From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

ARTIFICIAL INTELLIGENCE IN HISTOPATHOLOGY IMAGE ANALYSIS FOR CANCER PRECISION MEDICINE

Philippe Weitz

Karolinska Institutet

Stockholm 2023

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2023

ISBN 978-91-8017-148-9
Cover illustration: ACROBAT challenge logo, designed by Masi Valkonen

Artificial Intelligence in Histopathology Image Analysis for Cancer Precision Medicine

Thesis for Doctoral Degree (Ph.D.)

By

Philippe Weitz

The thesis will be defended in public at the lecture hall Atrium, Nobels väg 12B, Karolinska Institutet, 171 65 Solna, at 1:00 PM on the 27th of October 2023.

Principal Supervisor:
Dr. Mattias Rantalainen
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Co-supervisor(s):
Professor Johan Hartman
Karolinska Institutet
Department of Oncology-Pathology

Professor Martin Eklund
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Professor Henrik Grönberg
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Opponent:
Professor Anne Martel
University of Toronto
Department of Medical Biophysics

Examination Board:
Professor Gustaf Edgren
Karolinska Institutet
Department of Medicine
Division of Clinical Epidemiology

Professor Pernilla Wikström
Umeå University
Department of Medical Biosciences

Professor Claes Lundström
Linköping University
Department of Science and Technology

Dedicated to Moritz, Hannah, Birgit and Josef.

Abstract

In recent years, there have been rapid advancements in the field of computational pathology. This has been enabled through the adoption of digital pathology workflows that generate digital images of histopathological slides, the publication of large data sets of these images and improvements in computing infrastructure. Objectives in computational pathology can be subdivided into two categories, first the automation of routine workflows that would otherwise be performed by pathologists and second the addition of novel capabilities. This thesis focuses on the development, application, and evaluation of methods in this second category, specifically the prediction of gene expression from pathology images and the registration of pathology images among each other.

In Study I, we developed a computationally efficient cluster-based technique to perform transcriptome-wide predictions of gene expression in prostate cancer from H&E-stained whole-slide-images (WSIs). The suggested method outperforms several baseline methods and is non-inferior to single-gene CNN predictions, while reducing the computational cost with a factor of approximately 300. We included 15,586 transcripts that encode proteins in the analysis and predicted their expression with different modelling approaches from the WSIs. In a cross-validation, 6,618 of these predictions were significantly associated with the RNA-seq expression estimates with FDR-adjusted p-values <0.001. Upon validation of these 6,618 expression predictions in a held-out test set, the association could be confirmed for 5,419 (81.9%). Furthermore, we demonstrated that it is feasible to predict the prognostic cell-cycle progression score with a Spearman correlation to the RNA-seq score of 0.527 [0.357, 0.665].

The objective of Study II is the investigation of attention layers in the context of multiple-instance-learning for regression tasks, exemplified by a simulation study and gene expression prediction. We find that for gene expression prediction, the compared methods are not distinguishable regarding their performance, which indicates that attention mechanisms may not be superior to weakly supervised learning in this context.

Study III describes the results of the ACROBAT 2022 WSI registration challenge, which we organised in conjunction with the MICCAI 2022 conference. Participating teams were ranked on the median 90th percentile of distances between registered and annotated target landmarks. Median 90th percentiles for eight teams that were eligible for ranking in the test set consisting of 303 WSI pairs

ranged from 60.1 µm to 15,938.0 µm. The best performing method therefore has a score slightly below the median 90th percentile of distances between first and second annotator of 67.0 µm.

Study IV describes the data set that we published to facilitate the ACROBAT challenge. The data set is available publicly through the Swedish National Data Service SND and consists of 4,212 WSIs from 1,153 breast cancer patients.

Study V is an example of the application of WSI registration for computational pathology. In this study, we investigate the possibility to register invasive cancer annotations from H&E to KI67 WSIs and then subsequently train cancer detection models. To this end, we compare the performance of models optimised with registered annotations to the performance of models that were optimised with annotations generated for the KI67 WSIs. The data set consists of 272 female breast cancer cases, including an internal test set of 54 cases. We find that in this test set, the performance of both models is not distinguishable regarding performance, while there are small differences in model calibration.

List of scientific papers

I. **P. Weitz**, Y. Wang, K. Kartasalo, L. Egevad, J. Lindberg, H. Grönberg, M. Eklund, M. Rantalainen, "Transcriptome-wide prediction of prostate cancer gene expression from histopathology images using co-expression-based convolutional neural networks," *Bioinformatics*, vol. 38, no. 13, pp. 3462–3469, Jun. 2022.

II. **P. Weitz**, Y. Wang, J. Hartman, and M. Rantalainen, "An investigation of attention mechanisms in histopathology whole-slide-image analysis for regression objectives," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, IEEE, Oct. 2021.

III. **P. Weitz***, M. Valkonen*, L. Solorzano*, C. Carr, K. Kartasalo, C. Boissin, S. Koivukoski, A. Kuusela, D. Rasic, Y. Feng, S. Sinius Pouplier, A. Sharma, K. Ledesma Eriksson, S. Robertson, C. Marzahl, C. Gatenbee, A. Anderson, M. Wodzinski, A. Jurgas, N. Marini, M. Atzori, H. Müller, D. Budelmann, N. Weiss, S. Heldmann, J. Lotz, J. Wolterink, B. De Santi, A. Patil, A. Sethi, S. Kondo, S. Kasai, K. Hirasawa, M. Farrokh, N. Kumar, R. Greiner, L. Latonen, A. Laenkholm, J. Hartman, P. Ruusuvuori*, M. Rantalainen*, "The ACROBAT 2022 Challenge: Automatic Registration Of Breast Cancer Tissue",
Manuscript: doi.org/10.48550/arXiv.2305.18033

IV. **P. Weitz***, M. Valkonen*, L. Solorzano*, C. Carr, K. Kartasalo, C. Boissin, S. Koivukoski, A. Kuusela, D. Rasic, Y. Feng, S. Sinius Pouplier, A. Sharma, K. Ledesma Eriksson, L. Latonen, A. Laenkholm, J. Hartman*, P. Ruusuvuori*, M. Rantalainen*, "A Multi-Stain Breast Cancer Histological Whole-Slide-Image Data Set from Routine Diagnostics," *Scientific Data*, vol. 10, no. 1, p. 562, Aug. 2023.

V. **P. Weitz**, V. Sartor, B. Acs, S. Robertson, D. Budelmann, J. Hartman, M. Rantalainen, „Increasing the usefulness of already existing annotations through WSI registration",
Manuscript: doi.org/10.48550/arXiv.2303.06727

* Equal contribution.

TABLE OF CONTENTS

List of abbreviations

ACROBAT	AutomatiC Registration Of Breast cAncer Tissue
ADR	Active Data Repository
AI	Artificial Intelligence
anti-HER2	HER2-Targeted Therapy
AUROC	Area Under the Receiver Operating Characteristic Curve
BH	Benjamini-Hochberg
CCP	Cell-Cycle Progression
ChT	Chemotherapy
Clinseq	Clinical Sequencing of Cancer in Sweden
CNN	Convolutional Neural Network
CPH	Cox Proportional Hazards
DBA	Distance between First and Second Annotator
DCIS	Ductal Carcinoma In Situ
DICOM	Digital Imaging And Communications in Medicine
DL	Deep Learning
DRE	Digital Rectal Exam
ER	Estrogen Receptor
ESMO	European Society of Medical Oncology
ET	Endocrine Therapy
FCNN	Fully Connected Neural Network
FDR	False Discovery Rate
FFPE	Formalin-Fixed Paraffin-Embedded
FISH	Fluorescence In Situ Hybridisation
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
H&E	Haematoxylin & Eosin

HER2	Human Epidermal Growth Factor Receptor 2
HR	Hazard Ratio
HSV	Hue, Saturation, Value
IC	Invasive Cancer
IHC	Immunohistochemistry
INCA	Information Network for Cancer Care
LCIS	Lobular Carcinoma In Situ
LME	Linear Mixed Effect
MICCAI	Medical Image Computing and Computer Assisted Intervention
MIL	Multiple-Instance Learning
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
NHG	Nottingham Histological Grade
NKBC	National Quality Registry for Breast Cancer
PGR	Progesterone Receptor
PSA	Prostate Specific Antigen
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
RGB	Red, Green, Blue
ROI	Region of Interest
SCAN-B	Sweden Cancerome Analysis Network – Breast
SND	Swedish National Data Service
ST	Spatial Transcriptomics
TCGA	The Cancer Genome Atlas
WSI	Whole-Slide-Image

1 Introduction

The histopathological assessment of tissue samples is an essential component of both biomedical research, as well as clinical routine to establish diagnosis and prognosis of many diseases such as cancer. Historically, this assessment was based on the analysis of stained tissue slides with a microscope. However, during the last decade, digital pathology methods have been established as clinically non-inferior and found broader dissemination into pathology departments. In digital pathology workflows, the direct examination of slides with a microscope is at least partially replaced through scanning the slides and their inspection on screens. While this has potential benefits such as the possibility of remote consultations and the analysis of digital images with specialised software tools, the broad adoption of digital pathology has been slow. This might e.g. be due to high initial costs of slide scanners and unclear cost or time saving benefits.

In recent years, deep learning based on convolutional neural networks (CNNs), often also referred to as artificial intelligence (AI), has substantially advanced many areas of research and technology that require the analysis of complex non-tabular data. This data includes time series, images, tomography volumes and videos. The analysis of histopathology images with these tools and the application of current image processing techniques are often referred to as computational pathology. Computational pathology has the potential to both automate routine diagnostics, as well as to add new capabilities for pathological analysis. Particularly the latter may facilitate broader access to precision diagnostics.

Breast and prostate cancer are among the most common cancers both globally and in Sweden. For both of these cancers, there are several diagnostic tests that are based on the expression of genes that are associated with the respective cancer. These can provide prognostic and predictive information and guide treatment decisions. However, due to costs and complex logistics, these tests are not available to all patients, particularly globally. One of the aims of this thesis is therefore the development, application and evaluation of methods that predict gene expression from pathology images.

For many diseases, it is furthermore common to not only create a single diagnostic slide from a tissue sample but multiple, which are stained with different chemicals to reveal different types of clinically relevant information. However, during sample preparation, tissue can deform easily before it is scanned. The alignment of corresponding tissue regions to each other in images is referred to as registration.

An alignment of pathology images is desirable to facilitate research applications such as stain-guided learning and 3D-reconstruction. Clinical use cases include the identification of regions of interest for biomarker scoring and the investigation of resection margins with multiple stains. The registration of pathology images is particularly challenging due to the large image sizes and non-linear deformations of the tissue. This thesis therefore also aims at the application and evaluation of image registration methods for computational pathology.

