MIRA VALKONEN

Machine Learning Assisted Digital Pathology
ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Medicine and Health Technology
of Tampere University,
for public discussion in the Small Auditorium 1 (FA032)
of the Festia building, Korkeakoulunkatu 8, Tampere,
on 24 November 2022, at 11 o’clock.
ACKNOWLEDGEMENTS

As I get closer to finishing this whole PhD work, I find myself struggling to let go, yet at the same time, I am more than ready to close this chapter of my life and excited to see what’s next. This has definitely been a great journey and a learning process that would not have been possible without so many amazing people.

I want to express the deepest gratitude to my supervisor, Pekka Ruusuvuori. Thank you for guiding and encouraging me through the years of my thesis work. You have created a wonderful working environment, it truly feels like a second family, a safe place to learn, make mistakes and grow.

Likewise, my heartfelt thanks go to all of the current and former BIIT members. A great group of kind and talented people working as a team. I have been truly fortunate and privileged to work with you all. Thank you for your time and support in many challenges faced during this work, especially, Kaisa Liimatainen, Kimmo Kartasalo, Masi Valkonen, Umair Khan, and Antti Aho.

A special thanks goes to Leena Latonen for all the guidance and support over the years, you have been truly invaluable for many projects by sharing your deeper biological wisdom. Also, I want to thank Matti Nykter, my academic path started as a part of the Computational Biology group and it had a huge impact on where I am to this day. I also would like to warmly thank Teemu Tolonen and Anna Välimäki for collaboration and many insightful conversations related to pathology.

I also wish to acknowledge all other co-authors who have contributed to the original publications of this thesis, Gunilla Högnäs, Jorma Isola, Steven Bova, Tapio Visakorpi, Onni Ylinen, and Ville Muhonen, your input and expertise have been invaluable.

Finally, I would like to express my gratitude to all my friends for their moral support and warm encouragements. I am grateful to my parents and big brothers, thank you for always supporting whatever path I choose. Most importantly I wish to thank my teammate in life, Petteri. You are the source of my strength for overcoming
whatever obstacles there will be in life, be it floating in water or any other kind. :) Also I am grateful for our two amazing kids, Oskari and Kasperi, they have taught me what truly matters in life.

The research for this dissertation was carried out in the Bioimage informatics research group at Tampere University, Faculty of Medicine and Health Technology between 2015 and 2022. The research and doctoral studies were financially supported by TEKES, Academy of Finland, KAUTE Foundation, Orion Research Foundation, Finnish Foundation for Technology Promotion, Nokia Foundation, Instrumentarium Science Foundation, The Finnish Society of Electronics Engineers.

Mira Valkonen,
Kainuun korvessa,
17.7.2022
ABSTRACT

Histopathological tissue samples contain a vast amount of information on underlying biological mechanisms that contribute to disease manifestation and progression. Therefore, diagnosis from histopathological tissue samples has been the gold standard for decades. However, traditional histopathological assessment is a laborious task and prone to human errors, thereby leading to misdiagnosis or delayed diagnosis. The development of whole slide scanners for digitization of tissue glass slides has initiated the transition to a fully digital pathology workflow that allows scanning, interpretation, and management of digital tissue slides. These advances have been the cornerstone for developing intelligent algorithms and automated computational approaches for histopathological assessment and clinical diagnostics.

Machine learning is a subcategory of artificial intelligence and can be defined as a process of learning from data. In image analysis tasks, the raw pixel values are transformed into quantitative feature representations. Based on the image data representation, a machine learning model learns a set of rules that can be used to extract meaningful information and knowledge. Over the years, the field of machine learning based image analysis has developed from manually handcrafting complex features to the recent revolution of deep learning and convolutional neural networks. Histopathological assessment can benefit greatly from the ability of machine learning models to discover patterns and connections from the data. Therefore, machine learning holds great promise to improve the accuracy, reproducibility, and efficiency of clinical diagnostics in the field of digital pathology.

This thesis is focused on developing machine learning based methods for assisting in the process of histopathological assessment, which is a significant step in clinical diagnostics as well as in preclinical studies. The studies presented in this thesis show the effectiveness of feature engineering and machine learning in histopathological assessment related tasks, such as; tissue characterisation, metastasis detection, epithelial tissue detection, and nuclei detection. Moreover, the studies presented in this
thesis address the key challenges related to variation presented in histopathological data as well as the generalisation problem that need to be considered in order to integrate machine learning approaches into clinical practice. Overall, these studies have demonstrated the potential of machine learning for bringing standardisation and reproducibility to the process of histopathological assessment.
TIIVISTELMÄ


Tämä väitöstyö esittelee koneoppimiseen pohjautuvia menetelmiä jotka on kehitetty avustamaan kudosnäytteen histopatologista arviointia, vaihetta joka on merkityksellinen niin kliinisessä diagnostiikassa kuin prekliinisissä tutkimuksissa. Työssä esitetään piirteidenirroituksesta ja koneoppimisen tehokkuus histopatologiseen arvioin- tiin liittyvissä kuva-analyysitehtävissä kuten kudoksen karakterisoinnissa, sekä rintasyövän etäpesäkkeiden, epiteelikudoksen ja tumien tunnistuksessa. Menetelmien
lisäksi tässä väitöystössä on käsitelty keskeisiä haasteita jotka on huomioitava integroitaessa koneoppimismenetelmiä kliiniseen käyttöön. Ennen kaikkea nämä tukkimukset ovat kuitenkin osoittaneet koneoppimisen mahdollisuudet tulevaisuudessa parantaa patologian kliinisten rutinilähtöjä tehokkuutta ja toistettavuutta sekä diagnostiikan laatua.
CONTENTS

1 Introduction ............................................. 17

2 Background ............................................. 21
  2.1 Digital and computational pathology ....................... 21
    2.1.1 Histopathological analysis. .......................... 22
    2.1.2 Sample preparation ................................... 22
    2.1.3 Staining ............................................. 23
    2.1.4 Whole slide scanning ................................. 25
    2.1.5 Computational image analysis ......................... 26
  2.2 Machine learning in histopathological assessment .......... 28
    2.2.1 Feature-based histopathological assessment ............ 30
    2.2.2 Deep learning based histopathological assessment ...... 32
    2.2.3 Generalisation and domain adaptation ................. 34

3 Aims of the study ....................................... 37

4 Materials ............................................... 39
  4.1 Preneoplastic alterations in mouse models ................ 39
    4.1.1 Mouse prostate tissue samples. ....................... 39
  4.2 Breast cancer metastasis detection ........................ 40
    4.2.1 Lymph node sections ................................. 40
  4.3 Epithelial cell detection in assessment of Ki-67 score .... 41
    4.3.1 Primary breast cancer samples .................... 41
  4.4 Nucleus detection in histopathological examination ....... 42
    4.4.1 Radical prostatectomy prostate tissue samples ...... 43
    4.4.2 Multitissue samples .............................. 43
ABBREVIATIONS

AI  artificial intelligence
AUC  area under the curve
CNN  convolutional neural network
DL   deep learning
DP   dorsal prostate
ER   estrogen receptor
GLCM gray-level co-occurrence matrix
HE   hematoxylin and eosin
IHC  immunohistochemistry
IoU  intersection over union
LBP  local binary pattern
LP   lateral prostate
LRP  layerwise relevance propagation
ML   machine learning
mPIN mouse prostatic intraepithelial neoplasia
NN   nearest neighbors
PR   progesterone receptor
RF   random forest
ROC  receiver operating characteristic
ROI  region of interest
SIFT scale-invariant feature transform
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSIM</td>
<td>structural similarity</td>
</tr>
<tr>
<td>SSL</td>
<td>self-supervised learning</td>
</tr>
<tr>
<td>SVM</td>
<td>support vector machine</td>
</tr>
<tr>
<td>t-SNE</td>
<td>t-distributed stochastic neighbor embedding</td>
</tr>
<tr>
<td>VP</td>
<td>ventral prostate</td>
</tr>
<tr>
<td>WSI</td>
<td>whole slide image</td>
</tr>
</tbody>
</table>
ORIGINAL PUBLICATIONS


Equal contribution is denoted by *. The publications are cited in the text by their roman numerals. The original publications are reprinted with the permission of their copyright holders.
Author's contribution

Publication I  First author. The author was primarily responsible for implementation of the study and analysis of the results. Valkonen, Ruusuvuori and Latonen contributed equally to the drafting of the manuscript.

Publication II Joint first author with Kimmo Kartasalo (Tampere University). The author was primarily responsible for the machine learning steps. Kartasalo was primarily responsible for the whole slide image segmentation and preprocessing steps. Both authors contributed equally to the analysis of results and drafting of the manuscript.

Publication III First author. The author was primarily responsible for study design, method conceptualization, analysis of results, and drafting of the manuscript.

Publication IV First author. The author was primarily responsible for study design, method conceptualization, analysis of results, and drafting of the manuscript.
1 INTRODUCTION

Histopathological assessment aims to examine the structural changes in diseased tissue morphology and it is an indispensible step in studying and diagnosing many diseases, including cancer (Orchard and Nation, 2011). The examination is a time-consuming process performed with limited expert resources (Märkl et al., 2021; Metter et al., 2019) and is prone to human errors and subjectivity (Hollensead et al., 2004). In order to increase efficiency, the field of pathology is currently taking a significant step toward the digitalization of the entire workflow due to great advances in computational power and whole slide scanner technology (Figure 1.1 A-B). The availability of digital whole slide images has been the stepping stone for developing advanced computational algorithms and techniques for analysing tissue samples (Figure 1.1 C). The first attempt to assist a diagnosis process using computational methods can be traced back to the early 1990s and the use of digital mammography (Mendez et al., 1998). Since then, there has been a huge push toward building computer-assisted systems for diagnostic purposes, and recently, numerous studies have shown the significant potential of utilising machine learning in the context of diagnostic applications in digital pathology (Campanella et al., 2019; Esteva et al., 2017; Ström et al., 2020; Yu et al., 2016). In addition to clinical applications, histopathological examination is an essential step in preclinical studies on animal models and enables building knowledge on early pathogenesis and the histological characteristics that arise along with diseases (Park et al., 2018; Valkenburg and Williams, 2011).

Histopathological examination aligns strongly with the tasks of characterisation of tissue histology, and identification and classification of relevant regions and structures within the tissue. Computational analysis and machine learning can provide quantitative, objective and accurate tools for these tasks and furthermore, bring standardisation and reproducibility to the whole process of histopathological examination. Earlier work on computational approaches for analysing histopathological images has been focused on traditional approaches (Gurcan et al., 2009) based on
Figure 1.1 Transition from traditional histopathological assessment using microscopes (A) towards machine learning assisted digital pathology. The availability of digital slide scanners (B) have enabled the development of efficient machine learning based algorithms for analysing digital histopathological images (C).

Later, engineered features have mostly been replaced by machine learning algorithms that learn the relevant features automatically (Madabhushi and Lee, 2016; Srinidhi et al., 2021), such as deep neural networks (LeCun et al., 2015). Even though deep learning approaches have shown great success in the performance of many detection and classification tasks, the main requirement for deep learning is a massive amount of training data. The lack of labelled training data due to the laborious expert oriented annotation process, is one of the major challenges in deep learning.

Although computational analysis and machine learning approaches provide efficient tools for histopathological assessment, the unique characteristics of histopathological image data create many challenges from a computational point of view. First, analysing the large multigigapixel images poses challenges with respect to computational efficiency and required resources (X.-W. Chen and Lin, 2014; Zhang et al., 2018). Second, the overall appearance of histopathological images is texture-like containing a great amount of variation and subtle changes of heterogeneous patterns between different tissue types and structures (Tizhoosh and Pantanowitz, 2018). Moreover, the tissue preparation process introduces a great amount of variation originating from the underlying biological variation as well as technical variation from the staining and scanning process (Taqi et al., 2018). These image data character-
istics create a significant challenge for building a robust, efficient and generalisable analysis system (Tizhoosh and Pantanowitz, 2018). Generalisation to new data domains, such as data originating from different medical centers, is one of the major challenges that is unsolved and preventing widespread adaptation of machine learning approaches into routine clinical practice (Pocevičiūtė et al., 2022).

This thesis work is focused on developing and applying machine learning based methods for tasks such as feature extraction, identification of relevant regions, segmentation, and tissue classification, tasks that can assist in the process of histopathological assessment. In addition, the work presented here will address a few important factors that need to be considered while aiming to integrate these tools into real-world clinical environments to assist pathologists in their daily work.
2 BACKGROUND

2.1 Digital and computational pathology

Pathology refers to the study of disease by investigating the causes and effects of disease, and examining structural and functional changes in cells, tissues and organs (Sucaet and Waelput, 2014). As a subdiscipline of pathology, histopathology is focused on studying the structural changes in diseased tissue morphology (Orchard and Nation, 2011). Traditionally, this includes a microscopic examination of a processed tissue sample on a glass slide.

The first steps toward digital pathology date back to the 1960s and the earliest attempts at telepathology (Weinstein, 1986). Nevertheless, advances in computing power, storage capacity and scanner technologies during last two decades have substantially changed the workflow of pathology (Abels et al., 2019). These advances have induced the evolution of whole-slide imaging, allowed the digitalisation of tissue glass slides, and enabled the whole digital infrastructure, thus creating the newest of fields in medicine, digital pathology (Pantanowitz and Parwani, 2017).

The digitalization of pathology workflows has enabled the emergence of computational pathology, which refers to computational approaches focused on the extraction of information from digital histopathological samples (Abels et al., 2019). This thesis work is focused on computational approaches for image analysis tasks in the field of digital pathology. To broaden the understanding of challenges that need to be considered when building a computational approach for analysing digital histopathological samples, the following sections will introduce the steps that are taken when generating a digital image of a tissue sample.
2.1.1 Histopathological analysis

Morphological changes in tissues can be recognised for many medical conditions, such as infection, inflammation and cancer, even before obvious clinical signs and symptoms are absent (Orchard and Nation, 2011). Therefore, the histological findings of the examination have a high clinical value and contribute significantly to diagnoses, clinical decisions and treatment (Geller and Horowitz, 2014). In addition to its contribution to clinical applications, histopathological examination also has an important role in preclinical studies (Greaves, 2011). Preclinical animal models provide an essential approach to examine the mechanisms of disease and histological characteristics that arise along with the evolution of the disease. Animal models of human diseases aim to understand the early pathogenesis and have been successfully applied in the study of many diseases, including intestinal cancer (Jackstadt and Sansom, 2016), Alzheimer’s disease (Sasaguri et al., 2017), breast cancer (Park et al., 2018) and prostate cancer (Valkenburg and Williams, 2011). In addition, histopathological assessment holds a significant role in preclinical toxicity studies as a part of drug development (Greaves, 2011). Thus, histopathological examination is an invaluable tool in studying disease manifestation and progression with respect to clinical and preclinical applications.

Interpreting the morphological changes in tissue requires a high amount of expertise and is a time-consuming and subjective process. The examination process starts with surgery, biopsy, or postmortem examination. In surgical treatment, the aim is to remove all diseased tissue, whereas biopsy refers to only a small sample of tissue taken from the examined tissue or organ (Orchard and Nation, 2011). The removed tissue sample is then prepared and examined in order to provide biological interpretation based on histological findings.

2.1.2 Sample preparation

The preparation of a tissue slide begins with fixation of a tissue specimen. The purpose of this step is to preserve the tissue structures and prevent autolysis and bacterial decomposition (Qidwai et al., 2014). Fixation of tissues can be performed by using chemical or physical approaches. Physical approaches include frozen fixation of tissue, which typically provides excellent preservation of biomolecules but also disrupts the structure of the tissue (Orchard and Nation, 2011). Chemical fixation
approaches can be divided into two common classes: (1) Organic solvent methods, such as PAXgene fixation and, (2) a cross-linking method such as formaldehyde (formalin) fixative (Hobro and Smith, 2017; Qidwai et al., 2014). Formalin fixative is the most commonly used approach for processing tissues for histopathological diagnosis, however, the quality of biomolecules is inferior to that of frozen samples and limited for modern omics technologies (Evers et al., 2011). PAXgene fixative preserves both tissue morphology and biomolecule integrity and has been shown to be suitable for many tissues (Kap et al., 2011). After fixation, the tissue is embedded, most commonly using paraffin wax, to provide a support matrix that enables sectioning the tissue into very thin slices (Orchard and Nation, 2011). Finally, the tissue blocks are manually cut into approximately 5 \( \mu m \) thin sections and carefully transferred to warm water. From the water surface, the sections are placed on a glass slide.

The protocol to carry out these sample preparation steps varies to a great extent between different laboratories and between lab technicians. In larger laboratories, these steps have already been automatized, however, in smaller laboratories and in smaller sample sizes, the whole process can still be manual labor. Therefore, the introduced variability is clearly prominent (Taqi et al., 2018).

2.1.3 Staining

A staining step can be performed after the tissue section has been placed on a glass slide. Unstained tissue is nearly transparent, and therefore, the sections are stained to highlight certain cellular components and structures, and counterstained to provide contrast. In histopathology, the most common stains include hematoxylin and eosin (HE) and immunohistochemical (IHC) stains.

Hematoxylin and eosin (HE) staining

HE staining is the most common stain in routine clinical diagnosis and has been used for at least a century due to its ability to make visible various tissue types and morphologic changes and working well with multiple different tissue fixatives (Bancroft and Layton, 2012). The HE staining can highlight a wide range of cytoplasmic, nuclear, and extracellular matrix structures. Blue hematoxylin mainly stains the cell nuclei and magenta-red eosin stains the extracellular material and cytoplasm. These
qualities build the foundation for several routine clinical diagnostics tasks, such as cancer diagnostics.

There are many HE protocols available for histopathological assessment, and a particular approach is often selected based upon the needs of the histopathological question at hand. The main differences between different protocols mainly include the dye composition, staining time, and intensity of the blue stain. The achieved contrast for a particular tissue will differ based on the approach that is used. An example of tissue stained for HE can be seen in Figure 2.1.

![Figure 2.1](image)

**Figure 2.1** Example images of an HE stained tissue sample (left) and an IHC stained tissue sample (right) stained for Ki-67. In the HE stained tissue sample, blue hematoxylin mainly stains the cell nuclei, and magenta-red eosin stains extracellular material and cytoplasm. In the IHC stained tissue sample, the Ki-67 positive cells are stained brown and negative cells are stained blue.

Immunohistochemical (IHC) staining

Immunohistochemistry (IHC) is a commonly used method for visualising cellular components, such as proteins, in a tissue sample, and therefore serves as an important tool in clinical diagnostics as well as in research-oriented pathology (Taylor, 2014). The IHC staining method utilises antibodies to localize antigens in cellular components from histological tissue samples. The antibodies bind to the targeted antigens, and by utilising enzyme or fluorescent dye that is linked to the antibody, the targeted cellular components can be visualised (Orchard and Nation, 2011). Thus, the IHC method can assist in a diagnosis process by revealing the presence or absence of particular antigens.

IHC staining can be performed after the paraffin has been removed completely. Similarly to HE staining, multiple IHC staining protocols are available, as well as a large number of antigens that are diagnostically useful, such as Ki-67. Antigen KI-67 is a nuclear protein that can be used to determine tumor cell proliferation rate. Ki-67 IHC staining is commonly used in breast cancer diagnostics when determining the aggressiveness of cancer. An example of tissue stained for IHC can be seen in Figure 2.4.
2.1.

2.1.4 Whole slide scanning

With the emergence of digital pathology, the way of viewing pathological samples has gone through a fundamental change. Instead of a traditional microscopic examination, tissue samples can now be examined through a digital monitor. A whole slide scanner is the key component in the process of capturing glass slides into digital images. The first commercial whole slide scanner was designed by James Bacus in 1994 (Bacus Laboratories Inc., Slide Scanner, Bacus et al., 1995).

Figure 2.2 Whole slide scanning enables the examination of digital slides using a computer system without relying upon a traditional microscope. Whole slide scanning includes steps from physically moving the slides under the scanner camera to capture smaller regions tile by tile or via line-scanning and finally stitching the smaller images into one high-resolution image presenting the whole glass slide.

The features of a slide scanner vary widely between different models. However, five main components can be recognised from the process of glass slide digitization: slide handling, slide scanning, optics, detection, and acquisition software (Sucaet and Waelput, 2014). These steps include both hardware and software components. The components of the whole slide scanning process are illustrated in Figure 2.2. The scanner camera moves along the glass slide and captures images that are eventually stitched together to present a digital version of the whole glass slide. The slide handling step includes the processing of physical glass slides, and the capacity for slide-loading ranges from one to hundreds of slides. Slide scanning process can be based on tile scanning (Bacus et al., 1995) or line scanning (Soenksen, 2004). Both approaches capture smaller areas (tiles or strips) from the processed slide that are stacked into a whole slide image (WSI) via the acquisition software.

The whole slide scanner objective (e.g. 20x or 40x) defines the resolution of the generated WSI. A typical pathology slide scanned at 40x can produce a WSI that is several gigabytes large. Therefore, file formatting is required to handle the massive amount of raw data before storing the image data. Commonly, the captured images
are first compressed using one of the various compression schemes (e.g. JPEG2000) available that are either lossless or lossy. For storing the image, there is a large amount of different vendor-specific WSI file formats available. The generated file typically contains high-resolution scanned areas of the tissue and image information in the form of metadata. In addition, scaled versions (e.g. 20x or 40x) of the original image, called zoom levels, are stored in a pyramidal format as visualised in Figure 2.3. By using tile based pyramidal image format and storing precomputed scaled-down versions of the high-resolution image, random access can be optimised, and disk read operations can be minimized (Sucaet and Waelput, 2014).

Overall, whole slide scanning is a complex process with multiple different steps and technical specifications that can introduce variation and artefacts in the generated image data, such as out-of-focus (Kohlberger et al., 2019) or striping (Farahani et al., 2015). For further analysis of the image data and for building robust image analysis tools, it is important to bear in mind the whole process of generating a digital image from a tissue sample.

2.1.5 Computational image analysis

Traditional histopathological assessment is prone to interpretive variability and observer subjectivity, which has been shown in many studies (Cocker et al., 1968; Ismail et al., 1989; Stoler, Schiffman, et al., 2001). The availability of digital WSIs
enables a computational assessment of tissue histology. Computational methods enable quantitative, objective, and efficient image analysis and utilisation of all the information within the high-resolution multigigapixel tissue slides and acquisition of new knowledge beyond human limits (Morales et al., 2021). In addition, computational analysis can bring standardisation and reproducibility to the highly expert oriented process of histopathological assessment and clinical diagnostics (Bizzego et al., 2019; Pell et al., 2019).

The early attempts to extract quantitative measures from microscopy images date back to the 1960s when Prewitt and Mendelsohn (1966) presented a framework for image characterisation and object detection in cytology. However, the early work in computational image analysis was limited to small fields of view and low computational resources. The introduction of whole slide scanners in the 1990s (Bacus et al., 1995) and the availability of digital WSIs led to increased interest in applying image analysis and machine learning techniques to histopathology.

Since then, a great amount of work in histopathological image analysis techniques has been focused on traditional image processing tasks, such as color and illumination normalization or image quality enhancement, which often serve as preprocessing steps for further analysis (Belsare and Mushrif, 2012; Gurcan et al., 2009; Magee et al., 2009). Due to the complex work in sample preparation and variation in staining and scanning conditions, color and illumination normalization is an essential step for microscopy image analysis. This has led to the emergence of many color normalization algorithms, most of which are often based on color separation (Onder et al., 2014). Reinhard et al. (2001) proposed a template color matching method, that aligns the statistical color distribution of a source image with a reference image. Ruifrok et al. (2001) proposed a color deconvolution algorithm that separates presentation of stain components from the RGB channels of an image based on a reference stain vector. Later, Magee et al. (2009) presented an extension to Reinhard et al.’s color transfer technique, and both Magee et al. (2009) and Macenko et al. (2009) presented an automated extraction of stains as an extension to Ruifrok et al.’s color deconvolution work. However, many of the earlier works have shortcomings limiting their applicability rapidly growing data cohorts and WSIs. To address this limitation, a more recent study by Bejnordi et al. (2015) presented a whole-slide image color standardiser model. Recently, a majority of studies have utilised deep learning approaches such as deep generative models (de Bel et al., 2021; Shaban et
al., 2019; Zanjani et al., 2018). Other commonly studied traditional image analysis preprocessing tasks in histopathology are related to image enhancement, such as histogram equalization (Gonzalez et al., 2009) or noise reduction.

One of the prerequisites for histopathological assessment and diagnosis of a disease is detecting certain histological structures from histopathology images (Gartner, 2020). The histological and morphological appearance of structures such as epithelial tissue, cell nuclei or glands, are important diagnostic indicators. Over the years, an extensive number of computational algorithms and image analysis techniques have been developed for segmentation, detection and classification of different histological structures (Belsare and Mushrif, 2012; Komura and Ishikawa, 2018; Srinidhi et al., 2021). Machine learning approaches for histopathological assessment related tasks such as detection and segmentation will be more broadly discussed in the next section.

2.2 Machine learning in histopathological assessment

Machine learning is a subcategory of artificial intelligence and can be defined as a process of learning from data (Mitchell, 1997). A machine learning model can be trained to perform a predictive task by unveiling a possible hidden pattern from the data and building a set of rules for distinguishing these different patterns (Theodoridis, 2015). A machine learning model is a function that maps an input to an output, and the function weights are defined during a training process (Bishop and Nasrabadi, 2006). Machine learning approaches can be broadly categorized into supervised, unsupervised and reinforcement learning approaches (Anzai, 2012). In supervised learning, machine learning model weights are optimised based on labelled training data, that consists of input-output example pairs. Figure (2.4) illustrates the basic functioning of a supervised machine learning model. Employing machine learning based methods can be of great assistance in histopathological assessment by bringing data based decisions and objectivity to the process.

Earlier work has been focused on traditional approaches based on feature engineering and machine learning or other discrimination criteria (Gurcan et al., 2009). Feature engineering aims at mapping raw image pixels into a quantitative feature representation that can be used to extract information from the image (Nixon and Aguado, 2012). Generally, a feature is a primitive characteristic or an attribute of
Figure 2.4  The basic operations of a supervised machine learning model. Model parameters are optimised during the training process using labelled training data. The trained model can then be utilised to predict a label for unseen sample.

an image, that can be directly extracted from the image pixel values or that can be extracted by utilising some manipulation operation. Features can be extracted either locally as a patch based approach or globally on a whole image level. The most commonly used features are morphological, intensity-based, or image texture related features (G. Kumar and Bhatia, 2014; Reed and Dubuf, 1993). The process of feature engineering relies solely on expert knowledge of the discriminative properties that define the regions or objects of interest within an image. An ideal feature extractor would generate a feature representation that makes a classification, segmentation or detection problem trivial. A feature-based machine learning approach builds a predictive model by learning from the extracted feature data. Commonly utilised predictive models include models such as, random forest (RF) models (Breiman, 2001), support vector machines (SVMs) (Boser et al., 1992), linear regression models, and nearest neighbor (NN) classifiers.

