, eight pathologists were asked to select

circular hot-spot regions containing approximately 200 nuclei from digitized Ki-67-stained breast tumor slides. From these hotspots, areas containing staining or scanning errors as well as overlapping areas were removed resulting in a set of 101 hot-spots from 24 different slides and cases. The digitized slides were downsampled if necessary to a magnification of 20×, and for each hotspot, a subimage containing that region was extracted. Center marker annotations for all nuclei within the circular hot-spot regions were created by a trained expert and verified by an experienced breast pathologist. Figure 2 examplary shows annotated hot-spot regions.

The evaluation is based on the nuclei detection method described by Kost *et al*.[18] A random forest assigns a probability value to each input image pixel for being close to the center of a nucleus. The feature set comprises:

- The normalized H, S, and V color channels
- The box filtered S channel
- An approximation of the difference of Gaussian on the S channel using box filters
- The radial symmetry on the S channel
- The box filtered radial symmetry.

Then, an optimized gray scale watershed algorithm is used to find and separate the individual nuclear regions. The algorithm is configured to only include positions with probability values above 0.5 into the nuclear regions as lower values indicate that it is more likely that the position belongs to background than to a nucleus. Another random forest uses the H, S, and V color channels to classify the staining within the nuclear regions and performs a majority vote to decide whether a nucleus is Ki-67 positive or negative.

To train the second classifier, we used modified versions of the sample extraction methods. For each position x, y with $d(x, y) < t_{nuc}$, an additional training sample for the second classifier was generated. The class label of the training sample was set depending on whether m (x, y) corresponds to a center marker of a Ki-67 positive or negative nucleus. This way, one training set was produced for each classifier. The selection methods were then applied to both sets individually using the same parameters.

The quality of the nuclei detection was assessed by comparing the results to the center marker annotations. Each detected nucleus was assigned to the closest center marker, provided that the distance of their positions was sufficiently close. A threshold of 10 pixels was found to be adequate. It corresponds to the approximate radius of the nuclei in the images. A one-to-one match was then considered a true positive (TP), a detected nucleus without a matching annotation was considered a false positive (FP), and an annotation without a matching detected nucleus considered a false negative (FN). In case when multiple detected nuclei were matched with the same center marker, one of these was considered TP whereas the others were counted as FP. Based on these values, precision, recall, and the f1-measure were computed as overall quality measures.

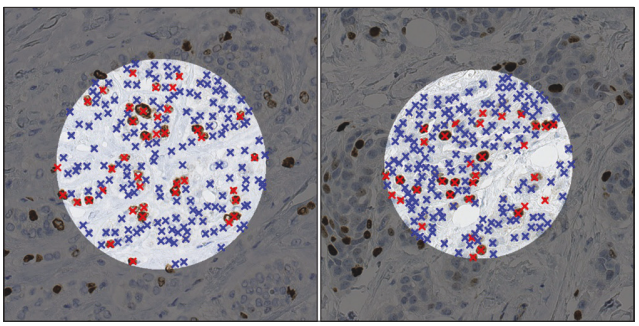


Figure 2: Visualization of the center marker annotations for two different images. The circle is scaled to contain approximately 200 nuclei. Inside the circle, all nuclei are annotated

For the experiments, we combined the sample extraction and selection methods in several ways to produce different training sets. The nuclei detection algorithm was then trained using these sample sets and the quality of the detection was assessed. To produce more robust results, the experiments were performed with 5-fold cross-validation. The folds were created in a way that images originating from the same slide were assigned to the same fold. This way, no training set is tested on the same slide it is created from. For all experiments, the same folds were used to ensure comparability. The quality measures of the individual folds were averaged to obtain the final measures.

Results

Experiment 1: Comparison of sample extraction and selection combinations

For the following experiment setups, distinct sample sets have been extracted by the distance-based and the Voronoi-based method. Then, different combinations of sample selection methods were applied to these sample sets. To obtain comparable results, all experiments, but (1a), which incorporates no selection method, produce a training set containing 2000 samples. This amount was found to produce adequate results while keeping the processing time for the classifier at an acceptable level. For the experiments involving two selection methods, the first one was applied per input image, leaving 256 samples per image. The latter one then was applied to the total of the remaining samples.

(a) No selection: As a base experiment, the outputs of the extraction methods were directly used to train the nuclei detection algorithm
(b) Random: The extracted samples were, as a whole, subjected to the stratified random subsampling selection method
(c) Kd-tree: The extracted samples were, as a whole, subjected to the Kd-tree subsampling selection method
(d) AL: The extracted samples were, as a whole, subjected to the active learning selection method
(e) Random + AL: The random subsampling selection method was applied per image. The final training set was then selected from the remaining samples using the active learning selection method
(f) Kd-tree + AL: The Kd-tree subsampling selection method was applied per image. The final training set was then

selected from the remaining samples using the active learning selection method

(g) AL + random: The effect of inverting the order of experiment (e) was examined. The active learning selection method was applied per image, followed by the stratified random subsampling selection method

(h) AL + Kd-tree: The effect of inverting the order of experiment (f) was examined. The active learning selection method was applied per image, followed by the Kd-tree subsampling selection method

(i) Kd-tree + random: In this experiment, both class balancing methods were combined. The Kd-tree subsampling selection method was first applied per image, and then the stratified random subsampling selection method was applied on the remaining samples afterward

(j) AL + AL: In this experiment, the active learning selection method was applied to both the per-image and the remaining samples.

Table 1 shows the results of the described experiment setups. First of all, we can state that the Voronoi-based extraction method yields quality measures slightly superior to the distance-based method in most experiment setups. Looking at the selection methods, we can see that the experiments that do not comprise a class balancing method lead to far worse quality measures. This can be observed in experiments (1d) and (1j), which only consist of active learning, and especially in experiment (1a), where no selection is performed at all. The best results are obtained by the combinations that include class balancing and active learning.

The tested sample extraction methods produce highly imbalanced training sets. On average, only 6.09% or 5.54% of the samples belong to the nuclear class for the distance-based and Voronoi-based extraction, respectively. The imbalance affects the resulting classification in a negative way. This can be observed in experiment (1a). The detection quality for the unprocessed training sets is low.

The usage of active learning alone, as shown in experiments (1d) and (1j), does improve the detection quality slightly but still yields results well inferior to other experiments. This indicates that active learning is not very well suited to deal with these large imbalances, which stems from the way active learning selects new samples. When there are mostly nonnuclear samples to choose from, the most uncertain samples are likely to be imbalanced toward nonnuclear samples as well. For this reason, a proper balancing of the samples is advisable.

The absence of class balancing in experiments (1a), (1d), and (1j) results in a strong bias of the classifier, which can be observed as a considerable difference in the precision and recall values. In experiment 2, precision-recall-curves (PR-curves) are analyzed to further examine this issue.

The stratified random subsampling and the Kd-tree-based selection seem to be equally suited for balancing as

Table 1: Quality measures of both proposed sample extraction methods combined with different sample selection methods

	Ki-67-positive nuclei			Ki-67-negative nuclei			All nuclei					
	TP	FP	FN	TP	FP	FN	TP	FP	FN	Precision	Recall	f1-measure
Distance-based												
(a) No selection	650.8	74.0	189.8	803.6	239.2	2024.0	1454.4	313.2	2213.8	0.823	0.396	0.530
(b) Random	716.8	156.8	123.8	2220.6	500.6	607.0	2937.4	657.4	730.8	0.817	0.801	0.806
(c) Kd-tree	721.6	174.0	119.0	2203.0	509.4	624.6	2924.6	683.4	743.6	0.811	0.797	0.801
(d) AL	688.6	66.2	152.0	1658.0	232.0	1169.6	2346.6	298.2	1321.6	0.887	0.640	0.740
(e) Random + AL	732.8	161.2	107.8	2290.4	556.2	537.2	3023.2	717.4	645.0	0.808	0.824	0.814
(f) Kd-tree + AL	732.6	151.2	108.0	2246.6	535.0	581.0	2979.2	686.2	689.0	0.813	0.812	0.810
(g) AL + random	706.2	118.6	134.4	2176.8	451.4	650.8	2883.0	570.0	785.2	0.835	0.786	0.807
(h) AL + Kd-tree	715.8	107.4	124.8	2202.8	460.4	624.8	2918.6	567.8	749.6	0.837	0.796	0.813
(i) Kd-tree + random	718.2	192.2	122.4	2188.6	516.6	639.0	2906.8	708.8	761.4	0.804	0.792	0.795
(j) AL + AL	683.4	76.6	157.2	1964.2	369.4	863.4	2647.6	446.0	1020.6	0.856	0.722	0.781
Voronoi-based												
(a) No selection	611.8	66.6	228.8	585.0	151.4	2242.6	1196.8	218.0	2471.4	0.846	0.326	0.467
(b) Random	746.0	206.4	94.6	2333.6	558.4	494.0	3079.6	764.8	588.6	0.801	0.840	0.817
(c) Kd-tree	750.4	192.6	90.2	2294.2	506.4	533.4	3044.6	699.0	623.6	0.813	0.830	0.819
(d) AL	628.8	50.8	211.8	1535.0	208.0	1292.6	2163.8	258.8	1504.4	0.893	0.590	0.711
(e) Random + AL	761.8	192.6	78.8	2364.8	561.8	462.8	3126.6	754.4	541.6	0.806	0.852	0.826
(f) Kd-tree + AL	763.2	188.0	77.4	2364.2	576.0	463.4	3127.4	764.0	540.8	0.804	0.853	0.825
(g) AL + random	728.8	119.4	111.8	2210.6	458.6	617.0	2939.4	578.0	728.8	0.836	0.801	0.815
(h) AL + Kd-tree	726.8	120.4	113.8	2195.8	444.0	631.8	2922.6	564.4	745.6	0.838	0.797	0.814
(i) Kd-tree + random	744.6	176.6	96.0	2314.8	541.4	512.8	3059.4	718.0	608.8	0.810	0.834	0.819
(j) AL + AL	694.8	91.4	145.8	1734.4	288.2	1093.2	2429.2	379.6	1239.0	0.865	0.662	0.747

TP: True positive, FP: False positive, FN: False negative, AL: Active learning

the comparison of the quality measures in experiments (1b) and (1c), (1e) and (1f), as well as (1g) and (1h) indicates. However, since stratified random subsampling is much simpler and improves the runtime performance when integrated into the extraction step, it is to be preferred over the Kd-tree-based approach. The best results were achieved by experiment setup (1e) being the combination of Voronoi-based extraction, stratified random subsampling, and active learning. Two example outputs are visualized in Figure 3. Another interesting approach is (1b), the solely applied stratified random subsampling. It is simple, yields good results, and has a good runtime performance due to the integrability into the sample extraction step. However, in general, the differences of the methods that use balancing are rather small. In contrast, the differences between the methods with and without balancing are major.

Experiment 2: Precision-recall-curves

As described in section 2.2, a cutoff value of 0.5 was used for the experiments, which is the natural threshold for a two-class problem. However, it is interesting to investigate how different cutoff values influence precision and recall.

For each approach described in experiment 1, the cutoff value was altered in 16 steps between 0 and 1, and at each step, precision and recall were determined. Figure 4 shows the PR-curves for all approaches in an overview graph. In the subsequent graphs, the curves are reduced and grouped to highlight different aspects. Furthermore, the axes are scaled to only show the most interesting quadrant of the graph.

In Figure 5, the PR-curves are divided into approaches that contain a sample selection method and approaches that do not, which is only the case for (1a) curves. It is clearly visible that the application of even the most basic sample selection methods improves the quality of the nuclei detection considerably. This is the case for both sample extraction methods.

Figure 6 shows the approaches that contain sample selection grouped according to their sample extraction methods. Here, it becomes apparent that the Voronoi-based extraction leads to better results than the distance-based extraction. This is especially the case for recall.

In Figure 7, only the approaches using the Voronoi-based sample extraction are plotted. We found that approaches consisting of two subsequent sample selection methods with at least one of them incorporating active learning perform especially well. These approaches are highlighted in this figure. Active learning per image followed by a selection that performs class balancing (1g) and (1h) leads to the best results for both extraction methods.

The PR-curves shown in this section have an unusual shape. Normally, with cutoff values becoming lower, the precision declines while the recall grows toward 1. In our case, the recall does not increase after a certain value but decreases again. The reason for this behavior is the watershed algorithm which is part of the nuclei detection method. This limits the number of detected nuclei. With a low cutoff value, more pixel positions

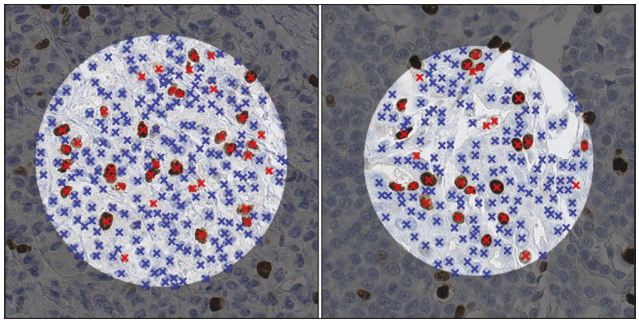


Figure 3: Example results of experiment (1e). The red and blue markers show Ki-67 positive and negative nuclei as detected by the algorithm, respectively

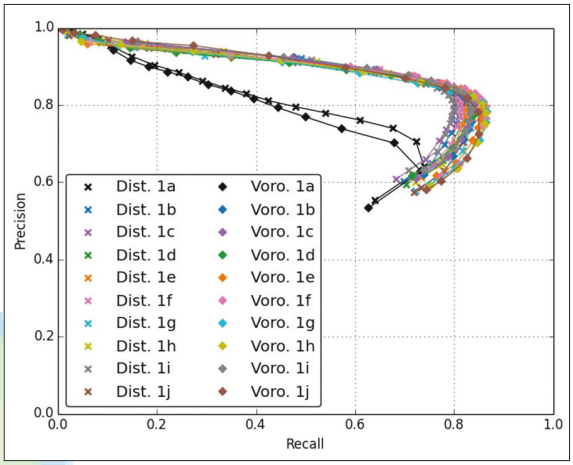


Figure 4: Overview plot of the precision-recall-curves for all training approaches and both distance-based (dist.) and Voronoi-based (voro.) sample extraction. The labels 1a–1j correspond to the notation in experiment 1

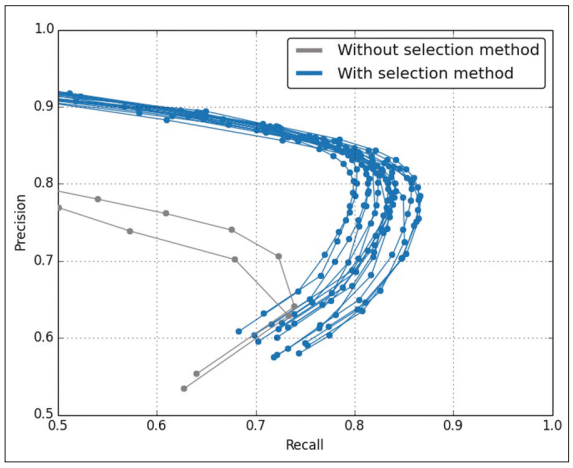


Figure 5: Precision-recall-curves showing the quality improvements when using a sample selection method (blue) compared to approaches without sample extraction (gray)

are being considered by the algorithm. Nevertheless, those are likely to be assigned to an existing nuclear region instead of constituting a new region. Another effect is that the nuclear

regions, which are segmented by the watershed algorithm, become larger. The nuclear positions are computed as the center points of the nuclear regions and the positivity of the nuclei is derived from the staining classification results within the nuclear regions. When such regions become unreasonably large, nuclear positions or their positivity might become incorrect. This effect causes the decrease of recall at low cutoff values.

Experiment 3: Impact of the training set size

The impact of the training set size on the quality of the nuclei detection was evaluated in experiment 3. Training sets of different sizes were produced using the Voronoi-based extraction method, followed by a selection method as described in (1b) and (1e), which appeared to be the most interesting approaches in experiment 1. For the latter approach, the active learning was parametrized to select 10% of the samples selected by the stratified random subsampling, which is comparable to the ratio in the above experiments. The overall f1-measure of the training sets was then assessed to compare the learning curves of these two methods. Training set sizes from 100 up to 5000 samples have been evaluated with an offset of 100 and up to 15,000 samples with an offset of 1000.

Figure 8 shows the results of experiment 3. Both learning curves have an approximately asymptotic shape. They rise steeply until about 2000 samples and ascend more slowly afterward. Nevertheless, the learning curve for the approach containing active learning shows superior quality values throughout all training set sizes. The experiment shows that the number of 2000 samples for a training set is a reasonable choice. Although more samples would slightly increase the quality of the nuclei detection, we consider this a good compromise between quality and runtime performance.