The overarching goal of this thesis is therefore the development, application and evaluation of computational pathology methods that expand the capabilities of current pathological diagnostics. This has the potential to advance the quality and access to precision diagnostics, with the associated potential benefits for patients.

2 Background

2.1 Cancer

The term cancer refers to a set of diseases that are characterised through uncontrollable growth and spread of cells. The malignant tumours of solid cancers differ from benign tumours in other neoplastic diseases through their capability to invade surrounding tissue. Hanahan and Weinberg suggested eight hallmarks of cancer and two enabling capabilities [1], [2], which provide a common model to aid in understanding this complex disease. These eight hallmarks of cancer cells are sustained proliferative signalling, the evasion of growth suppressors, replicative immortality, the activation of invasion and metastasis, the induction of angiogenesis and resistance to cell death, as well as the avoidance of immune destruction and the deregulation of cellular energetics. The two enabling characteristics are tumour-promoting inflammation and genome instability and mutation. Cancer is often referred to as a genetic disease, which is caused by mutations within a cell's DNA. These mutations can be inherited or occur during an organism's life span. While modern lifestyle factors such as exercise, diet, alcohol, and tobacco consumption, are associated with increases (or decreases) in an individual's risk of cancer, it is not only a disease that occurs in modern societies. One of the earliest samples of malignant neoplastic disease in a human ancestor dates back 1.6 to 1.8 million years. This hominin metatarsal (bone of the forefoot) specimen found in South Africa contains a malignant osteosarcoma [3]. Presently, cancer is a leading cause of death globally, with an estimated 10 million or one out of six deaths attributed to it [4]. During cancer diagnosis, solid cancers are often described with the TNM staging system, where T captures the size and extent of the primary tumour, N the number of local lymph nodes that contain metastases and M whether there are distant metastases. Besides staging, cancer diagnostics involves a histopathological assessment, which follows criteria that are specific to the tissue of origin of the tumour.

2.1.1 Histopathological Assessment

Histopathological assessment is a key component in cancer diagnostics both to establish a cancer diagnosis, characterise the cancer, as well as to guide treatment decisions. It requires a multi-step sample preparation process, which can vary between laboratories and therefore lead to inter-laboratory variability.

Once a tissue sample is obtained e.g. through a core needle biopsy or surgery, it is typically preserved by fixation in formalin or formaldehyde. This aims to preserve morphological structures and molecules. The sample is then embedded in a paraffin wax block. From this block, slices with a thickness of typically 5-15 micrometres can be cut. These are then mounted on glass slides and the wax is removed with heat and solvents such as xylene. Subsequently, a stain can be applied. The most common stain is haematoxylin and eosin (H&E). Haematoxylin is positively charged and binds to the negatively charged cell nuclei, colouring them in blue-purple. The negatively charged eosin binds to the positively charged extracellular matrix and cytoplasm, staining them pink. Other structures may bind both chemicals in varying proportions, leading to intermediate colours. This allows pathologists to identify morphological patterns and the distributions of cells in a tissue sample. H&E staining is the de-facto standard for morphological assessment of histological samples. Another important staining technique is immunohistochemistry (IHC). IHC staining is based on antibodies that bind to specific antigens, typically proteins. Stains that bind to these antibodies can then be used to visualise the presence and distribution of these specific proteins. This allows e.g. to assess whether a cancer-related protein is overexpressed, which can indicate specific targeted treatments that interact with this protein or provide information on the proliferation rate of the cancer cells.

2.2 Breast Cancer

Breast cancer refers to tumours that develop from the epithelial cells of the breast. Precancerous lesions, also referred to as carcinoma *in situ*, are categorized as ductal carcinoma in situ (DCIS) or lobular carcinoma in situ (LCIS), depending on their site of origin. When the tumour invades the surrounding tissue, it is considered invasive cancer (IC), signifying a more advanced stage of the disease. The majority of breast cancers are adenocarcinomas, originating from the ducts or lobules of the breast, and are classified as either ductal carcinoma or lobular carcinoma. Ductal carcinomas account for 70–75% and lobular carcinomas for 12–15% of breast cancers. The remaining cases can be categorised into 18 further rare histological subtypes [5].

2.2.1 Epidemiology

With an estimated 2.3 million new cases in 2020, breast cancer has now surpassed lung cancer as the most commonly diagnosed cancer globally [4]. Furthermore, breast cancer is the cancer with the highest age-standardised

incidence rate, with 47.8 cases per 100,000 person-years [4]. In Sweden, there were approximately 7,500 newly diagnosed breast cancer cases and 1,500 deaths attributable to the disease in 2020 [4]. While the incidence of breast cancer has been increasing in Sweden, the mortality rate associated with it has been on a decreasing trend, which might be attributed to the implementation of population-wide screening programs and improved treatments. Reproductive and hormonal risk factors include low age at menarche, high age at menopause, high age at first birth, low number of children, less breastfeeding, menopausal hormone therapy, and oral contraception. Lifestyle risk factors include alcohol consumption, excess body weight and physical inactivity [6]. While some studies found an association between smoking and breast cancer, particularly in pre-menopausal women, further research is needed to firmly establish smoking as a risk factor [7].

2.2.2 Screening and Diagnosis

Since 1997, mammography screening has been recommended for women aged between 40 and 74 in Sweden. The Swedish national standardised care workflow, as depicted in Figure 1, is initiated either if the patient seeks care due to symptoms or due to a suspicious lesion that is detected during screening. If there is a well-founded suspicion of breast cancer, a triple-diagnostic process is initiated. This process includes clinical examination, imaging diagnostics, and morphological diagnostics of biopsy and surgical specimens.

2.2.3 Histopathological Assessment

The histopathological assessment of biopsy and resection specimen is a key element of the diagnostic process of breast cancer that guides decisions, including everything from neoadjuvant treatment to surgery and adjuvant treatments. Postoperative pathological assessments should include the number, location and maximum diameter of resected tumours, the number of removed lymph nodes and how many of these were positive for metastases, an evaluation of resection margins that includes the minimum distance of the margin, vascular invasion, histological type, grade, and biomarker statuses [5]. There are different grading systems for carcinoma in situ and invasive breast cancer. Both grading systems are subject to inter-assessor variability, but particularly grading of in situ carcinomas has only moderate reproducibility [9]. Grading of invasive cancer in Sweden follows the guidelines of the Nottingham grading system [10].

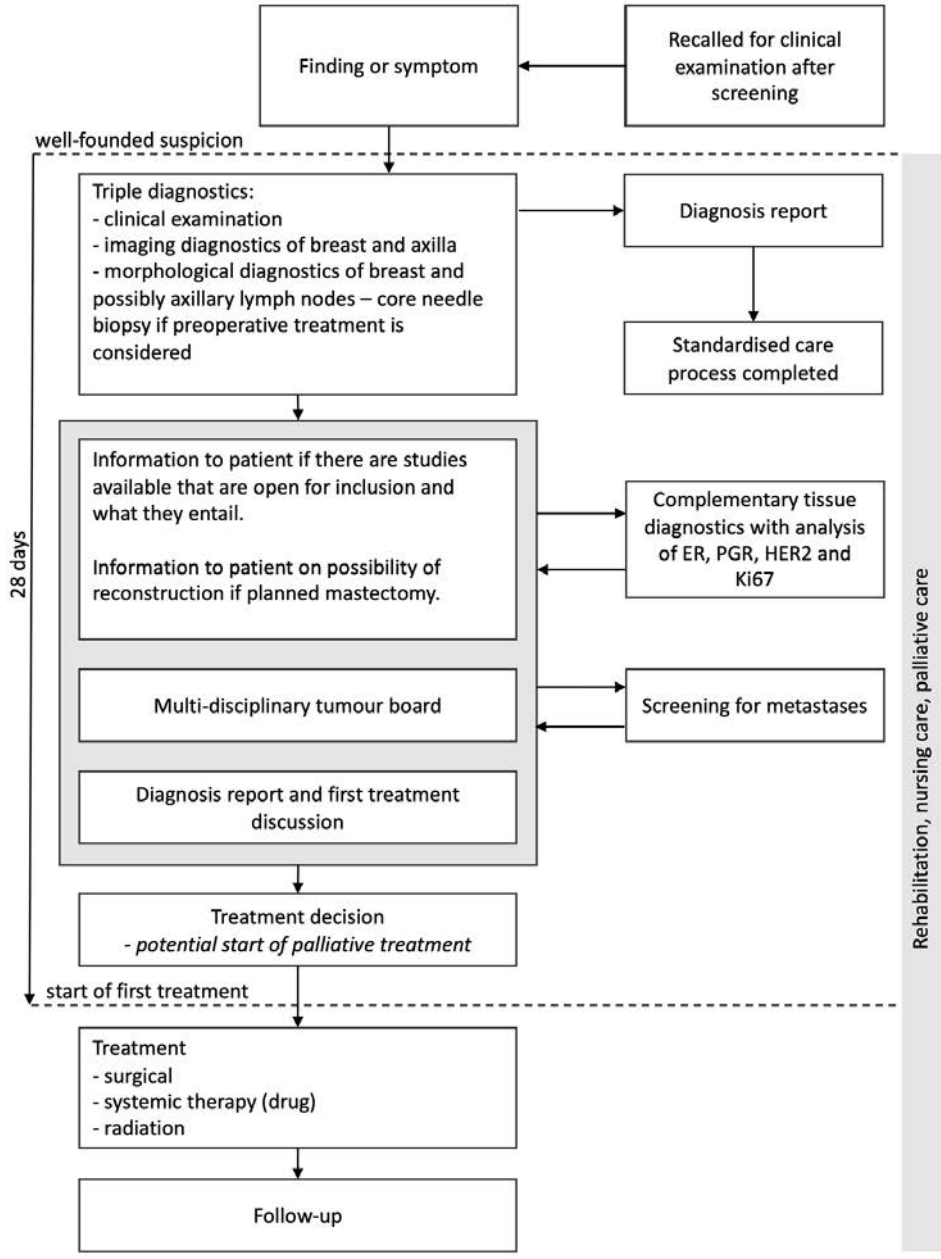

Figure 1. National Swedish standardised care workflow for breast cancer, translated from [8].