More recently, engineered features have been mostly replaced by machine learning algorithms that learn the relevant features automatically (Madabhushi and Lee, 2016; Srinidhi et al., 2021), such as deep neural networks. The popularity of traditional feature-based approaches has decreased significantly, since deep learning based methods have consistently outperformed traditional feature-based machine learning approaches in histopathological image analysis tasks (Bejnordi et al., 2017; Boumaraf et al., 2021; Sharma and Mehra, 2020).
Feature-based histopathological assessment enables the transformation of a tissue sample into quantitative characteristics via feature extraction. Quantitative representation of tissue histology is the basis for many further analyses in clinical as well as preclinical studies. One major factor for the approach’s wide range of applications is the applicability to extensive and small sample sizes. In addition, quantitative feature representation allows the utilisation of machine learning for tasks related to tissue classification, detection, and segmentation of structures. Many traditional machine learning models compute an importance weight for each feature related to the classification task, thus increasing the model explainability. This provides a possibility to potentially reveal the most relevant tissue characteristics related to the classification or detection task and allows exploration of novel biological information.

Feature-based analysis has been successfully utilised in the characterisation and detection of early changes preceding cancer in human tissue, as well as in mouse tissue (Ruusuvuori et al., 2016). Nandakumar et al. utilised morphological and textural features to characterise nuclear structure alterations associated with preneoplastic progression (Nandakumar et al., 2011). Another early study by Berman and Moore (1994) utilised morphological features for detecting preneoplastic lesions. The previous studies have however left room for studying and quantifying the histological variation in normal tissue and distinguishing subtle preneoplastic changes from the normal variation.

Numerous applications of traditional feature engineering combined with machine learning have been widely studied in breast cancer (Fusco et al., 2016; Yassin et al., 2018). Beck et al. (2011) discovered stromal features related to breast cancer morphology that were associated with survival. Petushi et al. (2006) discovered that the surface density of nuclei and their subsequent spatial positions are related to the grade of differentiation for breast cancer. Basavanhally et al. (2013) showed that graph-based features and Haralick features could be used for accurate breast cancer grading.

Feature-based analysis in prostate cancer is another widely studied application field. Early work by Wetzel et al. (1999) compared image features from a test image with similar but previously graded images from a database, using features related to the architectural arrangement of cells. Tabesh et al. (2007) performed Gleason
grading using features related to color, texture, and structural morphology. Likewise, Jafari-Khouzani and Soltanian-Zadeh (2003) were able to separate low and high Gleason grades using a series of wavelet and tissue texture features. In addition to breast cancer and prostate cancer, traditional feature-based approaches have been successfully used for detecting many other cancers, such as neuroblastoma (Sertel et al., 2009) and non-small cell lung cancer (Yu et al., 2016).

Cell or cell nuclei detection, segmentation and classification is also an important and broadly studied research topic, since nuclei count and type have a high significance in cancer diagnosis and grading (Irshad et al., 2013; Xing and Yang, 2016). Early work on detection was often based on traditional image processing, consisting of color component separation, thresholding, morphology operation and watershed (Beucher, 1979; Beucher and Meyer, 2018). The watershed approach is prone to over-segmentation, and therefore, marker-controlled watershed is often used with different methods for locating markers, such as the H-minima/maxima transform (Jung and Kim, 2010; Wählby et al., 2004) or radial symmetry transform (Veta et al., 2013). Cosatto et al. (2008) proposed a Hough transform for nuclei segmentation and utilised an active contour around each point. Ali and Madabhushi (2012) presented an active contour model with initialization based on watershed for nuclei detection in prostate and breast cancer biopsy images and an SVM model based on morphology for classification. Huang and Lai (2010) used a marker-controlled watershed transform and a snake model for nuclei segmentation from liver tissue and an SVM model based on texture and morphology for classification of the nuclei. Another valuable application for nuclei detection is computing the Ki-67 labelling index. The Ki-67 labelling index is a solid marker for predicting the aggressiveness of cancer and holds significant value in the diagnostic assessment of breast cancer. One related study is by Tuominen et al. (2010), who presented an application for counting the percentage of positively stained nuclear area by using a color deconvolution algorithm and adaptive thresholding for nuclear area segmentation. The previous studies have however left room for more standardised determination of the Ki-67 labelling index by developing methods for epithelial tissue detection, since the analysis should be restricted exclusively on malignant epithelial cells.

Other interesting study branches of feature-based histopathological assessment include extending the quantitative feature representation of tissue histology into multidimensional data, such as combination of genetic information with histology
(Pontén et al., 2008; Tomczak et al., 2015; Uhlén et al., 2015), or moving from 2-dimensional histopathological analysis into 3-dimensional analysis (Liimatainen et al., 2021; Magee et al., 2015; Ruusuvuori et al., 2022). Machine learning enables the discovery of hidden patterns from multidimensional data and therefore provides a powerful tool for discovering novel information. However, there is still room for developing more accurate methods to quantify subtle histological characteristics and to link these characteristics to specific genotypes. These multidimensional approaches hold a great potential for building deeper understanding of the pathogenesis and tumor growth patterns (Roberts et al., 2012), and eventually move the field closer to precision histology and profiling tumors on a genomic level (Bera et al., 2019; Djuric et al., 2017; Noble et al., 2022).

In the earlier phases during the transition from the traditional approaches toward deep learning methods, there was a lack of tools for explaining the reasons behind deep learning based model decisions. To fill this absence of explainability, a branch of methods has focused on integrating convolutional neural networks with interpretable engineered features (Valkonen et al., 2017; H. Wang et al., 2014). The main aim of these methods was to combine the high performance of deep learning methods while maintaining the explainability of feature-based approaches.

2.2.2 Deep learning based histopathological assessment

Deep learning based methods, such as deep convolutional neural networks, do not need any specific feature engineering, and deep neural networks learn both the relevant feature representation as well as the classification model (Goodfellow et al., 2016). Convolutional neural networks are a specialized kind of neural networks that use convolution operation in at least one of their layers and process images in a spatially connected way (Pujari et al., 2018). In addition to local connections, these models include multiple layers that allow the discovery of high-level abstractions and patterns within an image. In a supervised learning setting, the network parameters are optimised for the classification task based on labelled training data.

The history of deep learning can be traced back to the 1940s–1960s, at the time known as cybernetics and later connectionism in the 1980s–1990s (Goodfellow et al., 2016). The latest revolution of deep learning began in the early 2000s (LeCun et al., 2015). The final milestone for marking the beginning of deep learning revolution was when Krizhevsky et al. (2012) showed that convolutional neural networks
can outperform all previous machine learning approaches by classifying 1.2 million natural images. The seminal work by Cireșan et al. (2013) revolutionised the entire field of digital histopathology by showing how convolutional neural network can detect mitosis with remarkable accuracy from histopathological images. Since then, deep learning based approaches have rapidly become the most commonly used method for analysing medical images, such as histopathological images (Litjens et al., 2017; Srinidhi et al., 2021).

Deep learning approaches related to histopathological assessment can be divided into two main groups: 1) detecting, locating, segmenting or classifying relevant objects in order to assist in further analysis, and 2) direct image level predictions for diagnosis, prognosis or grading (Morales et al., 2021). Considering the first purpose, deep convolutional neural networks have reached great success in detecting preneoplastic lesions for the early detection of cancer (Sato et al., 2021; Sena et al., 2019), mitosis detection (Tellez et al., 2018), and nuclei segmentation and classification (Graham et al., 2019). Considering the second purpose, deep learning approaches have shown great clinical impact in methods such as, distinguishing biopsies of invasive breast cancer from benign biopsies (Bejnordi et al., 2018) or reaching experts comparable accuracy of detecting and grading prostate cancer biopsies (Ström et al., 2020).

One of the main limitations in developing more accurate deep learning models is the lack of labelled data. Generation of large annotated datasets is an expensive process since it requires a great amount of manual work, expertise and time of pathologists. Challenges and contests organized by research groups or as a part of
conferences have provided a great opportunity for the image analysis community to share and develop new best practices (Bejnordi et al., 2017; Bulten et al., 2022). Recent studies have also investigated the possibility of utilising crowdsourcing approaches in the label collection process (Hou et al., 2020; Marzahl et al., 2019). Another valuable approach for automated collection of labelled data is to exploit physical stains (Turkki et al., 2016). Stains can be used to highlight different structures in the tissue and by utilising image analysis methods, this information can be converted into image labels. This approach provides a wide range of applications to be further studied since there are a great amount of different biomarkers that can be utilised for labelling different tissue structures.

In case a large amount of data is lacking, transfer learning has become a common approach for reaching high accuracy in many detection and classification tasks even with smaller datasets or with limited computational resources. Transfer learning utilises a pretrained network that is already optimised to recognise high-level features from images, commonly used architectures include networks trained with ImageNet (Deng et al., 2009) dataset, such as VGG (Simonyan and Zisserman, 2014), Inception (Szegedy et al., 2015), or ResNet (He et al., 2016). Fine-tuning is a well-known example of transfer learning, where a pretrained model is further trained for a new prediction task. Transfer learning approaches have also been successfully utilised for analysing histopathological images, in tasks such as survival prediction of colorectal cancer (Kather et al., 2019), detecting breast cancer metastases (Y. Liu et al., 2017), and classification of renal cell carcinoma subtypes and survival prediction (Tabibu et al., 2019).

Overall, deep learning has significant potential to improve the efficiency, objectivity and accuracy of cancer diagnostics and many related studies have built the foundation for eventually applying these methods in clinical practice (Campanella et al., 2019; Chugh et al., 2021; Cui and Zhang, 2021; Ström et al., 2020).

2.2.3 Generalisation and domain adaptation

In histopathological image analysis, a high amount of variability present in histological images creates a challenge in building robust and generalisable analysis models. It is known that a majority of the variability is caused by the underlying biological variation, such as variation in texture of different tissues or in nuclei shape and size, but a great amount of variability is also caused by the sample preparation and
scanning procedure (Yagi, 2011). For instance, variation caused by different factors, such as different scanners and tissue fixatives, call for further comprehensive studies with more extensive datasets to assess their impact on model generalisation. Although several approaches have been proposed (BenTaieb and Hamarneh, 2017; Faryna et al., 2021; Ren et al., 2019) to tackle the problem of generalisation into different real-world clinical environments and domain adaptation between different source data domains, in the histopathological image domain the task of generalisation still remains a challenging task.

In the context of machine learning model generalisation in histopathological image analysis, introduced technical variation caused by staining and scanning and the underlying biological variation of different tissue types create a domain shift problem (Ben-David et al., 2010; Moreno-Torres et al., 2012; Stacke et al., 2019). This specifically refers to the shift in data distribution between source data and target data. As a concrete consequence that arises from the problem is e.g. a significant performance drop when analysing histopathological data from different laboratory than the model training data was collected from, or detecting cancer cells from prostate tissue images with a model trained on images of cancer cells from breast tissue. One straightforward solution is to collect a small amount of data from the target domain and fine-tune the model for a new domain. This approach is called domain adaptation and it is an active branch of machine learning (Z. Liu et al., 2020; Long et al., 2015; Saenko et al., 2010). In contrast to transfer learning, in domain adaptation the target domain labels are unavailable. Liimatainen et al. (2019) presented a method for detecting unseen cell lines from brightfield images by utilising iterative domain adaptation. In histopathological image analysis, Ciga et al. (2019) proposed a domain adaptation technique for training a multilevel domain-adversarial network for breast cancer classification. Ren et al. (2018) performed unsupervised domain adaptation through adversarial training of Siamese networks for the classification of Gleason scores. Both reported superior performance to existing state-of-the-art networks.

A main limitation to the domain adaptation method is the assumption of target data being accessible (Zhou et al., 2021). For example, if a domain shift occurs between different patient samples, it is highly impractical to fine-tune a model for each patient. Domain generalisation considers an arbitrary number of related domains and applies the learned knowledge to a completely unseen domain (Blanchard et al., 2011; Muandet et al., 2013). This task, however, is one of the major challenges lim-
iting the adaptation of machine learning approaches into routine clinical practice.
3 AIMS OF THE STUDY

The overall goal of this work was to investigate the possibilities and challenges of machine learning techniques in analysing histopathological images as a part of routine pathology. To address the overall goal, the concrete aims of the study were the following:

1. Develop methods for quantitative characterisation of tissue histology.
2. Develop machine learning approaches for histopathological image analysis, such as detection, segmentation, and tissue classification.
4. Study the tissue fixation as a source of variability present in histological images and its effects in training an accurate and robust machine learning models.

The aim I of the thesis study was addressed in Publications I and II. All of the Publications contributed in achieving the aim II of the thesis. The aim III of the thesis study was addressed in Publications III and the aim IV of the thesis study was addressed in Publications IV.
4 MATERIALS

4.1 Preneoplastic alterations in mouse models

In Publication I, we studied whether computational methods can be used to identify early changes preceding the malignant stages of cancer. We experimented with two genetic prostate cancer mouse models that are commonly used in cancer research, Pten\(^{+/-}\) (Cristofano et al., 1998) and Hi-Myc (Ellwood-Yen et al., 2003). Both of the models develop in situ lesions and therefore are well suited to study an early phase of the development of cancer. Our goal was to discover histological features that can differentiate early pathological lesions from normal prostate tissue, and separate genetically different types of early neoplastic changes from each other.

![Figure 4.1](image_url)

**Figure 4.1** Pten\(^{+/-}\) mouse prostate tissue samples. Manually collected regions of interest are shown in panel A, and a whole tissue section of mouse prostate and different lobular areas are shown in panel B. Adapted from Publication I.

4.1.1 Mouse prostate tissue samples

The mouse prostate tissue samples were fixed in PAXgene tissue fixative and 5 \(\mu m\) sections were cut, attached to glass slides, and stained with HE. An example of an HE stained normal mouse (Pten\(^{+/-}\)) prostate tissue section can be seen in Figure 4.1B.
The section presents a vast range of variation within the different lobular areas of normal prostate tissue, ventral prostate (VP), lateral prostate (LP) and dorsal prostate (DP). To study the normal variation between the three lobular areas as well as mouse prostatic intraepithelial neoplasia (mPIN), regions of interest (ROI) were manually selected from the VP, LP and DP areas for further analysis. ROIs presenting normal tissue included epithelial layers of prostate acini, and ROIs presenting pathological lesions included mPIN/neoplastic epithelium. The collected Hi-Myc samples were always within a single acinus each. Examples of the Pten+/− dataset and collected ROIs are shown in Figure 4.1A.

4.2 Breast cancer metastasis detection

Lymph node metastases occur in most cancer types, for example in breast cancer, prostate cancer, and colon cancer. In breast cancer, the detection of micro- and macro-metastases in lymph node sections has high clinical relevance since it is one of the most important prognostic variables. In Publication II, we developed a machine learning based method for detecting breast cancer metastases from whole tissue sections of lymphatic tissue.

4.2.1 Lymph node sections

Detection of metastatic regions was performed with data from the Camelyon 2016 contest (Bejnordi et al., 2017). The dataset consists of 270 whole slide images of sentinel lymph node sections with corresponding annotations of micro- and macro-metastases, collected from two different laboratories (Radboud University Medical Center and University Medical Center Utrecht, the Netherlands). Examples of whole slide images from the two different laboratories are shown in Figure 6.2A,C. The dataset included 160 normal lymph node sections and 110 cancerous lymph node sections, 97 of these cancerous lymph node sections were fully annotated, and 13 were partially annotated. In fully annotated sections, all cancerous areas were marked, whereas in partially annotated sections, some cancerous areas might be unannotated. The fully annotated sections were used to obtain both positive (cancerous tissue) and negative (normal tissue) training examples, and partially annotated sections were used to obtain only positive examples.
4.3 Epithelial cell detection in assessment of Ki-67 score

In breast cancer diagnostics, Ki-67 immunohistochemical staining of tissue sections is commonly used practice when determining the aggressiveness of cancer (Van Diest et al., 2004). Determination of the tumor cell proliferation rate defined by the Ki-67 labelling index is an important prognostic parameter. The labelling index is defined by the proportion of stained cells in a Ki-67 stained tissue section, and traditionally, the index is manually approximated. Clinical practice of scoring the Ki-67 labelling index varies significantly between different laboratories, yet there are a few guidelines that aim to achieve a more standardised methodology, such as, restricting the estimation of the Ki-67 labelling index to malignant epithelial cells only (Dowsett et al., 2011). In Publication III, we trained a convolutional neural network for epithelial tissue detection and demonstrated how a significant number of high tumor proliferation cases could be incorrectly defined as low tumor proliferation cases without using the epithelial tissue detection as part of the Ki-67 analysis.

4.3.1 Primary breast cancer samples

In Publication III, we analysed formalin-fixed and paraffin embedded (FFPE) samples of 152 invasive breast cancers. The samples were stained with two alternative immunohistochemical staining techniques, fluoro-chromogenic cytokeratin-Ki-67 double staining and sequential hematoxylin-IHC staining. The stainings are shown in Figure 5.2. In the fluoro-chromogenic double staining technique (Fig 5.2 I), the the brightfield image (Fig 5.2 IA) data were used for model training and the immunofluorescent (Fig 5.2 IB) pan-cytokeratin stained data were used for generating the epithelial tissue annotations. In the sequential staining method (Fig 5.2 II), the the brightfield image (Fig 5.2 IIA) data were used for model training and chromogenic (Fig 5.2 IIB) pan-cytokeratin stained data were used for generating the epithelial tissue annotations. Training data were randomly sampled and collected from these 152 whole sections. In total, the training set included 13344 images (256 x 256 pixels). The quality of the training images was manually verified to exclude out-of-focus images, or images with other significant artefacts such as failed staining. The test material consisted of sections stained for estrogen receptor (ER), progesterone receptor (PR), and Ki-67 with hematoxylin as a counterstain. In total, the test set
consisted of 366 images (1276 x 512 pixels) captured from 98 unseen whole slide images representing different patients.

Figure 4.2 The materials used in Publication IV included a prostate tissue dataset, a holdout test set of 5 different tissues and a publicly available MoNuSeg dataset. The prostate tissue dataset is shown in panel A and panel B consists of images collected from five different patients and presents 5 different tissue types. Reprinted from Publication IV.

4.4 Nucleus detection in histopathological examination

Histopathological examination often includes analysis of nuclear morphology, therefore, detection of the cell nucleus can be a fundamental step for many follow-up analyses. Analyses, such as determination of the Ki-67 labelling index, single-cell phenotyping or cancer grading, all include the step of nucleus detection. Nevertheless, the task is very challenging due to a high amount of biological and technical variation present in different tissues. In Publication IV, we trained a deep convolutional neural network for nucleus detection. The datasets used in Publication IV included a prostate tissue dataset (Högnäs et al., 2018), a holdout test set of 5 different tissues and a publicly available MoNuSeg dataset (N. Kumar et al., 2017).
4.4.1 Radical prostatectomy prostate tissue samples

Prostate tissue samples (Högnäs et al., 2018) (Fig 4.2 A) were radical prostatectomy prostate tissue samples collected from 16 men, and fixed using three different tissue preparation methods, fresh frozen, formalin-fixed paraffin-embedded, and PAXgene-fixed paraffin- embedded. The samples were HE stained and scanned using a pixel resolution of 0.23 $\mu m$. Annotations for the nuclei were manually collected for one 550x550 $\mu m$ image from each patient with three different fixatives, resulting in a total of 67,070 nuclei in 48 images.

4.4.2 Multitissue samples

The test dataset with 5 different tissues (Fig 4.2 B) was collected from five different patients, and the images represent metastatic tissue from 5 different tissue types. The tissue types presented in the dataset include periurethral tissue, bone tissue from rib, axillary lymphatic tissue, adrenal gland tissue, and pelvic lymphatic tissue. Each tissue is represented in a 550x550 $\mu m$ image randomly selected from a WSI. In total, 9011 nuclei coordinates were manually annotated.
5 IMAGE ANALYSIS METHODS

5.1 Preprocessing

The aim in image preprocessing is to improve and prepare the image data for further analysis. The improvements can include steps, such as, quality enhancements, distortion removal, normalization or segmentation. Figure 5.1 presents an overview of the image analysis workflow presented in Publication I and an example of possible preprocessing steps.

Figure 5.1 Overview of the image analysis workflow that includes masking of ROI, correcting color variation using histogram matching, excluding areas not including tissue, separating hematoxylin and eosin stains, segmenting cell nuclei, and extracting quantitative feature data. Adapted from Publication I.

5.1.1 Color normalization

In histopathological image data, color variation between samples can be significant due to the staining process. In order to remove the color variation and normalize the image data, color normalization is a commonly used preprocessing step. In Publications I and II, we performed histogram matching in order to generate comparable
image data for feature extraction. For histogram matching, we computed a reference histogram to match the whole image dataset. The reference histogram was computed as a mean histogram from a representative set of samples of the data in Publication I. In Publication II, a histogram from one WSI was selected as the reference histogram based on a visual examination.

5.1.2 Color deconvolution

In histopathological samples, the physical stains highlight different structures of the tissue. For example, hematoxylin mainly stains the cell nuclei, and eosin stains the cytoplasm. Separating these stains and the structures that they present as different color channels can be beneficial for further analysis. In Publications I and II, a color deconvolution algorithm by Ruifrok et al. (2001) was applied to convert the red, green, and blue channels into hematoxylin stain, eosin stain, and background. The algorithm is based on separating the stain components based on a reference stain vector.

5.1.3 Tissue segmentation

In Publication I, we wanted to restrict the analysis only to the effective tissue area, and therefore, we segmented the unwanted regions. The unwanted regions included secretion-filled regions and empty areas within each ROI. A mask for unwanted regions was obtained by subtracting different color channels and performing contrast-limited mapping. A final binary mask was computed using Otsu’s thresholding method (Otsu, 1979) and by applying morphological opening, closing, and filling.

In Publications I and II, we used traditional image processing methods for nuclei segmentation. The nuclei segmentation step was implemented for feature extraction purposes. Since hematoxylin mainly stains the cell nuclei, the hematoxylin channel was further processed to segment the cell nuclei. In Publication I, to obtain a map for high rate of hematoxylin, tophat filtering, maximum filtering, Gaussian filtering, and image intensity adjustment were performed. A binary mask for high rate of hematoxylin was computed using Otsu’s thresholding method (Otsu, 1979). Additionally, for blob detection, the maximally stable extremal regions (MSER) (Matas et al., 2004) method was applied to the hematoxylin channel. In Publication I, the final binary mask for cell nuclei was obtained from MSER regions that were over-
lapping with a mask for a high rate of hematoxylin. In Publication II, cell nuclei were segmented using Gaussian filtering, adaptive thresholding and watershed segmentation.

In Publication II, we performed tissue segmentation from WSI to reduce the amount of data by excluding the background and most of the adipose tissue. The tissue segmentation included, HSV transform, Gaussian filtering, thresholding using Otsu’s method, and applying morphological dilation and exclusion of objects.

Figure 5.2 Examples of the epithelial cell ground truth masks generated from fluoro-chromogenic (I) and sequential chromogenic stains (II). Column A shows the brightfield image, column B shows the immunofluorescent (IB) and chromogenic (IIB) pan-cytokeratin stains, column C shows the binary epithelial cell mask, and column D shows the epithelial cell mask overlaid on images of column A. Adapted from Publication III.

In Publication III, we performed tissue segmentation in order to generate training material for epithelial tissue detection. The training material is presented in Figure 5.2 and includes histological images stained with fluoro-chromogenic cytokeratin-Ki-67 double staining and sequential hematoxylin-IHC staining. In the fluoro-chromogenic double staining technique, the training material was generated by Gaussian filtering the fluorescent pan-cytokeratin images and thresholding with Otsu’s method. In the sequential staining method, to obtain corresponding areas from immunostained images and the cytokeratin stained images, the sequential images were first registered using the StackReg plugin in ImageJ. After registration, DAB stain was separated using a color deconvolution algorithm and then processed with a Gaussian filter and...
thresholded using Otsu’s method. These generated binary mask images were utilised as epithelial tissue annotations for the corresponding immunostained tissue areas.

5.1.4 Label collection

In machine learning, supervised learning algorithms are trained using labelled data. A large amount of labelled data is most often lacking in the histopathological image domain due to the laborious process of data annotation. In Publications I and II, we collected the labelled training data using histology expert annotated pixelwise labels. Deep learning approaches require a large amount of labelled training data in order to train an accurate model. In Publication III, we presented a convenient alternative for time-consuming manual annotation. The training data labels were collected using cytokeratin staining as described in the previous section and presented in Figure 5.2. In Publication IV nucleus locations as coordinates were manually collected by a histology expert.

5.2 Tissue characterisation

The main goal of tissue characterisation is to capture tissue properties and generate a numerical multidimensional representation of an image. In Publication I, tissue characterisation was performed on a ROI level and features were extracted around each nucleus within each ROI. In Publication II, tissue characterisation was performed on a patch level, thus, each WSI was analysed in smaller patches for computational reasons.

5.2.1 Feature extraction

In Publication I, the properties of each tissue ROI were characterised using over two hundred features related to the distribution of nuclei within the ROI and related to image texture within a local neighborhood around each nucleus. To include tissue characteristics in a multiscale manner, each feature was extracted from multiple neighborhood areas representing distinct scale levels of tissue histology. The features related to tissue texture included second-order statistical properties of the gray-level co-occurrence matrix (GLCM) (Haralick et al., 1973), which describes the spatial relations of similar gray tones. In addition, the features included MSER (Matas et
al., 2004) descriptors, local binary patterns (LBPs) (Ojala et al., 2002; Pietikäinen et al., 2000), and descriptors obtained by scale-invariant feature transform (SIFT) (Lowe, 2004). Features related to the distribution of cell nuclei were calculated from a nuclei location map and included features such as; nuclei density features, distance measures between neighboring nuclei, and nuclei locations with respect to each other described with angular statistics.

In Publication II, the image patches were described with over one hundred texture features extracted from both hematoxylin and eosin channels. The extracted features were a modification from the feature engineering presented in Publication I. The features related to tissue texture included properties of GLCM and descriptors obtained by SIFT, MSER, and LBP. Extracted nuclei density features included descriptors related to inter-nuclei distance within an image patch.

5.2.2 Feature visualisation

Visualisation methods are often used in the analysis and interpretation of feature data and dimensionality reduction is one commonly used visualisation method. In Publication I, we wanted to visualise the samples in feature space using dimension-reducing t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008). As a visualisation method, t-SNE is a useful technique for revealing structures at many different scales from high-dimensional data. In Publication I, we were able to demonstrate in two- and three-dimensional space the discriminative properties of the extracted feature data. These plots clearly presented a difference in tissue characteristics of different lobular areas and preneoplastic changes within the analysed tissue samples.