Experiment 4: Comparison of extraction methods with manual segmentations

To assess the quality of the sample extraction methods, we compared a manual nuclei segmentation with the proposed distance and Voronoi-based extraction methods. To produce a training set from segmentation annotations, nonnuclear samples were generated from all positions outside nuclear regions. Since the trained method should yield maximum nucleus probability at the center of the nuclei, the nuclear samples were only generated at the centers of the nuclei, equally to the extraction methods proposed.

As the selection method, we used stratified random subsampling per image followed by active learning (1e), which appeared to achieve the best results in experiment 1. We used ten images with exhaustive nuclei segmentation annotations which do not belong to the image set used in experiment 1–3. Samples were extracted from the annotations for each pixel. To compare those samples with the described extraction methods, the segmentation annotations were reduced to center markers by computing the center of gravity of each segment. For this experiment, we did not perform a cross-validation but tested the resulting training sets using the image set described above. The

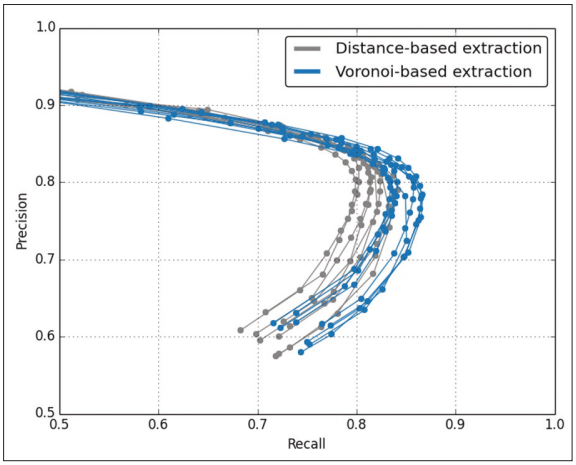


Figure 6: In most cases, the approaches using Voronoi-based sample extraction (blue) lead to better detection quality than those using distance-based sample extraction (gray)

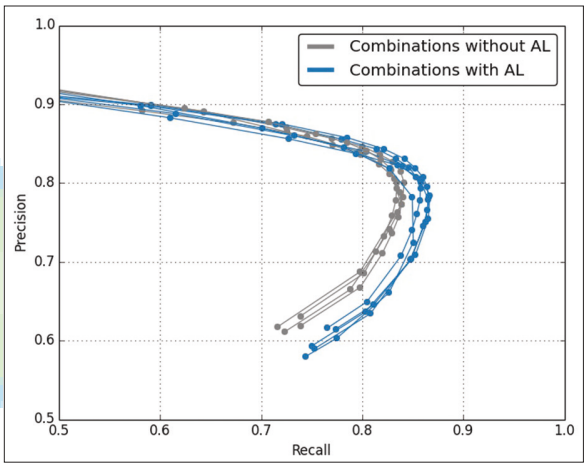


Figure 7: Most approaches that combine multiple selection methods and contain active learning (blue) yield better quality measures than others (gray), here shown for the approaches using Voronoi-based sample extraction

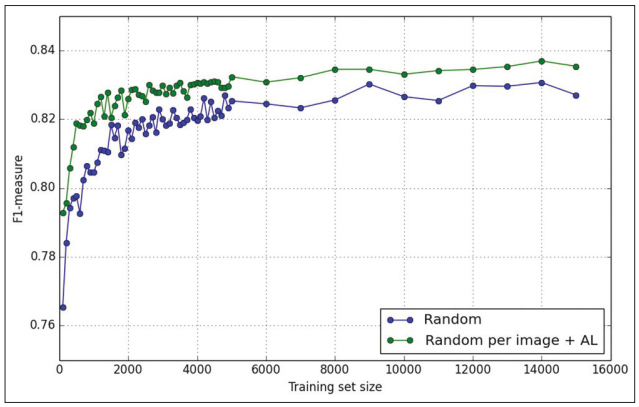


Figure 8: In experiment 3, training sets of different sizes were produced for the approaches described in experiment (1b) and (1e). The graph plots the f1-measures of the nuclei detection trained with these sets against their size

experiment was repeated five times to average out randomness in the used algorithms.

Table 2 shows the results of experiment 4. It appears that the manual extraction yields only marginally better results than the approaches based on center markers. Since the quality of the segmentation-based samples is certainly better than of those of the proposed methods, the low difference of the methods might seem unexpected. However, the produced training sets are all subject to the same selection process so that these quality differences manifest less noticeable in the final results.

Conclusions

The quality of machine learning-based nuclei detection methods is fundamentally dependent on the training data used. Ideally, training data should be generated from manual segmentations of a large number of nuclei, which is, however, a time-consuming and tedious task.

In this paper, we proposed and compared approaches to produce training sets from easy to generate center marker annotations. We divided the training set generation into a sample extraction and a sample selection step. The samples were extracted using a distance-based or a Voronoi boundary-based method. Training sets were selected from the resulting sample sets using different combinations of stratified random subsampling, Kd-tree subsampling, and active learning.

For evaluation, we trained a nuclei detection method with these training sets and assessed the resulting detection quality measures. In addition, we investigated the influence of the cutoff value on these measures. For a cutoff value of 0.5, the default threshold for two-class problems, class balancing had the largest positive impact on the detection quality. Independent of the cutoff value, the best results were obtained using training sets produced by Voronoi-based sample extraction and sample selection methods that incorporate active learning. We also evaluated the influence of the training set size on the detection quality. The quality increased quickly until approximately 2000 samples and more moderate afterward. In a fourth evaluation, a comparison revealed that the f1-measures obtained using the proposed extraction methods almost reached the values obtained using samples generated from manual segmentations.

We conclude that the usage of center marker annotations in conjunction with appropriate sample extraction and selection methods represents a valid alternative to conventionally produced training sets. In this manner, the effort for the creation of annotations can be greatly reduced.

In addition, while machine learning-based nuclei detection methods are usually trained on all training samples available, our study shows that subselecting samples can improve the detection quality considerably at no additional cost in terms of execution time or complexity of the nuclei detection method.

In future work, we will further evaluate the general applicability of the proposed approaches using different image datasets and detection algorithms, especially within the area of deep learning.

Table 2: Quality measures obtained by different extraction methods

Extraction	Precision	Recall	F1-measure
Distance-based	0.791	0.818	0.804
Voronoi-based	0.783	0.831	0.806
Manual segmentation	0.787	0.829	0.807

Financial support and sponsorship

This work was financially supported by Sectra AB, Linköping, Sweden. Part of this work was conducted under the QuantMed project funded by the Fraunhofer Society, Munich, Germany.

Conflicts of interest

There are no conflicts of interest.

References

1. Mahmoud SM, Paish EC, Powe DG, Macmillan RD, Grainge MJ, Lee AH, *et al.* Tumor-infiltrating CD8+ lymphocytes predict clinical outcome in breast cancer. J Clin Oncol 2011;29:1949-55.
2. Speirs V, Walker RA. New perspectives into the biological and clinical relevance of oestrogen receptors in the human breast. J Pathol 2007;211:499-506.
3. Tang LH, Gonen M, Hedvat C, Modlin IM, Klimstra DS. Objective quantification of the Ki67 proliferative index in neuroendocrine tumors of the gastroenteropancreatic system: A comparison of digital image analysis with manual methods. Am J Surg Pathol 2012;36:1761-70.
4. Vink JP, Van Leeuwen MB, Van Deurzen CH, De Haan G. Efficient nucleus detector in histopathology images. J Microsc 2013;249:124-35.
5. Arteta C, Lempitsky V, Noble JA, Zisserman A. Learning to detect cells using non-overlapping extremal regions. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2012. Berlin, Heidelberg: Springer; 2012. p. 348-56.
6. Kårsnäs A, Dahl AL, Larsen R. Learning histopathological patterns. J Pathol Inform 2011;2:S12.
7. Xing F, Xie Y, Yang L. An automatic learning-based framework for robust nucleus segmentation. IEEE Trans Med Imaging 2016;35:550-66.
8. Xu J, Luo X, Wang G, Gilmore H, Madabhushi A. A Deep Convolutional Neural Network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing 2016;191:214-23.
9. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J Pathol Inform 2016;7:29.
10. Gul-Mohammed J, Arganda-Carreras I, Andrey P, Galy V, Boudier T. A generic classification-based method for segmentation of nuclei in 3D images of early embryos. BMC Bioinformatics 2014;15:9.
11. Sirinukunwattana K, Raza S, Tsang YW, Snead D, Cree I, Rajpoot N. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans Med Imaging 2016;35:1196-206.
12. Chawla NV, Japkowicz N, Kotcz A. Editorial: Special issue on learning from imbalanced data sets. SIGKDD Explor Newsl 2004;6:1-6.

13. Pechenizkiy M, Puuronen S, Tsymbal A. The impact of sample reduction on PCA-based feature extraction for supervised learning. In: Proceedings of the 2006 ACM Symposium on Applied Computing (SAC). ACM: New York, USA; 2006. p. 553-8.
14. Bentley JL. Multidimensional binary search trees used for associative searching. Commun ACM 1975;18:509-17.
15. Omohundro SM. Efficient algorithms with neural network behavior. Complex Syst 1987;1:273-347.
16. Settles B. Active Learning Literature Survey. Computer Sciences Technical Report 1648. University of Wisconsin-Madison: Madison, USA; 2009.
17. Molin J, Bodén A, Treanor D, Fjeld M, Lundström C. Scale Stain: Multi-Resolution Feature Enhancement in Pathology Visualization, ArXiv Prepr. arXiv:1610.04141; 2016.
18. Kost H, Homeyer A, Bult P, Balkenhol MC, van der Laak JA, Hahn HK. A generic nuclei detection method for histopathological breast images. In: Medical Imaging 2016: Digital Pathology. Bellingham; Washington USA; 2016;9791:97911E-1 - 97911E-7.

Paper 3

Deep learning nuclei detection: A simple approach can deliver state-of-the-art results

Henning Höfener, André Homeyer, Nick Weiss, Jepser Molin, Claes F. Lundström, Horst K. Hahn.

Computerized Medical Imaging and Graphics 70, 43–52. 2018.

The original publication is available at:
https://doi.org/10.1016/j.compmedimag.2018.08.010

Computerized Medical Imaging and Graphics 70 (2018) 43–52

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Deep learning nuclei detection: A simple approach can deliver state-of-the-art results

Henning Höfener[a,*], André Homeyer[a], Nick Weiss[a], Jesper Molin[b], Claes F. Lundström[b,c], Horst K. Hahn[a,d]

[a] Fraunhofer MEVIS, Am Fallturm 1, 28359, Bremen, Germany
[b] Sectra AB, Teknikringen 20, 58330, Linköping, Sweden
[c] Center for Medical Image Science and Visualization, Linköping University, 58183, Linköping, Sweden
[d] Jacobs University, Campus Ring 1, 28759, Bremen, Germany

ARTICLE INFO

Article history:
Received 8 February 2018
Received in revised form 13 July 2018
Accepted 23 August 2018

Keywords:
Nuclei detection
Deep learning
PMap
Histology
Image analysis

ABSTRACT

Background: Deep convolutional neural networks have become a widespread tool for the detection of nuclei in histopathology images. Many implementations share a basic approach that includes generation of an intermediate map indicating the presence of a nucleus center, which we refer to as PMap. Nevertheless, these implementations often still differ in several parameters, resulting in different detection qualities.
Methods: We identified several essential parameters and configured the basic PMap approach using combinations of them. We thoroughly evaluated and compared various configurations on multiple datasets with respect to detection quality, efficiency and training effort.
Results: Post-processing of the PMap was found to have the largest impact on detection quality. Also, two different network architectures were identified that improve either detection quality or runtime performance. The best-performing configuration yields f1-measures of 0.816 on H&E stained images of colorectal adenocarcinomas and 0.819 on Ki-67 stained images of breast tumor tissue. On average, it was fully trained in less than 15,000 iterations and processed 4.15 megapixels per second at prediction time.
Conclusions: The basic PMap approach is greatly affected by certain parameters. Our evaluation provides guidance on their impact and best settings. When configured properly, this simple and efficient approach can yield equal detection quality as more complex and time-consuming state-of-the-art approaches.

1. Introduction

The quantification of cell nuclei in histological images is essential for many pathological assessments, including the determination of various biomarkers. Prominent examples in cancer diagnosis are the Ki-67 index or the progesterone and estrogen receptor status. Detecting nuclei also enables the quantification of tumor immune infiltrates, which have been shown to be of strong prognostic importance (Mahmoud et al., 2011), and are commonly assessed in immunotherapy trials (Denkert et al., 2016).

Such assessments are usually performed by visual estimation, which is labor- and time-intensive and can lead to high inter- and intra-observer variability (Andrion et al., 1995). The ongoing digitalization in pathology allows for automated analysis methods to support pathologists at such tasks and to increase the reliability of quantitative assessments.

However, the automatic detection of cell nuclei is challenging. The appearance of nuclei varies considerably with staining and tissue preparation conditions, as well as with different nuclear types and pathologies.

The first attempts to automate nuclei detection date back to the mid-1950s (Meijering, 2012), starting with static rule-based approaches from simple intensity thresholds to using intensity-derived features. Those approaches suffered from not being able to capture the complexity of the input data sufficiently well. The next generation of methods addressed the aforementioned variability by using hand-crafted features and applying machine-learning to build more complex and flexible rule sets (Arteta et al., 2012;

* Corresponding author.
E-mail addresses: henning.hoefener@mevis.fraunhofer.de (H. Höfener), andre.homeyer@mevis.fraunhofer.de (A. Homeyer), nick.weiss@mevis.fraunhofer.de (N. Weiss), Jesper.Molin@sectra.com (J. Molin), claes.lundstrom@liu.se (C.F. Lundström), horst.hahn@mevis.fraunhofer.de (H.K. Hahn).

https://doi.org/10.1016/j.compmedimag.2018.08.010

Kårsnäs et al., 2011; Vink et al., 2013). Recent developments mainly employ convolutional neural networks (CNN) (Jacobs et al., 2017; Janowczyk and Madabhushi, 2016; Sirinukunwattana et al., 2016; Wang et al., 2016; Xie et al., 2016, 2015a,b; Xing et al., 2016), as those tend to yield significantly better results. The first major breakthrough was reported by Cireşan et al. (2013), who were able to detect mitotic nuclei with an f1-measure of 0.782, while the closest competitors achieved 0.718.

Most deep learning–based nuclei detection methods employ CNNs to predict a value for each input image pixel. That value represents the proximity to a nucleus center or the probability of being close to one. Together, the values of all input image pixels constitute a map, which we refer to as PMap. Nuclei positions are afterwards determined by finding local maxima in the PMap. Predicting the PMap can be interpreted as either a classification or a regression problem. The classification problem is to distinguish *nucleus center* and *background* positions and to populate the PMap with the each position's probability to belong to the *nucleus center* class, whereas the regression problem is to map a position to a continuous value, which is dependent on the distance to the nearest nucleus center. This basic PMap approach will be described in more detail in Section 2.1.

The basic PMap approach is controlled by several parameters. Examples are the post-processing of the PMap before finding local maxima, the use of data augmentation or the use of dropout.

1.1. Related work

There are different variants of the basic PMap approach proposed in the literature, using CNN classification or regression, even if that term is not used.

As described above, Cireşan et al. (2013) have used CNN classification for the detection of mitoses. Their approach uses a 12 and a 10 layer deep network and achieves processing speeds between 0.01 and 0.03 megapixels per second at prediction time. CNN classification with the 8 layer deep AlexNet (Krizhevsky, 2010) has been used by Janowczyk and Madabhushi (2016) to detect lymphocytes in breast cancer images. They have achieved an f1-measure of 0.900. The 7 layer deep LeNet (LeCun et al., 1998) classification network has been applied by Wang et al. (2016) to detect nuclei for a subsequent cell subtype classification. For the detection, they have reported an f1-measure of 0.822. Khoshdeli et al. (2017) have used a 5 layer deep CNN classification for the detection of nuclei in Hematoxylin and Eosin (H&E) stained images of various tissue types. They have proposed to preprocess the input images by extracting the Hematoxylin channel using color deconvolution and applying a Laplacian of Gaussian filter. The result is then fed into to the network. An f1-measure of 0.722 has been reported. Jacobs et al. (2017) have used a 14 layer deep regression network to detect nuclei in H&E stained prostate cancer biopsies for a subsequent nucleus type classification. The authors have evaluated transfer learning for the application with limited training data. They have trained on colon images and have fine-tuned their model with the prostate images. They have reported f1-measures between 0.849 and 0.864, depending on the amount of training data for the fine-tuning, as well as a processing speed of 2.2 megapixels per second.