2.2.3.1 Nottingham Histological Grade

Breast cancer grading can be performed on H&E-stained FFPE-fixated tissue samples from biopsy or surgery. Determining the Nottingham Histological Grade (NHG) consists of the assessment of three subcomponents: mitotic count, nuclear pleomorphism and tubular formation. Each of these subcomponents is scored from 1 to 3. The mitotic count measures the number of mitoses within 10 high power fields. These high-power fields should be selected from the regions with the highest density of mitoses within the tumour. Cut-offs on the number of mitoses depend on the field area [10]. The selection of these power fields can lead to inter-assessor variability. Scoring nuclear pleomorphism is based on an assessment of nuclear atypia, which refers to differences between cancer and normal cells. This is based on the size, shape, vesicularity and presence of nucleoli with normal epithelium and should be performed in less differentiated tumour regions. Small, regular uniform cells are assigned a score of 1, a moderate increase in size and variability a score of 2 and marked variations a score of 3 [10]. Tubular formation assesses the proportion of cancer cells that follow tube-shaped structures with clear central lumina. If more than 75% of cells are arranged in tubule, a score of 1 is assigned, a score of 2 is indicated if 10-75% of cells display tubule formation and a score of 3 if less than 10% of cells follow these structures [10]. The final NHG is then based on a cut-off applied to the sum of the subgrades and also ranges from 1 to 3. Sums of the subcomponents of 3 to 5 are assigned NHG 1, 6 or 7 NHG 2 and 8 or 9 NHG 3. Higher NHGs are associated with worse prognosis. While the NHG system is well established, the proportions of grades can vary substantially between pathology departments in Sweden, indicating clinically relevant inter-laboratory variability that affects treatment decisions [11]. Based on a study of the Swedish National Breast Cancer Quality Registry (NKBC), out of 38,076 invasive breast cancer cases studied, 21.4% were assessed as NHG 1, 51.4% as NHG 2 and 27.2% as NHG 3, with information missing in 1.3% of cases [11].

2.2.3.2 Biomarker Assessment

Biomarkers in breast cancer are assessed mainly through IHC staining, either of biopsy or resection specimen. While it is possible to perform the assessment on samples collected with core needle biopsies, it is recommended to repeat the assessment with surgical specimen after resection due to intra-tumour heterogeneity [5], [12]. There are four biomarkers that are routinely assessed in Sweden. These are the estrogen receptor (ER), the progesterone receptor (PGR),

the human epidermal growth factor receptor 2 (HER2) and Ki67, which is a marker of proliferation.

There are two types of estrogen receptors, alpha (ERα) and beta (ERβ), which are both nuclear receptors that are activated by the sex hormone estrogen. ERα stimulates cell proliferation in breast tissue [13]. It is often present in early-stage breast cancer. ER receptor status evaluation is done through IHC staining. If at least 10% of cancer cells within a tumour are positively stained, the cancer is considered ER-positive, which is the case in 86.7-89.2% of breast cancer cases based on the NKBC [11]. Tumours with 1-10% positively stained cells are considered low-positive. ER-positivity is predictive for benefit from endocrine therapy [14].

The progesterone receptor PGR stimulates proliferation through modulation of ER activity [15]. Similar to ER, it is assessed with IHC and the same cut-off at 10% as with ER applies. It is a prognostic marker in ER-positive breast cancers [14]. In the investigation in [11], 70.9-74.8% of breast cancer cases in the NKBC were PGR positive.

The human epidermal growth factor receptor 2, HER2, is a receptor on breast cells whose overexpression leads to uncontrolled cell growth and division. HER2 status is primarily assessed with IHC, yielding a score from 0 to 3+. The scores 0 to 1+ are considered HER2-negative. A score of 3+ is considered HER2 positive, whereas 2+ is considered borderline. In this case, the HER2 status should be confirmed with silver in-situ hybridisation (SISH) or fluorescence in-situ hybridisation (FISH). HER2-positivity is associated with a worse prognosis if untreated and an indicator for HER2-targeted therapies such as Trastuzumab [16]. Based on data from the NKBC, 12.4-13.8% of breast cancer cases were assessed as HER2 positive [11].

KI67 is a nuclear protein that is associated with cell proliferation and RNA transcription. It is expressed in active phases of the cell cycle, but absent during quiescence. It is quantified via IHC staining and reported as the percentage of positively stained cells, either in hotspot regions or the entire invasive cancer region, depending on local guidelines. In Sweden, the recommendation recently changed from hotspot-based scoring to whole-tumour scoring [17]. Currently, the clinical utility of Ki67 is limited due to limitations in its analytical validity, which requires robust standards both for sample preparation and scoring [18], [19]. The most recent Swedish guidelines consider a cancer to have low KI67 expression if 5% or less of cells are positively stained and as highly expressed if 30% or more are positively stained, with a corresponding intermediate range of positively

stained cells that requires further diagnostics for risk stratification [8]. The median KI67 proliferation index is 22% in the NKBC [11]. Based on the statuses of these biomarkers, it is possible to assign cases to subtypes, as shown in Figure 2.

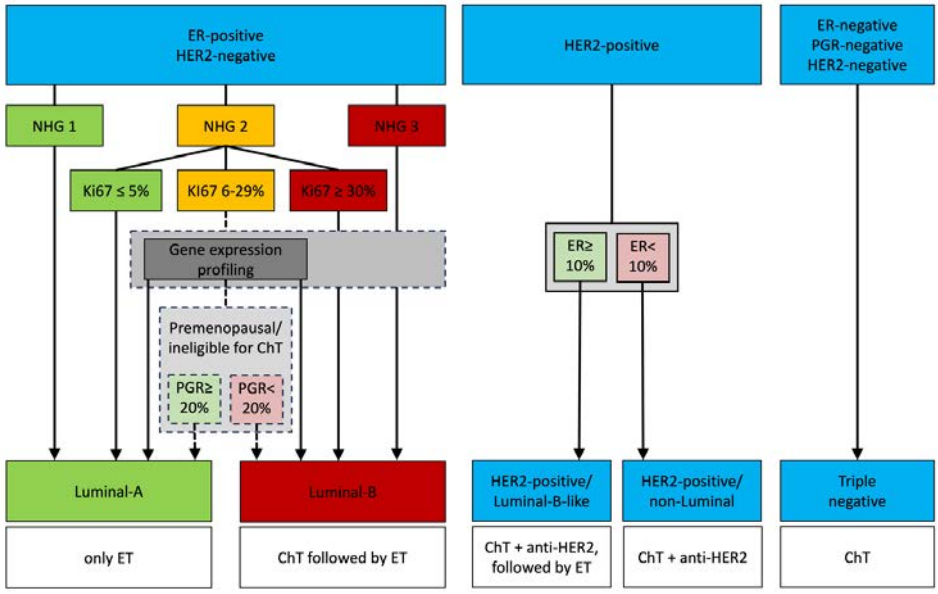

Figure 2. Overview of breast cancer subtypes, adapted and translated from the Swedish national care guidelines for breast cancer [20]. These guidelines also indicate that gene expression profiling should be performed for postmenopausal women with NHG2 ER-positive and HER2-negative cancer if there is uncertainty regarding the risk classification of the cancer. Furthermore, gene expression profiling can be considered if only IHC indicates a classification as Luminal-B rather than Luminal-A and therefore the need for chemotherapy. White boxes indicate ESMO treatment recommendations [5].

2.2.4 Molecular Profiling

There are several prognostic or predictive diagnostic tests that use gene expression profiling to further stratify breast cancer patients. These include MammaPrint by Agendia, the Oncotype DX Recurrence Score by Genomic Health, the Prosigna PAM50 by NanoString Technologies, the Breast Cancer Index by Biotheranostics and Endopredict by Myriad Genetics. Except for MammaPrint, all these tests are intended for ER-positive early disease [5]. The Prosigna test uses the expression of 50 genes, the PAM50 gene panel, to determine the molecular subtype of a cancer sample, as well as a risk-of-recurrence score. These subtypes are Luminal-A, Luminal-B, HER2-enriched, basal-like and normal-like [21]. Current Swedish national care guidelines recommend this test for postmenopausal women whose histopathological assessment resulted in a NHG 2 classification with an intermediate KI67 score, ER-positive and HER2-negative IHC if there is uncertainty regarding their risk classification. Furthermore, it can be considered if

IHC is the only indication for chemotherapy [8]. As shown in Figure 2, it is also possible to approximate these subtypes based on the biomarker statuses. Another common gene expression-based test is the Oncotype DX by Exact Sciences. It computes a recurrence score based on the expression of 21 genes. It has been shown to predict the risk of distant recurrence in ER-positive node-negative patients treated with tamoxifen [22].

2.2.5 Treatment

There is a variety of treatment options for breast cancer. The selection of a specific treatment depends on the size and location of the primary tumour, the number of lesions, the number of lymph nodes involved, histopathological grade, biomarkers, gene expression if available, menopausal status, as well as the patient's health status and preferences. It is recommended to take age into consideration only in the context of these other factors and not as the determining one. For premenopausal patients, fertility preservation may also be taken into consideration [5]. There are local and systemic treatments. Local treatments include surgical resection of tumours, mastectomy, and radiotherapy. Systemic treatments include chemotherapy (ChT), endocrine therapy (ET) and HER2 targeted therapies (anti-Her2), as well as further recently developed targeted therapies [23]. The selection of the appropriate systemic treatment can be based on the breast cancer subtype, as shown in Figure 2, which indicates the European Society for Medical Oncology (ESMO) treatment recommendation for each subtype [5].

2.3 Prostate Cancer

Prostate cancer refers to invasive tumour cell growth in the prostate, which is a walnut-sized gland that is a part of the male reproductive system. The prostate produces the seminal fluid, which sustains and transports sperm. More than 95% of prostate cancers are adenocarcinomas, which are cancers that originate from epithelial cells of glandular tissue structures.

2.3.1 Epidemiology

Globally, there were 1.4 million new prostate cancer cases in 2020, and prostate cancer is the most common cancer among men in Europe and North America. Northern Europe is the region with the globally highest age-standardised prostate cancer incidence of 83.4 per 100,000 person-years [4]. The high incidence of prostate cancer in Northern Europe is partially attributable to the age structure of

the populations, as well as to rigorous screening guidelines that allow for the detection even of early-stage tumours [24]. In Sweden, there are approximately 11,000 prostate cancer cases and 2,400 deaths annually, which is the third highest age-standardised incidence in Europe [25]. The aetiology of prostate cancer remains an active area of research. Established risk factors include age, ethnicity (particularly western African ancestry) and a family history of the disease. Furthermore, some mutations such as BRCA1 and BRCA2, as well as conditions such as Lynch syndrome are associated with an increased risk of prostate cancer. Lifestyle risk factors may include smoking, excess body weight and nutritional factors [4].

2.3.2 Screening and Diagnosis

Currently, the Swedish National Board of Health and Welfare does not recommend population-based prostate cancer screening to avoid overdiagnosis and treatment. However, there are currently several ongoing studies that investigate the effect of inviting all men in specific regions and birth cohorts for screening, which is referred to as organised prostate cancer testing. Depending on the outcome of these studies, this may result in a future national screening programme. For men with at least two first-degree relatives with a history of prostate cancer, testing is already recommended from the age of 40 [26]. Prostate cancer diagnosis is either initiated based on a palpable nodule in a digital rectal exam (DRE) or elevated prostate specific antigen (PSA) levels. A PSA serum value of 3 µg/l for men below 70 years of age, 5 µg/l for men between the age of 70 and 80 or higher than 7 µg/l for men older than 80 years is considered elevated. If any of these criteria are met, current guidelines recommend magnetic resonance imaging (MRI) for most patients. Based on the findings of DRE, PSA, and MRI, a prostate biopsy might be recommended. Prostate cancer biopsies are either systemic core needle biopsies or guided through transrectal ultrasound or MRI [26]. There are several blood, urine and tissue tests that can be used in order to reduce unnecessary biopsies, e.g. the Stockholm3 test is currently investigated in regional projects [27], [28].