Hierarchical clustering is another commonly used method for the analysis and visualisation of high-dimensional data. The method groups similar samples into clusters based on sample distance and linkage criterion. In Publication I we performed hierarchical cluster analysis of normalized feature values and were able to show that our tissue characterisation approach was able to discriminate between samples of two different genotypes, Hi-Myc and Pten+/−.
5.3 Tissue detection and classification

The main goal of tissue classification is to identify an object or structure from an image, and to define a label for it based on discriminative tissue characteristics and features. In traditional machine learning, the utilised features are handcrafted and engineered by an expert. In deep learning approaches, the model learns both the important features related to the detection task as well as the classifier model. Traditional approach was used in Publications I and II, and the deep learning approach was used in Publications III and IV. In all cases, a supervised learning approach was used.

5.3.1 Feature selection

Feature-based tissue classification approaches can generate hundreds or thousands of features and often many are irrelevant with respect to the classification task at hand. The main goal of feature selection is to remove redundant and irrelevant features as well as find the most relevant ones (Bolon-Canedo and Remeseiro, 2020). Therefore, feature selection is one of the most important steps in designing a traditional machine learning model and has a great impact on the performance of the developed model. In feature selection, a subset of features is selected automatically or manually for further analysis. Feature selection is used for several reasons, such as, dimensionality reduction, simplification and interpretation, or data compatibility improvement.

In Publication I, we compared quantitative histological characteristics of two different genotypes, Hi-Myc and Pten+/- . To ensure comparability of the data, we performed feature selection and selected features whose distributions did not show statistically significant differences (Kolmogorov-Smirnov, threshold $\alpha = 0.05$) within the normal groups of the two different genotypes.

5.3.2 Feature-based machine learning

In Publications I and II, we trained an RF model for tissue classification. The bootstrap aggregation technique was used, which is a machine learning ensemble algorithm that combines multiple versions of decision trees into a random forest model (Breiman, 1996). For both studies, the trained model was an ensemble of 50 decision trees. Each decision tree was trained with a bootstrap sampled dataset, and the
Figure 5.3 The analysis workflow for metastasis detection from lymph node tissue sections. The upper half presents the training phase, and the lower half presents the classification phase. As an output, the trained random forest model assigns an estimate for each analysed tissue patch that presents a probability to belong to the group of cancerous tissue. This probability value is assigned for each tissue patch to obtain a heatmap for the entire WSI as an output. Here, the ground truth annotations are overlaid in yellow on the probability map for reference. Reprinted from Publication II.

The number of features for each decision split was reduced to the square root of the total number of features. Together, the bootstrap method and ensemble approach can improve the stability and accuracy of a model, and avoid overfitting. In Publication II, we selected the random forest model based on a model comparison. In addition to an RF model, we trained a linear regression model, an SVM, an NN classifier with all features presented in Publication II, and an NN classifier with a subset of...
these features.

In Publication II, we built a full tissue classification model, as shown in Figure 5.3. This workflow illustration presents all the preprocessing steps from sampling to color channel separation in the task of metastasis detection from WSIs. The upper half presents the training phase, where the tissue is randomly sampled, and a set of features is extracted from the sampled patches. The extracted features are then used to train a random forest model. The bottom half presents a test sample analysis, where the whole image is processed and features are extracted patch-wise. The extracted features are then run through a trained random forest model, and the corresponding predictions are returned in the spatial tissue context as a heatmap representing areas with high probability of being metastatic tissue.

5.3.3 Deep learning

In Publication III, we trained a deep convolutional neural network for epithelial cell detection from immunostained histological images. A VGG-16 architecture pre-trained on the ImageNet (Deng et al., 2009) dataset was chosen for a baseline model mainly for its earlier success on ImageNet challenge, as well as the good spatial resolution with fairly simple fully convolutional structure that easily allowed modifications on its layers. The structure of the built neural network included the baseline VGG-16 architecture with two additional convolutional layers followed by max pooling and dropout. The output layer was equipped with sigmoid activation to provide a heatmap presenting tissue areas with high probability of being epithelial tissue. The optimisation was performed by using an Adam optimiser (Kingma and Ba, 2014) which is a stochastic gradient descent method. The convolutional neural network presented in Publication III is shown in Figure 2.5.

In Publication IV, we trained a deep convolutional neural network for nuclei detection from histopathological images. The structure of the built fully convolutional neural network included four base layers from a pretrained VGG-16 architecture. The VGG-16 architecture was chosen based on accurate results of our previous studies with histopathological samples as well as its good spatial resolution. The base layer weights were fixed during training. Four additional convolutional layers were added on top of the base layers followed by ReLu activation and dropout. The sigmoid activation function was used at the model output layer to provide a nuclei location probability map. The optimisation was performed by using an Adam optimiser
(Kingma and Ba, 2014) which is a stochastic gradient descent method. The convolutional neural network architecture presented in Publication IV is shown on the right of Figure 5.4.

![Figure 5.4](image_url)

Figure 5.4 Nuclei detection workflow. The upper half presents the baseline model training step which is followed by an option to utilise unsupervised domain adaptation to detect nuclei from an external dataset (new domain) without annotations. The convolutional neural network consists of four convolutional base layers from a pretrained VGG-16 network appended with four additional convolutional layers. Reprinted from Publication IV.

5.4 Generalisation and domain adaptation

In Publication IV, we studied the effect of sample fixation on the accuracy of a deep learning based nuclei detection model. There are multiple different tissue fixatives that can be used in the sample preparation process, and different fixatives cause ap-
parent variation in the quality of the scanned tissue image. In Publication IV, we showed with our experiments that this variation caused significant differences in nuclei detection model accuracies and that the most generalisable model was the one trained on data including multiple sample fixatives. Therefore, in Publication IV, we presented one solution for overcoming generalisation challenges by utilising data from multiple source domains in order to create a generalisable model.

Domain adaptation provides another potential solution for generalising from one data domain to another in histopathological image analysis. In Publication IV, we used the representations learned from labelled source data to generate pseudo-labels in an unlabelled data domain. The workflow for unsupervised domain adaptation using pseudo-labels is shown in Figure 5.4. Thus, we were able to create a nuclei detection model that can generalise to unseen target domains, such as new tissue types or different sample preparation fixatives, without the need for time-consuming manual annotation.

![Figure 5.5](image)

**Figure 5.5** Relative importance of the 10 most significant features selected by the random forest model (A). Example HE images of normal tissue (B) and metastatic tissue (C) are shown with the corresponding 10 most significant features. Reprinted from Publication II.

### 5.5 Model explainability

Model explainability refers to the process of investigating the connection between spatial patterns in input data and classifier outcome (Burkart and Huber, 2021). The features from which the model conclusions are drawn can provide important information on the possible deficiencies of the model or hidden knowledge from the data. Traditional feature-based analysis is obviously easier to interpret since the features are hand-crafted and many commonly used machine learning models provide
an importance weight for each feature in the analysis and thus are explainable by themselves. However, these methods cannot draw hidden patterns from the data since the features are fixed. On the other hand, deep learning based approaches learn both the classification model as well as the important features related to the task, and therefore, can extract novel information from the data. The set-back of these approaches is the risk of using unrelevant information for the classification or detection tasks, leading into unstable models. Therefore, model explainability methods provide means to identify model failures and verify the reasons behind model decisions and thus, can help building trust towards medical AI systems.

5.5.1 Feature-based analysis

![Figure 5.6](image)

**Figure 5.6** Examples of the important areas in an input image resulting in nucleus detection using the LRP method. Each example visualises a 32x32 image block around a detected nucleus and the corresponding probability map presenting the network output, and a heatmap for relevant areas provided by the LRP method. The red areas present positively relevant areas related to nucleus detection and the blue areas present the negatively relevant areas. Reprinted from Publication IV.

In Publication I and II, we visualised the most significant features provided by trained random forest models. An example of feature importance analysis is shown in Figure 5.5, which presents the most significant features related to metastasis detection from lymph node tissue sections.
5.5.2 Deep learning

In Publication IV, we used layerwise relevance propagation (LRP) in order to gain insight into the decisions of our deep learning based nuclei detection model. We visualised examples of detected nuclei and the areas in an input image that were considered to be important in the detection task. A few examples are shown in Figure 5.6.

5.6 Evaluation

Selection of a suitable evaluation metric for tissue classification and segmentation is a challenging task. In plain classification problems, the evaluation task is straightforward and common performance metrics of a binary classifier can be used, such as, sensitivity, specificity, precision or F-score. Most of the methods presented in the Publications of this thesis represent a setting where a classifier model predicts a probability for an analysed tissue area to belong to a class of a certain tissue type. This probability is then thresholded to generate a binary label mask for the corresponding tissue area. Often, this predicted pixelwise mask is compared with the pixelwise ground truth mask, leading to a problem of comparing the similarity of two binary images. Therefore, the process of model performance evaluation and selection of a suitable evaluation metric is not a straightforward task.

5.6.1 Classification performance

In Publication I, system performance evaluation was quite straightforward. We used a receiver operating characteristic (ROC) (Fawcett, 2006) curve, from which the area under the curve (AUC) measure is a common method for analysing the performance of a binary classifier. In Publication I, we evaluated the binary classification task between normal and preneoplastic tissues. A label was predicted for a ROI and compared with labels provided by an expert.

In Publication II, we considered the metastatic detection task as a patchwise classification problem and compared whether each image patch was correctly classified as metastatic or normal tissue compared to the ground truth metastatic tissue mask. Again, we computed the ROC curve for each WSI and evaluated the model performance based on the mean AUC value.
In Publication III, the epithelial tissue detection task was evaluated by multiple quantitative measures. First, we considered the detection task as a pixelwise classification problem and compared each pixel in the predicted mask with the corresponding pixel in the ground truth mask. Classification accuracy, sensitivity, and specificity were computed as statistical performance measures of a binary classifier. The AUC measure was also computed from a mean ROC curve.

In Publication IV, we needed to define a rule for a correctly detected nucleus, since the ground truth was a coordinate for each nucleus. We considered a predicted nucleus to be a true positive detection when a ground truth annotation was within a fixed radius from a predicted nucleus location. Based on this rule, we evaluated the model accuracy based on the F1-score, which rely on precision and recall. For the MoNuSeg dataset in Publication IV, ground truth nuclei segmentations were provided instead of coordinates, therefore, a detection that hits a segmented nuclear area was considered as a true positive.

5.6.2 Segmentation performance

In Publication III, we wanted to perform an image level quantitative comparison in addition to statistical performance measures. We compared each predicted binary mask with a corresponding ground truth cytokeratin mask and computed an intersection over union (IoU) measure and structural similarity (SSIM) index (Z. Wang et al., 2004). In addition, visual evaluation of the predicted epithelial cell masks was performed by two pathologists.
6 RESULTS

6.1 Quantification of normal tissue and preneoplastic alterations

In Publication I, we built a set of computational methods to characterise tissue and detect subtle changes in tissue histology. Our aim was to separate early neoplastic lesions from the normal variation in the prostate tissue of model mice. In addition, our work provides a proof-of-principle for linking specific tumor genotypes to quantitative histological characteristics.

Spatial variation in the epithelium of normal prostate tissue

In order to assess normal spatial variation in the histology of the mouse prostate, we utilise our feature extraction method to characterise and separate the three lobular areas within the normal mouse prostate. The analysed lobular areas include: VP, LP and DP. The normal appearance of the epithelium varies between the different lobular areas. We manually selected 227 regions within the three lobular areas and visualise the extracted feature characteristics of each region in a dimension-reducing t-SNE plot. The feature representation included descriptors related to tissue texture and spatial arrangement of nuclei. The feature representation was able to capture the characteristics of normal spatial variation, DP was more clearly distinguished from the other lobes, and VP and LP partly shared tissue characteristics. The results were well in agreement with the visual appearance of the selected regions.

Quantitative characteristics of mPIN lesions

To examine the changes in tissue histology preceding cancer, we wanted to compare normal tissue to mild pathological changes. We selected 199 early pathological lesions that present mPIN. We performed our feature extraction workflow and visualised the results in a t-SNE plot. The mPIN lesions of different prostate lobes were mixed
Figure 6.1  Quantitative characteristics of normal epithelium, Pten heterozygous mPIN, and Hi-Myc-induced preneoplasia in the mouse lateral prostate. (A) Examples of representative histologies of normal and preneoplastic LP epithelium from Pten+/− and Hi-Myc mouse model prostates. Scale bars 25 μm. (B) Distinct feature value patterns of Pten heterozygous PIN, Hi-Myc-induced early neoplasia, and normal epithelium presented in a heatmap after hierarchical clustering of normalized feature values. (C) Three-dimensional t-SNE visualisation shows distinctive patterns for the three histological populations. Reprinted from Publication I.

rather than separated into lobe-specific clusters, and thus, we concluded that the normal spatial heterogeneity reduced within the tissue as the preneoplastic changes occurred.

To further test the ability to reliably separate mPIN lesions from normal epithelium based on the feature representation, we trained a random forest model to
estimate the probability of a sample belonging to the group of mPIN lesions. Due to the clear phenotypic difference in the tissue histology, mPIN in LP was most accurately separated from normal epithelium (AUC 0.997). mPIN in VP and DP were more challenging, yet the presented feature representation was able to capture the histological differences and a very high classification accuracy was still reached (AUC 0.972).

**Computational distinction between histologies of different genetic groups**

In addition to quantification of normal tissue and preneoplastic alterations, we wanted to study whether our feature representation was able to capture differences within specific tumor genotypes. We compared early neoplasms of two different genetic prostate cancer mouse models, Pten+/- and Hi-Myc, to normal mouse prostatic tissue. Most of the tumors in these mouse models appear in the LP area, and therefore, we limited our experiments to these lobular areas. Examples of normal and neoplastic histologies are shown in Figure 6.1 A. In total, we had 137 normal epithelium and 145 mPIN samples from Pten+/- model mice, and 111 normal epithelium and 189 mPIN samples from Hi-Myc mice.
We performed quantitative image analysis to find computational features separating genotypes from each other and from the normal epithelium. In order to ensure consistency and comparability of the data from Hi-Myc and Pten+/− tissues we selected features whose distributions did not show statistically significant difference (Kolmogorov-Smirnov, threshold $\alpha = 0.05$) within the two normal groups. We performed hierarchical clustering utilising the selected feature values and discovered distinct signatures between analysed the three groups (Figure 6.1 B). The differences are also visible in a t-SNE plot (Figure 6.1 C) that clearly separates the genotypes from each other and from the normal epithelium.

Furthermore, we trained a random forest model on the selected subset of features to estimate the probability of a sample belonging to the group of normal epithelium, mPIN, or Hi-Myc early neoplasia. The model was able to accurately predict the correct class of a sample based on the histological features (AUC 0.997 for normal, 0.990 for Pten+/−, and 0.995 for Hi-Myc) using leave-one-out cross validation approach.

The results obtained in Publication I show that separation between different spatial locations within normal prostate tissue of model mice, as well as separation between histologies linked to different genetic backgrounds, can be performed using computational feature analysis. In addition, the numerical tissue characteristics provide a basis for using machine learning based classification of tissue regions with subtle alterations. Overall, the workflow presented in Publication I met the aims I and II of the thesis study.

### 6.2 Metastasis detection

In Publication II, we present a machine learning approach for detecting cancerous tissue from whole slide images of sentinel lymph node sections. The method is based on feature engineering and a random forest model and is an extension of the feature-based image analysis presented in Publication I. An RF model was selected after a model comparison between linear regression model, an SVM, and NN classifiers. The RF outperformed the other models in terms of correctly classified samples, sensitivity, and F-score.

Whole slide images were processed in smaller sections, block by block, and the properties of each tissue sample block were described with local descriptors related to
image texture, spatial structure, and distribution of nuclei. A random forest model was trained to predict a probability value indicating the likelihood of metastatic cells being present in the corresponding part of the image. Examples of the predicted heat maps are presented in Figure 6.2 B,D. We performed a blockwise receiver operating characteristic (ROC) analysis in order to evaluate the performance of our method. We obtained high accuracy in separating metastatic areas from normal tissue using patient-level leave-one-sample-out cross-validation (dataset from laboratory 1 mean AUC = 0.905, dataset from laboratory 2 mean AUC = 0.887, CI 95%). During training, the model has seen image data collected from both of the laboratories, even though the model has not seen image samples from the particular test patient at each fold.

Therefore, to evaluate the generalisation ability of our approach, we additionally evaluated the performance with completely independent data. First, we trained with the dataset from one laboratory and then tested with images from the other laboratory. The results were encouraging, our approach was able to separate metastatic areas from normal tissue with a mean AUC = 0.839 (CI 95%, training dataset from laboratory 1, test dataset from laboratory 2) and AUC 0.855 (CI 95%, training dataset from laboratory 2, test dataset from laboratory 1). A slight decrease in the performance is observed compared to the cross-validation results, however, the overall accuracy is still high. The presented approach ranked as a best feature-based approach in the CAMELYON16 contest and outperformed a few deep learning based methods, however, the best performing methods all relied on deep learning. The results from the contest clearly represent the development of the field at the time, when deep learning methods started to consistently outperform traditional feature-based approaches in many image analysis tasks.

The results obtained in Publication II show that machine learning provides a powerful tool in analysing and classifying histopathological tissue samples, and thus, meets the aims I and II of the thesis study. In addition, the results obtained in Publication II indicate the high potential of automating pathological routine work by using machine learning, however, the work also provides insight into the generalisation problem in the field. In order to eventually adapt and integrate machine learning tools into real-world clinical environments, more research is needed to discover the sources of variability present in histological images. Here, we used images from two different laboratories, and we could already see a great amount of variability in the
image data due to technical reasons. Thus, to some extent, the results of Publication II also address the aim IV of the thesis study. To conclude, there will be a pressing need for methods that can generalise to data from different sources, such as different clinical environments, and discover the underlying universal tissue characteristics from image data amongst the unimportant features, technical variation and noise.

Figure 6.3  Example outputs of deep learning based epithelial cell detection in Ki-67 IHC stained invasive carcinoma (A,D), ductal carcinoma in situ (B,E), and non-malignant ductal breast epithelium (C,F). Tissues recognised as epithelial by the deep learning algorithm are highlighted in D-F. Because the pan-cytokeratin staining was used as training material for the deep learning algorithm, epithelial cells in DCIS and normal-breast are included. Pathologists should exclude them in the areas to be measured for Ki-67%. Reprinted from Publication III.

6.3 Epithelial cell detection in Ki-67 analysis

In Publication III, we built a deep learning based system for epithelial cell detection from histological images. The labelled training data were collected using a cytokeratin staining based approach to avoid the laborious annotation step. The performance of the system was evaluated by comparing the generated masks to cytokeratin based epithelial cell masks (AUC of mean ROC = 0.93) and by visual evaluation of the masked brightfield images performed by two pathologists (4.01/5 and 4.67/5). A good discrimination of epithelial cells was achieved with our system.

In addition, we studied the importance of epithelial cell masking in immunohistochemistry of Ki-67 and determination of cancer cell proliferation rate. Our findings support the fact that the determination of the Ki-67 labelling index should be restricted to malignant epithelial cells only. From 98 of our test patient samples, 52 tumor images were initially classified as low proliferation without epithelial cell masking. After applying our deep learning based epithelial cell masking, these 52 samples
were re-classified as high proliferation cases. Similarly, Niazi et al., 2018 showed the importance of restricting determination of the Ki-67 labelling index to tumor area alone to standardize the analysis. However, their approach is based on traditional thresholding, and thus, can be quite sensitive to high color variation present in histopathological samples.

Figure 6.4 Examples (A-D) of predicted nuclei locations using three different values for radius $R$ ($R_1$-$R_3$), which is defined as the distance between a ground truth nucleus annotation and a predicted nucleus location. The ground truth annotation is marked with a light blue circular marker and the predictions are marked with a star shaped marker. The color coding for true positive (TP), false negative (FN) and false positive (FP) cases are shown in the upper left subfigure. Light green circle around a predicted nucleus (magenta star) visualises the different $R$ values; $R_1=6$, $R_2=10$, $R_3=14$. In addition, a white line is drawn between each ground truth coordinate annotation and the corresponding true positive prediction. The problems that originate from having coordinate annotations as ground truth detections and defining true predictions are also visible in the figure, such as having a seemingly conflicting annotations with FN and FP in a single nucleus (column E). Reprinted from Publication IV.

Our findings in Publication III indicate that machine learning provides an accurate and effective tool in epithelial cell detection, and an important addition to the Ki-67 analysis. Routine pathology involves a vast amount of time-consuming tasks that could be automated using machine learning. Our results also demonstrate that in routine pathology, there is an actual need for standardised and reproducible di-
agnostics. Here, we also provide an alternative to laborious manual annotation to collect a sufficient amount of labelled data for training a deep learning model and therefore meet the aim III of the thesis study.

6.4 Nucleus detection

In Publication IV, we built a deep learning based system for nucleus detection, and experimented with changes in model generalisation and accuracy caused by sample fixation. We showed that the tissue fixation variability in the training data causes significant variation in the detection accuracies of deep learning based models. Overall, our deep learning based system performed well on the task of nucleus detection from prostate tissue samples (F1-score 0.879). The obtained F1-score was a clear improvement when compared to the F1-score (0.78) reported by Högnäs et al., 2018 that was achieved using traditional watershed based nucleus detection. We evaluated the robustness and generalisation ability of our model using public dataset acquired from multiple hospitals and including tissue images from several patients, disease states, and organs (MoNuSeg, N. Kumar et al., 2017) and achieved good accuracy in nucleus detection (F1-score 0.775). In addition, we wanted to further improve our model generalisation and applied unsupervised domain adaptation to enable generalisation to images from unseen domains. We further trained our model with MoNuSeg dataset using pseudo-label approach and achieved similar results compared to models that were trained with labelled data (F1-scores: 0.7538 (CNN1), 0.8267 (CNN2) reported by N. Kumar et al., 2017). After domain adaptation the F1-score improved from 0.879 to 0.908 on the prostate tissue dataset and from 0.775 to 0.807 on the MoNuSeg dataset.

The key findings of Publication IV address one important factor in the goal of building a robust system for pathologist in the real-world clinical environments, that is, the sources of variability present in histological images. Here, our experiments showed that the data variability causes significant variation in the detection accuracies of deep learning based models and we presented unsupervised domain adaptation as one solution for these challenges. As discussed in the results of Publication II, there is a need for methods that can generalise to data in different clinical environments. The importance of understanding all possible sources of variability in the data, allows us to eventually discover meaningful tissue characteristics and to build a system that can
detect these characteristics while avoiding the data characteristics caused by technical variation. Overall, the study presented in Publication IV met the aims II and IV of the thesis study.
The past several years have shown exciting progress in artificial intelligence to improve the accuracy and efficiency of histopathological sample assessment. The work done for this thesis has been focused on developing and applying machine learning based methods for histopathological assessment, a standard process in clinical diagnostics and preclinical studies. Furthermore, the publications presented here have addressed some of the main challenges and limitations that need to be considered and overcome before applying these methods in clinical practice and for real-world clinical data.

The aim of this thesis work was focused on quantitative characterisation of tissue histology. A key part of histopathological assessment is to know what is normal in the tissue histology before being able to define whether something is missing or extra. Quantitative characterisation of tissue histology provides tools for assessing tissue histology in a numerical manner. The ability to characterise tissue histology and to establish a numerical representation of a tissue sample create a basis for further computational analysis, such as detection and classification of different structures. A commonly used technique for tissue characterisation prior to the widespread adoption of deep learning methods has been feature engineering. One major advantage of feature-based tissue characterisation is the applicability to small sample sizes, which is often the case in studies with animal models. In Publication I, we developed an image analysis workflow for extracting engineered features from histopathological images of mouse prostate tissue. With the proposed feature representation we were able to capture key characteristics within tissue histology and characterise mouse prostate tissue with its normal biological variation. In addition to the characterisation of normal variation within the tissue, we were able to detect subtle changes that precede the manifestation of cancer. As a key contribution of our work in Publication I, we were able to link genetic alterations and histological attributes by using the extracted feature representations. This study confirms that feature engineering can
be a powerful tool in the quantitative characterisation of histological attributes, and thus, meeting the aim of this thesis study. Feature engineering that is able to capture even subtle changes in tissue histology provides a potential for discovering new biological aspects of cancer tissue (Beck et al., 2011) and therefore is an important tool for histopathological assessment.

Even though feature engineering has been successfully used for the characterisation of tissue histology over the past decades, the approach has many limitations, especially related to histopathological image data. Many of the related studies have used statistical texture features derived from the image histogram and co-occurrence matrix that describe the gray level pixel arrangements. A major limitation of statistical texture features in histopathological assessment is their sensitivity to a high amount of variation presented in histopathological data. These limitations can be prevented to some extent by using normalization methods, but in a search for universal features of tissue characteristics that can generalise to data collected from real-world clinical environments, this is a major limitation. An alternative solution to address the variation challenge is gray level invariant features, such as LBP features, or engineering features related to structures, key-points, or edges extracted using methods, such as MSER, SIFT, or HOG operations. Another popular direction for feature engineering in histopathological image analysis has been nuclei count and nuclei arrangement related features, since nuclei count and morphology have clear clinical relevance and provide information on tissue histology. Based on our experiments in Publications I and II, these gray level invariant features as well as features related to certain structures and nuclei arrangement were often the most relevant features for histopathological assessment related detection and classification tasks. These findings emphasize the challenges inflicted by the variability present in histopathological image data on engineering relevant task fitting features.

In addition to the challenges caused by variability present in histopathological data, feature engineering has limitations related to the scalability of the computations. We adapted the image analysis workflow presented in Publication I and scaled up the analysis to metastasis detection from WSIs (CAMELYON16, Bejnordi et al., 2017) in Publication II. A subset of the engineered features related to nuclei arrangement were computationally too heavy for an increased amount of data, and therefore only part of the features presented in Publication I were selected to be used for metastasis detection from WSIs in Publication II. Even though we were able to
detect metastatic tissue with high accuracy using feature engineering, the study reflects one of the core challenges when analysing histopathological WSIs. In order to utilise all the available information of these multigigapixel WSIs with high data heterogeneity, high computational resources and efficient algorithms are required. Complex feature engineering is simply not efficient or scalable enough for increasing amounts of available data, such as large data cohorts consisting of WSIs.

Even though feature-based machine learning is a highly effective method for tasks of detection, segmentation and classification, the challenges and limitations related to the analysis of histopathological data have pushed the transition from feature engineering toward deep learning approaches. Moreover, when visual descriptors learned by a deep convolutional neural network trained on images from entirely different domains were able to outperform task and domain specific feature engineering, the popularity of using feature engineering was dramatically dropped. The emergence of deep learning provided highly accurate tools without the need for expert knowledge and hard manual work of handcrafting and developing domain specific features. Another relevant factor that has been accelerating the transition to deep learning approaches has been advances in graphics processing units (GPUs) aiming to speed up computational processes specifically for deep learning. Modern GPUs have enabled the development of deeper neural network models and scaling up the amount of processed data. Therefore, the benefits of using deep learning are not limited to the increased performance and accuracy, but also the computational effectiveness of being able to process vast amounts of data and learn the most relevant features from the data itself. Following these changes in the image analysis field, we extended our approach presented in Publication II with deep convolutional features and established the increased accuracy and effectiveness of deep learning in the task of metastasis detection (Valkonen et al., 2017). Later, we have developed deep learning approaches for detecting epithelial tissue (Publication III) and cell nuclei (Publication IV), and with these studies meeting the aim II of this thesis. Overall, earlier studies have shown the power of feature-based machine learning in many histopathological image analysis related tasks, however, the deep learning revolution has proven that even higher performance and accuracies can be reached.

