Tampere University

KIMMO KARTASALO

Machine Learning and 3D Reconstruction Methods for Computational Pathology

Tampere University Dissertations 415

Tampere University Dissertations 415

KIMMO KARTASALO

Machine Learning and 3D Reconstruction Methods for Computational 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 Jarmo Visakorpi auditorium
of the Arvo building, Arvo Ylpön katu 34, Tampere,
on 21 May 2021, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Medicine and Health Technology
Finland

Responsible supervisor or/and Custos	Docent Pekka Ruusuvuori Tampere University Finland	
Supervisors	Professor Matti Nykter Tampere University Finland	Professor Olli Yli-Harja Tampere University Finland
Pre-examiners	Docent Tuomas Mirtti University of Helsinki Finland	Honorary Clinical Associate Professor Darren Treanor University of Leeds United Kingdom
Opponent	Professor Johan Lundin Karolinska institutet Sweden	

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

Cover design: Roihu Inc.

ISBN 978-952-03-1952-6 (print)
ISBN 978-952-03-1953-3 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-1953-3

PunaMusta Oy – Yliopistopaino
Joensuu 2021

ACKNOWLEDGEMENTS

This doctoral research was conducted in the Bioimage informatics research group at Tampere University, Faculty of Medicine and Health Technology (Tampere University of Technology and University of Tampere until 2019) between 2015 and 2020.

The doctoral studies, including conference visits and research exchanges abroad, were made possible by much appreciated funding from Tampere University of Technology graduate school, Cancer Society of Finland, Doctoral Education Network on Intelligent Systems, Doctoral Programme on Biomedical Sciences and Engineering, Emil Aaltonen Foundation, Finnish Foundation for Technology Promotion, Finnish Society of Information Technology and Electronics, Industrial Research Fund of Tampere University of Technology, KAUTE Foundation, Orion Research Foundation, Päivikki and Sakari Sohlberg Foundation, Svenska sällskapet för bildanalys, Svenska tekniska vetenskapsakademien i Finland, Tampere University Foundation, Tampere University of Technology on World Tour program and Walter Ahlström Foundation (Tutkijat maailmalle -program).

The biggest thank you goes to my excellent supervisor, **Pekka Ruusuvuori**. Your way of running a research group has in only a few years resulted in just that - *a group*, in the true meaning of the word - with such team spirit and collaborative efficiency that neither the occasional digital nomad lifestyle of some of us nor a global pandemic have proven to be more than minor obstacles (if even that). To me it has always been clear that you have invested a lot of time and thought on figuring out how to help each one of us find our own way of reaching our full potential and feeling happy doing it. For any doctoral student, a healthy balance between having enough room for learning from mistakes and sufficient support for still being able to pull it off is something not to be taken for granted, but you have really hit the global optimum there. Even in those times when there have occasionally been more downs than ups, you have always kept up and protected the determination and positive spirit this group runs on, and ensured everyone feels included. True *sisu* at its best!

Most importantly, doing science the "BIIT way" has always been fun! I enjoyed the discussions on image analysis, life and global conspiracies we had at the office, over beers and online with you, **Kaisa Liimatainen**, **Mira Valkonen**, **Masi Valkonen**, **Gerardo Gonzáles**, **Hannu Hakkola** and everyone else who spent shorter periods in the group over the years. I would still be working on this thesis without the round-the-clock moral and debugging support. A special mention of course goes to **Leena Latonen**, who has always been there to ensure that we don't make ourselves a cancer biological laughing stock and would certainly be more than worthy of an honorary PhD degree in bioimage informatics by now! Your words of wisdom, appreciation of the subtleties of English punctuation rules and healthy (but enjoyably high) level of skepticism are highly valued by me.

I am grateful to **Matti Nykter**, who not only acted as a co-supervisor for this thesis, but guided my first steps in the world of science. The things you taught me when I was an undergraduate student still shape my approach to research and it's been a pleasure to be a part of the Computational biology group, first contractually and later socially, for almost a decade. Thanking everyone in the rapidly grown group who deserves it for their contribution to a fun and scientifically stimulating environment during these ten years would be impossible, but at the very least I would like to name **Ebrahim Afyounian**, **Matti Annala**, **Kirsi Granberg**, **Tomi Häkkinen**, **Sergei Häyrynen**, **Serafiina Jaatinen**, **Juha Kesseli**, **Suvi Lehtipuro**, **Simo-Pekka Leppänen**, **Gnanavel Mutharasu**, **Anssi Nurminen**, **Sofia Randelin**, **Tommi Rantapero**, **Ippa Seppälä** and **Francesco Tabaro**. All the current and former members of the biomedical research community in Tampere with whom I have studied, worked and spent time over the years would be too many to list here, but I am thankful to all of you. **Pasi Kallio** with the microsystem and stem cell friends deserve a special mention, though, as it is thanks to them I ended up working with microscopy image analysis during my Master's, which ultimately set me on the track towards this thesis.

Thanks to generous funding and the trust and confidence in me shown by Pekka and our friends in Sweden, I have had the chance of spending considerable amounts of time abroad as a guest doctoral student. Thank you for warmly welcoming me in your group in lovely Uppsala, **Carolina Wählby**, and trusting me with the project that since became a key part of my thesis. Spending time in such an environment with tens and tens of people who know what a Sobel filter is has been a unique

experience for me - not to mention the kindness and support this bunch of people excel in! Thank you, **Martin Eklund**, for without hesitation trusting me with the massive challenge and inviting me to join MEB at Karolinska Institute. I have already learnt a great deal from you and really value your determination for always going for what is fair and right, appreciation of scientific rigour, and dry humour. **Peter Ström** and **Henrik Olsson**, you have been the best of colleagues and friends, a real dream team! I could not have wished for better office mates and co-authors, our teamwork was like clockwork from day one and I'm so proud of what we managed to pull off together. You also deserve my gratitude for the never-ending patience required for reviving my Swedish - *tusen tack!*

Thank you, **Mattias Rantalainen**, for the chance to also contribute to your group's exciting projects at MEB plus for helping to keep up a credible level of geekiness in the otherwise medically oriented environment with your GPU rigs (...it's not noise, it's the sweet sound of science in the making!). Furthermore, thank you, **Lars Egevad**, for putting both your own and your global network's immense expertise on prostate pathology at our disposal, and for the seemingly infinite stock of witty, educational and invariably amusing anecdotes. I am also grateful to the entire Uppsala University image analysis community, especially **Axel Andersson**, **Eva Breznik**, **Ankit Gupta**, **Raphaela Heil**, **Chiara Költringer**, **Damian Matuszewski**, **Gabriele Partel**, **Nicolas Pielawski**, **Petter Ranefall**, **Sajith Sadanandan**, **Ida-Maria Sintorn**, **Leslie Solorzano**, **Amit Suveer**, **Elisabeth Wetzer**, **Håkan Wieslander** and **Johan Öfverstedt**, and all the other colleagues at MEB, especially **Alessio Crippa**, **Andrea Discacciati**, **Henrik Grönberg**, **Andreas Karlsson**, **Thorgerdur Palsdottir**, **Yinxi Wang** and **Philippe Weitz** for making me feel welcome and for their help and scientific contributions.

I wish to thank all the other co-authors who contributed to the publications in this thesis, including **Brett Delahunt**, **Hemamali Samaratunga**, **Toyonori Tsuzuki** and all other members of the **Pathology Imagebase expert panel**, **Johan Lindberg**, **Cecilia Lindskog**, **Jorma Vihinen** and **Tapio Visakorpi**, as well as **Olli Yli-Harja**, who acted as a member of the doctoral follow-up group. Moreover, **Mika Anttila**, **Juha Herrala** and the entire **Tampere Center for Scientific Computing**, as well as **CSC - IT Center for Science** are gratefully acknowledged for providing the computation capacity and associated support that have made this thesis possible.

Last, but definitely not least, I wish to warmly thank my mother **Kaarina**, fa-

ther **Juhani** and sister **Anna Kartasalo**. I don't think I would be writing a doctoral thesis now without having grown up in a home where I always felt loved, safe and supported, and I am very thankful for that. Dear **Miina**, I've been very fortunate to share the ups and downs of both life and science with you for all these years - thank you. My life would surely have been much more boring without you by my side on all of our adventures (many of which probably would have never got past the planning phase if it were down to me). Your family, **Emmi, Miisu, Leena** and **Timo Ojansivu**, have also been my second family - you are all very dear to me. I am also grateful to all of my friends and relatives for their love and support and for reminding me that there are other things in life apart from science.

Finally, I was planning to swim against the tide and not mention *you-know-what*, the constant front page news of the past six months or so. Considering my fascination for history and the thought that some day we will look back at this extraordinary period, I cannot resist the temptation to point out that this thesis was written entirely in relative isolation at home in Tampere and Stockholm during the Covid-19 pandemic. I believe and hope that in a few years time, these events will not seem as dramatic as they do now. After all, there is still more to science (and life) than the novel coronavirus.

Kimmo Kartasalo
Stockholm, 17 July 2020

ABSTRACT

Assessment of the microscopic anatomy of tissue samples forms the cornerstone of histopathological diagnostics. The current clinical practice is associated with challenges such as inter-observer variability and a global shortage of pathologists. Many fundamental aspects of pathology as a medical discipline have remained largely unchanged for decades, but the field is currently undergoing a transition into a digital discipline by replacing microscopes with whole slide scanners. Among other benefits, digital pathology unlocks the possibility of applying computational methods on the resulting image data. Some of the promises of computational pathology, such as improved efficiency and patient safety, take the advantages of digitization a step further, while others represent new types of analyses. This thesis focuses on two techniques in computational pathology: machine learning and 3D reconstructions.

Machine learning is a branch of computer science falling under artificial intelligence, which aims at emulating intelligent decision making. The field has progressed rapidly during the last decade due to the availability of larger datasets and improved computational resources. Deep learning in particular, representing a renaissance of artificial neural network algorithms, has demonstrated unprecedented performance across a range of problems and is seen as revolutionary for histopathology. By streamlining the work of pathologists, machine learning tools could potentially mitigate the issues with the unsustainable workload and inter-observer variability, and even enable the discovery of new image-based prognostic markers.

Digital imaging also enables 3D histology, where serially sectioned tissue samples are reconstructed computationally. Conventionally, 2D tissue sections representing only limited cross-sectional views of the original 3D samples are used. Studying tissue in 3D holds potential for obtaining a more comprehensive view of normal and pathological processes where the spatial arrangement of different tissue structures or cell types is of relevance. Compared to direct 3D imaging using specialized instruments, computational reconstruction allows applying various histological and

biochemical assays, while achieving subcellular resolution even for large tissue samples. The core methodological problem is how to align a sequence of 2D images to reconstruct a 3D volume without introducing distortions. Many algorithms have been proposed for the task, but an objective comparison of their performance has been lacking, complicating the application of 3D histology.

This thesis presents machine learning based systems for diagnostics of breast and prostate cancer, which represent a considerable fraction of all samples assessed in pathology departments worldwide. The system for assessing lymph node samples of breast cancer patients was based on extracting numerical features describing the tissue as input for random forest classifiers, and it was demonstrated to be capable of distinguishing between normal and metastatic tissue. This allows visually highlighting potentially malignant regions. The system for assessing prostate biopsies was based on deep neural networks and gradient boosted trees. It achieved clinically useful sensitivity and specificity in cancer detection, and cancer length estimates closely corresponding to those performed by a pathologist. In cancer grading, the system was comparable to a panel of specialized pathologists. This marks the first time that diagnostic performance comparable to specialists has been demonstrated on a large, clinically representative dataset of prostate biopsies.

The other two studies of the thesis present a framework for evaluating the quality of 3D reconstructions. The developed framework was applied to compare several publicly available algorithms and two commercial options. Moreover, the feasibility of automated hyperparameter tuning of reconstruction algorithms using Bayesian optimization was demonstrated for the first time. Algorithms relying on elastic transformation models capable of compensating for local tissue deformations were observed to achieve the most accurate reconstructions. Moreover, all of the studies in this thesis aimed at developing efficient ways of processing whole slide image data, resulting in a streamlined computational workflow utilizing parallel computing on graphics processing units on high-performance computer clusters.

Taken together, this thesis demonstrates that computational pathology techniques can achieve expert-level diagnostic performance, paving the way for the clinical adoption of such tools. The comparative results concerning 3D reconstruction algorithms highlight useful algorithmic features and hopefully promote further development of 3D histology from a prototype technique to a mainstream approach in biomedical research.

TIIVISTELMÄ

Kudosnäytteiden mikroskooppisen anatomian tarkastelu on histopatologisen diagnostiikan kulmakivi. Nykyisen kliinisen käytännön ongelmia ovat mm. eri patologien diagnoosien epäyhdenmukaisuus sekä maailmanlaajuinen patologipula. Patologia on monilta osin säilynyt vuosikymmeniä suhteellisen muuttumattomana, mutta ala käy nyt läpi digitaalista murrosta, jossa skannerit syrjäyttävät mikroskoopit. Digitaalinen patologia mahdollistaa muiden etujen lisäksi tuotetun kuvadatan laskennallisen käsittelyn. Laskennallinen patologia voi viedä joitakin digitalisaation hyötyjä kuten lisääntynyttä tehokkuutta ja potilasturvallisuutta entistäkin pidemmälle, mutta myös mahdollistaa aivan uudenlaista analytiikkaa. Tämä väitöskirja käsittelee kahta laskennallisen patologian tekniikkaa: koneoppimista ja 3D-rekonstruktioita.

Koneoppiminen on tekoälyn piiriin luettava tietotekniikan osa-alue, joka pyrkii jäljittelemään älykästä päätöksentekoa. Ala on kehittynyt nopeasti viimeisimmän vuosikymmenen aikana, pääasiassa kasvaneiden aineistojen ja nopeamman laskentakapasiteetin ansiosta. Erityisesti syväoppiminen, joka edustaa keinotekoisiksi neuroverkoiksi kutsuttujen algoritmien uutta aaltoa, on mahdollistanut ennennäkemättömät tulokset monissa eri ongelmissa. Tätä tekniikkaa pidetään mullistavana myös histopatologiaa ajatellen. Koneoppimiseen perustuvien työkalujen uskotaan voivan suoraviivaistaa patologien työtä ja siten helpottaa kestämätöntä työkuormaa ja parantaa diagnoosien yhdenmukaisuutta. Lisäksi ne voivat auttaa löytämään uusia, kuvapohjaisia tapoja ennustaa tautien kehittymistä.

Digitaalinen kuvantaminen mahdollistaa myös 3D-histologian, jossa sarjaleikattuja kudosnäytteitä rekonstruoidaan laskennallisesti. Tavanomaiset yksittäiset kudosleikkeet edustavat vain rajattua poikkileikkausta alkuperäisestä kolmiulotteisesta näytteestä. Kudoksen kolmiulotteinen tarkastelu voi auttaa kattavamman kuvan muodostamisessa sekä normaaleista että patologisista prosesseista, joissa erilaisten kudoksen rakenteiden ja solutyyppien keskinäisellä sijoittumisella on merkitystä. Suoraan 3D-kuvantamiseen verrattuna tavanomaisen mikroskopian pohjalta tehty

laskennallinen rekonstruktio sallii eri histologisten ja biokemiallisten tekniikoiden monipuolisen käytön ja mahdollistaa solutason erottelukyvyn suurillekin kudosnäytteille. Teknisesti tehtävä kiteytyy kysymykseen, kuinka sarja 2D-kuvia kohdistetaan toisiinsa ilman, että syntyvään 3D-rekonstruktioon muodostuu vääristymiä. Ongelmaan esitettyjä monia algoritmeja ei toistaiseksi ole kattavasti vertailtu, mikä hankaloittaa 3D-histologian käytännön soveltamista.

Tässä väitöskirjassa esitellään koneoppimisjärjestelmät rinta- ja eturauhassyöpien diagnostiikkaan. Nämä syövät edustavat kaikkialla maailmassa suurta osaa patologisista näytteistä. Rintasyöpäpotilaiden imusolmukenäytteiden arviointiin tarkoitettu järjestelmä perustuu suureen määrään kudosta kuvaavia numeerisia piirteitä sekä random forest -algoritmeihin, ja sen havaittiin kykenevän erottelemaan etäpesäkkeet normaalista kudoksesta. Järjestelmän avulla voidaan esittää visuaalisesti kunkin näytteen todennäköisesti pahanlaatuiset alueet. Eturauhaskoepaloja analysoiva järjestelmä perustuu syviin neuroverkkoihin ja gradient boosted tree -luokittelijoihin. Se saavutti kliinisesti käyttökelpoisen herkkyyden ja spesifisyyden syövän havaitsemisessa ja kykeni arvioimaan syöpäkudoksen pituuden kussakin koepalassa patologia tarkasti vastaavalla tavalla. Syövän pisteyttämisessä järjestelmä on verrattavissa joukkoon erikoistuneita patologeja. Kyseessä on ensimmäinen tutkimus, jossa on osoitettu asiantuntijoiden kanssa vertailukelpoinen diagnostinen suorituskyky laajalla, kliinisesti edustavalla eturauhaskoepala-aineistolla.

Muissa osatöissä esitellään 3D-rekonstruktioiden tarkkuuden arviointiin kehitetty menetelmä, jonka avulla vertailtiin useaa vapaasti saatavilla olevaa sekä kahta kaupallista rekonstruktiotyökalua. Lisäksi osoitettiin ensimmäistä kertaa bayesilaisen optimoinnin toimivuus rekonstruktioalgoritmien parametrien säätämisessä. Tarkimmat rekonstruktiot saavutettiin elastisia muunnoksia käyttävillä algoritmeilla, jotka kykenevät kompensoimaan kudoksen muodonmuutoksia. Kaikissa osatöissä tutkittiin myös tapoja digitaalipatologisen datan tehokkaaseen käsittelyyn ja kehitettiin laskentaklustereilla grafiikkaprosessoreilla suoritettavaa rinnakkaislaskentaa.

Yhteenvetona tämä väitöskirja osoittaa, että laskennallisen patologian keinoin voidaan saavuttaa asiantuntijatasoinen diagnostinen tarkkuus, mikä kannustaa vastaavien menetelmien kliiniseen käyttöönottoon. Eri 3D-rekonstruktiomenetelmien vertailu paljasti toimivia algoritmisia ratkaisuja ja voi toivon mukaan auttaa 3D-histologian jatkokehittämistä prototyypistä laajemmin käytetyksi biolääketieteellisen tutkimuksen menetelmäksi.

CONTENTS

ABBREVIATIONS

AI	artificial intelligence
ANN	artificial neural network
AR	augmented reality
ATRE	accumulated target registration error
AUC	area under the curve
Auto-ML	automated machine learning
BCa	breast cancer
CNN	convolutional neural network
CP	computational pathology
CV	cross validation
DL	deep learning
DNN	deep neural network
DP	digital pathology
FDA	Food and Drug Administration
GAN	generative adversarial network
GLCM	gray-level co-occurrence matrix
GPU	graphics processing unit
GS	Gleason score
HE	hematoxylin & eosin
HPC	high-performance computing
IHC	immunohistochemistry

ISUP	International Society of Urological Pathology
kNN	k-nearest neighbors
LBP	local binary pattern
MI	mutual information
ML	machine learning
MRI	magnetic resonance imaging
MSE	mean squared error
NCC	normalized cross correlation
NMI	normalized mutual information
PCa	prostate cancer
RF	random forest
RMSE	root mean squared error
RNN	recurrent neural network
ROC	receiver operating characteristics
ROI	region of interest
RP	radical prostatectomy
SIFT	scale-invariant feature transform
SURF	speeded up robust features
SVM	support vector machine
t-SNE	t-distributed stochastic neighbor embedding
TMA	tissue microarray
TRE	target registration error
WSI	whole slide image

ORIGINAL PUBLICATIONS

Publication I **Kartasalo, K.**, Latonen, L., Visakorpi, T., Nykter, M. and Ruusuvuori, P. (2016). Benchmarking of algorithms for 3D tissue reconstruction. *IEEE International Conference on Image Processing*, 2360–2364.

Publication II **Kartasalo, K.**, Latonen, L., Vihinen, J., Visakorpi, T., Nykter, M. and Ruusuvuori, P. (2018). Comparative analysis of tissue reconstruction algorithms for 3D histology. *Bioinformatics* 34.17, 3013–3021.

Publication III Valkonen, M.*, **Kartasalo, K.***, Liimatainen, K., Nykter, M., Latonen, L. and Ruusuvuori, P. (2017). Metastasis detection from whole slide images using local features and random forests. *Cytometry Part A* 91.6, 555–565.

Publication IV Ström, P.*, **Kartasalo, K.***, Olsson, H., Solorzano, L., Delahunt, B., Berney, D. M., Bostwick, D. G., Evans, A. J., Grignon, D. J., Humphrey, P. A., Iczkowski, K. A., Kench, J. G., Kristiansen, G., van der Kwast, T. H., Leite, K. R., McKenney, J. K., Oxley, J., Pan, C.-C., Samaratunga, H., Srigley, J. R., Takahashi, H., Tsuzuki, T., Varma, M., Zhou, M., Lindberg, J., Lindskog, C., Ruusuvuori, P., Wählby, C., Grönberg, H., Rantalainen, M., Egevad, L. and Eklund, M. (2020). Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet Oncology* 21.2, 222–232.

Equal contribution indicated by asterisk (*). The original publications and figures are reprinted with permission from the copyright holders.

Author's contribution

Publication I: First author. The author was primarily responsible for study design, implementation, analysis of results and drafting the manuscript.

Publication II: First author. The author was primarily responsible for study design, implementation, analysis of results and drafting the manuscript.

Publication III: Joint first author with Mira Valkonen (Tampere University). The author was primarily responsible for the whole slide image segmentation and pre-processing steps. Valkonen was primarily responsible for the machine learning steps. Both authors contributed equally to the analysis of results and drafting the manuscript.

Publication IV: Joint first author with Peter Ström (Karolinska Institute). The author was primarily responsible for whole slide image segmentation and pre-processing steps and for high-performance computing. Ström was primarily responsible for statistical analysis of results. Both authors contributed equally to study design, the design and implementation of the machine learning steps and to drafting the manuscript.

1 INTRODUCTION

Histological assessment of tissue samples, that is, the evaluation of their microscopic anatomy, is frequent in biomedical research and forms the cornerstone of histopathological diagnostics in clinical practice. Conventionally, these assessments are performed visually by an expert, such as a pathologist, with the help of a microscope. Many fundamental aspects of pathology as a medical discipline have remained largely unchanged for decades, but the field is currently undergoing a transition into a digital discipline (Griffin et al. 2017). In digital pathology (DP), the microscopy slides are scanned and examined on computer screens (Fig. 1.1). Digitization holds promise for a number of advantages, such as time savings, improvements in quality and patient safety aspects, and possibility for efficient remote consultation.

Figure 1.1 Microscopy slides (left), Aperio scanner[1] (center), a WSI of a prostate biopsy (right).

In spite of these factors, and the fact that whole slide image (WSI) scanners enabling the automated imaging of histological samples have been available since the 1990s, the adoption of digital workflows in clinical laboratories has been slow (Colling et al. 2019; Griffin et al. 2017). Some of the potential reasons for the hesitant adoption have been the requirement for considerable initial investment, technical shortcomings of digital systems, reluctance of pathologists to adopt digital workflows due to these shortcomings and their familiarity with conventional microscopes, and reg-

[1] https://www.leicabiosystems.com/digital-pathology/scan/aperio-cs2/, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=77363624.

ulatory reasons such as the lack of approval by the US Food and Drug Administration (FDA) until recently. While there is an increasing number of digital laboratories around the world, the potential benefits of DP have not yet been able to sufficiently outweigh the drawbacks to promote widespread clinical adoption of the technology.

Besides the logistical benefits offered by storage and examination of samples in digital format, DP also unlocks the possibility of applying computational image processing and analysis techniques on the resulting data (Griffin et al. 2017). In particular, WSI systems serve as an enabling technology for the application of machine learning (ML) based image analysis (Bera et al. 2019). Machine learning is a branch of computer science falling under artificial intelligence (AI), which aims at computationally emulating intelligent decision making. The field has seen rapid progress during the last decade, mainly attributable to the renaissance of biologically inspired artificial neural network (ANN) techniques in the form of deep learning (DL) using deep neural networks (DNN) (LeCun et al. 2015). Today, the terms machine learning, deep learning and AI are often used interchangeably. The unprecedented performance of DNNs across a wide range of problems has not gone unnoticed in medicine and there is widespread optimism around the application of ML in the medical domain (Rajkomar et al. 2019), including DP.

Some of the promises of AI for pathology mirror the advantages associated with digitization, but take them a step further (Bera et al. 2019; Niazi et al. 2019). For example, a digital workflow could improve safety by reducing human errors in the information management process, such as the handling of patient identifiers and slide labels. The use of AI could extend the quality assurance to the diagnosis itself, issuing a warning to the pathologist about potentially missed cancers. Similarly, digitization has been claimed to provide time savings by reducing the need to physically transport slides, and AI tools could potentially further improve efficiency for example by automatically excluding cases that are likely to be benign from the pathologist's review. By streamlining the work that pathologists currently do, AI-based tools could help manage the worldwide workload challenge associated with a decreasing number of practising pathologists and an increasing number of samples that need to be assessed. These added benefits offered by AI and other computational approaches might encourage more laboratories to go digital.

Besides automation of routine tasks, computational approaches are expected to outperform humans by learning from datasets containing more cases than any medi-

cal expert will assess during their career, and to enable completely new analyses (Bera et al. 2019). One such prospect enabled by computational techniques is 3D histology, where tissue samples are analyzed in their native three-dimensional setting (Roberts et al. 2012). Conventionally, histological assessments rely on 2D sections cut from the sample, offering only a view based on limited cross-sections of the underlying 3D structures. Imaging tissue in 3D is possible using specialized instruments, but using digital image processing, the 3D structure can be reconstructed based on a series of 2D tissue sections prepared conventionally (Pichat et al. 2018).

To describe the field focused on solving problems in pathology via computational methods, the term 'computational pathology' (CP) was coined by Fuchs and Buhmann (Fuchs 2010; Fuchs et al. 2011). I too have chosen to use this term to emphasize the computational focus of this work, as opposed to the broader topics encompassed by DP, and to consider the digital workflow as an established platform for computational approaches. In doing so, we ignore many questions that have been widely studied within the DP community, such as the concordance between clinical diagnoses performed using conventional microscopes and WSI systems, logistics and information management in a digital laboratory and most regulatory aspects. Instead, this work focuses on the algorithmic methods built on top of the digital workflow, representing the transition from digital to computational pathology.

The articles included in this thesis focus on two separate topics: 3D histology and ML for cancer diagnostics. The aim of Publication **I** was to reconstruct 3D histology from 2D serial sections, and to develop a framework for assessing the quality of the reconstructions. In Publication **II**, this framework was extended and applied to compare several reconstruction algorithms. In Publication **III**, a ML system for detecting breast cancer (BCa) metastases in lymph node samples was presented. In Publication **IV**, a DL-based system for prostate cancer (PCa) diagnostics using biopsies was developed. A further aim shared by all four studies was the development of high-performance computing (HPC) approaches for processing the WSI data. The following chapters will present a review of the literature (Ch. 2), define the aims of the work (Ch. 3), briefly introduce the methodology used (Ch. 4) and summarize and discuss the results (Ch. 5). The clinical relevance of the results and the future of CP are then discussed in Chapter 6, followed by concluding remarks in Chapter 7, and the original publications.

2 BACKGROUND

2.1 Digital and computational pathology

Whole slide imaging systems have now been available for more than 20 years (Pantanowitz et al. 2018). In 2013, the lack of comprehensive diagnostic validation of the technology was still pinpointed as the main barrier to large-scale adoption (Ghaznavi et al. 2013). Over the course of the past decade, guidelines for conducting validation studies have been published and several such studies have shown broad concordance between diagnoses made using digital and conventional pathology (Griffin et al. 2017). A milestone event recently took place, when the long-standing obstacle of lacking FDA approval of WSI systems was cleared by Philips with their IntelliSite Pathology Solution, which was shown to be non-inferior to microscopy for primary diagnosis in surgical pathology (Mukhopadhyay et al. 2018). This success has simplified the process for other manufacturers, leading to Leica obtaining approval for their Aperio AT2 DX scanner in 2019. Overcoming these hurdles is expected to promote the rapid adoption of DP in clinical use (Pantanowitz et al. 2018). While many open questions regarding the detailed clinical implementations still need to be answered, the scientific focus is already partly shifting towards the next steps of computational and AI-based approaches (Colling et al. 2019).

It is worth noting, that the application of CP does not necessarily require WSI systems and a fully digital workflow. Google Health proposed an augmented reality (AR) approach, where computation is integrated into an otherwise conventional microscope (P.-H. C. Chen et al. 2019). In this solution, a DNN-based diagnostic system was trained offline using WSI data and a microscope was retrofitted with a digital camera and an AR display. As a result, the user can operate the microscope as usual, while being aided by seeing the outlines of predicted malignant regions. Compared to a fully digital workflow, this approach can potentially harvest some of the benefits of CP while avoiding the costs of setting up scanning infrastructure. More-

over, pathologists would not be required to adapt to working digitally and could avoid some of the technical issues that still persist with most WSI systems, such as inability to examine different focus levels of the sample (Griffin et al. 2017).

While improvements in WSI systems have undeniably accelerated the progress of CP, computational image analysis has been applied irrespective of the imaging modality as long as digital images of samples have been available (Prewitt et al. 1966). Early studies on what would be called CP today relied on limited numbers of photomicrographs representing manually selected fields of view from the sample. In fact, the earliest CP algorithms did not even rely on digital images at all, but instead sought to formulate pathological knowledge and diagnoses into representations that would allow computational processing and decision support (Hamilton et al. 1994; Heathfield et al. 1991). As exemplified by these early approaches and the contemporary AR microscope, CP can be seen not only as the next step following full digitization of pathology, but as a branch of computer science and (bio)image analysis that has progressed in parallel with DP. It remains to be seen if a partially digital, but nevertheless computational, approach to pathology will amount to more than a curiosity, or if large-scale whole slide scanning remains the primary workhorse powering a transition to fully digital laboratories in the future. Currently, the development of CP methods builds almost exclusively on image analysis of WSI data (Abels et al. 2019), however, and plans for large-scale clinical implementation typically rely on the full digitization of pathology departments (Colling et al. 2019).

2.2 Histopathological diagnostics

Development of AI solutions for the diagnostics of prostate and breast cancer is particularly relevant due to the large volumes of these samples analysed worldwide. In many regions, including the Nordic countries, PCa is the most common cancer among men (Bray et al. 2018). Globally, it is the second most common cancer among men, second only to lung cancer. Breast cancer, on the other hand, is the most frequently diagnosed cancer and the most common cause of cancer death among women globally. The histopathological assessment of biopsies and resected specimens is crucial for the diagnostics and treatment choices. These assessments are formalized into various systems and guidelines, such as the so called TNM staging system (Edge et al. 2010). Classification of cancers following this system is based on

the size of the primary tumor (T), spread of cancer to regional lymph nodes (N), and the presence of distant metastasis (M). An important task in BCa staging is the assessment of sentinel lymph nodes for the presence of metastatic cells (Bejnordi et al. 2017). However, this is a time-consuming task, and sensitivity and inter-pathologist reproducibility often remain sub-optimal (Vestjens et al. 2012).

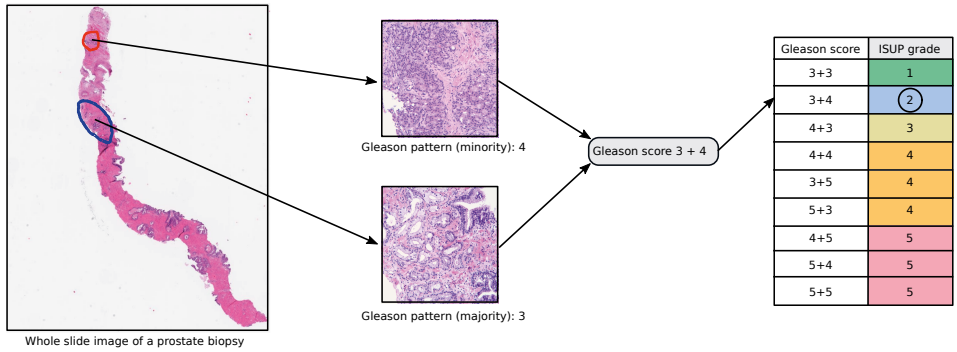

Figure 2.1 Simplified illustration of the Gleason grading process. The most common and second most common Gleason patterns in a prostate sample are combined into a Gleason score. The Gleason score is further converted into an ISUP grade group.

The diagnosis of PCa is based on examination of needle biopsies extracted from the prostate, graded following the Gleason grading system (Fig. 2.1), where higher Gleason scores (GS) are associated with a worse prognosis and a need for more radical treatment (Bulten et al. 2020). Slightly simplified, the GS is based on assigning a Gleason grade on a scale from 1 to 5 for the dominant and secondary morphology present in the sample, and calculating the sum of the two grades. In practice, only grades 3-5 are used. The grade itself is based on the growth patterns of the tumor, visually assessed by the pathologist. The International Society of Urological Pathology (ISUP) recommends reporting with an updated five-step system (ISUP grade groups), which takes into account the different prognostic values of GS 3+4 and 4+3 (Epstein et al. 2016). The difficulty of producing objective and reproducible assessments is reflected in high variability in Gleason grading performed by different pathologists, which in turn represents a key problem for the clinical management of PCa (Egevad et al. 2013). As the Gleason grading is the most important factor in view of treatment decisions, misclassifications can have severe consequences for both individual patients and health care systems. The use of reference databases (Egevad et al. 2017) and AI tools (Egevad et al. 2020) are hoped to mitigate these issues.

2.3 Image analysis in computational pathology

The methodological roots of CP lie in bioimage analysis (Meijering et al. 2016) and much of the early research in the field was based on an object-centered view built using classical image analysis operations (Gurcan et al. 2009). This means that the images are treated as collections of objects, such as cell nuclei, which in turn may serve as constituents of other objects, such as glands. This can be seen as a logical extension of the algorithms for analyzing images of individual cells developed e.g. for confocal microscopy (Schneider et al. 2012), simply scaled up to analyze the larger numbers of cells encountered in pathology. As an example, an early attempt at automated Gleason grading relied on detecting nuclei, modeling their geometrical arrangement with spanning trees and comparing the obtained trees to a database of samples with known Gleason grades (Wetzel et al. 1999). In line with this evolution, the initial focus of pathological image analysis tended to be on cytology, where isolated cells are examined instead of tissues (Gurcan et al. 2009). This simplified the analysis due to the absence of complex structures and the relative ease of segmenting individual cells or nuclei on the homogeneous background of cytological samples.

Most workflows in bioimage analysis and CP can be decomposed into a number of key image processing tasks (Kothari et al. 2013; Meijering et al. 2016) (Fig. 2.2). Initial pre-processing may involve quality control to e.g. detect artefacts such as poor focus or folded tissue (Janowczyk et al. 2019; Palokangas et al. 2007) and correction for uneven illumination, noise or optical blurring via deconvolution. In pathology, color is of particular interest (Clarke et al. 2017), and pre-processing is often employed to reduce variation due to differences in tissue staining and imaging. This may involve e.g. color normalization relative to a reference slide or stain deconvolution, i.e. separation of the signals representing different stains such as hematoxylin and eosin (HE) (Bejnordi et al. 2015; Khan et al. 2014; Macenko et al. 2009; Ruifrok et al. 2001). A key pre-processing operation for many applications is image registration (Sotiras et al. 2013; Zitova et al. 2003), where images representing e.g. different specimens, imaging modalities or histological stains are aligned into a shared coordinate frame to establish pixelwise correspondence between the images. Image registration allows e.g. joint analysis of multiple immunohistochemistry (IHC) stains (Borovec et al. 2020), cross-modality analysis between e.g. histology and magnetic resonance imaging (MRI) (Reynolds et al. 2015) and 3D reconstructions (Roberts et al. 2012).

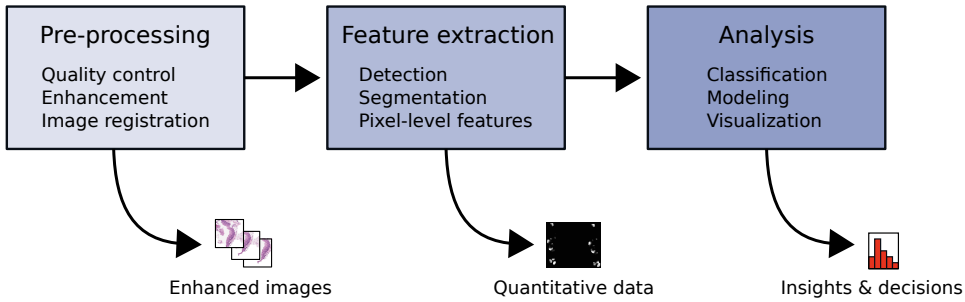

Figure 2.2 A typical bioimage analysis workflow. Pre-processing steps are often applied to produce enhanced versions of the input images, which may serve as the final output in some cases. Subsequently, numerical features are extracted to characterize the objects and the texture present in the images, yielding a quantitative representation of the input data. Finally, this representation can be further analyzed to gain insights of the process being studied, or to build predictive models.

Some bioimage analysis workflows may merely consist of pre-processing, producing enhanced and registered images as output for visual examination or archiving, but further image analysis steps are often required. Following the classical object-based approach, most of these steps can be formulated as either detection or segmentation (Gurcan et al. 2009). Detection consists in estimating if some objects of interest are present or absent, as well as often estimating their locations in the image. In segmentation, the aim is to group image pixels into sets that represent the objects of interest, allowing one to additionally quantify the sizes, shapes and other properties of the objects. Over the years, considerable effort has been devoted to the development of detection and segmentation algorithms tailored for specific cellular organelles, such as nuclei (Irshad et al. 2013). A number of numerical parameters (e.g. diameter, eccentricity or mean brightness) are then typically extracted to describe the objects quantitatively in the form of object-level features (Kothari et al. 2013). An alternative feature extraction approach is to compute pixel-level features, including e.g. histogram statistics, Haralick features based on gray-level co-occurrence matrices (GLCM) (Haralick et al. 1973) or local binary patterns (LBP) (Ojala et al. 2002; Pietikäinen et al. 2000). Pixel-level features can be employed to obtain a quantitative description of the image independently of any notion of objects but they can also be used as the basis for object detection or segmentation. One may also combine these approaches and compute pixel-level features for subsets of pixels belonging to individual segmented objects to describe nuclear texture, for example.

Analysis steps following feature extraction vary greatly depending on the application, but typically they represent some form of classification, modeling, visualization or exploratory analysis of the resulting data (Kothari et al. 2013; Meijering et al. 2016). One may for example study how different samples cluster in feature space with the aim of discovering morphological patterns in an exploratory setting (Valkonen, Ruusuvuori et al. 2017), or use the features as input for ML algorithms with the aim of predictive modeling of clinical diagnoses (Tabesh et al. 2007). In tasks like cell counting, the extracted features may already contain the desired information without further analysis. Most state-of-the-art approaches integrate the feature extraction and ML steps into a single, automated process using DNNs (Gupta et al. 2019).

In addition to the complexity of the patterns present in histological data, the large data volume, especially in the case of WSIs, represented a significant computational bottleneck for years (Gurcan et al. 2009). As an example, the data used in Publication **IV** contained over 8600 WSIs with dimensions of approximately 50 000 x 30 000 pixels each. This amounts to roughly 1.3×10^{13} pixels and, when expressed at 24 bits per pixel, 39 terabytes of raw image data. Although a considerable amount of these pixels can be excluded as background, a dataset of this scale would have been an insurmountable challenge for most of the computer systems successfully used to process cytological images a decade ago. In radiology, the data volumes have historically been even more manageable than in the case of cytology (Gurcan et al. 2009).

As a result of the increasing amounts of data and computation power becoming available, bioimage analysis in general, and CP in particular, have become data-intensive disciplines where statistical ML plays a central role (Fuchs et al. 2011; Meijering et al. 2016). Top-performing solutions to detection (Bejnordi et al. 2017), segmentation (Ronneberger et al. 2015), classification (Bulten et al. 2020) and even quality control (Kohlberger et al. 2019) tasks in CP today rely on ML, and more specifically DNNs, nearly without exception. In Section 2.4 we will focus on ML, with an emphasis on applications to PCa and BCa diagnostics in view of Publications **III-IV**. However, more classical image processing methods still remain important for many pre- and post-processing operations, and some tasks are less amenable to formulation as a ML problem than classification or segmentation, with image registration being a notable example. ML-based solutions to medical image registration have only recently started to emerge (Fan et al. 2019; Haskins et al. 2020; Miao et al. 2016), but applications to histology are still nearly non-existent (Awan et al. 2018). Image reg-

istration is the cornerstone of 3D histology (Pichat et al. 2018), which we will focus on in Section 2.5 in view of Publications **I-II**.

2.4 Machine learning in computational pathology

2.4.1 Feature-based learning for diagnostics

The early attempts of employing ML to detect cancer from histopathological images relied on tens to hundreds of photomicrographs representing manually picked regions of interest (ROI) (Jafari-Khouzani et al. 2003; Tabesh et al. 2007). These studies relied on general-purpose features popularized in the image processing community, such as those based on wavelet transforms, which provide a natural multi-scale framework for texture analysis (Laine et al. 1993), fractal geometry (Jacquin 1993) and GLCM statistics (Haralick et al. 1973), as well as tailored morphological features based on segmenting the image into various histological objects. Support vector machine (SVM) (Cortes et al. 1995), Gaussian and k-nearest neighbors (kNN) classifiers were trained in these studies to classify prostate tissue into benign or malignant, and further into different Gleason grades. Tabesh *et al.* reported accuracies of 96.7% and 81.0% for the tasks of classifying benign vs. malignant tissue ($n = 367$) and Gleason grading ($n = 268$), respectively, while Jafari-Khouzani and Soltanian-Zadeh reported 97% accuracy in Gleason grading ($n = 100$). While impressive for their time, one should note that the performance estimates only rely on cross validation (CV) on a limited number of ROI. With regards to classification models, it is worth noting that despite the popularity of SVM and random forest (RF) (Breiman 2001) classifiers prior to the 2010s DL era, neural networks in their shallow multilayer perceptron form have been applied in CP already decades ago (Cheng et al. 1995).

An early work by Diamond *et al.* (Diamond et al. 2004) is notable in its patch-based approach, where images of radical prostatectomy (RP) specimens were divided into sub-regions, each processed in turn to obtain predictions across a full slide. This has become a standard approach in contemporary WSI based studies (Bejnordi et al. 2017), including Publications **III-IV**. Another patch-wise method for WSI analysis used a multi-resolution, RF-based framework for classifying BCa samples following the modified Bloom-Richardson grading (Basavanhally et al. 2013). The features engineered in that study were mainly based on the segmentation of nuclei, and char-

acterization of their organization using graphs constructed by Voronoi diagrams, Delaunay triangulation and minimum spanning trees.

Likewise, the studies by Doyle *et al.* (Doyle et al. 2010; Doyle et al. 2006) are noteworthy as they attempted to analyze WSI of entire samples. The obvious challenges posed by the computational capacity at the time were circumvented by designing a multi-resolution approach inspired by the way pathologists work with the samples. The images were decomposed into a resolution pyramid, and the feature extraction and classification process was first applied at a low resolution to exclude regions that are likely to be benign. Potentially malignant regions were then re-analyzed at higher resolution, and the process was repeated iteratively over several resolution levels. The large pool of over 900 features, mined using an AdaBoost (Freund et al. 1996) ensemble method is also reminiscent of studies published years later (e.g. Publication **III**). On a dataset of 100 prostate biopsy WSIs from 58 patients, a receiver operating characteristics (ROC) analysis estimated an area under the curve (AUC) of 0.85 in pixel-wise PCa detection. Notably, Doyle *et al.* not only utilized a relatively large number of slides, but also performed patient-level CV, meaning that all biopsies from a given patient were included in either training or testing data, which has since been indicated as an important consideration when evaluating classifiers (Nir et al. 2019). Other early feature-based approaches have been reviewed in detail elsewhere (Gurcan et al. 2009; Kothari et al. 2013; Mosquera-Lopez et al. 2014).

A more recent study investigated feature selection approaches to improve generalization across histopathological data from multiple sites (Leo et al. 2018). In DNN-based approaches, the problem is typically attacked by applying data augmentation during training to encourage learning representations that are invariant to perturbations resembling the inter-site variation (Tellez, Litjens, Bandi et al. 2019). In the feature-based setting, one can explicitly select features that are not only discriminative but also robust to the inter-site variation. The study was based on a total of 212 RP samples from four institutions and used ROIs cropped from the WSIs to represent benign, Gleason grade 3 and Gleason grade 4 tissue. A set of 242 features was considered, including descriptors of gland morphology and Haralick features. The authors achieved an improvement of 4.38 % in AUC for the task of distinguishing between Gleason patterns 3 and 4, when compared to feature selection that only considers discriminative capacity of the features. However, in the classification of malignant versus benign tissue, no improvement was reported.

In contrast to object-level features designed to capture some biological property of the tissue, such as the organization of nuclei relative to glands (D. Wang et al. 2015), many pixel-level features, such as Gabor filter banks (Lee 1996), lack straightforward biological or physical interpretations (Kothari et al. 2013). Arguably, approaches relying on large numbers of such generic, low-level features represent a gradual transition from the classical methods, which utilized small numbers of biologically meaningful descriptors carefully tailored to a particular task, towards the current DNN-based paradigm, where feature engineering has become nearly obsolete. For example, in the work of Wang *et al.* (H. Wang et al. 2017), multi-scale features including LBP (Ojala et al. 2002; Pietikäinen et al. 2000), Gabor filter banks, Haralick features and Pyramid Histogram of Visual Words (Bosch et al. 2007) were extracted and used as input to train a RF classifier for detecting pixels that represent cancer using prostate MRI scans and digitized biopsies. Gertych *et al.* used a similar concept for PCa detection (Gertych et al. 2015): instead of segmenting objects, they extracted histogram features from deconvolved HE stain components to discriminate epithelial from stromal tissue using an SVM, followed by pixel-wise classification of the epithelial tissue into benign and malignant using LBP and local variance features as input for a RF classifier. Publication **III** is also an example of an approach combining elements from the earlier object-based tradition with the now dominant ideas of "blindly" employing large numbers of pixel-level features as the basis for classifiers with very high capacity. Deep learning can be seen as a further step in this direction in the evolution of ML-based computational pathology.