2.3.3 Histopathological Assessment

Histopathological assessment of prostate biopsies primarily serves the purpose of establishing a cancer diagnosis and grading the cancer areas to guide treatment decisions. Furthermore, the extent of the cancer might also be

estimated based on the biopsies, however, there is no consensus for a standardised approach.

2.3.3.1 Gleason Grading

The Gleason grading system categorises areas of prostate cancer cells in prostate biopsies into five grades [29]. These grades progress in severity from grade 1 with well differentiated cells to grade 4 and 5, which are considered poorly differentiated or anaplastic, as shown in Figure 3. Tumours with patterns of grade 4 and 5 are associated with a higher risk of cancer death compared to tumours with a pattern of grade 3 [30]. Typically, the most prevalent and second most prevalent or highest-grade patterns are reported and can be summed to obtain the Gleason sum. The International Society of Urological Pathology (ISUP) only considers Gleason grade patterns 3-5 as cancerous, whereas grades 1 and 2 are considered benign [31]. Assigned Gleason grades can diverge even among expert pathologists [32]. This inter-assessor variability can lead to under- or overtreatment. To further standardise prostate cancer grading, the ISUP also suggested the ISUP grading system, which is based on a combined grade of primary and secondary Gleason grades [33].

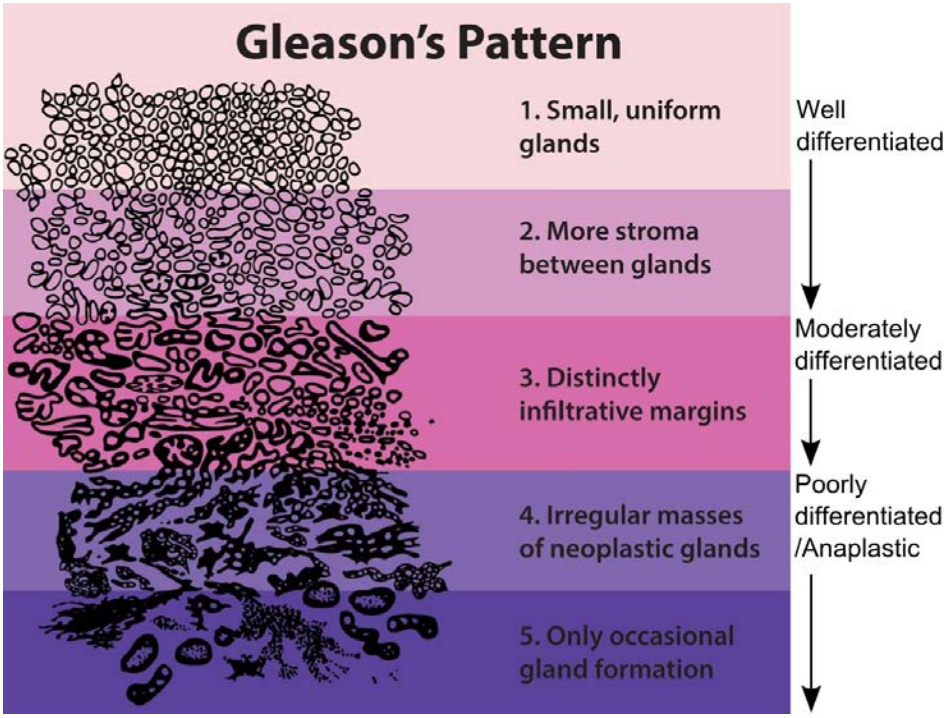

Figure 3. Schematic of tissue patterns that exemplify Gleason grades 1 to 5. Grade 1 and 2 are not considered cancerous, as opposed to grade 3 to 5. A higher grade is associated with worse prognosis.

2.3.4 Molecular Profiling

To guide treatment decisions and to reduce under- and overtreatment, multiple gene expression-based prostate cancer assays have been proposed. These assays quantify the expression of cancer-associated genes to risk-stratify patients. The cell-cycle progression (CCP) score developed by Polaris is the mean mRNA expression of 31 genes based on samples either from biopsies or prostatectomies. The CCP is associated with disease aggressiveness, the 10-year risk of metastasis after therapy, the risk of recurrence after prostatectomy and the disease-specific mortality under conservative management [34]–[36]. Other gene expression-based tests for prostate cancer include the Decipher Biopsy and Decipher postoperative scores [37]–[39] and the Oncotype DX genomic prostate score [40]–[44]. However, these tests remain costly and are not standard of care in Sweden.

2.3.5 Treatment

Depending on the risk assessment of the cancer and the patient's preference, there are several treatment options. These can be divided into conservative treatments, curative treatments, and non-curative life-prolonging treatments. Conservative treatments include active surveillance and watchful waiting. Active surveillance refers to an expectant management with frequent diagnostic investigations. Watchful waiting mainly relies on DREs and PSA testing [26]. Curative treatments include radical prostatectomy, which refers to the surgical removal of the prostate, and radiotherapy. Non curative treatments include hormone therapy, chemotherapy, and radiotherapy [26].

2.4 Digital and Computational Pathology

In recent years, there has been an increasing digitisation of pathology departments. Digitisation in this context refers to scanning slides with tissue samples and their subsequent analysis on screens, as well as their storage in digital image archives. The first commercial slide scanner was designed in 1994. It was commonly referred to as BLISS (Bacus Laboratories Inc., Slide Scanner) [45]. With this scanner, it took approximately 24h to scan a single slide. The first software to view pathology images was initially based on a software for satellite image retrieval and processing, called Active Data Repository (ADR) [46]. It was first used for virtual microscopy in 1996 and used ADR for spatial data retrieval at varying magnifications, which was quickly adapted to support pre-computed image pyramids, which are still a key component for WSIs today [45]. Over the last

two decades, there have been dramatic improvements regarding scanning speed and image quality, as well as in image storage and management systems. Today, several different companies produce WSI scanners that rely either on tile- or line-based scanning. These scanners typically generate WSIs in proprietary formats [45]. Open source software tools such as OpenSlide [47] can convert between some of these formats and the open TIFF format. There is also an effort to standardise the formats of WSIs that are based on tiles in image pyramids by the Digital Imaging and Communications in Medicine (DICOM) standards committee [48]. Today, the most essential components of WSI scanners typically include digital cameras, some of which are connected to a microscope with one or several objective lenses, robotics to facilitate movement of samples, and computers for rudimentary image processing, such as region-of-interest (ROI) detection and focusing [45]. However, broad adoption of digital pathology was initially hindered by several factors. Besides high initial costs of WSI scanners, these included concerns regarding the accuracy of diagnoses based on WSIs and regulatory approval of scanners. By now, there is a multitude of studies that establish a high concordance between diagnoses based on the inspection of tissue samples with a microscope and WSIs. An overview of these studies is available in [49]. The first WSI scanner that was granted FDA approval was the IntelliSite Pathology Solution by Philips. This decision was based on a non-inferiority clinical trial [50]. Several other WSI scanners have received regulatory approval since.

2.4.1 Computational Pathology

The generation of WSIs in routine clinical workflows, or research environments, allows for the application of automated or semi-automated computer-based image analysis tools, which is referred to as computational pathology. Particularly the thousands of WSIs and linked clinical, genetic and outcome data that were published by The Cancer Genome Atlas (TCGA) research consortium [51] were instrumental for the development of computational pathology tools during recent years. Computational pathology tools are typically either based on classical image analysis techniques such as feature extraction and their subsequent analysis by machine learning models, or deep learning models that directly learn features from data. Classical feature extraction in this context is based on expert-designed features such as measures of area, size, shape, texture, colour, as well as spatial relationships and distributions of structures of interest. These structures include micro-anatomic objects such as cell nuclei and morphological structures such as glands or tissue areas. However, this approach is limited to human-conceivable

features. Furthermore, it is often more vulnerable to differences in sample preparation such as cutting, fixation and staining than deep learning models. Therefore, deep learning has replaced these classical image analysis techniques in almost all applications of computational pathology [45]. The objectives of computational pathology tools range from the segmentation or classification of micro-anatomic objects, morphologies, and tissue to the generation of WSI-level classifications or regressions. These objectives can be broadly assigned to two categories. The first aims to automate routine diagnostic workflows, such as cancer detection, typing, grading in H&E WSIs and biomarker scoring in WSIs of immunohistochemically stained tissue. This has the potential to decrease inter-assessor variability and to reduce costs. The second category of tools aims to generate information that pathologists cannot obtain through visual assessment, such as the prediction of genetic alterations, outcomes, or treatment responses. Particularly this second category has the potential to contribute to the advancement and broad availability of precision medicine through image-based biomarkers, which can be cheaper and faster to generate than measurements of molecular biomarkers using e.g. DNA- or RNA-sequencing or other profiling methodologies.

2.4.2 Whole-Slide-Image Registration

WSI-registration is an active field of research within computational pathology that aims to align corresponding tissue regions between multiple WSIs from the same tissue specimen, as shown in Figure 4. Since tissue sections are mounted onto glass slides, the position and rotation of tissue even of consecutive sections on these glass slides and resulting WSIs will vary. Furthermore, the thickness of the sections only measures a few micrometres. Therefore, they are susceptible to deformations and tears. However, the alignment of these tissue regions allows combined analyses of information from H&E and IHC, which has applications both in research and diagnostics. In the research setting, this may be useful for stain-guided learning, virtual staining, the analysis of multiplex stained histology, 3D reconstruction and for the transfer of annotations or predictions between WSIs. In the clinical context, this may be useful in order to identify e.g. invasive cancer during biomarker scoring or for the investigation of suspicious lesions at resection margins. WSI registration is a particularly challenging registration task due to the gigapixel scale of WSIs, differences in the appearance of different stains, changes in structure and morphology resulting from sample preparation, which can also introduce artefacts, tears, and deformations. WSI registration algorithms typically

rely on extracting and matching features between WSIs or on optimising an intensity-based metric that quantifies the similarity between tissue regions [52]. Pre-processing steps often include tissue segmentation and histogram or colour matching between the WSIs of an image pair. In recent years, the application of deep learning both in the pre-processing and registration steps has become increasingly common.

Figure 4. Example of an H&E WSI registered to an IHC WSI.