The field of digital pathology can benefit significantly from the development of computational algorithms that can bring standardisation and reproducibility to histopathological assessment and clinical diagnostics (Bizzego et al., 2019; Pell et al.,
Since deep learning has become the mainstream methodology for analysing histopathological images, we have witnessed a vast amount of publications presenting a significant potential in the context of diagnostic applications (Campanella et al., 2019; Echle et al., 2021; Esteva et al., 2017; Ström et al., 2020; Yu et al., 2016). In Publication III, the detection of epithelial tissue as a part of determining Ki-67 labelling index is a concrete example of the benefits and need for quantitative approaches in clinical diagnostics. Patient outcome is significantly poorer with rapidly proliferating tumors, which can be determined with the Ki-67 labelling index (Petrelli et al., 2015). However, non-malignant cells should be excluded from the analysis to avoid underestimation of the Ki-67 labelling index, which is a highly error prone task when approximating the cell ratios manually. In Publication III, our experiments confirmed the importance of standardisation in order to avoid misclassification of between high and low proliferating cases. Undoubtedly, it can be stated that computational methods enable quantitative, objective, and efficient image analysis and can bring standardisation and reproducibility to the highly expert oriented process of histopathological assessment.

Most of the seminal work and high accuracy deep learning models have been trained in a supervised manner. However, one of the biggest limitations to training an accurate deep learning model in the field of digital pathology is the lack of labelled data due to the time-consuming annotation process. This challenge is even more prominent with the multigigapixel nature of WSIs. In Publication III, we developed an alternative approach for collecting labelled data for training a deep learning based model in a supervised manner by using cytokeratin staining for generating the pixel-wise tissue labels, and thus met the aim III of this thesis study. Immunohistochemical stainings provide a convenient alternative for time-consuming expert driven manual annotation in order to collect a vast amount of training data for training a deep learning model. While this approach provides an automated alternative for collecting labelled data, it requires a known target of interest and expensive extra stainings. Another alternative approach for overcoming the challenge of limited labelled data is crowdsourcing, which has been increasingly utilised in recent years. Moreover, the public datasets provided as a part of different challenges have proven valuable instruments for overcoming the lack of labelled data as well as accelerating the development of new methods (Hartman et al., 2020). However, there is a limit to how far the field of machine learning can go with supervised learning alone,
since it is impossible to label everything in the world, or in some cases, there simply is not enough labelled data. One example in the field of histopathological image analysis would be a rare cancer case, where there might only be few images available. Therefore, the future of machine learning approaches in the field of digital pathology cannot be dependent on the supervised approaches. One potential direction in the future is in self-supervised approaches combined with weak label strategies, which have recently been increasingly gaining attention in the field (Ciga et al., 2022).

Machine learning can be utilised to recognise difficult relationships and manage enormous amounts of data and find solutions to extraordinarily difficult and time consuming tasks beyond human capabilities. To eventually expand and scale the histopathological analysis to populations and multiple different clinical environments, the generalisation problem needs to be surpassed. The histopathological data differ significantly from many other data types. The high data heterogeneity, originating from technical variation as well as biological variation, poses many computational challenges. In Publication IV, we scratched the surface and studied the variability caused by tissue fixation. This factor alone had an effect on the accuracy of the trained deep learning models. The study provides a view of the large amount of work that remains to be completed and needs to be considered before integrating these systems into clinical practice. One solution is to guide deep convolutional neural networks to learn the right features, the ones that universally describe and characterise tissue histology, and avoid learning features related to technical variation yet, being able to learn visual attributes of normal biological variation and features arising from the malignancies. Domain generalisation methods provide tools for addressing this task.

The main goal in generalisation is to learn features from related domains and generalise to any out-of-distribution target domains. To overcome the domain generalisation problem, several approaches have been proposed, including methods, such as ensemble learning, domain alignment, self-supervised learning (SSL), data augmentation, and domain synthesisization (Zhou et al., 2021). SSL is a branch of study in the field of representation learning that aims to learn generic features by applying transformations to the image data. SSL methods, such as Siamese networks, adjust the model parameters by bringing the outputs of distorted versions of the same image closer together, therefore, these approaches are less prone to overfit to domain-specific biases. Recent studies by Chen et al. (2020) and Azizi et al. (2021) clearly
highlight the advantages of using domain related self-supervised pretraining instead of using networks pretrained with natural images. In addition, self-supervised approaches truly benefit from larger datasets, since their performance seems to keep increasing while training longer when compared to supervised approaches that reach the saturation point earlier. Closely related, domain synthesisation aims to increase the diversity of source domains and improve the out-of-distribution generalisation. Overall, all these approaches have a similar angle in solving the generalisation problem that also supports the findings of our study in Publication IV. In Publication IV, we showed that the best accuracy was achieved by having all the different tissue fixations presented in the training data, and by doing so, steering the model to learn the actual universal features of tissue histology rather than learning domain-specific biases from the data. However, further studies and more extensive datasets are needed to eventually discover how much data diversity is needed to obtain sufficient generalisation.

Despite the high potential of machine learning based diagnostic applications, the remaining major obstacles related to generalisation need to be addressed before clinical adoption. Even though many studies have reached a pathologist level performance for histopathological assessment related tasks, these approaches often suffer a performance drop when analysing data scanned on a different laboratory (Campanella et al., 2019). Therefore, it is essential to acquire larger datasets for the development of these approaches, utilise computational methods that can benefit from the massive datasets and all information within the multigigapixel WSIs and conduct a comprehensive clinical validation of these technologies to actually prove the generalisation to real-world data.
This doctoral research is focused on developing machine learning techniques for analysing histopathological images, and investigating the challenges of utilising these methods as a part of routine pathology workflow. The key findings and contributions of this thesis work can be summarized as follows:

1. Handcrafted features can be utilised in quantitative characterisation of tissue histology and can help in discovering new biological aspects. As the key contributions of our work in Publication I, we were able to detect subtle changes that precede the manifestation of cancer, and to link genetic alterations with histological attributes by using handcrafted quantitative histological characteristics. Furthermore, in Publication II, we were able to extract quantitative histological characteristics that separated metastatic tissue from normal lymph node tissue.

2. Histopathological assessment can significantly benefit from machine learning based methods for tasks of detection, segmentation and classification. As the key contributions of the studies presented in this thesis, we have shown how machine learning can accurately and efficiently be utilised in classification of preneoplastic lesions, in detection and segmentation of metastatic tissue and epithelial tissue, and in detection of cell nuclei.

3. Biomarker guided image labelling facilitates annotation process. In addition to annotation collection, we showed in publication III that biomarker guided cross-modality learning approaches combined with deep learning can be used to even replace expensive physical stains.

4. Tissue fixation is a notable source of variability present in histological images and needs to be considered when training a machine learning model. In publication IV, we showed that having multiple fixatives present in the training data allows to build more generalisable model.
REFERENCES


ki67 in breast cancer working group. *Journal of the National cancer Institute, 103*(22), 1656–1664.


Roberts, N., Magee, D., Song, Y., Brabazon, K., Shires, M., Crellin, D., Orsi, N. M., Quirke, R., Quirke, P., & Treanor, D. (2012). Toward routine use


ings of the IEEE International Conference on Computer Vision Workshops, 27–35.


92
PUBLICATIONS
Analysis of spatial heterogeneity in normal epithelium and preneoplastic alterations in mouse prostate tumor models

Valkonen, M., Ruusuvuori, P., Kartasalo, K., Nykter, M., Visakorpi, T., and Latonen, L.,

*Scientific reports*, 7, 44831

Publication reprinted with the permission of the copyright holders.
Analysis of spatial heterogeneity in normal epithelium and preneoplastic alterations in mouse prostate tumor models

Mira Valkonenb,*, Pekka Ruusuvuori1,2,*, Kimmo Kartasalo1,3, Matti Nykter2, Tapio Visakorpi1,4 & Leena Latonen1,4

Cancer involves histological changes in tissue, which is of primary importance in pathological diagnosis and research. Automated histological analysis requires ability to computationally separate pathological alterations from normal tissue with all its variables. On the other hand, understanding connections between genetic alterations and histological attributes requires development of enhanced analysis methods suitable also for small sample sizes. Here, we set out to develop computational methods for early detection and distinction of prostate cancer-related pathological alterations. We use analysis of features from HE stained images of normal mouse prostate epithelium, distinguishing the descriptors for variability between ventral, lateral, and dorsal lobes. In addition, we use two common prostate cancer models, Hi-Myc and Pten+/- mice, to build a feature-based machine learning model separating the early pathological lesions provoked by these genetic alterations. This work offers a set of computational methods for separation of early neoplastic lesions in the prostates of model mice, and provides proof-of-principle for linking specific tumor genotypes to quantitative histological characteristics. The results obtained show that separation between different spatial locations within the organ, as well as classification between histologies linked to different genetic backgrounds, can be performed with very high specificity and sensitivity.

Tissue histology is one of the main determinants in studying and diagnosing many pathologies, including cancer. Solid tumors change the structure of the tissue due to altered morphologies and localizations of tumor cells within the normal tissue. Histopathology is traditionally a very intuitive method, where decisions are most often based on visual inspection. Often, however, ability to objectively recognize and quantify pathological changes in tissue histology would be desired. Furthermore, gaining the decisive pathological information from basic histological stainings, most often hematoxylin and eosin (HE), would help to avoid using costly special stainings. Several current approaches aim to develop tools to help clinical pathologists to determine presence or state of a particular lesion from HE-stained images, e.g. to stage cancer for diagnostic and prognostic purposes1,2. Yet, a pressing need to diagnose cancer at earlier stages concerns several cancer types, e.g. prostate cancer and breast cancer; when tumors are still small and changes in them more benign, treatment options and prognoses are better. Ability to recognize and quantify small and subtle changes in tissue morphology are crucial also to basic and preclinical research aiming to identify early changes preceding and leading to malignant stages of cancer.

To be able to quantify changes in tissue morphology, measurable determinants of the pathology in question need to be determined. Key questions are how to differentiate between normal and pathological tissue, how to measure the stage of the pathological change, and even how to differentiate between several possible types of change, e.g. subtypes of cancer. For example, accurate separation of pathologies from normal tissue histology requires understanding and inclusion of all states and variables of the normal tissue, whether originating from

1Prostate Cancer Research Center, Faculty of Medicine and Life Sciences and BioMediTech, University of Tampere, Tampere, Finland. 2Tampere University of Technology, Pori, Finland. 3BioMediTech Institute and Faculty of Biomedical Sciences and Engineering, Tampere University of Technology, Tampere, Finland. 4Finlab Laboratories, Tampere University Hospital, Tampere, Finland. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to L.L. (email: leena.latonen@uta.fi)
In cancer research, recent years of next-generation sequencing have revealed the extent of genetic and gene expression alterations in cancer\textsuperscript{7,14}. However, the phenotypic effects of many genetic alterations and their combinations is still under research. The common goal is to be able to subtype tumors better for enhanced patient stratification and care in the future. Combination of genetic information with histology, however, requires that the histological information can be transformed into a quantitative, objective form. This can be achieved through digital imaging and computational analysis of the histological characteristics. While computational image informatics can provide a plethora of quantified descriptors of a given image, the challenge in histology is to sort out the relevant characteristics which can be presented in the form of useful feature representations.

Feature-based analysis combined with supervised learning has been a common approach in decision support systems and computer aided diagnosis based on whole slide images\textsuperscript{15-16}. Such approaches have been successfully used for quantitatively describing characteristics of prostate histology in neoplastic lesions both for a mouse model\textsuperscript{7} and for human tissue\textsuperscript{5}. However, previous studies have left room for improved feature engineering and classification performance.

We aim at improving histological recognition and quantification of pathological features in prostate cancer, and search for descriptors to differentiate early pathological lesions from each other. In here, we use two classical and popular genetic prostate cancer mouse models, namely heterozygous Pten\textsuperscript{−/−} and Hi-Myc\textsuperscript{+}, to perform quantitative image analysis on early neoplasms compared to normal mouse prostatic tissue. With a computational separation of hundreds of features from the whole slide images of histological tissue sections and a random forest based machine learning approach, we find a combination of tissue features able to distinguish between 1) normal epithelium, we applied machine learning. We developed a random forest based model and applied it in

The prostate is shown (Fig. 1C). While the VP shares characteristics within the range of LP, the DP is more clearly differentiated by prostate acini covered with an epithelial cell layer and are surrounded by a basement membrane and loose connective tissue. Between the lobes, subtle differences exist in the orientation of the acinar tubes, somewhat affecting the appearance and lumen size of acini in histological preparations. The epithelium is of specific relevance due to being the tissue component where prostate cancers originate from. The normal appearance of the epithelium varies between the different lobes (Fig. 1A). Epithelium in the DP is columnar, cytoplasm is relatively eosinophilic, and the nuclei are centrally to basally located. The epithelium can be tufted. LP epithelium has only sparse infoldings. The cells are cuboidal or low columnar, and the cytoplasm is less eosinophilic. Nuclei are small and basally located (Fig. 1A).

We manually selected 227 acini to represent variation in prostate epithelium in histological sections, including both the heterogeneity in normal appearance of the tissue, and the technical variance (e.g. acini cut in different orientations). For these images, we performed preprocessing to, for example, correct color variation across the tissue samples, and to exclude areas not including tissue (e.g. empty and secretion-containing areas inside the acini) (Fig. 1B). We applied color deconvolution to separate hematoxylin and eosin stains as separate color channels, and performed nuclear segmentation. The resulting image information we used to compute a compilation of 241 features (listed in Supplementary Table 1). These features included numerous descriptors related to tissue texture and local environments, as well as numeric representations of properties, spatial arrangement, and distribution of nuclei. When these 241 features are used to represent the samples in dimension-reducing t-Distributed Stochastic Neighbor Embedding (t-SNE)\textsuperscript{11} plot, the extent of the extractable spatial variation between the normal lobes of the prostate is shown (Fig. 1C). While the VP shares characteristics within the range of LP, the DP is more clearly distinguished from the other lobes.

**Quantitative characteristics of mPIN lesions.** To study distinction of small pathological changes from normal epithelium, we wanted to compare normal tissue to early pathological lesions. Mice heterozygous for tumor suppressor Pten form mouse prostatic intraepithelial neoplasia (mPIN) within 8–12 months\textsuperscript{7}. In here, we used prostate samples from Pten\textsuperscript{+/-} mice of 10-11 months, when recognizable mPIN is evident (Fig. 2A). We selected 199 areas of mPIN, and performed image processing and feature computation as above. A t-SNE plot (Fig. 2B) shows that the PIN lesions of different prostate lobes are mixed rather than separated as lobe-specific clusters. This indicates that, compared to normal epithelium, the spatial heterogeneity is decreased in mPIN (compare Fig. 2B to Fig. 1C). When comparing the relative distributions of normal epithelium and mPIN lesions, PIN areas are clearly separate from normal tissue areas by the computed compilation of features (Fig. 2C). When comparing the different prostatic lobes, it is evident that LP is furthest and DP closest to PIN lesions based on the computed feature profile, corresponding to the tissue characteristics observed by eye (Fig. 2A).

To test whether the computed feature characteristics can be used to reliably separate mPIN lesions from normal epithelium, we applied machine learning. We developed a random forest based model and applied it in leave-one-out cross validation (LOOCV) to estimate the probability of a sample to belong to the group of PIN lesions based on the feature data. Figure 3A shows the classification confidence given by the machine learning model for each sample to belong to the group of PIN. The accuracy of the estimations by the model was analysed using receiver operating characteristic (ROC) curve, from which the area under curve (AUC) measure can be
used for quantifying the separation between the normal and preneoplastic tissues (Fig. 3B, AUC 0.988). Predictor importances of 20 most influential features in the model to distinguish between the morphology types are shown in Fig. 3C. These feature importances were given by a model that was trained with all available samples.

From the LOOCV experiment, used to validate the robustness of our random forest model, we collected altogether 426 models from which the average importances are shown in Supplementary Figure 1. The averaged feature importance list contains a similar set of features as that given by the original model trained with all available samples (Fig. 3C). These include several types of texture features, such as LBPs and SIFT-features. Nuclear features include several descriptors of nuclear size, density and neighbourhood (NhoodMaxDist, NhoodStdDist, NhoodSkewness, meanNucSize, meanNucDistInNucNB, NhoodMeanDist). Another set of important features are the features describing the relative positions and orientations of the nuclei (NhoodNucAngleSkewABS, NhoodNucAngleKurtABS, NhoodNucAngleVar, NhoodNucAngleStd0). These capture the distinctive property of normal epithelial tissue, where nuclei are most often oriented as “beads in a row” as opposed to scattered distribution in a tumor (Supplemental Fig. 2).

We further tested the ability of the model to distinguish mPIN from normal epithelium in each lobe (Supplementary Figure 3). According to the different incidence of mPIN in the prostatic lobes in the Pten+/- model mice, the number of mPIN samples in the analysis varied between lobes being greatest in LP and lowest in VP (nVP = 37, nLP = 282, nDP = 107). PIN was distinguished from normal prostate most accurately in LP (AUC 0.997), likely due to the clear phenotypic difference in the histology. mPIN in VP and DP were similarly challenging, although with these, relatively small sample sets, a very high accuracy was still reached (AUC 0.972).

**Computational distinction between histologies of different genetic groups.** High expression of oncogenic Myc in the mouse prostate induces neoplastic lesions visible already at 1 month of age36. These lesions develop later on to adenocarcinoma, in contrast to the mPIN in the Pten+/- heterozygous mice which does not develop into carcinoma without additional genetic or carcinogenic manipulation7. We wanted to compare these two types of early neoplasms with genetic differences, and to find computational features separating them from each other and from the normal epithelium. As most of the tumors in these models form in the LP, we...
concentrated on this tissue area selecting only LP samples from $\text{Pten}^+/−$ model mice (normal epithelium $n = 137$, mPIN $n = 145$), and manually selected samples from LP of Hi-Myc mice (normal epithelium $n = 111$, mPIN $n = 189$). Examples of representative histologies on normal and preneoplastic LP epithelium are shown in Fig. 4A.

As the two preneoplastic change types occur at different ages, our samples from the two different models represented prostate epithelium from mice of very different age (10–11 months in $\text{Pten}^+/−$ mice compared to 1 month in Hi-Myc mice). Thus, to ensure consistency and comparability of the data from hiMyc and $\text{Pten}^+/−$ tissues, we selected features whose distributions did not show statistically significant difference (threshold $\alpha = 0.05$) according to Kolmogorov-Smirnov test in the two control groups to use in the further analysis. Altogether 59 features fulfilled the criterion (Supplementary Table 2, Supplementary Figure 4). It is evident already by the feature

---

**Figure 2. Quantitative characteristics of mouse PIN.** (A) Examples of prostatic acini containing normal epithelium and PIN lesions from the three lobular areas (VP, LP, DP). (B) Representation of the quantitative features of PIN lesions in three mouse prostate lobes (VP, LP, DP) in two-dimensional feature space using t-SNE. (C) A t-SNE visualisation reveals decreasing spatial heterogeneity in PIN lesions compared to normal epithelium in the mouse prostate.
mice are a popular model to study prostate cancer, as it is one of the few genetic mouse models forming adenocarcinoma in situ.

In the prostate, these mice develop mPIN, which is a type of lesion not reported to develop to invasive carcinoma without additional genetic or carcinogenic manipulation. Myc, on the other hand, is a transcription factor and a powerful oncogene overexpressed in many cancers, including prostate cancer. Hi-Myc mice are a popular model to study prostate cancer, as it is one of the few genetic mouse models forming adenocarcinoma, thus exhibiting the cancer type most common in human prostate. Both our models represent an early phase in the step-wise development of cancer, and thus can be used in studying the early pathological changes in tissue. We achieved a successful separation between the histologies provoked in these two models by their relative importances in the separation of mPIN from normal epithelium given by a classification model trained with all Pten+/− samples.

values that the three groups of samples have their own, distinct signatures (Fig. 4B). Furthermore, these groups are clearly distinguished in representation of the samples according to the feature values in t-SNE (Fig. 4B).

We developed a machine learning model based on the selected subset of features to estimate the probability of a sample to belong to any of the three groups (normal epithelium, Pten heterozygous mPIN, or Hi-Myc-induced early neoplasia). The model is successful in predicting accurately the histological classes of the samples, as shown in a confusion matrix of the predictions (Fig. 5A) and a ROC-curve of accuracy analysis (Fig. 5B; AUC 0.997 for normal, 0.990 for Pten+/−, and 0.995 for Hi-Myc). Similarly as in distinguishing mPIN from normal epithelium, the predictor importances of the most valuable features in this random forest model (Fig. 5C) include nuclear density and angle-related features (NhoodNucAngleKurtABS; NhoodMeanDist, numberOfNucInNucNB). As expected, however, the separation of two morphologically different neoplastic histologies brings features capturing more detailed textural information from the stain intensities and from nuclear density map to the top list of features (e.g. DensityLBP6, LBP4H). These remain in the list of most influential features even when averaging the predictor importances of 582 trainings of the random forest model (Supplementary Figure 5), and provide detailed, quantitative information of the differences between Hi-Myc and Pten+/−-provoked early neoplasms (Fig. 5D).

Discussion

Automatic recognition and quantitative analysis of tissue pathologies requires ways to computationally identify representative histological features separating normal and altered tissue. In this work, we extracted a set of 241 features from images of prostate tissue, and used them to analyze spatial heterogeneity within normal prostate, as well as to separate different types of early neoplastic changes from normal tissue and from each other. Our results show that separation between different spatial locations within the organ, as well as classification between histologies provoked by different genetic lesions, can be performed with very high specificity and sensitivity.

We applied traditional machine learning approach based on extraction of a large set of engineered features followed by a random forest ensemble classifier. Given the very high accuracy obtained for the relatively low number of samples and small size of regions of interest, this approach appears to be well justified. Recently, neural network based deep learning approaches have gained much attention in efforts to recognize cancerous tissue from normal tissue, and in developing pre-screening tools for pathologists to indicate suspect areas in tissue. The set-back of these approaches so far is the difficulty of examining the relevance and meaning of model properties, i.e. network weights and output layer values used in the decision flow in the context of the tissue. In order to interpret the model properties used computationally to a pathologist or a research biologist, meaningful and recognizable features of the tissue are preferred, especially if the information of the tissue features need to be combined with readouts obtained from other measurement modalities. Methods to avoid the interpretability issues in using deep learning are likely to follow, while combination approaches of deep learning and traditional, feature based machine learning are also being investigated.

In this study, we used two popular prostate cancer mouse models, to compare the histology of early pathologies originating from different genetic alterations. Pten is a tumor suppressor that functions by inhibiting Akt pathway, and it is often deleted in human prostate cancer. Mice heterozygous for Pten form cancer in many organs. In the prostate, these mice develop mPIN, which is a type of in situ lesion not reported to develop to invasive carcinoma without additional genetic or carcinogenic manipulation. Myc, on the other hand, is a transcription factor and a powerful oncogene overexpressed in many cancers, including prostate cancer. Hi-Myc mice are a popular model to study prostate cancer, as it is one of the few genetic mouse models forming adenocarcinoma, thus exhibiting the cancer type most common in human prostate. Both our models represent an early phase in the step-wise development of cancer, and thus can be used in studying the early pathological changes in tissue. We achieved a successful separation between the histologies provoked in these two models by their...
different genetic alterations with very high specificity and sensitivity. This gives promise for future aims to automatically link genetic and quantitative histological information for more varied genetic populations and tumor subtyping.

The method we presented here is generic, and the applicability is not limited to mouse tissue. Tumor material from human patients, however, includes higher genetic and phenotypic variability compared to genetically restricted mouse model material. Thus, well annotated, large enough datasets are required to further develop and validate automated image analysis pipelines for future use in clinical and research applications in human cancer. Another current challenge is to automate the detection step of regions of interest. Recently, machine learning has been applied in automated ROI detection in, e.g., metastasis detection from human breast cancer samples. Our approach for ROI classification could also be applied for ROI detection. This requires quantitative processing of not just the interest areas, but all the neighbouring tissue types and structures as well, including the basement membrane, stromal cells and fibers, nerves, vasculature, smooth muscle, and adipose tissue. Acquiring quantitative representations of these different tissue components and types will benefit applications of digital pathology also beyond cancer research.

In addition to tumor genotypes, data representing tumor phenotypes are desired to combine with quantitative representations of histology provided by methods such as the ones presented here. Recent advances in spatial transcriptomics integrated with computational histological analyses will undoubtedly provide understanding of spatial variation and evolution of tumors. Further, quantitative representation of tissue features and heterogeneity along with three-dimensional reconstructions of whole organs from serial sections, such as the prostate.

Figure 4. Quantitative characteristics of normal epithelium, Pten heterozygous mPIN, and Hi-Myc-induced preneoplasia in mouse lateral prostate. (A) Examples of representative histologies of normal and preneoplastic LP epithelium from Pten+/− and Hi-Myc mouse model prostates. Scale bars 25 μm. (B) Distinct feature value patterns of Pten heterozygous PIN, Hi-Myc-induced early neoplasia, and normal epithelium presented in a heatmap after hierarchical clustering of normalized feature values. (C) Three-dimensional t-SNE visualisation shows distinctive patterns for the three histological populations.
will enable intuitive visualization and provide novel insight into the spatial variation within tissue, as well as tumor growth patterns, in a natural context.

**Methods**

**Ethical permissions.** All animal experimentation and care procedures were carried out in accordance with guidelines and regulations of the national Animal Experiment Board of Finland, and were approved by the board of laboratory animal work of the State Provincial Offices of South Finland (licence numbers ESAVI/6271/04.10.03.02011 and ESAVI/5147/010.07.2015).

**Tissue samples.** FVB/N mice either heterozygous for Pten (Pten+/−) or transgenic for MYC oncogene (Hi-Myc) were used. Prostates were fixed in PAXgene tissue fixative according to manufacturer's protocol, and embedded in paraffin. 5 μm tissue sections were cut, attached to glass slides, and stained with hematoxylin and eosin.