2.4.2 Feature-based learning for quality control

Feature-based ML has also been applied for quality control. For example, Hashimoto *et al.* used linear regression to derive a quality index consisting of terms representing image sharpness and noise (Hashimoto et al. 2012). Training images annotated by pathologists according to their perceived quality were utilized to estimate weights for the individual features constituting the index. The feature used for sharpness quantification with the aim of detecting poor focus was based on the Canny edge detector (Canny 1986), whereas noise quantification was performed by a filtering process that enhances isolated pixels exhibiting large intensity gradients along all directions. Moles Lopez and colleagues used 48 000 patches extracted from 27 WSIs

containing heterogeneous tissue morphology stained with HE and IHC as training data and 3438 patches from 97 WSIs for evaluation (Lopez et al. 2013). A total of 16 features were used to estimate sharpness of the patches, including several gradient-based descriptors and GLCM features. Decision trees and SVM classifiers were then trained for classifying patches as sharp or blurred. As a result, the system was able to generate maps indicating poorly focused regions on whole slides.

2.4.3 Deep neural networks

Deep learning is a family of ML methods which can be classified as representation learning (LeCun et al. 2015). The fundamental difference compared to feature-based approaches is that DL bypasses the need of designing feature extraction steps, where the raw data are transformed into a representation capturing the patterns relevant for the task. Instead, learning a relevant representation is part of the training phase of DNNs. In DL, the representations are composed of a multi-level, or deep, hierarchy of relatively simple transformations, which together have the capacity to learn complex, non-linear functions. This approach has proven to be highly successful in discovering useful patterns from complex, high-dimensional data, such as images.

Modern DL builds on the research conducted on ANNs between the 1940s and 1990s (Schmidhuber 2015), initially based on neurobiological considerations and leading to the invention of the perceptron, a simple model of a neuron functioning as a linear classifier (Hebb 1949; McCulloch et al. 1943; Rosenblatt 1958). Multi-layer perceptrons allowed more complex representations to be learned by composing multiple layers of perceptrons, and the neocognitron (Fukushima 1980) introduced the concept of convolutional neural networks (CNN), which are central in modern DL. Backpropagation of errors and optimization of neural network parameters via gradient descent, which is the dominant algorithm for supervised training of DNNs today, was also introduced nearly four decades ago (Linnainmaa 1970; Rumelhart et al. 1986; Werbos 1982). Still, training ANNs consisting of several layers to discover useful features was considered difficult, and the interest of the ML and computer vision communities shifted to other types of models (Schmidhuber 2015).

Despite a period of domination by other approaches like SVMs (Cortes et al. 1995), a string of incremental improvements to ANN training took place during the 1990s and early 2000s (Schmidhuber 2015). As larger training datasets have become

available, these algorithmic improvements, such as stochastic gradient descent based training accelerated on graphics processing units (GPU), have enabled DNNs to efficiently utilize the increasing quantities of data. An event frequently quoted as the beginning of the renaissance of ANNs and the current DL era took place in 2012, when a CNN dominated the popular image classification competition ImageNet, which was based on one million images depicting 1000 different classes, achieving close to half the error rate of its closest competitors (Krizhevsky et al. 2012).

The distributed, hierarchical representations learned by CNNs make them a particularly good fit for the properties of many natural signals, most notably images (LeCun et al. 2015). In a CNN, weight sharing across different locations of the input space is achieved by performing convolution between the input signal and filters with learned weights. This has the effect of being able to learn representations of local patterns, such as objects of a specific type, irrespective of the location of the patterns in the input. By alternating between such convolutional layers, and pooling layers, where similar features are pooled together, CNNs are able to model hierarchical structure in data. As an example, it is typical for natural images to contain minute patterns, such as edges, which act as the building blocks of individual parts, which in turn form objects. A particular collection of objects can in turn have a more abstract semantic meaning. In contrast to ANNs of the earlier decades, deep CNNs and other types of DNNs can consist of tens or even hundreds of layers. Since the ImageNet success, DNNs in general, and CNNs in particular, have become the dominant approach for most detection, recognition and segmentation tasks.

The application of DL to medical images mirrors the wider adoption process of the technology, that is, there are sparse examples from earlier decades (Sahiner et al. 1996), but interest in solving medical problems using DNNs has only truly peaked after the breakthroughs in other image analysis tasks (Greenspan et al. 2016). The medical domain poses some special challenges: data collection is typically costly, patient privacy issues need to be considered, and labeling images for training requires medical experts, whose time is a scarce resource. Still, DL has gained dominance also within medical imaging and pathology (Janowczyk et al. 2016). Before, the designer had to closely examine the data and collect domain knowledge to recognize relevant patterns, and then design suitable features to capture them. Designing features that generalize well typically required manual examination of cases where the system failed, and iterative improvement of the features to cover the problematic parts of

the input space. After having optimized the design for a given dataset and task, the process typically had to be repeated at least partially on new data sources and tasks. Streamlining this process by DL allows scalable utilization of the growing datasets generated by WSI systems. Some of the first problems in CP attacked with modern DL included e.g. detection of mitotic cells (Roux et al. 2013; Veta et al. 2015) and invasive ductal carcinoma (Cruz-Roa et al. 2014) for BCa diagnostics.

2.4.4 Deep learning for diagnostics

The field of DL-based computational pathology has expanded tremendously during the last few years and several comprehensive reviews have already been published (Bera et al. 2019; Gupta et al. 2019; Niazi et al. 2019). We will highlight selected key studies, with a focus on PCa diagnostics. Much of the recent research has focused on pursuing performance comparable to, or even exceeding, that of medical experts in routine diagnostic tasks, with the aims of streamlining the pathology workflow by partial automation and improving patient safety. One of the first studies to convincingly demonstrate expert-level performance using DL was focused on skin cancer (Esteva et al. 2017). In that study, an Inception V3 CNN (Szegedy et al. 2016) was trained using close to 130 000 clinical images, and performance in classifying them to benign and malignant skin lesions was found comparable to 21 experts.

The popularity of the field has also led to the organization of challenges, where competitors try to find optimal ways of solving a specified task (Hartman et al. 2020). The contributed solutions are evaluated using the same criteria and data, allowing direct comparison of different approaches, which can be highly beneficial for the development of the field. One of the best-known challenges, CAMELYON16, aimed at the detection of metastatic BCa in lymph nodes and was based on a dataset of 399 HE stained WSIs. Approaches based on DL clearly outperformed leading feature-based contributions and represent another landmark result where DNNs were shown to achieve expert-level diagnostic performance (Bejnordi et al. 2017). The winner of CAMELYON16 achieved an AUC of 0.994 in classifying WSIs into benign or metastasis-containing, and the top 5 algorithms were all comparable to an expert analysing the samples without time constraints. In an experiment featuring time constraints, several top algorithms outperformed a panel of 11 pathologists. The associated large public dataset, extended for CAMELYON17 (Bándi et al. 2018;

Litjens et al. 2018), has also contributed to a number of other studies.

The availability of increased amounts of training data has been a major factor in these successes. For solutions aimed at PCa diagnostics, lack of data has been a limitation until recently. For example, the early study on PCa detection by Jafari-Khouzani (Jafari-Khouzani et al. 2003) relied on only 100 ROI and even the more recent attempts by Wang *et al.* (D. Wang et al. 2015) and Niazi *et al.* (Niazi et al. 2016) were based on only 300 and 131 ROI, respectively. Training classifiers to correctly capture the heterogeneous tissue morphologies encountered in PCa samples is probably impossible based on such limited data. Källén *et al.* tried to address this problem by transfer learning (Källén et al. 2016). They extracted the activations of several layers of a CNN pre-trained on photographic images and fed these as features to RF and SVM classifiers to train them to predict the Gleason grade of input patches. An accuracy of 81.1% was reported based on 10-fold CV on a small, but balanced, dataset containing approximately 50 ROI representing each class (benign and Gleason grades 3-5). In a follow-up study, a CNN was instead trained from scratch for the same task using the same dataset, with the help of data augmentation (Gummeson et al. 2017). In this case, the accuracy improved to 92.7% based on 4-fold CV.

Arvaniti *et al.* approached Gleason grading by training their system on 641 tissue microarray (TMA) spots, and evaluating on a test set of 245 spots (Arvaniti et al. 2018). Each spot represents a relatively small tissue region (3100 x 3100 pixels at 0.23 μm per pixel) obtained from a single patient following surgery. The spots were processed patch-wise to train a MobileNet CNN (Howard et al. 2017), resulting in a Cohen's quadratic kappa of 0.55 relative to the same pathologist who had graded the training data. Compared to a second pathologist, the kappa was 0.49, and the pathologist-to-pathologist value was 0.67. When considering GS on the TMA spot level, the system reached kappa values of 0.75 and 0.71 for the two pathologists, comparable to the inter-pathologist agreement of 0.71. Li *et al.* (J. Li et al. 2018) chose a semi-supervised semantic segmentation approach, where they trained a U-Net (Ronneberger et al. 2015) to classify pixels as benign, low grade or high grade cancer using 135 patches from RP samples, annotated in pixel-wise manner. They then applied the trained model on a larger dataset of 1800 samples having only image-level labels and utilized them as additional training data through an expectation-maximization process. On a test set of 289 patches, the system reached 74.79% pixel-wise accuracy and a mean intersection over union of 49.47% over all three classes.

In contrast to only using patches or TMA spots, processing entire slides provides more tissue for training and evaluation and is also more relevant in view of clinical use. Moreover, in the case of PCa, biopsies represent the majority of all diagnostic samples, and their WSI-based analysis is thus of considerable clinical interest. With this in mind, 225 pixel-wise annotated WSIs of prostate biopsies allowed training a CNN, which achieved an impressive AUC of 0.99 in detecting slides with cancer (Litjens et al. 2016). The slide-level prediction was obtained as the median probability across all pixels in a WSI. While this study marks an improvement over most of the earlier research, the data contained very few examples of high grade PCa (e.g. not a single sample with GS 5+5). This not only prevented training a system to perform grading, but also complicates evaluating the cancer detection performance, since it may vary between low grade and high grade cases.

Campanella *et al.* based their study on an unprecedented amount of 24,859 WSI of prostate biopsies, 9,962 WSI of skin samples, and 9,894 WSI with lymph node samples (Campanella et al. 2019). They relied on the multiple instance learning paradigm, which enabled patch-wise training of ResNet34 models in weakly supervised manner using only slide-level labels extracted from pathology reports. The underlying assumption is that in negative slides, all patches must be negative, and in positive slides, at least one patch has to be positive. The process then involves repeated predictions on the training data, followed by picking the top-ranking patch for each slide in terms of estimated cancer probability, and optimization of the loss function on these top-ranking patches. To perform slide-level classification, the patch-wise probabilities from each slide were summarized into pre-specified features and used to train a RF classifier. This approach is virtually identical to that of Publication **III** and closely related to the one used in Publication **IV**. The authors also evaluated using the DNN as a feature extractor, and training a recurrent neural network (RNN) using the DNN's representations of the patches as input sequences. This is similar to the approach proposed for outcome prediction in colorectal cancer (Bychkov et al. 2018). The RNN slightly outperformed the RF method and resulted in AUC values of 0.991, 0.988 and 0.966 for prostate (n=1,784), skin (n=1,575) and breast cancer (n=1,473) detection, respectively. Notably, when repeating the analysis on prostate biopsies scanned with a different scanner or prepared in different laboratories, relatively large decreases of 3% points and 6% points in AUC were observed.

Perhaps due to limited data and complications in multiple instance learning for

a multi-class problem, neither Litjens *et al.* nor Campanella *et al.* considered grading. The first study to present automated Gleason grading on large-scale WSI data used 1226 WSI of RP samples for training and 331 WSI for evaluation (Nagpal et al. 2019). An ensemble of Inception V3 CNNs was trained in patch-wise manner, employing hard-negative mining to improve sensitivity. The patch-level predictions of the CNNs were summarized into tissue fractions per Gleason grade for each WSI, and fed to a kNN classifier to perform slide-level classification. Relying on a second classifier to aggregate patch-level predictions into slide-level outputs resembles Publications **III-IV** and the study by Campanella *et al*. Compared to the grading by a genitourinary pathologist, the system achieved an accuracy of 0.70, higher than the mean accuracy of 0.61 among a panel of 29 non-specialist pathologists. The system outperformed 8 of the 10 pathologists who graded the entire test set.

Publication **IV** marks the first time a clinically representative WSI dataset was used to demonstrate automated expert-level Gleason grading of biopsies, as opposed to RP samples. Another study, independently performed at the same time, reported comparable results using 933 WSI with 4712 biopsies as training data (Bulten et al. 2020). First, malignant tissue was detected using a DNN (Litjens et al. 2016), followed by detection of epithelial tissue (Bulten et al. 2019) and assignment of pixel-wise grade labels for the malignant epithelium based on pathology reports. A U-Net was then trained in patch-wise manner based on the labels. Biopsy-level Gleason grade groups were obtained based on the percentages of tissue pixels classified into each grade. The system achieved an AUC of 0.990 in detection of biopsies with cancer, and a quadratic Cohen's kappa of 0.918 relative to the consensus grading of three urological pathologists on a test set of 550 biopsies. On a subset (n=100) of samples graded by 15 pathologists, the system outperformed 10 pathologists in terms of kappa measured against the grading by the specialists. On an external test set comprising the same 245 TMA spots used by Arvaniti *et al.*, the system reached kappa of 0.723 and 0.707 relative to two different pathologists. Despite slightly different methodology, the results of Bulten *et al.* are similar to those of Publication **IV**, although the kappa values are not directly comparable due to different weighting. In contrast to Publication **IV**, Bulten and colleagues did not explicitly consider estimation of tumor burden in the form of cancer length in each biopsy, even though that would most likely be feasible based on the pixel-wise output of their U-Net model.

2.4.5 Deep learning for quality control

In view of quality control, DL has been employed to detect out-of-focus tissue. The work of Senaras *et al.* (Senaras et al. 2018) relied on training data collected by physically perturbing the focus level while scanning slides, creating images with varying levels of out-of-focus blur. A DNN was then trained to perform binary classification between sharp and blurry image patches. In contrast, Kohlberger and colleagues (Kohlberger et al. 2019) asked pathologists to delineate regions in sharp focus based on visual evaluation, and then created simulated out-of-focus examples by applying low-pass filtering to the sharp patches. A DNN based multi-class classifier was then trained on the simulated examples to estimate the degree of blur on a 30-step scale. The main advantage of DNN based focus quality assessment compared to the classical approaches relying e.g. on a single gradient-based descriptor is that if the training data sufficiently cover different tissue morphologies, one can reliably estimate the degree of focus independently of the image content. The challenge with classical approaches is that different tissue types naturally exhibit differing amounts of sharp edges, which are typically used as an indirect measure of good focus.

The important topic of color variation has also been approached using DNNs. Recent studies have proposed using sparse autoencoders and generative adversarial networks (GAN) (Goodfellow et al. 2014) for standardizing histological staining across different WSI datasets (BenTaieb et al. 2017; Janowczyk et al. 2017). By using cycle-consistent GANs (Zhu et al. 2017), it is even possible to perform style transfer between unpaired images (Shaban et al. 2019). This allows one to map WSI data collected from a new site to a specified color model, harmonizing staining variation between different scanners or laboratories. Classical color standardization methods and data augmentation have also been studied in the context of DNNs (Tellez, Litjens, Bandi et al. 2019). As a potential sign of the field maturing towards clinical adoption, some studies have started to shed light on other aspects that may influence performance in a real-world setting but have been largely overlooked thus far, such as image compression (Y. Chen et al. 2020; Zanjani et al. 2019). Encouragingly, Zanjani *et al.* reported that their DNN-based system developed for the CAMELYON17 challenge was tolerant to relatively high JPEG2000 compression ratios.

2.5 3D reconstruction in computational pathology

2.5.1 From 2D to 3D histology

Histological samples are normally examined in 2D using thin sections cut from the tissue (Roberts et al. 2012). However, studying tissue in 3D holds potential for obtaining a more comprehensive view of normal and pathological processes where the spatial arrangement of different tissue structures or cell types is of relevance. A conceptually straightforward approach for achieving this is obtaining consecutive 2D slices of the 3D object via serial sectioning (Pichat et al. 2018). The difficulty of then mentally recreating the 3D structure by only examining the 2D sections was acknowledged much before the time of computer graphics and was initially tackled using wax models (Born 1883). More than a century later, 3D histology has been enabled by digital imaging of the 2D sections, followed by computational 3D reconstruction (Fig. 2.3). Typically, the reconstruction process consists in a series of image registration operations, where each pair of consecutive sections in the image stack are aligned to obtain a series of transformations (Magee et al. 2015). Since each transformation relates the previous section to its neighbor, all of the sections can be brought into a common coordinate system by serially applying a composite transformation for each section. As the end result, a coherent 3D volume is produced.

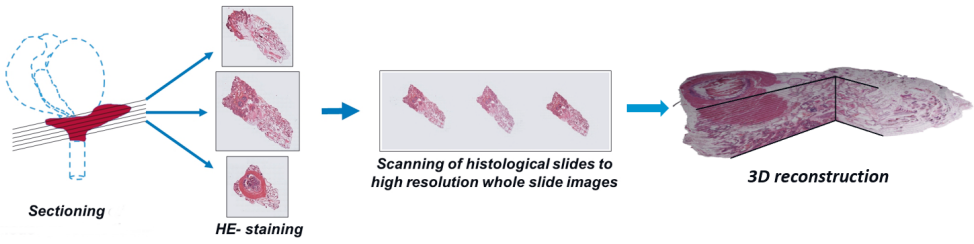

Figure 2.3 The principle of 3D histology. A sample is sectioned serially, the sections are stained and scanned, followed by computational 3D reconstruction.[1]

Direct 3D imaging of tissue can be accomplished using techniques such as MRI or computed tomography (Roberts et al. 2012). However, compared to these methods, using light microscopy has the advantage of achieving both subcellular resolution

[1]Figure panels courtesy of Pekka Ruusuvuori, Leena Latonen and Kaisa Liimatainen, modified from Ruusuvuori et al. 2016.

and large sample sizes at the same time. This holds true especially if the imaging of the serial sections is performed using WSI scanners. This allows e.g. the construction of high-resolution atlases of entire organs (Amunts et al. 2013; Johnson et al. 2010; Lein et al. 2007). Moreover, since standard histological stains can be used, pathologists can in principle rely on conventional interpretation techniques when visually examining the 3D reconstructions. Another advantage is that since sample preparation is identical to conventional histology, biochemical techniques such as IHC or *in situ* hybridization are fully compatible with 3D histology. This would even allow integration of spatially resolved genomic, transcriptomic or proteomic data within the 3D model (Koos et al. 2015; Mignardi et al. 2016; Ståhl et al. 2016). Such cell atlases are expected to provide new insights in e.g. cancer research (Ledford 2017).

The core methodological problem in 3D histology is how to accurately align a sequence of 2D images to reconstruct a 3D volume without introducing distortions (Pichat et al. 2018). The images cannot be merely stacked in the correct order, since random rotations and translational offsets are introduced during sample preparation and scanning. While conceptually simple, due to technical and anatomical variation from image to image the reconstruction process is generally a difficult task. The problem is further complicated by tissue deformations introduced during embedding and sectioning (Gibson, Gaed, Gómez et al. 2013b). Various algorithms have been proposed to address these issues (Pichat et al. 2018), and they will be reviewed in Sections 2.5.2 and 2.5.3. Evaluating if a 3D reconstruction is accurate is a non-trivial problem in itself and will be discussed in Section 2.5.4.

2.5.2 Image registration

Virtually all 3D reconstruction methods are based on image registration (Pichat et al. 2018), which has been traditionally divided into two main categories: area-based and feature-based (Zitova et al. 2003). In area-based image registration, detection and matching of features are combined into a single process. These methods do not attempt to detect particular salient structures, such as edges of objects, but rather try to directly match the images. For example, in a block-matching scheme, patches from the two images are compared and corresponding pairs are detected based on a similarity metric (e.g. correlation between pixel values). In the spectral approach, properties of the Fourier transform are utilized to estimate the translation, rotation

and scaling differences between the images. A third widely used approach is using numerical optimization to find transformation parameters that maximize some measure of similarity between the two images. The result of the matching is an estimated transformation relating coordinates in one image to those in the other image.

In feature-based image registration, feature detection and matching are separated into two steps (Zitova et al. 2003). First, features usually representing some salient structures are found using e.g. edge detectors. Scale-invariant feature transform (SIFT) (Lowe 2004) and Speeded up robust features (SURF) (Bay et al. 2008) are two popular types of features that are invariant to uniform scaling, rotation and even partially to affine distortion and changes in illumination. A number of 'keypoints' are then detected from each image, and corresponding keypoint pairs are found using a matching algorithm operating in the feature space. Based on the corresponding point pairs, a transformation relating the two images can then be estimated.

A key algorithmic choice common to all image registration methods is the type of transformation to use for modeling the image-to-image correspondences (Sotiras et al. 2013). For 3D histology, typically at least a rigid transformation is needed to accommodate for translation and rotation between tissue sections, and an affine transformation can further compensate for shrinking, swelling and shearing of the tissue by allowing scaling of the image. As opposed to global transformations that apply similarly to the entire image, so called elastic transformations can compensate for locally varying tissue deformation in different parts of the image.

It is worth noting, that new ML-based image registration methods, where the transformation parameters are in some cases even learned directly from the raw images, are blurring the line between the traditional classification into area- and feature-based algorithms (Haskins et al. 2020). Several reviews of medical image registration, which has been an active research topic for decades, provide a more comprehensive view of the field (Maintz et al. 1998; Oliveira et al. 2014; Viergever et al. 2016).

2.5.3 Three-dimensional reconstruction

The first computational 3D histology reconstructions were attempted already before the 1990s (Merickel 1988; Salisbury et al. 1993). Early methods often relied on partial interaction. For example, Kay *et al.* reconstructed microvessels of RP specimens by first segmenting edges and performing serial registration using a surface matching

algorithm (Kay et al. 1998). Misaligned sections were then manually adjusted, and reconstruction quality was assessed visually. Besides the need to compensate for tissue deformations, a common issue in 3D histology is the unintended straightening of curved structures, sometimes referred to as the banana-to-cylinder problem (Malandain et al. 2004) or the shear effect (Hughes et al. 2012). This happens if patterns visible in several consecutive cross-sections of a curved structure are forced into co-alignment by pairwise registration. Moreover, sections having e.g. missing tissue can lead to catastrophic failure of the entire reconstruction if one only relies on pairwise serial registration without considering the global 3D structure (Pichat et al. 2018). Since each composite transformation is formed by concatenating all of the preceding transformations, even a single erroneously estimated transformation can introduce significant distortions into the 3D reconstruction. Much of the algorithmic work to date has focused on designing automated means of mitigating these problems.

Algorithmic features since included in many other solutions, namely block-wise matching and multi-resolution processing, were proposed in the fully automated method of Ourselin *et al.* in application to brain tissue (Ourselin et al. 2001). They divided the images into blocks and searched for corresponding block pairs based on the correlation coefficient between pixel values. They then robustly estimated a global rigid transformation for each image pair based on the set of displacement vectors represented by the block-wise matches. Even if some parts of an image contain e.g. torn off tissue, the algorithm can tolerate such artefacts as long as a sufficiently large fraction of the blocks can be matched. The multi-resolution scheme consists in first performing registration at a coarse scale with low resolution images, and then proceeding to progressively finer scales. This not only speeds up computation, but can also increase the capture range of the algorithm, that is, the magnitude of translational and rotational offsets between a pair of images that can still be compensated for. Moreover, the initial low-pass filtering of the coarse levels may reduce the risk of the registration process converging prematurely to local minima of the loss function (Sorzano et al. 2005). The same reasoning can be extended to the type of transformations performed, starting from a global rigid transformation, followed by more refined registration using elastic transformations (Braumann et al. 2005), based e.g. on B-splines (Arganda-Carreras et al. 2006; Sorzano et al. 2005).

One aspect to consider is the choice of the so called reference section (Pichat et al. 2018). One of the images in the series is typically fixed, and all the other images are

brought into alignment relative to this reference section. Since registration errors accumulate with every pairwise step, it is beneficial to select the reference section from the middle of the stack and proceed serially towards both ends of the stack (Magee et al. 2008). This minimizes the number of transformations that need to be concatenated. In view of computational efficiency, this choice also allows running the two sequential processes in parallel, as they are independent of each other. Selecting the best reference section automatically, considering image quality, has also been proposed (Bagci et al. 2010). A special case of interest is multi-stain 3D reconstruction, where multiple histological stains, e.g. HE alternating with IHC, have been applied to the serial sections (Song et al. 2013). The question then is, whether to first reconstruct each stain separately, followed by aligning the two 3D stacks, or to e.g. use each HE-stained section as a reference for the adjacent IHC section, followed by serial pairwise registration of the HE-IHC pairs to obtain the 3D reconstruction.

Even if the robustness of the individual registration steps can be improved using the algorithmic features above, the quality of the overall reconstruction is still to a large extent dictated by the poorest pairwise result (Pichat et al. 2018). Pairwise methods are also prone to accumulation of errors. One way of mitigating the issue is to still rely on pairwise registration operations, but refine or regularize them to introduce some dependency across individual image pairs. For example, the transformation matrices obtained during the registration steps can be subjected to Gaussian filtering across the series, decreasing the effect of outlier pairs on the entire reconstructed volume (Ju et al. 2006). Another solution also relies on serial pairwise registration, but the process is performed several times according to a Gauss-Seidel iteration scheme (Gaffling et al. 2014). This has the effect of separating smoothly varying true anatomical changes throughout the stack from random, 'high-frequency' errors. Using a reconstruction smoothness metric that is directly incorporated into the loss function of the optimization process has also been proposed (Cifor et al. 2011).

Some algorithms consider the entire stack simultaneously or, as the method by Saalfeld *et al.*, align each section to several neighbors (Saalfeld et al. 2012). An initial alignment step relying on SIFT, followed by block-matching, is used to initialize a spring mesh system representing the entire volume. An optimization process is then performed to find a solution representing a compromise between exact matching of neighboring sections and overall coherence of the volume. The role of the spring model is thus to regularize the solution. One can also utilize an external 3D

reference, such as block-face photographs acquired during serial sectioning, to help estimate a reconstruction that is true to the original shape (Feuerstein et al. 2011). Casero *et al.* captured block-face images with a customized polarized light illumination system for pre-aligning the stack of histology images (Casero et al. 2017). They developed a reconstruction algorithm, conceptually related to that of Gaffling *et al.*, that refines the estimated transformations iteratively using a process called transformation diffusion, which is formulated following a physical analogue of heat diffusion. The algorithm is trivially parallelizable and produced reconstructions of heart tissue that are both smooth and closely correspond to the 3D shape of the block-face reference. The approach of Xu *et al.* is unique in that it utilizes natural histological landmarks - nuclei (Xu et al. 2015). The rationale of the method is that a fraction of the nuclei present on the sections get halved by the microtome blade such that the two halves are visible on two adjacent sections. By detecting these nuclei and using them as landmarks for the registration, the banana-to-cylinder effect can be avoided.

Numerous questions and tissue types have been studied using 3D histology, including e.g. analysis of tumor growth patterns of metastatic BCa in lymph nodes (Paish et al. 2009), ductal carcinoma in situ (Booth et al. 2015; Norton et al. 2012), PCa in RP samples (Hovens et al. 2017; Rojas et al. 2015; Tolkach et al. 2018) and adenocarcinoma of the lung (Onozato et al. 2012). Other examples include the study of vasculature in different tissues (Brown et al. 2015; Fónyad et al. 2015; Grothausmann et al. 2017; Liang et al. 2015). Studying histology in 3D has also shed light on the organization and functioning of stem cells responsible for the normal epithelial homeostasis in the human prostate (Moad et al. 2017). Several studies have also sought to integrate 3D histological data with other modalities such as MRI (Johnson et al. 2010; Reynolds et al. 2015; Stille et al. 2013). Nevertheless, 3D histology is still rarely used in clinical applications or mainstream biomedical research, and mainly remains a technology utilized in proof-of-concept studies. However, new imaging and tissue processing techniques may relieve some of the current bottlenecks of 3D histology in the future (Farahani et al. 2017).

2.5.4 Evaluation of reconstruction quality

As reconstructions are often performed in the absence of any ground truth on the true 3D structure, assessing their quality is challenging. Most of the proposed met-

rics measure the accuracy of the pairwise image registration operations as a substitute instead. A direct measure of image registration accuracy can be obtained by computing target registration error (TRE) (Fitzpatrick et al. 1998). TRE is simply the Euclidean distance between the locations of the same, known target in two images. Calculating TRE requires that the target points can be somehow detected or annotated. For annotating a large number of image pairs, the amount of manual labor can be substantial. However, TRE represents the only direct measure of registration accuracy and is in many cases more reliable than indirect metrics (Rohlfing 2011).

Indirect metrics of registration accuracy compare the values of corresponding pixels in the two images after registration (Rohlfing 2011). Widely used pixel-wise similarity metrics include root mean squared error (RMSE) or mean squared error (MSE), normalized cross correlation (NCC), mutual information (MI) and normalized mutual information (NMI). The basis for these metrics is the assumption that correctly aligned pixels should be similar to each other and the difference between the metrics listed above is the definition of similarity. An alternative is to perform segmentation of some objects of interest and compare the segmented regions in the two images in terms of the Jaccard index, which quantifies the relative overlap of the two sets of pixels. In 3D histology, different images represent tissue from slightly different locations in the sample, and perfect overlap or matching pixel values are thus not a likely consequence of successful reconstruction. On the other hand, these metrics have the advantage that they can be computed without any annotations or ground truth information. Moreover, TRE is usually computed only using a small set of landmarks while the pixel-wise metrics can be evaluated at all pixels.

There are also metrics proposed specifically for 3D reconstructions, considering not only the pairwise errors but also the global 3D shape. One direct measure of accumulated errors is a variation of TRE, where the error is not computed as the Euclidean distance between corresponding points on adjacent tissue sections but instead between the positions of each point in the reconstruction under evaluation and a reference reconstruction (Xu et al. 2015). We refer to this metric as the accumulated target registration error (ATRE). While ATRE allows directly quantifying the distortion of the reconstructed volume, it is dependent on the availability of a reference reconstruction. In the work of Xu *et al.*, a reference in principle free of accumulated distortions was obtained using bisected nuclei as landmarks. However, as in the case of TRE, the amount of work required to manually annotate or curate automatically

detected bisected nuclei can be a limiting factor for applying this metric.

In addition to pairwise registration, indirect metrics have also been proposed specifically for 3D reconstructions (Cifor et al. 2011; Gaffling et al. 2014). They are based on the assumption that a correct reconstruction should exhibit pixel values slowly changing from section to section. This requires that the distance between adjacent sections is shorter than the dimensions of anatomical structures within the tissue. If the sections are sampled sparsely, abrupt changes in pixel values due to anatomical differences between adjacent sections are likely. If the assumption holds, quantifying the smoothness of the intensity profile along the stack of sections can function as a surrogate measure for the coherency of the 3D volume. The proposed metrics rely on GLCM-based contrast or correlation descriptors computed along the direction across the slices. An accurate reconstruction featuring a smoothly varying intensity profile should then exhibit low contrast and high correlation.

Approaches based on physical fiducial markers have also been proposed. For example, catheters perfused with a mixture of cuttlefish ink and flour were inserted into fixed tissue prior to embedding in paraffin, resulting in fiducial markers that are visible both in blockface images acquired during sectioning and the resulting WSI (Shojaii et al. 2011). Gibson *et al.* used strand-shaped fiducials, visible in both MRI and histology, inserted into RP specimens (Gibson, Gaed, Gómez et al. 2013a), whereas Hughes *et al.* designed a device consisting of three needles fixed at known angles relative to each other, allowing the generation of a known pattern of holes through the sample (Hughes et al. 2012). In addition to numerical measures, it is also commonplace to visualize cross-sectional views of the reconstructed volume in order to visually evaluate quality (Cifor et al. 2011; Gaffling et al. 2014; Ju et al. 2006; Magee et al. 2015; Saalfeld et al. 2012; Song et al. 2013).

Many 3D reconstruction algorithms have been proposed, but very few studies have attempted any objective comparisons. Evaluations of 2D histological image registration methods (Borovec et al. 2020) can provide some information that is of relevance, but due to specific aspects such as the need to minimize accumulation of errors during the serial registration process, findings from these studies cannot be fully extrapolated to 3D histology. To the best of my knowledge, the previous study comparing 3D reconstruction algorithms was published over 10 years ago (Beare et al. 2008). In that study, five algorithms were compared for reconstructing mouse brains from serial sections. The algorithms included semi-automatic methods based

on fiducial markers, as well as automatic registration methods based on maximizing pixel-wise similarity or tissue section overlap. Feature-based registration using SIFT was also evaluated. By assuming that two holes drilled into the sample were straight, the residual errors between the locations of the holes relative to a linear fit could be used as a measure of reconstruction accuracy. This method served as an inspiration for Publication **II**, where a sample was pierced with an industrial laser prior to sectioning. The methods evaluated by Beare *et al.* can be considered as relatively simple baseline algorithms, and most of the more advanced algorithms aimed at 3D histology were not published yet at the time. Moreover, hyperparameter tuning was not considered. These limitations were addressed in Publications **I-II**.

3 AIMS OF THE STUDY

The main aims of the work were:

1. Develop machine learning techniques for breast cancer diagnostics.
2. Develop machine learning techniques for prostate cancer diagnostics.
3. Evaluate and compare 3D histology reconstruction algorithms.
4. Optimize computational approaches for efficient processing of WSI data.

4 MATERIALS AND METHODS

4.1 Data collection

4.1.1 Murine samples

Murine prostate and liver samples were prepared as described in Publication **I-II** at Tampere University, Finland, according to permits by Etelä-Suomen aluehallintovirasto (ESAVI/6271/04.10.03/2011, ESAVI/5147/04.10.07/2015). The liver was processed with an industrial laser after fixation to introduce four holes functioning as artificial landmarks. The tissue blocks were serially sectioned and HE stained. The slides were scanned using a system based on a Zeiss Axioskop40 microscope (Carl Zeiss Microimaging, NY, USA).

From each pair of consecutive images of the prostate sample, four pairs of corresponding landmark points were manually selected at the centers of nuclei bisected by the sectioning blade. The annotation was performed by one observer for Publication **I** and repeated independently by another observer for Publication **II**. For each image of the liver sample, the two observers independently marked the locations of the four artificial landmarks introduced with the laser. A total of 2448 landmarks were annotated by the two observers across the two samples.

4.1.2 Lymph node samples

The samples for Publication **III** were obtained by participation in CAMELYON16 and collected as described by the challenge organizers (Bejnordi et al. 2017). The dataset contained sentinel axillary lymph nodes retrospectively sampled from 399 BCa patients at Radboud University Medical Center, Nijmegen and University Medical Center Utrecht, The Netherlands, with one HE stained tissue section per WSI.

Among the 170 WSI from Nijmegen, 60 out of 70 WSI containing metastases were fully annotated and 10 were partially annotated. Among the 100 WSI from Utrecht, 37 out of 40 metastatic WSI were fully annotated and 3 were partially annotated. The annotations were provided in the form of binary masks indicating the pixels labeled as representing metastases by pathologists.

4.1.3 Prostate biopsies

Diagnostic prostate biopsies used in Publication **IV** were mainly collected in the prospective, population-based, screening-by-invitation trial STHLM3 conducted in Stockholm, Sweden (ISRCTN84445406) (Grönberg et al. 2015). The study protocol was approved by Stockholm regional ethics committee (2012/572-31/1, 2012/438-31/3 and 2018/845-32). For Publication **IV**, 8313 biopsy cores from 1222 randomly selected participants, stratified by ISUP grade, were digitized. We also obtained 271 cores from 93 men with high-grade disease from Capio St. Göran Hospital, Stockholm as additional training data. As an additional test set, we collected 87 cores from Pathology Imagebase, a reference collection launched by ISUP (Egevad et al. 2018). Furthermore, external validation data comprising 330 biopsies from 73 men was obtained from Karolinska University Hospital, Stockholm. All samples were formalin fixed and HE stained. The biopsies from STHLM3, Capio St. Göran Hospital and Imagebase were scanned using a Hamamatsu C9600-12 (Hamamatsu Photonics, Hamamatsu, Japan) at Karolinska Institute, Stockholm and an Aperio Scanscope AT2 (Leica Biosystems, Wetzlar, Germany) at SciLifeLab, Uppsala, Sweden. The external validation data were scanned using a Hamamatsu S360 C13220-01 at Karolinska University Hospital.

A single urological pathologist (Prof. Lars Egevad), blinded to the clinical characteristics of the patients, examined the slides using a microscope and graded the biopsies following the ISUP grading system (see Section 2.2). He also measured the linear cancer length for each biopsy core, and reported a summary ISUP grade and length for each patient. Additionally, he drew a line next to each cancerous area with a marker pen. Biopsies in the Pathology Imagebase were graded by 23 urological pathologists, including Prof. Egevad, independently of each other following the ISUP grading system. The assessment was performed based on digital images of the relevant tissue regions captured with a microscope.

4.2 Image processing

4.2.1 Tissue segmentation

Tissue segmentation was performed to exclude the background pixels in each WSI from further analysis, first manually (Publication **I**) and later automatically using an algorithm based on applying Otsu's thresholding (Otsu 1979) to the saturation component of the HSV transformed image (Publications **II-III**). An improved segmentation algorithm, based on Laplacian filtering to detect tissue based on local variance, followed by Otsu's thresholding of the filter response, was developed for Publication **IV**. By excluding regions lacking local variation, out-of-focus tissue was also removed from further analysis. Post-processing based on mathematical morphology and exclusion of pen markings and other artefacts using color-based rules was applied to refine the segmentation result. The results were stored in the form of binary images indicating tissue pixels, which we refer to as tissue masks.

4.2.2 Label extraction

Since most of the 3D reconstruction tools in Publication **II** do not allow applying the estimated transformations on numerical coordinates, landmarks had to be represented as images for the purpose of evaluating TRE and ATRE. This was accomplished by generating images containing disks of different colors at the landmark locations. The unique color of each disk allowed identification of each landmark after applying the transformations. The image processing workflow used in Publication **II** is depicted in Fig. 4.1. In Publication **I**, landmark images were not needed as the transformations could be applied directly to numerical coordinates.

To train ML models in supervised manner, labels for the tissue pixels in each WSI had to be extracted. For Publication **III**, labels were available as binary masks indicating metastatic tissue. For Publication **IV**, in addition to the labels reported on the biopsy core level, the slides had pen markings adjacent to malignant tissue regions (see Section 4.1.3). To convert these physical annotations into digital pixel-wise form, the markings were segmented using an algorithm relying on Otsu's thresholding and color-based rules. Each segmented marking was then mapped to the adjacent segmented tissue. The mapping algorithm involved constructing vectors defined by

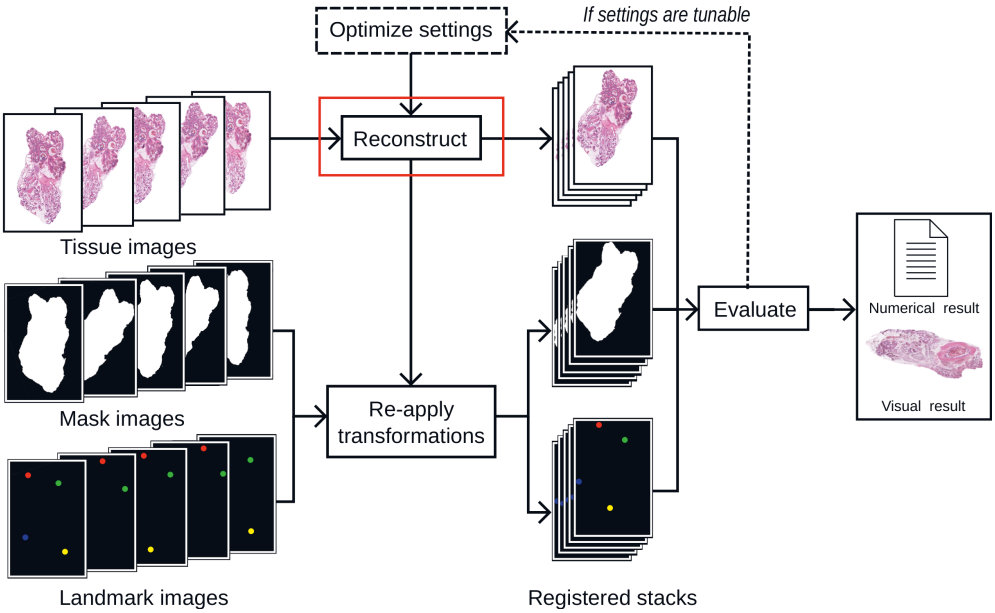

Figure 4.1 Evaluation framework for 3D reconstructions. Tissue images are input to the evaluated reconstruction algorithm and a stack of co-registered images is produced as output. The estimated transformations are re-applied to a series of mask images indicating the tissue regions and images defining landmark locations. The registered stacks of tissue, mask and landmark images are used to numerically and visually evaluate reconstruction quality. If the algorithm has tunable settings, they can be optimized iteratively. Reprinted from Publication **II**.

pairs of nearest neighbors between the marking and tissue region boundaries. Based on these vectors, sets of tissue boundary pixels were picked, and the areas enclosed by these pixels were assigned the label 'cancer'. Other tissue pixels were then assigned the label 'benign', and the results were stored as label masks.

4.2.3 Patch extraction

In Publications **I-II**, WSIs at different downsampling factors were stored in TIFF format and directly used as input for the evaluated reconstruction tools. In Publications **III-IV**, the images were processed patch-wise, allowing thousands of training samples to be extracted from each WSI. As a side note, alternatives to the popular patch-wise approach have also been proposed, such as neural compression (Tellez, Litjens, van der Laak et al. 2019), where lower-dimensional representations are constructed to allow treating an entire WSI as a single training sample.

In Publication **III**, the segmented tissue regions were first split into subimages (8192 x 8192 pixels) to reduce the memory usage and allow parallelization of the following steps. Histogram matching (Gonzales et al. 2002) relative to a reference image was performed for each subimage to reduce color variation, followed by color deconvolution to separate the HE stain components (Ruifrok et al. 2001). Nuclei were then segmented from the H channel using adaptive thresholding (Bradley et al. 2007). In the training phase, the final patches of 200 x 200 pixels at full resolution (approx. 49 x 49 μm) were randomly sampled from the subimages, with labels assigned according to the label masks. In the prediction phase, all of the subimages were fully tiled into patches. The overall system design is presented in Fig. 4.2.

In Publication **IV**, all tissue regions were directly split into patches. After evaluating several patch sizes and resolutions, we opted for 598 x 598 pixels (approx. 540 x 540 μm) and 50% overlap between neighboring patches. Patches were resampled using Lanczos interpolation to harmonize the pixel sizes of different scanners, and background pixels were masked based on the tissue masks to remove any pen markings from the patches. The pre-processing system is depicted in Fig. 4.3.

4.3 Feature-based learning

4.3.1 Feature extraction

In Publication **III**, features extracted from the H and E channels of each patch included 104 texture features, e.g. GLCM descriptors (Haralick et al. 1973), LBP (Ojala et al. 2002; Pietikäinen et al. 2000), histograms of oriented gradients (Dalal et al. 2005) and maximally stable extremal regions (Matas et al. 2004). Additionally, a number of features quantifying e.g. inter-nuclei distances and the number of nuclei in local neigborhoods were computed. The over 200 features extracted from each patch represent generic texture descriptors rather than application-specific tailored features. The nucleus-based features are specific to histological images but not tailored to a specific tissue type or task. This approach is thus more similar to DL or the more recent feature-based systems (Yu et al. 2016) than the ones relying on features carefully tailored to a specific task (Niazi et al. 2016).

4.3.2 Patch classification

Different classifier models (NN, SVM, logistic regression and RF) were evaluated using 10-fold CV for discriminating between benign and metastasis-containing patches based on the extracted features, and RF (Breiman 2001) was chosen in Publication **III** based on this experiment. The RF algorithm consists in training a set of decision trees to obtain an ensemble classifier via bootstrap aggregation, or bagging, by sampling the patches used for training each decision tree randomly with replacement (Hastie et al. 2009). The overall decision is then obtained based on voting among the trees, which decreases the variance of the overall model while having the same bias as the constituent classifiers. De-correlating the individual decision trees by training them on different samples of the data is important, since averaging highly-correlated classifiers would provide little improvement. A random subset of features is selected for consideration at each split when growing the decision trees to further reduce correlation between individual trees.

In Publication **III**, RF models consisting of 50 trees were built using the TreeBagger implementation in MATLAB (The MathWorks Inc., Natick, MA, USA). In the prediction phase, when a patch arrives at a leaf node of a decision tree, the number of training patches labeled with the corresponding class relative to the total number of training patches in that leaf node is computed. This fraction represents the probability associated with the decision for a given tree. After probabilities have been obtained from all the other trees in the RF model in similar manner, their mean is taken to represent the overall probability of malignancy for a patch.

4.3.3 Slide classification

The predicted patch-wise probabilities were collected into images we refer to as confidence maps, representing the estimated likelihood of malignant tissue being present across the input WSI. Confidence maps allow visual examination of high-probability regions, which might be sufficient as a diagnostic aid, but they can also be further post-processed to obtain WSI-level predictions. In Publication **III**, this was accomplished by generating confidence maps for the training slides, and re-applying the feature extraction and RF training procedure using the confidence maps as input instead of the original WSI. In this case, only the 104 texture features were extracted.

As a result, an RF model predicting whether a WSI contains metastatic tissue or not was obtained. Such a two step approach relying on a second ML stage to integrate patch-level predictions into slide-level predictions has become relatively common in the field (Bychkov et al. 2018; Campanella et al. 2019; Nagpal et al. 2019).