2.5 Machine Learning

Machine learning, often also referred to as Artificial Intelligence (AI), differs from classical optimization tasks in its objective. Optimization aims at solving a specific task given data optimally. Machine learning on the other hand aims at a transfer, in which the results from optimising a model on a training set are intended to approximate an optimal parametrization for unknown test data. The purpose of supervised machine learning is to find an approximation of a function that maps input data to outputs such that the error in these outputs as quantified by some loss function or metric is minimised.

2.5.1 Artificial Neural Networks and Deep Learning

Artificial neural networks are a type of machine learning model. The most basic form of an artificial neural network is a fully connected network (FCNN), which is

also referred to as multi-layer perceptron (MLP). Artificial neural networks consist of artificial neurons, also referred to as perceptrons [53], as depicted in Figure 5.

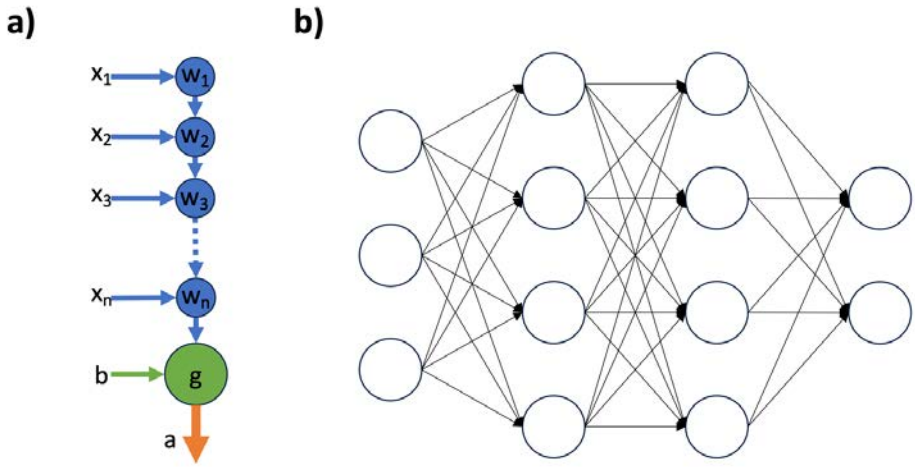

Figure 5. Schematic of a perceptron in a) and an artificial neural network in b).

MLPs are parametrized through their weights w and biases b. Typically, each neuron in a layer has an input x from all neurons in the previous layer and its activation is relayed as an input to all the neurons in the consecutive layer. The connections between these neurons are weighted with the weights w. Furthermore, the sum of all weighted inputs into a neuron is offset by a learnable bias b. After addition of this bias, an activation function g is applied, which is often either a sigmoid function or the rectified linear unit (ReLU) function. The output of this function is the activation a, which is the input to the neurons of the subsequent layer. The number of layers and neurons in a network is typically optimised as a hyperparameter. Deep learning (DL) refers to a technique where a large number of consecutive model layers is used to approximate complex functions through increasing complexity of the representations of these layers. In recent years, machine learning and particularly deep learning has made significant contributions to almost all areas of research and technology. This was not as much driven by novel theories or models, but rather through increased availability of large-scale data sets and improving computing infrastructure at decreasing costs. Even convolutional neural networks (CNNs), which are now one of the most common types of neural networks for the analysis of data with an evenly spaced grid topology such as images, have been conceptualised much earlier. The first neural network that was referred to as a CNN was proposed in 1998 by LeCun et

al. in [54]. A similar model, the neocognitron, had already been proposed by Fukushima et al. in 1980 in [55]. However, the publication by LeCun et al. in 1998 met more favourable conditions due to improvements in computing infrastructure and more importantly, an immediate application with the automated identification of handwritten digits in post codes and bank checks. Back-propagation, the algorithm now ubiquitously used to optimise neural networks, was already proposed in 1986 by Rumelhart et al. [56].

2.5.2 Optimisation of Deep Neural Networks

Loss functions are used in supervised machine learning to quantify the error between a prediction $\hat{y}$ and the true label y of a sample. Considering that the prediction $\hat{y}$ is a function $f(x, \theta)$ of the model inputs x and the model parameters θ, the loss $L(\hat{y}, y)$ can be formulated as $L(x, \theta, y)$, which becomes $L(\theta)$ for a specific combination of input x and label y. The model parameters θ are then updated based on this loss value. One or multiple loss functions can be applied. If multiple loss functions are used, the total loss is the weighted sum of the individual loss functions. These weights can be optimised as hyperparameters. It is common to add losses that do not directly aid the optimisation of the error between predictions and true labels but that are intended to aid generalisation to unseen data. The purpose of these losses is therefore to prevent overfitting the training data, which is referred to as regularisation. Figure 6 depicts an example of a decision boundary with and without regularisation.

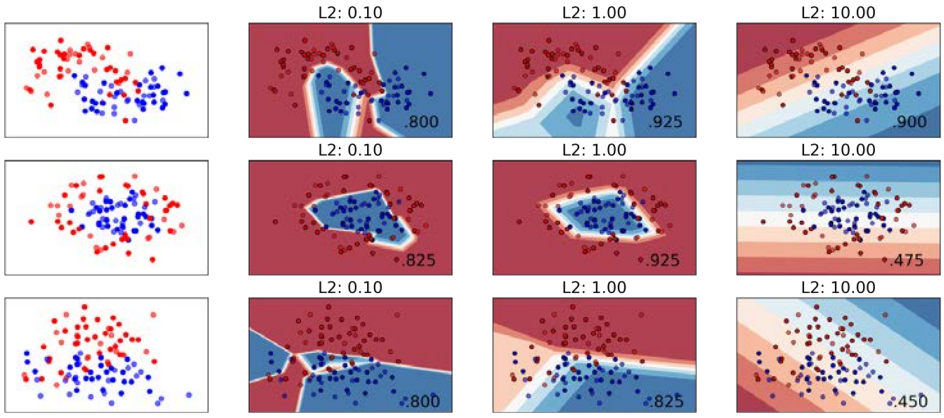

Figure 6. Decision boundary for different values of L2 regularisation for an artificial neural network. Blue and red dots indicate samples from two different classes. The lower right number in each plot indicates the accuracy for the respective L2 regularisation.

As can be seen, the decision boundary is smoother when regularisation is applied, which does not aid the prediction accuracy on the training data, but might be a

better decision rule on unseen test data. Common regularisations are L_1 and L_2 regularisation, based on the respective norm of the model parameters. Once the loss for a sample or set of samples has been computed, model parameters are updated based on the loss. In this context, this is typically phrased as the minimisation problem

$$\theta^* = \arg\min L(\theta)$$

for a fixed input x, where θ^* refers to the optimal model parameters. Since the number of parameters of neural networks is often in the millions, this optimization problem is not solved analytically but numerically with gradient descent. Gradient descent is an iterative method to solve an optimization problem such that it converges at least to a local minimum. Its application requires the loss function to be locally differentiable with regards to the current model parameters θ. Intuitively, gradient descent can be understood to modify the model parameters iteratively against the sign of the derivative of the loss. The parameter update can then be described as

$$\theta' = \theta - \epsilon \nabla_\theta L(\theta)$$

where θ' are the updated parameters and ϵ the step size, which is referred to as the learning rate in the context of machine learning. Particularly in deep learning, which is characterised by neural networks with many layers, it can be complex to compute the derivatives of the loss function for each parameter. The back-propagation algorithm [56] can be used to efficiently compute local derivatives through a highly efficient order of computations. Local gradients are computed solely based on locally available information with the simplifying assumption that only the individual weight to be updated changes while all other weights are held constant. Depending on the size and type of training data, it is often not possible to compute the gradient for all samples in the data set simultaneously due to computing hardware limitations. In this case, the gradient is estimated based on a randomly sampled subset of the training data set, referred to as a batch. This method is then called stochastic gradient descent. A common problem of training deep models with gradient descent and backpropagation is that gradients may converge to zero (vanish) or diverge (explode). Besides the depth of deep learning models, this might be because the assumption of backpropagation that all parameters are constant during parameter updates except for the updated parameter is not met. A technique that alleviates this is batch normalisation, which was first proposed by Ioffe et al. in [57]. In batch normalisation, the activations and

gradients are normalised to zero mean and unit variance, where the mean and standard deviation used for normalisation are learnable parameters that are updated based on the standard deviation and mean of batches that are passed through the network. This mean and standard deviation converge towards the mean and standard deviation of activations of the entire training data at the respective layer after passing a sufficient number of batches through the network. During CNN optimization, the average loss value for a pre-defined number of samples, referred to as an epoch, is monitored both for the training data and a tuning or validation data set. Model optimization is terminated when the loss on this validation data does not further improve for a specified number of epochs to avoid overfitting.

2.5.3 Convolutional Neural Networks

CNNs are a type of neural network that perform convolutions with learned filters on the input data in at least some of their layers. While there are graph-CNNs that can work with more complex input data, CNNs are typically applied to data that is structured in an evenly spaced grid, such as time series data, image data or videos. Convolutions can be expanded to an arbitrary number of dimensions, but the focus here will be on image data. RGB images are 3D matrices where one dimension corresponds to the width, one to the height, and one to the three colour channels red, green, and blue. The continuous one-dimensional convolution of two functions $f(t), g(t)$ is denoted as $f(t) * g(t)$ and defined through

$$f(t) * g(t) = \int_{\tau=-\infty}^{\infty} f(\tau)\, g(t-\tau)\, \mathrm{d}\tau.$$

Real signals are finite, which allows adjusting the bounds of the integral accordingly. On computers, this operation is not performed continuously but numerically, yielding

$$f(t) * g(t) = \sum_{\tau=-\infty}^{\infty} f(\tau)\, g(t-\tau).$$

Intuitively, this can be understood as a multiplication of the flipped filter with the data points that it overlaps with, shifting the filter along the axes and summing the products for a given location. Since the entire filter needs to overlap with data points to obtain a valid result, each convolutional layer shrinks the input to the consecutive layers, unless padding is applied. Padding appends zeros to the output of the convolution to preserve its size. Layers of a CNN have many filters

that are applied in parallel, which can be done efficiently with graphical processing units (GPUs). Each filter extracts a specific feature from the outputs of the previous layer, resulting in a corresponding feature map for each filter that is passed to the next layer. This sequence of a large number of filters allows the detection of complex structures.

The success of CNNs in image analysis is in part attributable to weight sharing. Weight sharing refers to the application of the same filter to all parts of an input or a feature map. This also results in a degree of translational invariance of CNNs, which means that a translational shift in the input of a layer results in a corresponding shift in the activations of that layer. Weight sharing of relatively small filters drastically reduces the number of parameters of a CNN and can therefore alleviate overfitting compared to FCNNs. This also reduces the number of computations drastically. Nevertheless, CNNs typically have millions of parameters, which require large image data sets to optimise.