**Imaging and ROI separation.** HE-stained slides were whole slide imaged with Zeiss Axiostar microscope (Carl Zeiss Micromining, NY, USA) with 20x objective and a CCD color camera (QICAM Fast; QImaging, Canada) and a motorized specimen stage (Märzhäuser Wetzlar GmbH, Germany). The automated image acquisition was controlled by the Surveyor imaging system (Objective Imaging, UK). Uncompressed bitmap output was converted by JVSdicom Compressor application to JPEG2000 WSI format. Snapshot images for Figures were obtained through JVSView virtual microscope (http://jvsmicroscope.uta.fi) and ImageJ software (National Institutes of Health, Bethesda, MD, USA). Regions of interest were manually marked using a freehand selection tool in ImageJ. The resulting binary mask was used for extracting the ROI from the full resolution original HE image for further processing. In normal tissue, epithelial layer of prostate acini was included, excluding other tissue components in the organ such as stroma, urethra, vessels and nerve bundles. Pathological lesion masks each included solely mPIN/neoplastic epithelium. In the case of Hi-Myc samples, these were always within a single acinus each. In the case of Pten+/− lesions, where one mPIN tumor could reach several acinar lumen within a certain section, all affected lumen were included in a single mask.

**Preprocessing of images.** The preprocessing of the images included histogram matching in order to remove the color variation between samples, exclusion of unwanted regions, color deconvolution to separate
hematoxylin and eosin stains, and nuclei segmentation. These steps were implemented for bounding box areas around each ROI.

Histogram matching was performed to balance staining variation between sections. For a reference histogram, a mean histogram was computed from a representative set of samples consisting of ROI images of neoplastic lesions and normal prostate epithelium from all three lobular areas. After this, histogram matching was performed by using a transform function computed between the image's histogram and the reference histogram.

To segment the effective tissue area within each ROI, a mask for secretion-filled regions and empty areas was obtained by subtracting different color channels and performing contrast limited mapping of the intensity values similarly as in Ruusuvuori et al.33. The final binary mask was obtained by thresholding using Otsu's method33. To smooth the binary segmentation mask, morphological opening, closing, and filling were performed.

A color deconvolution algorithm34 was applied to convert the red, green, and blue channels of each image into hematoxylin stain, eosin stain, and background. Hematoxylin stains mainly the cell nuclei and therefore, hematoxylin channel was further processed to segment cell nuclei. Maximally stable extremal regions (MSER)27 were extracted from the grayscale image of hematoxylin stain. MSER is a method for blob detection from an image. Set of detected regions were selected based on the size corresponding to potential nuclei size. Additionally, regions that did not contain high grayscale intensity of hematoxylin stain, were excluded.

To get the map for high rate of hematoxylin, tophat filtering, maximum filtering, Gaussian filtering, and image intensity adjustment was performed. Binary mask for high hematoxylin rate was obtained by thresholding using Otsu's method35. To get the final binary mask for cell nuclei, MSER regions that were overlapping with mask for high rate of hematoxylin were selected.

**Feature extraction.** Properties of each ROI were described with extraction of 241 features (Supplementary Table 1). These features included local descriptors related to image texture and distribution of nuclei. Texture features were extracted from local neighborhoods representing distinct levels of tissue architecture and, thus, measured as different features (e.g. Contrast-H, NhoodContrast-H, ROI-BlockContrast-H, ROIContrast-H), and also from both hematoxylin and eosin channels (e.g. Contrast-H, Contrast-E). The neighborhoods for texture features included bounding box of each segmented nucleus, 35 × 35 pixel neighborhood around each nucleus, non-overlapping 50 × 50 pixel neighborhoods within effective tissue area with unwanted regions included and excluded (e.g. mROI-BlockContrast-H, ROI-BlockContrast-H), and the whole bounding box image of the ROI. Nuclei distribution features were extracted from a 100 × 100 pixel neighborhood around each nucleus. To obtain a single feature vector for the whole ROI area, mean feature values were calculated from all blocks presenting one combination of certain feature and neighborhood. Details about each extracted feature and the applied local neighborhoods are presented in Supplementary Table 1.

**Texture features.** The extracted texture based features included, e.g., mean intensity value, contrast, correlation, and energy, calculated from gray level co-occurrence matrix (GLCM). Properties of the texture within each ROI were also extracted using local binary patterns (LBP)28-29 and scale-invariant descriptors obtained via the Scale-invariant feature transform (SIFT)30. Additionally, properties of MSER regions were used as features. VLFeat31 implementations of SIFT and MSER were used in this work.

**Nuclei distribution features.** Features related to distribution of cell nuclei were calculated from a nuclei location map, generated by marking the center point of each segmented nucleus. Features included descriptors related to inter-nuclei distance, nuclei locations with respect to each other described with angular statistics, number of nuclei within a neighborhood, and density features. The density features were calculated from a Gaussian filtered nucleus location map.

The angular statistic features were extracted using CircStat toolbox32. For each nucleus, an angle to all its neighbouring nuclei within a 100 × 100 pixel block was calculated. Supplementary Figure 2 presents an example polar histogram of these calculated angles from both normal epithelium sample and neoplastic lesion sample. The features related to angular statistics included properties of this polar histogram, such as, variance, standard deviation, skewness, and kurtosis.

**Feature selection.** To study the spatial variation in epithelium within different lobular areas of normal tissue, as well as in the comparison of normal tissue and neoplastic mPIN lesions, we used all the extracted 241 features. For the comparison of three groups (normal, \( Pten^+/- \), and \( Pten^-/- \)) and \( Hi-Myc \), a feature selection was performed by statistical testing between features extracted from both \( Hi-Myc \) and \( Pten^+/- \) normal prostate epithelium samples. Two-sample Kolmogorov-Smirnov test55 (significance threshold \( \alpha = 0.05 \)) was used to determine if the feature data extracted from these two normal sample groups were from the same continuous distribution. Altogether 59 features not showing statistically significant difference between the two normal populations were included in further analysis (Supplementary Table 2).

**Classification.** The feature representations of ROI samples were used to train a random forest model36. Random forest algorithm was chosen due to its capability to handle both high data dimensionality and varying sample sizes in a computationally efficient manner. Additionally, the algorithm assigns weights for input features based on their importance in the classification task, providing an interpretable classification model and additional insight in the contribution of features. Bootstrap aggregation, which is a machine learning algorithm that combines multiple versions of decision trees into a random forest model, was used to improve the stability and accuracy of the model. Each decision tree version is constructed from a randomly sampled dataset with replacement. The implemented model was an ensemble of 50 decision trees.

Leave-one-out cross-validation was used for estimating the classification performance of our random forest model. For example, when distinguishing mPIN from normal epithelium, we had 426 samples in total, and
therefore, we trained 426 random forest models. For each model, one sample was left out from the training phase and then the trained model was used to predict the probability for this excluded sample to belong to the group with early neoplastic changes.

From the LOOCV experiments, average feature importances and corresponding standard deviations were compiled. When distinguishing mPIN from normal epithelium, these were calculated from the feature weights given by each of the trained 426 models (Supplementary Figure 1). When distinguishing between Hi-Myc-induced early neoplasia, Pten heterozygous mPIN, and normal epithelium, the average feature importances of the 582 models were calculated (Supplementary Figure 5).

References

Acknowledgements
We thank Ms. Päivi Martikainen, Ms. Marja Pirinen, and Mrs. Marika Vähä-Jaakkola for skillful technical laboratory assistance. Grant support has been obtained from Finnish Funding Agency for Innovation (project Three dimensional histogenomic modeling of whole prostate), Academy of Finland (grant no. 279270), the Cancer Society of Finland, the Sigrid Juselius Foundation, the Medical Research Fund of Tampere University Hospital, Doctoral Programme of Computing and Electrical Engineering at Tampere University of Technology, and Emil Aaltonen Foundation.

www.nature.com/scientificreports/
Author Contributions
M.V. analyzed the data and was the lead developer of the computational methods. P.R. conceived the computational work and participated in implementation and writing. K.K. participated in implementation of computational tools. M.N. and T.V. conceived the study. L.L. conceived and implemented the experimental and histological work. L.L. and P.R. designed the study. L.L. and M.V. wrote the manuscript. All authors reviewed the manuscript.

Additional Information
Supplementary information accompanies this paper at http://www.nature.com/srep

Competing Interests: The authors declare no competing financial interests.


Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2017
Metastasis detection from whole slide images using local features and random forests

Valkonen, M.*, Kartasalo*, K., Liimatainen, K., Nykter, M., Latonen, L., and Ruusuvuori, P.,

*Cytometry Part A, 91(6), 555–565

Publication reprinted with the permission of the copyright holders.
Metastasis Detection from Whole Slide Images Using Local Features and Random Forests

Mira Valkonen,1,2 Kimmo Kartasalo,1,2 Kaisa Liimatainen,1,2 Matti Nykter,1,2 Leena Latonen,1 Pekka Ruusuvuori1,3*

1BioMediTech and Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland
2BioMediTech Institute and Faculty of Biomedical Science and Engineering, Tampere University of Technology, Tampere, Finland
3Faculty of Computing and Electrical Engineering, Tampere University of Technology, Pori, Finland

Grant sponsor: Academy of Finland, Grant number: 269474
Grant sponsor: Tekes - The Finnish Funding Agency for Innovation, Grant number: 26831/2015
Grant sponsor: Cancer Society of Finland, Sigrid Juselius Foundation and Doctoral Programme of Computing and Electrical Engineering, Tampere University of Technology

Additional Supporting Information may be found in the online version of this article.
Mira Valkonen and Kimmo Kartasalo contributed equally to this work.
*Correspondence to: Pekka Ruusuvuori; Tampere University of Technology, P.O. Box 300, 28101, Pori, Finland. E-mail: pekka.ruusuvuori@tut.fi
Published online 20 April 2017 in Wiley Online Library (wileyonlinelibrary.com)
DOI: 10.1002/cyto.a.23089
© 2017 International Society for Advancement of Cytometry

Abstract
Digital pathology has led to a demand for automated detection of regions of interest, such as cancerous tissue, from scanned whole slide images. With accurate methods using image analysis and machine learning, significant speed-up, and savings in costs through increased throughput in histological assessment could be achieved. This article describes a machine learning approach for detection of cancerous tissue from scanned whole slide images. Our method is based on feature engineering and supervised learning with a random forest model. The features extracted from the whole slide images include several local descriptors related to image texture, spatial structure, and distribution of nuclei. The method was evaluated in breast cancer metastasis detection from lymph node samples. Our results show that the method detects metastatic areas with high accuracy (AUC = 0.97–0.98 for tumor detection within whole image area, AUC = 0.84–0.91 for tumor vs. normal tissue detection) and that the method generalizes well for images from more than one laboratory. Further, the method outputs an interpretable classification model, enabling the linking of individual features to differences between tissue types.

Key terms
metastasis detection; digital pathology; computer aided diagnosis; whole slide images; machine learning; random forest; breast cancer; sentinel lymph nodes

Introduction
In recent years, improvements in computational power and whole slide digital scanners have allowed digitalization of histopathological tissue sections and enabled the development of digital pathology into a routine practice (1). Histopathological whole slide images (WSI) contain vast amounts of data, for which digital pathology enables quantitative analysis and the utilization of all available data, allowing for more information to be gained from the images (2,3). This has led to increased interest in the development of image analysis tools for tasks such as automatic detection of regions of interest (4), stain normalization (5), and nuclei detection (6). These advances hold great promise for providing clinical decision support systems for pathologists (7).

Breast cancer is the most common malignant disease in women worldwide (8). In less developed countries, it is the most frequent cause of cancer death in women, while in developed countries it is the second most common cause of cancer death after lung cancer (8). With over 1.7 million new cancer cases diagnosed annually, diagnosis, and treatment of breast cancer poses a humane as well as an economic challenge all over the world.

In breast cancer patients, the main cause of death is metastasis at distant sites of the body. Metastasis in sentinel lymph nodes is one of the most important prognostic
variables in breast cancer (9). Traditional histopathological diagnosis is, however, time-consuming as well as prone to misinterpretation and subjectivity. Automated detection of lymph node metastasis could facilitate the task of pathologists by reducing their workload in breast cancer diagnostics and overcome the subjective interpretation problem (10). Ideally, automated analysis would screen the samples and provide the detected regions for pathologist review, or even proceed directly to decisions. A more realistic scenario is to use automated analysis for pre-screening the images in order to give supporting information and to potentially exclude areas not relevant for diagnosis.

As diagnosis of cancer requires a significant amount of expertise—in practice, a pathologist—it is natural that any automated methods should be capable of incorporating or mimicking such knowledge in their decision making process. Certain qualitative decision rules apply in the diagnosis, and in order to automatize the process, such rules should be replaced by quantitative analysis of numerical data. Supervised machine learning provides a powerful tool for deriving decision rules based on example data. Traditionally, supervised learning involves the process of feature extraction from images prior to applying the learning algorithm. Thus, in addition to providing the teaching samples by outlining regions of tumor content and normal tissue, expert, and prior knowledge can be included in the feature generation step.

A number of studies available in the literature show the great potential of machine learning tools in digital pathology applications, such as in the detection of regions of interest (ROI), or in phenotype, cell type, or tissue type classification, see Refs. 11–15 for recent examples. In order to use learning based methods, a training dataset is required, that is, slides/images for which the ground truth segmentation/annotation of ROIs is available. Typically, this approach utilizes available training data both for determining the decision rules and for selecting the features to be used in the decision process, where the latter property may be either a separate step or belong intrinsically to the classifier design (16,17). Recently, methods relying on built-in automated feature extraction and deep learning, such as convolutional neural networks, have gained ground in classification and detection tasks (18–21). Using the deep learning approach, several breakthrough results in contest challenges and image classification tasks have been achieved (22–24). While appealing due to the high accuracy in tasks where a large amount of training data is available, methodology for interpreting a deep classifier model is currently lacking.

The requirement of a large and representative annotated dataset when applying machine learning for image segmentation poses a challenge in practice (2). Generation of such annotations is expensive, since it requires expertise and time of pathologists, and an extensive amount of manual work especially when considering pixelwise annotations. Thus, datasets of decent size paired with ground truth information are extremely valuable for the community developing the detection and segmentation methods. Recently, challenges and contests organized within conferences in the field of biomedical image analysis have gained interest from the community of image analysis developers. Such events facilitate the sharing of new ideas and best practices. More importantly, they provide annotated datasets for the use of the community. In this study, we use data from the Camelyon16 breast cancer metastasis detection challenge which was organized in conjunction with the IEEE International Symposium on Biomedical Imaging 2016 (http://camelyon16.grand-challenge.org). The challenge dataset contains altogether 270 images obtained at two separate laboratories, each equipped with a different scanner device. The set consists of images from 160 normal samples and 110 tumor samples with cancer metastases outlined by experts, providing a valuable resource for method development and validation purposes.

In this article, we present a method for automated detection of cancer hot-spots in hematoxylin and eosin (H&E) stained WSI of sentinel lymph node sections. Our method is based on feature engineering and machine learning, and it is an extension of the learning-based analysis presented in Ref. 25 into a fully automated WSI analysis pipeline. The proposed system also enables learning about tissue texture, potentially linking the extracted features with growth properties in normal and metastatic tissue. We evaluated the performance of the method in breast cancer metastasis detection via blockwise receiver operating characteristic (ROC) analysis.

**MATERIALS AND METHODS**

**Image Data**

The first dataset used in this study consists of 170 whole slide images of sentinel lymph node sections collected at the Radboud University Medical Center (Nijmegen, the Netherlands). A total of 100 WSIs presented normal lymph node sections and 70 WSIs contained micro- and macro-metastases. Altogether 60 of these cancerous lymph node sections were fully annotated and 10 partially annotated. The second dataset of 100 WSIs was collected at the University Medical Center Utrecht (Utrecht, the Netherlands) and it contains 60 WSIs of normal lymph node sections and 40 WSIs with lymph node metastases. Of the 40 cancerous slides, 37 were fully annotated and 3 partially annotated. Both datasets were provided for the Camelyon16 challenge (http://camelyon16.grand-challenge.org). The whole slide images and the corresponding annotation masks were provided as multi-resolution pyramids in Phillips BigTIFF format. The pixel size of the images at the full resolution level was 243 nm. We used the fully annotated slides to obtain both positive and negative training examples. The partially annotated slides were only used to obtain positive examples to avoid the risk of using unannotated metastatic regions as negative training data.

**System Overview**

An overview of the system presented in this study is shown in Figure 1. As preprocessing steps, we segment the tissue region, and apply color correction through matching the color space to that of a reference image. Color correction is needed for the purpose of generalizing the method to inputs with different characteristics due to scanner and staining...
protocols. The feature engineering phase is tailored to the extraction of a large set of quantitative descriptors of image texture, spatial structure, and distribution of nuclei. The machine learning module applies a random forest model learned from the annotated samples, which outputs confidence values indicating the likelihood of cancer cells being present in the corresponding part of the image. Depending on the exact application at hand, these maps of confidence values can be further refined to for example classify entire slides as negative or positive, visualize hotspots of cancer cells for the pathologist to focus on or numerically quantify the properties of detected lesions. Individual steps of the pipeline are described in more detail in the following sections.

**Figure 1.** The analysis workflow for training (upper half) and classification (lower half). During model training, the lymph node tissue (blue outline) is first segmented from the whole slide image containing annotated metastatic regions (yellow outline). The detected tissue sections are then divided into 8,192 × 8,192 pixel RGB subimages and subjected to an optional stain normalization step. Eosin and hematoxylin channels are separated from each subimage using a color deconvolution approach. Tissue blocks of 200 × 200 pixels are then randomly sampled from normal (boxes outlined in green) and cancerous (boxes outlined in red) regions from both channels. Features are extracted from each tissue block to get feature vector representations, which are fed to a random forest model as training data. During classification, the workflow proceeds similarly until the extraction of eosin and hematoxylin channels. Instead of random sampling, all 200 × 200 pixel blocks (boxes outlined in blue) are analyzed from each stain channel and fed to the feature extraction module. The trained random forest model is then used to classify each test block and as an output the model assigns a confidence value associated with its choice. Confidence value is an estimate of probability for a sample block to belong to the group of cancerous tissue. This confidence value is assigned for each tissue block to get a confidence map for the entire WSI as an output. Here, the ground truth annotations are overlaid in yellow on the confidence map for reference. Depending on the application, the confidence maps can be further refined to obtain different final outputs, such as binary classification of entire slides, visualizations of cancer hotspots or quantification of the properties of detected lesions.

**Tissue Segmentation**

In order to simplify the classification task and to reduce the amount of data, we first performed a rough segmentation step for each image to detect the lymph node tissue while excluding the background and most of the adipose tissue. The segmentation procedure applied to a single image consisted of the following steps:

1. Compute the HSV transform of the image.
2. Filter the S component using a circular Gaussian kernel (standard deviation = 50 pixels) to blur subcellular details which are not relevant for segmenting the tissue region.
3. Apply a threshold of $0.5 \times \text{I}_{\text{Otsu}}$ to $S$, where $\text{I}_{\text{Otsu}}$ is the value obtained using Otsu’s method (26), to obtain a binary image $B$.

4. Exclude objects in $B$ with aspect ratio (defined as major axis length per minor axis length) over 10 or mean value of the $V$ component under a fixed threshold (here: 0.3). These objects are dark and thin artifacts caused by cover-slip edges.

5. Perform dilation for $B$ using a disk-shaped structuring element with a radius of 50 pixels to obtain smooth object boundaries.

6. Fill holes within objects in $B$.

7. Exclude pixels close to the image’s edges in $B$. Pixels on the left and right side or the top and bottom are excluded if their distance from the closest edge is less than 2% of the image’s width or height, respectively.

8. Exclude objects in $B$ with area under a desired limit (500,000 pixels). These small objects represent remaining debris or very small torn-off pieces of tissue.

The value of 50 pixels (~12 μm) was selected for the smoothing operations in steps 2 and 3 based on the consideration that details smaller than this are mainly subcellular and can be neglected when detecting the gross boundaries of the tissue slice. A constant multiplier of 0.5 was introduced in step 3 to avoid losing faintly stained lymph node tissue, while still excluding the background and most of the weakly stained adipose tissue. The thresholds in steps 4, 7 and 8 were selected experimentally to exclude most of the debris and imaging artifacts present in the images. For the tissue segmentation, we used the fifth image in the resolution pyramid stored within the input TIF files. The images on this level had been downsampled by a factor of 16. All values given above in pixels are the input TIF files. The images on this level had been downsampled by a factor of 16. All values given above in pixels are used the fifth image in the resolution pyramid stored within the input TIF files. The images on this level had been downsampled by a factor of 16. All values given above in pixels are used the fifth image in the resolution pyramid stored within the input TIF files. The images on this level had been downsampled by a factor of 16. All values given above in pixels are.

558

Metastasis Detection From Whole Slide Images
Feature Extraction

The properties of each tissue sample block were described with 104 texture features extracted from both hematoxylin and eosin channels. See Supplementary Table for a full list of features with descriptions. Texture features included, for example, contrast, correlation and energy, calculated from the gray level co-occurrence matrix (GLCM). Spatial sampling parameters for the gray level co-occurrence matrix were distance of one pixel and 8 directions. More specifically, the co-occurrences of gray values were computed for all adjacent pixels including corner pixels at distance of one pixel. The texture of each tissue sample block was further described using local binary patterns (LBP) (31,32). This texture operator is a measure of the spatial structure of local image intensities. The basic idea of the LBP operator is to transform a local circular neighborhood into a binary pattern by thresholding the neighborhood with the gray value of the center pixel. Due to this thresholding, the features are robust in terms of gray scale variations caused by changes in illumination caused by, for example, different scanners. The circularly symmetric neighborhood is determined by assigning parameters that control the quantization of the angular space and radius of the neighborhood. In our method, we used radius of 2 pixels with angular space of 8 points. By applying a shift operation, the extracted LBP features are also rotation-invariant. Other extracted texture features were scale-invariant descriptors obtained via the Scale-invariant feature transform (SIFT) (33), the histogram of oriented gradients (HOG) descriptor (34,35), and maximally stable extremal regions (MSER) (36). In this work, the VLFeat (37) implementation of MSER and SIFT was used.

In addition to the texture features, six nuclei density features were extracted, calculated from a nuclei location map. This location map was generated by marking the center point of each segmented nuclei. Nuclei density features included descriptors related to inter-nuclei distance inside the sample block, such as mean, maximum, minimum and standard deviation. Also density and number of nuclei inside the sample block were calculated. The density feature was the mean value of the Gaussian filtered sample block from the nuclei location map.

Model Comparison

For selecting the learning algorithm, we compared the performance of a number of different models for classifying the sample blocks as either normal or tumor tissue based on the extracted features. Approximately 1,000,000 sample blocks were randomly selected and used to train a linear regression model, a support vector machine (SVM), a random forest model and two nearest neighbor (NN) classifiers, one using all the features and one using a subset of 28 manually selected features which roughly corresponds to the feature set in Ref. 38 in single resolution. The trained regression model is a generalized linear regression model for the binomial distribution using logit link function. The SVM model utilizes a nonlinear radial basis function as a kernel function and grid-search was used to find the optimal values for kernel size and soft margin. NN classifiers utilize kd-tree search to find the Euclidean distance to the closest neighbor.

Sensitivity, specificity, F-score and the percentage of correctly classified samples are shown for each method in Table 1. The random forest model outperformed the other models in terms of correctly classified samples, sensitivity, and F-score. The specificity of the NN classifier was higher than that of the random forest (96.8% vs. 93.3%). However, as this was at the expense of much lower sensitivity (85.7% vs. 92.6%), and the random forest model had a higher percentage of correctly classified samples (93.0% vs. 91.3%), and a higher F-score (0.93 vs 0.91), we selected the random forest model as the learning algorithm for our system.

Random Forest Model

We used the feature representations of tissue samples to train a random forest model (17). The model was an ensemble of 50 classification trees. The number of features selected randomly for each decision split was the square root of the total number of features. Bootstrap aggregation was used to improve the stability and accuracy of the model. Bootstrap aggregation is a machine learning algorithm that combines multiple versions of decision trees into a random forest model. Each decision tree version is constructed from a randomly sampled dataset with replacement. The trained model was then used to evaluate the test images. About 214 features were

<table>
<thead>
<tr>
<th>Model</th>
<th>CORRECTLY CLASSIFIED SAMPLES (%)</th>
<th>SENSITIVITY (%)</th>
<th>SPECIFICITY (%)</th>
<th>F-SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>87.0</td>
<td>86.4</td>
<td>87.6</td>
<td>0.87</td>
</tr>
<tr>
<td>NN</td>
<td>82.8</td>
<td>74.4</td>
<td>91.0</td>
<td>0.81</td>
</tr>
<tr>
<td>NN feature subset</td>
<td>91.3</td>
<td>85.7</td>
<td>96.8</td>
<td>0.91</td>
</tr>
<tr>
<td>Random forest</td>
<td>93.0</td>
<td>92.6</td>
<td>93.3</td>
<td>0.93</td>
</tr>
<tr>
<td>SVM</td>
<td>88.3</td>
<td>85.9</td>
<td>90.6</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Approximately 1,000,000 sample blocks were classified using the following models: logistic regression, nearest neighbor (NN) using either all or a subset of features, support vector machine (SVM) and a random forest model. Percentage of correctly classified samples, sensitivity, specificity, and F-score are shown for each model.
extracted from each 200 × 200 pixel block in test subimages. The confidence for being either a normal tissue block or a tumor tissue block was predicted with the trained random forest model. These subimage confidences were stored in unsigned 8-bit integer format and pieced together to form a metastasis confidence image for each test WSI. Since a single confidence value is predicted for each 200 × 200 pixel block, the size of the resulting confidence images corresponds to a 1:200 downsampling of the original WSIs along each dimension.

Training of one random forest model with 700,000 training samples takes approximately 90 minutes. To classify a new WSI with a trained random forest model, our method takes approximately 130 minutes. The processing time varies of course depending on the amount of tissue in WSI. These computation times for training the random forest model and processing of one WSI are obtained using parallel computing with 95 GB of memory and two six-core Intel X5660 processors.

RESULTS

Detection of metastatic regions from whole slide images was evaluated with the data from the Camelyon 2016 contest. First, we determined the performance for a set of 170 images from a single scanner, eliminating the variability of source images due to technical reasons. Leave-one-out cross-validation (LOOCV) was used to assess the performance of our random forest classification approach. Each sample from one WSI not used in training was scored with confidence levels using a random forest model trained with all the samples from 169 other images.