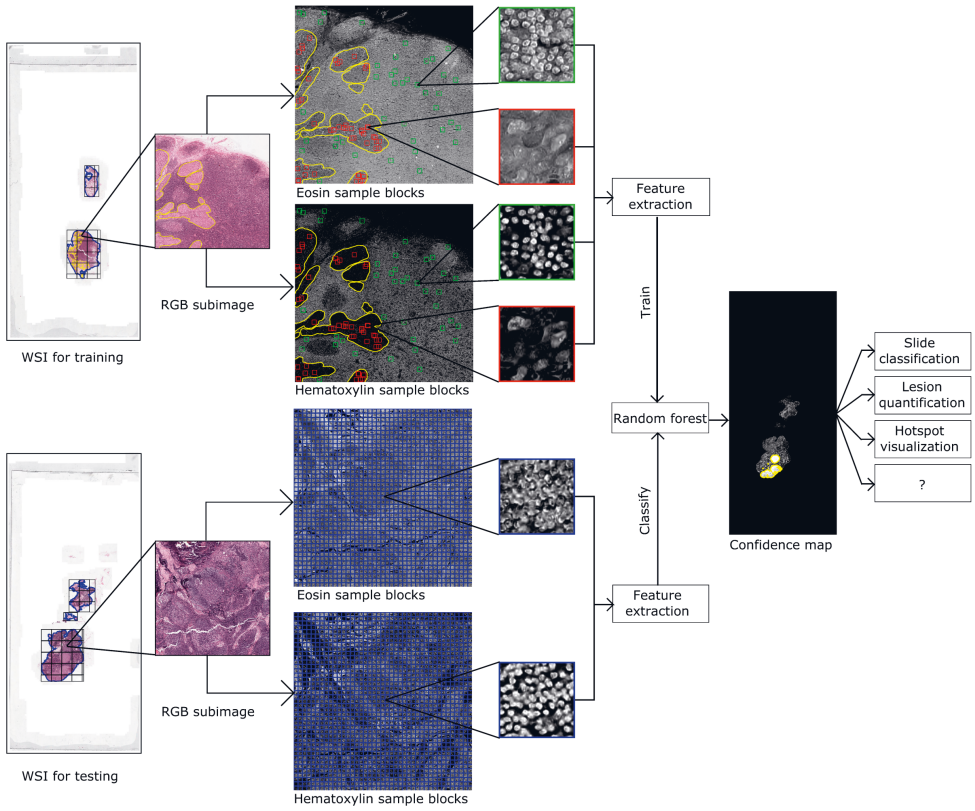

Figure 4.2 Overview of the breast cancer detection system. The tissue is segmented (left, blue outline) and split into subimages. Color deconvolution is applied to obtain channels corresponding to hematoxylin and eosin. In the training phase (top), patches representing malignant (red) tissue are randomly sampled from annotated regions (yellow outline). Benign patches (green) are sampled from other regions. Features are extracted from the patches and used to train RF classifiers. In the prediction phase (bottom), the probability of cancer is predicted for all patches using the trained model. Confidence maps of the probabilities across the WSI are generated and can be further used for different applications. Reprinted from Publication **III**.

4.3.4 Feature analysis

An advantage of feature-based ML is that the relevance of different features for the decision can be analysed. Features used by RF models can be ranked in terms of their relative importance by calculating how much the split criterion used for constructing the decision trees improves at each split. An overall importance measure for a feature can then be obtained as the average improvement over all splits in the RF model where the feature appears (Hastie et al. 2009). In Publication **III**, the outputs of each feature relative to the input image were additionally examined visually.

4.4 Deep learning

4.4.1 Patch classification

In Publication **IV**, the extracted patches were used to train DNN classifiers. The following DNN architectures were evaluated on a fixed validation split on the training data: Inception V3 (Szegedy et al. 2016), ResNet50 (He et al. 2016), Inception-ResNet V2 (Szegedy et al. 2017) and Xception (Chollet 2017), and based on this Inception V3 was selected. A key problem was the lack of pixel-level annotations for Gleason grades (see Section 4.2.2). For samples with a single Gleason pattern (e.g. GS 3+3), all malignant regions represent the pattern reported for the slide. In cases with multiple Gleason patterns (e.g. GS 3+4), there is no information on which malignant regions represent which pattern. Several approaches for utilizing the multi-pattern samples in training were evaluated, but ultimately training only on the single-pattern WSI was found to result in the best performance. However, this wastes training data that could still serve as examples of PCa tissue, even if the grade is unknown. For this reason, two separate models were used to perform: 1) cancer detection and 2) Gleason grading. The detection model was trained using all WSI, whereas the grading model was trained only using WSI with a single Gleason pattern.

Another problem was the class imbalance present in the training data, especially in terms of Gleason grades. That is, patches representing benign tissue are much more abundant than patches representing cancer, and particularly high grades. Failure to compensate for class imbalance can lead to poorly performing DNN models (Buda et al. 2018). The adopted solution consisted in picking all patches of the

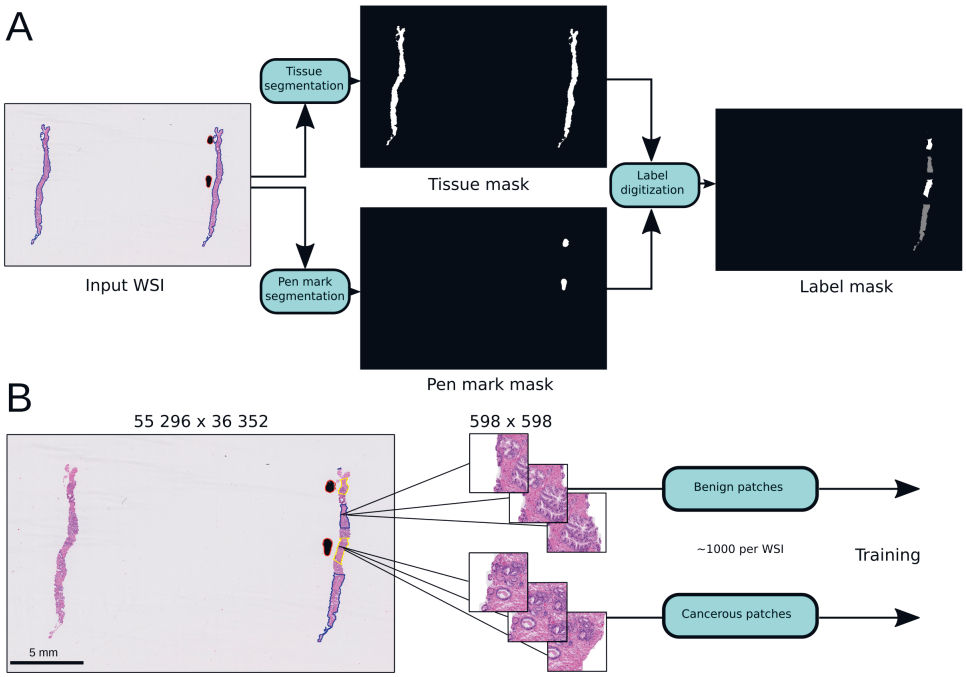

Figure 4.3 Image pre-processing for the prostate biopsy analysis system. **(A)** From left: Tissue (blue outline) and annotations (red outline) are segmented from each WSI and stored as binary masks. The annotations are projected onto adjacent tissue to obtain label masks indicating malignant (white), benign (grey) and unknown or background pixels (black). **(B)** From left: Each WSI is split into patches extracted from the tissue regions and labeled according to the label mask. Non-tissue pixels in the patches are masked with a constant white value. Approximately 1000 patches are obtained per WSI and input to DNNs for training (benign and malignant patches) or prediction (all tissue patches). Reprinted from Publication **IV**.

minority class and randomly sampling an equal amount of patches from all other classes, running one 'epoch' of training on these balanced data, and then repeating the sampling for the next 'epoch'. Overfitting against the minority class patches was decreased by applying data augmentation consisting of random rotations and flips every time a patch was drawn. As a result, the DNN was presented with modified versions of the minority class patches on every 'epoch'. Moreover, one can speculate that this approach, where random sampling is performed repeatedly during the training process, perturbing the set of training samples, may reduce the risk of the optimization getting stuck in local minima in a manner analogous to e.g. simulated annealing (Kirkpatrick et al. 1983). This aspect was not studied in Publication **IV**, but the approach was found to perform well empirically.

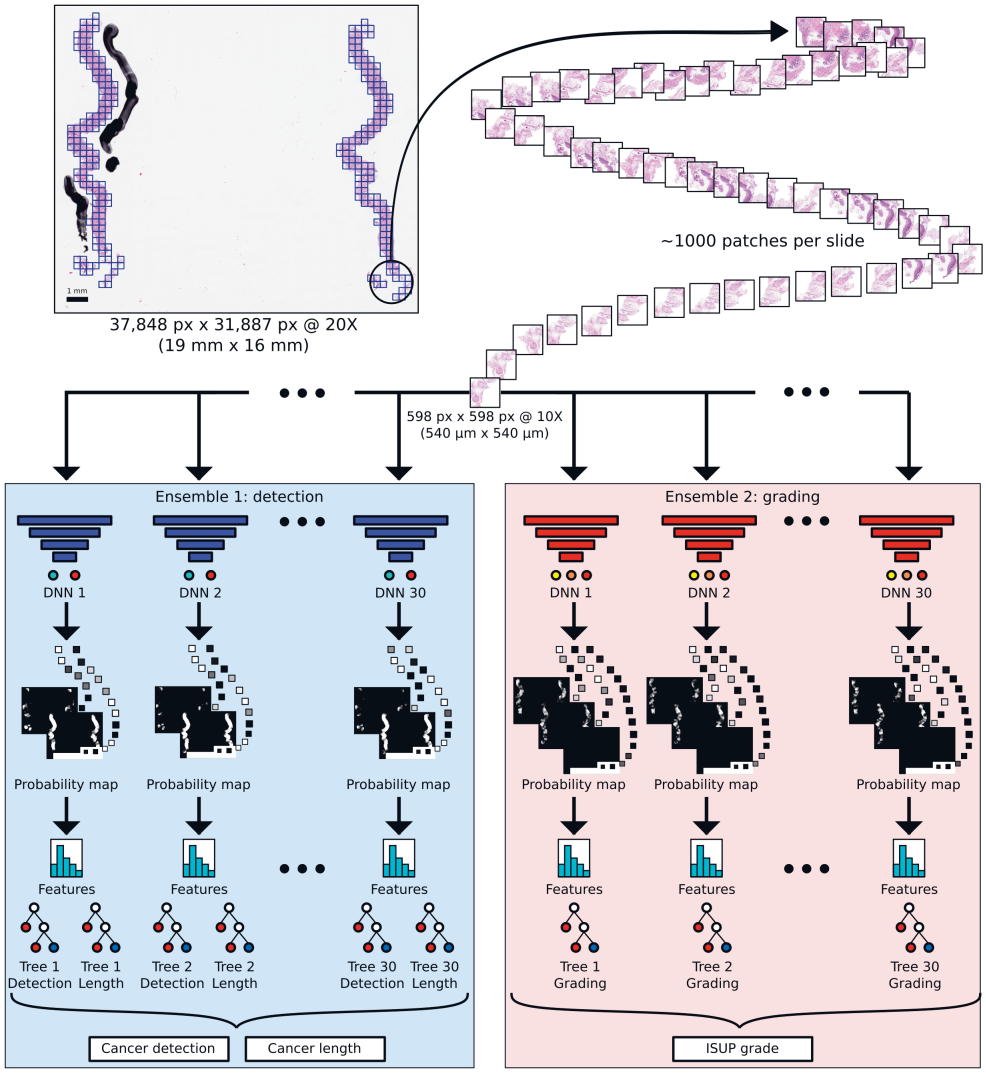

Figure 4.4 Overview of the prostate biopsy analysis system. Patches obtained from a WSI (top) are input to ensembles of 30 DNNs, trained to detect malignant patches (left box, top row) and estimate Gleason grade (right box, top row). The class-wise probabilities output by each DNN for each patch (grayscale squares) in the prediction phase are used to construct a probability map for each WSI and each class (both boxes, middle row). The probability maps are summarized using a number of statistical features, which are used for training ensembles of boosted trees for predicting cancer presence and length (left box, bottom row) and ISUP grade (right box, bottom row). The predictions from the boosted trees in each of the three ensembles are averaged to obtain the final prediction for a WSI. Reprinted from Publication **IV**.

We used an ensemble of multiple DNNs as the final classifier, with the same aim as in the case of the RF model, that is, to reduce the variance by averaging the output of multiple partially correlated classifiers. Similarly to decision trees, this can be beneficial also for neural networks due to their training process being stochastic (Hastie et al. 2009). We observed that the training process of the individual DNNs was relatively unstable, which carries a risk of converging to a poor, local optimum. Moreover, due to considerable label noise in the patch-level annotations, measuring patch-level performance during training was deemed unreliable in view of estimating when to stop training to obtain optimal slide-level performance. These issues can be mitigated by ensembling, and we chose to use 30 DNNs for cancer detection and 30 DNNs for Gleason grading. In the prediction phase, class-wise probabilities were assigned to each patch and composed into confidence maps similarly to Publication **III**. The overall ML system is depicted in Fig. 4.4.

4.4.2 Slide classification

As in Publication **III**, another ML stage was employed to obtain slide-level predictions. One XGBoost gradient boosted tree classifier (T. Chen et al. 2016) was trained per DNN, using features computed from the confidence maps as input. The features included statistics such as the median patch-wise probability estimated by the DNN for each WSI. Three ensembles of boosted trees were trained using the confidence maps generated for all training slides: one for detecting slides containing cancer, one for estimating cancer length, and one for estimating the ISUP grade of a slide. Since the first two tasks only depend on detecting patches with cancer irrespective of grade, the corresponding boosted trees were trained using the predictions from the cancer detection DNNs as input. The boosted trees performing grading were trained using the Gleason grading DNNs' predictions as input. The final slide-level predictions for each of the three tasks were obtained by taking the mean of the predictions produced by the boosted trees forming the corresponding ensemble.

4.4.3 Feature analysis

The representations learned by the DNN models were analyzed by two techniques. By extracting the activations from the second to last layer of the Inception V3 net-

work, a feature representation was obtained for each patch. These 2048-dimensional representations were condensed into two dimensions by performing dimensionality reduction with the t-distributed stochastic neighbor embedding (t-SNE) algorithm (van der Maaten 2014; van der Maaten et al. 2008). This allows visualizing where individual patches are located relative to each other in the feature space learned by the DNN, and allows some interpretation of the way the model has learned to represent the different morphologies. In addition, we analyzed the patterns in the input space that are relevant for the classification using the deep Taylor decomposition technique (Montavon et al. 2017) implemented in the iNNvestigate toolbox (Alber et al. 2019). This is a relevance backpropagation algorithm that produces estimates of the relative importance of pixels in the input images for the final output for the classifier, and allows visual examination of these estimates in the form of heatmap images.

4.5 3D reconstruction

The murine prostate and liver samples (see Section 4.1.1) were 3D reconstructed using the following algorithms to allow comparative analysis in Publications **I-II**.

- **Landmark-based reconstruction:** As a manual baseline method, an affine transformation was estimated for each pair of images representing adjacent tissue sections by minimizing MSE of the displacements between corresponding manually annotated landmarks.
- **Optimization-based reconstruction:** As an automated baseline method representing optimization-based registration, an affine transformation was estimated for each pair of consecutive images by maximizing pixel-wise MI (Publication **I**) or minimizing MSE (Publications **I-II**). A regular step gradient descent algorithm was used for the optimization task.
- **Feature-based reconstruction:** As an automated baseline method representing feature-based registration, an affine transformation was estimated for each pair of consecutive images using SURF (Bay et al. 2008) (Publication **I**) and SIFT (Lowe 2004) (Publications **I-II**) to obtain keypoints between the two images. Robust fitting of affine transformations to the keypoint pairs was performed using the Random Sample Consensus algorithm (Fischler et al. 1981).
- **HyperStackReg:** The HyperStackReg plugin (Ved P. Sharma, Albert Ein-

stein College, New York, https://github.com/ved-sharma/HyperStackReg) was used in Publication **II**. HyperStackReg is a multi-channel implementation of the optimization-based StackReg registration algorithm (Thevenaz et al. 1998). Global affine transformations were used as the transformation model.

- **Register Virtual Stack Slices:** The RegisterVirtualStackSlices and TransformVirtualStackSlices plugins (Arganda-Carreras et al. 2006) are based on the bUnwarpJ algorithm, which allows elastic registration based on the optimization of a loss function which combines SIFT and optimization based registration with regularization terms.

- **Elastic Stack Alignment:** The ElasticStackAlignment plugin (Saalfeld et al. 2012), a part of the TrakEM2 package (Cardona et al. 2012), implements a multi-step, elastic reconstruction process relying on optimization-based initialization, SIFT-based establishment of corresponding landmarks across the image stack, and a final optimization process relying on a physical model.

- **Medical Image Manager:** The 3D pathology add-on of Medical Image Manager (HeteroGenius Ltd, Leeds, UK) relies on an algorithm consisting in multi-resolution optimization of a spline-based elastic transformation. The volumes reconstructed using a trial version of the software were converted from MHD to TIFF format using the University of Leeds Volume Viewer.

- **Voloom:** Voloom (microDimensions GmbH, Munich, Germany) is a commercial software based on an elastic transformation model. A trial version of the software was used to perform 3D reconstructions.

In the case of the landmark-, optimization- and feature-based baseline methods implemented in MATLAB, the estimated pairwise transformations were concatenated to obtain a composite transformation for each image. The transformations were applied to each image using bilinear interpolation and to each corresponding tissue mask and landmark image using nearest neighbor interpolation. The ImageJ (Schneider et al. 2012) plugins were used in Fiji (Schindelin et al. 2012) and run via a Jython script through the ImageJ-MATLAB interface (Hiner et al. 2016). Medical Image Manager and Voloom represent standalone software.

4.6 Hyperparameter optimization

Simple grid search was used in Publication **I** to study the effect of hyperparameters on TRE, and in Publication **IV** to tune the hyperparameters of the DNN system. No systematic hyperparameter tuning was performed in Publication **III**. In Publication **II**, the hyperparameters of the reconstruction methods were optimized using Bayesian optimization, which lends it well to problems where a single iteration required for evaluating the objective function is computationally costly (Shahriari et al. 2015; Snoek et al. 2012). Moreover, Bayesian optimization is well-suited for non-convex objective functions. As TRE can be considered as the most reliable quality metric (Rohlfing 2011), mean pairwise TRE calculated over the entire stack of sections was chosen as the objective function to minimize.

4.7 Software and computing

The computations for Publications **I-III** were implemented in MATLAB apart from a number of reconstruction tools (see Section 4.5). For Publication **IV**, the image pre-processing steps were implemented in MATLAB and Python, the DL algorithms in Python using Keras (Chollet et al. 2015) with the TensorFlow backend (Abadi et al. 2016) and the gradient boosted trees using XGBoost (T. Chen et al. 2016) for Python. In Publications **I-III**, the images were stored in TIFF or JPEG2000 format (Tuominen et al. 2010). In Publication **IV**, the images were stored and accessed in the native WSI formats of the scanners using OpenSlide (Goode et al. 2013). This simplifies the overall processing pipeline by reducing unnecessary conversion steps, and avoids repeated image compression, which could introduce artefacts. Visualization of the confidence maps produced in Publication **IV** was performed using TissUUMaps, a web-based WSI viewer platform (Solorzano et al. 2020).

Resources provided by Tampere Center for Scientific Computing and CSC, Finland, were utilized for HPC. The computations for Publications **I-III** relied on parallel CPU computing, whereas Publication **IV** relied mainly on GPU computing. The GPU computing was performed on Nvidia Tesla P100 and V100 accelerators (Nvidia, Santa Clara, CA, USA) on the Narvi, Taito and Puhti compute clusters. Training was performed on multiple GPUs using the data parallelism strategy provided by Keras, where the DNN model is replicated on each GPU, and each mini-

batch is divided across the GPUs and processed in parallel. The results are then concatenated on the CPU and the weights of the model replicas are updated in synchronized manner. In addition to training each DNN in parallel on multiple GPUs, the training process of the ensemble in Publication **IV** was parallelized, such that all of the constituent classifiers were trained simultaneously. Moreover, efficient GPU computation required a buffering approach, where input patches are constantly read from a fast solid state drive and prepared on a multi-core CPU in parallel with the GPUs processing the preceding minibatch.

4.8 Statistical analysis

4.8.1 Evaluation of machine learning models

Splits between training and testing data were performed on patient-level meaning that all patches or biopsy cores from a given patient were used either for training or evaluation, which provides a more realistic setup in view of clinical use (Nir et al. 2019). In Publication **III**, all the evaluations were based on CV, i.e. there was no independent test set. The reason for this was that the study was prepared as a contribution to CAMELYON16, using data provided by the organizers, and in the context of a competition it was not reasonable to sacrifice training data for building an independent test set. In Publication **IV**, the evaluation was performed on an independent test set that was held out during the entire project and only used once for evaluation. Performance was additionally evaluated on the Pathology ImageBase (Egevad et al. 2018) and an external test set (see Section 4.1.3).

Cancer detection performance was evaluated using ROC analysis and quantified using AUC by comparing the output of the classifier to the pathologist's diagnosis. In Publication **III**, the analysis was performed primarily on the level of patches, since accurate patch-level annotations were available and highlighting the correct regions containing metastatic tissue was considered the main aim of the study. In Publication **IV**, the pixel-level annotations contained considerable label noise, which is why the analysis was performed on the level of biopsy cores and patients. Moreover, sample-level analysis was considered more relevant in view of the clinical focus of the study. In Publication **IV**, cancer detection performance was additionally evaluated in terms of specificity at multiple operating points corresponding to a range of sensi-

tivity values that were considered acceptable for clinical use. Cancer length estimates were compared to the pathologist's measurements in terms of the linear correlation coefficient. The analysis was performed both for individual biopsy cores and summarized on patient-level. Moreover, the evaluation was performed both using all cores, and only the cores indicated as positive by the pathologist.

Gleason grading performance was evaluated by comparing the ISUP grades estimated by the system in Publication **IV** to those assigned by the study pathologist (internal and external test set) or each of the 23 pathologists in the ImageBase panel. Cohen's kappa with linear weights was used as the metric, as it penalizes more for larger disagreements on the ordinal ISUP scale (Egevad et al. 2018). In the case of ImageBase, we calculated the mean of all pairwise Cohen's kappa values obtained for each pathologist and the DNN system when comparing them against all the other panel members and the system. All observers (including the DNN system) were then ranked in terms of their mean Cohen's kappa. Observers that tend to be most consistent with all other observers obtain the highest ranking in this evaluation.

4.8.2 Evaluation of 3D reconstructions

Manually selected landmarks were used to assess pairwise TRE (Fitzpatrick et al. 1998) between each pair of adjacent sections, followed by calculating mean TRE and other statistics across all section pairs to evaluate the entire reconstructed volume. Quantifying the accumulated distortion of the 3D shape using ATRE was performed slightly differently depending on the sample. In the case of the prostate, ATRE was calculated based on the displacement of the landmarks relative to their locations in the distortion-free, landmark-based reconstruction (see Section 4.5) (Publication **I**). This approach was modified in Publication **II** to allow estimating ATRE even if the volumes do not share the same coordinate system. In this case, the pairwise displacements of each landmark were treated as vectors and averaged to obtain the mean displacement of each tissue section. Distortion through the stack of sections was then evaluated by cumulatively summing the mean displacement vectors. In the case of the liver sample in Publication **II**, ATRE was estimated directly based on the displacement of the landmarks relative to the linear 3D trajectories they were expected to follow (see Section 4.1.1 and Fig. 4.5).

Pixel-wise similarity was evaluated for each pair of adjacent sections in the re-

constructed volume using RMSE, NCC and MI, calculated only on tissue pixels overlapping between the two sections. The mean and other statistics across all pairs of sections were calculated to evaluate the entire reconstruction. Reconstruction smoothness was quantified as suggested before (Cifor et al. 2011; Gaffling et al. 2014) by computing a GLCM for each pair of sections considering only the set of overlapping pixels, followed by summing the resulting matrices over the entire volume, and computing the *f2* contrast and *f3* correlation metrics.

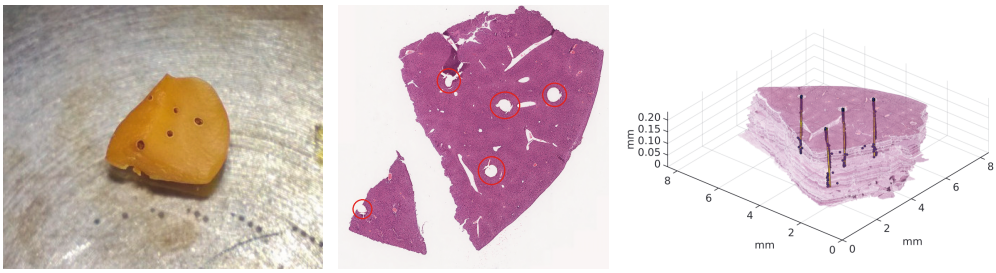

Figure 4.5 Evaluation of accumulated errors using fiducial markers. The liver sample was penetrated with an industrial laser (left), creating holes visible in the resulting WSIs (center). The deviations of the holes' locations on each section (blue dots) from linear trajectories (yellow lines) were evaluated post-reconstruction (right, reprinted from Publication **II**).

As control measures, we also calculated the Jaccard index (Rohlfing 2011) between each pair of adjacent sections and the relative tissue shrinkage compared to the original tissue section areas. A low Jaccard index, that is, a low degree of overlap between sections may indicate that the values of the pixel-wise similarity metrics are inconclusive. If not properly regularized, 3D reconstruction algorithms may also lead to excessive shrinkage of the tissue sections, which may lead to overly optimistic results in terms of pixel-wise similarity or TRE.

5 RESULTS

5.1 Feature-based learning for breast cancer detection

In Publication **III**, a feature-based ML method for detection of metastatic tissue in lymph node samples of BCa patients was presented. The system was shown to discriminate malignant from normal image patches with mean AUC values (95% CI) of 0.905 ([0.886, 0.925]) and 0.887 ([0.869, 0.905]) when analysed separately on samples from Radboud University Medical Center and University Medical Center Utrecht, respectively, each one evaluated using sample-level leave-one-out-cross-validation. That is, in these experiments, the training and testing data were always collected at the same site. When training on all samples from one institution and evaluating on all samples from the other institution to assess generalization performance, mean AUC values (95% CI) of 0.839 ([0.821, 0.856]) and 0.855 ([0.831, 0.879]) were obtained. Alternatively, the tissue segmentation step can be considered a part of the detection task, in which case the evaluation can be performed based on the entire WSI area instead of tissue areas only. Since tissue segmentation was considered a non-trivial step in view of metastasis detection, the AUC values corresponding to this problem formulation were also presented in Publication **III**. However, taking into account the ease at which tissue segmentation can be performed today and the performance of the improved segmentation algorithm developed for Publication **IV**, considering the segmentation step in the evaluation is of limited interest.

Publication **III** represents a contribution to CAMELYON16 (Bejnordi et al. 2017), and this context is reflected in some of the limitations of the study. The study was only based on data provided by the challenge organizers and the aim of the study was to obtain maximal metastasis detection performance within the setting of the competition. This means that all of the available data were used for model training and development, resulting in the lack of a held-out test set. As a result, performance estimates reported for the method of Publication **III** as well as those of other challenge

participants are likely to be overly optimistic in view of generalization to completely unseen data. One must therefore be cautious in drawing conclusions, as is the case for many challenge results in general (Maier-Hein et al. 2018).

Moreover, the system was initially designed to obtain optimal performance measured on the metric used in the challenge, based on free-response ROC analysis. This required providing one coordinate point with an associated probability per each metastatic region, and has a number of complications. Firstly, it does not evaluate if the shape or size of the detected region corresponds to the annotations. Secondly, it requires a post-processing algorithm for condensing regional pixel or patch-level predictions into a single coordinate. The performance of this post-processing step may have a considerable impact on the results, but its practical relevance is questionable in view of using such a system as a detection aid for highlighting suspicious regions on WSIs. For these reasons, conventional patch-wise ROC analysis was chosen in Publication **III**. However, as the system was not initially designed with this metric in mind, the reported performance may be somewhat pessimistic compared to what could be achieved with a similar system fully optimized for the task of choice.

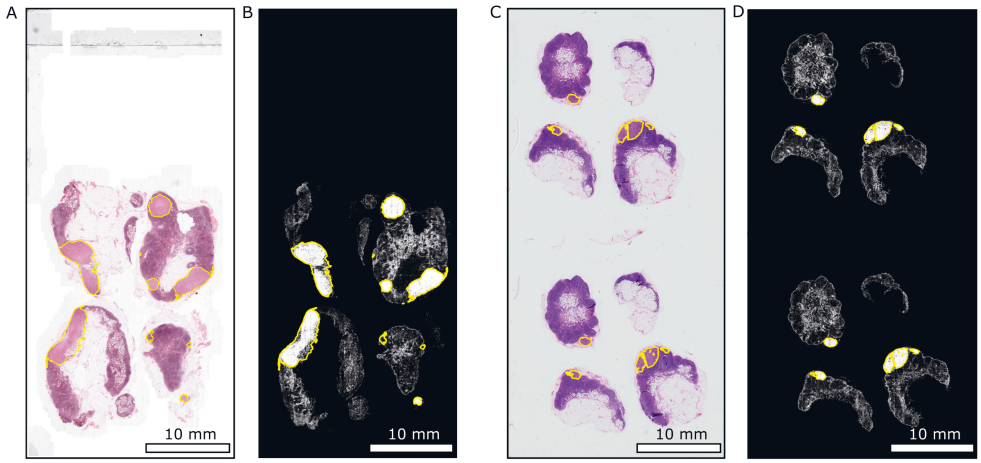

Figure 5.1 Output of the metastasis detection system for example WSIs from Radboud University Medical Center **(A)** and University Medical Center Utrecht **(C)** as confidence maps **(B)** and **(D)**, respectively. The brightness corresponds to the estimated probability of malignancy. Annotated metastatic regions are indicated with yellow outlines. Reprinted from Publication **III**.

Due to the above limitations, the presented method can be mainly considered a proof-of-concept, demonstrating that detection of metastatic tissue using feature-based learning on whole-slide scale is feasible both in terms of computation and in

terms of promising classification performance. The result met Aim **I** of the research, but the work on pre-processing and handling WSI datasets of considerable size was relevant also in view of Aim **IV**. Being able to produce visual outputs highlighting the approximate regions of each WSI most likely to contain metastasis (Fig. 5.1) could already be useful for speeding up the work of pathologists if utilized as a semi-automated diagnostic aid. However, considering the rapid progress of the field, the main value of the work presented in Publication **III** is perhaps the point of reference it provides as the top-ranked (11th) feature-based method in CAMELYON16. The proposed method outperformed a number of DNN-based contributions, and reached performance comparable to some of the top 10 methods, all of which relied on DL. However, the top-performing DNN-based methods were clearly superior, exemplifying the wide performance gap to even the best feature-based approaches of the time, similar to other image analysis tasks.

5.2 Deep learning for prostate cancer grading

In Publication **IV**, a DL system for assessing prostate biopsies was presented. In terms of detecting biopsies with PCa, the system achieved an AUC (95% CI) of 0.997 ([0.994, 0.999]) on a held out test set sampled from the same data source as the training data, and an AUC of 0.986 ([0.972, 0.996]) on an external test set. Cancer length in each biopsy could be estimated with linear correlation coefficients of 0.96 ([0.95, 0.97]) and 0.87 ([0.84, 0.90]) on the internal and external test sets, respectively, when compared to the values reported by the study pathologist. In terms of Gleason grading, the system reached linearly weighted Cohen's kappa values of 0.83 (internal test) and 0.70 (external test) relative to the study pathologist. On the Pathology ImageBase dataset graded by 23 experienced uropathologists, the DNN system reached a mean pairwise kappa of 0.62, whereas the pathologists in the panel had corresponding values ranging from 0.60 to 0.73. An example prediction is shown in Fig. 5.2.

The demonstrated performance would most likely be sufficient for such a system to be useful in clinical practice. For example, at a sensitivity of 99.3 % for detecting malignant biopsy cores, which would have meant correctly detecting every patient with cancer in the internal test set, the specificity of the system was estimated to be 88.9 %. This corresponds to positive and negative predictive values of 87.6 % and 99.4 %, respectively. In the clinical setting, this could be leveraged by automatically

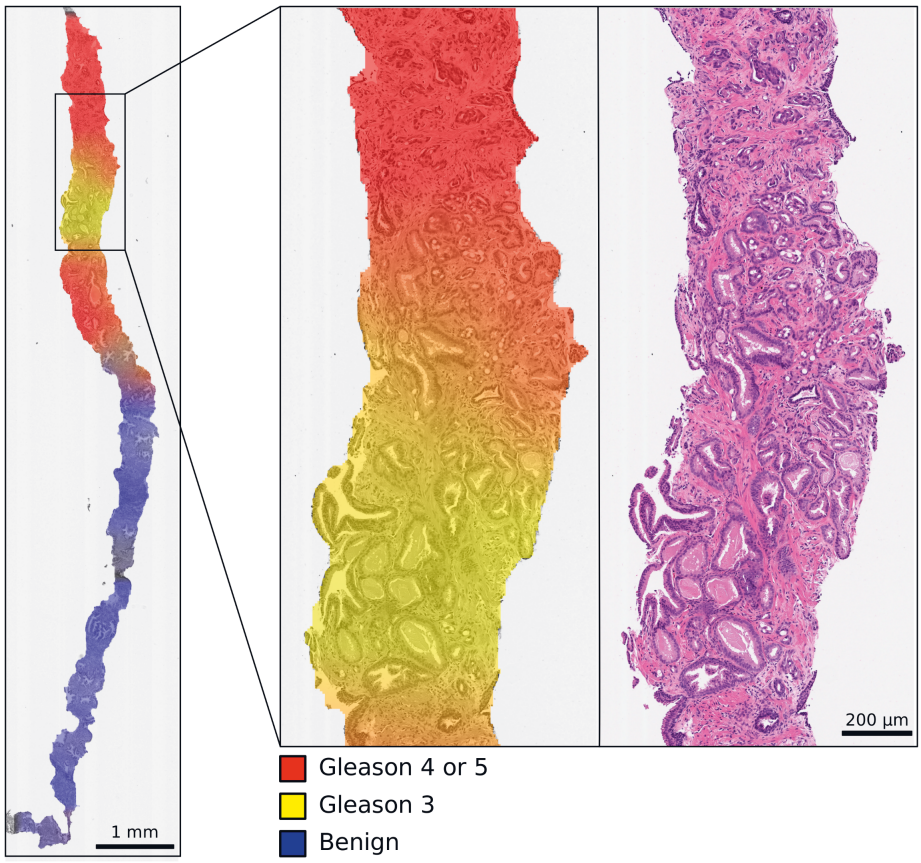

Figure 5.2 Output of the prostate biopsy analysis system visualized for a case from the test set. The intensity of the colors indicates the estimated probability of benign (blue) and Gleason pattern 3 (yellow) or 4-5 (red) across the biopsy, combined into RGB values. A magnified tissue region (right) exemplifies a predicted transition between low and high Gleason grades (center). The sample had GS 4+3 according to the pathologist. Reprinted from Publication **IV**.

pre-screening all the samples, after which only those predicted to be positive would be assessed by a pathologist. As benign biopsies constitute the majority of all samples, this would greatly decrease the workload in pathology departments. A lower risk alternative would be to re-analyze all samples indicated as benign by the pathologist, potentially improving safety by detecting cancers that were initially missed. Moreover, automating cancer length measurements appears feasible in view of the obtained results and could lead to time savings for pathologists. The question of how automated grading could be used in the clinic is less straightforward. One option would be to consider it as decision support especially for inexperienced pathol-

ogists, or as a means of providing expertise to regions where pathologists are not available. However, further studies in a clinical setting are needed to understand how pathologists would interact with automated decision support systems and how these systems would influence clinical decisions.

Publication **IV** marks the first time automated uropathologist-level grading of prostate biopsies has been demonstrated based on a sizable WSI dataset representing a well-defined clinical cohort. A particular strength of the study is that the samples were collected within the STHLM3 clinical trial (Grönberg et al. 2015) and represent a population-based sample of patients, which allowed clinically meaningful estimation of performance metrics. Moreover, the data contain the full spectrum of prostate tissue morphology encountered in clinical practice, for example, difficult to diagnose cancer variants and benign mimickers of cancer. The fact that similar results were independently achieved by another research group based on a different dataset (Bulten et al. 2020) supports the finding that DNN models can indeed assess prostate biopsies comparably to experienced pathologists. Taken together, these findings met Aim **II** of the research. In addition, the computations required for conducting the study necessitated the use of HPC, and as another product of the study, a streamlined computational workflow utilizing parallel GPU computing on compute clusters was designed. Together with the initial work conducted for Publication **III**, this met Aim **IV** of the research.

5.3 Comparison of 3D reconstruction algorithms

In Publication **I**, a panel of quality metrics for evaluating 3D histology reconstructions was designed and demonstrated on a set of baseline algorithms used to reconstruct a murine prostate sample. The main outcome of the study was the evaluation framework itself, which represents a collection of metrics and evaluation techniques which have been used variably in earlier 3D histology studies. The algorithms based on SIFT and SURF features outperformed optimization-based algorithms in terms of TRE but not in terms of pixel-wise similarity metrics or reconstruction smoothness. Compared to the result of the manual landmark-based method, which is in principle free of global distortions, the optimization-based approaches exhibited some undesired straightening of curved structures. Since these algorithms consider all tissue pixels in their optimization process, they are more prone to distorting elon-

gated structures that are curved or located at an angle relative to the sectioning plane. This explains the seemingly better pixel-wise similarity results of these methods and underlines the importance of considering multiple complementary metrics when interpreting reconstruction quality results. Moreover, the effect of hyperparameter selection on TRE was found to be considerable for all of the algorithms and it should not be overlooked in comparative studies.

In Publication **II**, the evaluation framework was applied to compare a comprehensive selection of 3D reconstruction algorithms, including two commercial software products. Moreover, in addition to the prostate sample, a liver sample was prepared, and both samples were independently annotated by two observers. Moreover, artificial fiducial markers allowing direct evaluation of accumulated distortions were introduced into the liver sample using a novel method, an industrial laser. In terms of hyperparameter tuning, most of the tunable parameters of each algorithm were now considered and Bayesian optimization was employed instead of parameter sweeps in view of computational feasibility.

Based on the evaluation, algorithms using elastic transformation models were found superior to methods that only use global affine transformations. Interestingly, the leading algorithms outperformed even the baseline reconstruction that relied on the manually selected landmarks. This finding, while not overly surprising, confirms that compensating for local tissue deformations is required for obtaining optimal 3D reconstructions. Moreover, all of the top algorithms feature some sort of a multi-resolution scheme and a transformation model where different parts of the image are registered independently of each other, which increases robustness against regions containing artefacts. Differences between the evaluated algorithms were more subtle on the prostate dataset, but the more challenging liver tissue caused more considerable variation in the results (see Fig. 5.3). The two commercial solutions, Medical Image Manager and Voloom, and the free ElasticStackAlignment (Saalfeld et al. 2012) performed most robustly even when facing artefacts such as torn tissue.

To the best of my knowledge, Publication **II** marks the first time Bayesian optimization has been utilized to successfully tune the hyperparameters of 3D reconstruction algorithms. This supports earlier findings demonstrating the utility of Bayesian optimization for other WSI analysis tasks (Teodoro et al. 2016). While this was not the main aim of the study, it may serve as an example encouraging automated hyperparameter optimization in CP applications. Hyperparameter optimization

was performed on the prostate dataset, and the same parameters were used for the liver sample. Encouragingly for practitioners of 3D histology, the top methods performed well on both datasets without tissue-specific parameter tuning. Moreover, the results collected during the optimization process provide a semi-quantitative view of the behaviour of each reconstruction algorithm in response to parameter adjustments. This revealed differences which have practical utility for the users of these algorithms: for example, RegisterVirtualStackSlices (Arganda-Carreras et al. 2006) was found to be highly sensitive to parameter choices and difficult to tune, whereas the optimization and feature-based baseline algorithms reacted in a predictable and stable manner. This aspect is often overlooked in comparative studies of algorithms, but Publication **II** underlines its importance. Taken together, the results presented in Publications **I-II** met Aim **III** of the research.

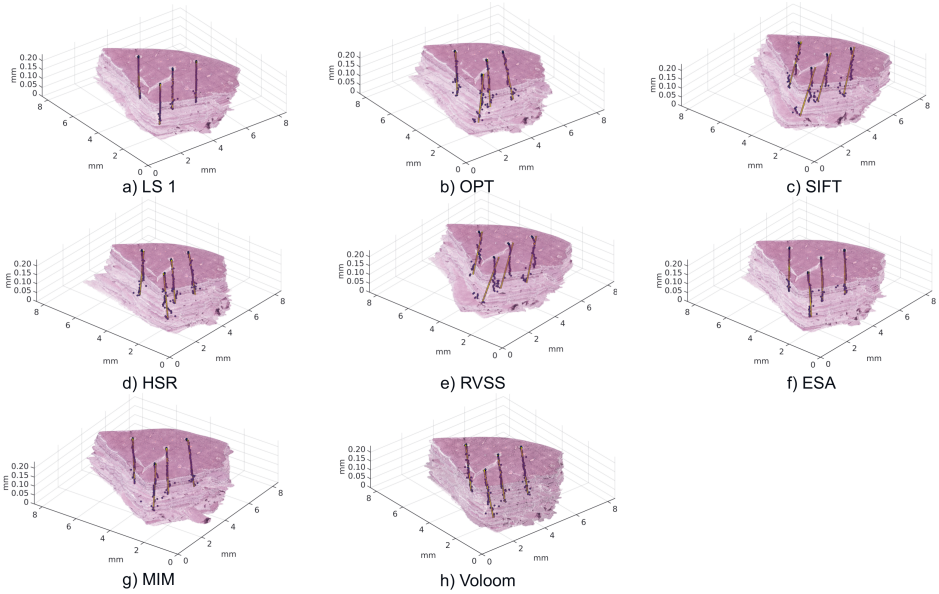

Figure 5.3 The liver sample reconstructed using **(a)** manual landmarks, **(b)** optimization, **(c)** SIFT features, **(d)** HyperStackReg, **(e)** RegisterVirtualStackSlices, **(f)** ElasticStackAlignment, **(g)** Medical Image Manager and **(h)** Voloom, all using optimized hyperparameters. The manually annotated locations of the landmark holes on each section are indicated with dots and lines of best fit to the landmarks are shown as yellow lines. Reprinted from Publication **II**.

6 DISCUSSION

6.1 Clinical adoption of machine learning based diagnostics

The field of ML-based histopathological diagnostics has progressed rapidly during the last few years, a process that is clearly mirrored by the publications included in this thesis as well. In as late as 2014, the commercial availability of clinically applicable computer-aided diagnostics for histopathology was seen as a prospect of the distant future (Mosquera-Lopez et al. 2014). This conclusion referred to the assessment of prostate biopsies, but the diagnosis of PCa is hardly a special case. Even how to computationally process entire WSIs was still considered an open question at the time. Much of the work behind this thesis focused on overcoming such practical obstacles as well. Deep learning, efficient GPU computing and the exponentially increasing amount of WSI data have, however, quickly led to results such as those presented in Publication **IV** and by others (Bulten et al. 2020; Campanella et al. 2019; Esteva et al. 2017), demonstrating performance comparable to medical experts on data that are representative of the clinical reality.

With the emergence of breakthrough results, within the span of only a few years, computer aided diagnostics for histopathology has turned to a realistic candidate for imminent clinical adoption, with even commercial software such as Galen Prostate (IBEX Medical Analytics Ltd, Tel Aviv-Yafo, Israel) (Pantanowitz et al. 2020) and INIFY Prostate (ContextVision AB, Stockholm, Sweden) (Fraggetta 2019) becoming available. The recent developments have also been met with optimism by pathologists. In a recent survey conducted among ISUP members, 71% of the respondents believed ML will have a role in screening, decision support and efficiency improvements in PCa histopathology during the current decade (van Leenders et al. 2020). Moreover, 31% of the respondents had themselves been involved in ML projects.

While the optimism is well motivated by the impressive recent results, there have been voices of caution regarding some of the current practices in the field,

which might jeopardize successful clinical adoption of the technology. Shortcomings pointed out in the way that medical AI research is currently being carried out include incomplete reporting that hampers the reproducibility of the studies, lack of rigour in the way algorithms are evaluated (Liu et al. 2019), and a trend towards only publishing results as preprints without a peer-review process ("AI diagnostics need attention" 2018). Very few prospective clinical trials have so far evaluated the use of ML tools in actual clinical workflows (Rajkomar et al. 2019). Reliable evaluation of the algorithms is, however, crucial if they are to be deployed in the clinic. Evaluation metrics need to be chosen carefully to reflect the intended diagnostic task, as opposed to e.g. only evaluating performance in terms of low-level segmentation metrics (Gurcan et al. 2009). Ultimately, the focus has to be on patients rather than pixels or patches. In line with this, it has been shown to be important to perform splits between training and testing data at the patient level, as was done in Publications **III-IV** (Nir et al. 2019). While a somewhat obvious requirement, evaluation of algorithms has rarely been performed on external validation data, which would provide more realistic performance estimates than CV experiments (Liu et al. 2019). One reason for neglecting this aspect has previously been the limited amount of WSI data, but with the collection of larger datasets, construction of sufficiently large independent test sets becomes feasible.

Encouragingly, there are signs that the focus of the field is now shifting from demonstrating the accuracy of prototype ML algorithms to devising concrete plans and guidelines on how to implement the technology in health care, taking into account the associated regulatory and administrative steps as well as health economics (Colling et al. 2019). For example, the World Health Organization and the International Telecommunication Union recently established a focus group with the aim of developing a benchmarking process for AI tools in healthcare (Wiegand et al. 2019). Initiatives like this will hopefully enable safe, evidence-based adoption of ML in clinical diagnostics, as well as in other medical applications.

6.2 Handling real-world variability in WSI data

One question crucial for the clinical use of CP systems is how to handle the variability present in real-world WSI data, such as that introduced by different scanners (Fig. 6.1). The way factors like scanner characteristics, different laboratories and patient

populations affect ML algorithms is still relatively poorly understood. Moreover, different healthcare providers may have varying preferences in terms of the operating characteristics of the systems, such as sensitivity, perhaps necessitating site-specific calibration of classifiers. In the few studies which have presented external validation on samples that originate from a different source than the training data, it has not been uncommon to observe considerable degradation in performance compared to internal test sets. For example, Campanella *et al.* reported drops of 3% points and 6% points in AUC for a different scanner and different laboratories, respectively, in the task of PCa detection (Campanella et al. 2019). While these results are far from catastrophic and regarded by the authors as an indication of successful generalization, the clinical implications of such a drop in performance would not be negligible. Similarly, a drop of 1.1% points in AUC was observed in Publication **IV** when applying the system to samples prepared in a different laboratory and scanned on a device different from the training and internal test data.

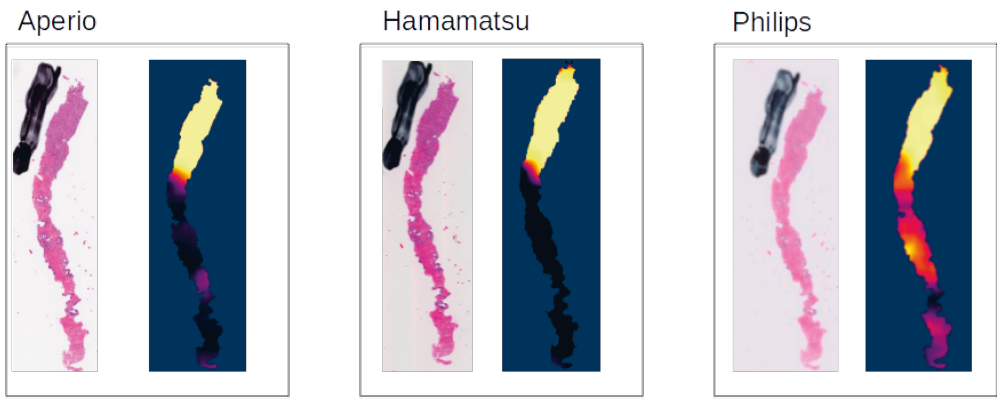

Figure 6.1 An example prostate biopsy digitized with Aperio, Hamamatsu and Philips scanners. Probability of prostate cancer presence estimated by the DNN system of Publication **IV** is visualized as a heatmap alongside each image. Differences in both the appearance of the tissue and in the resulting predicted probabilities are visible.