To reduce the amount of required training data, it is common to apply data augmentation. Data augmentations are transformations of the input data that do not alter the relationship between image and label. Common augmentations for image data are rotations, mirroring and slight shifts in colour, contrast, saturation, and brightness. Furthermore, CNNs can be pre-trained on large publicly available data sets such as ImageNet [58] to start the optimization at a parameterization that is better than random. This is referred to as transfer learning. These pre-trained models can then be fine-tuned with data from the specific application domain. The combination of all these methods allows the efficient optimisation of deep CNNs with limited data even on personal computers within hours.

3 Research Objectives

The objective of this doctoral thesis was the development, application, and evaluation of computational pathology methods with applications both in research and pathology diagnostics.

- **Study I**: The objective of this study was the development and evaluation of a computationally efficient approach to predict mRNA gene expression from WSIs of H&E-stained prostatectomy specimens. Specifically, we wanted to explore whether the co-expression of genes can be leveraged to cluster genes both to reduce the computational cost, as well as to improve prediction performance.

- **Study II**: The objective of this study was to evaluate whether attention-based multiple-instance-learning models have the potential to improve regression prediction objectives. In this case, we wanted to investigate whether the prediction of gene expression from WSIs of H&E-stained breast resection specimens can be improved with this technique, after it had been shown to lead to increased performance for classification tasks.

- **Study III**: The objective of this study was to assess the current state-of-the-art in multi-stain WSI registration. To this end, we conducted the ACROBAT WSI registration challenge.

- **Study IV**: The objective of this study was to publish and describe the data set that we published to facilitate the ACROBAT challenge, such that it might be used by the research community to further improve future registration algorithms or to increase its usefulness in other research contexts beyond registration.

- **Study V**: The objective of this study was to assess whether with current WSI registration methods, the usefulness of annotations that already existed for sections in one stain could be transferred to sections with other stains. We then aimed to evaluate whether cancer detection models trained with these registered annotations are inferior to those trained with annotations that were directly generated for the target stain.

4 Materials and methods

4.1 Image Processing

WSIs were processed for analysis in Study I, II and V. The pre-processing consists of several steps, starting with tissue detection, followed by tiling, stain normalisation, cancer detection, and in Study II feature extraction.

In Study I and II, tissue detection is performed with the level of the H&E WSI image pyramid that is closest to a downsampling factor of 32 compared to 40X. This image was then transformed to the HSV colour space, where an Otsu threshold [59] was applied to the saturation channel. The resulting binary mask was then compared to a binary mask with the criterion that the hue channel should have a value < 0.75 with a logical AND operation. We then applied morphological opening and closing to remove salt-and-pepper noise from the mask, yielding the final tissue mask. In Study V, tissue masks were generated from KI67 WSIs with the method described in [60], at a resolution of 3.64 µm/pixel and all other parameters as suggested by Bándi et al.

Tissue regions were then tiled. In Study I, tiles were generated at 40X, 20X and 10X magnification. For Study II and Study V, tiles were only generated at 20X magnification. Since WSIs originate from different scanners, tiling was performed at the most common microns-per-pixel in the SöS data set, which are 0.252, 0.504 and 0.904 microns-per-pixel at 40X, 20X and 10X respectively. If the respective MPP did not exist for a WSI, the next higher resolution was accessed, and tiles were downsampled using Lanczos interpolation. Only tiles containing more than 50% tissue based on the tissue masks were included. Tiles were generated with a size of 598 x 598 pixels for Study II and Study V. In Study I, STHLM3 WSIs were tiled with 598 x 598 pixels and TCGA WSIs with 500 x 500 pixels, as we used random cropping to 500 x 500 pixels for training the STHLM3 cancer detection model.

H&E tiles were then normalised with an adaptation of the method proposed by Macenko et al. in [61]. This normalisation normalises tiles to a predefined target stain vector, which was obtained by randomly sampling 3000 tiles from the target training data set. The stain vector for each WSI to be normalised was then obtained by randomly sampling 100 tiles from each WSI. Each tile was then normalised individually based on the respective WSI-level stain vector and the target stain vector.

Cancer detection was performed with an Inception CNN model that was trained with data from the Clinseq study. For each tile, we generated a binary prediction that indicates whether the tile contains cancer. Resulting tile predictions were then transformed into a cancer mask at a downsampling factor of 32 compared to 40X. We then performed morphological opening and closing to remove salt-and-pepper noise. Only tiles that originate from areas within the resulting cancer masks were included in the analysis for Study I and II.

4.2 Machine Learning

4.2.1 Data Splits and Hyperparameter Tuning

When optimising machine learning models, particularly models as complex as CNNs, it is common that not only the parameters of the model itself need to be optimised, but also parameters that define model design choices or the optimization process. These parameters are referred to as hyperparameters. Examples of hyperparameters are the learning rate, the number of layers and neurons in each layer, or even the CNN architecture. When hyperparameters are to be optimised, it is not sufficient to subdivide available data into development and test data. The development data needs to be further subdivided into data used for model parameter optimisation, and data that can be used to compare different hyperparameter combinations. Common techniques for this are cross-validation (CV) or nested CV. All hyperparameters and models optimised in the studies that comprise this thesis were optimised in either a CV or nested CV. Generated data splits were stratified by relevant clinical covariates such that the distribution of these were approximately the same in the different data partitions.

4.2.2 Inception Networks

The first Inception network [62], [63], also known as GoogleNet, was developed by Szegedy et al. in 2014, winning the ImageNet top-5 classification ranking that year. The CNN has approximately 4 million parameters in 22 learnable layers. Inception networks consist of Inception modules. The main idea of Inception modules is to apply convolutional filters of different sizes in parallel. This has the advantage to not constrain the network to a specific filter size and to allow for different weighting of different filter sizes throughout the network. To ensure matching feature map sizes for different filter sizes, the network uses zero padding. The convolutions with 1x1 filters are used to reduce the dimensionality of feature maps before convolving with large kernels. Inception CNNs consist of a stem, Inception modules and different classifier stages, which are intended to improve gradient

flow through the network during training. The lower classification outputs are not used during deployment. However, during training, the respective model outputs are subject to loss computation and backpropagation of this loss. Inception networks remain one of the most common CNN architectures.

4.2.3 Residual Networks

Another very common CNN architecture are residual networks or ResNets. A ResNet is the first network that might have outperformed humans on the ImageNet top-5 classification competition in 2015, with an error rate of 3.57%. The ResNet architecture was developed at Microsoft Research by He et al. in [64]. The main innovation of ResNets is skip connections in residual blocks. Like Inception networks, this can be considered as the combination of convolutions at different input scales. Furthermore, skip connections allow the training of considerably deeper networks, which might be due to an improved preservation of gradients during back-propagation through the skip connections. While the Inception network that won the ImageNet competition in 2014 had 22 trainable layers, the ResNet that won the competition in 2015 had 152 layers. However, ResNets were also the first CNNs in the ImageNet competition that used batch normalisation, and it is unclear how much residual blocks and this normalisation contributed to the improvements in performance.

4.2.4 Attention-based Multiple-Instance-Learning

Supervised machine learning models require a label for model optimization. In the context of computational pathology, it is common to divide WSIs into smaller image patches, referred to as tiles, to circumvent current computing hardware limitations. Labels are then assigned to these individual tiles. Some labels, such as areas of invasive cancer, can be trivially transferred to individual tiles based on their coordinates. However, for WSI-level or patient-level labels, such as an outcome or a treatment response, the contribution of each tile to the label is not defined since the label only exists on the WSI (or patient)-level. Often, the WSI-level label is still naively assigned to each tile, which can produce satisfactory prediction results [65], [66]. Another method that has been shown to be effective in this setting is multiple-instance-learning (MIL) [67]–[69]. Multiple instance learning refers to a setting where a label is only known for a set or bag of instances, without knowledge regarding which of these instances contribute to the bag-level label. Attention-based MIL has been proven effective in some contexts to solve tasks of this structure. Neural networks that use attention mechanisms typically

consist of two subnetworks. One of these subnetworks, the attention module, predicts how much each instance should be weighted. Another subnetwork, the prediction module, then generates a bag-level prediction based on the features and weight of each instance. The predicted weights can also be used as an indication of how much each instance contributes to the bag-level label. In the context of computational pathology, this can be used to identify relevant regions in WSIs.

4.3 Statistical Analysis and Performance Metrics

4.3.1 Spearman Correlation

The Spearman correlation can be used to quantify the association between two variables. It is a non-parametric measure that does not assume a linear relationship between variables. It is defined as the Pearson correlation of the ranks of the variables. The Spearman correlation between X, Y is in [-1, 1], where 0 indicates no correlation and higher absolute values a monotonic association. It can be computed with

$$r_s = \frac{\mathrm{cov}(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}},$$

where cov is the covariance, σ denotes the standard deviation and R the rank.

4.3.2 Sensitivity and Specificity

Sensitivity and specificity quantify the accuracy of a binary classification. The sensitivity is defined as the probability of a positive class prediction if the observation's true label is positive. The specificity is defined as the probability of a negative class prediction if the observation's true label is negative. The sensitivity is also referred to as recall or the true positive rate. The specificity is also referred to as the true negative rate [70].

4.3.3 Precision and Recall

Similarly to sensitivity and specificity, precision and recall can be used to quantify the accuracy of a binary classifier. Precision, also referred to as positive predictive value, is defined as the proportion of true positives out of all positive predictions. The definition of recall is equivalent to the definition of sensitivity.

4.3.4 Area under the Receiver Operating Characteristic Curve

The area under the receiver operating characteristic curve (AUROC, AUC) is a quantitative measure of the discriminative ability of a binary classifier. The receiver operating characteristic (ROC) curve is generated by plotting the sensitivity against 1 - specificity for varying classification thresholds. A value of 0.5 indicates that the outputs of the classifier are not related to the true class labels. In this case, all points of the curve are on a diagonal line with no offset. A value of 1 indicates a perfect classification [70].

4.3.5 Sørensen-Dice Coefficient and Jaccard Index

In the context of machine learning for image analysis, the Sørensen-Dice coefficient, or Dice coefficient, and the Jaccard index are often used to quantify the accuracy of semantic segmentations. The Dice coefficient can be understood as twice the intersection between two sets divided by the sum of the number of elements in each set. In the context of image analysis, this can be understood as twice the overlap between two areas, e.g. predicted and true area, divided by the sum of the two areas. It can be computed with

$$D = \frac{2TP}{2TP + FP + FN},$$

where TP indicates true positives, FP false positives and FN false negatives. The Jaccard index follows the same formula, but does not include multiplication with 2 in the numerator or denominator. For Boolean variables, the Dice coefficient is equivalent to the F1-score.