To interpret our random forest model, we visualized predictor importance weights assigned by the model for each feature. These weights are higher for the features that have higher impact on the correct classification result. Weight estimates for every feature are based on changes in the mean squared error due to splits in every decision tree. The averaged feature importance’s of the 10 most significant features for the LOOCV experiment are shown in Figure 2A. An example area of normal (Fig. 2B) and tumor tissue (Fig. 2C), as well as the feature values for the same areas, are shown in Figures 2B1–10 and 2C1–10. The majority of the ten most significant features were calculated from the hematoxylin channel, excluding the NumberOfNuc-feature, which is based on the binary image of segmented nuclei and e-LBP9, which is calculated from the eosin channel. Differences in feature values between normal and tumor samples are clearly visible for most of the ten features. LBP-3, number of nuclei, LBP-7, contrast, LBP-8, correlation, LBP-6, and LBP-4 all tend to be higher in normal lymph node tissue than in cancerous areas (Figs. 2B1–10 and 2C1–10). Of these, LBP-3 (Figs. 2B1 and 2C1) and correlation (Figs. 2B8 and 2C8) seem most robust in tolerating the follicular material in addition to lymph node cortex, representing the normal variation in the lymph node tissue. eLBP-9 (Figs. 2B4 and 2C4) and kurtosis (Figs. 2B7 and 2C7) signals were higher in cancerous material than in the normal tissue. Contrast (Figs. 2B5 and 2C5) is especially low in cytoplasm-rich cancer cells and high in lymph node cortex and helpful in finding especially large areas of metastases.

An example of a classification result for a WSI is shown in Figure 3. An original image of a tumor sample with pathologist’s annotations overlaid in yellow is presented in Figure 3A. The corresponding confidence values given by the random forest classifier are shown as an image in Figure 3B. The higher confidence values are concentrated in areas marked as tumor by the pathologist, while confidences in normal tissue area are generally lower, with occasional higher hits scattered around the tissue. The visual appearance of the example result in Figures 3A and 3B suggests that the classifier is able to detect the metastatic areas.

In order to evaluate the performance of our system numerically, we collected all confidence values within normal and tumor tissue areas for all 170 images of the first dataset in

Figure 2. Relative importance of the 10 most significant features selected by the random forest model (A). Example H&E images of normal tissue (B) and metastatic tissue (C) are shown with the corresponding features computed from the hematoxylin (H) or eosin (E) channel: local binary pattern 3 (H) (B1 and C1), number of nuclei (B2 and C2), local binary pattern 7 (H) (B3 and C3), local binary pattern 9 (E) (B4 and C4), contrast (H) (B5 and C5), local binary pattern 8 (H) (B6 and C6), kurtosis of the intensity distribution (H) (B7 and C7), correlation (H) (B8 and C8), local binary pattern 6 (H) (B9 and C9) and local binary pattern 4 (H) (B10 and C10). The intensity scales in 1–10 are comparable between each feature pair B and C. [Color figure can be viewed at wileyonlinelibrary.com]
the LOOCV experiment (Fig. 4A) and calculated the blockwise ROC curve both for the whole image area (Fig. 4B) and for the lymph node tissue areas with the background excluded (Fig. 4C). Next, we applied the same computational pipeline to the second image dataset containing 100 WSIs scanned with another device to obtain the corresponding confidence WSIs. We again collected all confidence values within normal and tumor tissue areas (Fig. 4D) and calculated the blockwise ROC curves for all blocks and tissue blocks only (Figs. 4E and 4F, respectively). Partially annotated images were excluded from all numerical evaluations. The mean area under the curve (AUC) value for metastatic tumor versus all image blocks including background was 0.983 for the first image set (Fig. 4B) and 0.975 for the second set (Fig. 4E). For metastatic tumor versus normal tissue, the mean AUC value was 0.905 for first image set (Fig. 4C) and 0.887 for the second set (Fig. 4F). The numerical results in Figure 4 support the conclusions drawn from the visual example in Figure 3.

In order to determine the generalizability of our approach to datasets with more variability, containing images originating from different laboratories and imaged with different scanners, we combined the two datasets. Although representing the same tissue and in principle processed with a similar H&E staining procedure, the visual appearance of the tissues differs between the images from the two laboratories, as can be seen from the example images in Figures 3A and 3C. We trained our RF model with 700,000 samples from the combined dataset and conducted the LOOCV experiment for all of the 270 images. The confidence values from normal and metastatic tumor tissue areas (Fig. 5G) and the blockwise ROC curves from all image blocks (Fig. 5H, mean AUC = 0.985) or tissue blocks only (Fig. 5I, mean AUC = 0.902) indicate that the method generalizes well to datasets containing images from different laboratories. The effect of metastasis size on the detection accuracy was examined by separately considering tissue blocks from metastatic regions larger and smaller than the median area (0.1867 mm²) of all regions in the LOOCV ROC analysis of the combined dataset. In line with the approach adopted in the Camelyon16 competition, we considered all regions annotated in the ground truth masks with area larger than that of a circle having a radius of 100 μm. This analysis resulted in AUC values of 0.801, 95% CI [0.787, 0.814] and 0.906, 95% CI [0.896, 0.916] for the small and large metastatic regions, respectively.

Finally, we used the two independent image sets in turn as a training set and as a testing set to determine if the system is capable of handling the situation where the testing data are markedly different from the data used for training. First, we trained our RF model with 350,000 samples collected from the first set of 170 WSIs and evaluated the 100 WSIs from the second set. Then, we trained the RF model with 350,000 samples collected from the second set of 100 WSIs and evaluated the 170 WSIs from the first set. The results of this experiment are presented in Figures 5J–5L for the former and in Figures 5M–5O for the latter case. The distributions of confidence values and the ROC analysis for all image blocks (mean AUC = 0.970 and mean AUC = 0.978) and tissue blocks only (mean AUC = 0.839 and mean AUC = 0.855) indicate that classification accuracy remains relatively high even when the testing data are completely independent of the training data and have different characteristics, although a slight decrease in performance is observed compared with the LOOCV results.

Most false positive signals were detected where normal lymph node medulla was misinterpreted as cancerous tissue (Fig. 5A). The reticular cells forming the lymph node stroma have partly similar color tones and size of nuclei as certain breast cancer cell phenotypes, especially in areas surrounding lymph node trabeculae and/or vasculature. False positive signals were occasionally resulting also from nerve bundles cut in

![Figure 3.](image-url)

An example whole slide image from the first dataset (A) with the corresponding confidence map (B) and an example whole slide image from the second dataset (C) with the corresponding confidence map (D). Ground truth annotations are shown in yellow.
Figure 4. Results obtained using leave-one-out cross validation for dataset 1 (A–C), dataset 2 (D–F) or the combined dataset (G–I) and for a classifier trained on dataset 1 and evaluated on dataset 2 (J–L) or for a classifier trained on dataset 2 and evaluated on dataset 1 (M–O). Distribution of confidence values for all normal and tumor tissue blocks in the dataset is shown in (A, D, G, J, M). The red line represents the median, the edges of the blue box correspond to the 25th and 75th percentiles and the length of the whiskers is 1.5 times the interquartile range. Outliers beyond this limit are shown in red. Blockwise ROC curves are shown for all blocks in (B, E, H, K, N) and for tissue blocks only in (C, F, I, L, O). The solid lines represent the mean and the dashed lines represent the pointwise 95% confidence interval. Corresponding AUC values are shown above each ROC curve. The total number of classified blocks was 85,545,658 (dataset 1, all blocks), 6,393,412 (dataset 1, tissue blocks), 29,660,702 (dataset 2, all blocks), or 5,301,888 (dataset 2, tissue blocks). [Color figure can be viewed at wileyonlinelibrary.com]
such an orientation that an approximately similar ratio of blue nuclei to surrounding light pink material was created, where myelin sheets in nerve bundles resembled the appearance of the cytoplasm of cancer cells (Fig. 5B). Some out-of-focus image areas also resulted in false positive signals (Fig. 5C). False negative signals were detected in especially infiltrative areas (Fig. 5D) or small metastases, where single or only a few cancer cells are surrounded by lymphocytic cells.

The blockwise confidence output can be used as a starting point for other tasks. Ideally, automated analysis would screen the WSIs and for example provide the detected cancerous regions for pathologist’s review or perform slide-level classification to exclude some slides as completely negative for cancer. To provide an example of further analyzing the WSI confidence maps and to determine the generalization capability of our computational pipeline, we finally used our approach for slide-level binary classification. We used the same feature extraction and random forest classification approach as in the earlier experiments but this time, the input to the classifier was the WSI confidence map (in other words, the output from the classification model for an H&E WSI) instead of the underlying tissue image. The same 104 texture features, which were extracted from each hematoxylin or eosin sample block, were now extracted from the WSI confidence map. These features were then used to train our RF model to separate the normal WSIs from the WSIs containing metastasis. LOOCV was used to determine one confidence value for each of the 270 WSIs indicating the likelihood for the whole slide to contain any metastatic tissue. We collected all whole slide confidence values and calculated the image-wise ROC curve and obtained a mean AUC value of 0.73 for metastasis-containing WSIs versus normal WSIs. This example demonstrates the generic nature of the features used in our system and exemplifies one possible approach for utilizing the WSI confidence maps for downstream analysis, such as for slide-level classification between cancer versus normal.

**DISCUSSION**

Automated processing of whole slide images and detection of regions of interest is an open challenge in digital pathology based cancer diagnosis (14). Herein, we developed a method for automated detection of hot-spot regions in whole slide images. The feature based classification approach presented here is generic and can be applied to a variety of segmentation and detection tasks. We evaluated the performance of the method in detection of breast cancer metastases in lymph node sections from H&E stained WSI. This detection task represents an interesting challenge for digital pathology, since one of the major factors in breast cancer prognostics is metastasis of cancer cells to sentinel lymph nodes (9). The diagnostic procedure for pathologists is currently tedious and time-consuming, as well as prone to misinterpretation. Automated detection of lymph node metastases has great potential to help the pathologist to improve diagnostics as well as to reduce both the workload and costs. Our anticipation is that the method presented in this study is useful for the detection of hot-spots, including the task of separating regions of metastatic breast cancer cells from normal lymphatic tissue composed of lymphocytes. Qualitative (Fig. 3) and quantitative (Fig. 4) results support this anticipation.

From the pathologist’s viewpoint, the sensitivity of the method (Table 1) and the confidence map provided by the method of the possible hotspots in each slide are the most useful parameters for pre-screening the slides to help focus on
suspect areas. In addition to the hot-spot (here: metastatic tumor tissue) detection, our method enables linking the differences between tissue types in hot-spot areas versus normal tissue to specific features describing the tissue properties. This can potentially provide insights into the tissue type characteristics or even suggest differences in growth patterns. The average random forest model obtained in the cross validation study was illustrated in Figure 2A. The top ten most important features contributing to the classifier model are in practice the descriptors which behave differently in normal and metastatic tumor tissue areas. While part of them are not straightforwardly interpretable, there are also some features that either support existing knowledge (e.g., nuclear count in local neighborhood, Figs. 2B2 and 2C2) or stand out as candidates for straightforward computational readouts (e.g., local contrast, Figs. 2B5 and 2C5).

Evaluating the performance of methods for cancer detection from digitized slides is a non-trivial task (2). Obtaining ground truth annotations can be a very laborious process and represents a significant bottleneck in the development of new methods. Even if this issue can be overcome to obtain large, annotated datasets, as in the case of the Camelyon16 challenge, the problem of designing a relevant performance metric remains. The selection of a suitable evaluation metric depends heavily on the way the method is intended to be used in a practical setting. If the aim is to, for example, classify entirely whole slide images (WSIs) as either normal or tumor containing, it is sensible to evaluate performance using slide-level receiver operating characteristic (ROC) analysis. This approach was adopted by us in our slide-level classification experiment and as the first metric in the Camelyon16 challenge. If, on the other hand, the intention is to use the method to pinpoint suspicious areas in the images to speed up the work of pathologists, as in the case of metastasis detection from lymph node sections, performance must be evaluated in a pixelwise, blockwise, or region-based manner for each WSI. As an example, for the second evaluation metric of the Camelyon16 challenge, participants of the competition had to provide a single coordinate and a confidence value for each metastatic region detected from the images. Coordinates located within annotated tumor regions were considered as correct detections and the teams were ranked according to the area under the curve (AUC) metric computed based on free-response receiver operating characteristic (FROC) analysis. This metric relies on scoring a single coordinate point per region as either a hit or a miss, instead of evaluating the identification of the actual regions. However, accurate detection of the boundaries of metastatic areas is a prerequisite for further computational analysis of their size, shape and numerous other characteristics. Moreover, selecting a single coordinate to represent the entire cancerous region in a meaningful way is problematic, especially for regions with a complicated shape featuring, for example, protrusions.

Considering the above, in this study we treated the metastasis detection task as a blockwise classification problem and evaluated the performance of our method by ROC analysis applied to the 200 × 200 pixel blocks. A similar approach has been used for example to evaluate the performance of classifiers applied to non-small cell lung cancer samples (39). In comparison to the Camelyon16 measure, blockwise or pixelwise metrics take into consideration the entire tumor regions and avoid the artificial coordinate selection step. The downside of blockwise evaluation is that larger tumor regions attain more weight in the final scoring, as they consist of a larger number of pixels than smaller lesions.

This is problematic in the sense that examining the slides for micrometastases or individual tumor cells can be very time-consuming for the pathologist, while large macrometastases can often be spotted more easily. In the context of computer aided diagnosis, the capability to accurately detect small tumor regions should thus not be neglected during evaluation. Still, in the absence of a universal evaluation metric suitable for all intended applications, the blockwise metrics represent a straightforward application-independent approach for quantifying detection performance in a task that can be seen as the basis for all further steps—discrimination between target and non-target areas in an image. Good performance in this task is a prerequisite for the consequent delineation of entire metastatic regions, binary classification of entire WSIs and other more refined analysis steps, and should thus be a common characteristic of all well-performing methods.

In addition to performing large-scale numerical evaluation using the entire dataset, we also visually examined examples of different normal and metastatic tissue areas, which had been either successfully or unsuccessfully detected. Normal lymph nodes are composed of primarily lymphocytic cells and follicles structured along a supportive reticular network. The appearance of cancer cells of epithelial origin is most often well distinguishable from especially the lymph cell component of lymph nodes with their relatively large size, prominent presence of cytoplasm and light staining of nuclei. However, there are phenotypically various cancer cell types, and the growth pattern within the lymph node may affect the classification outcome. Most nodular metastatic lesions are easily distinguishable from our method. In contrast, especially small metastatic lesions with only a few cells and especially with an invasive growth pattern alongside normal tissue structures are more challenging for the method to detect.

False positives occasionally emerged at certain areas of normal lymph node medulla. This seems to be due to that the reticular cells forming the lymph node stroma have partly similar color tones and size of nuclei as certain breast cancer cell phenotypes, especially in areas surrounding lymph node trabeculae. Another source of error was out-of-focus image areas, emphasizing the importance of consistently high technical quality of the images. False negative signals were mainly associated to small metastases with a small number of cancer cells or especially infiltrative metastatic growth patterns. In these cases, cancer cells appeared as single cells, or small groups of cells were surrounded by lymphocytic cells. A probable reason for the weaker performance observed in such tissue regions is that many of the analyzed subimages in these regions contain some normal tissue in addition to cancer cells. The feature values computed from such subimages partly resemble those obtained from entirely normal tissue,

**Original Article**

Evaluating the performance of methods for cancer detection from digitized slides is a non-trivial task (2). Obtaining ground truth annotations can be a very laborious process and represents a significant bottleneck in the development of new methods. Even if this issue can be overcome to obtain large, annotated datasets, as in the case of the Camelyon16 challenge, the problem of designing a relevant performance metric remains. The selection of a suitable evaluation metric depends heavily on the way the method is intended to be used in a practical setting. If the aim is to, for example, classify entirely whole slide images (WSIs) as either normal or tumor containing, it is sensible to evaluate performance using slide-level receiver operating characteristic (ROC) analysis. This approach was adopted by us in our slide-level classification experiment and as the first metric in the Camelyon16 challenge. If, on the other hand, the intention is to use the method to pinpoint suspicious areas in the images to speed up the work of pathologists, as in the case of metastasis detection from lymph node sections, performance must be evaluated in a pixelwise, blockwise, or region-based manner for each WSI. As an example, for the second evaluation metric of the Camelyon16 challenge, participants of the competition had to provide a single coordinate and a confidence value for each metastatic region detected from the images. Coordinates located within annotated tumor regions were considered as correct detections and the teams were ranked according to the area under the curve (AUC) metric computed based on free-response receiver operating characteristic (FROC) analysis. This metric relies on scoring a single coordinate point per region as either a hit or a miss, instead of evaluating the identification of the actual regions. However, accurate detection of the boundaries of metastatic areas is a prerequisite for further computational analysis of their size, shape and numerous other characteristics. Moreover, selecting a single coordinate to represent the entire cancerous region in a meaningful way is problematic, especially for regions with a complicated shape featuring, for example, protrusions.

Considering the above, in this study we treated the metastasis detection task as a blockwise classification problem and evaluated the performance of our method by ROC analysis applied to the 200 × 200 pixel blocks. A similar approach has been used for example to evaluate the performance of classifiers applied to non-small cell lung cancer samples (39). In comparison to the Camelyon16 measure, blockwise or pixelwise metrics take into consideration the entire tumor regions and avoid the artificial coordinate selection step. The downside of blockwise evaluation is that larger tumor regions attain more weight in the final scoring, as they consist of a larger number of pixels than smaller lesions.

This is problematic in the sense that examining the slides for micrometastases or individual tumor cells can be very time-consuming for the pathologist, while large macrometastases can often be spotted more easily. In the context of computer aided diagnosis, the capability to accurately detect small tumor regions should thus not be neglected during evaluation. Still, in the absence of a universal evaluation metric suitable for all intended applications, the blockwise metrics represent a straightforward application-independent approach for quantifying detection performance in a task that can be seen as the basis for all further steps—discrimination between target and non-target areas in an image. Good performance in this task is a prerequisite for the consequent delineation of entire metastatic regions, binary classification of entire WSIs and other more refined analysis steps, and should thus be a common characteristic of all well-performing methods.

In addition to performing large-scale numerical evaluation using the entire dataset, we also visually examined examples of different normal and metastatic tissue areas, which had been either successfully or unsuccessfully detected. Normal lymph nodes are composed of primarily lymphocytic cells and follicles structured along a supportive reticular network. The appearance of cancer cells of epithelial origin is most often well distinguishable from especially the lymph cell component of lymph nodes with their relatively large size, prominent presence of cytoplasm and light staining of nuclei. However, there are phenotypically various cancer cell types, and the growth pattern within the lymph node may affect the classification outcome. Most nodular metastatic lesions are easily distinguishable from our method. In contrast, especially small metastatic lesions with only a few cells and especially with an invasive growth pattern alongside normal tissue structures are more challenging for the method to detect.

False positives occasionally emerged at certain areas of normal lymph node medulla. This seems to be due to that the reticular cells forming the lymph node stroma have partly similar color tones and size of nuclei as certain breast cancer cell phenotypes, especially in areas surrounding lymph node trabeculae. Another source of error was out-of-focus image areas, emphasizing the importance of consistently high technical quality of the images. False negative signals were mainly associated to small metastases with a small number of cancer cells or especially infiltrative metastatic growth patterns. In these cases, cancer cells appeared as single cells, or small groups of cells were surrounded by lymphocytic cells. A probable reason for the weaker performance observed in such tissue regions is that many of the analyzed subimages in these regions contain some normal tissue in addition to cancer cells. The feature values computed from such subimages partly resemble those obtained from entirely normal tissue,
which leads to false negatives. Improved performance in these kinds of regions could possibly be achieved by using a multi-scale approach, where the size of the analysis window would be varied over a certain range, and/or by utilizing superpixels (4).

In conclusion, the machine learning based approach for detecting metastatic tissue regions presented in this article performs well in blockwise detection of breast cancer metastases from lymph node tissue sections. The method was applied to whole slide images of H&E stained tissue obtained using two different scanners at two separate laboratories. Even though H&E images were used here, the presented method is generic in nature, and the information extracted from other histological images can be included in our analysis pipeline in a straightforward manner. The method is extendable also in the sense that it allows the incorporation of any number of new features that can be extracted from H&E images and, when available, other measurements from the same spatial location, such as images of immunohistochemically stained samples. Other potential places for improvement and further study include applying more advanced strategies for training, such as using misclassification from the cross validation step for boosting the classifier in a re-training step. Furthermore, deep learning based methods have been used in similar tasks with very high detection accuracy (40, 41). The presented classification pipeline could benefit from complementing the feature extraction phase with convolutional neural networks or the classification pipeline could benefit from complementing the feature extraction phase with convolutional neural networks or autoencoders, gaining the benefits of deep learning methods while preserving also the interpretable features.

ACKNOWLEDGMENT

The authors declare that there are no conflicts of interest.

LITERATURE CITED

Cytokeratin-supervised deep learning for automatic recognition of epithelial
cells in breast cancers stained for er, pr, and ki-67

Valkonen, M., Isola, J., Ylinen, O., Muhonen, V., Saxlin, A., Tolonen, T.,
Nykter, M., and Ruusuvuori, P.,

*IEEE transactions on medical imaging*

Publication reprinted with the permission of the copyright holders.
Cytokeratin-supervised deep learning for automatic recognition of epithelial cells in breast cancers stained for ER, PR, and Ki-67

Mira Valkonen, Jorma Isola, Onni Ylinen, Ville Muhonen, Anna Saxlin, Teemu Tolonen, Matti Nykter, and Pekka Ruusuvuori

Abstract—Immunohistochemistry (IHC) of ER, PR, and Ki-67 are routinely used assays in breast cancer diagnostics. Determination of the proportion of stained cells (labeling index) should be restricted on malignant epithelial cells, carefully avoiding tumor infiltrating stroma and inflammatory cells. Here, we developed a deep learning based digital mask for automated epithelial cell detection using fluoro-chromogenic cytokeratin-Ki-67 double staining and sequential hematoxylin-IHC staining as training material. A partially pre-trained deep convolutional neural network was fine-tuned using image batches from 152 patient samples of invasive breast tumors. Validity of the trained digital epithelial cell masks was studied with 366 images captured from 98 unseen samples, by comparing the epithelial cell masks to cytokeratin images and by visual evaluation of the brightfield images performed by two pathologists. A good discrimination of epithelial cells was achieved (AUC of mean ROC = 0.93; defined as the area under mean receiver operating characteristics), and well in concordance with pathologists’ visual assessment (4.01/5 as the area under mean ROC = 0.93). The effect of epithelial cell masking on the Ki-67 labeling index was substantial. 52 tumor images initially classified as low proliferation (Ki-67 < 14%) without epithelial cell masking were re-classified as high proliferation (Ki-67 ≥ 14%) after applying the deep learning based epithelial cell mask. The digital epithelial cell masks were found applicable also to IHC of ER and PR. We conclude that deep learning can be applied to detect carcinoma cells in breast cancer samples stained with conventional brightfield IHC.

Index Terms—Deep learning, image segmentation, breast cancer, histopathology, digital pathology.

I. INTRODUCTION

IMMUNOHISTOCHEMICAL stainings of estrogen receptor (ER), progesterone receptor (PR), and proliferation antigen Ki-67 are routinely used in the diagnostic assessment of breast cancer [1]. Although the analytical quality of ER and PR assays has been debated for decades, their role is well-established in current diagnostics. The tumor cell proliferation rate, as defined by Ki-67 immunostaining, is an additional tool for defining patient prognosis [2]. Patients with rapidly proliferating tumors are predicted to endure poorer outcomes than patients with tumors exhibiting low proliferation rate [3]. The Ki-67 labeling index has been suggested as a surrogate marker to discriminate of luminal type A and type B tumor subtypes. A cut-point of 14 % has been suggested as an optimal cut-point to define these low and high tumor proliferation types [4].

An increasing number of pathologists and laboratories analyze ER, PR, and Ki-67 by using digital image analysis, either with commercially available or public domain image analysis software (see, e.g. [5]–[7]). The software are often based on differentiation of DAB chromogen stained (brown) and counterstained (hematoxylin, blue) cell nuclei using color deconvolution algorithm [8]. The percentage of positively (brown) stained nuclei (or pixels) over total nuclei (brown nuclei + counterstain only nuclei) is reported as the labeling index [5].

When using these software pathologists are well aware that the labeling index should be analyzed only from the invasive malignant epithelial cells, excluding stromal fibroblasts and inflammatory cell infiltrates. Since non-malignant cells are usually non-proliferating, Ki-67 labeling index is usually underestimated and biased if non-malignant are not excluded from the analysis. Similarly, non-malignant cells are typically ER and PR negative causing bias to the results. When using digital image analysis (DIA) software, a pathologist usually define the tissue area to be analyzed by drawing graphically a region of interest (ROI). Even with careful delineation of invasive carcinoma, the defined ROI usually contains varying proportions of stromal and inflammatory cells. Attempts to exclude small-sized (non-malignant) and spindle-shaped (stromal) cell nuclei from the analysis by their shape and size, has not yielded fully satisfactory results. One recently reported solution is the Virtual Double Staining (VDS) method, which is based on staining of adjacent serial sections with an epithelial cell marker (cytokeratin) and applying the cytokeratin mask to the image captured from the Ki-67-stained adjacent slide [9], [10]. Although technically possible, this method requires availability of adjacent tissue sections on separate microscope slides.

We have modified immunohistochemical double staining to allow both cytokeratin and Ki-67 IHC done on one slide using a fluoro-chromogenic method [11]. In this method the two stains do not interfere with each other, because the epithelial marker is fluorescent and invisible to the brightfield image showing a typical ER, PR, or Ki-67 with DAB chromogen and
hematoxylin counterstain. Image of the cytokeratin staining is captured using fluorescence illumination and used to create an accurate epithelial cell mask. The described immunohistochemical stainings provide a convenient alternative for time-consuming expert driven manual annotation in order to collect large amount of groundtruth training data, which is the prerequisite for applying machine learning methods, such as deep learning. In recent years, deep learning methods, particularly convolutional networks, have rapidly become the main approach for analyzing medical images [12], [13]. Pre-trained convolutional networks combined with transfer learning and problem-specific training data enable the training of high performing models with lower amount of training data. These methods have achieved remarkable results in many medical image analysis tasks, such as, skin cancer classification [14], breast cancer detection and classification [15], [16], brain lesion segmentation [17], and other tasks [18]–[20]. There is, however, still need for deep learning in a cytokeratin-supervised manner to aid the image analysis of immunostained sections in order to achieve more accurate breast cancer prognostic parameters.

As an alternative approach for estimating cell proliferation in breast cancer, counting of mitotic cells has been commonly used technique. Also here, automated image analysis using machine learning, particularly with deep learning methods, holds great promise [21]. As a related approach, Aubreville et al. [22] presented two deep learning based methods to the selection of a field of interest for computer-aided mitotic count. Although related, these studies represent tools for an alternative path towards computer-aided breast cancer grading and are based on using different staining (hematoxylin & eosin).

The aim of this study was to generate improved DIA software for ER, PR, and Ki-67 to restrict the analysis specifically on epithelial tumor cells, but without the need of performing cytokeratin staining. We introduce a new innovation of using the described tumor mask stainings as training material for a deep learning algorithm, which was developed to generate a digital pan-cytokeratin (panCK) epithelial cell detection based on single color immunohistochemical stain with hematoxylin counterstain only. Thus, a novel approach of using label-based transfer learning of a neural network was developed for automated epithelial cell masking from histological images of the most common subtype of breast cancer, ductal carcinoma.