The simplest way of mitigating the issues due to variability in the data is to collect heterogeneous training material. This brute-force approach was relatively successful in Publication **IV**, where the training images were obtained using scanners from two different vendors. However, including all possible sources of variability in the training data is difficult. Moreover, new sources of variation may arise over time, for example as the scanner used in a particular laboratory ages. Handling such variability

would then require always collecting more and more training material and periodically retraining the classifier. This would be problematic not only in terms of the workload, but also in view of regulatory aspects. Various computational solutions such as color normalization or augmentation (Tellez, Litjens, Bandi et al. 2019), the stain-GAN (Shaban et al. 2019) or domain-adversarial training (Lafarge et al. 2019) have been proposed but thus far there have been few comparative studies on these methods. Evaluating these methods in terms of their effects on the clinical operating characteristics of AI systems will be a key question to solve in the near future.

Another way of looking at the issue of generalization to data that differs from the training material is automated quality control, which can be used to ensure that the classifiers only operate on data they are suited for. In some sense, this seems a more realistic concept than attempting to train classifiers that are robust against all known and unknown input perturbations. Moreover, incorporating estimates of the uncertainty associated with a classifier's decision could be helpful in recognizing cases where intervention by a human expert is necessary, and could help pathologists to build trust on computational diagnostic aids (Colling et al. 2019; Rajkomar et al. 2019). To this end, techniques such as conformal prediction and Bayesian DL are currently being studied (Gupta et al. 2019).

6.3 The issue of explainable decisions

There is an ongoing debate on whether machine-based decisions should always be explainable in high-risk domains such as healthcare, or if performance should be prioritized even at the expense of interpretability (Rudin 2019). Arguably, if a diagnostic classifier has been reliably shown to be accurate, and it can be monitored to ensure it retains its performance over time, then the process used to arrive at the decisions is irrelevant in view of the desired output of accurate diagnoses. On the other hand, clinical application of models whose operating principles are not fully understood can be problematic in terms of medical device regulations (Bera et al. 2019). Troubleshooting a system whose decision process is unknown is also difficult. Perhaps even more importantly, blind reliance on machine-generated decisions without any associated explanations may even lead to the deterioration of the clinical skills of pathologists over time (Colling et al. 2019). If taking this worrying thought experiment to the extreme, application of non-explainable AI over several decades could

eventually lead to the loss of the human skill of visually assessing and understanding tissue morphology. Of course, many skills have faced the same fate over the course of history. Perhaps a more likely consequence at least in the medium term is that changes in pathologist education are required to train 'computational pathologists' with a skillset combining medical and algorithmic expertise (Colling et al. 2019).

One way of obtaining explainable decisions is to use models designed with this in mind from the start, such as feature-based ML instead of DL. For example, several relatively recent studies aimed at automated Gleason grading using features engineered to capture spatial properties of nuclei and glands, designed based on existing histopathological knowledge (Niazi et al. 2016; Nir et al. 2018; D. Wang et al. 2015). The authors claimed that the visual interpretability of the features aids pathologists in validating and accepting the results (Niazi et al. 2016). Combinations of DL and feature engineering have also been proposed, where predefined features are extracted and provided as input to a DNN in addition to the raw images (Sadanandan et al. 2016) or fused with DNN-based features at the final classification stage (Valkonen, Kartasalo et al. 2017). However, considering the widening performance gap between DL and feature-based approaches, it seems unlikely that a return to classical methods solely for the sake of explainability will happen. Moreover, even feature-based models can be difficult to explain, if the features used do not have clear histological interpretations, such as most of the features used in Publication **III**.

An alternative is designing additional algorithms for explaining the decisions of black box classifiers. Publication **IV** relied on such DNN interpretation methods (Alber et al. 2019) that help highlight input patterns relevant for the classifier (Fig. 6.2). However, these methods do not explain *what* a particular pattern represents for the model, and *why* it is important for the decision. Interpreting such visual explanations is qualitative and subjective at best. Moreover, these methods have been criticized as they can merely approximate the model they are trying to explain (otherwise an explanation would not be needed in the first place) and it is difficult to judge whether the explanation is true to the original model (Rudin 2019). These techniques can, however, help a human designer to spot hidden biases in the underlying data (Lapuschkin et al. 2019). Such hidden stratification, where a model learns to recognize undesired patterns in the data, instead of those truly relevant for the diagnosis can have severe consequences (Oakden-Rayner et al. 2020). In the context of CP, an example are subtle patterns introduced into the images by different scanners.

If the data are biased such that samples representing some condition of interest are more frequently scanned on one scanner than other samples, the classifier may learn to associate the patterns representing that scanner with the condition. Even crude model interpretation techniques may help in detecting such effects. Ensuring that the trained model is not relying on any such biases in the underlying data was the main rationale for applying model interpretation algorithms in Publication **IV**.

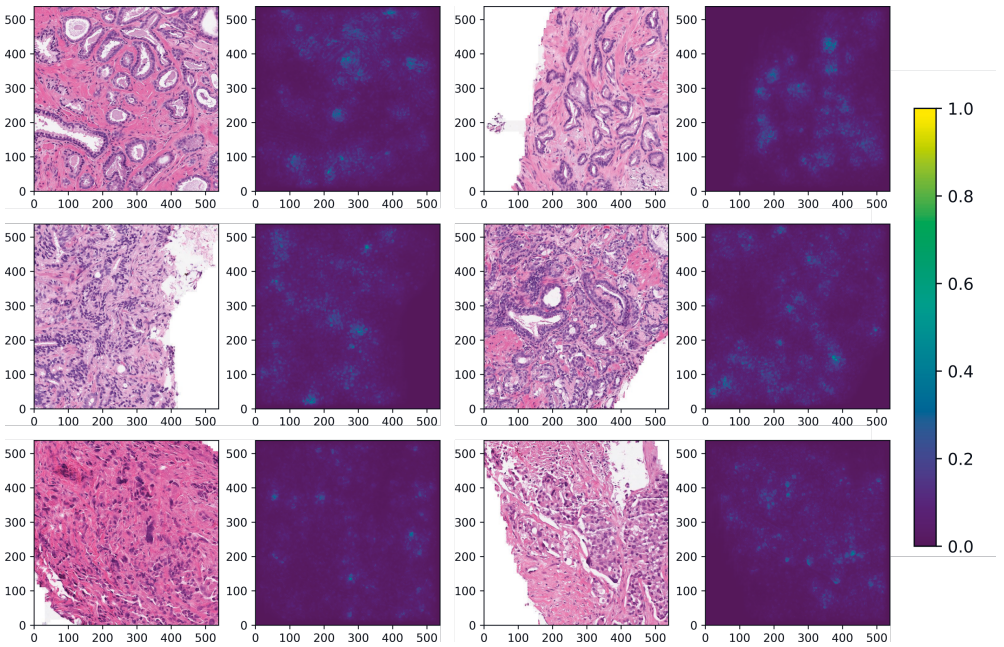

Figure 6.2 Randomly selected tiles representing Gleason patterns 3 (top row), 4 (middle row) and 5 (bottom row) with estimates of the relative importance of input image pixels for the DNN's decisions. The analysis was performed using deep Taylor decomposition. Nuclei and the edges of glands are highlighted as important input patterns. Axis units are in μm. Modified from Publication **IV**.

Studying the phenomenon of adversarial examples, frequently criticized as a weakness of DNN models, provides another viewpoint on explainability (Ilyas et al. 2019). The introduction of subtle, artificial perturbations, such as the addition of a minute amount of noise to input data has been shown to often fool DNNs to produce erroneous predictions. This has been interpreted as an indication of the tendency of DNNs to overfit to irrelevant patterns in the data, and as proof for the necessity of using model interpretation techniques to diagnose such undesired behaviour. Interestingly, Ilyas *et al.* showed that this interpretation may merely be the result of a

biased human-centric view, which is built on the assumption that patterns used by humans in their decisions are true and robust, whereas patterns imperceptible to humans are irrelevant artefacts, and a classifier relying on such patterns is not robust. Classification errors caused by artificial perturbation of these subtle patterns in a way that is imperceptible to a human observer, but might never happen naturally in the real-world data, do not prove that the patterns are irrelevant for the decision task. The root cause of this issue may well be fundamental differences in the way that humans and current machine intelligence work. Forcing ML models to only utilize the same patterns that humans recognize inevitably means that all the potential performance that lies beyond human cognition is never achieved. However, these observations should not be taken as a confirmation of some sort of natural law dictating that a trade-off between performance and interpretability is always unavoidable (Rudin 2019). In other words, there are no guarantees that increasingly complex black box models will always be superior to human-comprehensible solutions, even if that is currently the case in image analysis, but there are neither guarantees of new interpretable models emerging that could challenge the performance of DNNs.

6.4 Large-scale 3D histology

Even though studies such as Publications **I-II** have shown that 3D reconstructions can be performed with subcellular registration accuracy, and many studies over the years have demonstrated the potential value of the technology, 3D histology has not become a mainstream approach. By far the biggest obstacle to clinical or widespread research use is the considerable manual work needed to prepare serial sections. Scanning of the prepared slides is no longer an issue due to the high throughput of WSI systems, but cutting of the samples remains a manual process performed by skilled technicians. Automated sectioning machines have been demonstrated to produce high-quality serial sections and these kind of devices could potentially remove the tissue preparation bottleneck (Fu et al. 2018; Onozato et al. 2013). However, the task is difficult and requires sophisticated robotics. So far these systems have not been widely adopted in histology laboratories, but should they become sufficiently reliable, they would represent a breakthrough for 3D histology.

Microscopy techniques that directly image the sample in 3D have also been proposed (Farahani et al. 2017). For example, knife edge scanning microscopy integrates

the sectioning and imaging process using a line scanner built into the sectioning blade (Mayerich et al. 2008). Scanning is performed while cutting the sample, section by section, producing a 3D volume directly. The technology has been commercialized by 3Scan, Inc. (San Francisco, CA, USA). A very similar device called micro-optical sectioning tomography has also been proposed (A. Li et al. 2010). The same destructive layer-by-layer imaging principle can be coupled with other sources of contrast, such as multiphoton imaging (Ragan et al. 2012) or optical coherence microscopy (Min et al. 2020). Optical clearing of tissue followed by multi-photon imaging (Torres et al. 2014) or confocal microscopy (van Royen et al. 2016) has been suggested as an alternative, allowing 3D imaging of uncut specimens without the destructive serial sectioning. However, all of these techniques have the limitation that tissue staining is challenging, as it has to be done by perfusing the uncut sample prior to sectioning. Serial sectioning followed by computational 3D reconstruction does not share this limitation, as the tissue sections can be processed with any histological or biochemical techniques prior to scanning. Therefore, while these new imaging techniques can undoubtedly produce exciting new data, they do not fully answer the problems associated with achieving high-throughput 3D histology. One interesting future prospect is, however, using non-stained tissue material coupled with DL-based virtual staining (Rivenson et al. 2019) to circumvent the problems with staining uncut samples.

Another hurdle for 3D histology is the analysis and examination of the resulting volumetric data. This is of course routine in other types of medical imaging, such as MRI, but the considerably higher resolution means that most tools developed for 3D medical imaging cannot be directly applied to 3D histology. For example, at full resolution, the mouse prostate reconstructed in Publications **I-II** contains more than 50 gigavoxels, corresponding to over 150 GB of image data. The size of a corresponding human organ would of course be vastly larger. In addition to the technical issues, visualizing and analysing the data in a manner that allows intuitive interpretation is a challenge. Three-dimensional examination of tissue morphology at microscopical resolution is quite simply a task that experts are not accustomed to. For this reason, the ability of 3D histology based on serial computational reconstruction to utilize standard techniques like HE staining can be considered an advantage, as the tissue itself appears similar to the 2D images pathologists are already used to. Many of the computational analytics can also be extended from 2D to 3D in a relatively straight-

forward manner (Liang et al. 2015). Virtual reality has been proposed as a way of exploring 3D microscopy data in a manner that is intuitive to humans (Cali et al. 2016; Theart et al. 2017). There is ongoing work focused on applying virtual reality to the 3D histology data generated using the methods of Publication **II**, allowing one to examine the morphology of the murine prostates at full resolution as if one was inside the sample (Liimatainen et al. 2020). As an alternative, physical models produced by 3D printing have also been studied (Liimatainen et al. 2019).

6.5 From imitating to surpassing human experts

Besides emulating pathologists to match or exceed human performance in routine tasks (e.g. Publications **III-IV**), and performing quality control and enhancement, attempts are increasingly made to extend CP to new kinds of analyses (Bera et al. 2019; Niazi et al. 2019). The first step in this direction is quantification of pathological features that are difficult and time consuming for humans to assess. An example is the detection and further spatial analysis of tumor infiltrating lymphocytes, which has prognostic significance across several cancers (Saltz et al. 2018). Precise quantification of such features, requiring the assessment of large numbers of individual cells and their relative spatial locations, may be better suited for computational systems than humans. Another example is using DL to discover patterns that, while not perhaps completely overlooked by pathologists, are not formally part of structured grading schemes. For example, DNNs have been used to identify tumor-associated patterns in stromal tissue in breast biopsies (Bejnordi et al. 2018). The diagnosis of BCa is mainly based on the morphology of epithelial tissue, but the associated stroma appears to also contain patterns that are potentially relevant for the diagnostics.

An obvious way to try and leverage the capacity of ML to detect patterns in data is to train models directly with clinical outcomes to derive image-based prognostic predictors as opposed to producing readouts following current pathological reporting systems, such as Gleason grading (Gurcan et al. 2009). This would avoid some of the issues associated with the subjectivity of pathologist-derived ground truth labels (Colling et al. 2019). While there have been attempts at performing prognostics based on engineered features as well (Yu et al. 2016), the key advantage of DL in these applications is that it can potentially allow data-driven discovery of novel image-based biomarkers, since the features used do not need to be fixed by a human

designer *a priori*. An example is stratification of colorectal cancer patients into low- and high-risk cases based on TMA samples (Bychkov et al. 2018). In that study, relying on transfer learning, generic features were extracted from the tissue samples using a VGG-16 CNN (Simonyan et al. 2014) pre-trained on ImageNet (Russakovsky et al. 2015). The features extracted from patches across each slide were then used as input to a long short-term memory network (Hochreiter et al. 1997), a type of RNN, to predict the probability of 5-year disease specific survival. The system outperformed experts when the cases were dichotomized into low- and high-risk groups.

Following a similar line of thought, generic features were extracted from more than 13 000 images of RP samples in an unsupervised manner using deep autoencoders (Yamamoto et al. 2019). This resulted in a 100-dimensional feature representation for each case, and these representations were then used to train a SVM to predict 1-year and 5-year biochemical recurrence of PCa. The authors reported AUC values of 0.820 and 0.721 for 1 and 5 year predictions, respectively. This outperformed predictions based on GS (0.744 and 0.695). Interestingly, combining both the autoencoder features and GS resulted in even better predictions. Nagpal *et al.* also assessed the ability of their system to predict risk of biochemical recurrence of PCa and reported improved prognostic performance compared to a genitourinary specialist (Nagpal et al. 2019). However, due to the fact that the grading performed as part of the clinical routine had already affected the treatment decisions of these patients, drawing such a conclusion is not straightforward (Eklund et al. 2019). Avoiding such pitfalls when comparing the prognostic utility of new markers to those in routine use requires careful study design.

Integrated analysis of morphology and molecular markers has been proposed as a means of extracting more useful information from histopathological samples (Gurcan et al. 2009). Mobadersany *et al.* used CNNs to predict survival times in glioma cases (Mobadersany et al. 2018). The model, termed survival-CNN by the authors, features a combination of a CNN and a Cox proportional hazards model layer and was trained using 1061 WSIs of glioma cases from The Cancer Genome Atlas. The performance of the algorithm in predicting survival was comparable to current models based on visual assessment and molecular subtypes. By adding genomic information on mutations and deletions as input to the CNN, the authors reported a further improvement in performance. Another way of blurring the line between image-based morphological data and the underlying molecular information is image-based

modeling of molecular features, such as the presence of mutations. Besides accurately diagnosing and classifying lung cancer samples into subtypes, Coudray *et al.* also attempted to train an Inception V3 model to predict the mutational status of the most commonly mutated genes in lung adenocarcinoma based only on the image data, succeeding in six out of ten genes (Coudray et al. 2018). As another example, Jain *et al.* were able to predict the mutational burden of lung cancer samples based on WSI data, achieving high concordance with whole exome sequencing (Jain et al. 2020). Performing such analyses, combining many different inter-related data types, is considered to be beyond human cognition and thus an exciting opportunity for a new kind of AI-enabled pathology (Bera et al. 2019; Niazi et al. 2019).

6.6 Scalability of computational pathology development

The recent improvements in CP have, to a large extent, relied on increasingly larger datasets being analyzed using increasingly complex models, run on increasingly faster computers. Whether this process is scalable, allowing further development at an increasing pace, is thus a key question in view of the future of the field. A key requirement for this progress is the accessibility of data. There are unfortunate technical hurdles in the form of lacking standards for WSI formats, which complicates data access in comparison to e.g. radiology, and has required CP developers to resort to reverse engineering (Goode et al. 2013). Efforts are ongoing to incorporate WSI data into the DICOM standard (Colling et al. 2019) but it remains to be seen if time is finally ripe for this long-standing aim (Tuominen et al. 2010) to become reality.

A bigger hurdle in terms of data availability is that relatively few institutions are capable of collecting the kind of datasets used in Publication **IV** or other recent studies (Bulten et al. 2020; Campanella et al. 2019) in-house, and only some of these institutions are willing and allowed to share the data. One consequence of data not being openly available is that comparison of proposed algorithms is difficult, since each study typically relies on a different dataset and different performance metrics (Gurcan et al. 2009). The role of comparative studies such as Publication **II**, challenges (Hartman et al. 2020) and openly available data for benchmarking algorithms, such as the CAMELYON datasets (Litjens et al. 2018), are therefore considered central for the advancement of the field. All of the data used for this thesis has either been openly shared by other institutions (Publication **III**), deposited to public reposito-

ries (Publications **I-II**) or made available through the organization of a challenge[1] (Publication **IV**). It is easy to claim that all data should always be shared in the name of open science, but realizing this in practice in a fair manner is a whole different matter. Collection and annotation of large medical datasets is costly and time-consuming and getting adequate return on these investments is not an unreasonable requirement from the data owner. In the case of medical data, further questions arise in terms of who should actually own the data and what are the legal and patient information security constraints on data sharing. These are important questions in view of the future of CP as a clinically and commercially utilized set of technologies.

Even if data are available, they are typically not useful for development in their raw form. A much discussed bottleneck for the development of ML algorithms is the annotation of data by medical experts (Gurcan et al. 2009). Performing the kind of detailed annotations used in all publications of this thesis cannot be scaled up indefinitely to larger datasets. It has been suggested, that ML in the medical domain ultimately cannot rely only on supervised learning (Greenspan et al. 2016). The use of weakly supervised learning, relying only on sample-level labels rather than pixel-level annotations, has recently been advocated as a more scalable alternative (van der Laak et al. 2019) and demonstrated to result in excellent performance when applied to data encompassing tens of thousands of WSI (Campanella et al. 2019). The advantage of weakly supervised learning is that sample-level labels are typically obtained as a by-product of routine clinical practice instead of requiring medical experts to devote time to labeling data. However, one can question if the process of extracting this information from often unstructured and even non-digital pathology reports and patient records is truly scalable either. Suggested approaches for collecting annotations in a more automated manner include tracking the microscope usage of pathologists (Gecer et al. 2018) or using IHC to obtain biochemically derived pixel-wise labels (Sadanandan et al. 2017; Turkki et al. 2016; Valkonen et al. 2019). However, the former includes considerable uncertainty as the time spent by pathologists examining a particular region may not always correlate reliably with the final label, and the latter requires that a biomarker indicative of the target of interest is known and can be stained. The application of new imaging techniques such as scanning Raman microspectroscopy may provide an alternative way of obtaining labels, but this will require more research to establish the correspondence of e.g. Raman spectra with

[1] https://www.kaggle.com/c/prostate-cancer-grade-assessment

known histological features (Hollon et al. 2020).

Another less discussed bottleneck may be the AI development process itself. Designing increasingly complex systems in a scalable manner may require accelerating the design process itself via algorithmic solutions. While systematic design approaches are emerging and replacing trial and error, the role of optimal design choices and parameter tuning has been largely ignored in CP. Considerable improvements in many ML tasks were achieved when feature engineering was replaced by DL (LeCun et al. 2015). However, the requirement for "craftsmanship" has arguably merely shifted to the tasks of designing DNN architectures, pre-processing and data augmentation algorithms, and tuning hyperparameters (Hutter et al. 2019). Besides accelerating development of AI systems for new tasks, reducing the involvement of human designers may lead to further performance improvements. The emerging field of automated machine learning (Auto-ML) aims at automating system design in a data-driven manner. Applications of Auto-ML to the design of CP algorithms have so far been nearly non-existent (Koohbanani et al. 2018). However, there is no reason to assume Auto-ML would not be effective also within this field. Publications **I-II** and **IV** addressed hyperparameter tuning in a systematic manner, using Bayesian optimization and grid searches, but algorithmic approaches for designing entire AI pipelines from low-level building blocks are now emerging (Real et al. 2020). Ultimately, designing increasingly complex AI systems by hand to model pathology even at the level currently considered by human experts, let alone exceeding that, may not be feasible. Adoption of Auto-ML algorithms in the future may help to not only surpass human medical experts but also human ML designers and data scientists.

The third ingredient to the recent progress besides more data and more capable ML models is HPC. Thus far, CP developers, and the AI field in general, have to a large extent settled on a brute-force mentality, where the growing datasets and neural networks are powered by increasing amounts of parallel computation (B. Chen et al. 2019). The studies presented in this thesis are no different, and Publication **IV** in particular greatly benefited from parallel GPU computing, which enabled training large DNN ensembles. At the moment, it seems that this trend of hardware investments dominating over algorithmic improvements is set to continue. For example, the European HPC Joint Undertaking is investing in compute clusters such as LUMI[2], featuring thousands of GPU accelerators. However, there have recently

[2]https://www.lumi-supercomputer.eu/

been voices of concern over the energy consumption and the resulting CO_2 emissions caused by the inflating computations (Strubell et al. 2019). Using large DNN models for natural language processing with expensive hyperparameter tuning as an extreme example, Strubell *et al.* claimed that the CO_2 footprint of the entire process of training such a model once can exceed that of an average car in the USA during its entire lifespan. Data center design resulting in carbon neutral operation can arguably mitigate the problem in terms of CO_2 emissions, and development of more efficient hardware customized for DL to replace GPUs may allow scaling up the computational capacity for long into the future. Still, it is unclear if neglecting algorithmic development and focusing on further scaling up existing solutions will stop yielding performance improvements at some point, and if so, when this will happen. Encouragingly, recent studies have presented some progress on more efficient DNN architectures (Tan et al. 2019) and approximate algorithms for accelerating computations even on CPUs (B. Chen et al. 2019). It would be easy to imagine that despite the many successes highlighted in this thesis, this period of "brute-force machine learning" will be seen as rather primitive in the future.

7 CONCLUSIONS

The studies presented in this thesis handled methodology for 3D histology and machine learning based diagnostics in computational pathology. The results of these studies can be summarized as follows:

- In Publication **I**, a quality evaluation framework for 3D histology reconstruction algorithms was designed. Quality metrics measuring pixel-wise similarity and reconstruction smoothness were found to complement those quantifying target registration error based on manually selected landmarks. The effects of hyperparameter tuning, assessed using grid search, were found to be considerable for all of the evaluated methods, indicating that parameter adjustments should not be ignored when comparing reconstruction algorithms.
- In Publication **II**, the 3D histology evaluation framework was extended to achieve compatibility with most reconstruction tools and then applied to compare several publicly available algorithms and two commercial options. The feasibility of automated hyperparameter tuning of 3D reconstruction algorithms using Bayesian optimization was demonstrated for the first time, and the process resulted in improved performance compared to default parameters for all of the evaluated algorithms. Algorithms relying on elastic transformation models capable of compensating for local tissue deformations achieved the most accurate reconstructions. The commercial tools Medical Image Manager 3D Pathology and Voloom, as well as the freely available Elastic Stack Alignment plugin for ImageJ exhibited the best overall performance.
- In Publication **III**, a machine learning based solution was developed for detecting metastatic tissue in lymph node samples of breast cancer patients. A tissue segmentation algorithm and a patch-wise processing method were developed to efficiently handle entire whole slide images. The system, based on extracting a large number of features quantifying texture characteristics and spatial

properties of cell nuclei as input for random forest classifiers, was shown to have good discriminatory performance for distinguishing between patches of normal and metastatic tissue. This allowed generating visualizations highlighting potentially malignant regions on the slides, which could potentially speed up the work of pathologists.

- In Publication **IV**, a deep learning based system for detecting and grading prostate cancer in biopsies was presented. Pre-processing methods were designed for tissue segmentation and digitization of annotations manually drawn on the slides, allowing training of ensembles of DNN classifiers to detect malignant patches and predict their Gleason grade. Slide-level predictions of cancer presence, cancer length and ISUP grade were obtained by using ensembles of gradient boosted trees operating on the patch-wise outputs of the DNN classifiers. Using an independent test set reflecting a population-based, clinically representative sample of patients, the system was shown to achieve over 99% sensitivity at approximately 89% specificity in cancer detection and cancer length estimates were shown to closely correspond to those performed by a pathologist. The grading performance of the system was comparable to a panel of experienced urological pathologists. External validation on samples prepared in a different laboratory and scanned on a different scanner also resulted in acceptable performance. This marks the first time that diagnostic performance comparable to specialized pathologists has been demonstrated on a large, clinically representative dataset of prostate biopsies.

REFERENCES

"AI diagnostics need attention" (2018). *Nature* 555.7696, 285.

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv preprint 1603.04467*.

Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M. D., van der Laak, J., Bui, M. M., Vemuri, V. N., Parwani, A. V., Gibbs, J., Agosto-Arroyo, E., Beck, A. H. and Kozlowski, C. (2019). Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *The Journal of Pathology* 249.3, 286–294.

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S. and Kindermans, P.-J. (2019). iNNvestigate neural networks!: *Journal of Machine Learning Research* 20.93, 1–8.

Amunts, K., Lepage, C., Borgeat, L., Mohlberg, H., Dickscheid, T., Rousseau, M.-É., Bludau, S., Bazin, P.-L., Lewis, L. B., Oros-Peusquens, A.-M., Shah, N. J., Lippert, T., Zilles, K. and Evans, A. C. (2013). BigBrain: an ultrahigh-resolution 3D human brain model. *Science* 340.6139, 1472–1475.

Arganda-Carreras, I., Sorzano, C. O., Marabini, R., Carazo, J. M., Ortiz-de-Solorzano, C. and Kybic, J. (2006). Consistent and elastic registration of histological sections using vector-spline regularization. *International Workshop on Computer Vision Approaches to Medical Image Analysis*, 85–95.

Arvaniti, E., Fricker, K. S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P. J., Rüschoff, J. H. and Claassen, M. (2018). Automated Gleason

grading of prostate cancer tissue microarrays via deep learning. *Scientific Reports* 8.1, 12054.

Awan, R. and Rajpoot, N. (2018). Deep Autoencoder Features for Registration of Histology Images. *Annual Conference on Medical Image Understanding and Analysis*, 371–378.

Bagci, U. and Bai, L. (2010). Automatic best reference slice selection for smooth volume reconstruction of a mouse brain from histological images. *IEEE Transactions on Medical imaging* 29.9, 1688–1696.

Bándi, P., Geessink, O., Manson, Q., van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B., Paeng, K., Zhong, A., Li, Q., Zanjani, F. G., Zinger, S., Fukuta, K., Komura, D., Ovtcharov, V., Cheng, S., Zeng, S., Thagaard, J., Dahl, A. B., Lin, H., Chen, H., Jacobsson, L., Hedlund, M., Çetin, M., Halıcı, E., Jackson, H., Chen, R., Both, F., Franke, J., Küsters-Vandevelde, H., Vreuls, W., Bult, P., van Ginneken, B., van der Laak, J. and Litjens, G. (2018). From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging* 38.2, 550–560.

Basavanhally, A., Ganesan, S., Feldman, M., Shih, N., Mies, C., Tomaszewski, J. and Madabhushi, A. (2013). Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides. *IEEE Transactions on Biomedical Engineering* 60.8, 2089–2099.

Bay, H., Ess, A., Tuytelaars, T. and van Gool, L. (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding* 110.3, 346–359.

Beare, R., Richards, K., Murphy, S., Petrou, S. and Reutens, D. (2008). An assessment of methods for aligning two-dimensional microscope sections to create image volumes. *Journal of Neuroscience Methods* 170.2, 332–344.

Bejnordi, B. E., Litjens, G., Timofeeva, N., Otte-Höller, I., Homeyer, A., Karssemeijer, N. and van der Laak, J. A. (2015). Stain specific standardization of whole-slide histopathological images. *IEEE Transactions on Medical Imaging* 35.2, 404–415.

Bejnordi, B. E., Mullooly, M., Pfeiffer, R. M., Fan, S., Vacek, P. M., Weaver, D. L., Herschorn, S., Brinton, L. A., van Ginneken, B., Karssemeijer, N., Beck, A. H., Gierach, G. L., van der Laak, J. A. and Sherman, M. E. (2018). Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies. *Modern Pathology* 31.10, 1502–1512.

Bejnordi, B. E., Veta, M., van Diest, P. J., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A. W. M., Hermsen, M., Manson, Q. F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M. C., Bult, P., Beca, F., Beck, A. H., Wang, D., Khosla, A., Gargeya, R., Irshad, H., Zhong, A., Dou, Q., Li, Q., Chen, H., Lin, H.-J., Heng, P.-A., Haß, C., Bruni, E., Wong, Q., Halici, U., Öner, M. Ü., Cetin-Atalay, R., Berseth, M., Khvatkov, V., Vylegzhanin, A., Kraus, O., Shaban, M., Rajpoot, N., Awan, R., Sirinukunwattana, K., Qaiser, T., Tsang, Y.-W., Tellez, D., Annuscheit, J., Hufnagl, P., Valkonen, M., Kartasalo, K., Latonen, L., Ruusuvuori, P., Liimatainen, K., Albarqouni, S., Mungal, B., George, A., Demirci, S., Navab, N., Watanabe, S., Seno, S., Takenaka, Y., Matsuda, H., Ahmady Phoulady, H., Kovalev, V., Kalinovsky, A., Liauchuk, V., Bueno, G., Fernandez-Carrobles, M. M., Serrano, I., Deniz, O., Racoceanu, D. and Venâncio, R. (2017). Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA* 318.22, 2199–2210.

BenTaieb, A. and Hamarneh, G. (2017). Adversarial stain transfer for histopathology image analysis. *IEEE Transactions on Medical Imaging* 37.3, 792–802.

Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. and Madabhushi, A. (2019). Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology* 16.11, 703–715.

Booth, M. E., Treanor, D., Roberts, N., Magee, D. R., Speirs, V. and Hanby, A. M. (2015). Three-dimensional reconstruction of ductal carcinoma in situ with virtual slides. *Histopathology* 66.7, 966–973.

Born, G. (1883). Die plattenmodellirmethode. *Archiv für mikroskopische Anatomie* 22.1, 584–599.

Borovec, J., Kybic, J., Arganda-Carreras, I., Sorokin, D. V., Bueno, G., Khvostikov, A. V., Bakas, S., Chang, E. I., Heldmann, S., Kartasalo, K., Latonen, L., Lotz, J., Noga, M., Pati, S., Punithakumar, K., Ruusuvuori, P., Skalski, A., Tahmasebi, N., Valkonen, M., Venet, L., Wang, Y., Weiss, N., Wodzinski, M., Xiang, Y., Xu, Y., Yan, Y., Yushkevic, P., Zhao, S. and Muñoz-Barrutia, A. (2020). ANHIR: Automatic Non-rigid Histological Image Registration Challenge. *IEEE Transactions on Medical Imaging*.

Bosch, A., Zisserman, A. and Munoz, X. (2007). Image classification using random forests and ferns. *IEEE International Conference on Computer Vision*, 1–8.

Bradley, D. and Roth, G. (2007). Adaptive thresholding using the integral image. *Journal of Graphics Tools* 12.2, 13–21.

Braumann, U.-D., Kuska, J.-P., Einenkel, J., Horn, L.-C., Loffler, M. and Hockel, M. (2005). Three-dimensional reconstruction and quantification of cervical carcinoma invasion fronts from histological serial sections. *IEEE Transactions on Medical Imaging* 24.10, 1286–1307.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A. and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* 68.6, 394–424.

Breiman, L. (2001). Random forests. *Machine Learning* 45.1, 5–32.

Brown, P. J., Toh, E.-W., Smith, K. J., Jones, P., Treanor, D., Magee, D., Burke, D. and Quirke, P. (2015). New insights into the lymphovascular microanatomy of the colon and the risk of metastases in pT1 colorectal cancer obtained with quantitative methods and three-dimensional digital reconstruction. *Histopathology* 67.2, 167–175.

Buda, M., Maki, A. and Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* 106, 249–259.

Bulten, W., Bándi, P., Hoven, J., van de Loo, R., Lotz, J., Weiss, N., van der Laak, J., van Ginneken, B., Hulsbergen-van de Kaa, C. and Litjens, G. (2019). Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. *Scientific Reports* 9.1, 1–10.

Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C. and Litjens, G. (2020). Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology*.

Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P. E., Verrill, C., Walliander, M., Lundin, M., Haglund, C. and Lundin, J. (2018). Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific Reports* 8.1, 3395.

Cali, C., Baghabra, J., Boges, D. J., Holst, G. R., Kreshuk, A., Hamprecht, F. A., Srinivasan, M., Lehväslaiho, H. and Magistretti, P. J. (2016). Three-dimensional immersive virtual reality for studying cellular compartments in 3D models from

EM preparations of neural tissues. *Journal of Comparative Neurology* 524.1, 23–38.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S. and Fuchs, T. J. (July 2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 679–698.

Cardona, A., Saalfeld, S., Schindelin, J., Arganda-Carreras, I., Preibisch, S., Longair, M., Tomancak, P., Hartenstein, V. and Douglas, R. J. (2012). TrakEM2 software for neural circuit reconstruction. *PLOS One* 7.6, e38011.

Casero, R., Siedlecka, U., Jones, E. S., Gruscheski, L., Gibb, M., Schneider, J. E., Kohl, P. and Grau, V. (2017). Transformation diffusion reconstruction of three-dimensional histology volumes from two-dimensional image stacks. *Medical Image Analysis* 38, 184–204.

Chen, B., Medini, T., Farwell, J., Gobriel, S., Tai, C. and Shrivastava, A. (2019). Slide: In defense of smart algorithms over hardware acceleration for large-scale deep learning systems. *arXiv preprint 1903.03129*.

Chen, P.-H. C., Gadepalli, K., MacDonald, R., Liu, Y., Kadowaki, S., Nagpal, K., Kohlberger, T., Dean, J., Corrado, G. S., Hipp, J. D., Mermel, C. H. and Stumpe, M. C. (2019). An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nature Medicine* 25.9, 1453–1457.

Chen, T. and Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 785–794.

Chen, Y., Janowczyk, A. and Madabhushi, A. (2020). Quantitative Assessment of the Effects of Compression on Deep Learning in Digital Pathology Image Analysis. *JCO Clinical Cancer Informatics* 4, 221–233.

Cheng, H.-D., Chen, C. and Freimanis, R. I. (1995). A neural network for breast cancer detection using fuzzy entropy approach. *IEEE International Conference on Image Processing*. Vol. 3, 141–144.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *IEEE Conference on Computer Vision and Pattern Recognition*, 1251–1258.

Chollet, F. et al. (2015). *Keras*. `https://keras.io`.

Cifor, A., Bai, L. and Pitiot, A. (2011). Smoothness-guided 3-D reconstruction of 2-D histological images. *NeuroImage* 56.1, 197–211.

Clarke, E. L. and Treanor, D. (2017). Colour in digital pathology: a review. *Histopathology* 70.2, 153–163.

Colling, R., Pitman, H., Oien, K., Rajpoot, N., Macklin, P., CM-Path AI in Histopathology Working Group, Snead, D., Sackville, T. and Verrill, C. (2019). Artificial intelligence in digital pathology: A roadmap to routine use in clinical practice. *The Journal of Pathology* 249.2, 143–150.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning* 20.3, 273–297.

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N. and Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine* 24.10, 1559–1567.

Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J. and Madabhushi, A. (2014). Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. *International Society for Optics and Photonics Medical Imaging 2014: Digital Pathology*. Vol. 9041, 904103.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. *IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 1, 886–893.

Diamond, J., Anderson, N. H., Bartels, P. H., Montironi, R. and Hamilton, P. W. (2004). The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. *Human Pathology* 35.9, 1121–1131.

Doyle, S., Feldman, M., Tomaszewski, J. and Madabhushi, A. (2010). A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. *IEEE Transactions on Biomedical Engineering* 59.5, 1205–1218.

Doyle, S., Madabhushi, A., Feldman, M. and Tomaszeweski, J. (2006). A boosting cascade for automated detection of prostate cancer from digitized histology. *International Conference on Medical Image Computing and Computer-assisted Intervention*, 504–511.

Edge, S. B. and Compton, C. C. (2010). The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Annals of Surgical Oncology* 17.6, 1471–1474.

Egevad, L., Ahmad, A. S., Algaba, F., Berney, D. M., Boccon-Gibod, L., Compérat, E., Evans, A. J., Griffiths, D., Grobholz, R., Kristiansen, G., Langner, C., Lopez-Beltran, A., Montironi, R., Moss, S., Oliveira, P., Vainer, B., Varma, M. and Camparo, P. (2013). Standardization of Gleason grading among 337 European pathologists. *Histopathology* 62.2, 247–256.

Egevad, L., Cheville, J., Evans, A. J., Hörnblad, J., Kench, J. G., Kristiansen, G., Leite, K. R. M., Magi-Galluzzi, C., Pan, C.-C., Samaratunga, H., Srigley, J. R., True, L., Zhou, M., Clements, M., Delahunt, B. and The ISUP Pathology Imagebase Expert Panel (2017). Pathology Imagebase—a reference image database for standardization of pathology. *Histopathology* 71.5, 677–685.

Egevad, L., Delahunt, B., Berney, D. M., Bostwick, D. G., Cheville, J., Comperat, E., Evans, A. J., Fine, S. W., Grignon, D. J., Humphrey, P. A., Hörnblad, J., Iczkowski, K. A., Kench, J. G., Kristiansen, G., Leite, K. R. M., Magi-Galluzzi, C., McKenney, J. K., Oxley, J., Pan, C.-C., Samaratunga, H., Srigley, J. R., Takahashi, H., True, L. D., Tsuzuki, T., van der Kwast, T., Varma, M., Zhou, M. and Clements, M. (2018). Utility of Pathology Imagebase for standardisation of prostate cancer grading. *Histopathology* 73.1, 8–18.

Egevad, L., Ström, P., Kartasalo, K., Olsson, H., Samaratunga, H., Delahunt, B. and Eklund, M. (2020). The utility of artificial intelligence in the assessment of prostate pathology. *Histopathology* 76.6, 790–792.

Eklund, M., Kartasalo, K., Olsson, H. and Ström, P. (2019). The importance of study design in the application of artificial intelligence methods in medicine. *npj Digital Medicine* 2.1, 1–2.

Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B., Srigley, J. R. and Humphrey, P. A. (2016). The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *The American Journal of Surgical Pathology* 40.2, 244–252.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542.7639, 115–118.

Fan, J., Cao, X., Yap, P.-T. and Shen, D. (2019). BIRNet: Brain image registration using dual-supervised fully convolutional networks. *Medical Image Analysis* 54, 193–206.

Farahani, N., Braun, A., Jutt, D., Huffman, T., Reder, N., Liu, Z., Yagi, Y. and Pantanowitz, L. (2017). Three-dimensional imaging and scanning: current and future applications for pathology. *Journal of Pathology Informatics* 8.

Feuerstein, M., Heibel, H., Gardiazabal, J., Navab, N. and Groher, M. (2011). Reconstruction of 3-D histology images by simultaneous deformable registration. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 582–589.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* 24.6, 381–395.

Fitzpatrick, J. M., West, J. B. and Maurer, C. R. (1998). Predicting error in rigid-body point-based registration. *IEEE Transactions on Medical Imaging* 17.5, 694–702.

Fónyad, L., Shinoda, K., Farkash, E. A., Groher, M., Sebastian, D. P., Szász, A. M., Colvin, R. B. and Yagi, Y. (2015). 3-dimensional digital reconstruction of the murine coronary system for the evaluation of chronic allograft vasculopathy. *Diagnostic Pathology* 10.1, 16.

Fraggetta, F. (2019). Clinical-grade computational pathology: Alea iacta est. *Journal of Pathology Informatics* 10.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. *International Conference on Machine Learning*. Vol. 96, 148–156.

Fu, X., Klepeis, V. and Yagi, Y. (2018). Evaluation of an Automated Tissue Sectioning Machine for Digital Pathology. *Diagnostic Pathology* 4.1.

Fuchs, T. J. (2010). Computational pathology: a machine learning approach. PhD thesis. ETH Zurich.

Fuchs, T. J. and Buhmann, J. M. (2011). Computational pathology: Challenges and promises for tissue analysis. *Computerized Medical Imaging and Graphics* 35.7-8, 515–530.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* 36.4, 193–202.

Gaffling, S., Daum, V., Steidl, S., Maier, A., Köstler, H. and Hornegger, J. (2014). A Gauss-Seidel iteration scheme for reference-free 3-D histological image reconstruction. *IEEE Transactions on Medical Imaging* 34.2, 514–530.

Gecer, B., Aksoy, S., Mercan, E., Shapiro, L. G., Weaver, D. L. and Elmore, J. G. (2018). Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. *Pattern Recognition* 84, 345–356.

Gertych, A., Ing, N., Ma, Z., Fuchs, T. J., Salman, S., Mohanty, S., Bhele, S., Velásquez-Vacca, A., Amin, M. B. and Knudsen, B. S. (2015). Machine learning approaches to analyze histological images of tissues from radical prostatectomies. *Computerized Medical Imaging and Graphics* 46, 197–208.

Ghaznavi, F., Evans, A., Madabhushi, A. and Feldman, M. (2013). Digital imaging in pathology: whole-slide imaging and beyond. *Annual Review of Pathology: Mechanisms of Disease* 8, 331–359.

Gibson, E., Gaed, M., Gómez, J. A., Moussa, M., Romagnoli, C., Pautler, S., Chin, J. L., Crukley, C., Bauman, G. S., Fenster, A. and Ward, A. D. (2013a). 3D prostate histology reconstruction: An evaluation of image-based and fiducial-based algorithms. *Medical Physics* 40.9, 093501.

Gibson, E., Gaed, M., Gómez, J. A., Moussa, M., Pautler, S., Chin, J. L., Crukley, C., Bauman, G. S., Fenster, A. and Ward, A. D. (2013b). 3D prostate histology image reconstruction: quantifying the impact of tissue deformation and histology section location. *Journal of Pathology Informatics* 4.

Gonzales, R. C. and Woods, R. E. (2002). *Digital image processing*.

Goode, A., Gilbert, B., Harkes, J., Jukic, D. and Satyanarayanan, M. (2013). OpenSlide: A vendor-neutral software foundation for digital pathology. *Journal of Pathology Informatics* 4.1, 27.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2672–2680.

Greenspan, H., van Ginneken, B. and Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging* 35.5, 1153–1159.

Griffin, J. and Treanor, D. (2017). Digital pathology in clinical use: where are we now and what is holding us back?: *Histopathology* 70.1, 134–145.

Grönberg, H., Adolfsson, J., Aly, M., Nordström, T., Wiklund, P., Brandberg, Y., Thompson, J., Wiklund, F., Lindberg, J., Clements, M., Egevad, L. and Eklund, M. (2015). Prostate cancer screening in men aged 50-69 years (STHLM3): A prospective population-based diagnostic study. *The Lancet Oncology* 16.16, 1667–1676.

Grothausmann, R., Knudsen, L., Ochs, M. and Mühlfeld, C. (2017). Digital 3D reconstructions using histological serial sections of lung tissue including the alveolar capillary network. *American Journal of Physiology-Lung Cellular and Molecular Physiology* 312.2, L243–L257.

Gummeson, A., Arvidsson, I., Ohlsson, M., Overgaard, N. C., Krzyzanowska, A., Heyden, A., Bjartell, A. and Åström, K. (2017). Automatic Gleason grading of H and E stained microscopic prostate images using deep convolutional neural networks. *Medical Imaging 2017: Digital Pathology*. Vol. 10140, 101400S.

Gupta, A., Harrison, P. J., Wieslander, H., Pielawski, N., Kartasalo, K., Partel, G., Solorzano, L., Suveer, A., Klemm, A. H., Spjuth, O., Sintorn, I.-M. and Wählby, C. (2019). Deep Learning in Image Cytometry: A Review. *Cytometry Part A* 95.4, 366–380.

Gurcan, M. N., Boucheron, L., Can, A., Madabhushi, A., Rajpoot, N. and Yener, B. (2009). Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering* 2, 147.

Hamilton, P., Anderson, N., Bartels, P. and Thompson, D. (1994). Expert system support using Bayesian belief networks in the diagnosis of fine needle aspiration biopsy specimens of the breast. *Journal of Clinical Pathology* 47.4, 329–336.

Haralick, R. M., Shanmugam, K. and Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics* 6, 610–621.

Hartman, D. J., van der Laak, J. A., Gurcan, M. N. and Pantanowitz, L. (2020). Value of public challenges for the development of pathology deep learning algorithms. *Journal of Pathology Informatics* 11.

Hashimoto, N., Bautista, P. A., Yamaguchi, M., Ohyama, N. and Yagi, Y. (2012). Referenceless image quality evaluation for whole slide imaging. *Journal of Pathology Informatics* 3.

Haskins, G., Kruger, U. and Yan, P. (2020). Deep learning in medical image registration: A survey. *Machine Vision and Applications* 31.1, 8.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Heathfield, H., Bose, D. and Kirkham, N. (1991). Knowledge-based computer system to aid in the histopathological diagnosis of breast disease. *Journal of Clinical Pathology* 44.6, 502–508.