4.3.6 Time-to-Event Analysis

Time-to-event analysis refers to the analysis of the time interval starting e.g. at diagnosis or treatment until an event of interest, e.g. recurrence of cancer or death, occurs. It is common that the event has not or not yet occurred for all observations at the end of their follow-up period. This is referred to as right-censoring, which is common in the context of medical studies. To obtain correct inferences, these observations need to be included in the analysis. One of the most common models for time-to-event analysis that are suited for partially censored data is the Cox proportional hazards (CPH) model [71]. CPH models quantify the association between one or more covariates and the event. CPH models are semi-parametric, since they do not require choosing a distribution of survival times or hazards. In CPH models, the hazard h is defined through

$$h(t) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n),$$

where h_0 is the baseline hazard, β the model coefficients and x denotes the n exposure variables. A common application of CPH models is to compute the hazards ratio between two groups. Often, one of these groups is e.g. treated or has a distinct prognostic marker, indicated through $x_i = 1$, compared to the reference or baseline group, indicated through $x_i = 0$. The hazards ratio then represents the change in risk of occurrence of the event. The hazards ratio HR between these two groups can be computed with

$$HR = \frac{h_0(t) \exp(\beta_i)}{h_0(t)} = \exp(\beta_i).$$

A limitation of CPH models is the proportional hazards assumption, which assumes that the effect of covariates is constant over time. Furthermore, the relationship between covariates is assumed to be linear and additive.

4.3.7 Linear Mixed Effects Models

Linear Mixed Effects (LME) models can be used to estimate effects when there are observations that originate from the same statistical units and are therefore not independent [72], [73]. LME models can be represented as

$$y = X\beta + Zu + \varepsilon,$$

where y represents the response variable, β the fixed effects coefficients and u the random effects coefficients. X and Z are matrices with the values of observations. ε captures the residuals. Fixed effects are assumed to be independent, whereas random effects originate from statistical units, such as the same tissue specimen, patient, or the same clinic. The assumptions of LME models are met if the explanatory variables have a linear relationship with the response and the errors have constant variance, are normally distributed and independent.

4.3.8 Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is a non-parametric hypothesis test [74]. It can be used to test whether matched samples originate from the same distribution through testing the null hypothesis whether the distribution of differences is symmetric around zero. The paired test can be performed by computing the differences X between pairs of observations. These differences are then ranked, such that the smallest value has a rank R of one and the highest rank corresponds to the number of observations N. The test statistic T can be computed with

$$T = \sum_{i=1}^{N} \text{sign}(X_i) R_i.$$

The p-value can then be looked up based on the test statistic.

4.3.9 Controlling the False Discovery Rate with Benjamini-Hochberg's Method

Statistical hypothesis tests aim to evaluate null hypotheses. The p-value corresponds to the probability that a value as extreme as the observed one arose by chance. This null hypothesis is then rejected if the p-value is below a pre-specified threshold, which is often chosen to be 0.05 or 0.01. There are two possible types of error in hypothesis testing, type I and type II errors. Type I error refers to the rejection of a true null hypothesis, and therefore a false discovery, whereas a type II error refers to a failure to reject a false null hypothesis, which is a missed discovery. When conducting a large number of hypothesis tests simultaneously, it is essential to adjust p-values such that the rate of false discoveries is controlled. The false discovery rate (FDR) is the proportion of false positives in the set of false and true positives. The FDR can be controlled with the method described by Benjamini and Hochberg [75] at a level α. With this method, given m p-values sorted in ascending order, the null hypothesis up to the k-th element can be rejected for which

$$p(k) \leq \alpha \frac{k}{m}.$$

4.4 Data Sets

The WSIs included in the studies that comprise this thesis originate from six studies, STHLM3, TCGA-PRAD, TCGA-BRCA, Clinseq, SCAN-B, and SöS. All included WSIs were digitised at 40X magnification, which corresponds to approximately 0.25 µm/pixel.

4.4.1 STHLM3

The STHLM3 study is a prospective population based diagnostic trial that was conducted between May 2012 and December 2014. Prostate biopsies were conducted on patients with a PSA ≥3ng/mL or a PSA ≥1ng/mL and a S3M probability of high-grade prostate cancer >10%. A subset of the generated biopsies originating from 1,136 patients, selected with stratified sampling on ISUP grade, was then digitised with a Hamamatsu NanoZoomer XR WSI scanner. Besides

ISUP grade, detailed cancer annotations are available for all WSIs. WSIs from this study were included in Study I.

4.4.2 TCGA-PRAD

The TCGA-PRAD study was conducted by The Cancer Genome Atlas Research Network [76]. The study includes 403 prostate cancer patients who underwent radical prostatectomy. WSIs in this study contain these prostatectomy specimens and were generated with Aperio WSI scanners. WSIs, genomic data, patient demographics, clinical characteristics, and outcome data are available from the GDC data portal (https://portal.gdc.cancer.gov/). Data from TCGA-PRAD was included in Study I.

4.4.3 TCGA-BRCA

The TCGA-BRCA study was conducted by The Cancer Genome Atlas Research Network [77]. The study is based on breast cancer resection specimens from 1,098 primary breast cancer patients. The TCGA WSIs, genomic data, as well as patient demographics, clinical characteristics and outcome data are available from the GDC data portal (https://portal.gdc.cancer.gov/). Patients whose slides were scanned at 20X resolution were excluded. All WSIs in TCGA-BRCA were digitised with Aperio WSI scanners. Data from this study was included in Study II.

4.4.4 Clinseq

The Clinseq (Clinical Sequencing of Cancer in Sweden) study consists of 307 female breast cancer patients from the Libro-1 and KARMA studies [78], [79]. The Libro-1 study retrospectively included breast cancer patients younger than 80 years who underwent surgery at Karolinska Universitetssjukhuset between 2001 and 2008. KARMA patients were enrolled prospectively from Stockholm's Södersjukhuset in 2012. Clinical characteristics and outcome information was obtained from the Stockholm-Gotland Regional Breast Cancer quality register, which contains historical data up to 2007, and the Information Network for Cancer Care (INCA), which contains data of breast cancer patients diagnosed between 2007 and 2018. WSIs of surgical resection specimens were digitised with Hamamatsu NanoZoomer XR and S360 WSI scanners. The Clinseq data set was used in Study II.

4.4.5 SCAN-B/ABiM

The SCAN-B (Sweden Cancerome Analysis Network - Breast) study is a multi-center study comprising seven hospital sites in South Sweden. We had access to

a subset of the SCAN-B study materials, consisting of 1,262 prospectively enrolled breast cancer patients that were diagnosed in Lund from 2010 to 2019 [80]. Patient information, including tumour characteristics such as NHG and biomarker statuses, as well as treatment and outcome information were obtained from INCA. Slides with resection specimens were digitised with a Hamamatsu NanoZoomer XR WSI scanner. A subset of patients from the SCAN-B cohort has RNA-seq data available. This subset is referred to as the ABiM cohort and was used in Study II.

4.4.6 SöS

The SöS cohort consists of 2,421 patients whose slides were retrieved from hospital archives and who were retrospectively enrolled. Both WSIs of biopsy and resection specimen are available, however, only WSIs of resection specimen were used here. Included patients were diagnosed with breast cancer at Stockholm's Södersjukhuset between April 2012 and May 2018. Associated clinical information was obtained from the Swedish national quality registry for breast cancer (NKBC) [81]. NKBC includes information on patient demographics, tumour characteristics such as NHG and biomarker statuses, as well as on treatments and outcomes. All slides were digitised with Hamamatsu NanoZoomer XR or S360 WSI scanners. The resulting WSIs are included in Study III-IV.

5 Ethics

All materials that were used in the studies that comprise this thesis were retrieved from archives, registries, or data bases. No interventions were performed. WSIs and clinical information were stored on servers with restricted access at the Department for Medical Epidemiology and Biostatistics (MEB) at Karolinska Institutet. These servers are only accessible from the internal network at MEB. Data that were published to conduct the ACROBAT challenge were fully anonymised before publication. Data management is therefore compliant with the General Data Protection Regulation (GDPR) and the Swedish Data Protection Act. The following ethical permits apply:

- **Study I**: The TCGA PRAD data are fully anonymized and publicly available through the US-American NIH National Cancer Institute GDC data portal and therefore require no ethical permits. For STHLM3, the ethical permits DNR 2012/572-31/1, DNR 2012/438-31/3, DNR 2013/981-32, DNR 2018/845-32 apply.

- **Study II**: The TCGA BRCA data are fully anonymized and publicly available through the US-American NIH National Cancer Institute GDC data portal and therefore require no ethical permits. For Clinseq and SöS, the ethical permits DNR 2017/2106-31 with amendments DNR 2018/1462-32, DNR 2019-02336 apply. For SCAN-B, the ethical permits DNR 2009/658, DNR 2009/659 with amendment DNR 2015/277 apply.

- **Study III-V**: For SöS, the ethical permits DNR 2017/2106-31 with amendments DNR 2018/1462-32, DNR 2019-02336 apply.

6 Results

6.1 Study I

In Study I, I developed and evaluated a computationally efficient approach to predict tumour average gene expression from WSIs of H&E-stained prostatectomy specimens. The underlying assumption is that co-expressed genes are associated with similar morphological changes, which might improve prediction performance. We predicted expression values for the whole transcriptome, selected genes for which the predictions were significantly associated with RNA-seq estimates in the development data and then validated those genes in a test set. Furthermore, we evaluated whether the prognostic cell-cycle progression (CCP) score can be predicted from WSIs. We used the Spearman correlations between predictions and RNA-seq estimates as the primary performance metric.

Study I is based on the TCGA PRAD study, which consists of 403 patients that underwent radical prostatectomy and that originate from 27 cancer centres. Out of these 403 patients, 370 were included in this study, based on the prostate cancer subtype, the availability of matching prostatectomy WSIs and RNA-seq data, prior systemic treatment or synchronous malignancies and a minimum detected contiguous tumour area of 1 mm^2. To identify regions of invasive cancer in the WSIs, we developed a cancer detection model based on prostate biopsies from the STHLM3 study. Only tiles that were predicted to contain invasive cancer were included in the gene expression prediction analysis. Out of the 370 included patients, we randomly selected 92 as a held-out test set.

In TCGA PRAD, there is expression data for 19,601 transcripts available. We included genes with at least three counts in at least 10% of patients, resulting in 15,586 selected genes. We then proceeded to cluster these genes into 50 clusters based on their co-expression, which we quantified through the average absolute Spearman correlation of transcripts in the development set. Out of these 50 clusters, we randomly selected 10 clusters with 2,636 genes for model optimization and selection.