II. MATERIALS AND METHODS

A. Immunohistochemistry and whole slide scanning

Routinely formalin-fixed and paraffin embedded (FFPE) samples of 152 invasive breast cancers were stained with two alternative immunohistochemical staining techniques. In the fluoro-chromogenic double staining technique, de-paraffinization (3 x 2 min in n-hexane) and antigen retrieval (15 min in Tris- EDTA buffer, pH 9.0 at 98°C) were followed by sequential incubations with Ki-67 antibody (clone MIB-1, dilution 1:200, 30 min), anti-mouse peroxidase polymer (Histofine, Nichirei, ready-to-use, 30 min) and DAB. After washing, the slides were incubated with anti-pan-cytokeratin antibody cocktail (Jilab Inc, Tampere, Finland; diluted 20 ug/ml) and anti-mouse Cy2 fluorochrome conjugate (Jackson Immunotech, dilution 1:150). The slides were washed, counterstained with hematoxylin, and mounted with DPX. Whole slide scanning was performed with sequential brightfield and epifluorescence illuminations using SlideStrider scanner (Jilab, Tampere, Finland). Image layers were stored as multilayer JPEG2000 files [23] and viewed with SlideVantage software, which includes ImmunoRatio software for ER, PR, and Ki-67 image analysis.

In the sequential staining method, slides were stained first with hematoxylin and mounted with DPX. Slides were scanned with brightfield illumination. After whole slide scanning, coverslips were removed and slides were washed with TBS buffer. Thereafter a standard immunohistochemical staining was performed using the same pan-cytokeratin antibody as described above (diluted 0.5 ug/ml) using DAB as chromogen. Slides were scanned again, and the two scans were overlaid on each other in a multilayer JPEG2000 image file.

Fig. 1. Examples of the epithelial cell groundtruth masks generated with fluoro-chromogenic (I) and sequential chromogenic stains (II). Column A shows the brightfield image, column B the immunofluorescent (IB) and chromogenic (II) pan-cytokeratin stains, column C shows the binary epithelial cell mask, and column D shows the epithelial cell mask overlaid on images of column A.

The test material consisted of conventional single color immunohistochemically stained slides stained for ER, PR, and Ki-67 in the routine diagnostic setting. Hematoxylin was used as counterstain.

B. Epithelial cell mask detection using deep learning

In this study, we developed cytokeratin-supervised deep learning based algorithm for automated epithelial cell detection. The training materials included histological images stained with fluoro-chromogenic cytokeratin-Ki-67 double staining and sequential hematoxylin-IHC staining. In the fluoro-chromogenic double staining technique, the groundtruth
binary epithelial cell masks were generated by blockwise processing of the fluorescent pan-cytokeratin images, first smoothing the intensities using Gaussian filter and then applying Otsu’s thresholding method. In the sequential staining method, the sequential images were first registered using StackReg plugin in ImageJ [25]. Color deconvolution algorithm was applied to the sequentially DAB stained cytokeratin image in order to separate the hematoxylin and DAB stain channels. Similarly as before, the separated DAB channel was processed blockwise, filtered with Gaussian filter, and thresholded using Otsu’s method to generate the groundtruth binary epithelial cell masks for training.

A convolutional neural network model was built for predicting epithelial cell mask from immunostained images (Figure 2). A convolutional neural network (CNN) consists of a sequence of layers that maps an input vector \( x \) to an output vector \( y \).

\[
y = f(x, w),
\]

where \( w \) is the weight and bias vector that define the network layers. During the training phase, these network variables are estimated by solving an optimization problem. Given a training set of input vectors \( x_n \) with corresponding target vectors \( t_n \), where \( n = 1, ..., N \), the optimization problem can be defined as

\[
\arg\min_w \frac{1}{N} \sum_{i=1}^{N} L(f(x_i, w), t_i)
\]

where, \( L \) is a task-fitting loss function. Here, we used binary cross-entropy as a loss function.

\[
L(y, t) = -\frac{1}{N} \sum_{i=1}^{N} (t_i \log(y_i) + (1 - t_i) \log(1 - y_i))
\]

The optimization was performed by using an Adam optimizer [26] which is a stochastic gradient descent method. Instead of training a network from scratch, transfer learning approach was used. It utilises a pre-trained network that is already optimized to recognize high-level features from images of some other domain. Here, using Python programming language and Keras [27] module with Tensorflow [28] backend, we built a fully convolutional neural network that consists of a VGG-16 architecture pre-trained on the ImageNet dataset [24], two additional convolutional layers, and an output layer. The last two convolutional layers both include 100 convolutional filters with 3x3 receptive field followed by leaky rectified linear unit, max pooling with filter size of 2x2 and dropout. Sigmoid activation function was used at the final output layer. As an output the network generates a probability map for the epithelial cell mask.

C. Datasets

Training data included image batches from 152 patient samples stained with fluoro-chromogenic cytokeratin-Ki-67 double staining or sequential hematoxylin-IHC. Quality control of the training images was done manually in order to get accurate and representative training data. In total, the training set included 13344 (256 x 256 pixels) images. Excluded images consisted of significantly out-of-focus images, or images with other prominent artefacts such as failed staining. The training was performed on a workstation with Nvidia Titan V GPU (128 GB RAM) and Intel i7-6900K 8-core 3.2 GHz CPU. Training was done in two phases, first training for 5 epochs with a batch size of 20, and the second phase for 5 epochs with a batch size of 50. With these specifications, training of the model
Fig. 3. Examples of the deep learning based epithelial cell detection in an invasive carcinoma (A,D), ductal carcinoma in situ (B,E) and non-malignant ductal breast epithelium (C,F) in Ki-67 IHC staining. Tissues recognized as epithelial by the deep learning algorithm are highlighted in D-F. Because the pan-keratin staining was used as training material for the deep learning algorithm, epithelial cells in DCIS and normal-breast are included. Pathologists should exclude them in the areas to be measured for Ki-67%.

took approximately one hour.

Using the trained model, epithelial cell masks were predicted for 366 image fields of view (1276 x 512 pixels) collected from 98 unseen whole slide images (WSI), that constitute the test set used in this study. Our test set included only samples from the most common histologic subtype of breast cancer, ductal carcinoma. Pixel size of the images was 0.275 µm at full resolution. For analysing the images, a downsample factor of 4 was used. The datasets for training and testing are available at https://github.com/BioimageInformaticsTampere/DigitalPanCK.

D. Performance evaluation

The validity of the predicted epithelial cell masks was tested by visual evaluation performed by two pathologists and by quantitatively comparing to groundtruth masks generated from the cytokeratin images. The visual evaluation was performed by showing each original brightfield image together with the masked image to the pathologists. Pathologists gave scores ranging from 0 to 5, zero being entirely inaccurate and five presenting a perfect epithelial cell masks. See Supplementary Table 1 for description of the criteria used for pathologist evaluation.

The quantitative evaluation to the groundtruth cytokeratin masks was performed by considering the mask prediction as a pixelwise classification problem and by calculating binary classification accuracy metrics for the classification results. Each pixel in the predicted mask was compared to the corresponding pixel in the groundtruth mask to verify whether the pixel is correctly classified as a pixel from epithelial tissue or not. Classification accuracy, sensitivity, and specificity were calculated as statistical performance measures of a binary classifier. Moreover, AUC measure was computed, which is defined as the area under receiver operating characteristics (ROC) curve. AUC measure gives a numerical value for the performance of a binary classifier. Here, we computed AUC from mean ROC.

In addition, image level quantitative comparison to the groundtruth cytokeratin masks was performed between each predicted binary mask and the corresponding groundtruth mask, thus, intersection over union (IoU) and structural similarity (SSIM) index [30] were computed. IoU, also known as the Jaccard index, is defined as the intersection of the two masks divided by the size of their union. The SSIM index separates the influence of illumination and image contrast and measures local image structure by sample cross-correlation. Therefore, it is well suited for comparing structures of binary images.

E. Quantification of epithelial cells stained for ER, PR and Ki-67

Test set was analyzed for the epithelial cell content with the developed deep learning algorithm, and analyzed for the immunostaining positive (brown) and - negative (blue) cell content using a modified ImmunoRatio software. The ImmunoRatio software outputs the immunostaining positive and negative cell counts. The results were expressed as Ki-67 (or ER/PR) ratio (labeling index), which is defined as follows:

$$r = \frac{\text{positive cells}}{\text{positive + negative cells}} \times 100,$$

In addition, the following error measures were computed:

$$e_{Ki67_{gtm, unm}} = r_{gtm} - r_{unm}.$$

$$e_{Ki67_{gtm, dlm}} = r_{gtm} - r_{dlm},$$

where \(r_{gtm}\) is the Ki-67 ratio defined by ImmunoRatio from the groundtruth masked images. Correspondingly, \(r_{dlm}\) and \(r_{unm}\) are the Ki-67 ratios computed from the deep learning (DL) based digital panCK masked images and unmasked images.

The accuracy of Immunoratio (IR) was recently evaluated for in vitro diagnostic use at Fimlab Laboratories. One pathologist (A.S.) manually counted Ki-67 proliferative indexes from 100 breast cancer images, which was considered ground truth. The images were independently analyzed by four experienced pathologists by semiquantitative estimation (eyeballing)
method, and by IR. Both methods correlated well to the ground truth, but IR was more accurate (r=0.99 for IR and r=0.93-0.95 for pathologists).

F. Implementation of the epithelial cell deep learning algorithm to ImmunoRatio WSI software

The epithelial cell deep learning algorithm was implemented as part of the ImmunoRatio image analysis software (ver. 3.0), which can be tested freely at http://wiser.jlab.fi/demo/wsi-deep-learning-immunoratio-2018 as part of a whole slide image viewing system.

Fig. 4. The visual validity assessment of the generated masks performed by two pathologists. Pathologists gave mask quality scores ranging from 0 to 5. The results for Ki-67 test set are presented in panel A and the results for additional ER and PR stained test set are presented the panel B.

### III. RESULTS

We developed a deep learning based digital panCK mask for automated epithelial cell detection using fluoro-chromogenic staining (Figure 1) and sequential chromogenic staining (Figure 1A) as training material. Since the pan-cytokeratin staining was used as training material for the deep learning algorithm, epithelial cells in ductal carcinoma in situ (DCIS) and normal-breast are included. Therefore, these areas were excluded from the test images that were intended to be used for Ki-67% analysis. Each test image was driven through the trained convolutional neural network and a digital panCK confidence map was predicted, as shown in Figure 2. The confidence map presents a probability of a certain tissue region to contain epithelial tissue predicted by the trained network. A probability threshold of 0.5 was used to convert the predicted epithelial cell confidence maps into binary digital panCK masks. Examples of the predicted digital panCK masks are presented in Figure 3. The examples include an invasive carcinoma (Figure 3A), ductal carcinoma in situ (Figure 3B), and non-malignant ductal breast epithelium (Figure 3C).

A. Validation of the predicted digital epithelial cell masks

First, the validity of the predicted epithelial cell masks was assessed visually by two pathologists. Pathologists gave mask quality scores ranging from 0 to 5, zero being entirely inaccurate and five presenting a perfect epithelial cell mask. Mean score set by the first pathologist was 4.01 (σ = 0.78) and 4.67 (σ = 1.11) set by the second pathologist. These visual validation results are shown in the panel A of Figure 4. Examples of the pathologists’ scores ranging from 0 to 5 are shown in Supplementary Figure 1 and the scoring criteria in Supplementary Table 1.

Next, the validity of the predicted epithelial cell masks was assessed quantitatively. The predicted masks were compared to the cytokeratin based groundtruth masks by assessing accuracy (mean acc. = 0.88), sensitivity (mean sensitivity = 0.82), specificity (mean specificity = 0.88), mean ROC (AUC = 0.93), IoU (mean IoU = 0.69) and SSIM (mean SSIM = 0.76) metrics.

The quantitative validation metrics obtained for the test set are presented as mean ROC curve in the Figure 5B and as boxplot in the Figure 5C. Overall, a good discrimination of epithelial cells was achieved. Figure 5A visualises two example images with high validation scores, where the predicted mask and the groundtruth mask are presented one over the other and color-coded to emphasize the overlapping and non-overlapping areas.

In order to determine the generalizability of our approach, epithelial cell masks were predicted for an additional small set of ER and PR stained images using the model that was trained on the Ki-67 stained images. Validity of these predicted epithelial cell masks was assessed visually by two pathologists and the results are shown in the panel B of Figure 4.

B. Ki-67 analysis using deep learning epithelial cell masks

The main aim of this study was to improve the cancer cell specificity of the immunohistochemical Ki-67 image analysis. The Ki-67 labeling indexes were computed for unmasked and masked test images using ImmunoRatio software. The groundtruth epithelial cell masks and the predicted digital epithelial cell masks were used. The Ki-67 analysis results are presented in the Figure 6. The Ki-67 ratios computed using the groundtruth masks were compared with Ki-67 ratios computed from unmasked images and with epithelial cell masked images. The Ki-67 ratio errors computed as shown in equations (5) and (6) are presented as boxplots in panel 6A and as distributions in panel 6B over the whole test set. The mean Ki-67 ratio error between groundtruth masked images and unmasked images was 3.64 with 5.71 standard deviation, and the mean Ki-67 ratio error between groundtruth masked images and digital panCK masked images was 0.15 with 2.33 standard deviation. In addition, we found that there was a statistically significant difference between the unmasked Ki-67 ratios when compared with masked Ki-67 ratios (reject the null hypothesis of equal averages with p-values; \( p_{unm, unm} = 5.62 \times 10^{-8} \) and \( p_{unm, gtm} = 1.39 \times 10^{-7} \), and yet no statistically significant difference between the DL masked and groundtruth masked Ki-67 ratios \( (p_{gtm, unm} = 0.85) \). Based on these results it can be concluded that the Ki-67 ratio difference between groundtruth masked images and digital panCK masked images is minor, compared to the Ki-67 ratio difference between groundtruth masked images and unmasked images. Therefore, the need of an epithelial
Fig. 5. Visual and quantitative validation results of the predicted epithelial cell masks. The panel A visualises two example images with high validation scores, where the predicted mask and the groundtruth mask are presented one over the other and color-coded to emphasize the overlapping and non-overlapping areas. The quantitative validation metrics are presented as mean ROC curve in the panel B and as boxplot in the panel C. In the panel C, the evaluation metrics range from 0 to 1, where higher value corresponds to higher similarity between the predicted mask and the groundtruth mask.

cell mask is considerable, and in addition, the deep learning based epithelial cell masking is comparable method with the cytokeratin staining based masking.

Because the Ki-67 labeling index result is often interpreted with binary classification using 14 % as cut-point, we analyzed how large fraction of the tumor images were re-classified when epithelial cell masks were used. Panel C in the Figure 6 visualises the fraction of high tumor proliferation cases in the test set when the epithelial cell masks are used or not. The test set included 14.8 % high tumor proliferation cases when no epithelial cell mask is used and the amount increased to 29.0 % when the deep learning epithelial cell mask was applied. The increase in high tumor proliferation cases can be also seen as a small shift in the histogram of unmasked ratios to DL masked ratios presented in panel E of the Figure 6. The total amount of high tumor proliferation cases in the test set was 105 when the groundtruth epithelial mask was used and 106 with the deep learning based digital panCK mask. The differently classified cases between groundtruth and digital panCK are presented in the Supplementary Figure 2. An example of the graphical output of the implemented system within WSI software is shown in Figure 7.

IV. DISCUSSION

Cell proliferation rate, defined by Ki-67 immunostaining, is considered as an important prognostic factor in breast cancer diagnostics. Although Ki-67 analysis has clinical validity, in practice, assessing the Ki-67 labeling index is prone to intra- and inter-observer variability. Non-malignant stromal or inflammatory cells can cause a significant bias, if they are not excluded from the analysis. Digital image analysis methods provide a quantitative and more reproducible computational alternative to manual assessment of Ki-67 labeling index. Computational methods are less time-consuming and more accurate, and therefore, enable standardized diagnostics. Standardized determination of the Ki-67 labeling index should be restricted on malignant epithelial cells, carefully avoiding tumor infiltrating stroma and inflammatory cells. Therefore, a method for masking malignant epithelial cells is needed.

One solution is the Virtual Double Staining (VDS) method, which is based on staining of adjacent serial sections with a tumor marker and using the marker stain as a mask to assess the Ki-67 labeling index [9], [10]. However, this method requires availability of adjacent tissue sections on separate microscope slides and is costly. We have developed a fluoro-chromogenic double staining method for the same purpose. It requires only one tissue section, but the microscope or whole slide scanner must be equipped with epifluorescence illumination. The aim of the present study was to generate improved DIA software for ER, PR, and Ki-67 analysis based on epithelial tumor cells only, without the need of performing additional cytokeratin staining. In order to avoid the fluorescence staining step, we presented here a cytokeratin-supervised deep learning based digital epithelial mask for automated epithelial cell detection. The generated epithelial cell masking can be applied to slides stained with conventional single-color brightfield IHC. The deep learning model, which was implemented as part of ImmunoRatio image analysis software, was found to be an accurate replacement for fluorescence cytokeratin staining, and further, to enable enhanced Ki-67 analysis.

In this study, quantitative evaluation of the predicted epithelial cell masks was considered as a pixelwise classification
Fig. 6. The Ki-67 analysis results computed using ImmunoRatio software. The Ki-67 labeling index errors of unmasked images and deep learning (DL) masked images compared to the groundtruth (GT) masked images are presented as boxplots in panel A and as distributions in panel B. The panel C pie charts represent the amount of high proliferation (Ki-67 >14%) cases without epithelial cell masking compared to the amount after applying the deep learning based epithelial cell mask. The same information as in panel C is also visualized in the panel E as histograms. The increased amount of high proliferation cases can be seen as a small shift of the Ki-67 ratio histogram of unmasked compared to DL masked. Mask examples (I,II,III,IV) of similar and significantly different Ki-67 ratios of unmasked images and deep learning masked images compared to the groundtruth masked images are presented in the panel D.

evaluation problem. Each pixel in the predicted mask was compared to the corresponding pixel in the groundtruth mask and binary classification metrics, such as AUC, sensitivity, and specificity, were calculated. Class imbalance in the test data, however, can cause bias in these results. Therefore, we also included metrics that consider these generated masks as a whole, such as IoU, and SSIM. Nevertheless, these metrics can be easily influenced by minor structural differences even though the main structure would be similar. Such strict comparison with the groundtruth would require flawless groundtruth masks. Yet, these groundtruth masks can be partly inaccurate, since they are indirect fluorescence measurements of the underlying biological sample, and affected by variation in imaging and staining, which affect the binary mask generation. In order to take these factors into account, it was necessary to include a pathologists’ visual evaluation as well, which can be considered as the clinically most relevant groundtruth.

For the most part, the generated digital panCK masks yield equal accuracy as the ground truth masks in detecting carcinoma cells and epithelial structures as shown in Figure 6. However, there were cases where the Ki-67 ratio error is high between a digital panCK mask and the corresponding ground truth mask (e.g. Figure 6 D III and IV). One reason for this is that in some cases the model still falsely detects some inflammatory cells which effects the calculated Ki-67 ratio. Another possible reason for the high Ki-67 ratio error cases is that DAB chromogen stain suppresses the fluorescence measurements and therefore the ground truth mask struggles to segment deeply brown areas within the tissue. Consequently, lower DAB positive cell count has an effect on the computed Ki-67 ratio.

The calculated Ki-67 scores support the hypothesis that the assessment of Ki-67 labeling index needs to be standardized and limited to malignant epithelial cells. The results presented in the Figure 6 show that the computed Ki-67 scores differ considerably between unmasked images and masked images. The difference in Ki-67 labeling indexes between unmasked and deep learning-based epithelial cell masked images ranged from minimal to highly significant, the latter in tumors displaying with abundant intratumoral non-epithelial cell infiltration. Additionally, the Ki-67 labeling index has been suggested as a surrogate marker to discriminate of luminal type A and type B tumor subtypes. A cut-point of 14 % has been suggested as an optimal cut-point to define these low and high tumor proliferation types. Considering this cut-point, significant amount of high tumor proliferation types could be incorrectly defined as low tumor proliferation type without
using the epithelial cell mask. When using the deep learning epithelial cell detection, the users must bear in mind that it detects also non-malignant ductal epithelium as well as ductal carcinoma in situ. Since Ki-67 should be analyzed from invasive carcinoma, the users should select the areas to be analyzed free of unwanted epithelial cell structures.

Immunohistochemical staining of ER and PR are routinely used in breast cancer diagnostics. In order to determine the generalizability of our approach, epithelial cell masks were predicted for an additional small set of ER and PR stained image samples using the model that was trained on the Ki-67 stained images. Based on visual assessment, the deep learning based epithelial cell detection performed equally well on ER and PR stained single color brightfield IHC. Since the amount of DAB chromogen stain (brown) varies quite significantly within these different stains; ER, PR and Ki-67, this additional test might indicate that the deep learning model has learned other characteristics of the tissue, such as, morphological properties, and that the epithelial cell masking is not purely based on the color intensity properties of the stained nuclei. This approach calls for more research, and specifically, tools to link model decision to biological information are needed. Our test set included only samples from the most common histologic subtype of breast cancer, ductal carcinoma. Therefore, additional validation and more research need to be conducted in order to determine the generalizability of our digital panCK mask to the second most common histologic subtype, invasive lobular carcinoma.

In addition, epithelial cell masks were predicted for an additional test set of 41 tissue microarray samples (Supplementary Figure 3). Based on visual assessment, the epithelial cell detection performed equally well also on this additional set. The epithelial cell detection algorithm was implemented as part of whole slide image viewing system, and for further validation and experimentation, we encourage to test the method online at [http://wriserver.jilab.fi/demo/wsi-deep-learning-immunoratio-2018](http://wriserver.jilab.fi/demo/wsi-deep-learning-immunoratio-2018). The implementation is modular, enabling efficient re-training and updating of the cell detection classifier with, e.g., challenging cases.

Our study presents a proof-of-concept implementation on masking epithelial cells from immunostained tissue using deep learning. While the results are very promising, it is likely that data originating from different laboratories will include staining variability and potentially other sources of technical variation, such as differences caused by scanning devices, as well as potential differences in sample characteristics. All these differences may also lead to variation in the obtained results. Validation studies with large populations, eventually in clinical practice, will be needed to determine the need for additional training data to generalize the method to all sources of variation.

In conclusion, our findings indicate that the deep learning based epithelial cell detection is an accurate and effective alternative for fluorescence based panCK masking, and an important addition to the Ki-67 analysis. The method can be used for reliable detection of epithelial cells from slides stained with conventional single-color brightfield IHC. A deep learning model, which has been implemented as part of the image analysis software, was found to provide significant improvement in Ki-67 image analysis when compared to the use of adjacent step sections, which is the prevailing method in current practise.

REFERENCES


Generalized fixation invariant nuclei detection through domain adaptation based deep learning

Valkonen, M., Högäs, G., Bova, G. S., and Ruusuvuori, P.,

*IEEE Journal of Biomedical and Health Informatics, 25*(5), 1747–1757

Publication reprinted with the permission of the copyright holders.
Abstract—Nucleus detection is a fundamental task in histological image analysis and an important tool for many follow up analyses. It is known that sample preparation and scanning procedure of histological slides introduce a great amount of variability to the histological images and poses challenges for automated nucleus detection. Here, we studied the effect of histopathological sample fixation on the accuracy of a deep learning based nucleus detection model trained with hematoxylin and eosin stained images. We experimented with training data that includes three methods of fixation; PAXgene, formalin and frozen, and studied the detection accuracy results of various convolutional neural networks. Our results indicate that the variability introduced during sample preparation affects the generalization of a model and should be considered when building accurate and robust nucleus detection algorithms. Our dataset includes over 67,000 annotated nuclei locations from 16 patients and three different sample fixation types. The dataset provides excellent basis for building an accurate and robust nucleus detection model, and combined with unsupervised domain adaptation, the workflow allows generalization to images from unseen domains, including different tissues and images from different labs.

Index Terms—Deep learning, digital pathology, domain adaptation, formalin-fixed, frozen section, nuclei detection, PAXgene-fixed, tissue fixation.

I. INTRODUCTION

HISTOPATHOLOGICAL examination is an important step in diagnosis of many diseases. Examination usually includes analysis of nuclei morphology, thus, nucleus detection is a fundamental step for many follow up analyses, such as phenotyping on a single-cell level [1], or cancer grading [2]. Machine learning based image analysis provides an efficient, quantitative, and objective way to perform histopathological examination and nuclei detection in a fully automated manner [3]. Nevertheless, building a robust and generalizable nuclei detection model is a challenging task due to the high amount of variability present in histological images. This variability is caused by the underlying biological variation, such as variation in nuclei shape, size, and texture of different tissue types and also by the technical variation introduced during the tissue preparation, such as in fixation process, and scanning procedure [4].

Preparation of tissue depends largely on the type of analysis the tissue is intended for, and for tissue fixation there exists alternatives which have effect on the appearance of tissue components, such as nuclei [5]. While freezing tissue typically provides excellent preservation of biomolecules, freezing also disrupts the structure of the tissue and is therefore not used for routine morphologic analysis. Tissue fixation with formalin is the standard in surgical pathology laboratories due to its low cost and excellent preservation of tissue morphology. Formalin preserves tissue structure by forming crosslinks between molecules [6] and is routinely used for almost all tissue types. PAXgene is an alcohol-based fixative that, in contrast to formalin, simultaneously preserves both tissue morphology and biomolecule integrity. PAXgene has been shown to be suitable for many tissues [7], although artefacts such as increased nuclear staining and tissue shrinkage have been reported. PAXgene-fixed tissue also stains more avidly with eosin, giving H&E (hematoxylin & eosin) stained sections a more intense pink hue compared to formalin fixed specimens. In our earlier study, we studied the feasibility of PAXgene fixation for molecular and diagnostic studies [5], but the fixation effect on modern deep learning based analytics remains unknown.

Deep learning methods, particularly convolutional networks combined with transfer learning [8], [9], have an outstanding ability to learn task specific feature representations and have rapidly become the main approach for microscopy image analysis tasks, including nuclei detection [10], [11]. However, most
of the existing methods have been optimised for a specific problem domain using a narrow dataset and fail to generalize to new domains, such as images from different labs or different tissues due to the high variability present in the histological images. A lot of work has been done in order to address the generalization challenge using techniques such as staining normalization [12]–[14], extensive data augmentation [15], [16], or utilization of datasets containing high variability such as multi-tissue datasets [17]. Although these approaches have achieved prominent results, they have still left room for further research, and some variability sources are yet to be studied, such as different tissue preparation techniques. To eventually generalize extensively to diverse patient populations and real world clinical environments, the effects of wider range of variability sources need to be covered.