Hebb, D. O. (1949). *The organization of behavior: a neuropsychological theory*. J. Wiley; Chapman & Hall.

Hiner, M. C., Rueden, C. T. and Eliceiri, K. W. (2016). ImageJ-MATLAB: a bidirectional framework for scientific image analysis interoperability. *Bioinformatics* 33.4, 629–630.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation* 9.8, 1735–1780.

Hollon, T. C., Pandian, B., Adapa, A. R., Urias, E., Save, A. V., Khalsa, S. S. S., Eichberg, D. G., D'Amico, R. S., Farooq, Z. U., Lewis, S., Petridis, P. D., Marie, T., Shah, A. H., Garton, H. J., Maher, C. O., Heth, J. A., McKean, E. L., Sullivan, S. E., Hervey-Jumper, S. L., Patil, P. G., Thompson, B. G., Sagher, O., McKhann II, G. M., Komotar, R. J., Ivan, M. E., Snuderl, M., Otten, M. L., Johnson, T. D., Sisti, M. B., Bruce, J. N., Muraszko, K. M., Trautman, J., Freudiger, C. W., Canoll, P., Lee, H., Camelo-Piragua, S. and Orringer, D. A. (2020). Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nature Medicine* 26.1, 52–58.

Hovens, M. C., Lo, K., Kerger, M., Pedersen, J., Nottle, T., Kurganovs, N., Ryan, A., Peters, J. S., Moon, D., Costello, A. J., Corcoran, N. M. and Hong, M. K. (2017). 3D modelling of radical prostatectomy specimens: Developing a method to quantify tumor morphometry for prostate cancer risk prediction. *Pathology-Research and Practice* 213.12, 1523–1529.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint 1704.04861*.

Hughes, C., Rouviere, O., Mege-Lechevallier, F., Souchon, R. and Prost, R. (2012). Robust alignment of prostate histology slices with quantified accuracy. *IEEE Transactions on Biomedical Engineering* 60.2, 281–291.

Hutter, F., Kotthoff, L. and Vanschoren, J. (2019). *Automated Machine Learning*. Springer.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B. and Madry, A. (2019). Adversarial examples are not bugs, they are features. *arXiv preprint 1905.02175*.

Irshad, H., Veillard, A., Roux, L. and Racoceanu, D. (2013). Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential. *IEEE Reviews in Biomedical Engineering* 7, 97–114.

Jacquin, A. E. (1993). Fractal image coding: A review. *Proceedings of the IEEE* 81.10, 1451–1465.

Jafari-Khouzani, K. and Soltanian-Zadeh, H. (2003). Multiwavelet grading of pathological images of prostate. *IEEE Transactions on Biomedical Engineering* 50.6, 697–704.

Jain, M. S. and Massoud, T. F. (2020). Predicting tumour mutational burden from histopathological images using multiscale deep learning. *Nature Machine Intelligence* 2.6, 356–362.

Janowczyk, A., Basavanhally, A. and Madabhushi, A. (2017). Stain normalization using sparse autoencoders (StaNoSA): application to digital pathology. *Computerized Medical Imaging and Graphics* 57, 50–61.

Janowczyk, A. and Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics* 7.

Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M. and Madabhushi, A. (2019). HistoQC: an open-source quality control tool for digital pathology slides. *JCO Clinical Cancer Informatics* 3, 1–7.

Johnson, G. A., Badea, A., Brandenburg, J., Cofer, G., Fubara, B., Liu, S. and Nissanov, J. (2010). Waxholm space: an image-based reference for coordinating mouse brain research. *Neuroimage* 53.2, 365–372.

Ju, T., Warren, J., Carson, J., Bello, M., Kakadiaris, I., Chiu, W., Thaller, C. and Eichele, G. (2006). 3D volume reconstruction of a mouse brain from histological sections using warp filtering. *Journal of Neuroscience Methods* 156.1-2, 84–100.

Källén, H., Molin, J., Heyden, A., Lundström, C. and Åström, K. (2016). Towards grading gleason score using generically trained deep convolutional neural networks. *IEEE International Symposium on Biomedical Imaging*. Vol. 2016-June, 1163–1167.

Kay, P. A., Robb, R. A. and Bostwick, D. G. (1998). Prostate cancer microvessels: a novel method for three-dimensional reconstruction and analysis. *The Prostate* 37.4, 270–277.

Khan, A. M., Rajpoot, N., Treanor, D. and Magee, D. (2014). A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering* 61.6, 1729–1738.

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220.4598, 671–680.

Kohlberger, T., Liu, Y., Moran, M., Chen, P.-H. C., Brown, T., Hipp, J. D., Mermel, C. H. and Stumpe, M. C. (2019). Whole-slide image focus quality: Automatic assessment and impact on AI cancer detection. *Journal of Pathology Informatics* 10.

Koohbanani, N. A., Qaisar, T., Shaban, M., Gamper, J. and Rajpoot, N. (2018). Significance of Hyperparameter Optimization for Metastasis Detection in Breast Histology Images. *Computational Pathology and Ophthalmic Medical Image Analysis*. Springer, 139–147.

Koos, B., Kamali-Moghaddam, M., David, L., Sobrinho-Simões, M., Dimberg, A., Nilsson, M., Wählby, C. and Söderberg, O. (2015). Next-generation pathology - surveillance of tumor microecology. *Journal of Molecular Biology* 427.11, 2013–2022.

Kothari, S., Phan, J. H., Stokes, T. H. and Wang, M. D. (2013). Pathology imaging informatics for quantitative analysis of whole-slide images. *Journal of the American Medical Informatics Association* 20.6, 1099–1108.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105.

Lafarge, M., Pluim, J., Eppenhof, K. and Veta, M. (2019). Learning domain-invariant representations of histological images. *Frontiers in Medicine* 6, 162.

Laine, A. and Fan, J. (1993). Texture classification by wavelet packet signatures. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15.11, 1186–1191.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W. and Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications* 10.1, 1096.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature* 521.7553, 436–444.

Ledford, H. (2017). Cell atlases race to map the body. *Nature* 542.7642, 404–405.

Lee, T. S. (1996). Image representation using 2D Gabor wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18.10, 959–971.

Lein, E. S. et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445.7124, 168.

Leo, P., Elliott, R., Shih, N. N., Gupta, S., Feldman, M. and Madabhushi, A. (2018). Stable and discriminating features are predictive of cancer presence and Gleason grade in radical prostatectomy specimens: a multi-site study. *Scientific Reports* 8.1, 1–13.

Li, A., Gong, H., Zhang, B., Wang, Q., Yan, C., Wu, J., Liu, Q., Zeng, S. and Luo, Q. (2010). Micro-optical sectioning tomography to obtain a high-resolution atlas of the mouse brain. *Science* 330.6009, 1404–1408.

Li, J., Speier, W., Ho, K. C., Sarma, K. V., Gertych, A., Knudsen, B. S. and Arnold, C. W. (2018). An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies. *Computerized Medical Imaging and Graphics* 69, 125–133.

Liang, Y., Wang, F., Treanor, D., Magee, D., Teodoro, G., Zhu, Y. and Kong, J. (2015). Liver whole slide image analysis for 3D vessel reconstruction. *IEEE International Symposium on Biomedical Imaging*, 182–185.

Liimatainen, K., Latonen, L., Kartasalo, K. and Ruusuvuori, P. (2019). 3D-Printed Whole Prostate Models with Tumor Hotspots Using Dual-Extruder Printer. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2867–2871.

Liimatainen, K., Latonen, L., Valkonen, M., Kartasalo, K. and Ruusuvuori, P. (2020). Virtual reality for 3D histology: multi-scale visualization of organs with interactive feature exploration. *arXiv preprint 2003.11148*.

Linnainmaa, S. (1970). The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. *Master's Thesis (in Finnish), Univ. Helsinki*, 6–7.

Litjens, G., Bandi, P., Bejnordi, B. E., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., Manson, Q. F., Stathonikos, N., Baidoshvili, A., van Diest, P., Wauters, C., van Dijk, M. and van der Laak, J. (2018).

1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience* 7.6, giy065.

Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-van de Kaa, C., Bult, P., van Ginneken, B. and van der Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports* 6, 26286.

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A. and Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health* 1.6, e271–e297.

Lopez, X. M., D'Andrea, E., Barbot, P., Bridoux, A.-S., Rorive, S., Salmon, I., Debeir, O. and Decaestecker, C. (2013). An automated blur detection method for histological whole slide imaging. *PLOS One* 8.12.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60.2, 91–110.

Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., Schmitt, C. and Thomas, N. E. (2009). A method for normalizing histology slides for quantitative analysis. *IEEE International Symposium on Biomedical Imaging*, 1107–1110.

Magee, D., Song, Y., Gilbert, S., Roberts, N., Wijayathunga, N., Wilcox, R., Bulpitt, A. and Treanor, D. (2015). Histopathology in 3D: From three-dimensional reconstruction to multi-stain and multi-modal analysis. *Journal of Pathology Informatics* 6.

Magee, D., Treanor, D. and Quirke, P. (2008). A new image registration algorithm with application to 3D histopathology. *Microscopic Image Analysis with Applications in Biology. New York, NY*.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., Feldmann, C., Frangi, A. F., Full, P. M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B. A., März, K., Maier, O., Maier-Hein, K., Menze, B. H., Müller, H., Neher, P. F., Niessen, W., Rajpoot, N., Sharp, G. C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A. A., van der Sommen, F., Wang, C.-W., We-

ber, M.-A., Zheng, G., Jannin, P. and Kopp-Schneider, A. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications* 9.1, 5217.

Maintz, J. A. and Viergever, M. A. (1998). A survey of medical image registration. *Medical Image Analysis* 2.1, 1–36.

Malandain, G., Bardinet, E., Nelissen, K. and Vanduffel, W. (2004). Fusion of autoradiographs with an MR volume using 2-D and 3-D linear transformations. *Neuroimage* 23.1, 111–127.

Matas, J., Chum, O., Urban, M. and Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing* 22.10, 761–767.

Mayerich, D., Abbott, L. and McCormick, B. (2008). Knife-edge scanning microscopy for imaging and reconstruction of three-dimensional anatomical structures of the mouse brain. *Journal of Microscopy* 231.1, 134–143.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics* 5.4, 115–133.

Meijering, E., Carpenter, A. E., Peng, H., Hamprecht, F. A. and Olivo-Marin, J.-C. (2016). Imagining the future of bioimage analysis. *Nature Biotechnology* 34.12, 1250.

Merickel, M. (1988). 3D reconstruction: the registration problem. *Computer Vision, Graphics, and Image Processing* 42.2, 206–219.

Miao, S., Wang, Z. J. and Liao, R. (2016). A CNN regression approach for real-time 2D/3D registration. *IEEE Transactions on Medical Imaging* 35.5, 1352–1363.

Mignardi, M., Ishaq, O., Qian, X. and Wählby, C. (2016). Bridging histology and bioinformatics—computational analysis of spatially resolved transcriptomics. *Proceedings of the IEEE* 105.3, 530–541.

Min, E., Ban, S., Lee, J., Vavilin, A., Baek, S., Jung, S., Ahn, Y., Park, K., Shin, S., Han, S., Cho, H., Lee-Kwon, W., Kim, J., Lee, C. J. and Jung, W. (2020). Serial optical coherence microscopy for label-free volumetric histopathology. *Scientific Reports* 10.1, 1–8.

Moad, M., Hannezo, E., Buczacki, S. J., Wilson, L., El-Sherif, A., Sims, D., Pickard, R., Wright, N. A., Williamson, S. C., Turnbull, D. M., Taylor, R. W., Greaves, L., Robson, C. N., Simons, B. D. and Heer, R. (2017). Multipotent basal stem cells,

maintained in localized proximal niches, support directed long-ranging epithelial flows in human prostates. *Cell Reports* 20.7, 1609–1622.

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Velázquez Vega, J. E., Brat, D. J. and Cooper, L. A. D. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences* 115.13, E2970–E2979.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W. and Müller, K. R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*.

Mosquera-Lopez, C., Agaian, S., Velez-Hoyos, A. and Thompson, I. (2014). Computer-aided prostate cancer diagnosis from digitized histopathology: a review on texture-based systems. *IEEE Reviews in Biomedical Engineering* 8, 98–113.

Mukhopadhyay, S., Feldman, M. D., Abels, E., Ashfaq, R., Beltaifa, S., Cacciabeve, N. G., Cathro, H. P., Cheng, L., Cooper, K., Dickey, G. E., Gill, R. M., Heaton, R. P., Kerstens, R., Lindberg, G. M., Malhotra, R. K., Mandell, J. W., Manlucu, E. D., Mills, A. M., Mills, S. E., Moskaluk, C. A., Nelis, M., Patil, D. T., Przybycin, C. G., Reynolds, J. P., Rubin, B. P., Saboorian, M. H., Salicru, M., Samols, M. A., Sturgis, C. D., Turner, K. O., Wick, M. R., Yoon, J. Y., Zhao, P. and Taylor, C. R. (2018). Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (pivotal study). *The American Journal of Surgical Pathology* 42.1, 39.

Nagpal, K., Foote, D., Liu, Y., Chen, P.-H. C., Wulczyn, E., Tan, F., Olson, N., Smith, J. L., Mohtashamian, A., Wren, J. H., Corrado, G. S., MacDonald, R., Peng, L. H., Amin, M. B., Evans, A. J., Sangoi, A. R., Mermel, C. H., Hipp, J. D. and Stumpe, M. C. (2019). Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *npj Digital Medicine*.

Niazi, M. K. K., Yao, K., Zynger, D. L., Clinton, S. K., Chen, J., Koyutürk, M., LaFramboise, T. and Gurcan, M. (2016). Visually meaningful histopathological features for automatic grading of prostate cancer. *IEEE Journal of Biomedical and Health Informatics* 21.4, 1027–1038.

Niazi, M. K. K., Parwani, A. V. and Gurcan, M. N. (2019). Digital pathology and artificial intelligence. *The Lancet Oncology* 20.5, e253–e261.

Nir, G., Hor, S., Karimi, D., Fazli, L., Skinnider, B. F., Tavassoli, P., Turbin, D., Villamil, C. F., Wang, G., Wilson, R. S., Iczkowski, K. A., Lucia, M. S., Black, P. C.,

Abolmaesumi, P., Goldenberg, S. L. and Salcudean, S. E. (2018). Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Medical Image Analysis* 50, 167–180.

Nir, G., Karimi, D., Goldenberg, S. L., Fazli, L., Skinnider, B. F., Tavassoli, P., Turbin, D., Villamil, C. F., Wang, G., Thompson, D. J., Black, P. C. and Salcudean, S. E. (2019). Comparison of Artificial Intelligence Techniques to Evaluate Performance of a Classifier for Automatic Grading of Prostate Cancer From Digitized Histopathologic Images. *JAMA Network Open* 2.3, e190442.

Norton, K.-A., Namazi, S., Barnard, N., Fujibayashi, M., Bhanot, G., Ganesan, S., Iyatomi, H., Ogawa, K. and Shinbrot, T. (2012). Automated reconstruction algorithm for identification of 3D architectures of cribriform ductal carcinoma in situ. *PLOS One* 7.9.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G. and Ré, C. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proceedings of the ACM Conference on Health, Inference, and Learning*, 151–159.

Ojala, T., Pietikäinen, M. and Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 7, 971–987.

Oliveira, F. P. and Tavares, J. M. R. (2014). Medical image registration: a review. *Computer Methods in Biomechanics and Biomedical engineering* 17.2, 73–93.

Onozato, M. L., Hammond, S., Merren, M. and Yagi, Y. (2013). Evaluation of a completely automated tissue-sectioning machine for paraffin blocks. *Journal of Clinical Pathology* 66.2, 151–154.

Onozato, M. L., Klepeis, V. E., Yagi, Y. and Mino-Kenudson, M. (2012). A role of three-dimensional (3D)-reconstruction in the classification of lung adenocarcinoma. *Analytical Cellular Pathology* 35.2, 79–84.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* 9.1, 62–66.

Ourselin, S., Roche, A., Subsol, G., Pennec, X. and Ayache, N. (2001). Reconstructing a 3D structure from serial histological sections. *Image and Vision Computing* 19.1-2, 25–31.

Paish, E. C., Green, A. R., Rakha, E. A., Macmillan, R. D., Maddison, J. R. and Ellis, I. O. (2009). Three-dimensional reconstruction of sentinel lymph nodes with

metastatic breast cancer indicates three distinct patterns of tumour growth. *Journal of Clinical Pathology* 62.7, 617–623.

Palokangas, S., Selinummi, J. and Yli-Harja, O. (2007). Segmentation of folds in tissue section images. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 5641–5644.

Pantanowitz, L., Quiroga-Garza, G. M., Bien, L., Heled, R., Laifenfeld, D., Linhart, C., Sandbank, J., Shach, A. A., Shalev, V., Vecsler, M. et al. (2020). An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *The Lancet Digital Health* 2.8, e407–e416.

Pantanowitz, L., Sharma, A., Carter, A. B., Kurc, T., Sussman, A. and Saltz, J. (2018). Twenty years of digital pathology: An overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives. *Journal of Pathology Informatics* 9.

Pichat, J., Iglesias, J. E., Yousry, T., Ourselin, S. and Modat, M. (2018). A survey of methods for 3D histology reconstruction. *Medical Image Analysis* 46, 73–105.

Pietikäinen, M., Ojala, T. and Xu, Z. (2000). Rotation-invariant texture classification using feature distributions. *Pattern Recognition* 33.1, 43–52.

Prewitt, J. M. and Mendelsohn, M. L. (1966). The analysis of cell images. *Annals of the New York Academy of Sciences* 128.3, 1035–1053.

Ragan, T., Kadiri, L. R., Venkataraju, K. U., Bahlmann, K., Sutin, J., Taranda, J., Arganda-Carreras, I., Kim, Y., Seung, H. S. and Osten, P. (2012). Serial two-photon tomography for automated ex vivo mouse brain imaging. *Nature Methods* 9.3, 255.

Rajkomar, A., Dean, J. and Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine* 380.14, 1347–1358.

Real, E., Liang, C., So, D. R. and Le, Q. V. (2020). AutoML-Zero: Evolving Machine Learning Algorithms From Scratch. *arXiv preprint 2003.03384*.

Reynolds, H. M., Williams, S., Zhang, A., Chakravorty, R., Rawlinson, D., Ong, C. S., Esteva, M., Mitchell, C., Parameswaran, B., Finnegan, M., Liney, G. and Haworth, A. (2015). Development of a registration framework to validate MRI with histology for prostate focal therapy. *Medical Physics* 42.12, 7078–7089.

Rivenson, Y., Wang, H., Wei, Z., Haan, K. de, Zhang, Y., Wu, Y., Günaydın, H., Zuckerman, J. E., Chong, T., Sisk, A. E. et al. (2019). Virtual histological staining

of unlabelled tissue-autofluorescence images via deep learning. *Nature biomedical engineering* 3.6, 466–477.

Roberts, N., Magee, D., Song, Y., Brabazon, K., Shires, M., Crellin, D., Orsi, N. M., Quirke, R., Quirke, P. and Treanor, D. (2012). Toward routine use of 3D histopathology as a research tool. *The American Journal of Pathology* 180.5, 1835–1842.

Rohlfing, T. (2011). Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE Transactions on Medical Imaging* 31.2, 153–163.

Rojas, K. D., Montero, M. L., Yao, J., Messing, E., Fazili, A., Joseph, J., Ou, Y., Rubens, D. J., Parker, K. J., Davatzikos, C. and Castaneda, B. (2015). Methodology to study the three-dimensional spatial distribution of prostate cancer and their dependence on clinical parameters. *Journal of Medical Imaging* 2.3, 037502.

Ronneberger, O., Fischer, P. and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* 65.6, 386.

Roux, L., Racoceanu, D., Loménie, N., Kulikova, M., Irshad, H., Klossa, J., Capron, F., Genestie, C., Le Naour, G. and Gurcan, M. N. (2013). Mitosis detection in breast cancer histological images An ICPR 2012 contest. *Journal of Pathology Informatics* 4.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1.5, 206–215.

Ruifrok, A. C., Johnston, D. A. et al. (2001). Quantification of histochemical staining by color deconvolution. *Analytical and Quantitative Cytology and Histology* 23.4, 291–299.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323.6088, 533–536.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115.3, 211–252.

Ruusuvuori, P., Valkonen, M., Nykter, M., Visakorpi, T. and Latonen, L. (2016). Feature-based analysis of mouse prostatic intraepithelial neoplasia in histological tissue sections. *Journal of Pathology Informatics* 7.

Saalfeld, S., Fetter, R., Cardona, A. and Tomancak, P. (2012). Elastic volume reconstruction from series of ultra-thin microscopy sections. *Nature Methods* 9.7, 717.

Sadanandan, S. K., Ranefall, P., Le Guyader, S. and Wählby, C. (2017). Automated training of deep convolutional neural networks for cell segmentation. *Scientific Reports* 7.1, 1–7.

Sadanandan, S. K., Ranefall, P. and Wählby, C. (2016). Feature augmented deep neural networks for segmentation of cells. *European Conference on Computer Vision*, 231–243.

Sahiner, B., Chan, H.-P., Petrick, N., Wei, D., Helvie, M. A., Adler, D. D. and Goodsitt, M. M. (1996). Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. *IEEE Transactions on Medical Imaging* 15.5, 598–610.

Salisbury, J. R. and Whimster, W. F. (1993). Progress in computer-generated three-dimensional reconstruction. *The Journal of Pathology* 170.3, 223–227.

Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Vu, N., Samaras, D., Shroyer, K., Zhao, T., Batiste, R., van Arnam, J., Shmulevich, I., Rao, A., Lazar, A., Sharma, A. and Thorsson, V. (2018). Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Reports* 23.1, 181–193.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P. and Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nature Methods* 9.7, 676.

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

Schneider, C. A., Rasband, W. S. and Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* 9.7, 671.

Senaras, C., Niazi, M. K. K., Lozanski, G. and Gurcan, M. N. (2018). DeepFocus: Detection of out-of-focus regions in whole slide digital images using deep learning. *PLOS One* 13.10, e0205387.

Shaban, M. T., Baur, C., Navab, N. and Albarqouni, S. (2019). Staingan: Stain style transfer for digital histological images. *IEEE International Symposium on Biomedical Imaging*, 953–956.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. and De Freitas, N. (2015). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE* 104.1, 148–175.

Shojaii, R., Karavardanyan, T., Yaffe, M. and Martel, A. L. (2011). Validation of histology image registration. *International Society for Optics and Photonics Medical Imaging 2011: Image Processing*. Vol. 7962, 79621E.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint 1409.1556*.

Snoek, J., Larochelle, H. and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 2951–2959.

Solorzano, L., Partel, G. and Wählby, C. (2020). TissUUmaps: Interactive visualization of large-scale spatial gene expression and tissue morphology data. *Bioinformatics*.

Song, Y., Treanor, D., Bulpitt, A. J. and Magee, D. R. (2013). 3D reconstruction of multiple stained histology images. *Journal of Pathology Informatics* 4.Suppl.

Sorzano, C. O. S., Thévenaz, P. and Unser, M. (2005). Elastic registration of biological images using vector-spline regularization. *IEEE Transactions on Biomedical Engineering* 52.4, 652–663.

Sotiras, A., Davatzikos, C. and Paragios, N. (2013). Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging* 32.7, 1153.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J. and Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353.6294, 78–82.

Stille, M., Smith, E. J., Crum, W. R. and Modo, M. (2013). 3D reconstruction of 2D fluorescence histology images and registration with in vivo MR images: application in a rodent stroke model. *Journal of Neuroscience Methods* 219.1, 27–40.

Strubell, E., Ganesh, A. and McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650.

Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. *Thirty-First AAAI Conference on Artificial Intelligence*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016). *Rethinking the Inception Architecture for Computer Vision*.

Tabesh, A., Teverovskiy, M., Pang, H.-Y., Kumar, V. P., Verbel, D., Kotsianti, A. and Saidi, O. (2007). Multifeature prostate cancer diagnosis and Gleason grading of histological images. *IEEE Transactions on Medical Imaging* 26.10, 1366–1378.

Tan, M. and Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *International Conference on Machine Learning*, 6105–6114.

Tellez, D., Litjens, G., Bandi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F. and van der Laak, J. (2019). Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *arXiv preprint 1902.06543*.

Tellez, D., Litjens, G., van der Laak, J. and Ciompi, F. (2019). Neural Image Compression for Gigapixel Histopathology Image Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Teodoro, G., Kurç, T. M., Taveira, L. F., Melo, A. C., Gao, Y., Kong, J. and Saltz, J. H. (2016). Algorithm sensitivity analysis and parameter tuning for tissue image segmentation pipelines. *Bioinformatics* 33.7, 1064–1072.

Theart, R. P., Loos, B. and Niesler, T. R. (2017). Virtual reality assisted microscopy data visualization and colocalization analysis. *BMC bioinformatics* 18.2, 64.

Thevenaz, P., Ruttimann, U. E. and Unser, M. (1998). A pyramid approach to subpixel registration based on intensity. *IEEE Transactions on Image Processing* 7.1, 27–41.

Tolkach, Y., Thomann, S. and Kristiansen, G. (2018). Three-dimensional reconstruction of prostate cancer architecture with serial immunohistochemical sections: hallmarks of tumour growth, tumour compartmentalisation, and implications for grading and heterogeneity. *Histopathology* 72.6, 1051–1059.

Torres, R., Vesuna, S. and Levene, M. J. (2014). High-resolution, 2-and 3-dimensional imaging of uncut, unembedded tissue biopsy samples. *Archives of Pathology and Laboratory Medicine* 138.3, 395–402.

Tuominen, V. J. and Isola, J. (2010). Linking whole-slide microscope images with DICOM by using JPEG2000 interactive protocol. *Journal of Digital Imaging* 23.4, 454–462.

Turkki, R., Linder, N., Kovanen, P. E., Pellinen, T. and Lundin, J. (2016). Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples. *Journal of Pathology Informatics* 7.

Valkonen, M., Isola, J., Ylinen, O., Muhonen, V., Saxlin, A., Tolonen, T., Nykter, M. and Ruusuvuori, P. (2019). Cytokeratin-Supervised Deep Learning for Automatic Recognition of Epithelial Cells in Breast Cancers Stained for ER, PR, and Ki-67. *IEEE Transactions on Medical Imaging* 39.2, 534–542.

Valkonen, M., Kartasalo, K., Liimatainen, K., Nykter, M., Latonen, L. and Ruusuvuori, P. (2017). Dual structured convolutional neural network with feature augmentation for quantitative characterization of tissue histology. *IEEE International Conference on Computer Vision*, 27–35.

Valkonen, M., Ruusuvuori, P., Kartasalo, K., Nykter, M., Visakorpi, T. and Latonen, L. (2017). Analysis of spatial heterogeneity in normal epithelium and preneoplastic alterations in mouse prostate tumor models. *Scientific Reports* 7, 44831.

van der Laak, J., Ciompi, F. and Litjens, G. (2019). No pixel-level annotations needed. *Nature Biomedical Engineering* 3.11, 855–856.

van der Maaten, L. (2014). Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research* 15.1, 3221–3245.

van der Maaten, L. and Hinton, G. E. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* 9.Nov, 2579–2605.

van Leenders, G., van der Kwast, T., Grignon, D., Evans, A., Kristiansen, G., Kweldam, C., Litjens, G., McKenney, J. K., Melamed, J., Mottet, N., Paner, G., Samaratunga, H., Schoots, I., Simko, J., Tsuzuki, T., Varma, M., Warren, A., Wheeler, T., Williamson, S. and Iczkowski, K. (2020). The 2019 International Society of Urological Pathology (ISUP) Consensus Conference on Grading of Prostatic Carcinoma. *The American Journal of Surgical Pathology*.

van Royen, M. E., Verhoef, E. I., Kweldam, C. F., van Cappellen, W. A., Kremers, G.-J., Houtsmuller, A. B. and van Leenders, G. J. (2016). Three-dimensional microscopic analysis of clinical prostate specimens. *Histopathology* 69.6, 985–992.

Vestjens, J., Pepels, M., de Boer, M., Borm, G., van Deurzen, C., van Diest, P., van Dijck, J., Adang, E., Nortier, J., Rutgers, E., Seynaeve, C., Menke-Pluymers, M., Bult, P. and Tjan-Heijnen, V. (2012). Relevant impact of central pathology review on nodal classification in individual breast cancer patients. *Annals of Oncology* 23.10, 2561–2566.

Veta, M., van Diest, P. J., Willems, S. M., Wang, H., Madabhushi, A., Cruz-Roa, A., Gonzalez, F., Larsen, A. B., Vestergaard, J. S., Dahl, A. B., Cireşan, D. C., Schmidhuber, J., Giusti, A., Gambardella, L. M., Tek, F. B., Walter, T., Wang, C.-W., Kondo, S., Matuszewski, B. J., Precioso, F., Snell, V., Kittler, J., Campos, T. E. de, Khan, A. M., Rajpoot, N. M., Arkoumani, E., Lacle, M. M., Viergever, M. A. and Pluim, J. P. (2015). Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical Image Analysis* 20.1, 237–248.

Viergever, M. A., Maintz, J. A., Klein, S., Murphy, K., Staring, M. and Pluim, J. P. (2016). A survey of medical image registration – under review. *Medical Image Analysis* 33, 140–144.

Wang, D., Foran, D. J., Ren, J., Zhong, H., Kim, I. Y. and Qi, X. (2015). Exploring automatic prostate histopathology image gleason grading via local structure modeling. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2649–2652.

Wang, H., Viswanath, S. and Madabhushi, A. (2017). Discriminative scale learning (DiScrn): Applications to prostate cancer detection from MRI and needle biopsies. *Scientific Reports* 7.1, 1–13.

Werbos, P. J. (1982). Applications of advances in nonlinear sensitivity analysis. *System Modeling and Optimization*. Springer, 762–770.

Wetzel, A. W., Crowley, R., Kim, S., Dawson, R., Zheng, L., Joo, Y., Yagi, Y., Gilbertson, J., Gadd, C., Deerfield, D. and Becich, M. J. (1999). Evaluation of prostate tumor grades by content-based image retrieval. *27th AIPR Workshop: Advances in Computer-Assisted Recognition*. Vol. 3584, 244–252.

Wiegand, T., Krishnamurthy, R., Kuglitsch, M., Lee, N., Pujari, S., Salathé, M., Wenzel, M. and Xu, S. (2019). WHO and ITU establish benchmarking process for artificial intelligence in health. *The Lancet* 394.10192, 9–11.

Xu, Y., Pickering, J. G., Nong, Z., Gibson, E., Arpino, J.-M., Yin, H. and Ward, A. D. (2015). A method for 3D histopathology reconstruction supporting mouse microvasculature analysis. *PLOS One* 10.5, e0126817.

Yamamoto, Y., Tsuzuki, T., Akatsuka, J., Ueki, M., Morikawa, H., Numata, Y., Takahara, T., Tsuyuki, T., Tsutsumi, K., Nakazawa, R., Shimizu, A., Maeda, I., Tsuchiya, S., Kanno, H., Kondo, Y., Fukumoto, M., Tamiya, G., Ueda, N. and Kimura, G. (2019). Automated acquisition of explainable knowledge from unannotated histopathology images. *Nature Communications* 10.1, 1–9.

Yu, K.-H., Zhang, C., Berry, G. J., Altman, R. B., Ré, C., Rubin, D. L. and Snyder, M. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications* 7, 12474.

Zanjani, F. G., Zinger, S., Piepers, B., Mahmoudpour, S., Schelkens, P. and de With, P. (2019). Impact of JPEG 2000 compression on deep convolutional neural networks for metastatic cancer detection in histopathological images. *Journal of Medical Imaging* 6.2, 027501.

Zhu, J.-Y., Park, T., Isola, P. and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *IEEE international Conference on Computer Vision*, 2223–2232.

Zitova, B. and Flusser, J. (2003). Image registration methods: a survey. *Image and Vision Computing* 21.11, 977–1000.

PUBLICATIONS

PUBLICATION

I

Benchmarking of algorithms for 3D tissue reconstruction

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

IEEE International Conference on Image Processing. Ed. by 2016, 2360–2364

BENCHMARKING OF ALGORITHMS FOR 3D TISSUE RECONSTRUCTION

Kimmo Kartasalo[1,2,3], *Leena Latonen*[2,3,4], *Tapio Visakorpi*[2,3,4], *Matti Nykter*[1,2,3], *Pekka Ruusuvuori*[2,3,5]

[1]Department of Signal Processing, Tampere University of Technology, BioMediTech, Tampere, Finland
[2]University of Tampere, BioMediTech, Tampere, Finland
[3]Prostate Cancer Research Center, University of Tampere, Tampere, Finland
[4]Fimlab Laboratories, Tampere University Hospital, Tampere, Finland
[5]Tampere University of Technology, Pori, Finland

ABSTRACT

Studying tissue structure in 3D is beneficial in many applications. Reconstructing the structure based on histological sections has the advantages of high resolution and compatibility with conventional staining and interpretation techniques. However, obtaining an accurate 3D reconstruction based on a sequence of 2D sections is a difficult task. Evaluating the accuracy of such reconstructions is also challenging and it is often performed based only on visual inspections or a single indirect numerical measure. Here, we present a benchmarking framework composed of a panel of complementary metrics for assessing the quality of 3D reconstructions. We then apply the framework to evaluate the performance of several popular image registration algorithms in this context.

***Index Terms*—** Image registration, 3D reconstruction, histology, digital pathology, benchmark

1. INTRODUCTION

Studying tissues in 3D at the microscopic scale can provide new insights into many normal and pathological processes [1, 2]. Visualizing tissue directly in 3D is possible using techniques such as magnetic resonance imaging. However, conventional histology based on light microscopy has the important advantages of higher resolution and the possibility of applying techniques such as immunohistochemistry or *in situ* hybridization. Reconstructing the original 3D structure from a sequence of 2D histological images combines these advantages with the capability to examine the tissue in 3D.

The reconstruction process typically consists of a series of pairwise image registration operations to bring the sequence of images into alignment [2, 3]. This is generally a difficult task due to anatomical and technical variation from image to image. Evaluating the quality of the obtained reconstruction is equally challenging, especially in the absence of ground truth data concerning the true 3D structure of the sample. The evaluation is often based only on indirect measures of registration accuracy, which can produce misleading results [4].

In this paper, we demonstrate the use of a collection of measures for evaluating the quality of 3D tissue reconstructions. We constructed a reference least-squares solution based on manually selected point pairs (REF), compared to which we evaluated a number of popular algorithms representing the two main approaches to image registration [5]: area-based registration via optimization of mean squared error (MSE) or mutual information (MI), and registration using SIFT [6] or SURF [7] features coupled with model fitting by Random Sample Consensus (RANSAC) [8].

2. MATERIALS AND METHODS

2.1. Material

A prostate of a 14 month old male FVB/N mouse was fixed in PAXgene molecular fixative (PreAnalytiX GmbH, Hombrechtikon, Switzerland) according to manufacturer's recommendations, and embedded in paraffin. The tissue block was sectioned through, with 3x3 5μm sections used for hematoxylin-eosin (HE) staining, and every 10th section saved for other purposes. The slides were scanned with a Zeiss Axioskop40 microscope (Carl Zeiss MicroImaging, NY, USA) with 20x objective, a CCD color camera (QICAM Fast; QImaging, Canada) and a motorized specimen stage (Märzhäuser Wetzlar GmbH, Germany). Image acquisition was controlled by the Surveyor imaging system (Objective Imaging, UK). The pixel size was 0.46 μm.

2.2. Dataset preprocessing

Uncompressed bitmap output was converted by JVSdicom Compressor to JPEG2000 WSI format [9] and further processed using MATLAB R2015a (The MathWorks Inc., Natick, MA, USA). The dataset contained 260 images with one tissue section per image. Four corresponding points located

This work was supported by Academy of Finland (269474), Tekes – The Finnish Funding Agency for Innovation(269/31/2015), Cancer Society of Finland, Sigrid Juselius Foundation and Doctoral Programme of Computing and Electrical Engineering, Tampere University of Technology.

preferably at the centers of bisected nuclei were manually selected from each pair of adjacent sections. From each image, manual delineation of the tissue section was also performed, while excluding the background and torn-off pieces of tissue.

Variation in image appearance was compensated for by using histogram matching separately for each color channel based on a selected reference image [10]. The background pixels in each image were then set to the mean value of the tissue pixels, eliminating the effect of non-tissue pixels during the registration process. The images were subsampled using the JPEG2000 wavelet decomposition scheme to obtain a pixel size of 7.36 μm, which is close to the diameter of nuclei. This leads to the exclusion of small subcellular details which are not present on multiple slices and are not useful for registration. The images were converted to grayscale format following the NTSC standard by computing the weighted sum of the RGB components: 0.2989R+0.5870G+0.1140B.

2.3. Three-dimensional reconstruction

A 3D reconstruction of the sample was formed by serially registering each pair of consecutive sections via affine transformations, starting from the first section in image I_1. The pairwise transformations were estimated as described in Sections 2.4 - 2.6. The quality of the first image was verified visually. Let T_i denote the pairwise affine transformation that warps the image I_i to the image of the neighboring slice, I_{i-1}. The overall transformation for image I_i was obtained by concatenating the pairwise transformations as follows:

$$T_i^* = T_1 \circ T_2 \circ \cdots \circ T_{i-1} \circ T_i \quad (1)$$

where $\circ$ is composition. The transformations were applied via bilinear interpolation using MATLAB's *imwarp* function.

2.4. Least-squares image registration

An optimal reference reconstruction in the least-squares sense was formed by fitting an affine transformation to the manually selected points for each pair of images. Since the points represent mostly bisected nuclei appearing on only two consecutive sections, the reference reconstruction is in principle unaffected by the accumulation of errors over multiple sections [11].

2.5. Area-based image registration

Area-based registration was performed using MATLAB's *imregtform* function. A regular step gradient descent optimization algorithm was used to optimize the value of either the MSE or MI metric. As area-based techniques require an initial transformation which is close to the correct local optimum, we used a simple translation for initialization. The translation was computed as the displacement of the centroid of the tissue region in image I_{i-1} compared to image I_i.

We used the following parameters: number of multi-resolution pyramid levels 5, gradient magnitude tolerance 10^{-5}, minimum step length 10^{-6}, maximum step length 10^{-3}, maximum number of iterations 1000 and relaxation factors of 0.1 for MSE and 0.9 for MI. The number of pyramid levels was the largest possible given the image resolution. Maximum step length was chosen to be as large as possible without causing divergence. The gradient magnitude tolerance was set so low and the maximum number of iterations so high that convergence was essentially controlled by the minimum step length only. The minimum step length and relaxation factor were chosen based on an evaluation of values ranging from 10^{-8} to 10^{-2} and from 0.1 to 0.9, respectively.

2.6. Feature-based image registration

Feature-based registration was performed by computing SIFT and SURF keypoints for each image pair using the implementations in the VLfeat package [12] and the Image Alignment Toolbox (IAT) [13], respectively. Corresponding keypoints were established using the algorithm suggested by Lowe [6] and implemented in the VLfeat function *vl_ubcmatch*. An affine transformation was fitted to the keypoints using RANSAC, implemented in the IAT function *iat_ransac*.

We used default parameters for computing the SIFT features and a descriptor length of 64 for the SURF features. A minimum ratio of 1.25 between the distances to second closest and closest matches was used. For RANSAC, we used the following parameters: maximum number of iterations 100 000, probability to pick a minimum sample set with no outliers 0.99, maximum number of invalid set picks 1000 and maximum error tolerances of 0.01 for SIFT and 0.06 for SURF. The minimum distance ratio and the maximum error tolerances were chosen based on an evaluation of values ranging from 1.125 to 3 and from 0.001 to 1, respectively.

2.7. Evaluation of Target Registration Error

Pairwise target registration error (TRE) was quantified for each point and pair of images as the Euclidean distance between the location of manually selected point *j* in the image I_{i-1} and the location of the corresponding point in image I_i after applying the transformation T_i [14].

In addition to the pairwise errors, we quantified the accumulated error relative to the reference reconstruction [11]. Accumulated target registration error (ATRE) was quantified for each manually selected point *j* in each image I_i as the Euclidean distance obtained by comparing the point's location after applying either the overall reference transformation ${T^*}_{i,REF}$ or the overall transformation ${T^*}_i$ under evaluation.

2.8. Evaluation of tissue overlap

Let *A* denote the set of pixels belonging to the tissue region of image I_{i-1} and *B* denote the set of pixels belonging to the

tissue region of image I_i after the transformation T_i has been applied. Tissue overlap was quantified for each image pair as the Jaccard index [4], which is defined for the two sets of pixels, A and B, as

$$J_{A,B} = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

Due to anatomical and technical differences between consecutive sections, a perfect overlap is usually not the target of registration and Jaccard index alone is not a reliable measure of accuracy [4]. However, a low value can still be indicative of a poor registration result. The Jaccard index can also be considered a quality measure for the pixel-wise metrics described in Section 2.9, as computing them based on a small number of overlapping pixels can provide misleading results.

2.9. Evaluation of pixel-wise similarity

For each pair of images, we evaluated the similarity of corresponding pixels of image I_{i-1} and image I_i after applying the transformation T_i to the latter. The following measures were computed: root mean squared error (RMSE), normalized cross correlation (NCC), mutual information (MI) and normalized mutual information (NMI) [15]. We only considered overlapping tissue pixels, that is, the set of pixels $A \cap B$. Post-registration similarity of corresponding pixels is frequently used to evaluate registration accuracy, even though this indirect approach has been shown to be unreliable [4]. On the other hand, the more direct TRE measure is typically computed based only on a small set of landmark points. Pixel-wise measures provide at least some indirect information concerning regions located far from the landmarks.

2.10. Evaluation of reconstruction smoothness

The smoothness of the reconstructed volume was quantified using contrast and correlation features computed based on gray-level co-occurrence matrices (GLCMs) [16]. After applying the overall transformations to all images, we computed the GLCM for each pair of registered images by again considering only the set of overlapping tissue pixels $A \cap B$. The pixel values were not quantized, that is, we computed 256x256 GLCMs for our 8-bit images. The GLCM for the whole volume was obtained by summing the pairwise GLCMs. This approach is equivalent to computing the GLCM along the direction going across slices using a distance of 1, as suggested earlier [2, 3, 17]. Based on the combined GLCM, we computed the contrast feature f_2 and the correlation feature f_3. The rationale behind using these measures is that pixel values should change slowly when moving from slice to slice through the reconstructed volume [2, 3, 17]. A smooth reconstruction should thus exhibit low contrast and high correlation. In addition, we adopted the usual practice of visualizing cross-sectional views of the volume.

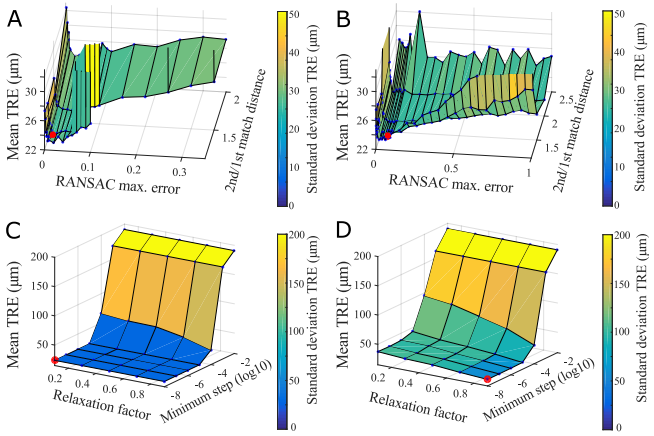

Fig. 1. Mean (surface height) and standard deviation (color) of TRE for SIFT (a), SURF (b), MSE (c) and MI (d) with different combinations of the RANSAC maximum error and minimum 2nd/1st match distance (a, b) or the relaxation factor and minimum step length (c, d). Minima of the mean TRE are marked with red dots. Parameter combinations producing failed or highly erroneous reconstructions have been omitted.

3. RESULTS AND DISCUSSION

3.1. Sensitivity to parameter selection

The mean and standard deviation of the TRE computed over all images and points are shown in Fig. 1 for different methods and combinations of parameter values. For SIFT and SURF, a low threshold of 1.25 for the second closest/closest match distance ratio coupled with a strict RANSAC maximum error tolerance (0.01 for SIFT, 0.06 for SURF) was optimal in this experiment. This produces a large number of putative matches, which are then strictly filtered for inliers by RANSAC. SURF produced satisfactory results for most parameter combinations, while SIFT failed to estimate valid transformations for many combinations with the RANSAC maximum error tolerance exceeding 0.3 and/or the distance ratio exceeding 2. Distance ratios over 2.5 also caused SURF to fail due to an insufficient number of putative matches fulfilling this criterion in some image pairs.

In the case of the MSE and MI methods, the error decreased rapidly as the value of the minimum step length was lowered down to 10^{-6}. Lowering the value further had only a negligible effect on the error while greatly increasing computation time and we therefore chose the value of 10^{-6}. Tuning the relaxation factor had a less dramatic effect on the results, but the more stable convergence obtained with higher values was beneficial for the MI method. This effect was not observed with the MSE method. Overall, the effect of tuning the parameters of the MSE and MI methods was more predictable than in the case of SIFT and SURF, which is certainly desirable when using the algorithms in practice.

Table 1. Pairwise errors (mean ± std, n = 259) for different methods and metrics. TRE is given in μm.

	REF	MSE	MI	SIFT	SURF
TRE	15.19±16.00	23.87±30.53	26.20±42.51	22.72±26.55	23.07±25.23
Jaccard	0.97±0.02	0.97±0.02	0.97±0.02	0.97±0.02	0.97±0.02
RMSE	47.46±5.48	45.31±5.43	45.58±5.74	46.03±5.58	46.02±5.56
NCC	0.53±0.11	0.57±0.10	0.56±0.11	0.56±0.10	0.56±0.10
MI	0.45±0.13	0.51±0.13	0.50±0.14	0.49±0.14	0.49±0.13
NMI	1.03±0.01	1.04±0.01	1.04±0.01	1.04±0.01	1.04±0.01

3.2. Pairwise errors

Values of pairwise metrics computed over all image pairs are shown in Table 1. According to the pixel-wise metrics, there are hardly any differences between the automated methods. Interestingly, REF had the worst pixel-wise values but the lowest mean TRE. The latter is not surprising, as the same landmarks are used by the REF method and for evaluating TRE. The nonzero TRE values of the REF method represent residual errors which cannot be corrected using a global affine model. The residual errors and the high standard deviations of TRE are probably caused by local deformations or errors in landmark selection. The contradicting TRE and pixel-wise metrics hint towards "overfitting" of pixel values by the automated methods but could also be explained by limitations of the reference such as relying on only four points per section.