Some approaches leave out the extraction of local maxima from the PMap. Xie et al. (2016) have estimated the nuclei count in an image region by integrating the PMap over that region. They have applied a 9 layer deep network. Xing et al. (2016) have applied a threshold to the PMap and have used the connected regions as initialization for nuclei segmentation. The generation of the PMap has been performed with 0.008 megapixels per second.

In other publications, the basic PMap approach has been used as a baseline algorithm to compare the proposed methods with. Xie et al. (2015a) have mapped each pixel of the input image to a 2D-vector pointing to the nearest nucleus center, using an 8 layer deep network. At prediction time, the positions, where the vectors point to, are accumulated to form a PMap. On Ki-67 stained neuroendocrine tumor (NET) images they have reported an f1-measure of 0.815 and a processing speed of 0.007 megapixels per second. They have compared their method with the basic PMap approach using CNN classification, which has yielded an f1-measure of 0.784. In another publication, the same authors have used a 7 layer deep network to predict a small region of the PMap at once instead of a single pixel value (Xie et al., 2015b). As before, they have accumulated the predictions to generate the PMap. They have evaluated the approach on H&E stained breast tumor images, Ki-67 stained NET images and phase contrast images of HeLa cervical cancer cells. F1-measures of 0.913, 0.906 and 0.957 have been reported, respectively. Processing speed has been 0.01 megapixels per second. A comparison with both CNN classification and regression according to the basic PMap approach has been conducted, but no quantitative measures have been given. Sirinukunwattana et al. (2016) have proposed a similar approach of predicting a region of the PMap. In contrast to (Xie et al., 2015b), the first 6 layers of their network are followed by a parameter estimation layer and a spatially constrained regression layer. They have reported an f1-measure of 0.802 on H&E stained images of colorectal adenocarcinomas and processing speed of 0.02 megapixels per second. They have compared their method with the basic PMap approach using CNN regression, for which an f1-measure of 0.692 has been reported. Xu et al. (2016) have used multiple stacked auto-encoders to learn feature representations of the input images in an unsupervised manner. The features are then fed into a softmax classifier, which classifies each input patch as either nuclear or non-nuclear. The softmax classifier has been trained supervisedly and the authors have reported an f1-measure of 0.845 on H&E stained breast cancer images. They compare their method with the basic PMap approach using CNN classification. There, an f1-measure of 0.820 and processing speed of 0.04 megapixels per second have been reported.

We want to stress here that the reported f1-measures should not be compared directly. Most approaches have been evaluated using different datasets with different nuclear types, varying quality and tissue complexity. Additionally, the hardware used to perform the experiments, especially the usage of GPUs, has a great influence on the processing speed. Although only comparable to a very limited extent, we listed processing speeds if available for completeness. Only few of the approaches above explicitly focused on processing time, although speed is critical when aiming at applying these methods in clinical routine.

In summary, for nuclei detection using deep learning, the basic PMap approach is widely used in the literature. Even if not termed basic PMap approach, numerous publications describe such methods either as the proposed or as alternative solutions for nuclei detection tasks. However, there are some parameters of these methods that differ from case to case. Most of the papers above only present a single configuration of them. There is no systematic evaluation of the influence and importance of the individual parameters.

The main contribution of this work is a systematic listing, evaluation and comparison of these parameters. We assess the impact of the individual parameters with respect to detection quality, efficiency and training effort. By doing so, we give guidance on which parameters to focus on when optimizing nuclei detection with the basic PMap approach. The second contribution is to combine those parameter settings, which perform best in our experiments and to evaluate this configuration. We show that the basic PMap approach delivers state-of-the-art results when parameterized well.

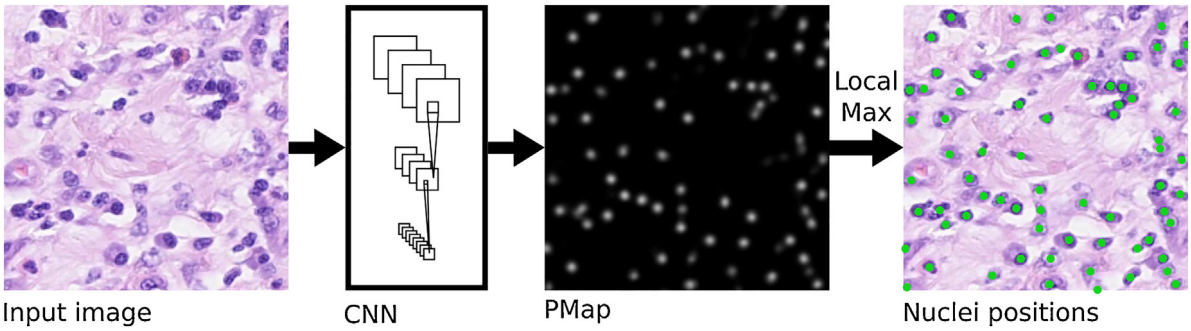


Fig. 1. The prediction workflow of the basic PMap approach.

2. Materials and methods

2.1. Basic PMap approach

In the vast majority of nuclei detection methods based on machine-learning, each pixel of an input image is assigned a value representing either the probability of being close to a nucleus center or the proximity to the nearest one. The entirety of these values for one input image is often termed probability map or proximity map and can be represented as a gray-scale image. While these two terms are semantically different, they share the same core properties: Intensities are high (towards 1.0) near nuclei center positions, lower at their periphery and lowest (towards 0.0) in background areas. In this paper, we refer to it as PMap.

When using PMaps, a popular and very straightforward approach is to first extract either a set of hand-crafted features or to crop an image patch for each input pixel. In case of CNNs, patches of the same extent as the network's receptive field are extracted. Secondly, reference or target PMap values are generated from reference annotations. Both are then fed to a machine learning–algorithm to learn the mapping. This way, the algorithm can afterwards predict PMap values for yet unseen input image pixels.

Training a machine learning–algorithm to generate a PMap can be formulated as either a classification or a regression problem. Considering the classification problem, there are two classes, namely *nucleus center* and *background*. For training, the target takes the discrete value 1 or 0, representing the class. At prediction time, the probabilities for the *nucleus center* class are used as the PMap values. Regarding the regression problem, the proximity to the nearest nucleus center is used to generate continuous target values between 0 and 1. The predictions of the machine learning–algorithm are then directly used as PMap values.

Usually, after generating a PMap from an input image, the positions of the nuclei centers are determined by finding local maxima in the PMap that exceed a certain threshold. A local maximum in this context is a region of equal values where all neighboring pixels have lower values. That region may be as small as a single pixel. We refer to the procedure of CNN-based PMap generation, followed by maxima finding as the basic PMap approach, which is depicted in Fig. 1.

2.2. Evaluation metric

The quality of nuclei detection is assessed by comparing the resulting nuclei positions with annotated nuclei center markers. For each detected nucleus position, the nearest center marker annotation is found. If the distance between these positions is at most r_{nuc}, which is the approximate nuclei radius of the dataset, it is considered a match. Otherwise, the detected position is evaluated as false positive (FP). Afterwards, for each annotation, if there is exactly one match, it is evaluated as true positive (TP). If there is more than a single match, one match is a TP and all additional matches are FP. If there is no match at all for a reference annotation, it is evaluated as false negative (FN). From these values, we derive the f1-measure as an overall quality measure.

The borders of the input images need special treatment, as in these areas the receptive field of the CNN partially lies outside the image bounds. Hence, within a margin of half the receptive field size, no nucleus can be detected. We exclude those margins from evaluation. To determine all matches for an annotation marker, the entire region of r_{nuc} around that marker must be evaluated. This is not possible if the region is partially or completely located within that margin. Thus, all annotation markers with such regions and all detected nuclei inside such regions are excluded from evaluation. Fig. 2 visualizes the excluded annotation markers and areas.

2.3. Fully convolutional neural network

In the literature, most variants of the basic PMap approach for nuclei detection are trained and applied patch-wise (Cireşan et al., 2013; Janowczyk and Madabhushi, 2016; Khoshdeli et al., 2017; Sirinukunwattana et al., 2016; Xie et al., 2015a,b; Xing et al., 2016; Xu et al., 2016): A patch of certain size is cropped from the input image and fed into the CNN. The output of the CNN is a scalar value and is interpreted as the PMap value belonging to the center position of the patch. Usually, network architectures use one or two pairs of small convolutional layers followed by max-pooling layers. Afterwards, dense layers are applied. The last layer outputs either one or two scalar values depending on being a regression or classification network.

In clinical routine, processing time is a major limiting factor when considering the use of automated analyses. Fully convolutional networks (FCN) (Long et al., 2014) are a variant of CNNs that reduce processing time considerably by eliminating the large amount of redundancy that patch-based sliding window approaches exhibit. The basic idea of FCN is to interpret dense layers as convolutional layers that cover the entire input region of that layer. By doing so, size constraints for the input images are removed. Instead of patches, entire images can be processed by the FCN resulting in an output image which is equivalent to the output of a patch-based sliding window. Patch-wise CNN architectures can be converted into FCNs such that the resulting network is equivalent to the original CNN and returns the exact same results. (An exception is the upsampling FCN architecture, explained in Section 2.4.1)

The model architecture used in this paper is the conversion of the described patch-wise architecture into an FCN. We chose a receptive field size of 33×33 pixels, which covers most nuclei in our datasets. As proposed by Xie et al. (2016), we double the channel dimension after each max-pooling layer to compensate for the reduction of the spatial dimensions. The model architectures are shown in Table 1 and 2.

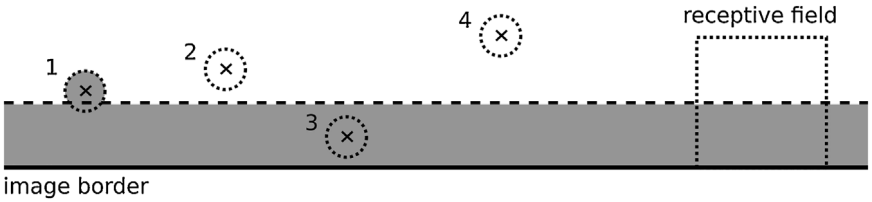


Fig. 2. Border handling for evaluation. Annotation markers 1 and 3 are excluded from evaluation. Detected nuclei in gray areas will be ignored.

Table 1
Architecture of the dilation FCN architecture. W, H are input image dimensions. N is number of output channels (1 for regression, 2 for classification).

#	Type	Filter Size	Dilation	Output Size
	Input			W × H × 3
1	Convolution	5 × 5	1 × 1	W - 4 × H - 4 × 32
2	Max-Pooling	2 × 2	1 × 1	W - 4 × H - 4 × 32
3	Convolution	3 × 3	2 × 2	W - 8 × H - 8 × 64
4	Max-Pooling	2 × 2	2 × 2	W - 8 × H - 8 × 64
5	Convolution	7 × 7	4 × 4	W - 32 × H - 32 × 128
6	Dropout 50%			W - 32 × H - 32 × 128
7	Convolution	1 × 1	4 × 4	W - 32 × H - 32 × 128
8	Dropout 50%			W - 32 × H - 32 × 128
9	Convolution	1 × 1	4 × 4	W - 32 × H - 32 × N

Table 2
Architecture of the upsampling FCN architecture. W, H are input image dimensions. N is number of output channels (1 for regression, 2 for classification).

#	Type	Filter Size	Strides	Output Size
	Input			W × H × 3
1	Convolution	5 × 5	1 × 1	W - 4 × H - 4 × 32
2	Max-Pooling	2 × 2	2 × 2	W/2 - 2 × H/2 - 2 × 32
3	Convolution	3 × 3	1 × 1	W/2 - 4 × H/2 - 4 × 64
4	Max-Pooling	2 × 2	2 × 2	W/4 - 2 × H/4 - 2 × 64
5	Convolution	7 × 7	1 × 1	W/4 - 8 × H/4 - 8 × 128
6	Dropout 50%			W/4 - 8 × H/4 - 8 × 128
7	Convolution	1 × 1	1 × 1	W/4 - 8 × H/4 - 8 × 128
8	Dropout 50%			W/4 - 8 × H/4 - 8 × 128
9	Convolution	1 × 1	1 × 1	W/4 - 8 × H/4 - 8 × N
10	Transposed Conv	8 × 8	4 × 4	W - 32 × H - 32 × N

The datasets for the training of the network consist of RGB images and associated center marker annotations. From the annotations, target PMaps are generated by applying a function $t(d)$ at each position, which is dependent only on the distance d to the nearest annotation marker. Along with the target PMap a weight map can be given to the training process, which the calculated loss is multiplied with. This way, positions in the input images can be weighted or masked out to be not considered during training.

2.3.1. Classification vs. regression network

Classification networks are trained to output class probabilities of the classes *nucleus center* and *background* for each input position. Therefore, the last layer of the FCN model has two output channels, one for each class, followed by a softmax activation function. The PMap is generated from the first channel of the activation's output.

For training, we use categorical cross-entropy loss and class balancing. The balancing is necessary as most of the pixels in the target PMaps belong to the *background* class. As described in (Long et al., 2014), with an FCN, class balancing can be achieved either by loss weighting or loss sampling. Both can be implemented using the weight map described in Section 2.3. For loss weighting, the weight map value for a position is chosen according to the frequency of that position's target class. For loss sampling, the weight map is set to 1.0 for all positions belonging to the *nucleus center* class and the same amount of positions belonging to the *background* class. For all further *background* positions, the weight is set to 0.0. In our experiments, loss weighting is used in order not to discard information.

In regression networks, the last layer only has a single output channel, which directly forms the PMap. Here, the mean squared error is used as the loss function. We experimented with balancing by performing kernel density estimation on the target PMap values and calculating the weights accordingly, which did not impact the performance of the approach. Therefore, we decided to exclude it from this paper.

Kainz et al. (2015) strongly advocate to formulate nuclei detection as a regression problem rather than as a classification problem. They compared both approaches using random forests.

2.3.2. Choice of best model

Usually, CNNs are trained for several epochs. After each epoch, the quality of the resulting model is assessed by determining one or multiple quality measures with an independent validation set. Finally, the model with the best quality is chosen as the result of the whole training process. A typical quality measure is the loss calculated on the validation data, which is used by several approaches (e.g. (Jacobs et al., 2017; Sirinukunwattana et al., 2016)). However, we eventually aim for an optimal f1-measure, which is not necessarily correlated with the loss. Thus, like in (Xu et al., 2016), we use the current network to generate PMaps of the validation images, determine nuclei positions and calculate the f1-measure after each epoch. This way, a reliable quality measure of the current model is obtained. Using an independent validation set suppresses the effect of overfitting. As soon as the model overfits the training data, quality on the validation data decreases.

When extracting the nuclear positions from the PMap, only those local maxima are considered, whose intensities exceed a certain threshold. The optimal threshold is highly dependent on the current network and the training data. Therefore, like (Cireşan et al., 2013; Janowczyk and Madabhushi, 2016; Sirinukunwattana et al., 2016), we do not use a fixed threshold but optimize it using the validation data such that the f1-measure is maximized. This is done for each epoch. The optimal threshold is stored alongside the network's weights and applied when determining the f1-measure on the test set.

2.4. Parameters

The basic PMap approach is governed by a number of parameters. The main goal of this paper is to identify such parameters and to evaluate their individual impact on detection quality, efficiency and training effort. In this section, the parameters are explained and the different settings are described, which will be compared against each other.

2.4.1. Fully convolutional network type

Many CNN architectures comprise layers with strides > 1, leading to a decrease of the spatial dimensions of that layer's output by a factor equal to the stride value. The most prominent example is the max-pooling layer, which is usually used with a stride of 2.

In our network, we want to generate an output value for each pixel in the input image (despite a margin described in Section 2.2).

Therefore, we have to compensate the decrease caused by the two max-pooling layers. To achieve this, two approaches are proposed in (Long et al., 2014), which we briefly describe here. The first is called filter rarefacation. In each layer, the stride is set to 1 in order to avoid the downsampling of the input. To maintain the original behavior of the network, the dilation rate of each layer is multiplied with the new dilation rate and the original stride of the previous layer. For our network architecture this leads to the architecture depicted in Table 1.

The second approach is to add an upsampling layer to the network to recover the original spatial dimensions of the PMap using interpolation. The upsampling factor is the product of all strides used throughout the network. For this, a transposed convolutional layer is appended to the network. We follow Long et al.'s proposal to initialize the kernel to perform bilinear interpolation but to leave the layer trainable. Of course, one must be aware that the conversion is not equivalent anymore when using the upsampling variant. Table 2 shows the resulting network. It can be seen that both architectures have equal output dimensions. They differ from the input image extent only by a constant margin. We compare both the dilation and the upsampling architecture.