In digital pathology, labeled training data is not largely available due to the time-consuming manual annotation process performed by an expert. Nuclei detection as an annotation task is particularly laborious due to the large amount of target objects. Consequently, in digital pathology, collecting a vast amount of labeled training data adds to the challenges posed by the heterogeneous nature of histological image data. However, in order to train a deep convolutional neural network in a supervised manner, labeled training data is an absolute requirement. Therefore, in order to build an accurate and robust nuclei detection algorithm that can generalize from one problem domain to another, alternative data labeling methods are needed.

Domain adaptation provides tools for overcoming the requirements of labeled data. In domain adaptation, the representations learned from labeled source data are utilized in a classification problem in an unlabeled target domain [18]. As the goal in domain adaptation is to enable domain shift from the problem domain of the original training data into a new domain from which no labeled data for re-training exists, it is a potential solution for generalizing histopathological image analysis methods from one tissue domain to another. Domain adaptation can be utilized in histopathological classification problems to address the requirement of labeled data in unsupervised [19], [20] or weakly supervised [21] manner. We have also shown in our previous studies how domain adaptation can be successfully utilised in model generalization to unseen cell lines from brightfield images in an unsupervised manner [22].

Deep learning methods have shown great success in many machine vision tasks, however their lack of transparency has attracted an increasing amount of attention and research [23]–[25]. Especially in medical domain applications, interpretability and transparency are seen as a necessity, since these can provide insights into the functioning of a deep learning model and can be used to verify and comprehend the predictions by a human expert. One aspect particularly of interest in classifier interpretability is the contribution of patterns in specific spatial locations in input data to classifier decision or outcome. For deep neural networks, methods such as Layer-wise Relevance Propagation (LRP) [26] have recently enabled such analysis, and their availability as tools for explainable AI [27] help giving insight in the classifier decision process. Interpretability tools can also enable verifying that a model learns relevant and similar information from training data even with subtle differences present in the data, caused for example by differences in staining or fixation.

In this study, we trained a convolutional neural network baseline model for nuclei detection using supervised transfer learning. As the second step of the workflow, we applied unsupervised domain adaptation to allow generalization to images from unseen domains, including different tissues and images from different laboratories, without the need for labeled data. Our main contributions are, 1) to study the effect of sample fixation on the accuracy of nuclei detection by using hematoxylin and eosin (H&E) stained training images prepared with three fixation methods; PAXgene, formalin and frozen, and 2) to provide the presented extensive multi-fixation dataset with manually obtained annotations for cell locations that allows further method development. The code and data are available at https://github.com/BioimageInformaticsTampere/NucleiDetection. The implemented workflow allows generalisation and adaptation to external unlabeled datasets in the field of digital pathology utilising unsupervised domain adaptation based on pseudo-labels and hard positive mining.

II. MATERIALS AND METHODS

A. Data

The materials used in this study included prostate tissue (PT) dataset [5], a holdout test set (5-tissue), and a publicly available Multi-Organ Nuclei Segmentation (MoNuSeg) [17] dataset. Examples from the PT dataset and 5-tissue test set are shown in the Fig. 1. In addition, number of annotated nuclei, number of different tissue types and number of patients in each
dataset are shown in the Table I. The following sections will describe the used materials in more detail.

**Prostate Tissue (PT) Dataset:** The image data was collected from radical prostatectomy prostate tissue samples from 16 men. The samples were collected and studied under Tampere University Hospital Ethical Committee Approval R03203. From each patient, three cores were collected from posterior side of the prostate. Each of the three cores was fixed using one of three different tissue preparation methods: fresh frozen, formalin-fixed paraffin-embedded, and PAXgene-fixed paraffin-embedded. The tissue sections were stained with H&E and scanned with a Hamamatsu Photonics Nano Zoomer XR C12000 automated scanner, using pixel resolution 0.23 μm. More detailed data acquisition is presented in a study by Högnäs et al. [5]. From every H&E stained core whole-slide image (WSI) one 500×500 μm image was randomly selected and nuclei were manually annotated, resulting in a total of 67 070 nuclei in 48 images. Compared to the original study where 50 images were used, two images were excluded in order to balance the number of nuclei in each dataset with respect to number of patients and fixations.

Manual annotation was carried out by an experienced histologist using the Cell Counter plugin in ImageJ software [28] by visually inspecting the images and by manually clicking coordinates of every nucleus in the image. The coordinates for each nucleus in were saved in xml files, one file per image. Further, we validated the accuracy of the manual annotation for a randomly chosen image by reproducing coordinate markings by two independent persons, yielding very high agreement both in terms of number of nuclei and numerical accuracy (F-score 0.9 for both annotators with the ground truth). The detailed validation results as well as visual representation of the validation image are provided as Supplemental Fig. 1 and Supplemental Table I. This dataset, along with the annotations, is available at: https://github.com/BioimageInformaticsTampere/NucleiDetection/

**5-Tissue Dataset:** The 5-tissue dataset included formalin-fixed paraffin-embedded H&E stained metastatic tissue images from 5 different tissue types. The tissue types included periurethral tissue, bone tissue from rib, axillary lymphatic tissue, adrenal gland tissue, and pelvic lymphatic tissue. The samples were collected from 5 patients and from each WSI a 500×500 μm image was randomly selected. All nuclei were annotated from the images using Multi-point Tool in ImageJ software [28] in a similar, fully manual fashion as described for the prostate dataset. In total, the test set included 5 images and 9011 annotated nuclei coordinates.

Multi-Organ Nuclei Dataset: Publicly available multi-organ dataset (MoNuSeg) [17] included 30 images captured from 30 different H&E stained WSIs. These images were collected from The Cancer Genome Atlas and originally prepared in 18 different hospitals. The manually annotated nuclei represent a diversity of nuclear appearances from several patients, disease states, and organs. The dataset consisted of seven different tissue types, including breast, liver, kidney, prostate, bladder, colon, and stomach tissue. In total, the dataset included 16 966 nuclei mask annotations. As the images are publicly available from https://monuseg.grand-challenge.org, this dataset enables benchmarking and provides further insight about generalization of the methods.

**B. Nuclei Detection Model**

A convolutional neural network model was built to detect cell nuclei locations from histopathological images. To achieve better detection accuracy and to reduce the computational costs of optimisation, transfer learning approach was used. Transfer learning allows the utilisation of a pre-trained network that is already optimised to classify images from some other domain [8]. The lower level features of a pre-trained network tend to be more generic and the later layer features become more specific to the details of the original classification task in the training data domain. Therefore, we utilised four base layers from VGG-16 architecture pre-trained on the ImageNet dataset [29]. These base layer weights were fixed during training, and four additional convolutional layers were added on top of them. Each convolutional layer was followed by Rectified Linear Unit (ReLU) activation and every two convolutional layers were followed by dropout in order to avoid overfitting of the model [30]. Sigmoid activation function was used at the model output layer to provide a nuclei location confidence map. The nuclei detection model was implemented using Python programming language and Keras [31] module with Tensorflow [32] backend. The model workflow and architecture are visualized in the Fig. 2.

Our choice of the base deep neural network architecture was motivated by the relatively simple adaptability of the VGG-16 architecture, which can be done simply by adding problem domain specific layers into a generic VGG-16 network. We have successfully used similar transfer learning strategy of extending the generic VGG-16 network into a specific problem domain in histopathology earlier in a study where image-to-image transform from immunohistochemical staining to cytokinin staining was done using VGG-16 based architecture [33]. In the current study, the use of a generic, well tested architecture underlines the applicability of the proposed domain adaptation approach. The use of other, more developed implementations or different architectures, is possible in a similar manner, but optimizing their accuracy for the nuclei detection task is out of the scope of this study.

**Model Training Specifications:** A convolutional neural network (CNN) consists of a sequence of layers that maps an input vector \( x \) to an output vector \( y \).

\[
y = f(x, w),
\]

where \( w \) is the weight and bias vector that define the network layers. During the training phase, the network variables are
Fig. 2. Nuclei detection workflow. Upper half presents the baseline model training step which is followed by an option to utilise unsupervised domain adaptation to detect nuclei from an external dataset (new domain) without annotations. The convolutional neural network consists of four convolutional base layers from a pre-trained VGG-16 network appended with four additional convolutional layers. 

The optimization was performed by using an Adam optimizer [34] which is a stochastic gradient descent method. For Adam algorithm, learning rate was initialized to 0.0001, the exponential decay rates for the moment estimates ($\beta_1, \beta_2$) were set to 0.9 and 0.999, respectively, and the fuzz factor was set to $1 \times 10^{-8}$ to prevent null division.

A set of model hyperparameters was optimised by using grid-search to maximize the nuclei detection accuracy in the training phase using pixel size of 0.5 μm. Test data was never used in hyperparameter optimization. These hyperparameters included number of epochs ($n_{epochs} = 2$), learning rate ($\text{lr} = 0.0001$), batch size ($bs = 16$) and dropout for regularization ($\text{drop}_-\text{rate} = 0.5$). Also optimal input image size (64x64 pixels) and nuclei location mask structuring element shape and size

estimated by solving an optimization problem. In supervised learning, where labeled training data is a prerequisite, a set of input vectors $x_n$ have corresponding target vectors $t_n$ ($n = 1, \ldots, N$). In which case, the optimization problem can be defined as

$$\arg\min_w \frac{1}{N} \sum_{i=1}^{N} L(f(x_i, w), t_i)$$  \quad (2)$$

where, $L$ is a task-fitting loss function. Here, we used binary crossentropy as a loss function.

$$L(y, t) = -\frac{1}{N} \sum_{i=1}^{N} (t_i \log(y_i) + (1 - t_i) \log(1 - y_i))$$  \quad (3)$$
(round, \( R = 4 \)) were searched to maximize the nuclei detection accuracy on the PT dataset.

In order to reduce variation caused by the different staining procedures and to focus on the morphological differences in the tissue caused by the different fixations, a data augmentation step was included. The step included color augmentation on HSV space and adding Gaussian noise \((\mu = 0, \sigma = 0.01)\). Color augmentation was implemented by first converting the RGB image into HSV space and then adding a constant value to the hue channel. The constant value was randomly drawn from normal distribution with mean of \( \mu = 0.1 \) and standard deviation of \( \sigma = 0.01 \). After the hue shift, the sample image was converted back to RGB space. Every third input image block was left in the original form and the other samples were augmented using either HSV shift or by adding noise.

**Prediction:** The trained nuclei detection model can be used to predict the nuclei locations of an input image with pixel size of 0.5 \( \mu \)m. The model takes an arbitrary sized RGB image as an input and predicts a confidence map as an output, where values close to 1 denote a high probability for a nucleus location and values close to 0 indicate a background pixel in the corresponding location of the input image.

The confidence map is post-processed in order to find single pixel locations for each detected nucleus. The confidence map is first converted into a binary image using threshold value of 0.5. Different objects in the binary image are labeled and the connectivity of the objects is defined by a centrosymmetric 3x3 structuring element. Finally, the center of mass from each object is selected as the coordinate for a detected nucleus.

**Ground Truth Generation From Annotations:** The ground truth nuclei location masks were generated from nucleus coordinates annotated manually by an expert. The coordinates were read from a csv file and first scaled to match the operating resolution of the model. A binary mask image was generated with a single pixel representing the coordinates of one annotated nucleus. This mask image was then dilated using a circular structuring element with radius \( R = 4 \) in order to expand each nucleus location area. This process corresponds to adding computationally a small level of uncertainty in the annotation coordinates, as the manual marking is practically impossible to be done on a pixel level accuracy for thousands of objects. The value of \( R \) was defined experimentally to be small enough such that the whole ground truth marker would remain inside the nuclei and that no overlap between ground truth objects occur. Examples of ground truth nuclei location masks can be seen from the workflow Fig. 2.

**Unsupervised Domain Adaptation Using Pseudo-Labels:** The ground truth nuclei location masks for an external dataset without any annotations can be generated by using the trained baseline model. Any dataset can be run through the trained baseline model and from the predicted nuclei location confidence maps a set of positive examples can be extracted based on a thresholding rule. These hard positive examples can be then used as new training examples to adapt the nuclei detection model to a new data domain.

The thresholding rule includes two different thresholds; higher for determining hard positives, which are confident detections, and lower for detecting other nuclei around the hard positives. To generate new training samples, an image block of the size 64 x 64 pixels is extracted around each hard positive example. In order to generate the ground truth nuclei location mask for the whole training sample block and not to miss any nuclei within the image, the second threshold (the lower one) is applied. The detection controlled by the two thresholds is an elemental part of the pseudo-label domain adaptation step. Thus, we examined the effect of the threshold values on detection accuracy in a grid search (Supplementary Table II), and chose the threshold values (0.8 higher, 0.5 lower) based on this experiment. From the generated binary image, the nuclei locations are generated similarly as in the prediction step and the final mask image is generated similarly as in the ground truth generation from the annotation step as explained in previous paragraphs.

### C. Evaluation

**Accuracy Metrics:** For numerical evaluation of the nuclei detection model accuracy, F1-score was used that rely on precision and recall. Precision measures the fraction of correctly classified positive instances among all retrieved positive instances. Precision is also referred to as the positive predictive value and it can be defined using true positive (TP) and false positive (FP) counts.

\[
\text{Precision} = \frac{TP}{TP + FP}
\]  

(4)

Recall measures the fraction of correctly classified positive instances among the actual positive instances, and it is defined using true positive (TP) and false negative (FN) counts.

\[
\text{Recall} = \frac{TP}{TP + FN}
\]  

(5)

The F1-score is defined as the harmonic mean of precision and recall.

\[
F1\text{-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]  

(6)

In order to analyse the accuracies of the nuclei detection models, a rule was needed to compare the ground truth coordinate annotations with the predicted nuclei locations. A predicted nuclei location was considered as a true positive detection when a ground truth annotation was within a certain radius from

<table>
<thead>
<tr>
<th>( R )</th>
<th>F1-score</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.800</td>
<td>0.823</td>
<td>0.780</td>
</tr>
<tr>
<td>8</td>
<td>0.861</td>
<td>0.885</td>
<td>0.839</td>
</tr>
<tr>
<td>10</td>
<td>0.879</td>
<td>0.904</td>
<td>0.857</td>
</tr>
<tr>
<td>12</td>
<td>0.886</td>
<td>0.911</td>
<td>0.864</td>
</tr>
<tr>
<td>14</td>
<td>0.891</td>
<td>0.917</td>
<td>0.869</td>
</tr>
<tr>
<td>16</td>
<td>0.895</td>
<td>0.921</td>
<td>0.873</td>
</tr>
<tr>
<td>18</td>
<td>0.898</td>
<td>0.924</td>
<td>0.876</td>
</tr>
<tr>
<td>20</td>
<td>0.902</td>
<td>0.928</td>
<td>0.880</td>
</tr>
<tr>
<td>22</td>
<td>0.905</td>
<td>0.931</td>
<td>0.883</td>
</tr>
</tbody>
</table>
the prediction. An optimal radius (R) was selected based on analysing the accuracy results of PT dataset using multiple different R values. The R values and corresponding F1-scores are presented in the Table II. However, the F1-score alone was not sufficient measure for selecting optimal R value, since higher values of R generated persistently higher F1-score. The reason for this is fundamentally in the inability to select completely correct metric for true detections while having wide scale of different sizes and shapes of nuclei and a coordinate annotation for a ground truth detection. Therefore, additional visual examination of different R values was performed to discover the optimal value for R. Three different R values are visualised in Fig. 3. Overall analysis resulted in selecting R=10. This decision and the problems concerning coordinate annotations are further addressed in the discussion chapter. All of the accuracy results presented in this paper are calculated using the selected optimal R value, excluding the results in Table II.

For MoNuSeg dataset, the groundtruth nuclei segmentations were provided instead of coordinates, therefore, a true positive was considered to be a detection that hits a segmented nuclear area.

Model Evaluation and Interpretation: In order to ensure that the model decisions are based on meaningful patterns in the input data, we utilised Layer-wise Relevance Propagation [26], [35]. Here we used the LRP implementation provided by the iNNvestigate toolbox [27]. The LRP is a technique for propagating the prediction backward in a neural network based on certain propagation rules. The method provides a heatmap that visualises positively and negatively relevant areas in the input image with respect to the classification task. We randomly selected a set of 64x64 sized fields of views around a detected nuclei from each of the processed image datasets. The set of sample images were analysed using LRP method and the generated relevance heatmaps were visually assessed in order to shed light on the meaningful patterns related to model decisions according to LRP analysis.

III. EXPERIMENTAL RESULTS

A. Nuclei Detection From Prostate Tissue

First, we consider the deep learning based nuclei detection from prostate tissue, for which we have an extensive annotated dataset of 67070 nuclei.

Each of the different fixation models were trained on the data that is defined by the model name. PAXgene model was trained on PAXgene fixed image data, and similarly formalin and frozen
models were trained on their respective image data. In total 17 models were trained, each model by leaving out all the data from one of the 16 patients and a final fixation model trained on the whole dataset of one fixation data. The 16 leave-one-patient-out models were used for calculating the results for a left-out image in a cross validation manner, and the reported result is the average from the 16 test images. The final fixation model trained with all data from the fixation type was then used to analyse the accuracy when detecting nuclei from other fixations.

We trained a baseline model with the whole PT dataset. The baseline model was used as an initial model in the domain adaptation step. In addition to the baseline model, 16 models were trained by leaving out all the data from one of the 16 patients, the nuclei from this left out patient data was then predicted using this leave-one-patient-out-model (LOPO-model). In total, 17 models were trained.

The numerical accuracy results are collected in Table III. Each row presents the detection accuracy of one model and the corresponding training and test data are specified in the following columns.

In rows 1-4, we present the results for the whole prostate tissue dataset with fixation specific models and with the baseline model. The F1-scores for fixation specific models range from 0.843-0.873, whereas the model trained with multiple fixations reaches 0.879 on the PT dataset. When comparing to the F1-score 0.78 reported in [5] for a two-step segmentation algorithm, all of the the deep learning based results presented here show clear improvement.

It can be observed that better F1-score is achieved when nuclei detection model training is carried out using image data with multiple fixations compared to the fixation specific models.

The precision and recall values for each model and test set are presented in the last two columns in the Table III. It can be noted that the models trained with visually better quality images (formalin, PAXgene) reach higher precision than a model trained with noisy frozen data. The precision values for PT data for formalin and PAXgene models are 0.925 and 0.903 respectively, whereas the frozen model precision reaches only 0.780. However, model recall is higher for a frozen model (0.924) compared to the formalin model (0.810) and the PAXgene model (0.850) recall values.

### B. Generalization to Multi-Organ Datasets

In order to test the generalization ability of our approach, we conducted experiments on a publicly available MoNuSeg dataset [17], and on another dataset with tissue from five organs (5-tissue dataset).

The results for the multi-organ data are shown in lines 14-17 of Table III. When testing the detection performance with MNS
data, the F1-scores are 0.730-0.799 and 0.802, revealing a clear drop in accuracy when compared to the prostate tissue dataset. For precision and recall, a similar pattern can be seen when testing with the MNS dataset as for the PT dataset; frozen model yields lower precision and higher recall when comparing to those by formalin and PAXgene models. Further, we applied a division to train and test sets according to the split applied in [17], and the results obtained with the baseline prostate model are listed in Table III, rows 18-19. The divergence in these results show that the dataset has significant variation in image characteristics.

Next, we further tested the generalization to other tissue types with our 5-tissue dataset. The results are shown in Table III lines 20–23. When testing with 5-tissue set, the F1-scores are 0.816-0.862 for the fixation specific models, and 0.883 for the baseline model. Again, similar patterns are observable in precision and recall, but this time the baseline model does not outperform PAXgene and frozen models in F1-score.

**Improved Generalization Through Unsupervised Pseudo-Label Domain Adaptation**

To enhance generalization of the deep learning models from PT dataset to other domains, we applied pseudo-label domain adaptation step. The baseline model trained on all PT data was used as a starting point for domain adaptation to the MoNuSeg data domain. The MoNuSeg data was divided into train and test datasets based on the division on the original paper. From the MoNuSeg train set, hard positive examples were collected using the detections using baseline model. The training data was generated as described previously in the Unsupervised generation of training samples from confidence map -section. Thus, the annotations provided with the dataset were merely utilized in the evaluation of the model - the DA step was fully unsupervised.

The resulting model after domain adaptation from PT to the MoNuSeg dataset is called MNS DA, and the results for all three datasets are listed on rows 24-26 in Table III. The results show that the unsupervised pseudo-label domain adaptation step enhances detection accuracy in all three cases. Specifically, the sensitivity is improved through domain adaptation; as the model gets samples from the new domain, it starts to detect more nuclei (i.e., the sensitivity increases). While the adaptation to the MoNuSeg data domain could be expected to improve accuracy for the MoNuSeg dataset, it was also the case for the 5-tissue dataset, and perhaps surprisingly, also for PT dataset.

For the sake of clarity and comparability, the F1-scores are also presented as boxplots in the Fig. 4. The nuclei counts predicted by different models were plotted against the manual counts and are shown in the Fig. 5, where (A) baseline and MNS data models show clear correlation with manual ground truth, (B) fixative-specific models yield more variation and divergence from manual ground truth, and (C) reference result using the two-step segmentation from [5] shows increased variance and bias which does not exist in the deep learning based results.

**Effect of Cell Density on Detection Accuracy**

Further, in order to show that the accuracy is not severely limited by the challenge caused by areas densely populated by nuclei, we conducted the following experiment: the prostate tissue dataset was divided pixelwise into five groups based on spatial density of nuclei, and the detection accuracy for baseline model as well as for baseline + MNS DA model was determined for each density group. The results, shown in Fig. 6 reveal there is only a minor drop in overall accuracy (F1-score) when moving from low-density (<20 nuclei per 50 μm) to high-density (>40 nuclei per 50 μm) areas. While the recall drops due to part of nuclei not being detected, the precision increases as there are
In order to generalize to images from new domains, including different tissues and images from different labs. In addition, we studied the effect of three histopathological sample fixation types on the accuracy of a nuclei detection trained with H&E stained images.

In order to allow further model development, we have shared the workflow implementation and the dataset, these are available at https://github.com/BioimageInformaticsTampere/NucleiDetection. The number of annotated nuclei in the provided dataset is considerably high compared to other public datasets. In addition, it includes three types of tissue fixation and processing and therefore enables development of fixation agnostic nuclei detectors or further study of the topic.

In order to study the effect of sample fixation on the accuracy of a nuclei detection, we trained multiple convolutional neural networks with varying training data. The numerical results are collected in the Table III. The detection results suggest that better
accuracy can be achieved when more variation in the sample fixation is present in the training data. This can be concluded when comparing F1-scores of the baseline model, trained with all three fixation types, and the models trained only with a single fixation data. The effect is similar when testing with each of the dataset. Correspondingly, after applying the unsupervised domain adaptation step using MoNuSeg dataset, the F1-score continues to increase.

Furthermore, the similarity between training and test data fixation can be perceived based on the results in the Table III. Detection in images from both PAXgene fixed and formalin fixed tissue sections is distinctly of better in quality compared to the images from frozen tissue sections. The similarity of image quality improves generalization of a machine learning algorithm. Consequently, PAXgene and formalin models detect nuclei nearly equally well nuclei from both PAXgene and formalin fixed tissue section images, yet, these models score low accuracies on the frozen tissue section image data. Similar effect can be concluded based on the precision and recall values. The model trained with noisy frozen data detects quite accurately nuclei from better quality images (PAXgene, formalin). Yet, conversely the PAXgene and formalin models fail (low recall) to detect the majority of the nuclei in images from frozen tissue sections. However, high precision indicates that when the nuclei is found it most often is a true positive. In the PT dataset, frozen model scores lower F1-scores compared to the formalin and PAXgene models. However, a contrary effect can be seen when testing with the 5-tissue data and MoNuSeg data. This seems to also indicate that the formalin and PAXgene fixed tissue sample images are quite similar concerning the image quality and therefore model generalization between these two datasets is decent. It also indicates that in order to generalise across tissue types, a frozen fixed tissue section images provide more variability in the image data domain.

When directly comparing the numerical results of the study, the problem with the metric itself, as well as the challenge caused by using a single coordinate as the ground truth, should be kept in mind. This was considered when analysing the effect of selected radius in the final accuracy values. The Table II presents how the F1-score reaches higher values when the R is increased and thus based on the Table alone one might argue that an even larger value for R should be utilized. The reality however is very different when looking into the examples of different R values in the Fig. 3. In the figure, white line is connecting the ground truth coordinate and the corresponding true positive detection. Here, a chain effect can be seen when one false negative is falsely detected as true positive by a prediction that is actually a signal from closely nuclei (see example D - R3). Thus, when selecting an optimal R, instead of looking at F1-score, the physical size and shape of nuclei present in the datasets should be considered.

In addition to selecting an optimal value of R, the location of a ground truth coordinate needs to be considered. An obvious challenge can be seen in Fig. 3 (E), where a benign elongated nucleus is shown. The location of a ground truth coordinate annotation can vary throughout the dataset, complicating the evaluation of true positive samples. Similar problem is faced when ground truth annotation is marked on the edge of a nucleus (see Fig. 3 (B) for an example).

Overall results confirm that sample fixation is a significant factor in the variability present in histological image data, and this should be considered when developing a robust and generalizable nuclei detection methods. Based on our results, increased cross domain generalization is achieved when multiple sample fixation methods are present in the training data. Our proposed pseudo-label based unsupervised domain adaptation step was shown to be beneficial for detection accuracy. However, the fully unsupervised domain adaptation step contains the risk of failed adaptation in cases where false positives are detected by the baseline model. In the experiments presented here, such problem did not occur. However, with ambiguous domain changes, e.g. when moving from HE to immunohistochemical staining, the unsupervised pseudo-labeling step may not provide adequate support for the new domain (see Supplementary Fig. 3 for such examples using the algorithm presented here and IHC data from [33].

In order to interpret the deep learning models and to discover the reasons behind model decisions, we visually assessed the model decisions using LRP method implemented in the iNNvestigate toolbox [27]. Few examples are shown in the Fig. 7. Based on these examples and visual assessment of multiple similar samples, the nuclei detection model seems to find reasonable areas important related to nucleus detection, such as nuclei edges. In addition, if nuclei are visible in case of a vesicular nucleus, those are often detected as important areas related to nuclei detection. Overall, the relevant areas provided by the LRP algorithm seem to correspond to the areas that a human observer would find relevant as well.

To conclude, this study addresses the question on the importance of the variability present in histological image data that is caused by the tissue fixation process, and the effects this variability has on the accuracy of a nuclei detection algorithm. We have shown with our experiments that the tissue fixation variability in the training data can cause significant differences between nuclei detection accuracies obtained by deep neural network models. The results of study are encouraging, and therefore, call for further research. A suitable next step would be to conduct experiments with bigger and more diverse datasets. In addition, more quantitative analysis of model explanation and interpretation methods are needed to build trust on the deep learning based approaches on these important biological questions. These steps will eventually enable development of a model that can generalize to real world clinical environments.

ACKNOWLEDGMENT

We are grateful to Noora Salokorpi for her skillful assistance in validation.

REFERENCES