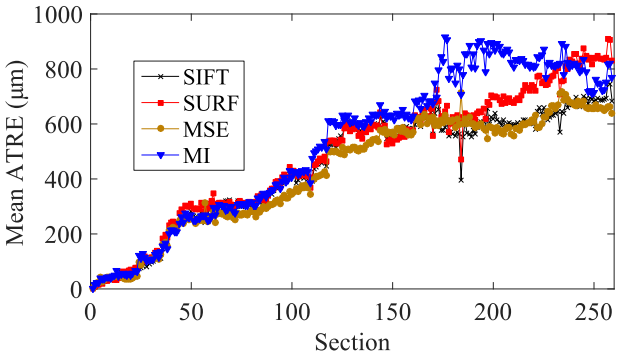

Fig. 2. Mean ATRE by section for different methods.

3.3. Accumulated error

Mean ATRE of the four landmarks for each pair of sections is visualized in Fig. 2 for each method. Accumulation of errors is also manifested in cross-sectional views of the volume in Fig. 3 as the distortion of structures relative to the reference reconstruction when proceeding from the beginning of the stack towards the end. However, it is important to note that cross-sections cannot capture the true 3D distortion of the volume. ATRE, on the other hand, is computed based on landmarks in different parts of the tissue and thus reflects the degree of distortion in 3D. In terms of ATRE, differences between the methods started to arise in the second half of the stack with MSE and SIFT outperforming MI and SURF.

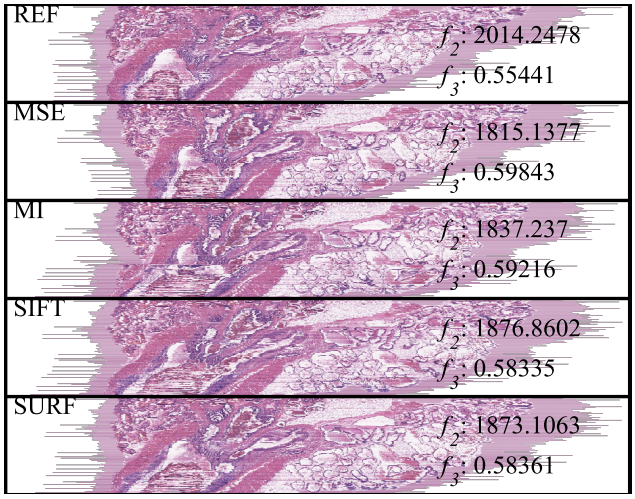

Fig. 3. Cross sections of volumes reconstructed using different methods. The top row in each cross section corresponds to I_1 and the bottom row corresponds to I_{260}. The contrast (f_2) and correlation (f_3) measures are shown for each case.

3.4. Reconstruction smoothness

Results concerning reconstruction smoothness are shown in Fig. 3 for each method. It is again important to note that the visual examination is limited to a single cross-sectional slice and also that the visual impression is dominated by salient structures. In contrast, the numerical measures are computed based on the entire volume, complementing the visual evaluation. Interestingly, the automated methods outperformed the REF method in terms of the contrast and correlation values. Especially in the case of the best-performing MSE method, this was accompanied by some visible straightening of curved structures. Failure to differentiate between the desired straightening of jagged edges and this type of "over-smoothing" might be a shortcoming of these metrics.

4. CONCLUSIONS

We developed a framework for assessing the quality of 3D tissue reconstructions using a panel of metrics and demonstrated it by comparing several algorithms. An evaluation of different parameter choices indicated that their effect on the results can be substantial and should not be neglected. Contradictions between different metrics underlined the necessity of using multiple complementary metrics. In the future, the framework will allow us to perform a comprehensive comparison of different approaches to 3D reconstruction while extending our dataset with other tissue specimens. The computational test bench and the associated data will be made available to the scientific community to support the evaluation of future algorithms. It will also be crucial to develop methods for obtaining ground truth data in an automated and reliable manner.

5. REFERENCES

[1] N. Roberts, D. Magee, Y. Song, K. Brabazon, M. Shires, D. Crellin, N. M. Orsi, R. Quirke, P. Quirke, and D. Treanor, "Toward routine use of 3d histopathology as a research tool," *Am. J. Pathol*, vol. 180, no. 5, pp. 1835–1842, 2012.

[2] S. Gaffling, V. Daum, S. Steidl, A. Maier, H. Kostler, and J. Hornegger, "A gauss-seidel iteration scheme for reference-free 3-d histological image reconstruction," *IEEE Trans. Med. Imaging*, vol. 34, no. 2, pp. 514–530, 2015.

[3] A. Cifor, L. Bai, and A. Pitiot, "Smoothness-guided 3-d reconstruction of 2-d histological images," *NeuroImage*, vol. 56, no. 1, pp. 197–211, 2011.

[4] T. Rohlfing, "Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable," *IEEE Trans. Med. Imag.*, vol. 31, no. 2, pp. 153–163, 2012.

[5] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image Vision Comput.*, vol. 21, no. 11, pp. 977–1000, 2003.

[6] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision.*, vol. 60, no. 2, pp. 91–110, 2004.

[7] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Comput. Vision Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[8] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[9] V. Tuominen and J. Isola, "Linking whole-slide microscope images with dicom by using jpeg2000 interactive protocol," *J. Digit. Imaging.*, vol. 23, no. 4, pp. 454–462, 2010.

[10] R. Gonzalez and R. Woods, *Digital Image Processing*, Addison-Wesley Longman Publishing Co., Inc, Boston, MA, USA, 2nd edition, 2001.

[11] Y. Xu, J. G. Pickering, Z. Nong, E. Gibson, J. M. Arpino, H. Yin, and A. D. Ward, "A method for 3d histopathology reconstruction supporting mouse microvasculature analysis," *PLoS One*, vol. 10, no. 5, pp. e0126817, 2015.

[12] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proc. ACM Int. Conf. Multimed.* 2010, pp. 1469–1472, ACM.

[13] G. Evangelidis, "Iat: A matlab toolbox for image alignment," 2013.

[14] J. M. Fitzpatrick, J. B. West, and C. R. Maurer Jr., "Predicting error in rigid-body point-based registration," *IEEE Trans. Med. Imag.*, vol. 17, no. 5, pp. 694–702, 1998.

[15] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3d medical image alignment," *Pattern Recognit.*, vol. 32, no. 1, pp. 71–86, 1999.

[16] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man. Cybern.*, vol. SMC-3, no. 6, pp. 610–621, 1973.

[17] S. Baheerathan, F. Albregtsen, and H.E. Danielsen, "Registration of serial sections of mouse liver cell nuclei," *J. Microsc.*, vol. 192, no. 1, pp. 37–53, 1998.

PUBLICATION

II

Comparative analysis of tissue reconstruction algorithms for 3D histology

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

Bioinformatics 34.17 (2018), 3013–3021

Bioinformatics, 34(17), 2018, 3013–3021
doi: 10.1093/bioinformatics/bty210
Advance Access Publication Date: 19 April 2018
Original Paper

Bioimage informatics

Comparative analysis of tissue reconstruction algorithms for 3D histology

Kimmo Kartasalo[1,2,3], Leena Latonen[1,3,4], Jorma Vihinen[5], Tapio Visakorpi[1,3,4], Matti Nykter[1,2,3] and Pekka Ruusuvuori[1,3,6,*]

[1]Faculty of Medicine and Life Sciences, University of Tampere, Tampere 33014, Finland, [2]Faculty of Biomedical Sciences and Engineering, Tampere University of Technology, Tampere 33101, Finland, [3]BioMediTech Institute, Tampere 33014, Finland, [4]Fimlab Laboratories, Tampere University Hospital, Tampere 33101, Finland, [5]Faculty of Engineering Sciences and [6]Faculty of Computing and Electrical Engineering, Tampere University of Technology, Tampere 33101, Finland

*To whom correspondence should be addressed.
Associate Editor: Robert Murphy

Received on November 29, 2017; revised on March 1, 2018; editorial decision on March 24, 2018; accepted on April 18, 2018

Abstract

Motivation: Digital pathology enables new approaches that expand beyond storage, visualization or analysis of histological samples in digital format. One novel opportunity is 3D histology, where a three-dimensional reconstruction of the sample is formed computationally based on serial tissue sections. This allows examining tissue architecture in 3D, for example, for diagnostic purposes. Importantly, 3D histology enables joint mapping of cellular morphology with spatially resolved omics data in the true 3D context of the tissue at microscopic resolution. Several algorithms have been proposed for the reconstruction task, but a quantitative comparison of their accuracy is lacking.
Results: We developed a benchmarking framework to evaluate the accuracy of several free and commercial 3D reconstruction methods using two whole slide image datasets. The results provide a solid basis for further development and application of 3D histology algorithms and indicate that methods capable of compensating for local tissue deformation are superior to simpler approaches.
Availability and implementation: Code: https://github.com/BioimageInformaticsTampere/RegBenchmark. Whole slide image datasets: http://urn.fi/urn:nbn:fi:csc-kata20170705131652639702.
Contact: pekka.ruusuvuori@tut.fi
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Digitalization of pathology has been accelerated by improvements in technology allowing acquisition of whole slide images (WSI) (Ghaznavi *et al.*, 2013; Griffin and Treanor, 2017). Besides computer-aided facilitation of pathologists' tasks, digital pathology can enable new approaches like 3D histology, where three-dimensional reconstructions of samples are formed *in silico* based on serial sections (Magee *et al.*, 2015; Roberts *et al.*, 2012). While other techniques allow imaging directly in 3D, they are currently incapable of matching the subcellular resolution and throughput of whole slide imaging. Examples of potential applications include construction of data-driven computer models and improved diagnostics of diseases associated with changes in the 3D microarchitecture of tissue. Moreover, 3D histology is compatible with established histopathological interpretation techniques and biochemical assays such as immunohistochemistry or *in situ* hybridization. This raises interesting prospects in view of recent advances in spatially resolved omics (Mignardi *et al.*, 2017; Ståhl *et al.*, 2016). Pairing imaging with genomic, epigenomic, transcriptomic and proteomic data in the spatial context of tissue holds great promise for pathology and other fields (Koos *et al.*, 2015). Taking a step further, this could be performed in 3D to truly probe the relationships between structural and functional features as well as the heterogeneity and interplay between different cell types in tumors, and significant projects are

now pursuing these goals (Ledford, 2017; Rusk, 2016). These kind of approaches have already led to the creation of brain atlases (Amunts *et al.*, 2013; Johnson *et al.*, 2010; Lein *et al.*, 2007). Such high-dimensional data also represent an exciting challenge for new ways of scientific visualization based e.g. on virtual reality techniques (Calì *et al.*, 2016; Ledford, 2017; Theart *et al.*, 2017).

Despite earlier computational and image acquisition bottlenecks (Roberts *et al.*, 2012), several algorithmic 3D histology solutions were already proposed before the recent developments in digital pathology (Ju *et al.*, 2006; Wang *et al.*, 2015). The key methodological problem is how to accurately register a sequence of 2D images to produce a 3D volume. Simply stacking the images does not result in a coherent volume due to differences between the relative locations and rotation angles of the sections and tissue deformations introduced during embedding and sectioning (Gibson *et al.*, 2013). Algorithms for image registration (Sotiras *et al.*, 2013) constitute the methodological basis of 3D histology. These algorithms are used to sequentially register each image with its neighbors to bring the entire series into alignment (Magee *et al.*, 2015; Wang *et al.*, 2015). Registration is accomplished by estimating transformations relating the images. Rigid transformations only allow translation and rotation of the entire image, while affine transformations are additionally able to model anisotropic scaling. Locally varying transformations, also called elastic models, can compensate for deformations on a local scale. Considering several nearby sections together (Saalfeld *et al.*, 2012) or applying regularization may be needed to obtain smooth, continuous 3D volumes (Casero *et al.*, 2017; Cifor *et al.*, 2011; Gaffling *et al.*, 2015; Ju *et al.*, 2006). After estimating the transformations, they need to be applied to the images via interpolation, which is possibly followed by postprocessing such as 3D visualization. Our focus is on the reconstruction step, which is usually the most difficult and crucial part of the image processing chain. Numerous approaches have been reported, relying on manual alignment (Onozato *et al.*, 2012; Paish *et al.*, 2009), semi-automatic methods using artificial landmarks (Hughes *et al.*, 2013; Rojas *et al.*, 2015) and automated algorithms (Arganda-Carreras *et al.*, 2010; Braumann *et al.*, 2005; Casero *et al.*, 2017; Cifor *et al.*, 2011; Ju *et al.*, 2006; Magee *et al.*, 2015; Saalfeld *et al.*, 2012; Song *et al.*, 2013; Stille *et al.*, 2013; Xu *et al.*, 2015).

Despite the widely acknowledged need for objective assessment of algorithms (Meijering *et al.*, 2016), an evaluation of modern computational methodology for 3D histology is lacking. Moreover, the common practice of relying only on visual inspections or a single indirect metric is insufficient (Rohlfing, 2012). The previous comparison of algorithms was published a decade ago and only included three basic approaches (Beare *et al.*, 2008). We have previously demonstrated a framework (Kartasalo *et al.*, 2016) based on a panel of indirect metrics and manually annotated landmarks allowing direct quantification of reconstruction accuracy (Rohlfing, 2012). In this study, we applied an extended version of the framework (see Fig. 1) to address the problem of comparing algorithms for 3D histology. As the basis of our evaluation, we used two WSI datasets representing two different tissue types. One obstacle complicating both the application and fair comparison of most algorithms is sensitivity to various settings or hyperparameters, which typically have to be selected by the user based on rules of thumb and tuned via trial and error. Encouraged by their recent application in the context of digital pathology, we employed automated hyperparameter selection methods to adjust tunable parameters (Shahriari *et al.*, 2016; Teodoro *et al.*, 2017).

As a baseline, we evaluated three basic methods: a least-squares fit to landmarks (LS), an optimization-based approach (OPT) and a method based on the Scale Invariant Feature Transform (SIFT) (Lowe, 2004). More advanced methods included the Fiji/ImageJ (Schindelin *et al.*, 2012; Schneider *et al.*, 2012) plugins HyperStackReg (HSR), which is an extension of StackReg (Thevenaz *et al.*, 1998), RegisterVirtualStackSlices (RVSS), which is based on bUnwarpJ (Arganda-Carreras *et al.*, 2006), and ElasticStackAlignment (ESA) (Saalfeld *et al.*, 2012), which is part of the TrakEM2 package (Cardona *et al.*, 2012). In addition, we evaluated two commercial tools: Medical Image Manager (MIM) (HeteroGenius Ltd, Leeds, UK) and Voloom (microDimensions GmbH, Munich, Germany). While LS, OPT, SIFT and HSR are based on global transformations, RVSS, ESA, MIM and Voloom use elastic models which make it possible to account for local tissue deformations. For a summary of the evaluated tools, see Supplementary Table S1.

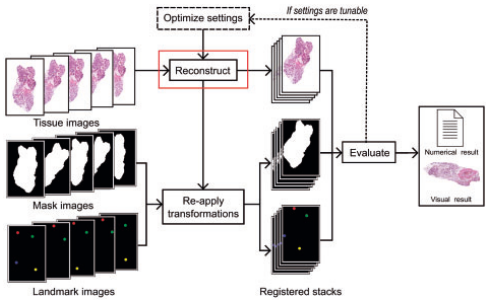

Fig. 1. Evaluation framework. A series of tissue images is input to a reconstruction method for registration. The transformations estimated by the method are re-applied to masks defining the tissue region and images containing landmarks. The registered tissue, mask and landmark images are used to evaluate reconstruction accuracy based on numerical metrics and visual examination. Moreover, tunable settings can be optimized. (Color version of this figure is available at *Bioinformatics* online.)

2 Materials and methods

2.1 Data collection and preprocessing

A murine prostate and a liver were fixed in PAXgene™ (PreAnalytiX GmbH, Hombrechtikon, Switzerland) and formalin, respectively, embedded in paraffin, and cut into serial 5 µm sections. The liver was processed with a laser prior to embedding in order to introduce artificial landmarks into the otherwise homogeneous tissue. Four holes were successfully introduced into the sample. The sections were hematoxylin-eosin (HE) stained and scanned at 20× (pixel size 0.46 µm) to obtain 260 (prostate) and 47 (liver) RGB images. The images were processed in MATLAB R2016b (The MathWorks Inc., Natick, MA, USA) to segment tissue from background and store the results as binary masks.

A total of 2448 landmarks were manually annotated. In the prostatic tissue, four corresponding points preferably at the centers of bisected nuclei were selected by two observers from each pair of adjacent sections. For the liver, the four holes in each image were marked by the same two observers. Most of the evaluated methods do not allow direct application of transformations to coordinates but support re-applying them to another stack of images. Therefore, we stored the landmarks as images with four disks placed at the landmark locations, each consisting of red, green, blue or yellow pixels. Color is invariant to the applied transformations, allowing

post-registration detection of the disks. The tissue, mask and landmark images were downsampled to different resolutions and stored as TIF. See Supplementary Methods for details.

2.2 Evaluation of reconstruction accuracy

2.2.1 Target registration error

Pairwise target registration error (TRE) (Fitzpatrick *et al.*, 1998), a direct measure of registration accuracy (Rohlfing, 2012), was quantified for each pair of adjacent sections. From the landmark images, we detected each landmark based on the colors of the disks and obtained their coordinates as the centroids of the detected pixels. For N pairs of sections, TRE was measured for each point ($j = \{1, 2, 3, 4\}$) and section pair ($i = \{1, 2, \ldots, N\}$) as:

$$TRE_{j,i} = \|\boldsymbol{X}_{j,i} - \boldsymbol{X}_{j,i+1}\| \tag{1}$$

that is, the Euclidean distance between the location $\boldsymbol{X_{j,i}}$ of point j on the section i and the location of the corresponding point on section $i+1$.

2.2.2 Accumulated error

Accumulated target registration error (ATRE) was calculated to quantify distortion accumulated through the stack, referred to as 'the banana problem' (Malandain *et al.*, 2004) or 'the shear effect' (Hughes *et al.*, 2013). Each landmark of the prostate dataset is only present on two consecutive sections and pairwise errors on different sections should thus be independent of each other. However, in the presence of accumulated errors, the error vectors on nearby sections are correlated (Beare *et al.*, 2008). We quantified this effect by treating the displacement of each landmark ($j = \{1, 2, 3, 4\}$) for each pair of sections ($i = \{1, 2, \ldots, N\}$) in vector form as $\boldsymbol{X}_{j,i} - \boldsymbol{X}_{j,i+1}$ and averaging the four vectors to obtain the mean displacement of each entire section. We then computed the cumulative sum of these mean vectors, proceeding from section 1 to section N. For section k, ATRE was defined as the Euclidean norm of the cumulative displacement vector:

$$ATRE_k = \left\| \sum_{i=1}^{k} \sum_{j=1}^{4} \frac{\boldsymbol{X}_{j,i} - \boldsymbol{X}_{j,i+1}}{4} \right\| \tag{2}$$

For the liver, a more direct quantification of ATRE was possible due to the landmarks extending through the sample. Ideally, the landmarks should lie on four parallel lines. In practice, parallelism could be violated due to slight movement of the sample between repeated applications of the laser. In a distorted volume, the landmarks deviate from the linear trajectories when proceeding through the stack. To measure this, we fitted a line in 3D to each of the four series of landmarks, minimizing mean squared error on the image plane. ATRE was then quantified for section i and landmark j as the Euclidean distance between the location of the landmark $\boldsymbol{X_{j,i}}$ and that of the fitted line $\boldsymbol{Y_{j,i}}$, on the image plane:

$$ATRE_{j,i} = \|\boldsymbol{X}_{j,i} - \boldsymbol{Y}_{j,i}\| \tag{3}$$

2.2.3 Tissue shrinkage and overlap

As certain reconstruction methods tend to shrink the tissue, relative change in tissue area (ΔA-%) was computed based on the tissue masks for each section. Overlap was quantified based on the masks for each section pair using the Jaccard index (Rohlfing, 2012). The Jaccard index can be considered a quality measure for pixel-wise metrics, as computing them for a pair of sections with little overlap can provide misleading results. Let A denote the set of tissue pixels of section i and B the set of tissue pixels of section $i+1$. The Jaccard index is defined as:

$$Jaccard_i = \frac{|A \cap B|}{|A \cup B|} \tag{4}$$

2.2.4 Pixel-wise similarity

For each section pair, we evaluated the similarity of corresponding pixels. After conversion to grayscale we computed the following measures: root mean squared error (RMSE), normalized cross correlation (NCC), mutual information (MI) and normalized mutual information (NMI) (Studholme *et al.*, 1999). Only the set of overlapping tissue pixels $A \cap B$ was considered. These indirect metrics provide information from the entire tissue area and complement the TRE evaluation.

2.2.5 Reconstruction smoothness

We quantified the smoothness of the reconstruction using contrast f_2 and correlation f_3 based on gray-level co-occurrence matrices (GLCMs) (Cifor *et al.*, 2011; Gaffling *et al.*, 2015; Haralick and Shanmugam, 1973). Low contrast and high correlation indicate a smooth reconstruction. We formed the GLCM for each pair of grayscale images based on pixels $A \cap B$ and summed them to obtain a single GLCM for the whole volume.

2.3 3D reconstruction

- LS: Least-squares fitting of an affine transformation to the landmarks was implemented in MATLAB R2016b. The result is in principle unaffected by error accumulation (Xu *et al.*, 2015).
- OPT: Optimization-based reconstruction implemented in MATLAB R2016b was used to estimate pairwise affine transformations by minimizing the value of pixel-wise MSE.
- SIFT: Feature-based reconstruction was performed by computing SIFT keypoints (Lowe, 2004) for each image pair, establishing putative matches and robustly fitting an affine transformation to the point pairs (Fischler and Bolles, 1981). We used the RegisterVirtualStackSlices (Arganda-Carreras *et al.*, 2006) implementation in Fiji, also used as an initial step in RVSS and ESA.
- HSR: HyperStackReg v. 5 (Ved P. Sharma, Albert Einstein College, https://sites.google.com/site/vedsharma/imagej-pluginsmacros/hyperstackreg) was run in Fiji to perform reconstruction using affine transformations.
- RVSS: Elastic reconstruction based on the bUnwarpJ algorithm, which is a combination of SIFT and optimization based methods, was applied using the RegisterVirtualStackSlices plugin in Fiji.
- ESA: The algorithm implemented in the ElasticStackAlignment plugin (Saalfeld *et al.*, 2012) was run via the TrakEM2 package (Cardona *et al.*, 2012) in Fiji to perform elastic reconstruction based on a combination of SIFT and optimization methods.
- MIM: Medical Image Manager, trial v. 0.94, was applied using images subsampled by a factor of 4 (magnification of 5×) as input. Sections 130 and 24 were used as references for the prostate and liver, respectively. We varied the initial magnification (0.3125×, 0.625×, 1.25× or 2.5×) and the number of nonrigid levels (1, 2, 3 or 4), thus modifying the image resolution used.
- Voloom: Trial v. 2.7.1 was used for elastic 3D reconstruction.

Fiji (Schindelin *et al.*, 2012; Schneider *et al.*, 2012) (v. 1.51h) plugins were run via ImageJ-MATLAB interface (v. 0.7.1) (Hiner *et al.*, 2016). Transformations were re-applied to the mask and landmark

images. Output was saved as TIF. See Supplementary Methods for details.

2.4 Parameter optimization

In the case of MIM, which had to be operated interactively, we evaluated each combination of tunable values by a parameter sweep. Tunable parameters of the other methods were optimized via Bayesian optimization (Shahriari *et al.*, 2016; Snoek *et al.*, 2012), which is well-suited for such problems, where the objective function is computationally expensive to evaluate, nonconvex, multimodal, and typically has low to moderate dimensionality. Bayesian optimization has been shown to perform favorably in comparison to other global optimization algorithms on benchmarking functions (Jones, 2001) as well as on real WSI data (Teodoro *et al.*, 2017). We used MATLAB's *bayesopt* implementation (https://www.mathworks.com/help/stats/bayesian-optimization-algorithm.html) with mean pairwise TRE as the objective function. We utilized a Gaussian process model of the objective function and an automatic relevance determination (ARD) Matérn 5/2 kernel (Snoek *et al.*, 2012) with 'expected-improvement-plus' as the acquisition function (Bull, 2011). Reconstructions with output image dimensions over fivefold compared to the input due to extreme error accumulation were considered failures. The number of variables to optimize was 2 (OPT), 4 (SIFT), 7 (RVSS) or 15 (ESA). We first optimized SIFT alone and used the optimal values for the SIFT step of RVSS and ESA. See Supplementary Table S1 for descriptions of the parameters. The number of seed points was set to twice the number of variables. We ran 30 iterations for OPT due to its simple objective function (Kartasalo *et al.*, 2016) and 100 iterations for the other tools. We used the prostate images subsampled by factors of 8 and 16, except for ESA, for which optimization was only feasible using the factor 16. Parameters optimized for ESA using the lower resolution were scaled to be used with the high resolution images. Computations were run on a workstation with Intel Xeon E5-1660 v3 3 GHz and 64 GB of RAM (low resolution) and a cluster node with Intel Xeon E5-2680 v3 2.5 GHz and 128 GB of RAM (high resolution).

3 Results

3.1 Effect of image resolution on evaluation metrics

First, we analyzed whether our metrics depend on image resolution (see Supplementary Results). TRE, ATRE, Jaccard and ΔA-% are essentially invariant to image resolution. They can be compared across different datasets and resolutions, as long as the accumulation of interpolation errors is avoided. RMSE, NCC, MI, NMI, f_2 and f_3 depend both on resolution and image content, and these metrics should thus only be compared within the same dataset and resolution. In all following analyses, we used images subsampled to pixel sizes of 7.36 and 3.68 μm, referred to as low and high resolution, respectively. The pixel sizes are close to the 5 μm section spacing and metrics computed from these images are not distorted by interpolation errors. Furthermore, we will only present RMSE as a measure of pixelwise similarity and f_2 as a measure of reconstruction smoothness due to their strong correlations with NCC, MI, NMI and f_3 (see Supplementary Table S1 for details).

3.2 Automated parameter tuning

Of the evaluated methods, LS, HSR and Voloom do not have tunable parameters. For OPT, SIFT, RVSS, ESA and MIM, we tuned the parameters automatically, minimizing the mean TRE computed for the prostate dataset. Parameter optimization took approximately 1500 hours in total to compute, producing 23 terabytes of data.

The optimization mostly converged close to the final solution in a handful of iterations (see Supplementary Results). By inspecting the variation in mean TRE values obtained during the process it is possible to reach a semi-quantitative view of the sensitivity of each method towards parameter adjustments. OPT and SIFT produced similar results for most parameter combinations while ESA, MIM and especially RVSS exhibited more sensitivity to parameter tuning.

We evaluated possible connections between accuracy and computation time, which might require the user to make a trade-off when selecting parameters (see Supplementary Results). The time taken by OPT varied only by a few minutes, except for the single inaccurate solutions where the parameters have not allowed proper convergence of the algorithm. For SIFT, there were no signs of a connection between accuracy and computation time. The differences in computation time between the fastest and slowest iterations of RVSS were roughly twofold and the fastest iterations were generally the ones with the highest error, indicating that minimizing the computation time of RVSS would sacrifice accuracy. In the case of ESA, the effect of parameter tuning was dramatic, leading to variation from approximately 12 min to more than 41 h. However, any clear relationship between computation time and accuracy was not observed.

3.3 Comparison of algorithms based on the prostate dataset

Results for the prostate dataset are listed in Table 1. The TRE values of LS based on landmarks by the two observers (LS1 and LS2) establish a baseline of accuracy. The case where the same landmarks were used for reconstruction and for calculating errors (LS1) is an optimistic estimate, representing the best accuracy reachable using an affine model. The errors calculated based on landmarks not used for reconstruction (LS2) represent a more realistic estimate of the accuracy of LS, serving as a cross-validation experiment between the two observers. The discrepancy between the optimistic and cross-validation results indicates that the LS solutions represent overfitting to the landmarks. Therefore, any methods with accuracy approaching LS can be regarded as highly accurate, since the other methods are not provided with any information concerning the landmarks. The systematic difference between TRE and ATRE calculated based on the two sets of landmarks (see Supplementary Table S1) is due to the fact that the two observers were free to select different landmarks and the error is generally not constant over the entire tissue section. However, using either set of landmarks leads to the same conclusions regarding the relative accuracy of the methods, confirmed by linear correlation coefficients of approximately 0.999 for mean TRE, 0.995 for maximum TRE, 0.888 for mean ATRE and 0.901 for maximum ATRE between the two sets of landmarks for the low resolution reconstructions. This also holds for the high resolution with corresponding values of 0.999, 0.986, 0.894 and 0.922. This indicates that even though four landmarks per section pair represent a relatively sparse sampling of the entire tissue section area, this number of landmarks is sufficient for reliable error estimation.

All methods benefited from parameter tuning on both image resolutions based on most of the metrics, using either set of landmarks for evaluation (see Table 1 and Supplementary Results). Of the top three methods, MIM and RVSS obtained better accuracy using high resolution images and ESA worked better on the low resolution images. ESA and MIM reached similar mean TRE values, slightly better than RVSS and approaching or exceeding the accuracy of LS.

Table 1. Evaluation results for the prostate data at low (top) and high resolution (bottom)

Prostate, low resolution

Algorithm	TRE1 μ	TRE1 max	TRE1 σ	ATRE1 μ	ATRE1 max	ATRE1 σ	RMSE μ	RMSE σ	Jaccard μ	Jaccard σ	Contrast f_2	ΔA-% μ	ΔA-% σ
Unregistered	489.26	2392.19	444.68	1153.08	2528.76	728.66	64.29	6.58	0.72	0.23	4260.86	0.00	0.00
LS 1	15.60	133.84	15.84	3.55	7.94	1.45	44.87	8.66	0.97	0.02	2150.63	5.28	8.89
LS 2	36.81	426.21	44.47	318.71	523.71	172.64	44.96	8.48	0.97	0.02	2126.81	31.75	22.22
OPT default	74.39	840.69	103.75	1207.72	2009.45	613.59	48.92	9.48	0.94	0.04	2538.84	−0.19	7.68
OPT optimal	23.89	350.99	28.67	417.90	648.24	206.70	42.83	8.65	0.97	0.02	1954.89	6.52	7.33
SIFT default	24.74	362.78	30.43	442.32	645.14	183.04	43.96	9.16	0.97	0.02	2066.20	−6.77	13.20
SIFT optimal	22.90	383.45	28.62	474.01	680.56	204.64	43.31	8.79	0.97	0.02	2001.13	−1.40	8.84
HSR	24.02	664.22	36.11	450.51	752.32	245.11	46.26	8.64	0.96	0.02	2280.25	3.18	5.32
RVSS default	93.96	4805.50	281.03	1228.69	2659.39	741.15	45.63	10.15	0.93	0.11	2072.08	−33.09	21.13
RVSS optimal	32.18	850.09	67.36	954.97	1353.44	431.53	42.46	8.89	0.96	0.04	1843.81	−8.99	5.44
ESA default	368.07	2278.21	442.01	834.71	1982.43	557.07	57.53	9.22	0.78	0.25	3127.28	0.01	0.10
ESA optimal	15.81	476.33	35.67	414.62	602.38	184.81	38.41	9.87	0.98	0.02	1603.96	2.34	2.73
MIM default	29.91	401.78	32.29	518.58	934.15	242.96	57.71	7.70	0.97	0.02	3449.70	0.01	2.38
MIM optimal	24.38	395.29	29.57	551.12	780.07	231.99	56.03	8.05	0.97	0.02	3266.80	−0.62	2.46
Voloom	39.18	730.44	48.39	713.29	1232.42	408.67	53.99	7.13	0.96	0.03	2988.03	−3.61	3.38

Prostate, high resolution

Algorithm	TRE1 μ	TRE1 max	TRE1 σ	ATRE1 μ	ATRE1 max	ATRE1 σ	RMSE μ	RMSE σ	Jaccard μ	Jaccard σ	Contrast f_2	ΔA-% μ	ΔA-% σ
Unregistered	489.25	2392.11	444.69	1152.97	2526.57	728.25	69.73	6.61	0.72	0.23	5021.08	0.00	0.00
LS 1	15.49	134.48	15.88	3.08	5.21	1.27	52.81	8.40	0.97	0.02	2939.94	4.91	8.77
LS 2	36.70	426.91	44.52	315.36	515.91	169.75	52.81	8.26	0.97	0.02	2908.40	31.28	22.08
OPT default	74.95	904.92	103.59	1327.22	2013.98	634.53	57.02	9.21	0.94	0.05	3404.82	−21.75	9.76
OPT optimal	24.25	345.68	29.46	402.79	633.01	201.36	50.75	8.43	0.97	0.02	2713.34	1.73	5.04
SIFT default	62.17	5451.71	319.97	577.46	1458.02	256.04	52.51	8.87	0.95	0.11	2838.59	−13.44	15.28
SIFT optimal	22.32	376.04	26.36	382.36	591.61	177.19	51.24	8.47	0.97	0.02	2763.28	−1.44	6.76
HSR	23.91	660.05	36.35	436.81	733.85	239.31	53.26	8.37	0.97	0.02	2990.32	1.03	5.60
RVSS default	34.35	1158.20	69.18	351.61	1070.20	148.22	50.26	9.51	0.96	0.06	2550.30	−28.06	13.25
RVSS optimal	19.49	446.90	28.31	352.14	579.83	162.65	48.92	8.56	0.97	0.02	2470.84	−4.28	3.62
ESA default	383.59	2278.27	441.44	934.43	2228.70	640.98	64.59	8.52	0.77	0.25	4043.04	0.02	0.08
ESA optimal	21.54	565.31	48.32	623.90	984.22	310.58	46.81	10.45	0.97	0.03	2346.21	1.21	2.30
MIM default	29.51	465.77	45.50	683.88	1105.29	290.42	56.74	8.12	0.96	0.03	3329.95	−0.37	3.00
MIM optimal	15.17	456.13	24.97	493.14	706.91	211.23	53.03	8.29	0.98	0.02	2944.42	−0.76	3.40
Voloom	43.35	684.11	56.28	687.46	1236.27	401.57	62.32	6.69	0.96	0.03	3945.05	−4.29	3.23

Note: Results for the unregistered images, LS based on landmarks by observer 1 (LS1) or 2 (LS2) and the automated methods (OPT, SIFT, HSR, RVSS, ESA, MIM, Voloom) using default or optimized parameters. Mean (μ), maximum (max) and standard deviation (σ) over all sections are shown. TRE and ATRE based on landmarks by observer 1 are in μm. In the online version, columns with TRE, ATRE, RMSE, *f*2 and ΔA-% are colored from low (blue) to high values (red). Columns with Jaccard are colored from high (blue) to low values (red). (Color version of this table is available at *Bioinformatics* online.)

In terms of maximum TRE and ATRE, the three methods were comparable, but RVSS reached slightly lower ATRE than ESA or MIM. Among all tools, ESA and MIM also obtained the highest Jaccard index values. The RMSE and f_2 metrics do not allow comparison across different image resolutions and one should note that MIM's output was always stored at the lower resolution for technical reasons. Considering these limitations, we can observe that ESA performed best in terms of these metrics on both image resolutions ahead of RVSS. Changes in tissue area introduced by ESA, MIM and RVSS were moderate. Behind the top three, most other tools reached accuracy comparable to each other. The worst results were obtained using default parameters and for some methods, most notably ESA and RVSS, they were even comparable to the unregistered original images.

Visual examination in 3D revealed differences in the geometry of the reconstructions formed using each of the methods (Fig. 2). Compared to the undistorted reference (LS1), the distortions introduced by OPT, SIFT, HSR, ESA and MIM were a manifestation of the typical 'banana-into-cylinder' issue. This gradual straightening of curved structures is most clearly seen here in the displacement of the urethra at the top of the stacks. As indicated by the numerical ATRE values, the overall magnitude of this effect was rather similar across the tools. The distortions caused by RVSS and Voloom were more complex, representing clockwise twisting of the sample when seen from the top.

3.4 Comparison of algorithms based on the liver dataset

Results for the liver dataset are listed in Table 2. The four artificial landmarks were annotated by both observers and the two sets of TRE and ATRE values can be treated as replicates. This is reflected by linear correlation coefficients of approximately one (ranging from 0.99993 to 0.99998) for mean TRE, maximum TRE, mean ATRE and maximum ATRE calculated based on the two sets of landmarks (see Supplementary Table S1). In this case, LS thus represents an optimistic estimate of the accuracy reachable with a global affine model. Compared to the prostate sample, this dataset is more challenging to reconstruct due to the more homogeneous appearance of the tissue and the presence of deformations such as folded and torn tissue. This is reflected by the metrics, which generally indicate higher errors, except for RMSE and f_2 which are lower due to the more homogeneous image content. Ideally, it would be convenient to process different datasets without having to readjust parameters. With this in mind, we reused the parameters optimized for the prostate dataset, treating the evaluation on the liver dataset as an independent validation experiment. Based on most metrics, the optimized parameters generally resulted in an improvement over the default parameters also when applied to the liver dataset (see Table 2 and Supplementary Results).

As with the prostate, the lowest TRE values among the automated methods were achieved by ESA on the lower resolution and MIM on the high resolution data with RVSS being the third best method. The other methods reached TRE values comparable to each other. In terms of maximum TRE and ATRE, the conclusion was less clear. Voloom performed better on the lower resolution, reaching a maximum TRE second only to LS, while ESA and OPT also reached comparable values. On this dataset, MIM suffered from larger maximum errors compared to the higher quality prostate sample. The lowest mean ATRE values among all automated methods were obtained by ESA, MIM and Voloom, while in terms of maximum ATRE Voloom was superior to ESA and MIM. ESA was the top method in terms of RMSE and f_2, and MIM obtained the highest Jaccard index. Again, the poorest results were obtained when using the default values of tunable parameters.

Visualization in 3D supported the numerical results (Fig. 3). ESA, MIM and Voloom formed reconstructions with landmarks concentrated on four roughly parallel lines as expected, but some

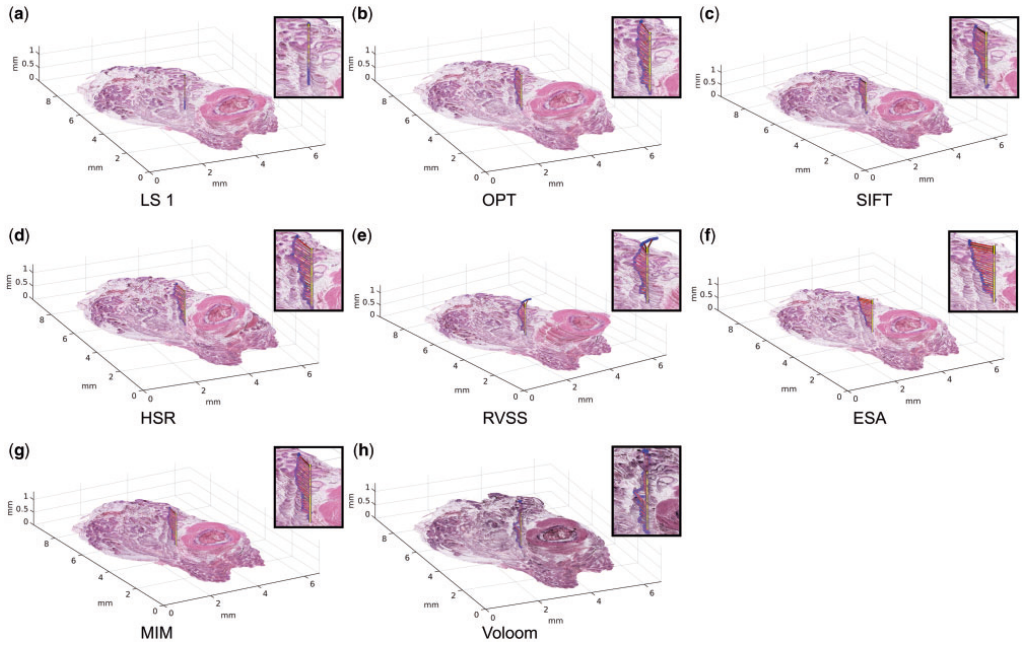

Fig. 2. Reconstructions using (**a**) LS based on landmarks by observer 1, (**b**) OPT, (**c**) SIFT, (**d**) HSR, (**e**) RVSS, (**f**) ESA, (**g**) MIM and (**h**) Voloom. Optimized parameters and the most suitable resolution were used for each method. The dots represent the trajectory of accumulated target registration error from section to section. The horizontal lines indicate the direction and magnitude of the cumulative mean displacement of each section relative to the ideal error-free trajectory (vertical line). Magnified views are shown next to each reconstruction. Viewing the high-resolution color version of the Figure online is recommended. (Color version of this figure is available at *Bioinformatics* online.)

Table 2. Evaluation results for the liver data at low (top) and high resolution (bottom)

Liver, low resolution

Algorithm	TRE1 μ	TRE1 max	TRE1 σ	ATRE1 μ	ATRE1 max	ATRE1 σ	RMSE μ	RMSE σ	Jaccard μ	Jaccard σ	Contrast f_2	ΔA-% μ	ΔA-% σ
Unregistered	726.81	2558.97	528.95	543.56	1706.62	298.02	44.90	5.03	0.67	0.15	2031.62	0.00	0.00
LS 1	27.30	396.78	55.62	25.87	314.15	35.96	34.69	6.39	0.90	0.07	1225.25	6.15	8.94
LS 2	33.52	401.27	55.70	29.52	318.41	36.55	34.76	6.41	0.90	0.07	1230.75	7.55	9.10
OPT default	200.11	1120.63	197.43	189.74	933.68	154.81	39.70	5.90	0.86	0.08	1663.83	−40.28	21.10
OPT optimal	84.86	617.62	112.51	97.28	482.65	80.44	35.26	6.44	0.92	0.06	1293.17	−10.76	8.69
SIFT default	178.38	3900.82	383.37	729.60	2096.57	511.87	36.28	7.08	0.86	0.12	1327.28	−6.61	10.43
SIFT optimal	173.15	3755.45	453.05	668.41	2837.41	572.90	35.07	6.91	0.87	0.14	1258.35	−0.78	7.44
HSR	86.99	718.85	117.16	118.15	407.31	83.00	38.27	6.26	0.92	0.06	1520.41	−15.99	9.96
RVSS default	330.02	3764.99	600.79	656.13	2186.17	494.23	36.85	7.46	0.92	0.08	1338.65	−13.23	14.70
RVSS optimal	252.32	2689.75	436.63	855.53	1677.06	334.83	35.20	7.45	0.85	0.16	1261.35	−0.39	3.31
ESA default	717.22	2558.97	539.55	538.28	1702.38	302.25	44.44	6.07	0.67	0.16	1992.03	0.00	0.01
ESA optimal	46.32	618.27	92.03	63.72	599.97	68.07	32.23	7.03	0.90	0.08	1075.18	−0.44	2.27
MIM default	121.44	2241.90	327.01	380.34	1500.07	370.61	42.83	5.70	0.90	0.11	1857.95	0.41	3.49
MIM optimal	79.74	1767.90	169.53	75.82	1233.78	108.02	42.58	5.59	0.92	0.08	1841.03	2.34	6.68
Voloom	90.98	555.46	103.81	80.12	382.78	71.12	37.69	5.39	0.91	0.07	1444.09	1.87	5.51

Liver, high resolution

Algorithm	TRE1 μ	TRE1 max	TRE1 σ	ATRE1 μ	ATRE1 max	ATRE1 σ	RMSE μ	RMSE σ	Jaccard μ	Jaccard σ	Contrast f_2	ΔA-% μ	ΔA-% σ
Unregistered	726.87	2559.07	528.92	543.55	1706.53	298.04	48.79	4.90	0.67	0.15	2396.69	0.00	0.00
LS 1	27.25	398.01	55.60	25.82	314.38	35.95	39.21	5.87	0.90	0.07	1554.89	5.87	8.92
LS 2	33.53	401.34	55.62	29.51	317.90	36.54	39.28	5.88	0.90	0.07	1560.83	7.27	9.08
OPT default	202.50	1115.20	198.27	185.80	961.31	154.84	43.85	5.48	0.86	0.08	2000.94	−40.49	20.46
OPT optimal	83.68	625.48	112.30	97.24	481.94	79.82	39.75	5.90	0.92	0.06	1628.50	−14.25	9.50
SIFT default	145.16	1388.05	173.41	223.89	1052.81	146.44	41.91	6.28	0.88	0.08	1782.81	−6.94	6.81
SIFT optimal	84.94	1026.27	130.96	157.17	630.95	117.20	39.51	6.01	0.90	0.08	1590.79	0.18	4.62
HSR	88.08	1117.63	133.55	153.43	598.88	120.99	42.24	5.73	0.92	0.07	1836.69	−19.07	10.87
RVSS default	179.82	1097.54	165.98	332.02	1052.27	165.93	42.31	5.84	0.92	0.06	1813.05	−7.96	8.40
RVSS optimal	79.26	1135.00	135.65	167.36	602.79	123.38	38.97	6.17	0.90	0.08	1548.98	−1.57	3.64
ESA default	693.75	2559.07	544.51	538.73	1711.11	301.12	47.90	6.70	0.68	0.16	2315.71	0.00	0.02
ESA optimal	60.60	929.16	142.25	56.58	832.23	99.19	37.68	6.44	0.90	0.09	1448.05	0.44	1.20
MIM default	95.74	1150.34	156.76	150.75	866.23	134.37	43.27	5.98	0.90	0.09	1896.02	0.85	3.79
MIM optimal	65.42	1060.78	122.46	66.54	646.40	78.31	42.00	5.70	0.92	0.07	1792.75	3.38	6.73
Voloom	144.08	3335.29	399.41	113.82	3159.53	274.36	42.77	4.84	0.91	0.07	1848.66	1.45	5.41

Note: Results for the unregistered images, LS based on landmarks by observer 1 (LS1) or 2 (LS2) and the automated methods (OPT, SIFT, HSR, RVSS, ESA, MIM, Voloom) using default or optimized parameters. Mean (μ), maximum (max) and standard deviation (σ) over all sections are shown. TRE and ATRE based on landmarks by observer 1 are in μm. In the online version, columns with TRE, ATRE, RMSE, *f*2 and ΔA-% are colored from low (blue) to high values (red). Columns with Jaccard are colored from high (blue) to low values (red). (Color version of this table is available at *Bioinformatics* online.)