2.4.2. Dropout

Using dropout is a common way to prevent the network from overfitting the training data. Dropout works by randomly muting neurons and thus forcing the network to learn multiple independent representations of patterns, intending to achieve better generalization of the network. In the literature, dropout is applied by some approaches (Sirinukunwattana et al., 2016; Xie et al., 2015a,b). Others (Janowczyk and Madabhushi, 2016; Xu et al., 2016) report that no benefit was gained from using dropout. Usually, dropout layers are inserted after dense layers. Accordingly, we add them after the converted dense layers 5 and 7. We use a dropout probability of 50%, which provides the highest level of regularization (Baldi et al., 2013). We compare the quality of networks with and without the usage of dropout layers.

2.4.3. Data augmentation

Neural networks need lots of training data to adjust their weights in a meaningful way. Having small training datasets can cause the network to overfit. Often, the desired amount of data is much larger than what is feasible to acquire. This particularly applies to medical data. In such cases, data augmentation can be used to generate additional artificial data by modifying the existing dataset. For the basic PMap approach, typical modifications are rotation (Janowczyk and Madabhushi, 2016; Xing et al., 2016), rotation plus mirroring (Cireşan et al., 2013; Jacobs et al., 2017) and additional slight color adaptions (Sirinukunwattana et al., 2016). However, the modifications need to be limited such that the artificial data remains to be realistic. For deformation and color adaptions, parameters need to be carefully set to stay within that limit, which is out of the scope of this paper. Thus, we augment the training data only by mirroring and rotation by multiples of 90 degrees, resulting in an 8-fold increase of the training data. We compare the quality of the networks trained with and without data augmentation.

2.4.4. Training target

The training data consists of pairs of images and corresponding exhaustive nuclei center annotations. Target PMaps are derived from the annotations by assigning each pixel a value depending on the distance to the nearest annotation d and a target function $t(d)$. The networks learn to map an input image to a target PMap. Depending on the target function, the difficulty of this mapping can vary.

For the regression approach we compare two definitions of $t(d)$. The first approach aims at directly regressing the proximity to the nearest nucleus center. We define proximity as being 1.0 for $d = 0$, and decreasing with increasing d. In background areas the proximity should be 0.0. Thus, we define a maximum distance d_{max} such that most positions with $d > d_{max}$ reside in background areas. This results in a target function ("dist")

$$t(d) = \begin{cases} 1 - \dfrac{d}{d_{max}} & for\ d \le d_{max} \\ 0 & otherwise \end{cases}.$$

We set $d_{max} = r_{nuc}$. The second approach takes into account that the center annotations may not be completely accurate. To counteract this, we employ a Gaussian-based function ("gauss"), as the values are changing slowly for small values of d. Scaled to meet the PMap's requirements, it is

$$t(d) = e^{-\frac{d^2}{2\sigma^2}}.$$

The approach is proposed by Xie et al. (2016). We set $\sigma = 2$ for comparability of both approaches.

For classification, distinct classes (0 and 1) are required. Like in (Janowczyk and Madabhushi, 2016; Xing et al., 2016), a step function is applied

$$t(d) = \begin{cases} 1 & for\ d \le d_{max} \\ 0 & otherwise \end{cases}$$

We compare $d_{max} = r_{nuc}$ ("large") and $d_{max} = 0.5 \cdot r_{nuc}$ ("small").

2.4.5. PMap post-processing

In theory, PMaps generated by the FCN should resemble the target PMaps. In practice however, predicted PMaps are noisy and exhibit outliers. This is because of the nature of CNNs that allow very different output values even for neighboring positions that share almost the entire receptive field. Thus, post-processing the PMaps may lead to superior nuclei detection quality. Only two of the referenced publications describe smoothing of the PMaps (Cireşan et al., 2013; Janowczyk and Madabhushi, 2016). Both convolve the PMaps with disk-shaped kernels. To compensate both outliers and noise, an alternative approach, which we found useful, is to first apply a small 3×3 median filter followed by Gaussian smoothing. As above, we set $\sigma = 2$ for the Gaussian kernel. We compare the nuclei detection quality with no post-processing, the disk-shaped kernel convolution with disk radius $r = r_{nuc}$ ("disk") and our proposed post-processing ("proposed").

2.5. Experimental setup

2.5.1. Datasets

We train and test the presented configurations of the basic PMap approach on two datasets. The first dataset was made publicly available by the Tissue Image Analytics Lab at Warwick, UK and described in (Sirinukunwattana et al., 2016). It comprises 100 field of view images of H&E stained colorectal adenocarcinoma tissue sections. The images have an extent of 500×500 pixels and a resolution equivalent to $20\times$ optical magnification. Center marker annotations exist for all nuclei in each image. The annotations were all either generated or validated by an experienced pathologist. Using this publicly available dataset improves comparability of the presented methods with the work of Sirinukunwattana et al. (2016) and potential future efforts using the same dataset.

The second dataset originates from a study described in (Molin et al., 2016). The dataset was generated by pathologists selecting 101 circular hot-spot regions from 24 digitized Ki-67 stained

Table 3
Parameter optimization configurations for regression. Bold entries mark the differences from configuration 0. The unit for processing speed is megapixels per second.

Config	Target	Augment	Dropout	Post-Processing	FCN type	Speed in MP/s	Best iteration	F1
0	gauss	yes	50%	disk	upsampling	3.74	21120	0.787
1	**dist**	yes	50%	disk	upsampling	3.74	22416	0.786
2	gauss	**no**	50%	disk	upsampling	3.74	7392	0.757
3	gauss	yes	**none**	disk	upsampling	3.74	14448	0.788
4	gauss	yes	50%	**none**	upsampling	6.94	22704	0.806
5	gauss	yes	50%	**proposed**	upsampling	4.18	22656	0.816
6	gauss	yes	50%	disk	**dilation**	2.41	21552	0.785
7	gauss	yes	50%	**none**	**dilation**	3.12	21648	0.703
8	gauss	yes	50%	**proposed**	**dilation**	2.43	22512	0.828
9	gauss	yes	**none**	**proposed**	upsampling	4.15	14976	0.816
10	gauss	yes	**none**	**proposed**	**dilation**	2.39	19536	0.827

Table 4
Parameter optimization configurations for classification. Bold entries mark the differences from configuration 0. The unit for processing speed is megapixels per second.

Config	Target	Augment	Dropout	Post-Processing	FCN type	Speed in MP/s	Best iteration	F1
0	small	yes	50%	disk	upsampling	3.69	22464	0.776
1	**large**	yes	50%	disk	upsampling	3.69	20256	0.749
2	small	**no**	50%	disk	upsampling	3.69	7488	0.767
3	small	yes	**none**	disk	upsampling	3.72	17424	0.784
4	small	yes	50%	**none**	upsampling	6.84	19344	0.735
5	small	yes	50%	**proposed**	upsampling	4.88	20304	0.794
6	small	yes	50%	disk	**dilation**	2.40	21408	0.775
7	small	yes	50%	**none**	**dilation**	3.40	3168	0.692
8	small	yes	50%	**proposed**	**dilation**	2.76	16944	0.786
9	small	yes	**none**	**proposed**	upsampling	4.17	16992	0.797
10	small	yes	**none**	**proposed**	**dilation**	2.76	21792	0.808

breast tumor tissue sections. Sub-images were extracted containing one hot-spot region each. The images also have a resolution equivalent to 20× optical magnification and an extent of approximately 450 × 450 pixels. Center marker annotations exist for all nuclei within the circular hot-spot regions. They have been validated by an experienced breast pathologist. The Ki-67 positivity of the nuclei is not considered. All nuclei are treated equally. For further analysis, once the nuclei are detected, it is easy to determine their staining. This is, however, beyond the scope of this paper.

In (Kost et al., 2017), different configurations of a random forest–based PMap approach have been evaluated with the Ki-67 dataset. The paper aims at generating optimal training sets for the random forest. With the best-performing configuration an f1-measure of 0.826 has been achieved.

Both datasets exhibit staining variability, artifacts and out-of-focus regions that are representative for real-world data.

The reference annotations have been generated carefully and with high temporal expenditure by experts. However, histological images have inherent ambiguity (cf. Fig. 5), even if the image quality is high. This ambiguity must be kept in mind when interpreting the performance of methods trained and tested using these annotations.

2.5.2. Implementation details

As described in Section 2.2, detected nuclei, which are within r_{nuc} distance to a reference marker, are counted as matches. Sirinukunwattana et al. set this threshold to 6 pixels for their dataset. For the Ki-67 dataset, the threshold used in (Kost et al., 2017) is 10 pixels. To obtain comparability in this paper, we set r_{nuc} to the more strict value of 6 pixels for both datasets.

All experiments are performed with cross-validation to produce robust evaluation results. Each dataset is randomly divided into 5 disjoint folds of equal size. In each round of the cross-validation, 3 folds are used for training, 1 for validation and 1 for testing. This results in 60 training images and 20 images each for both validation and test in each run of the cross-validation. TP, FP and FN values are summed up for all test images in all rounds and an f1-measure is calculated.

Training is conducted for 24,000 iterations in each round, which corresponds to 100 epochs with and 800 epochs without data augmentation. The optimization is performed using the Adam optimizer with the default parameter settings as proposed by Kingma and Ba (2014).

The implementation of the basic PMap approach uses Keras (Chollet, 2015) with Tensorflow backend (Abadi et al., 2015). All experiments are performed on a machine equipped with an Intel(R) Core(TM) i7-7700 CPU @ 3.60 GHz and an Nvidia GeForce GTX 1080 graphics card.

3. Results

3.1. Evaluation of parameters

Evaluating all possible combinations of the described parameters would have taken a very long time, as there are 48 possible combinations each for regression and for classification. Each evaluation requires a full cross-validation and takes about 18 h on the hardware described above.

Although all parameters are interdependent to some extent, we assume these dependencies to be sufficiently low so that each of them can be optimized separately. The only exception is the FCN type and the PMap post-processing, for which all combinations were included in the experiments. As described in Section 2.4.1, the upsampling FCN type performs an interpolation as well, so that FCN type and PMap post-processing are mutually dependent.

To optimize the parameters, we started with a baseline configuration 0, which used the parameter settings that are most common in the literature. For regression, this was a combination of the Gaussian based target, data augmentation, smoothing with a disk-shaped kernel and dropout. For classification, the combination was the same except that the target was the small binary target. Upsampling was chosen as the FCN type for the baseline configurations, as recommended by Long et al. (2014).

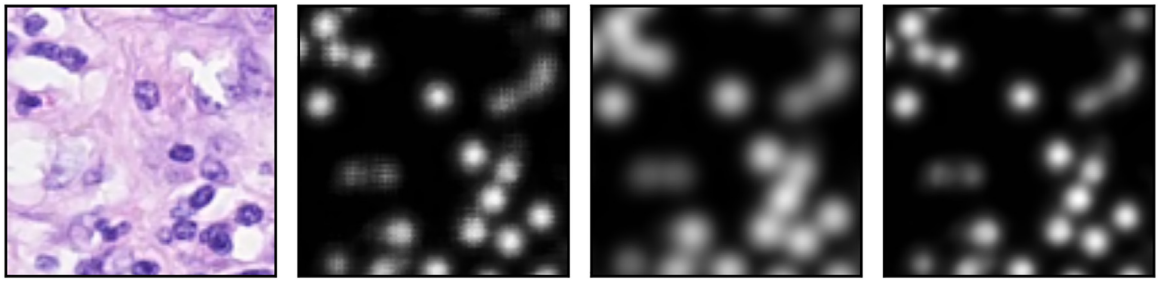


Fig. 3. PMaps after applying the different variants of post-processing. From left to right: example image, PMap without post-processing, PMap after disk smoothing, PMap after proposed post-processing. The PMap was generated using the trained network of regression configuration 7.

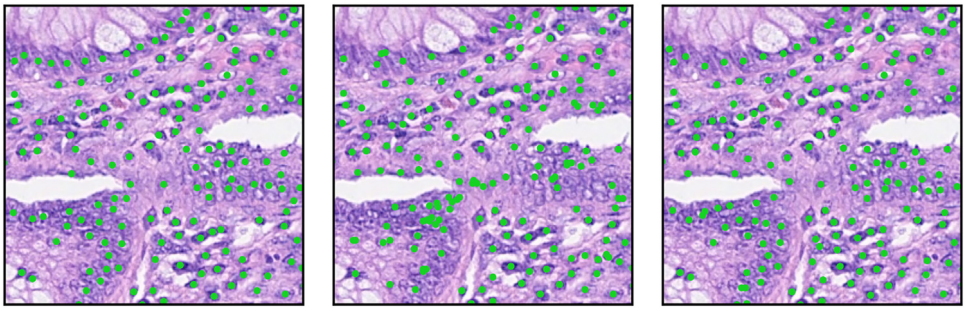


Fig. 4. Visualization of results for nuclei detection in an example image detail of the H&E dataset. Left: Reference markers. Center: Detected nuclei of classification configuration 7. Right: Detected nuclei of regression configuration 10.

Detection quality was determined for each experiment using the test set. Processing speed was measured excluding reading the images into memory. Additionally, the number of iterations was assessed that was required to train the network until its best validation quality was achieved.

The experiment results for the parameter optimization are listed in Table 3 for regression and Table 4 for classification. Both tables show the results for the baseline configuration 0 in the first line. In each of the configurations 1–4 a single parameter was varied. In configurations 5–8, both the PMap post-processing and the FCN type were varied.

3.1.1. Training target

Configuration 1 of the regression experiments shows that the choice between the two targets did not have an impact on the detection quality. For classification however, the usage of the smaller target was preferable.

3.1.2. Data augmentation

In configuration 2, no data augmentation was performed. Although the best iterations were reached much earlier, the achieved f1-measures were considerably lower.

3.1.3. Dropout

In configuration 3, the dropout was disabled. Dropout had a slightly negative effect on the detection qualities. It did, furthermore, increase the training effort, as the best iterations were reached far earlier without dropout.

3.1.4. PMap post-processing

Since the FCN type and the PMap post-processing are mutually dependent, all combinations of these parameters were evaluated. These parameters had the largest impact on the detection quality. For both regression and classification, the best detection quality was achieved using the proposed post-processing. Not applying post-processing at all yielded good results when combined with the upsampling FCN type. However, combined with the dilation FCN type, the f1-measure was the lowest among all configurations. Fig. 3 shows the impact of PMap post-processing on an example image. Execution time was shortest without post-processing, followed by the proposed method. Due to the large kernel size, most time was required when using disk smoothing. The training effort was not affected by the PMap post-processing.

3.1.5. FCN type

With disk smoothing, the FCN type did not affect the detection quality. When combined with the proposed post-processing, dilation yielded best detection quality for regression, whereas for classification the f1-measure of upsampling was higher. Therefore, both FCN types are interesting options. The approach executed faster when using the upsampling FCN type, as the tensors between layers 2 and 9 are considerably smaller.

3.1.6. Resulting configurations

For all parameters, we combined the best settings. As both FCN types have their advantages, this results in the two configurations 9 and 10. For both regression and classification, dilation yielded the best detection quality, whereas upsampling performed considerably faster while having only slightly lower f1-measures.

The very low best iteration value for classification configuration 7 in Table 4 is noteworthy. Although training and validation loss decreased over the iterations, the f1-measure did not improve. However, it varied considerably in the first iterations and became more stable afterwards. This is why the best iteration was observed that early in the training process.

Fig. 4 shows the range of achieved qualities for an example image of the H&E dataset.

3.2. Evaluating learning curve with datasets

After the first experiment, the resulting configurations for both the regression and the classification approach were further eval-

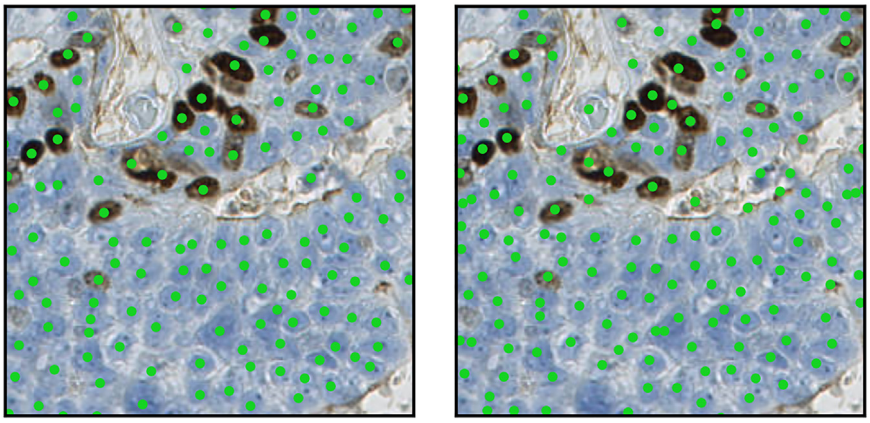


Fig. 5. Visualization of results for nuclei detection in an example image detail of the Ki-67 dataset. Left: Reference markers. Right: Detected nuclei of regression configuration 10.