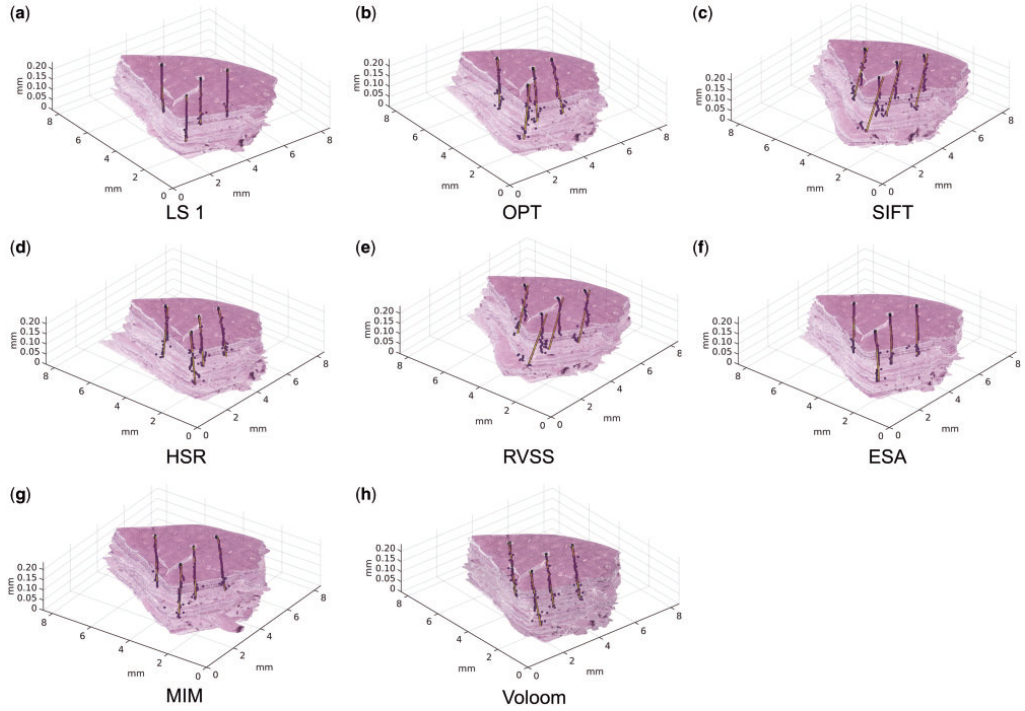

Fig. 3. Reconstructions using (**a**) LS based on landmarks by observer 1, (**b**) OPT, (**c**) SIFT, (**d**) HSR, (**e**) RVSS, (**f**) ESA, (**g**) MIM and (**h**) Voloom. Optimized parameters and the most suitable resolution were used for each method. The locations of the four landmark points on each section are indicated with dots, shown together with lines of best fit to each of the four series of points. Note that the scale of the vertical axis is different from the horizontal axes in the visualization. Viewing the high-resolution color version of the Figure online is recommended. (Color version of this figure is available at *Bioinformatics* online.)

distortion is visible at the bottom part of the stack reconstructed by MIM. These kind of distortions were more severe in the case of OPT, SIFT, HSR and RVSS.

4 Discussion

Based on this study, methods utilizing locally varying transformations (ESA, MIM, RVSS, Voloom) were superior to those constrained to global affine models (OPT, SIFT, HSR). ESA was the only method to consistently outperform or match the other approaches on two datasets based on the majority of metrics. In the case of the higher quality prostate dataset, differences in accuracy between the tools were rather subtle. All three top-performing methods on this dataset incorporate an elastic transformation model: MIM and RVSS use a B-spline grid and ESA is based on a piecewise linear mesh. While methods relying on a global transformation model also performed reasonably well, the additional accuracy offered by elastic transformations could be crucial when microstructure at the cellular scale is of interest. In the case of the liver sample, more profound differences between the methods were observed, likely due to the more challenging tissue content and the presence of deformations, which cannot be compensated for using a global model. ESA, MIM and Voloom stood out from the other methods. While Voloom appeared to be less accurate on average compared to ESA and MIM based on mean TRE, it demonstrated the lowest maximum and accumulated errors of all automated methods, indicating capability to avoid propagation of errors even in the presence of considerable deformations. The ability of the algorithms to tolerate such deformations is a significant benefit. Due to the mostly manual nature of histological sectioning and brittleness of the thin tissue sections, deformations in the form of folds and tears often occur. This challenge is especially encountered in 3D histology, when uninterrupted sequences of sections are desired.

Another important property of algorithms to consider is sensitivity to adjustable parameters. Even an algorithm that produces highly accurate results with a carefully selected set of parameter values will be useless if the user has little chance of finding this set of values. Comparing algorithms from this perspective is difficult. Each algorithm has a different set of parameters and the range of values to evaluate has to be selected for each parameter, which can in turn affect the amount of variation observed in the results. Nevertheless, this study still provides a semi-quantitative view of the sensitivity of the studied algorithms against parameter adjustments. Of the evaluated methods, LS, HSR and Voloom are the most convenient due to their lack of tunable parameters. OPT and SIFT also produced similar results with most parameter values. The results produced by ESA varied greatly depending on parameters, but we discovered numerous combinations leading to almost optimal results. In the case of MIM, there are only a handful of tunable parameters and they are relatively easy to tune. Moreover, ESA and MIM appear to

be well-behaving in the sense that parameters optimized for the prostate dataset also suited the liver dataset. In contrast, RVSS was found to be difficult to optimize and even though its accuracy using optimized settings was close to ESA and MIM on the prostate dataset, reaching this level of accuracy without automated parameter tuning would be challenging.

An open question common to all of the methods is how image resolution affects reconstruction accuracy. A pixel size close to the section spacing is often recommended (Amunts *et al.*, 2013; Braumann *et al.*, 2005; Dauguet *et al.*, 2007; Ju *et al.*, 2006; Kartasalo *et al.*, 2016; Saalfeld *et al.*, 2012) based on the assumption that objects smaller than this are only visible on a single section and are thus not useful for registration, and may even introduce errors (Beare *et al.*, 2008). However, suitably oriented elongated structures such as blood vessels can be observed on several sections even if their diameter on the image plane is smaller than the section spacing. In principle, some algorithms might thus benefit from a smaller pixel size. We evaluated reconstruction accuracy using pixel sizes of 3.68 and 7.36 μm. Based on the rule of thumb above, it is unclear which one of these should be preferred given a section spacing of 5 μm. Our results indicate that using a pixel size close to the section spacing is a reasonable starting point, but the optimal image resolution depends on the algorithm and also somewhat on the image content. Furthermore, we cannot rule out the possibility that algorithms which performed better on the high resolution images, most notably MIM, might benefit from an even smaller pixel size. In conclusion, the image resolution thus needs to be selected experimentally for each application and algorithm.

The two samples selected for this study are markedly different in their histological composition. The fact that the top methods performed well on both the prostate and the liver dataset without any retuning of parameters indicates that these methods are not overly sensitive to tissue appearance, and that the results obtained in this study are not specific to a single dataset. However, some variation in the relative performance of the algorithms on the two datasets was still observed. Thus, collecting and annotating additional datasets representing diverse tissue types and other histological stainings, such as immunohistochemistry, remains an important goal for future studies.

While we evaluated a comprehensive set of methods for 3D histology, it might be worthwhile to adapt general-purpose image registration algorithms to this context. Another opportunity, not supported by any of the methods here, could be the exploitation of additional data obtained e.g. by magnetic resonance imaging or in the form of blockface images (Amunts *et al.*, 2013; Casero *et al.*, 2017; Dauguet *et al.*, 2007; Gibson *et al.*, 2013; Johnson *et al.*, 2010; Stille *et al.*, 2013). Furthermore, although advances in image acquisition and processing have enabled the first steps towards 3D histology, sample preparation still constitutes a significant bottleneck. In the future, emerging technologies for automated sample preparation (Onozato *et al.*, 2011) or integrated sectioning and imaging (Li *et al.*, 2010; Ragan *et al.*, 2012) might potentially transform 3D histology into a high-throughput process.

Acknowledgements

We thank Ignacio Arganda-Carreras, Martin Groher, Derek Magee, Stephan Saalfeld and Ved Sharma for their helpful advice. Katja Liljeström, Marja Pirinen and Marika Vähä-Jaakkola are acknowledged for skillful technical assistance.

Funding

This work was supported by Academy of Finland [269474]; Tekes [269/31/2015]; Cancer Society of Finland; Emil Aaltonen Foundation; Finnish Foundation for Technology Promotion; KAUTE Foundation; and Orion Research Foundation.

Conflict of Interest: none declared.

References

Amunts,K. *et al.* (2013) BigBrain: an ultrahigh-resolution 3D human brain model. *Science*, **340**, 1472–1475.

Arganda-Carreras,I. *et al.* (2006) Consistent and elastic registration of histological sections using vector-spline regularization. In: *International Workshop on Computer Vision Approaches to Medical Image Analysis*, pp. 85–95.

Arganda-Carreras,I. *et al.* (2010) 3D reconstruction of histological sections: application to mammary gland tissue. *Microsci. Res. Technol.*, **73**, 1019–1029.

Beare,R. *et al.* (2008) An assessment of methods for aligning two-dimensional microscope sections to create image volumes. *J. Neurosci. Methods*, **170**, 332–344.

Braumann,U. *et al.* (2005) Three-dimensional reconstruction and quantification of cervical carcinoma invasion fronts from histological serial sections. *IEEE Trans. Med. Imaging*, **24**, 1286–1307.

Bull,A.D. (2011) Convergence rates of efficient global optimization algorithms. *J. Mach. Learn. Res.*, **12**, 2879–2904.

Calì,C. *et al.* (2016) Three-dimensional immersive virtual reality for studying cellular compartments in 3D models from EM preparations of neural tissues. *J. Comp. Neurol.*, **524**, 23–38.

Cardona,A. *et al.* (2012) TrakEM2 software for neural circuit reconstruction. *PloS One*, **7**, e38011.

Casero,R. *et al.* (2017) Transformation diffusion reconstruction of three-dimensional histology volumes from two-dimensional image stacks. *Med. Image Anal.*, **38**, 184–204.

Cifor,A. *et al.* (2011) Smoothness-guided 3-D reconstruction of 2-D histological images. *Neuroimage*, **56**, 197–211.

Dauguet,J. *et al.* (2007) Three-dimensional reconstruction of stained histological slices and 3D non-linear registration with in-vivo MRI for whole baboon brain. *J. Neurosci. Methods*, **164**, 191–204.

Fischler,M.A. and Bolles,R.C. (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, **24**, 381–395.

Fitzpatrick,J.M. *et al.* (1998) Predicting error in rigid-body point-based registration. *IEEE Trans. Med. Imaging*, **17**, 694–702.

Gaffling,S. *et al.* (2015) A Gauss-Seidel iteration scheme for reference-free 3-D histological image reconstruction. *IEEE Trans. Med. Imaging*, **34**, 514–530.

Ghaznavi,F. *et al.* (2013) Digital imaging in pathology: whole-slide imaging and beyond. *Annu. Rev. Pathol.-Mech.*, **8**, 331–359.

Gibson,E. *et al.* (2013) 3D prostate histology image reconstruction: quantifying the impact of tissue deformation and histology section location. *J. Path. Inform.*, **4**, 31.

Griffin,J. and Treanor,D. (2017) Digital pathology in clinical use: where are we now and what is holding us back? *Histopathology*, **70**, 134–145.

Haralick,R.M. and Shanmugam,K. (1973) Textural features for image classification. *IEEE Trans. Syst. Man Cybern.*, **3**, 610–621.

Hiner,M.C. *et al.* (2016) ImageJ-MATLAB: a bidirectional framework for scientific image analysis interoperability. *Bioinformatics*, **33**, 629–630.

Hughes,C. *et al.* (2013) Robust alignment of prostate histology slices with quantified accuracy. *IEEE Trans. Biomed. Eng.*, **60**, 281–291.

Johnson,G.A. *et al.* (2010) Waxholm space: an image-based reference for coordinating mouse brain research. *Neuroimage*, **53**, 365–372.

Jones,D.R. (2001) A taxonomy of global optimization methods based on response surfaces. *J. Global. Optim.*, **21**, 345–383.

Ju,T. *et al.* (2006) 3D volume reconstruction of a mouse brain from histological sections using warp filtering. *J. Neurosci. Methods*, **156**, 84–100.

Kartasalo,K. *et al.* (2016) Benchmarking of algorithms for 3D tissue reconstruction. In: *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2360–2364.

Koos,B. *et al.* (2015) Next-generation pathology—surveillance of tumor microecology. *J. Mol. Biol.*, **427**, 2013–2022.

Ledford,H. (2017) Cell atlases race to map the body. *Nature*, **542**, 404–405.

Lein,E.S. *et al.* (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, **445**, 168–176.

Li,A. *et al.* (2010) Micro-optical sectioning tomography to obtain a high-resolution atlas of the mouse brain. *Science*, **330**, 1404–1408.

Lowe,D.G. (2004) Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, **60**, 91–110.

Magee,D. *et al.* (2015) Histopathology in 3D: from three-dimensional reconstruction to multi-stain and multi-modal analysis. *J. Path. Inform.*, **6**, 6.

Malandain,G. *et al.* (2004) Fusion of autoradiographs with an MR volume using 2-D and 3-D linear transformations. *Neuroimage*, **23**, 111–127.

Meijering,E. *et al.* (2016) Imagining the future of bioimage analysis. *Nat. Biotechnol.*, **34**, 1250–1255.

Mignardi,M. *et al.* (2017) Bridging histology and bioinformatics—computational analysis of spatially resolved transcriptomics. *Proc. IEEE*, **105**, 530–541.

Onozato,M.L. *et al.* (2011) Evaluation of a completely automated tissue-sectioning machine for paraffin blocks. *J. Clin. Pathol.*, 200205.

Onozato,M.L. *et al.* (2012) A role of three-dimensional (3D)-reconstruction in the classification of lung adenocarcinoma. *Anal. Cell. Pathol.*, **35**, 79–84.

Paish,E.C. *et al.* (2009) Three-dimensional reconstruction of sentinel lymph nodes with metastatic breast cancer indicates three distinct patterns of tumour growth. *J. Clin. Pathol.*, **62**, 617–623.

Ragan,T. *et al.* (2012) Serial two-photon tomography for automated ex vivo mouse brain imaging. *Nat. Methods*, **9**, 255–258.

Roberts,N. *et al.* (2012) Toward routine use of 3D histopathology as a research tool. *Am. J. Pathol.*, **180**, 1835–1842.

Rohlfing,T. (2012) Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE Trans. Med. Imaging*, **31**, 153–163.

Rojas,K.D. *et al.* (2015) Methodology to study the three-dimensional spatial distribution of prostate cancer and their dependence on clinical parameters. *J. Med. Imaging*, **2**, 037502.

Rusk,N. (2016) Genomics: spatial transcriptomics. *Nat. Methods*, **13**, 710–711.

Saalfeld,S. *et al.* (2012) Elastic volume reconstruction from series of ultra-thin microscopy sections. *Nat. Methods*, **9**, 717–720.

Schindelin,J. *et al.* (2012) Fiji: an open-source platform for biological-image analysis. *Nat. Methods*, **9**, 676–682.

Schneider,C.A. *et al.* (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods*, **9**, 671.

Shahriari,B. *et al.* (2016) Taking the human out of the loop: a review of bayesian optimization. *Proc. IEEE*, **104**, 148–175.

Snoek,K. *et al.* (2012) Practical Bayesian optimization of machine learning algorithms. *Adv. Neurol. Int.*, 2951–2959.

Song,Y. *et al.* (2013) 3D reconstruction of multiple stained histology images. *J. Path. Inform.*, **4**, 7.

Sotiras,A. *et al.* (2013) Deformable medical image registration: a survey. *IEEE Trans. Med. Imaging*, **32**, 1153–1190.

Ståhl,P.L. *et al.* (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, **353**, 78–82.

Stille,M. *et al.* (2013) 3D reconstruction of 2D fluorescence histology images and registration with in vivo MR images: application in a rodent stroke model. *J. Neurosci. Methods*, **219**, 27–40.

Studholme,C. *et al.* (1999) An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognit.*, **32**, 71–86.

Teodoro,G. *et al.* (2017) Algorithm sensitivity analysis and parameter tuning for tissue image segmentation pipelines. *Bioinformatics*, **33**, 1064–1072.

Theart,R.P. *et al.* (2017) Virtual reality assisted microscopy data visualization and colocalization analysis. *BMC Bioinformatics*, **18**, 64.

Thevenaz,P. *et al.* (1998) A pyramid approach to subpixel registration based on intensity. *IEEE Trans. Image Process*, **7**, 27–41.

Wang,Y. *et al.* (2015) Three-dimensional reconstruction of light microscopy image sections: present and future. *Front. Med.*, **9**, 30–45.

Xu,Y. *et al.* (2015) A method for 3D histopathology reconstruction supporting mouse microvasculature analysis. *PloS One*, **10**, e0126817.

PUBLICATION III

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 (2017), 555–565

Cytometry
PART A
Journal of the International Society for Advancement of Cytometry

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 Ruusuvuori[1,3]*

[1]BioMediTech and Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland

[2]BioMediTech Institute and Faculty of Biomedical Science and Engineering, Tampere University of Technology, Tampere, Finland

[3]Faculty 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: 269/31/2015

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

• **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. © 2017 International Society for Advancement of Cytometry

• **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

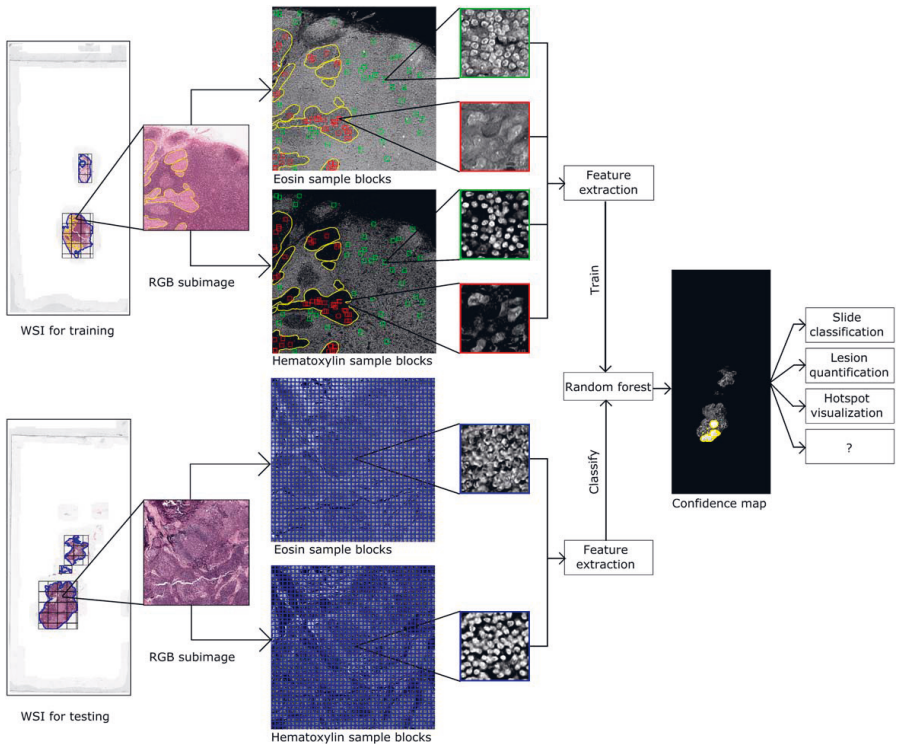

Figure 1. The analysis workflow for training (upper half) and classification (lower half). During model training, the lymph node tissue (blue outline) is first segmented from the whole slide image containing annotated metastatic regions (yellow outline). The detected tissue sections are then divided 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.

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.

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 $0.5 \times t_{Otsu}$ to S, where t_{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 coverslip 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 5 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 reported relative to the full resolution and were scaled accordingly and rounded to the nearest integer. The numerical parameter values are given as applied for the data in this study, and should be modified when data with a different resolution or different characteristics is processed.

Color Normalization

We used histogram matching, applied separately to each color channel, to correct for color variation across the WSIs (27). The training image Tumor_015.tif was selected as the reference based on visual examination, and the histograms of the image were used as templates for the other images. For each WSI and color channel, we computed the mapping function required for matching the original histogram to the template histogram. When estimating the histograms and the resulting mapping functions, we again used images downsampled with a factor of 16 and only considered lymph node tissue pixels detected in the previous step (i.e., the pixels indicated by TRUE in B). As a result, a mapping function was obtained for each color channel of each WSI.

Data Handling and Storage

After the detection of lymph nodes in an image, we computed the bounding box for each of the remaining objects in B and merged any overlapping bounding boxes into larger boxes. Regions of the WSIs corresponding to each bounding box were then retrieved at full resolution. The histogram mapping functions estimated in the previous step were then applied to the lymph node tissue pixels of the full resolution bounding box image. Each region was saved into a separate file first in uncompressed BigTiff format using a tile size of 1,024 × 1,024, followed by conversion into JPEG2000 format with a compression ratio of 50 using the JP2 WSI Converter (28).

In addition to the actual images, we also saved the segmentation masks in B corresponding to each bounding box. The masks were scaled up to full resolution by nearest-neighbor interpolation and saved in BigTiff format using one bit per sample, a tile size of 1,024 × 1,024 and PackBits compression. In the case of training images containing tumor, the parts of the ground truth masks corresponding to each bounding box were retrieved at full resolution and saved in separate image files using the same format as the segmentation masks.

For convenient handling of the image data during model training and classification, we further divided the images obtained in the previous step into smaller subimages and stored them in JPEG2000 format. Each resulting subimage had dimensions of 8,192 × 8,192, except for partial subimages at the edges of the bounding boxes. The segmentation and ground truth masks were processed similarly and saved in TIF format. The location of each subimage relative to the corresponding full-resolution WSI was also stored.

Preprocessing of Subimages

Color deconvolution and nuclei segmentation steps were applied to each train and test subimage. A color deconvolution algorithm (29) was used to convert the image's RGB channels into hematoxylin stain, eosin stain and background. In H&E stained images, hematoxylin stains mainly the cell nuclei and therefore the hematoxylin channel was used in the nuclei segmentation. The hematoxylin channel was filtered with a 10 × 10 pixel Gaussian filter (standard deviation = 5 pixels) and then an adaptive thresholding method was applied to get the binary image. The applied adaptive thresholding method (30) separates the cell nuclei from the background based on an individual threshold for each pixel. The individual threshold is selected based on the mean intensity in 20 × 20 pixel local neighborhood. Watershed segmentation was used to separate the overlapping and touching nuclei from each other. The separation lines of the watershed segmentation were computed from the distance transform of the binary image using 8-connected neighborhood.

Sampling

Random sampling was performed to reduce the amount of training data. Approximately 200,000 sample blocks were randomly selected from the subimages containing normal tissue and 200,000 sample blocks from subimages containing tumor. Sample blocks were selected from the full resolution subimages and the block size was 200 × 200 pixels. These negative and positive samples were selected only from the segmented lymph node tissue mask area and ground truth mask

area, respectively, while excluding the background. As all provided tissue area from all training images was covered, this led to approximately 200 sample blocks per tumor subimage and 15 sample blocks per normal tissue subimage. From each sample block, 214 descriptive features related to image texture, spatial structure, and distribution of nuclei were extracted.

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 1. Results concerning the performance of different classification models

	CORRECTLY CLASSIFIED SAMPLES (%)	SENSITIVITY (%)	SPECIFICITY (%)	F-SCORE
Logistic regression	87.0	86.4	87.6	0.87
NN	82.8	74.4	91.0	0.81
NN feature subset	91.3	85.7	96.8	0.91
Random forest	93.0	92.6	93.3	0.93
SVM	88.3	85.9	90.6	0.88

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.

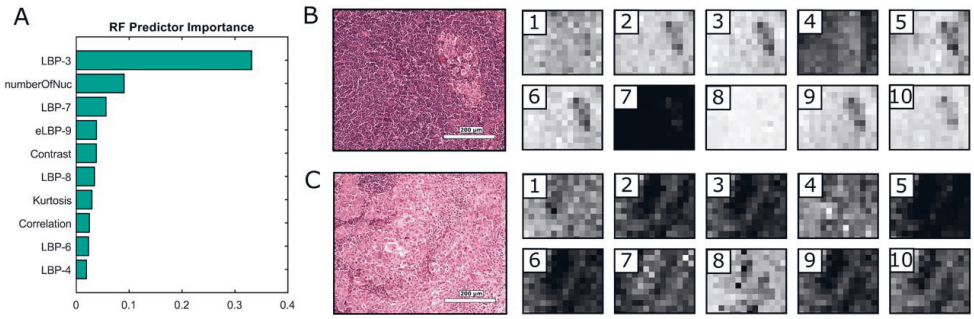

Figure 2. Relative importance of the 10 most significant features selected by the random forest model (A). Example H&E images of normal tissue (B) and metastatic tissue (C) are shown with the corresponding features computed from the hematoxylin (H) or eosin (E) channel: local binary pattern 3 (H) (B1 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]

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. 2B and 2C). 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

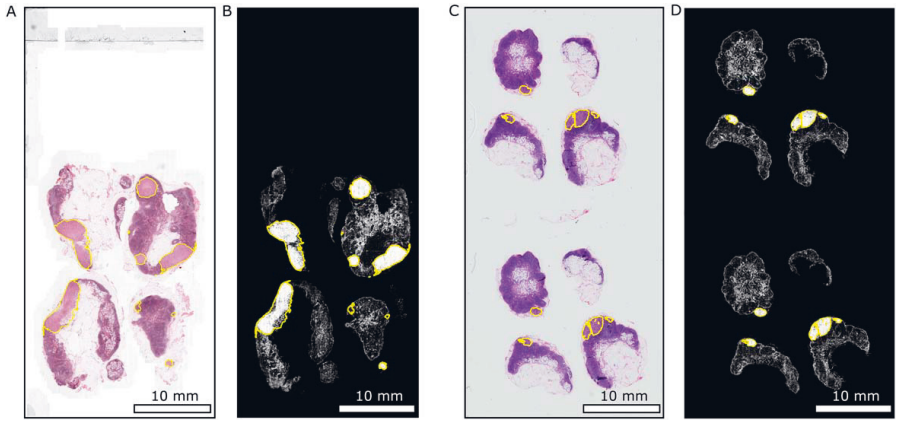

Figure 3. An example whole slide image from the first dataset (A) with the corresponding confidence map (B) and an example whole slide image from the second dataset (C) with the corresponding confidence map (D). Ground truth annotations are shown in yellow.

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 mm^2) 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

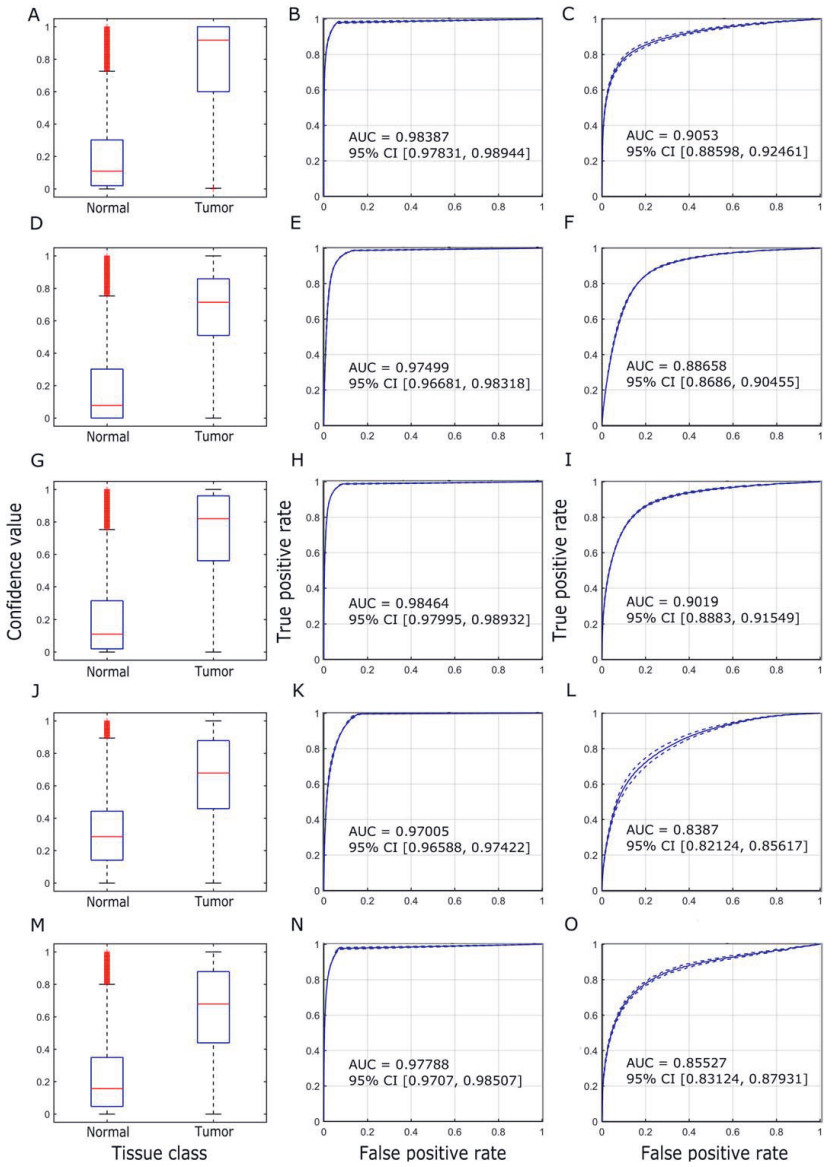

Figure 4. Results obtained using leave-one-out cross validation for dataset 1 (A–C), dataset 2 (D–F) or the combined dataset (G–I) and for a classifier trained on dataset 1 and evaluated on dataset 2 (J–L) or for a classifier trained on dataset 2 and evaluated on dataset 1 (M–O). Distribution of confidence 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]

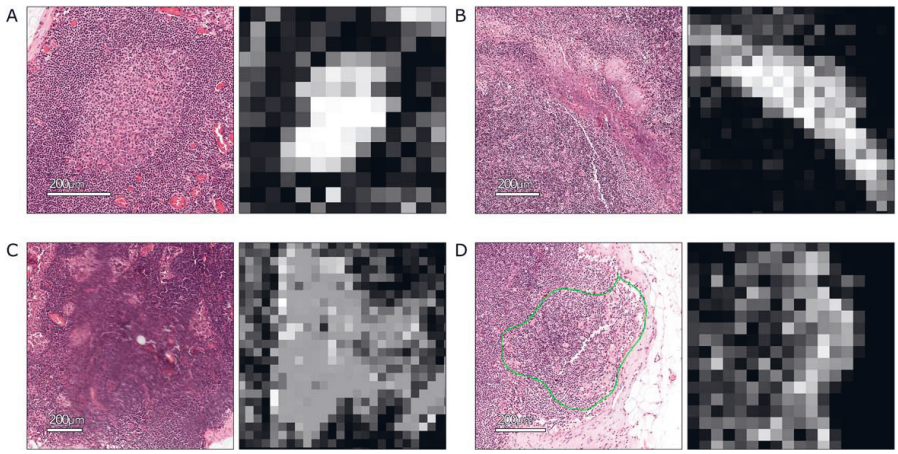

Figure 5. Examples of false positives caused by normal tissue texture resembling metastatic tissue (A, B) or an out-of-focus region (C) and an example of a false negative where a small lesion has been falsely detected as normal tissue (D). The H&E images are shown on the left and the corresponding confidence maps on the right. The ground truth annotation in (D) is shown as a green outline.

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 entire WSIs as either normal or tumor containing, it is sensible to evaluate performance using slide-level 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 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 with 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 kind 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 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

1. Pantanowitz L, Valenstein PN, Evans AJ, Kaplan KJ, Pfeifer JD, Wilbur DC, Collins LC, Colgan TJ. Review of the current state of whole slide imaging in pathology. J Pathol Inform 2011;2:36.
2. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B. Histopathological image analysis: A review. IEEE Rev Biomed Eng 2009;2:147–171.
3. Ghaznavi F, Evans A, Madabhushi A, Feldman M. Digital imaging in pathology: Whole-slide imaging and beyond. Annu Rev Pathol 2013;8:331–359.
4. Bejnordi BE, Litjens G, Hermsen M, Karssemeijer N, van der Laak JA. A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images. In: Proceedings of the International Society for Optics and Photonics (SPIE) Conference on Medical Imaging, Orlando, FL, USA; 2015. p 94200H-94200H.
5. Khan AM, Rajpoot N, Treanor D, Magee D. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. IEEE Trans Biomed Eng 2014;61:1729–1738.
6. Veta M, van Diest PJ, Kornegoor R, Huisman A, Viergever MA, Pluim JP. Automatic nuclei segmentation in H&E stained breast cancer histopathology images. PLoS One 2013;8:e70221.
7. Kothari S, Phan JH, Stokes TH, Wang MD. Pathology imaging informatics for quantitative analysis of whole-slide images. J Am Med Inform Assoc 2013;20:1099–1108.
8. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer; 2013. Available from: http://globocan.iarc.fr, accessed on 27/4/2016.
9. Ran S, Volk L, Hall K, Flister MJ. Lymphangiogenesis and lymphatic metastasis in breast cancer. Pathophysiology 2010;17:229–251.
10. Veta M, Pluim JP, van Diest PJ, Viergever MA. Breast cancer histopathology image analysis: A review. IEEE Trans Biomed Eng 2014;61:1400–1411.
11. Niwas SI, Palanisamy P, Sujathan K, Bengtsson E. Analysis of nuclei textures of fine needle aspirated cytology images for breast cancer diagnosis using complex Daubechies wavelets. Signal Process 2013;93:2828–2837.
12. Abas FS, Gokozan HN, Goksel B, Otero JJ, Gurcan MN. Intraoperative neuropathology of glioma recurrence: Cell detection and classification. In: Proceedings of the International Society for Optics and Photonics (SPIE) Conference on Medical Imaging, San Diego, CA, USA; 2016. p 979109-979109.
13. Kornaropoulos EN, Niazi M, Lozanski G, Gurcan MN. Histopathological image analysis for centroblasts classification through dimensionality reduction approaches. Cytometry Part A 2014;85A:242–255.
14. Doyle S, Feldman M, Tomaszewski J, Madabhushi A. A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. IEEE Trans Biomed Eng 2012;59:1205–1218.
15. Turkki R, Linder N, Kovanen PE, Pellinen T, Lundin J. Identification of immune cell infiltration in hematoxylin-eosin stained breast cancer samples: Texture-based classification of tissue morphologies. In: Proceedings of the International Society for Optics and Photonics (SPIE) Conference on Medical Imaging, San Diego, CA, USA; 2016. p 979110-979110.
16. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33:1–22.
17. Breiman L. Random forests. Mach Learn 2001;45:5–32. 1
18. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA; 2012. pp 1097–1105.
19. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–444.
20. Sirinukunwattana K, Raza SEA, Tsang YW, Snead D, Cree IA, Rajpoot NM. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans Med Imaging 2016;35:1196–1206.
21. Wang H, Cruz-Roa A, Basavanhally A, Gilmore H, Shih N, Feldman M, Tomaszewski J, Gonzalez F, Madabhushi A. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. J Med Imaging 2014;1:034003.
22. Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Berlin, Heidelberg: Springer; 2013. pp 411–418.
23. Chen H, Qi X, Yu L, Heng PA. DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation. arXiv preprint arXiv 2016; 1604.02677.
24. Sirinukunwattana K, Pluim JPW, Chen H, Qi X, Heng P-A, Guo YB, Wang LY, Matuszewski BJ, Bruni E, Sanchez U, et al. Gland Segmentation in Colon Histology Images: The GlaS Challenge Contest. arXiv preprint arXiv 2016; 1603.00275.
25. Ruusuvuori P, Valkonen M, Nykter M, Visakorpi T, Latonen L. Feature-based analysis of mouse prostatic intraepithelial neoplasia in histological tissue sections. J Pathol Inform 2016;7:5.
26. Otsu N. A threshold selection method from gray-level histograms. Automatica 1975; 11:23–27.
27. Gonzalez RC, Woods RE. Digital Image Processing, 2nd ed. New Jersey, Upper Saddle River: Prentice hall, Inc.; 2002.
28. Tuominen VJ, Isola J. Linking whole-slide microscope images with DICOM by using JPEG2000 interactive protocol. J Digit Imaging 2010;23:454–462.
29. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. Anal Quant Cytol Histol 2001;23:291–299.
30. Bradley D, Roth G. Adaptive thresholding using the integral image. J Graph Gpu Game Tools 2007;12:13–21.
31. Ojala T, Pietikäinen M, Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 2002;24:971–987.
32. Pietikäinen M, Ojala T, Xu Z. Rotation-invariant texture classification using feature distributions. Pattern Recognit 2000;33:43–52.
33. Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2004;60:91–110.
34. Ludwig O, Delgado D, Goncalves V, Nunes U. Trainable Classifier-fusion Schemes: An Application to Pedestrian Detection. In: Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems, St. Louis, MO, USA; 2009. pp 432–437.
35. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 1st IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA; 2005. pp 886–893.
36. Matas J, Chum O, Urban M, Pajdla T. Robust wide-baseline stereo from maximally stable extremal regions. Image Vis Comput 2004;22:761–767.
37. Vedaldi A, Fulkerson B. VLFeat: An Open and Portable Library of Computer Vision Algorithms. In: Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy; 2010. pp 1469–1472.
38. Sertel O, Kong J, Shimada H, Catalyurek UV, Saltz JH, Gurcan MN. Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development. Patt Recogn 2009;42:1093–1103.
39. Yu KH, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, Snyder M. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun 2016;7:12474.
40. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep learning for identifying metastatic breast cancer. arXiv preprint 2016; arXiv:1606.05718.
41. Chen R, Jing Y, Jackson H. identifying metastases in sentinel lymph nodes with deep convolutional neural networks. arXiv preprint 2016; arXiv:1608.01658.

PUBLICATION
IV

Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study

Ström, P.*, **Kartasalo, K.***, Olsson, H., Solorzano, L., Delahunt, B., Berney, D. M., Bostwick, D. G., Evans, A. J., Grignon, D. J., Humphrey, P. A., Iczkowski, K. A., Kench, J. G., Kristiansen, G., van der Kwast, T. H., Leite, K. R., McKenney, J. K., Oxley, J., Pan, C.-C., Samaratunga, H., Srigley, J. R., Takahashi, H., Tsuzuki, T., Varma, M., Zhou, M., Lindberg, J., Lindskog, C., Ruusuvuori, P., Wählby, C., Grönberg, H., Rantalainen, M., Egevad, L. and Eklund, M.

The Lancet Oncology 21.2 (2020), 222–232

Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study

Peter Ström, Kimmo Kartasalo*, Henrik Olsson, Leslie Solorzano, Brett Delahunt, Daniel M Berney, David G Bostwick, Andrew J Evans , David J Grignon, Peter A Humphrey, Kenneth A Iczkowski, James G Kench, Glen Kristiansen, Theodorus H van der Kwast, Katia R M Leite, Jesse K McKenney, Jon Oxley, Chin-Chen Pan, Hemamali Samaratunga, John R Srigley, Hiroyuki Takahashi, Toyonori Tsuzuki, Murali Varma, Ming Zhou, Johan Lindberg, Cecilia Lindskog, Pekka Ruusuvuori, Carolina Wählby, Henrik Grönberg, Mattias Rantalainen, Lars Egevad, Martin Eklund*

Lancet Oncol 2020; 21: 222–32

Published Online January 8, 2020 https://doi.org/10.1016/S1470-2045(19)30738-7

This online publication has been corrected. The corrected version first appeared at thelancet.com/oncology on January 30, 2020

See Comment page 187

*These authors contributed equally

Department of Medical Epidemiology and Biostatistics (P Ström MSc, H Olsson MSc, J Lindberg PhD, Prof H Grönberg MD, M Rantalainen PhD, M Eklund PhD) **and Department of Oncology and Pathology** (Prof L Egevad MD), **Karolinska Institutet, Stockholm, Sweden; Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland** (K Kartasalo MSc, P Ruusuvuori PhD)**; Centre for Image Analysis, Department of Information Technology** (L Solorzano MSc, Prof C Wählby PhD) **and Department of Immunology, Genetics, and Pathology** (C Lindskog PhD), **Uppsala University, Uppsala, Sweden; Department of Pathology and Molecular Medicine, Wellington School of Medicine and Health Sciences, University of Otago, Wellington, New Zealand** (Prof B Delahunt MD)**; Barts Cancer Institute, Queen Mary University of London, London, UK** (Prof D M Berney MD)**; Bostwick Laboratories, Orlando, FL, USA** (Prof D G Bostwick MD)**; Laboratory Medicine Program, University Health Network, Toronto General Hospital, Toronto, ON, Canada** (A J Evans MD, Prof T H van der Kwast MD)**; Department of Pathology and**

Summary

Background An increasing volume of prostate biopsies and a worldwide shortage of urological pathologists puts a strain on pathology departments. Additionally, the high intra-observer and inter-observer variability in grading can result in overtreatment and undertreatment of prostate cancer. To alleviate these problems, we aimed to develop an artificial intelligence (AI) system with clinically acceptable accuracy for prostate cancer detection, localisation, and Gleason grading.

Methods We digitised 6682 slides from needle core biopsies from 976 randomly selected participants aged 50–69 in the Swedish prospective and population-based STHLM3 diagnostic study done between May 28, 2012, and Dec 30, 2014 (ISRCTN84445406), and another 271 from 93 men from outside the study. The resulting images were used to train deep neural networks for assessment of prostate biopsies. The networks were evaluated by predicting the presence, extent, and Gleason grade of malignant tissue for an independent test dataset comprising 1631 biopsies from 246 men from STHLM3 and an external validation dataset of 330 biopsies from 73 men. We also evaluated grading performance on 87 biopsies individually graded by 23 experienced urological pathologists from the International Society of Urological Pathology. We assessed discriminatory performance by receiver operating characteristics and tumour extent predictions by correlating predicted cancer length against measurements by the reporting pathologist. We quantified the concordance between grades assigned by the AI system and the expert urological pathologists using Cohen's kappa.

Findings The AI achieved an area under the receiver operating characteristics curve of 0·997 (95% CI 0·994–0·999) for distinguishing between benign (n=910) and malignant (n=721) biopsy cores on the independent test dataset and 0·986 (0·972–0·996) on the external validation dataset (benign n=108, malignant n=222). The correlation between cancer length predicted by the AI and assigned by the reporting pathologist was 0·96 (95% CI 0·95–0·97) for the independent test dataset and 0·87 (0·84–0·90) for the external validation dataset. For assigning Gleason grades, the AI achieved a mean pairwise kappa of 0·62, which was within the range of the corresponding values for the expert pathologists (0·60–0·73).

Interpretation An AI system can be trained to detect and grade cancer in prostate needle biopsy samples at a ranking comparable to that of international experts in prostate pathology. Clinical application could reduce pathology workload by reducing the assessment of benign biopsies and by automating the task of measuring cancer length in positive biopsy cores. An AI system with expert-level grading performance might contribute a second opinion, aid in standardising grading, and provide pathology expertise in parts of the world where it does not exist.

Funding Swedish Research Council, Swedish Cancer Society, Swedish eScience Research Center, EIT Health.

Introduction

Histopathological evaluation of prostate biopsies is crucial to the clinical management of men suspected of having prostate cancer. However, the histopathological diagnosis of prostate cancer is associated with several challenges. More than one million men undergo prostate biopsy in the USA annually.[1] With the standard biopsy procedure resulting in 10–12 needle cores per patient, more than 10 million tissue samples need to be examined by pathologists. The increasing incidence of prostate cancer in an ageing population means that the number of biopsies is likely to further increase. Additionally, a global shortage of pathologists exists. For example, China has only one pathologist per 130 000 population, and in many African countries the ratio is in the order of one per million.[2,3] Western countries are facing similar problems, with an expected decline in the number of practicing pathologists due to retirement.[4] Gleason grade is a strong prognostic factor for the survival of patients with prostate cancer and is crucial for treatment decisions.

Research in context

Evidence before this study

We did a literature search in PubMed, searching the title, abstract, and keywords of peer-reviewed, English-language journal and conference articles published between database inception and May 17, 2019, using the terms "prostate cancer" AND "histo*" AND ("machine learning" OR "deep learning" OR "artificial intelligence"). We also examined the reference lists of relevant publications. Contemporary studies using whole slide imaging of entire histopathological slides and deep learning techniques have shown promising results for detection of prostate cancer, and attempts at grading in prostatectomies and tissue microarrays. These previous studies have not shown experienced urological pathologist-level consistency in grading or investigated grading of needle biopsies, which is the diagnostic sampling method used in routine clinical practice. Moreover, automated estimation of tumour burden in biopsies has not been reported. None of the previous studies have relied on a well defined sample cohort, which allows for clinically meaningful estimation of diagnostic performance metrics, such as sensitivity and specificity.

Added value of this study

To the best of our knowledge, we present for the first time an algorithm that reaches a performance comparable to experienced urological pathologists in the detection, tumour burden estimation, and grading of prostate cancer in needle biopsies. The AI system was developed and evaluated on a population-based dataset prospectively collected within a clinical trial, which included standardised biopsy procedures, centralised pathology reporting, and blinding to clinical characteristics, such as PSA. This dataset represents a broad spectrum of malignant morphologies of prostatic tissue encountered in clinical practice.

Implications of all the available evidence

Use of AI to assist pathologists could substantially decrease their workload by pre-screening cases and by automatically estimating tumour burden, improve patient safety by alarming about potentially missed cancers, and reduce variability in grading by providing decision support. Our results warrant prospective validation in clinical trials to confirm the potential benefits of AI-assisted prostate histopathology in routine clinical practice.

Gleason grade is based on morphological examination and is recognised as subjective. This subjectivity is reflected in high intrapathologist and interpathologist variability in reported grades, as well as both underdiagnosis and overdiagnosis of prostate cancer.[5,6]

A possible solution to these challenges is the application of artificial intelligence (AI) to prostate cancer histopathology. The development of an AI system to identify benign biopsies with high accuracy could decrease the workload of pathologists and allow them to focus on difficult cases. Furthermore, an accurate AI could assist the pathologist with the identification, localisation, and grading of prostate cancer among those biopsies not excluded in the initial screening process, thus providing a safety net to protect against potential misclassification of biopsies. AI-assisted pathology assessment could reduce inter-observer variability in grading, leading to more consistent and reliable diagnoses and better treatment decisions.

By use of high resolution scanning, tissue samples can be digitised to whole slide images and used as the input for the training of deep neural networks (DNNs), an AI technique that has achieved state-of-the-art accuracy in many classification problems across various fields, including medical imaging.[7–10] However, little work has been undertaken in prostate diagnostic histopathology.[11–16] Attempts at grading prostate biopsies by DNNs have been limited to small datasets or subsets of Gleason patterns, and they have not analysed the clinical implications of the introduction of AI-assisted prostate pathology. In this study, we aimed to develop an AI system with clinically acceptable accuracy for prostate cancer detection, localisation, and Gleason grading.