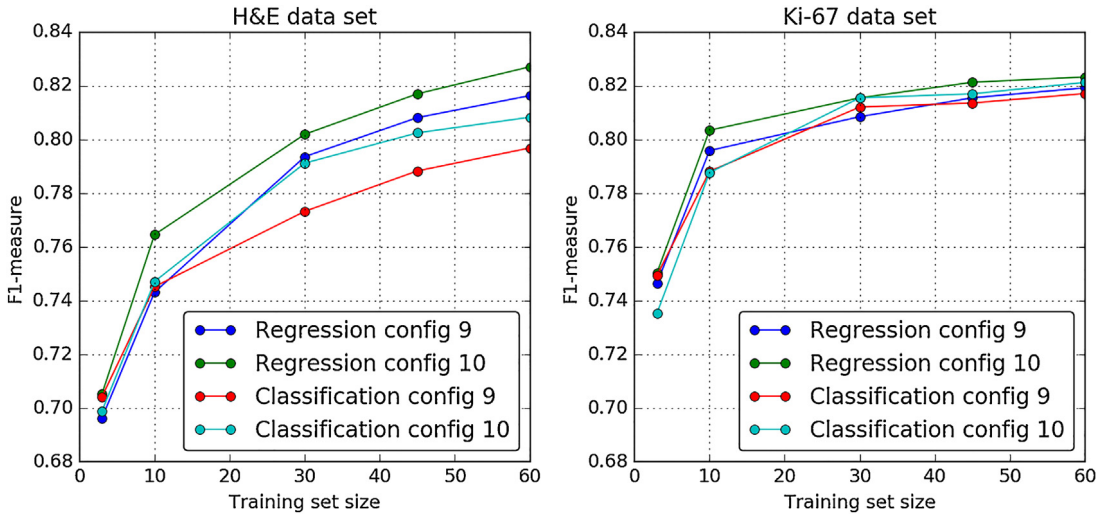


Fig. 6. Learning curves for regression and classification configurations 9 and 10 for both datasets.

uated. In this experiment, their dependency on the size and type of the training data was investigated. We evaluated the configurations with varying sizes of training data by subsampling training and validation data for each run of the cross-validation. We chose to use 100%, 75%, 50%, 17% and 5% of the training and validation data, resulting in 60, 45, 30, 10 and 3 training images and 20, 15, 10, 3 and 1 validation images for each run of the cross-validation, respectively. This evaluation was performed on both the H&E and the Ki-67 dataset. Fig. 5 shows the detected nuclei in an example image of the Ki-67 dataset.

Fig. 6 shows the learning curve experiment results of the configurations for the H&E and the Ki-67 dataset, respectively. For the Ki-67 dataset, the f1-measures achieved with the full training set ranged from 0.817 to 0.823. Overall, the selected configurations achieved similar results for both datasets. The regression configuration 10 yielded slightly superior results on both datasets for almost all training set sizes. Also, this configuration dealt particularly well with small training datasets, which can be observed in training set size of 17% (10 training images).

For most configurations, there is still substantial improvement of the f1-measure from 75% to 100% training set size. This implies that they had not yet reached their maximum.

4. Discussion

We evaluated several configurations of the basic PMap approach for nuclei detection. The f1-measures of the configurations ranged from 0.692 to 0.828. This implies that the performance of the basic PMap approach is greatly affected by its parameters.

The experiments show that PMap post-processing is the most important parameter. For the dilation FCN type, smoothing the PMap was essential. The upsampling FCN type, on the contrary, already performs an interpolation itself. Here, additional smoothing was less important, but still improved detection quality.

The need for smoothing primarily arises from noise that is especially present in the PMaps generated with the dilation FCN type. As the nuclei positions are determined by finding local maxima of the PMap, noise leads to overdetection. Instead of smoothing the PMaps, noise can probably also be reduced by using larger training datasets. However, as described earlier, training data is usually a very limited resource.

A second effect of smoothing is that information from a larger neighborhood is integrated into the current PMap value. Thus, the receptive field is subsequently increased, which is equivalent to adding an additional, fixed layer to the FCN. Such a layer could also

be trainable, but increasing the number of trainable weights in the network usually again requires more training data.

The proposed post-processing method is well suited to serve both of these purposes. The median filtering is used for outlier removal followed by a Gaussian smoothing which incorporates information from neighboring PMap values. For both FCN types, this post-processing method achieved the best results.

The setting of the FCN type is interesting as well. The best results were achieved with the dilation approach, whereas the upsampling approach shows better runtime performance.

The statement that dropout does not improve detection quality, had already been made in other papers (Janowczyk and Madabhushi, 2016; Xu et al., 2016). From our experiments, we additionally noticed an increased training effort, when applying dropout.

For both the regression and the classification architectures, configurations were found that yielded state-of-the-art detection quality on the H&E dataset. However, the regression approach yielded slightly better detection quality. This corresponds to the observations made in (Kainz et al., 2015) and indicates that learning smooth, continuous targets for nuclei detection is preferable to learning discrete targets.

In (Sirinukunwattana et al., 2016), the basic PMap approach has been used to compare their proposed method with, and has achieved an f1-measure of 0.692. Our regression configuration 7 is fairly similar to the configuration described in their paper. Thus, especially with a proper PMap post-processing, a detection quality similar to that of their proposed algorithm (f1-measure: 0.802) could have been achieved.

We further evaluated the two resulting configurations each for both regression and classification on a second dataset. Again, state-of-the-art detection quality was achieved, showing that the same approaches can be applied to different datasets without additional parameter adjustments.

For these configurations and both datasets, we have also investigated the influence of the size of the training data on the detection quality. The regression-based FCN with dilation architecture performed best overall. It was also able to achieve good f1-measures already with small training sets. The plots in Fig. 6 imply that at least some of the compared configurations have not reached their maxima at 60 training images.

In general, adding more variability to the training data helps decreasing the generalization error, as it allows the models to better distinguish real and spurious correlations. Thus, larger training sets that cover more of the expected variability, may further improve the detection quality of these configurations. Due to the lack of more training data, we were not able to prove that hypothesis at this point, and have to leave it for future research.

Overall, we propose the regression configuration 9, as the f1-measures were only slightly lower compared to configuration 10, while the processing speed of 4.15 megapixels per second was considerably better. For the application of the nuclei detection in an end-user software, runtime is a critical factor regarding user acceptance.

For the experiments, PMap post-processing has been performed on the CPU. Moving this step to the GPU and further improvements of runtime are the next steps to take. Future work also includes evaluation of the approaches with more datasets, including further stainings, as well as training a generic nuclei detector using multiple datasets combined.

5. Conclusions

The basic PMap approach is commonly used for the detection of nuclei. However, the implementations differ in a number of parameters, leading to different detection qualities. In this study, we evaluated the impact of the individual parameters on the performance of the approach.

Detection quality, efficiency and training effort of the basic PMap approach are strongly dependent on the parameter settings. When configured properly, performance equal or superior to state-of-the-art approaches can be achieved. Being simple and straightforward, the basic PMap approach constitutes a good choice for nuclei detection tasks.

Conflict of interest statement

JM and CFL are with Sectra AB, Linköping, Sweden. CFL is shareholder of Sectra AB, Linköping, Sweden.

Acknowledgements

This work was conducted under the QuantMed project funded by the Fraunhofer Society, Munich, Germany. Additional funding support was received from Vinnova grant 2014-04257.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.

Andrion, A., Magnani, C., Betta, P.G., Donna, A., Mollo, F., Scelsi, M., Bernardi, P., Botta, M., Terracini, B., 1995. Malignant mesothelioma of the pleura: interobserver variability. J. Clin. Pathol. 48, 856–860.

Arteta, C., Lempitsky, V., Noble, J.A., Zisserman, A., 2012. Learning to detect cells using non-overlapping extremal regions. In: Medical Image Computing and Computer-Assisted Intervention 2012. Springer, pp. 348–356.

Baldi, P., Sadowski, P.J., Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., 2013. Understanding dropout. In: Weinberger, K.Q. (Ed.), Advances in Neural Information Processing Systems 26. Curran Associates, Inc., pp. 2814–2822.

Chollet, F., 2015. Keras [WWW Document]. URL https://github.com/fchollet/keras.

Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks. In: Medical Image Computing and Computer-Assisted Intervention 2013. Springer, pp. 411–418.

Denkert, C., Wienert, S., Poterie, A., Loibl, S., Budczies, J., Badve, S., Bago-Horvath, Z., Bane, A., Bedri, S., Brock, J., Chmielik, E., Christgen, M., Colpaert, C., Demaria, S., Van den Eynden, G., Floris, G., Fox, S.B., Gao, D., Ingold Heppner, B., Kim, S.R., Kos, Z., Kreipe, H.H., Lakhani, S.R., Penault-Llorca, F., Pruneri, G., Radosevic-Robin, N., Rimm, D.L., Schnitt, S.J., Sinn, B.V., Sinn, P., Sirtaine, N., O'Toole, S.A., Viale, G., Van de Vijver, K., de Wind, R., von Minckwitz, G., Klauschen, F., Untch, M., Fasching, P.A., Reimer, T., Willard-Gallo, K., Michiels, S., Loi, S., Salgado, R., 2016. Standardized evaluation of tumor-infiltrating lymphocytes in breast cancer: results of the ring studies of the international immuno-oncology biomarker working group. Mod. Pathol. 29, 1155–1164, http://dx.doi.org/10.1038/modpathol.2016.109.

Jacobs, J.G., Brostow, G.J., Freeman, A., Alexander, D.C., Panagiotaki, E., 2017. Detecting and classifying nuclei on a budget. In: Cardoso, M.J., Arbel, T., Lee, S.-L., Cheplygina, V., Balocco, S., Mateus, D., Zahnd, G., Maier-Hein, L., Demirci, S., Granger, E., Duong, L., Carbonneau, M.-A., Albarqouni, S., Carneiro, G. (Eds.), Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. Springer International Publishing, Cham, pp. 77–86, http://dx.doi.org/10.1007/978-3-319-67534-3_9.

Janowczyk, A., Madabhushi, A., 2016. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J. Pathol. Inform. 7, 29, http://dx.doi.org/10.4103/2153-3539.186902.

Kainz, P., Urschler, M., Schulter, S., Wohlhart, P., Lepetit, V., 2015. You should use regression to detect cells. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention 2015. Springer International Publishing, Cham, pp. 276–283, http://dx.doi.org/10.1007/978-3-319-24574-4_33.

Kårsnäs, A., Dahl, A.L., Larsen, R., 2011. Learning histopathological patterns. J. Pathol. Inform. 2, 12, http://dx.doi.org/10.4103/2153-3539.92033.

Khoshdeli, M., Cong, R., Parvin, B., 2017. Detection of nuclei in H E stained sections using convolutional neural networks. 2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI). Presented at the 2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI), 105–108, http://dx.doi.org/10.1109/BHI.2017.7897216.

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. CoRR abs/1412.6980.

Kost, H., Homeyer, A., Molin, J., Lundström, C., Hahn, H., 2017. Training nuclei detection algorithms with simple annotations. J. Pathol. Inform. 8, 21, http://dx.doi.org/10.4103/jpi.jpi_3_17.

Krizhevsky, A., 2010. Convolutional Deep Belief Networks on cifar-10.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE vol. 86, 2278–2324, http://dx.doi.org/10.1109/5.726791.

Long, J., Shelhamer, E., Darrell, T., 2014. Fully Convolutional Networks for Semantic Segmentation. CoRR abs/1411.4038.

Mahmoud, S.M.A., Paish, E.C., Powe, D.G., Macmillan, R.D., Grainge, M.J., Lee, A.H.S., Ellis, I.O., Green, A.R., 2011. Tumor-infiltrating CD8+ lymphocytes predict clinical outcome in breast Cancer. J. Clin. Oncol. 29, 1949–1955, http://dx.doi.org/10.1200/JCO.2010.30.5037.

Meijering, E., 2012. Cell segmentation: 50 years down the road [Life sciences]. IEEE Signal Process. Mag. 29, 140–145, http://dx.doi.org/10.1109/MSP.2012.2204190.

Molin, J., Bodén, A., Treanor, D., Fjeld, M., Lundström, C., 2016. Scale Stain: Multi-resolution Feature Enhancement in Pathology Visualization. ArXiv Preprint arXiv:1610.04141.

Sirinukunwattana, K., Raza, S., Tsang, Y.-W., Snead, D., Cree, I., Rajpoot, N., 2016. Locality sensitive deep learning for detection and classification of nuclei in routine Colon Cancer histology images. IEEE Trans. Med. Imaging 35, 1196–1206, http://dx.doi.org/10.1109/TMI.2016.2525803.

Vink, J., Van Leeuwen, M., Van Deurzen, C., De Haan, G., 2013. Efficient nucleus detector in histopathology images. J. Microsc. 249, 124–135, http://dx.doi.org/10.1111/jmi.12001.

Wang, S., Yao, J., Xu, Z., Huang, J., 2016. Subtype cell detection with an accelerated deep convolution neural network, in: medical image computing and computer-assisted intervention 2016, lecture notes in computer science. Presented at the International Conference on Medical Image Computing and Computer-Assisted Intervention, 640–648, http://dx.doi.org/10.1007/978-3-319-46723-8_74.

Xie, Y., Kong, X., Xing, F., Liu, F., Su, H., Yang, L., 2015a. Deep voting: a robust approach toward nucleus localization in microscopy images. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention 2015. Springer International Publishing, Cham, pp. 374–382.

Xie, Y., Xing, F., Kong, X., Su, H., Yang, L., 2015b. Beyond classification: structured regression for robust cell detection using convolutional neural network. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention 2015. Springer International Publishing, Cham, pp. 358–365.

Xie, W., Noble, J.A., Zisserman, A., 2016. Microscopy cell counting and detection with fully convolutional regression networks. Comput. Methods Biomech. Biomed. Eng. Imaging Vis., 1–10, http://dx.doi.org/10.1080/21681163.2016.1149104.

Xing, F., Xie, Y., Yang, L., 2016. An automatic learning-based framework for robust nucleus segmentation. IEEE Trans. Med. Imaging 35, 550–566, http://dx.doi.org/10.1109/TMI.2015.2481436.

Xu, J., Xiang, L., Liu, Q., Gilmore, H., Wu, J., Tang, J., Madabhushi, A., 2016. Stacked sparse autoencoder (SSAE) for nuclei detection on breast Cancer histopathology images. IEEE Trans. Med. Imaging 35, 119–130, http://dx.doi.org/10.1109/TMI.2015.2458702.

Paper 4

Automated density-based counting of FISH amplification signals for HER2 status assessment

Henning Höfener, André Homeyer, Mareike Förster, Hans-Ulrich Schildhaus, Horst K. Hahn.

Computer Methods and Programs in Biomedicine 173, 77-85. 2019

The original publication is available at:
https://doi.org/10.1016/j.cmpb.2019.03.006

Computer Methods and Programs in Biomedicine 173 (2019) 77–85

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Automated density-based counting of FISH amplification signals for HER2 status assessment

Henning Höfener[a,*], André Homeyer[a], Mareike Förster[b], Norbert Drieschner[b], Hans-Ulrich Schildhaus[c,d], Horst K. Hahn[a,e]

[a] Fraunhofer MEVIS, Am Fallturm 1, 28359 Bremen, Germany
[b] ZytoVision GmbH, Fischkai 1, 27572 Bremerhaven, Germany
[c] Institute of Pathology, University Hospital Göttingen, Robert-Koch-Straße 40, 37075 Göttingen, Germany
[d] Institute of Pathology, University Hospital Essen, Hufelandstraße 55, 45147 Essen, Germany
[e] Jacobs University, Campus Ring 1, 28759 Bremen, Germany

ARTICLE INFO

Article history:
Received 20 December 2018
Revised 14 February 2019
Accepted 13 March 2019

Keywords:
Fluorescence *in situ* hybridization
HER2
Deep learning
Histology
Image analysis

ABSTRACT

Background: Automated image analysis can make quantification of FISH signals in histological sections more efficient and reproducible. Current detection-based methods, however, often fail to accurately quantify densely clustered FISH signals.

Methods: We propose a novel density-based approach to quantifying FISH signals. Instead of detecting individual signals, this approach quantifies FISH signals in terms of the integral over a density map predicted by Deep Learning. We apply the density-based approach to the task of counting and determining ratios of ERBB2 and CEN17 signals and compare it to common detection-based and area-based approaches.