Methods

Study design and participants

Between May 28, 2012, and Dec 30, 2014, the prospective, population-based, screening-by-invitation STHLM3 study (ISRCTN84445406) evaluated a diagnostic model for prostate cancer in men aged 50–69 years residing in Stockholm, Sweden.[17,18] STHLM3 participants had 10–12-core transrectal ultrasound-guided systematic biopsies if they had prostate-specific antigen (PSA) concentration of 3 ng/mL or more or a Stockholm3 test score of 10% or more. Urologists who participated in the study and the study pathologist were blinded to the clinical characteristics of the patients. A single pathologist (LE) graded all biopsy cores according to the International Society of Urological Pathology (ISUP) grading classification (where Gleason scores 6, 3+4, 4+3, 8, and 9–10 are reported as ISUP grade 1 to 5, also referred to as Gleason Grade Groups).[19] LE also delineated cancerous areas using a marker pen and measured the linear cancer extent.

The biopsy cores were formalin fixed and stained with haematoxylin and eosin. A random selection of 8571 biopsies from 1289 STHLM3 participants stratified by ISUP grade was digitised (figure 1). The cases were chosen to represent the full range of diagnoses, with an overrepresentation of high-grade disease. To further enrich the data with high-grade cases, 271 slides from 93 men with ISUP 4 and 5 prostate cancers were obtained from outside STHLM3 (figure 1; appendix p 3). These slides were regraded by LE, digitised, and used for training purposes only. We used 1631 cores from a random selection of 246 (19·1%) men to evaluate the performance of the AI (the independent test set), and the rest were used

Laboratory Medicine, Indiana University School of Medicine, Indianapolis, IN, USA (Prof D J Grignon MD); **Department of Pathology, Yale University School of Medicine, New Haven, CT, USA** (Prof P A Humphrey MD); **Department of Pathology, Medical College of Wisconsin, Milwaukee, WI, USA** (Prof K A Iczkowski MD); **Department of Tissue Pathology and Diagnostic Oncology, Royal Prince Alfred Hospital and Central Clinical School, University of Sydney, Sydney, NSW, Australia** (Prof J G Kench MD); **Institute of Pathology, University Hospital Bonn, Bonn, Germany** (Prof G Kristiansen MD); **Department of Urology, Laboratory of Medical Research, University of São Paulo Medical School, São Paulo, Brazil** (Prof K R M Leite MD); **Pathology and Laboratory Medicine Institute, Cleveland Clinic, Cleveland, OH, USA** (J K McKenney MD); **Department of Cellular Pathology, Southmead Hospital, Bristol, UK** (J Oxley MD); **Department of Pathology, Taipei Veterans General Hospital, Taipei, Taiwan** (C Pan MD); **Aquesta Uropathology and University of Queensland, Brisbane, QLD, Australia** (Prof H Samaratunga MD); **Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada** (Prof J R Srigley MD); **Department of Pathology, Jikei University School of Medicine, Tokyo, Japan** (H Takahashi MD); **Department of Surgical Pathology, School of Medicine, Aichi Medical University, Nagakute, Japan** (Prof T Tsuzuki MD); **Department of Cellular Pathology, University Hospital of Wales, Cardiff, UK** (M Varma MD); **Department of Pathology, UT Southwestern Medical Center, Dallas, TX, USA** (Prof M Zhou MD); **BioImage Informatics Facility of SciLifeLab, Uppsala, Sweden** (Prof C Wählby); **and Department of Oncology, St Göran Hospital, Stockholm, Sweden** (Prof H Grönberg)

Correspondence to:
Dr Martin Eklund, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm SE-171 77, Sweden
martin.eklund@ki.se

See Online for appendix

for model training. All biopsies from a given participant were assigned to either the training or the test dataset.[20]

Because slides from different pathology labs differ in appearance and quality due to differences in slide preparation and because the characteristics and appearance of whole slide images vary by scanner, assessment of the performance of DNN models on external labs and scanners (ie, images of slides from different pathology labs and scanners than the images on which the model was trained) from a real-world clinical setting is crucial. We therefore obtained 330 slides (73 men) from the Karolinska University Hospital and digitised them on the scanner available at the hospital's pathology laboratory to replicate their entire workflow of processing and slide digitisation (the external validation dataset; figure 1). The selection of slides was enriched for higher ISUP grades to permit evaluation of predictions for these uncommon grades. LE graded all biopsies in the external test dataset to avoid confounding from introducing a different reporting pathologist and a different laboratory and scanner workflow simultaneously.

As an additional test dataset, we digitised 87 cores from the Pathology Imagebase, a reference database launched by ISUP to promote the standardisation of reporting of urological pathology (figure 1).[21] These cases were independently reviewed by 23 highly experienced urological pathologists (the ISUP Imagebase panel). The experts were selected on the basis of their international reputation and scientific production. A Medline search informed that they had authored an average of 105 papers on prostate pathology (range 21–321), with an average of 39 first-author or last-author papers (5–190) at the time of

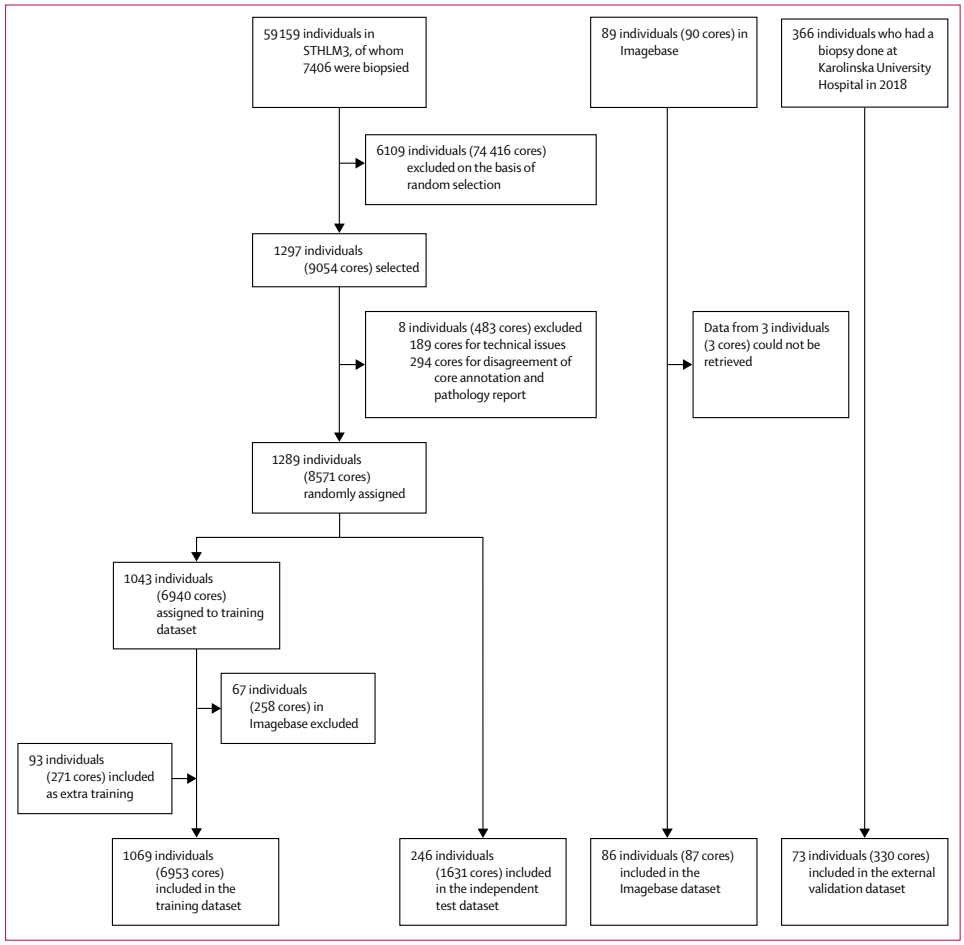

Figure 1: **Study profile**

	STHLM3	Participants (n=1454)				
	Biopsied (n=7406)	Training (n=976)	Extra training (n=93)	Test (n=246)	Imagebase (n=86)	External (n=73)
Age, years						
<49	45 (0·6%)	4 (0·4%)	0	1 (0·4%)	0	2 (2·7%)
50–54	639 (8·6%)	76 (7·8%)	2 (2·2%)	11 (4·5%)	10 (11·6%)	5 (6·8%)
55–59	1221 (16·5%)	136 (13·9%)	4 (4·3%)	44 (17·9%)	8 (9·3%)	10 (13·7%)
60–64	2027 (27·4%)	255 (26·1%)	5 (5·4%)	67 (27·2%)	23 (26·7%)	12 (16·4%)
65–69	3294 (44·5%)	482 (49·4%)	14 (15·1%)	115 (46·7%)	44 (51·2%)	15 (20·5%)
≥70	180 (2·4%)	20 (2·0%)	48 (51·6%)	8 (3·3%)	1 (1·2%)	29 (39·7%)
Missing	0	3 (0·3%)	20 (21·5%)	0	0	0
Previous negative biopsy						
Yes	505 (6·8%)	33 (3·4%)	0	13 (5·28%)	7 (8·1%)	..
No	6901 (93·2%)	940 (96·3%)	0	233 (94·72%)	79 (91·9%)	..
Missing	0	3 (0·3%)	93 (100·0%)	0	0	..
Prostate-specific antigen						
<3 ng/mL	1933 (26·1%)	228 (23·4%)	2 (2·2%)	43 (17·48%)	13 (15·1%)	..
3–<5 ng/mL	3458 (46·7%)	428 (43·9%)	2 (2·2%)	100 (40·65%)	48 (55·8%)	..
5–<10 ng/mL	1612 (21·8%)	213 (21·8%)	13 (14·0%)	73 (29·67%)	16 (18·6%)	..
≥10 ng/mL	403 (5·4%)	104 (10·7%)	47 (50·5%)	30 (12·2%)	9 (10·5%)	..
Missing	0	3 (0·3%)	30 (32·3%)	0	0	..
Digital rectal examination						
Abnormal	680 (9·2%)	133 (13·6%)	46 (49·5%)	39 (15·85%)	12 (14·0%)	..
Normal	6726 (90·8%)	840 (86·1%)	8 (8·6%)	207 (84·15%)	74 (86·0%)	..
Missing	0	3 (0·3%)	39 (41·9%)	0	0	..
Prostate volume						
<35 mL	2701 (36·5%)	425 (43·5%)	19 (20·4%)	92 (37·4%)	42 (48·8%)	..
35–<50 mL	2494 (33·7%)	319 (32·7%)	14 (15·1%)	82 (33·33%)	36 (41·9%)	..
≥50 mL	2211 (29·9%)	229 (23·5%)	19 (20·4%)	72 (29·27%)	8 (9·3%)	..
Missing	0	3 (0·3%)	41 (44·1%)	0	0	..
Cancer length						
No cancer	4605 (62·2%)	142 (14·5%)	0	35 (14·23%)	0	16 (21·9%)
>0–1 mm	545 (7·4%)	133 (13·6%)	2 (2·2%)	35 (14·23%)	4 (4·7%)	1 (1·4%)
>1–5 mm	922 (12·4%)	258 (26·4%)	10 (10·8%)	61 (24·8%)	20 (23·3%)	10 (13·7%)
>5–10 mm	449 (6·1%)	135 (13·8%)	17 (18·3%)	28 (11·38%)	20 (23·3%)	6 (8·2%)
>10 mm	885 (11·9%)	308 (31·6%)	64 (68·8%)	87 (35·37%)	42 (48·8%)	40 (54·8%)
Cancer grade*						
Benign	4605 (62·2%)	142 (14·5%)	0	35 (14·2%)	..	16 (21·9%)
ISUP 1 (3+3)	1558 (21·0%)	413 (42·3%)	1 (1·1%)	104 (42·3%)	..	12 (16·4%)
ISUP 2 (3+4)	761 (10·3%)	200 (20·5%)	1 (1·1%)	53 (21·5%)	..	12 (16·4%)
ISUP 3 (4+3)	253 (3·4%)	96 (9·8%)	1 (1·1%)	16 (6·5%)	..	16 (21·9%)
ISUP 4 (4+4, 3+5, and 5+3)	101 (1·4%)	63 (6·5%)	19 (20·4%)	21 (8·5%)	..	8 (11·0%)
ISUP 5 (4+5, 5+4, and 5+5)	128 (1·7%)	62 (6·4%)	71 (76·3%)	17 (6·9%)	..	9 (12·3%)

Data are n (%). No cancer grade information is shown for Imagebase, because the grading of this set of samples was done independently by multiple observers. Imagebase cancer length was assessed by LE. ISUP=International Society of Urological Pathology. *Numbers in brackets are the Gleason scores associated with the ISUP grades.

Table 1: **Baseline characteristics**

recruitment to Imagebase.[21] Cores from the men in the three test datasets were not part of model development and were excluded from any analysis until the final evaluation.

The study protocol was approved by Stockholm regional ethics committee (permits 2012/572-31/1, 2012/438-31/3, and 2018/845-32). Additional details concerning data collection are in the appendix (p 3).

Test methods

We processed the whole slide images with a segmentation algorithm based on Laplacian filtering to identify the regions corresponding to tissue sections and annotations drawn adjacent to the tissue. We then extracted digital pixel-wise annotations, indicating the locations of cancerous tissue of any grade, by identifying the tissue

	STHLM3	Digitised biopsy slides (n=8980)				
	Biopsied (n=83 470)	Training (n=6682)	Extra Training (n=271)	Test (n=1631)	Imagebase (n=87)	External (n=330)
Cancer length						
No cancer	73595 (88·2%)	3724 (55·7%)	1 (0·4%)	910 (55·8%)	0	108 (32·7%)
>0–1 mm	3307 (4·0%)	915 (13·7%)	7 (2·6%)	203 (12·4%)	8 (9·2%)	33 (10·0%)
>1–5 mm	4135 (5·0%)	1239 (18·5%)	41 (15·1%)	295 (18·1%)	44 (50·6%)	77 (23·3%)
>5–10 mm	1822 (2·2%)	591 (8·8%)	85 (31·4%)	150 (9·2%)	24 (27·6%)	75 (22·7%)
>10 mm	611 (0·7%)	213 (3·2%)	111 (41·0%)	73 (4·5%)	11 (12·6%)	37 (11·2%)
Missing	0	0	26 (9·6%)	0	0	0
Cancer grade						
Benign	73595 (88·2%)	3724 (55·7%)	1 (0·4%)	910 (55·8%)	..	108 (32·7%)
ISUP 1 (3+3)	5664 (6·8%)	1530 (22·9%)	1 (0·4%)	349 (21·4%)	..	65 (19·7%)
ISUP 2 (3+4)	2051 (2·5%)	538 (8·1%)	1 (0·4%)	142 (8·7%)	..	63 (19·1%)
ISUP 3 (4+3)	903 (1·1%)	261 (3·9%)	2 (0·7%)	66 (4·0%)	..	49 (14·8%)
ISUP 4 (4+4, 3+5, and 5+3)	689 (0·8%)	424 (6·3%)	45 (16·6%)	92 (5·6%)	..	19 (5·8%)
ISUP 5 (4+5, 5+4, and 5+5)	568 (0·7%)	205 (3·1%)	221 (81·5%)	72 (4·4%)	..	26 (7·9%)

Data are n (%). No cancer grade information is shown for Imagebase, because the grading of this set of samples was done independently by multiple observers. Imagebase cancer length was assessed by LE. ISUP=International Society of Urological Pathology. *Numbers in brackets are the Gleason scores associated with the ISUP grades.

Table 2: **Baseline characteristics of included biopsy cores**

region corresponding to each annotation. To obtain training data representing the morphological characteristics of Gleason patterns 3, 4, and 5, we extracted multiple partially overlapping smaller images, or patches, from each whole slide image. We used patch dimensions of 598×598 pixels (around 540×540 µm) at a resolution corresponding to 10× magnification (pixel size around 0·90 µm). The process resulted in around 5·1 million patches usable for training a DNN (appendix p 24).

We used two convolutional DNN ensembles, each consisting of 30 Inception V3 models pretrained on ImageNet, with classification layers adapted to our outcome.[22,23] The first ensemble performed binary classification of image patches into benign or malignant, while the second ensemble classified patches into Gleason patterns 3–5. To reduce label noise in the second ensemble, we trained it on patches extracted from cores containing only one Gleason pattern (ie, cores with Gleason score 3+3, 4+4, or 5+5). The test data still contained cores of all grades to provide a real-world scenario for evaluation. Each DNN in the first and the second ensemble thus predicted the probability of each patch being malignant, and whether it represented Gleason pattern 3, 4, or 5 (appendix p 25).

Once the probabilities for the Gleason pattern at each location of the biopsy core were obtained from the DNN ensembles, we mapped them to core-specific characteristics (ISUP grade and cancer length) using boosted trees, a machine learning algorithm based on decision tree models and gradient boosting.[24] All cores in the training data were used for training the boosted trees. Specifically, aggregated features from the patch-wise probabilities predicted by each DNN for each core were used as input to the boosted trees, and the clinical assessment of ISUP score and cancer length were used as outcomes. The ISUP grade group was assigned based on a Bayesian decision rule of the core-level classifier to obtain ISUP predictions at a clinically relevant operating point (appendix p 14).

Statistical analysis

No formal sample size calculation was done. We summarised the operating characteristics of the AI system in a receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC), both on core-level and patient-level. We then specified a range of acceptable sensitivities for potential clinical use and evaluated achieved specificity when compared to the pathology report. The enrichment of high-grade disease in the independent test data and the external validation data might inflate the estimated AUC values, because high grades might be easier to discriminate from benign cases compared with ISUP 1 and 2. Therefore, we also estimated the AUC when ISUP 3–5 cases were removed from the independent test and the external validation datasets.

We predicted cancer length in each core and compared it with the cancer length described in the pathology report. The comparison was done with individual and aggregated cores (ie, total cancer length) for each participant. Linear correlation was assessed in all cores and participants, as well as limited to positive cores and men.

Cohen's kappa with linear weights was used for evaluation of the AI's performance against the 23 experienced urological pathologists on the Imagebase test dataset. Linear weights emphasise a higher level of disagreement of ratings further away from each other on the ordinal ISUP scale, in accordance with previous publications on the Imagebase study.[21] Each of the 87 slides in Imagebase

was graded by each of the 23 Imagebase panel pathologists and by the AI system. To evaluate how well the AI system agreed with the pathologists, we calculated all pairwise kappas and summarised the mean for each of the 23 raters. Additionally, we estimated the kappa with a grouping of the Gleason scores in ISUP grades (grade groups) 1, 2–3, and 4–5. We further estimated Cohen's kappa against the study pathologist's ISUP grading of the independent test dataset and the external validation dataset. For the external validation dataset, we also estimated Cohen's kappa after calibrating the probabilities (ie, scaling the ISUP probabilities before assigning the predicted class).

We used t-distributed stochastic neighbour embedding and the deep Taylor decomposition to interpret the representation of the image data learned by the DNN models (appendix p 17).[25]

We excluded cores in which the on-slide annotations did not match the pathology report and cores with technical issues. Participants with missing patient characteristic data were not excluded, because these variables were not used in the statistical analysis.

All CIs are two-sided with 95% confidence and calculated from 1000 bootstrap samples. DNNs were implemented in Python (version 3.6.4) using TensorFlow (version 1.11), and all boosted trees using the Python interface for XGBoost (version 0.72; appendix p 5).

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Among the 59 159 STHLM3 participants, 7406 (12·5%) underwent systematic biopsy according to a standardised protocol consisting of 10 or 12 needle cores, with 12 cores being taken from prostates larger than 35 mL (figure 1; tables 1, 2). Among the biopsied participants, we randomly selected 1297, stratified by ISUP score, to be included in this study. After excluding slides with mismatched annotations or technical issues, we randomly split the remaining participants into training and test datasets, resulting in 6682 STHLM3 cores to be used for training of the AI system. We added another 271 cores from outside the study to the training dataset. The data are representative for a screening by invitation setting and include various diagnostically challenging cancer variants encountered in clinical practice (appendix p 35).

The AUC representing the ability of the AI system to distinguish malignant from benign cores was 0·997 (95% CI 0·994–0·999) for the independent test dataset (benign=910, malignant=721) and 0·986 (0·972–0·996) for the external validation dataset (benign=108, malignant=222; figure 2). When ISUP 3–5 cases were removed, AUC values were 0·996 (0·992–0·999) for the independent test dataset and 0·980 (0·959–0·995) for the external validation dataset (appendix p 27). The performance of the AI system for cancer detection is summarised in table 3.

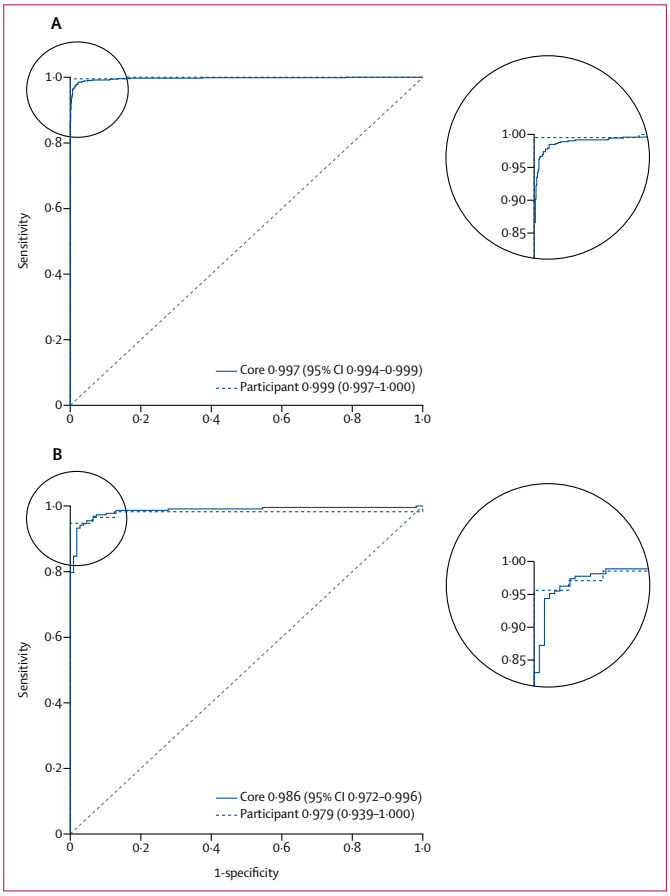

Figure 2: **Receiver operating characteristic curves and AUC for cancer detection in individual cores and individual participants**
(A) Independent test dataset. (B) External validation dataset. Dashed grey lines represent the baseline curve corresponding to random guessing. AUC=area under the curve.

A visualisation of the estimated localisation of malignant tissue for an example biopsy is presented in the appendix (p 33) and the correlation between the cancer length estimates of the AI system and the measurements of the pathologist is presented in figure 3. The correlation between cancer length predicted by the AI and assigned by the reporting pathologist was 0·96 (95% CI 0·95–0·97) for the independent test dataset and 0·87 (0·84–0·90) for the external validation dataset. Further randomly selected example biopsies can be

	Avoided benign biopsy cores, n (specificity)	Detected cancer biopsy cores, n (sensitivity)	Missed cores with cancer by ISUP score, n(%)					Missed men with cancer, n (%)
			ISUP 1	ISUP 2	ISUP 3	ISUP 4	ISUP 5	
Independent test dataset								
Example operating point 1—sensitivity ≥99·9	570 (62·6%)	720 (99·9%)	0	1 (0·7%)	0	0	0	0
Example operating point 2—sensitivity ≥99·6	788 (86·6%)	718 (99·6%)	2 (0·6%)	1 (0·7%)	0	0	0	0
Example operating point 3—sensitivity ≥99·3	809 (88·9%)	716 (99·3%)	4 (1·1%)	1 (0·7%)	0	0	0	0
Example operating point 4—sensitivity ≥99·0	864 (94·9%)	714 (99·0%)	4 (1·1%)	2 (1·4%)	0	0	1 (1·4%)	1 (0·5%)
External validation								
Example operating point 1—sensitivity ≥99·5	49 (45·4%)	221 (99·5%)	1 (1·5%)	0	0	0	0	1 (1·8%)
Example operating point 2—sensitivity ≥99·1	78 (72·2%)	220 (99·1%)	2 (3·1%)	0	0	0	0	1 (1·8%)
Example operating point 3—sensitivity ≥98·6	94 (87·0%)	219 (98·6%)	3 (4·6%)	0	0	0	0	1 (1·8%)
Example operating point 4—sensitivity ≥97·7	97 (89·8%)	217 (97·7%)	3 (4·6%)	1 (1·6%)	1 (2·0%)	0	0	1 (1·8%)

Presented for each operating point are the number of benign biopsy cores that could be discarded from further consideration (specificity), the number of correctly detected malignant biopsy cores needing pathological evaluation (sensitivity), the number of missed malignant cores by ISUP score (percentage of all cores with the given ISUP score), and the number of missed men (percentage of all men with cancer). ISUP=International Society of Urological Pathology.

Table 3: **Sensitivity and specificity at selected points on the receiver operating characteristic curves for cancer detection**

For **TissUUmaps** see https://tissuumaps.research.it.uu.se/sthlm3/

inspected using TissUUmaps, an online tool for interactive examination of predictions alongside the core tissue. Results of model interpretation are shown in the appendix (pp 31–32).

For Gleason grading, the mean pairwise kappa achieved by the AI system on the 87 Imagebase cases was 0·62. The pathologists had values ranging from 0·60 to 0·73, and the study pathologist (LE) had a kappa of 0·73. When considering a narrower grouping of ISUP grades (ISUP 1, 2–3, and 4–5), which often forms the basis for primary treatment selection, the AI system scored higher than when considering all ISUP grades (figure 4A). The grades assigned by the panel and the AI to each Imagebase case are shown in the appendix (p 26).

The kappa obtained by the AI system relative to the pathology report in the independent test dataset of 1631 cores was 0·83 (figure 4B). The kappa on the external validation dataset was 0·70 (figure 4C). By scaling the ISUP probabilities before assigning the predicted class (calibrating to the new site), the kappa increased to 0·76 on the external validation data (figure 4D). Moreover, we compared the predictions of the AI system and the pathologist in terms of PSA relapses among the participants in the test dataset who underwent radical prostatectomy (appendix pp 22,36)

Discussion

We have shown that an AI system based on DNNs can achieve excellent discrimination between benign biopsy cores versus cores containing cancer and that the time-consuming task of measuring cancer length can be automated with high precision. Moreover, we have shown that an AI system can grade prostate biopsies within the performance range of highly experienced urological pathologists.

Owing to the poor discriminative ability of the PSA test and the systematic biopsy protocol of 10–12 needle cores, which is still in common use, most biopsies encountered in clinical practice are of benign tissue. To reduce the workload of assessing these samples, we evaluated the AI system's potential to assist the pathologist by prescreening benign from malignant cores. Because the pathology report was used as gold standard for this evaluation, the AI system, by design, cannot achieve a higher sensitivity than the reporting pathologist. However, the sensitivity of the AI system could in fact be higher, because some malignant cores might be overlooked by the pathologist but detected by the AI. For example, Ozkan and colleagues[5] evaluated the agreement of two pathologists in the assessment of cancer in biopsy cores. Following examination of 407 cases, one pathologist found cancer in 231 cases, and the other found cancer in 202 cases. This finding suggests that an AI system could not only streamline the workflow, but also improve sensitivity by detecting cancer foci that would otherwise be accidentally overlooked.

The first attempt to use DNNs for the detection of cancer on prostate biopsies was reported by Litjens and colleagues.[15] Using an approach similar to ours, but based on a small dataset, they could safely exclude 32% of benign

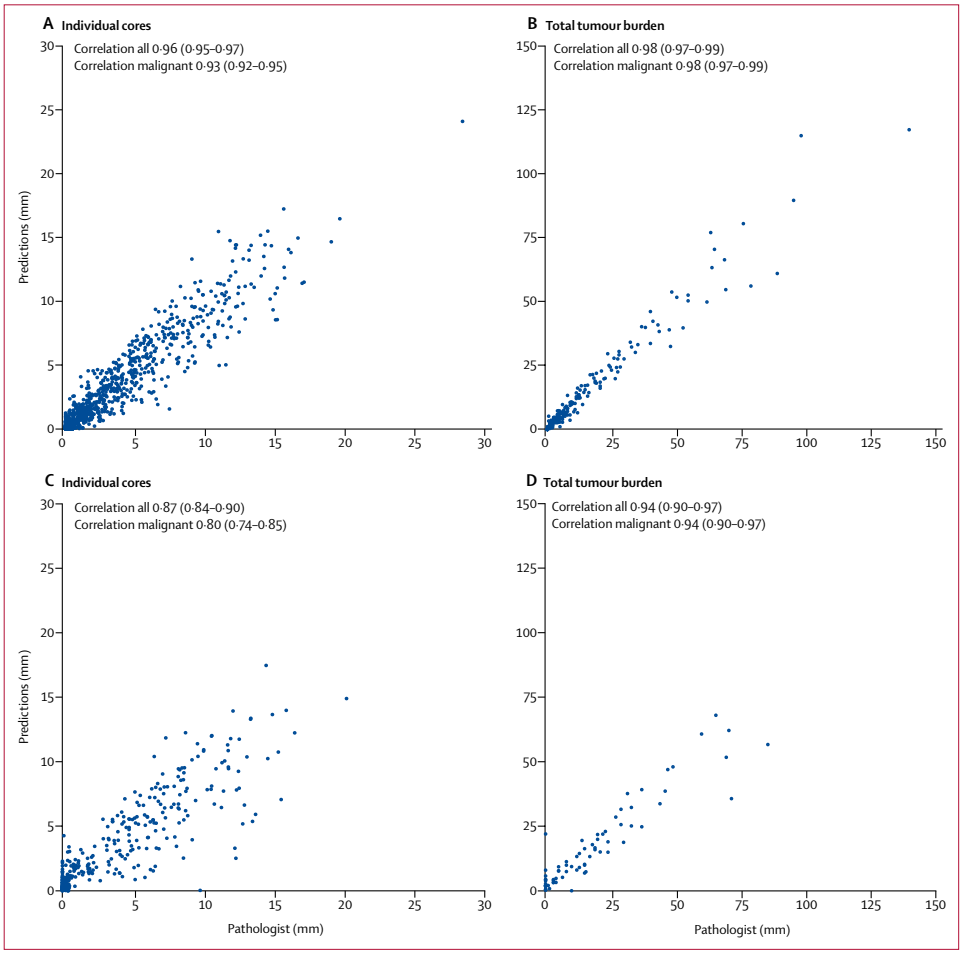

Figure 3: **Concordance between cancer lengths estimated by the AI system and the pathologist**
(A) Individual cores in the independent test dataset. (B) Total tumour burden (per participant) in the independent test dataset. (C) Individual cores in the external validation dataset. (D) Total tumour burden (per participant) in the external validation dataset. Corresponding linear correlation coefficients computed for all cores and malignant cores only are shown in each plot. Datapoints in the left plot are scattered along the x-axis for clarity.

cores. Campanella and colleagues[16] showed an AUC of 0·991 for cancer detection on an independent test dataset and 0·943 on external validation data. Attempts at grading of prostate tissue derived from prostatectomy or based on tissue microarrays have also been made.[14,26] None of these studies achieved expert urological pathologist-level consistency in Gleason grading, estimated tumour burden, or investigated grading on needle biopsies, which is notable because this type of sampling is used for diagnosis and grading in virtually every pathology laboratory worldwide. To the best of our knowledge, no previous study has used a well defined cohort of samples to estimate the clinical implications, with respect to key medical operating characteristic metrics, such as sensitivity and specificity.[27]

The strengths of our study include the use of well controlled data collected within the STHLM3 trial, which included standardised biopsy procedures, centralised pathology reporting, and blinding of both the urologists and the pathologist to clinical characteristics, such as PSA. The prospectively collected, population-based data cover a large random sample of men. Prostate cancers diagnosed in STHLM3 are representative for a screening-by-invitation setting, and the data include cancer variants that are difficult to diagnose (pseudohyperplastic and atrophic carcinoma), slides that required immunohistochemistry,

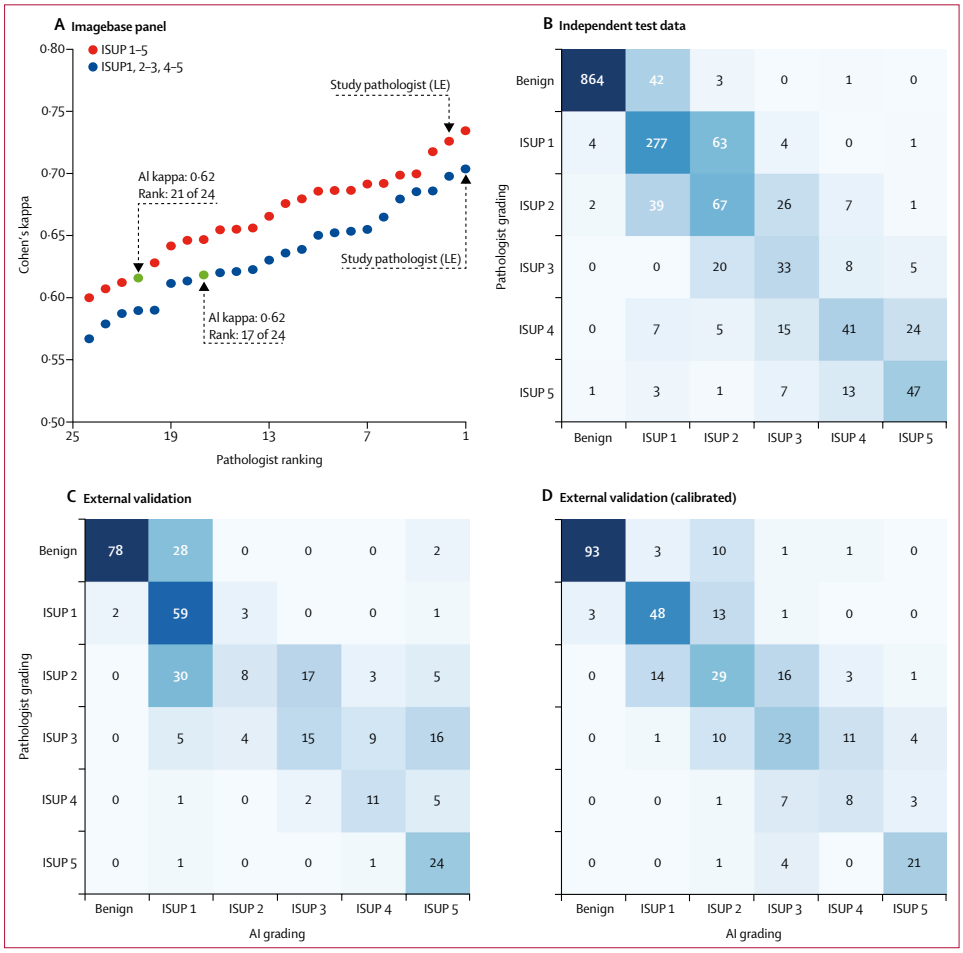

Figure 4: **Gleason grading performance on test data**
(A) Cohen's kappa for each pathologist ranked from lowest to the highest. Each kappa value is the average pairwise kappa for each of the pathologists compared with the others. To account for the natural order of the ISUP scores, we used linear weights. The AI is highlighted with a green dot and an arrow. The study pathologist (LE) is highlighted with an arrow. Values computed based on all five ISUP scores are plotted in red, whereas values based on a grouping of ISUP scores commonly used for treatment decision are shown in blue. (B) A confusion matrix on the independent test data of 1631 slides. (C) A confusion matrix on the external validation data of 330 slides. (D) Results on external validation data following calibration of the slide-level model. The blue shading represents the number of cores in each cell of the matrix. This procedure did not involve any model retraining. The results are presented for an operating point achieving a minimum cancer detection sensitivity of 99%. AI=artificial intelligence. ISUP=International Society of Urological Pathology.

benign mimickers of cancer, slides with thick cuts, and fragmented cores and poor staining. Despite these difficult cases, the AI system achieved excellent diagnostic concordance with the study pathologist. Furthermore, we confirmed that the enrichment of high-grade cases in our datasets did not result in optimistic estimates of discriminative performance. The study was subjected to a strict protocol, in which the splitting of cases into training and test datasets was performed at a patient level and all analyses were prespecified before the evaluation of the independent test dataset, including code for producing tables, figures, and result statistics. A further strength is the use of Imagebase, which is a unique dataset for testing the performance of the AI against highly experienced urological pathologists.

We trained the AI system using annotations from a single, highly experienced urological pathologist (LE). The decision to rely on a single pathologist for model

training was done to avoid presenting the system with conflicting labels for the same morphological patterns and to thereby achieve more consistent predictions. The study pathologist has shown high concordance with other experienced urological pathologists in several studies,[28,29] and therefore represents a good reference for model training. For model evaluation, however, it is crucial to assess performance against multiple pathologists.

Technical variability is introduced during slide preparation and scanning, which might affect the predictions of the AI system. Given the sensitivity of DNNs to differences in input data, differences across labs and scanners could invalidate any discriminatory capacity of a DNN.[30] Here, we showed that the capacity of the AI in discriminating between benign and malignant biopsies decreased, but remained excellent, in the external validation data compared with the independent test dataset. We did, however, observe some reduction in performance with respect to cancer length predictions and overall Gleason grading. By contrast with cancer detection, in which only a handful of correctly predicted patches might be sufficient, cancer length estimation relies on all patches being correctly predicted. Thus, imperfect generalisation is likely to first manifest itself in the length estimates. The reduction in grading performance was most notable for ISUP 2 grades. However, by scaling the AI's predictions for the different classes (ie, calibrating five scalar parameters to the new site), the results were more similar to the results achieved on the independent test data. This is a key observation, because it suggests that although some fine tuning to a new site or scanner is likely required to achieve optimal performance, this tuning is lightweight and can be done using little data. Notably, it does not require redevelopment or retraining of either the DNN models or the slide-level models, which would be infeasible both from a practical and regulatory perspective. Albeit a limitation of the method, requirement for such calibration is not uncommon when using a diagnostic test at a new site (eg, calibrants are routinely used in laboratory diagnostics to diagnose and prevent site-specific differences and variation in test results over time) and is unlikely to present a major hurdle for the clinical application of AI-based diagnostics.

A limitation of this study is the absence of exact pixel-wise annotations, because the annotations might highlight regions that include a mixture of benign and malignant glands of different grades. To address this issue, we trained the algorithm on slides with pure Gleason grades, used a patch size large enough to cover glandular structures, but small enough to minimise the presence of mixed grades within a patch, and we focused our attention on core and patient performance metrics, which avoids caveats of patch-level evaluation and is clinically more meaningful. Another limitation is the difficulty of using a subjective measure like ISUP grade as ground truth for AI models. We approached this problem by evaluating the ISUP grade assigned by the AI against a panel of experienced pathologists. We also confirmed that the classifications of the AI did not substantially differ from the pathologist's when evaluating PSA relapses among the operated men in the trial.

We believe that the use of an AI system like the one presented in this Article could increase sensitivity and promote patient safety by focusing the attention of the pathologist on regions of interest, reduce pathology workload by automated culling of benign biopsies, and reduce the high intra-observer variability in the reporting of prostate histopathology by producing reproducible decision support for grading. A further benefit is that AI can provide diagnostic expertise in regions where it is unavailable.

Contributors

ME had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. PS and KK contributed equally to algorithmic design, implementation, and drafting the manuscript. Additionally, PS was mainly responsible for statistical analysis of results and KK was mainly responsible for high-performance computing. HO was mainly responsible for data management and participated in algorithmic design and implementation, and in drafting the manuscript. LS developed the online viewer application allowing visual examination of results. BD was involved in drafting the manuscript. BD, DMB, DGB, LE, AJE, DJG, PAH, KAI, JGK, GK, THVDK, KRML, JKMK, JO, C-CP, HS, JRS, HT, TT, MV, and MZ did grading of the Imagebase dataset and provided pathology expertise and feedback. CL was involved in data collection. JL was involved in study design. PR and CW contributed to design and supervision of the study and to algorithmic design. Additionally, PR contributed to high-performance computing and CW contributed to designing the online viewer. HG contributed to the conception, design and supervision of the study. MR contributed to the conception, design and supervision of the study and to algorithmic design. LE graded and annotated all the data used in the study, contributed to the conception, design, and supervision of the study, and helped draft the manuscript. ME was responsible for the conception, design and supervision of the study, and contributed to algorithmic design, analysis of results and drafting the manuscript. All authors participated in the critical revision and approval of the manuscript.

Declaration of interests

PS and KK are named on a pending patent (1900061-1) related to cancer diagnostics quality control. HG has five patents (WO2013EP74259 20131120, WO2013EP74270 20131120, WO2018EP52473 20180201, WO2015SE50272 20150311, and WO2013SE50554 20130516) related to prostate cancer diagnostics pending, and has patent applications licensed to Thermo Fisher Scientific. ME has four patents (WO2013EP74259 20131120, WO2013EP74270 20131120, WO2018EP52473 20180201, and WO2013SE50554 20130516) related to prostate cancer diagnostics pending, and has patent applications licensed to Thermo Fisher Scientific, and is named on a pending patent (1900061-1) related to cancer diagnostics quality control. Karolinska Institutet collaborates with Thermo Fisher Scientific in developing the technology for the STHLM3 study. All other authors declare no competing interests.

Acknowledgments

Funding was provided by the Swedish Research Council, Swedish Cancer Society, Swedish Research Council for Health, Working Life, and Welfare, Swedish eScience Research Center, Walter Ahlström Foundation, Tutkijat maailmalle programme, Academy of Finland (313921), Cancer Society of Finland, Emil Aaltonen Foundation, Finnish Foundation for Technology Promotion, Industrial Research Fund of Tampere University of Technology, KAUTE Foundation, Orion Research Foundation, Svenska Tekniska Vetenskapsakademien i Finland, Tampere University Foundation, Tampere University graduate school, The Finnish Society of Information Technology and Electronics, TUT

on World Tour programme, the European Research Council (grant ERC-2015-CoG 682810), and EIT Health. The Tampere Center for Scientific Computing and CSC-IT Center for Science, Finland are acknowledged for providing computational resources. ME and MR report funding from the Swedish Research Council and Swedish Cancer Society. ME reports funding from the Swedish Research Council for Health, Working Life, and Welfare, and Swedish eScience Research Center. The Saint Göran Hospital, Stockholm, is acknowledged for providing additional high-grade slides as training data. Carin Cavalli-Björkman, Britt-Marie Hune, Astrid Björklund, and Olof Cavalli-Björkman have been instrumental in logistical handling of the glass slides. Hannu Hakkola, Tomi Häkkinen, Leena Latonen, Kaisa Liimatainen, Teemu Tolonen, Masi Valkonen, and Mira Valkonen are acknowledged for their helpful advice. We thank the participants in the Stockholm-3 study for their participation.

References

1 Loeb S, Carter HB, Berndt SI, Ricker W, Schaeffer EM. Complications after prostate biopsy: data from SEER-Medicare. *J Urol* 2011; **186:** 1830–34.

2 Egevad L, Delahunt B, Samaratunga H, et al. The International Society of Urological Pathology Education web—a web-based system for training and testing of pathologists. *Virchows Arch* 2019; **474:** 577–84.

3 Adesina A, Chumba D, Nelson AM, et al. Improvement of pathology in sub-Saharan Africa. *Lancet Oncol* 2013; **14:** e152–57.

4 Robboy SJ, Weintraub S, Horvath AE, et al. Pathologist workforce in the United States: I. Development of a predictive model to examine factors influencing supply. *Arch Pathol Lab Med* 2013; **137:** 1723–32.

5 Ozkan TA, Eruyar AT, Cebeci OO, Memik O, Ozcan L, Kuskonmaz I. Interobserver variability in Gleason histological grading of prostate cancer. *Scand J Urol* 2016; **50:** 420–24.

6 Melia J, Moseley R, Ball RY, et al. A UK-based investigation of inter- and intra-observer reproducibility of Gleason grading of prostatic biopsies. *Histopathology* 2006; **48:** 644–54.

7 Bejnordi BE, Veta M, Van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017; **318:** 2199–210.

8 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542:** 115–18.

9 Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. *Nature* 2016; **529:** 484–89.

10 Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; **316:** 2402–10.

11 Gummeson A, Arvidsson I, Ohlsson M, et al. Automatic Gleason grading of H and E stained microscopic prostate images using deep convolutional neural networks. In: Gurcan MN, Tomaszewski JE, eds. Proceedings of SPIE. Volume 10140. Medical Imaging 2017: Digital Pathology. Bellingham, WA: SPIE, 2017.

12 Källén H, Molin J, Heyden A, Lundström C, Åström K. Towards grading gleason score using generically trained deep convolutional neural networks. 2016 IEEE 13th International Symposium on Biomedical Imaging; Prague; 2016 (1163–167).

13 Jiménez del Toro O, Atzori M, Otálora S, et al. Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score. In: Gurcan MN, Tomaszewski JE, eds. Proceedings of SPIE. Volume 10140. Medical Imaging 2017: Digital Pathology. Bellingham, WA: SPIE, 2017.

14 Arvaniti E, Fricker KS, Moret M, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Sci Rep* 2018; **8:** 12054.

15 Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep* 2016; **6:** 26286.

16 Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019; **25:** 1301–09.

17 Grönberg H, Adolfsson J, Aly M, et al. Prostate cancer screening in men aged 50–69 years (STHLM3): a prospective population-based diagnostic study. *Lancet Oncol* 2015; **16:** 1667–76.

18 Ström P, Nordström T, Aly M, Egevad L, Grönberg H, Eklund M. The Stockholm-3 model for prostate cancer detection: algorithm update, biomarker contribution, and reflex test potential. *Eur Urol* 2018; **74:** 204–10.

19 Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. The 2014 International Society of Urological Pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma definition of grading patterns and proposal for a new grading system. *Am J Surg Pathol* 2016; **40:** 244–52.

20 Nir G, Karimi D, Goldenberg SL, et al. Comparison of artificial intelligence techniques to evaluate performance of a classifier for automatic grading of prostate cancer from digitized histopathologic images. *JAMA Netw Open* 2019; **2:** e190442.

21 Egevad L, Delahunt B, Berney DM, et al. Utility of Pathology Imagebase for standardisation of prostate cancer grading. *Histopathology* 2018; **73:** 8–18.

22 Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition; Las Vegas; 2016 (2818–26).

23 Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition; Miami; 2009 (248–55).

24 Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco; 2016 (785–94).

25 van der Maaten L, Hinton GE. Visualizing high-dimensional data using t-SNE. *J Mach Learn Res* 2008; **9:** 2579–605.

26 Nagpal K, Foote D, Liu Y, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med* 2019; **2:** 48.

27 Nature. AI diagnostics need attention. *Nature* 2018; **555:** 285.

28 Kweldam CF, Nieboer D, Algaba F, et al. Gleason grade 4 prostate adenocarcinoma patterns: an interobserver agreement study among genitourinary pathologists. *Histopathology* 2016; **69:** 441–49.

29 Egevad L, Cheville J, Evans AJ, et al. Pathology Imagebase—a reference image database for standardization of pathology. *Histopathology* 2017; **71:** 677–85.

30 Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. *arXiv* 2015; published online March 20. 1412.6572 (preprint).