Results: The ratios determined by our approach were strongly correlated with results obtained by manual annotation of individual FISH signals (Pearson's $r = 0.907$). In addition, they were highly consistent with cutoff-scores determined by a pathologist (balanced concordance = 0.971). The density-based approach generally outperformed the other approaches. Its superiority was particularly evident in the presence of dense signal clusters.

Conclusions: The presented approach enables accurate and efficient automated quantification of FISH signals. Since signals in clusters can hardly be detected individually even by human observers, the density-based quantification performs better than detection-based approaches.

1. Introduction

The status of the human epidermal growth factor 2 (ERBB2/HER2) is an important prognostic and predictive biomarker for breast cancer and other cancers. It is negatively associated with disease relapse and overall survival [27] and used as a basis for specific therapy decision [28]. Pathologists determine the HER2 status either by assessing the expression of the HER2 protein using immunohistochemistry (IHC) or by quantifying the amplification of the ERBB2 gene on chromosome 17 using *in situ* hybridization (ISH) [30]. However, there is controversy regarding the reliability of the IHC assessment [7,22,23]. ISH has the strong advantage of being quantitative, in contrast to the semi-quantitative results of IHC. Of the ISH methods, fluorescence *in situ* hybridization (FISH) is considered the gold standard [24]. FISH uses fluorescent probes that bind to specific sections of the chromosome like the ERBB2 gene. The number of gene copies can then be assessed using a fluorescence microscope.

Interpretation of HER2 status is described in the clinical practice guidelines of the American Society of Clinical Oncology (ASCO) and the College of American Pathologists (CAP). For FISH, the latest update from 2018 [30] recommends the process as follows: FISH is either performed using a single probe for the ERBB2 gene or a dual probe additionally highlighting the chromosome 17 centromer (CEN17) in a different color. A further 4′,6-diamidino-2-

* Corresponding author.
E-mail addresses: henning.hoefener@mevis.fraunhofer.de (H. Höfener), andre.homeyer@mevis.fraunhofer.de (A. Homeyer), foerster@zytovision.com (M. Förster), drieschner@zytovision.com (N. Drieschner), hans-ulrich.schildhaus@med.uni-goettingen.de (H.-U. Schildhaus), horst.hahn@mevis.fraunhofer.de (H.K. Hahn).

https://doi.org/10.1016/j.cmpb.2019.03.006

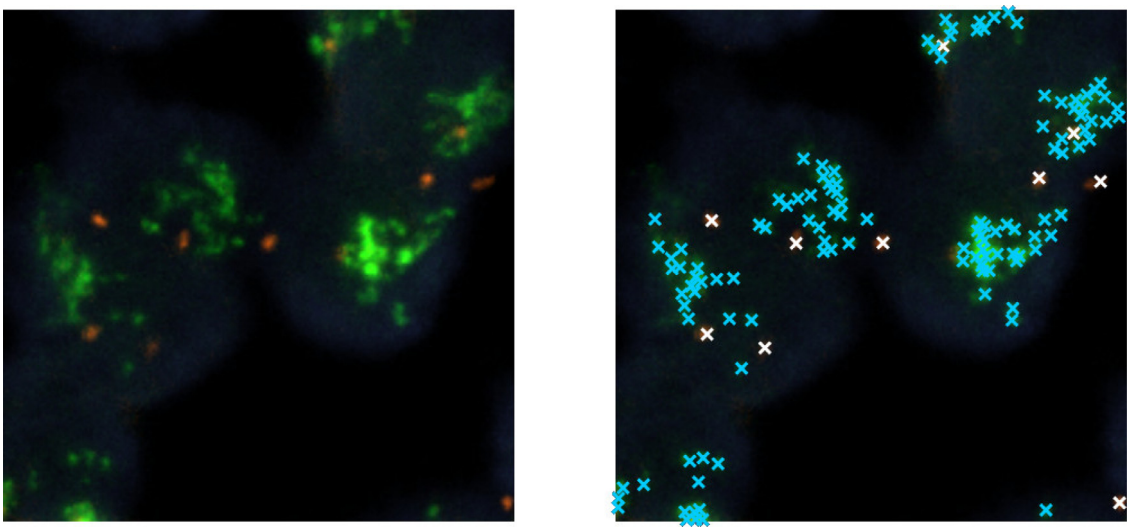


Fig. 1. Example of ERBB2 signals clusters (green), in which single signals cannot be identified unambiguously. Left: Original image. Right: Image with annotation markers visualized.

phenylindole (DAPI) counter stain is applied, binding to all DNA regions and thus highlighting nuclei. For the dual probe, ERBB2 and CEN17 signals have to be counted for at least 20 nuclei. The primary distinction between positive and negative amplification status is made on the basis of the ERBB2/CEN17 ratio. Afterwards, the average ERBB2 copy number per nucleus is assessed. The amplification status is reported as negative if the ERBB2/CEN17 ratio is smaller 2.0 and less than 4.0 ERBB2 signals are present per nucleus. It is considered positive, on the other hand, if the ERBB2/CEN17 ratio is larger or equal to 2.0 and there are 4.0 or more ERBB2 signals per nucleus. In other cases, additional IHC assessment is required to determine the amplification status.

The ASCO/CAP 2018 guidelines require signal counting on a per-nucleus basis. In histological samples, however, nuclei are often clustered such that signals cannot be associated with one nucleus unambiguously. Such clusters are usually not considered for the evaluation, which reduces the amount of analyzable nuclei. Additionally, excluding aggregated nuclei may introduce bias [16]. López et al. [15] assessed ERBB2/CEN17 ratios by considering either all signals in a given tissue region or only signals within nuclei boundaries. The scores almost perfectly agreed, implying that nuclei boundaries do not need to be taken into account.

In addition to the nuclei, ERBB2 and other amplified FISH signals also regularly form clusters of signals, in which single signals cannot be identified unambiguously (cf. Fig. 1 left). Pathologists are recommended to estimate the number of signals in such clusters [20]. They can rely on their experience for patterns and morphology to perform those estimations. Nevertheless, such estimations are always prone to errors and hardly reproducible. Automated signal quantification may thus improve both accuracy and reproducibility.

Manual assessment of the ERBB2 amplification status is very time-consuming. Automated methods can dramatically increase the time efficiency of the analysis. Due to faster execution of automated methods, larger areas of tissue can be analyzed, leading to higher statistical reliability. Moreover, ERBB2 genetic heterogeneity can be captured better when larger portions of the tissue are analyzed. The importance of assessing heterogeneity is increasingly recognized. Patients with ERBB2 genetic heterogeneity were found to have shorter disease-free survival time [25]. Automated methods for assessing the ERBB2 amplification status are already commercially available and have been evaluated [4,9,18].

A number of publications describe approaches for automated analysis of FISH signals for ERBB2 amplification status assessment and other applications. Most of them used classical image processing to enhance the fluorescence signals.

The approaches by Wang et al. [29], Raimondo et al. [19] and Grigoryan et al. [5] were based on top-hat transform. While Grigoryan et al. [5] applied thresholding and morphological filtering to the top-hat output in order to separate and count the signals, Raimondo et al. [19] and Wang et al. [29] used machine learning to distinguish between real and spurious signals. Konsti et al. [10] and Fontenete et al. [3] used difference of Gaussians (DoG) and Laplacian of Gaussians for signal enhancement, respectively, and extracted signals after thresholding. Lerner et al. [13] identified subsignals as areas being brighter than the background at a low resolution and then repeatedly refined using increasing resolutions. Gudla et al. [6] presented two different approaches to FISH spot detection. The first one was a combination of undecimated multiscale wavelet transform (UMSWT) and a low thresholding to obtain signal candidates. The approach then used a random forest to classify these as true or false signals. The second approach used a convolutional neural network (CNN) to generate signal probability maps from the fluorescence channels. The CNN was trained using corrected output of the UMSWT.

As described above, fluorescence signals often form clusters and are thus not always identifiable as dots. Similar problems also occur with other counting tasks like crowd counting or vehicle counting using surveillance camera images. For such tasks, alternatives to detection-based counting are often used. Those are regression-based and density-based counting. In regression-based counting only the number of objects in an image is predicted by a machine-learning regressor. In density-based counting the object density at each position in an image is inferred also using regression. For the latter, the number of objects in an image region equals the spatial integral of the density map over that region.

Since regression-based approaches map to a single global count for each image, they require more training data and discard useful spatial information. Density-based counting was first proposed by Lempitsky and Zisserman [12] and overcomes the disadvantage of regression-based counting. Compared to detection-based counting, density-based counting does not require accurate detection of the objects, which is beneficial when occlusion or densely packed objects are present. For tasks like crowd or vehicle counting but also

cell counting, it has been shown that density-based approaches are often most accurate [12,17].

In their proposal for density-based counting, Lempitsky and Zisserman [12] introduced *Maximum Excess over SubArrays* (MESA-distance) to measure the difference between output and reference density maps. It is defined as the largest absolute difference of the two maps over all possible rectangle subregions. The authors formulated the counting problem as a minimization of a regularized risk quadratic cost function.

Over the last years, CNNs have proven to be well-suited for diverse quantification tasks in images, like counting nuclei in histological sections [8,31]. Approaches using CNNs for density-based counting have since been proposed in several papers. Sindagi and Patel [26] provided a comprehensive overview of the current state-of-the-art particularly regarding CNN-based approaches.

All described approaches for the automated counting of fluorescence signals perform detection-based quantification. With the exception of Lerner et al.'s approach [13], they additionally assume dot-shaped signals. Since fluorescence signals tend to form dense clusters, these approaches are likely to lead to suboptimal quantification results. In this paper, we propose a density-based approach for the quantification of fluorescence signals by the example of ERBB2 amplification. We compare the approach to other detection-based and area-based approaches and evaluate its robustness against uncertainties in the training data.

2. Materials and methods

2.1. Dataset

The Dataset consists of two parts comprising 22 and 14 samples of invasive breast cancer, respectively. For each of the 36 cases, one slide was stained using the Zyto*Light* ERBB2/CEN17 Dual Color Probe (ZytoVision GmbH, Bremerhaven, Germany), which highlights the chromosome 17 centromer with an orange fluorochrome and the ERBB2 gene with a green fluorochrome. The slides were digitized using an Axio Scan.Z1 slide scanner (Carl Zeiss Microscopy GmbH, Jena, Germany) with a 40× objective and a resolution of 164 nm per pixel. Z-stacks consisting of 16 focal planes were acquired and afterwards projected to a single z-plane using the "Contrast" method of the "Extended depth of focus" tool of the ZEN software (Carl Zeiss Microscopy GmbH, Jena, Germany). Assessment and scanning of HER2 fluorescence *in situ* hybridization was carried out as part of routine evaluation for customers. Scans of FISH slides were used with permission of the Sponsor. The remaining 14 cases were archived anonymized clinical routine cases. An informed consent was available for all cases.

Of each slide, 6 field-of-view images (FOV) were extracted, resulting in a total of 216 FOV images. The FOV images were randomly extracted from analyzable areas of the slides. The extent of the FOV images was 30 × 30 µm, resulting in 182 × 182 px. This size enabled extraction and annotation of multiple FOV images from each slide in order to better sample the variability of the slides for both training and test of the approaches.

In all 216 FOV images, all fluorescence signals were annotated by a trained observer and validated by an experienced biologist. ERBB2 signals can form clusters, in which single signals cannot be identified unambiguously and pathologists are recommended to quantify the signal count for such clusters by estimation. For such clusters in our dataset, the number of the involved signals was estimated and the corresponding number of annotation markers were placed inside the cluster such that they reflect the cluster's appearance. Fig. 1 shows an example of cluster annotations.

Additionally, all slides of the first part of the dataset, comprising 22 cases, were assessed by an experienced pathologist according to the ASCO/CAP guidelines. ERBB2 signals between 0.9 and 15.0 per nucleus (median 1.4), CEN17 signals between 1.6 and 34.5 per nucleus (median 2.9) and ERBB2/CEN17 ratios between 1.4 and 11.4 (median 2.2) have been reported. Of the 22 cases, 7 were reported amplified according to the guidelines, 5 of which exhibited ERBB2/CEN17 ratios larger than 2.0.

The dataset exhibits some artifacts that are common in real-world data. Varying preprocessing conditions can lead to background staining, where signals of one color are also visible in other fluorescence channels or whole nuclei are slightly stained with the orange or green fluorochrome. The probes can also accumulate in stroma fibers creating bright structures that can be misinterpreted as signals. Having all color channels at hand, the actual FISH signals are recognizable and distinguishable for the human observer. Considering only a single channel at a time however, some signals become hardly distinguishable from artifacts (cf. Fig. 2).

2.2. Density-based counting with convolutional neural network

Our goal is to develop an approach for the generic quantification of FISH signals, independent of a specific application. The approach consists of a trained CNN that produces density maps from images containing a single fluorescence channel. The number of signals in any region of the image is then determined by integrating the density map over that region.

In this work, the approach is applied for quantifying ERBB2 amplification. The ERBB2 and CEN17 channels are treated as independent input images. In the following, the approach is referred to as *CNN-density*.

2.2.1. Network architecture

The CNN used to generate density maps is a fully convolutional neural network (FCN) [14]. For an input image of arbitrary size, FCNs generate an output image of the same spatial extent (except for a margin explained below). For the signal counting task, the

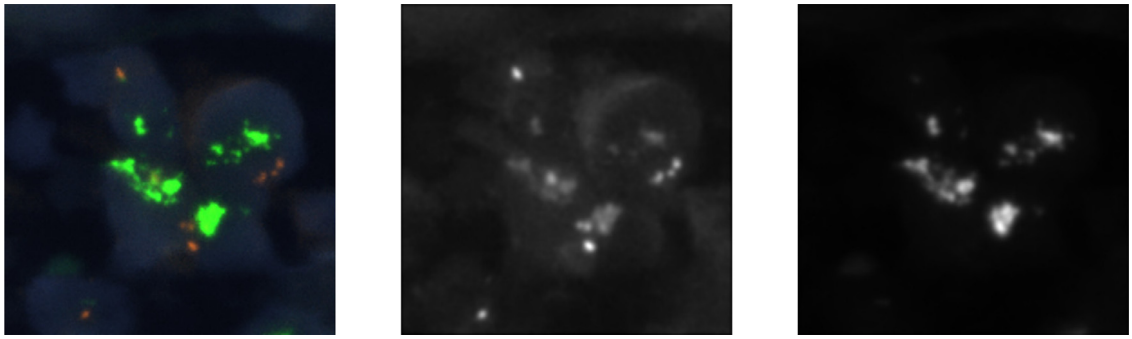


Fig. 2. Example of background staining artifact in FISH images. Left: Original image. Center and Right: CEN17 and ERBB2 channels of left image, respectively. ERBB2 signal is visible in CEN17 channel.

Table 1
Architecture of the network architecture. W and H are input image dimensions.

#	Type	Filter size	Strides	Output size
	Input			$W \times H \times 1$
1	Convolution	5×5	1×1	$W-4 \times H-4 \times 32$
2	Max-Pooling	2×2	2×2	$W/2-2 \times H/2-2 \times 32$
3	Convolution	3×3	1×1	$W/2-4 \times H/2-4 \times 64$
4	Max-Pooling	2×2	2×2	$W/4-2 \times H/4-2 \times 64$
5	Convolution	7×7	1×1	$W/4-8 \times H/4-8 \times 128$
6	Convolution	1×1	1×1	$W/4-8 \times H/4-8 \times 128$
7	Convolution	1×1	1×1	$W/4-8 \times H/4-8 \times 1$
8	Transposed Conv	8×8	4×4	$W-32 \times H-32 \times 1$

output image is a signal density map. The FCN contrasts with classical CNNs, where an image patch of fixed size is fed to the network and a single density value would be predicted for the center pixel of the patch. Although full density maps can also be obtained with a patch-based approach by using a sliding-window procedure, the FCN is much more efficient.

Table 1 shows the network architecture of the *CNN-density* approach. The architecture is straightforward and based on AlexNet [11]. Convolutional layers with small kernels are followed by max pooling layers. To convert it into a FCN, the fully connected layers of AlexNet are replaced by convolutional layers. The channel dimension is doubled after each max pooling layer to compensate for the reduction of the spatial dimensions, as proposed by Xie et al. [31].

To obtain an image of the same size as the input image, upsampling was added to the end of the network, implemented as a transposed convolution. As proposed by Long et al. [14], the kernel weights of the transposed convolution are initialized to perform bilinear upsampling but left trainable to allow optimization during training.

The receptive field of the network is 33 × 33 px. Therefore, the output size of the network is smaller than the input size (cf. Table 1). To obtain density maps of the same size as the input images, an additional margin of 16 px is added to all sides of the FOV images. The margin is filled with zero values (zero padding).

2.2.2. Training procedure

Input data is preprocessed by extracting the ERBB2 and CEN17 channels, which are then treated as independent input images. To compensate for different visual characteristics of the fluorochromes and varying lighting conditions, the channel images are afterwards normalized by subtracting the mean and dividing by standard deviation. A margin of 16 px is added as described earlier.

Target density maps of the same size as the original input images are created. For each annotation marker, the pixel at the corresponding position is set to 1, all other pixels are set to 0. The signal density is then approximated by smoothing the target maps with a small Gaussian filter ($\sigma = 1.33$), reflecting the common signal size.

Many counting methods use the mean absolute error (MAE) as quality and optimization measure. In order to tolerate larger errors when the number of signals is high and only small errors when the number of signals is low, we employ mean normalized absolute error (MNAE). Normalized absolute error (NAE) is defined as $|c_R - c_P|/(c_R + 1)$, where c_R and c_P are the reference and predicted signal counts, respectively. Adding 1 in the denominator avoids division by zero, in case $c_R = 0$.

For the loss function, the NAE can be calculated on the basis of regions of different sizes and then averaged over the image. Those regions range from per-pixel calculation over sliding windows up to a single global NAE for the whole image. If the NAE values are calculated in large regions, only few NAE values are available. In the case of the whole image being that region, only a single value can be calculated, transforming the density-based approach to a regression-based approach. Calculating the NAE values in smaller regions increases the number of values and leads to more robust training. For small regions like a single pixel, however, the individual NAE values become less meaningful.

We address this dilemma by using an increasing sliding window. Training starts using the MNAE with a small window of the same size as the network's receptive field (33 × 33). Every 20 epochs, the window size is increased by 16 px in both dimensions, until it covers the whole density map. This way, the network can use more spatial information in the beginning to robustly optimize towards learning the density map and gradually shift towards a global NAE, training the network for its actual task. After training, the network of the epoch with the lowest MNAE value on the validation set is chosen as the final network.

Different solutions were proposed by Zhang et al. [32] and Lempitsky and Zisserman [12]. Zhang et al. [32] alternated between windowed and global loss, Lempitsky and Zisserman [12] used the MESA-distance as loss, as described in Section 1. We implemented and tested both approaches, but with inferior results.

To estimate signal counts independently of their color, the network is trained with both the ERBB2 and the CEN17 images. The resulting network is chosen to be the network with the minimum validation loss.

2.3. Alternative approaches

We compared the *CNN-density* approach to alternative approaches. Two alternatives, referred to as *DoG-detection* and *CNN-detection*, were implemented on the basis of descriptions in Konsti et al. [10] and Gudla et al. [6]. The third alternative, *CNN-accumulate*, is an advancement of *CNN-detection*.

Konsti et al. [10] proposed a classical image processing approach. Based on their proposal, our *DoG-detection* approach consists of smoothing the input channels with a 3×3 Gaussian kernel filter ($\sigma = 0.5$), followed by signal enhancing using a 7×7 DoG filter ($\sigma_1 = 1.0$, $\sigma_2 = 1.5$). The result is thresholded and connected components are determined. The threshold value is optimized using the training set. The number of fluorescence signals corresponds to the number of connected components.

The second alternative, *CNN-detection*, is based on Gudla et al. [6]. They used a convolutional neural network to generate a probability map. The probability map is defined as a single channel image in which each pixel value represents the probability of the corresponding pixel in the original image belonging to a fluorescence signal. In our *CNN-detection* approach, to reduce the influence of noise, the probability map is smoothed using a Gaussian filter ($\sigma = 0.67$). Before detecting the fluorescence signals by finding local maxima, a threshold is applied to the map suppressing low probability values. The threshold is optimized using the validation set. As network architecture, a reduced version of the U-Net [21] and the dice loss [2] are used, as proposed by Gudla et al. [6].

To generate target probability maps for training the network, Gudla et al. [6] proposed a special processing pipeline. They applied UMSWT combined with a threshold to detect candidate FISH spots. This procedure enhances bright spots and suppresses their surroundings. In a second step, they manually labeled the candidates as either true or false spots. Afterwards, they generate binary target probability maps by removing the false candidates.

In our experiments, we found that the UMSWT approach is not well suited in the presence of signal clusters, as parts of the clusters are being suppressed. We modified the target probability map generation by using the dot-annotations as described in Section 2.1. The map is set to 1 where the distance to the closest signal annotation was ≤ 2 px.

Detection-based approaches have difficulties when the structures to be analyzed are too close to each other. In such cases, local maxima may no longer represent the centers of the structures well. To counter that drawback, *CNN-accumulate* estimates a signal count from the signal regions in the output probability map, instead of only counting the number of local maxima. To do so, a post-processing step was added to the approach. After the probability map is predicted, its values are accumulated over the image region. Then, a scaling factor is applied to calculate the signal count estimation from this value. The scaling factor is optimized using the validation set.

2.4. Evaluation

Signal counting quality was quantified by measuring the normalized absolute error. Normalization ensures comparability amongst the FOV images despite the large variation of the signal count (1 to 253 annotations on a single FOV image). For each approach, the MNAE (cf. Section 2.2.2) was evaluated as the average of NAE values per FOV image for the CEN17 signals, the ERBB2 signals, and all signals combined.

To measure the diagnostic impact of errors in the automated counting approaches, ERBB2/CEN17 ratios were calculated using the resulting counts. The FOV images were classified as amplified or non-amplified and the concordance of the classification between the approaches and the reference was assessed. This classification was determined by applying a cut-off value of 2.0 to the ERBB2/CEN17 ratio, which is a simplification of the ASCO/CAP 2018 guidelines that works without considering nuclei boundaries. For each approach, classification was performed for each FOV image using the predicted signal counts. Additionally, classification was performed for each FOV image using the reference signal counts. The concordance rate is then defined as the ratio of FOV images that have been classified equally by the approach and the reference. To compensate for the imbalance of amplified (70) and non-amplified (146) FOV images, the concordance rate was weighted accordingly. Additionally, as was done by Fontenete et al. [3] and Konsti et al. [10], Pearson correlation coefficients of the ERBB2/CEN17 ratios between the approaches and the reference were determined.

For all approaches we additionally present Bland–Altman plots [1] of the signal counts. In the Bland–Altman plots, the difference between an automated counting approach and the reference counts is plotted on the vertical axis against the mean value of them on the horizontal axis. With these diagrams, the quality of the approaches with respect to the number of signals present in a FOV image can easily be visually assessed. Additionally, median difference and the limits of agreement are visualized. Limit of agreement is the range of differences containing 95% of the FOV images, represented by the 2.5th and 97.5th percentiles. Systematic bias and the relation of error to the magnitude of the measurements can be assessed this way.

For the *CNN-density* approach, we furthermore evaluated the balanced concordance rate between the approach and the pathologist's assessment on a slide level. These assessments are available for the first part of the dataset containing 22 cases, as they were produced within the clinical study (cf. Section 2.1). To do so, we summed up the signal counts for all FOV images of each slide. For comparability, again only the ERBB2/CEN17 ratio was considered to determine the amplification status for both the automated approach and the pathologist's assessment.

Particularly in the presence of artifacts or signal clusters, FISH images always exhibit inherent ambiguity. Therefore, the center of the signals in some cases cannot be identified exactly. Even though the annotations were created with care and great expense of time, they contain uncertainties. To assess the robustness of the *CNN-density* approach against inaccurate annotation placement, in another experiment, the markers were randomly displaced by small amounts. To do so, for each marker, we defined a circular region of given radius around the original position. A new position was then chosen (with uniform distribution) from all pixel positions within that region. The MNAE was evaluated for position alterations with radii of 3, 7 and 11 px.

To produce robust evaluation results, we employed cross-validation. For this, the 216 FOV images were divided into 5 disjoint folds, such that the number of FOV images and reference signals is approximately evenly distributed throughout the folds. Additionally, FOV images originating from the same case were assigned to the same fold to obtain independent training and test data. For the experiments, 3 folds were used for training and one fold each for validation and testing. After cross-validation, each fold has once been used as test set.

3. Results

Our evaluations show that *CNN-density* performed best among the compared approaches. *DoG-detection*, on the other hand, performed worst with regard to all evaluated quality measures. *CNN-detection* and *CNN-accumulate* improved the signal counting error considerably, compared with *DoG-detection*.

CNN-accumulate was superior to *CNN-detection* in most quality measures. However, the counting error of *CNN-detection* is lower for CEN17 signals and also slightly lower for signals in total.

In the evaluation of the *DoG-detection* approach, some images exhibited an undefined ERBB2/CEN17 ratio, as there were no CEN17 signals detected. In the presence of undefined ratios, cut-off concordance and ratio correlation could not be determined. The evaluation results are summarized in Table 2. Output images of all approaches for an example input image are shown in Fig. 3.

Fig. 4 shows Bland–Altman plots for all approaches. On the horizontal axes, from left to right, the number of signals increases. The more signals there are in the FOV images, the more likely they are to form clusters. We can thus assume that the number of signal clusters also increases over the horizontal axes.

All approaches perform worse with increasing numbers of signals per FOV. This can be observed from the fact that at certain positions on the horizontal axes an increasing underestimation

Table 2
Evaluation results of the compared methods. Mean normalized absolute error (MNAE) of CEN17 signals, ERBB2 signals, and all signals combined. Cut-off concordance is based on the amplification status when cut-off value of 2.0 is applied. Measures with undefined values were excluded. The proposed CNN-density performs best for all evaluated measures. (For all correlations: $\boldsymbol{p} < 1\boldsymbol{e}-24$).

Method	MNAE CEN17	MNAE ERBB2	MNAE total	Cut-off concordance	Ratio correlation
DoG-detection	0.606	0.569	0.587	N/A	N/A
CNN-detection	0.347	0.214	0.281	0.829	0.630
CNN-accumulate	0.378	0.186	0.282	0.961	0.837
CNN-density	**0.247**	**0.184**	**0.215**	**0.965**	**0.907**

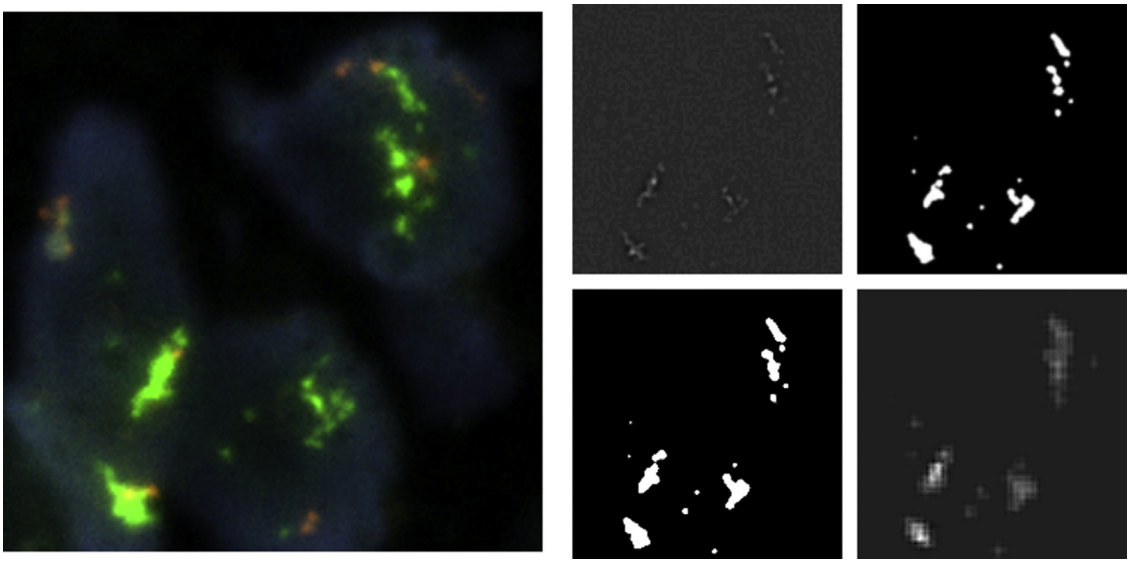


Fig. 3. Example input image (left) with downsampled outputs of the approaches (right). ERBB2 fluorescence channel (green) is examined, reference count is 69 signals. Predicted signal counts are: DoG-detection: 17, CNN-detection: 29 (top row), CNN-accumulate: 65.99, CNN-density: 71.04 (bottom row).

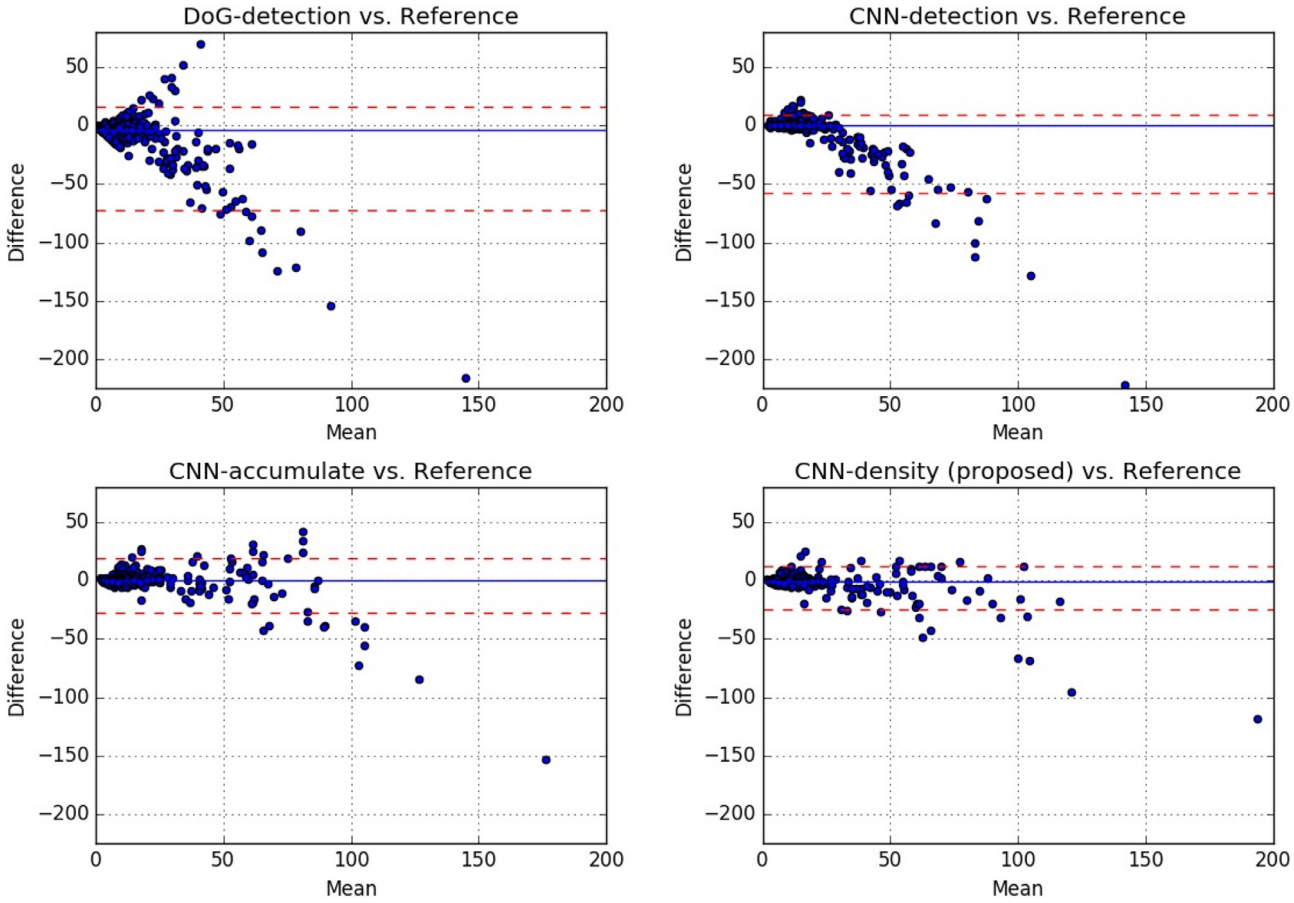


Fig. 4. Bland-Altman plots of signal counts. For all approaches and FOV images, mean and difference of the predicted and the reference count are plotted. Solid blue lines show median difference, dashed red lines visualize the limits of agreement as 2.5th and 97.5th percentile. With increasing signal counts, detection-based approaches misestimate severely. CNN-density shows least misestimation and narrowest limits of agreement.

occurs. However, these positions as well as the severity of the underestimation are different for the approaches. Both detection-based approaches start underestimating signal counts already at mean counts of about 30 signals. *DoG-detection* additionally produces substantial overestimation in a number of cases. For *CNN-accumulate* underestimation starts at mean counts of about 70 signals. *CNN-density* shows least underestimation. It can moreover be observed that the limits of agreement are narrowest for *CNN-density*. Here, 95% of all FOV images have a signal count difference between −25.1 and +12.0.

The amplification status derived from the signal counts of the pathologist and the *CNN-density* approach show good correspondence. Out of 5 cases classified as amplified using the pathologist's signal counts, all were classified equally by the approach. For the

Table 3
Evaluation of CNN-density approach with added uncertainty to annotation marker positions. Markers were repositioned inside circular region of given radius around original position. The approach is robust against uncertainty, ERBB2 signals even show better performance with uncertainty added.

Annotation position alteration radius	MNAE CEN17	MNAE ERBB2	MNAE Total
3 px	0.238	0.152	0.195
7 px	0.247	0.166	0.206
11 px	0.237	0.191	0.212

non-amplified cases, 16 out of 17 were classified equally. This corresponds to a balanced concordance rate of 0.971.

The speed of the *CNN-density* approach was 30.95 megapixels per second excluding image loading and z-projection. For a typical whole slide image with an analyzable area of 20,000 × 20,000 pixels, estimating the signal count would thus take 12.3 s per analyzed color channel. Runtime was measured on a machine equipped with an Intel(R) Core(TM) i7-7700 CPU @ 3.60 GHz and an Nvidia GeForce GTX 1080 graphics card.

As an additional experiment, we introduced uncertainty to the training and validation annotation markers. Displacing the markers by small amounts did not deteriorate the total counting error, but even slightly improved the total MNAE. Table 3 shows the impact of the added uncertainty on the quality measures.

4. Discussion

We evaluated four different approaches to automated fluorescence signal counting. The results show that the proposed density-based counting outperformed the other approaches in all evaluated metrics.

Quantification of CEN17 and ERBB2 signals is impeded mainly by two factors. The first is that ERBB2 signals regularly form dense clusters where single signals cannot be detected easily. Even human observers are unable to identify the individual signals unambiguously. In the presence of such clusters, the limitations of detection-based approaches became apparent. They tend to interpret clusters as a single or very few signals, leading to severe underestimation. Human observers, in contrast, are able to estimate the number of signals in a cluster by considering their size, shape and intensity distribution. Density-based quantification approaches can also take advantage of this information to provide robust signal counts while not being reliant on exact signal detections.

The second impediment for automated signal quantification is artifacts in the FOV images. Background staining artifacts can occur in the CEN17 channel, as described in Section 2.1 and appear to be a problem for all evaluated approaches. *DoG-detection* is only capable of distinguishing between signals and artifacts to a limited extend, leading to overestimated signal counts. *CNN-detection* and *CNN-accumulate* were able to ignore background staining artifacts to some degree. The main difference between these approaches is that a single artifact results in only one or a few false positives in *CNN-detection*, whereas it results in a larger area of false positive in *CNN-accumulate*. Therefore, results for CEN17 signals were better for *CNN-detection*. The *CNN-density* approach on the other hand was best able to tell real and spurious fluorescence signals apart.

Pathologists have much less of a problem distinguishing between artifacts and signals, as they can examine all fluorescence channels. In this work, we deliberately chose to perform signal quantification only on a single given fluorescence channel. The training is performed using both fluorescence channels, but only one at a time. This way, the model is exposed to the variability of both channels, enabling more robust quantification. Furthermore, by being able to conduct reliable signal counting in a single channel, application independence is maintained. In future work, we plan to perform experiments regarding the integration of information from both the ERBB2 and CEN17 channels, for example by feeding both channels to the network and producing two density maps simultaneously. This could improve FISH signal quantification specifically for ERBB2 amplification status assessment with the probe used in the experiments at the expense of becoming application-dependent. Another option would be to directly infer the ratio of ERBB2 / CEN17 signals. With this procedure, however, undefined results are obtained when there are no CEN17 signals. Also, information about absolute signal numbers is disregarded.

Considerable discrepancy can be observed between the results of the *DoG-detection* approach and the corresponding publication. Konsti et al. [10] correlated ERBB2/CEN17 ratios determined using their approach with such determined manually for 42 cases stained with a dual-color FISH probe and reported a Pearson correlation coefficient of 0.98. When classifying the cases using different cut-off values, they reported concordance rates between and 0.90 and 0.95. The mismatch between Konsti et al.'s [10] and our evaluation results can probably be attributed to the datasets used. We assume that their dataset contained no or only few signal clusters. The manual ERBB2/CEN17 ratios ranging from 0.45 to 8.11 for their and from 0.50 to 53.00 for our dataset support this assumption.

The balanced cut-off concordance was evaluated to estimate the extent to which quantification errors effect diagnoses. The concordance rates on a per-FOV basis and on a per-case basis were 96.5% and 97.1%, respectively, showing that the impact of the quantification error on the binary decision between amplified and non-amplified was very small. This suggests that automated, density-based methods can therefore be considered suitable for quantifying ERBB2 amplification status accurately and on a large scale. This in turn leads to the observation that more robust cut-off values could be established on the basis of such methods in future. Automated methods can also better capture heterogeneity, as larger areas of tissue are analyzed. Therefore, they could provide means to find more informative measures for the ERBB2 status that include spatial heterogeneity.

To assess readiness of the approaches for clinical use, they must be compared to multiple human observers. The approaches need to be evaluated against each of the human observers individually or against a reference annotation, derived from all individual annotations. Approaches that lie within the inter-observer agreement, can be considered non-inferior to human observers. However, this requires significant effort and the involvement of multiple experts and is therefore left as future work.

Due to the inherent ambiguity within the images, exact positions of the signals can hardly be determined. To assess the impact of this uncertainty, we further evaluated the robustness of the *CNN-density* approach against small random displacements in the annotation marker positions. The loss function used in the approach calculates the loss in a window of increasing size and is thus not tightly coupled to the exact positions of the annotation markers. Therefore, the signal quantification is robust against small displacements of the reference annotation markers. Instead, small uncertainty even improved the quality for ERBB2 signals. This can be explained by the fact that ERBB2 annotations have substantial uncertainty due to signal clusters. During training, the network is optimized to reproduce the resulting target density maps as good

as possible. Small random displacements prevent the network from aligning too much with the training data and instead more generally learn the concept of clusters. When displacement becomes too large, however, markers will regularly be positioned too far away from the annotated signals, canceling out the described benefit. In order to exploit this advantage, small random marker displacement could be deliberately introduced as a data augmentation technique in future advancements of the approach.

To assess the robustness of the approach against inter-observer variance more thoroughly, reference annotations should be collected from multiple observers. Multiple CNNs could be trained with the reference data of one observer each, and compared to each other. Also, a single CNN could be trained using combined reference data of all observers, in order to cancel out inter-observer variance. Insights into the robustness of the approach could then be gained by comparing the resulting CNN to the aforementioned CNNs.

Density-based counting has already proven to be promising for various tasks and shown its superiority over detection-based approaches. In this work, we showed the applicability of the approach for quantification of FISH signals. Density-based quantification can bring its advantage to bear especially in the presence of clustered structures, which are common in histology. Therefore, the approach could be well suited for quantification of various structures in microscopic images, which would have to be shown in future work.

5. Conclusion

Density-based quantification using CNN appears to be an accurate and efficient method to count fluorescence signals. Compared to existing approaches, it accurately quantifies signal clusters and is robust against artifacts. The results show good agreement with both per-FOV-image annotations and per-case pathologist assessments.

Acknowledgments

The authors would like to thank Dr. Brian Paul Lockhart and Nolwen Guigal-Stephan (Institut de Recherche Servier, Paris, France) for kindly granting permission to use their digitized FISH image data with associated scores. This work was supported by the Federal Ministry of Economics and Technology of Germany under the ZIM grant I³ Life Sciences - Ampli-FISH (16KN054621). Additional funding support was received from the QuantMed project funded by the Fraunhofer Society, Munich, Germany.

References

[1] J.M. Bland, D.G. Altman, Statistical methods for assessing agreement between two methods of clinical measurement, Lancet 327 (1986) 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8.
[2] L.R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (1945) 297–302. https://doi.org/10.2307/1932409.
[3] S. Fontenete, D. Carvalho, A. Lourenço, N. Guimarães, P. Madureira, C. Figueiredo, N.F. Azevedo, FISHji: new ImageJ macros for the quantification of fluorescence in epifluorescence images, Biochem. Eng. J. 112 (2016) 61–69. https://doi.org/10.1016/j.bej.2016.04.001.
[4] D. Furrer, S. Jacob, C. Caron, F. Sanschagrin, L. Provencher, C. Diorio, Validation of a new classifier for the automated analysis of the human epidermal growth factor receptor 2 (HER2) gene amplification in breast cancer specimens, Diagn Pathol 8 (2013) 17. https://doi.org/10.1186/1746-1596-8-17.
[5] A.M. Grigoryan, E.R. Dougherty, J. Kononen, L. Bubendorf, G. Hosteter, O. Kallionienmi, Morphological spot counting from stacked images for automated analysis of gene copy numbers by fluorescence in situ hybridization, JBO JBOPFO 7 (2002) 109–123. https://doi.org/10.1117/1.1428292.
[6] P.R. Gudla, K. Nakayama, G. Pegoraro, T. Misteli, SpotLearn: convolutional neural network for detection of Fluorescence In Situ Hybridization (FISH) signals in high-throughput imaging approaches, Cold Spring Harbor Symposia on Quantitative Biology, 2017 https://doi.org/10.1101/sqb.2017.82.033761.
[7] E.H. Hammond, A.C. Wolff, D.F. Hayes, J.N. Schwartz, Reply to G. Sauter et al., JCO 27 (2009) e153–e154. https://doi.org/10.1200/JCO.2009.24.0366.
[8] H. Höfener, A. Homeyer, N. Weiss, J. Molin, C.F. Lundström, H.K. Hahn, Deep learning nuclei detection: a simple approach can deliver state-of-the-art results, Comput. Med. Imaging Graph. 70 (2018) 43–52. https://doi.org/10.1016/j.compmedimag.2018.08.010.
[9] B. Kajtár, G. Méhes, T. Lörch, L. Deák, M. Kneifné, D. Alpár, L. Pajor, Automated fluorescent in situ hybridization (FISH) analysis of t(9;22)(Q34;Q11) in interphase nuclei, Cytometry 69A (2006) 506–514. https://doi.org/10.1002/cyto.a.20260.
[10] J. Konsti, J. Lundin, M. Jumppanen, M. Lundin, A. Viitanen, J. Isola, A public-domain image processing tool for automated quantification of fluorescence in situ hybridisation signals, J. Clin. Pathol. 61 (2008) 278–282. https://doi.org/10.1136/jcp.2007.048991.
[11] A. Krizhevsky, 2010. Convolutional deep belief networks on CIFAR-10.
[12] V. Lempitsky, A. Zisserman, Learning to count objects in images, in: Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1, NIPS'10, Curran Associates Inc., USA, 2010, pp. 1324–1332.
[13] B. Lerner, L. Koushnir, J. Yeshaya, Segmentation and classification of dot and non-dot-like fluorescence in situ hybridization signals for automated detection of cytogenetic abnormalities, IEEE Trans. Inf. Technol. Biomed. 11 (2007) 443–449. https://doi.org/10.1109/TITB.2007.894335.
[14] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965.
[15] C. López, B. Tomás, A. Korzynska, R. Bosch, M.T. Salvadó, M. Llobera, M. Garcia-Rojo, T. Alvaro, J. Jaén, M. Lejeune, Is it necessary to evaluate nuclei in HER2 FISH evaluation? Am. J. Clin. Pathol. 139 (2013) 47–54. https://doi.org/10.1309/AJCPPXLYJVFGOV8I.
[16] G. Pajor, B. Kajtár, L. Pajor, D. Alpár, State-of-the-art FISHing: automated analysis of cytogenetic aberrations in interphase nuclei, Cytometry 81A (2012) 649–663. https://doi.org/10.1002/cyto.a.22082.
[17] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, R. Okada, in: COUNT Forest: CO-Voting Uncertain Number of Targets Using Random Forest for Crowd Density Estimation, IEEE, 2015, pp. 3253–3261. https://doi.org/10.1109/ICCV.2015.372.
[18] M.J.D. Prins, J.P. Ruurda, P.J. van Diest, R. van Hillegersberg, F.J.W. ten Kate, Evaluation of the HER2 amplification status in oesophageal adenocarcinoma by conventional and automated FISH: a tissue microarray study, J. Clin. Pathol. 67 (2014) 26–32. https://doi.org/10.1136/jclinpath-2013-201570.
[19] F. Raimondo, M.A. Gavrielides, G. Karayannopoulou, K. Lyroudia, I. Pitas, I. Kostopoulos, Automated evaluation of her-2/neu status in breast tissue from fluorescent in situ hybridization images, IEEE Trans. Image Process. 14 (2005) 1288–1299. https://doi.org/10.1109/TIP.2005.852806.
[20] E.A. Rakha, S.E. Pinder, J.M.S. Bartlett, M. Ibrahim, J. Starczynski, P.J. Carder, E. Provenzano, A. Hanby, S. Hales, A.H.S. Lee, I.O. Ellis, Updated UK Recommendations for HER2 assessment in breast cancer, J. Clin. Pathol. 68 (2015) 93–99. https://doi.org/10.1136/jclinpath-2014-202571.
[21] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, 2015.
[22] G. Sauter, J. Lee, J.M. Bartlett, D.J. Slamon, M.F. Press, Guidelines for human epidermal growth factor receptor 2 testing: biologic and methodologic considerations, J. Clin. Oncol. 27 (2009) 1323–1333. https://doi.org/10.1200/JCO.2007.14.8197.
[23] G. Sauter, J. Lee, D.J. Slamon, M.F. Press, Reply to E.H. Hammond et al., JCO 27 (2009) e155–e157. https://doi.org/10.1200/JCO.2009.24.1463.
[24] A. Sáez, F.J. Andreu, M.A. Seguí, M.L. Baré, S. Fernández, C. Dinarés, M. Rey, HER-2 gene amplification by chromogenic in situ hybridisation (CISH) compared with fluorescence in situ hybridisation (FISH) in breast cancerA study of two hundred cases, Breast 15 (2006) 519–527. https://doi.org/10.1016/j.breast.2005.09.008.
[25] H. Seol, H.J. Lee, Y. Choi, H.E. Lee, Y.J. Kim, J.H. Kim, E. Kang, S.-W. Kim, S.Y. Park, Intratumoral heterogeneity of HER2 gene amplification in breast cancer: its clinicopathological significance, Mod. Pathol. 25 (2012) 938–948. https://doi.org/10.1038/modpathol.2012.36.
[26] V.A. Sindagi, V.M. Patel, A survey of recent advances in CNN-based single image crowd counting and density estimation, Pattern Recognit. Lett. Video Surveill.-Orient. Biom. 107 (2018) 3–16. https://doi.org/10.1016/j.patrec.2017.07.007.
[27] D.J. Slamon, G.M. Clark, S.G. Wong, W.J. Levin, A. Ullrich, W.L. McGuire, Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene, Science 235 (1987) 177–182. https://doi.org/10.1126/science.3798106.
[28] D.J. Slamon, B. Leyland-Jones, S. Shak, H. Fuchs, V. Paton, A. Bajamonde, T. Fleming, W. Eiermann, J. Wolter, M. Pegram, J. Baselga, L. Norton, Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2, N. Engl. J. Med. 344 (2001) 783–792. https://doi.org/10.1056/NEJM200103153441101.
[29] X. Wang, X. Chen, Y. Li, H. Liu, S. Li, R.R. Zhang, B. Zheng, Fluorescence in situ hybridization (FISH) signal analysis using automated generated projection images, Anal. Cell. Pathol. 35 (2012) 395–405. https://doi.org/10.3233/ACP-2012-0068.
[30] A.C. Wolff, M.E.H. Hammond, K.H. Allison, B.E. Harvey, P.B. Mangu, J.M. Bartlett, M. Bilous, I.O. Ellis, P. Fitzgibbons, W. Hanna, R.B. Jenkins, M.F. Press, P.A. Spears, G.H. Vance, G. Viale, L.M. McShane, M. Dowsett, Human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists Clinical practice guideline focused update, Arch. Pathol. Lab. Med. 142 (2018) 1364–1382. https://doi.org/10.5858/arpa.2018-0902-SA.

[31] W. Xie, J.A. Noble, A. Zisserman, Microscopy cell counting and detection with fully convolutional regression networks, Comput. Methods Biomech. Biomed. Eng. (2016) 1–10. https://doi.org/10.1080/21681163.2016.1149104.

[32] C. Zhang, H. Li, X. Wang, X. Yang, Cross-scene crowd counting via deep convolutional neural networks, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 833–841. https://doi.org/10.1109/CVPR.2015.7298684.