Models were optimised in a nested CV and compared based on the predictions for the outer CV validation folds. The proposed modelling approach consists of CNNs that predict clusters of genes simultaneously. If a cluster consists of n genes, the respective CNN will have n outputs, one for each transcript. This model

is referred to as *corr clusters*. We compared this modelling approach to several baseline models. The first baseline model, referred to as *rnd clusters*, is based on multi-output CNNs that predict randomly grouped transcripts. The second baseline model, referred to as *lgbm*, consists of boosting models that were optimised to predict a single transcript per model based on ResNet18 extracted ImageNet features with the package LightGBM [82]. As a further baseline model referred to as *all gene*, we optimised a CNN to predict all 15,586 included transcripts and selected the 2,636 development gene predictions for comparisons. As an additional baseline, we randomly selected 50 genes out of the 2,636 development transcripts and optimised CNNs that each predicted a single gene. This model is referred to as *single gene*. ResNet18 was selected as the model architecture for all CNNs.

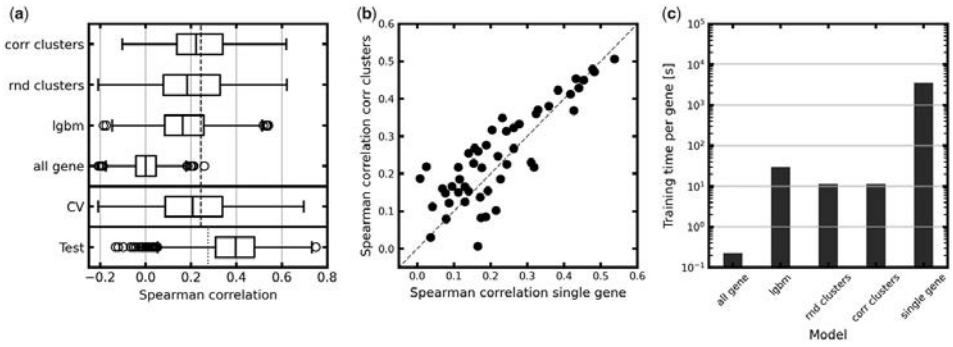

Figure 7. Performance overview from [66]. a) shows boxplots that summarise the distributions of Spearman correlations for 2,636 transcripts of the compared modelling approaches. The boxplot denoted with CV describes the distribution of Spearman correlations for the 15,586 transcripts for the proposed method corr clusters. Test indicates the boxplot that describes the distribution of Spearman correlations of the 6,618 transcripts that were selected for validation in the test set. b) shows a scatterplot between the Spearman correlations of 50 randomly selected transcripts for models trained with the corr clusters approach and single-gene prediction CNNs. c) shows a comparison of model optimisation times.

Figure 7 shows a comparison of the baseline model performances, as well as of the computational times for model optimization. As can be seen from Figure 7a), the proposed corr cluster modelling has higher Spearman correlations between predictions and RNA-seq estimates. In the evaluation of all 15,586 transcripts in the CV, 6,618 had a BH-adjusted p-value below 0.001. These were then evaluated in the test set, where 5,419 out of the 6,618 transcripts were significantly associated with predictions with BH-adjusted p-values below 0.001. Figure 7b) shows a scatterplot between the Spearman correlations of the single gene model and the corr clusters model. The p-value from a paired one-sided Wilcoxon rank sum test is below 0.01, indicating that the Spearman correlations are higher for the

corr clusters model. Figure 7c) indicates that the corr clusters model is substantially faster to optimise than single gene or lgbm models.

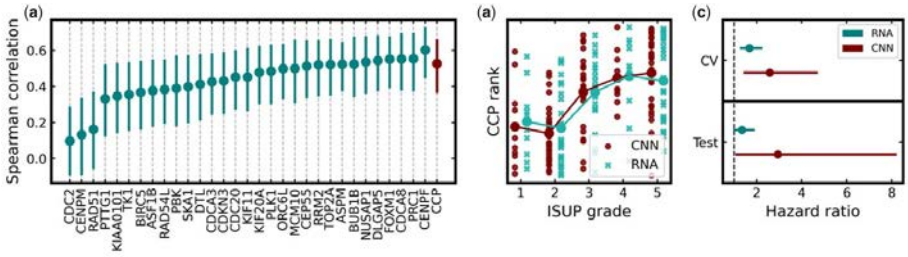

Figure 8. Comparison between the CNN and RNA-seq-based cell cycle progression (CCP) score, modified from [66]. a) shows the Spearman correlations between CNN-predicted and RNA-seq gene expression in the test set with bootstrapped 95% confidence intervals. b) displays the Ranked CCP scores for each ISUP grade both for the CNN and RNA-seq-based CCP score. c) shows the univariate HR of the CNN-predicted and RNA-seq-based CCP score for the CV and test set. In the CV data, the HR for the CNN predictions is 2.579 [1.412, 4.713] and 1.68 [1.256, 4.713] for RNA-seq. In the test set, the respective HRs are 2.943 [1.055, 8.212] for CNN predictions and 1.351 [0.956, 1.909] for RNA-seq.

Furthermore, we investigated the prediction of the CCP score, which was obtained by averaging the respective predictions of the transcripts included in the CCP score. The Spearman correlation for the CCP score based on CNN predictions and its RNA-seq counterpart is 0.527 [0.357, 0.665]. We also evaluated the hazards ratio for the time to biochemical recurrence in the CV data and the test set, as shown in Figure 8c). Due to a limited number of events in the test set, only univariable analysis was possible. In the test data, the RNA-seq-based CCP has a HR of 1.351 [0.956, 1.909], whereas the CNN-based CCP has a HR of 2.943 [1.055, 8.212].

This study indicates that the expression of many genes, including genes relevant in the context of prostate cancer, such as the ones comprising the CCP score, can be predicted from H&E WSIs to some degree. Predicting expression in clusters of co-expressed genes is a computationally efficient approach that can improve performance, but this conclusion may require further validation, particularly in external data. While prediction performances might not be sufficient to replace molecular assays, they may be sufficient to select patients who could benefit from further molecular testing. Furthermore, gene expression prediction from H&E WSIs could be useful as a cost-effective way to obtain some molecular profiling information in large-scale research studies, particularly on archived tissue materials.

6.2 Study II

In Study II, we investigated whether attention-based MIL, which has recently been shown to improve performance in some classification tasks, also improves the performance of regression objectives. We investigated this with a simulation and the prediction of gene expression from WSIs of H&E-stained breast cancer resection specimens. This analysis includes a data set with local expression estimates from spatial transcriptomics. For bulk gene expression prediction, we used the Spearman correlation between predictions and RNA-seq expression estimates as the primary performance metric. Spatial transcriptomics predictions were additionally evaluated with the proportion of variance explained by the fixed effects in LME models.

The simulation experiments in Study II are based on the MNIST data set [54], which consists of 28 x 28 pixel binary black and white images with handwritten digits in [0, 9]. The MNIST training data has 60,000 images, which we split into 48,000 images for model optimization and 12,000 images as a validation set for hyperparameter tuning. The MNIST test set consists of 12,000 images. Images were assigned to bags of size 32, which includes randomly generated images containing noise. The proportion of noisy images varied with 0%, 25%, 50% and 75% in different experiments. As a label for the bag, we set the mean of all MNIST images in the bag. This emulates a setting in which some instances in a bag do not contribute to the true label of the bag. All simulations were repeated 100 times.

The real-world application in this study is the prediction of bulk gene expression from WSIs of H&E-stained breast cancer resection specimens. WSIs originate from four different studies. From the first study, Clinseq, we included 270 patients. Furthermore, we selected 721 patients from the TCGA BRCA study for which sufficiently complete clinical information was available. 697 patients from these two data sets were randomly selected for model training. From the remaining patients, 122 were selected as a validation set and 172 as an internal test set. 350 patients from the ABiM study were chosen as an external test set, additionally to 22 patients with spatial transcriptomics data. We included 125 randomly selected transcripts out of 1,011 for which an association between predictions from WSIs and RNA-seq estimates had previously been established [65]. 25 transcripts were used for hyperparameter tuning and 100 for testing and model comparisons. ImageNet features were extracted for each tile containing invasive cancer using a ResNet18 model.

Both in the simulation and the gene expression prediction, we compared four modelling approaches. In the gene expression prediction, a WSI is considered as a bag of tiles, which are the instances in this case. The first model uses an attention-weighted average of instance features (AF) to generate bag-level representations, on which predictions are based. The second model uses an attention-weighted average of predictions (AP) to generate bag-level predictions from instance-level features. The third model uses the mean of all instance features (MF) as an input to a prediction module to generate bag-level predictions. The fourth model uses the mean of all instance-level predictions (MP) as the bag-level prediction. This model is conceptually comparable to the approach chosen in Study I, although models in Study I were trained end-to-end with some exceptions.

In the simulation experiments, we observed that models that are based on means outperform attention-based models if there are few images with noise in the respective bags. However, if a higher proportion of images containing noise is present, attention-based models perform better.

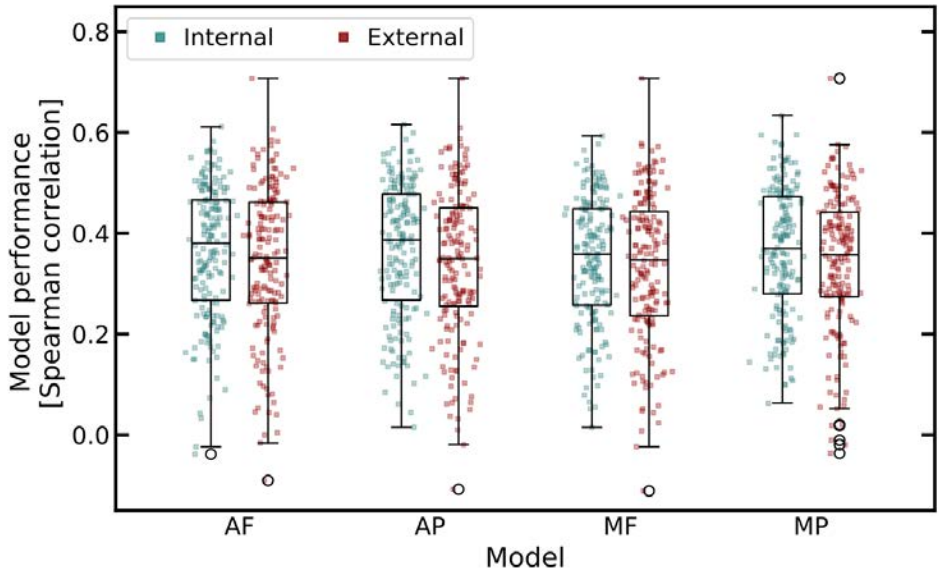

Figure 9. Boxplots with distributions of Spearman correlations for each modelling approach for the internal and external test data. From [83].

Figure 9 depicts boxplots and the distributions of Spearman correlations for bulk gene expression prediction in the internal and external test set. Paired one-sided Wilcoxon signed rank tests indicate that all models outperform the MF model with BH-adjusted p-values < 0.05. The median Spearman correlation of all models is

marginally higher in the internal test set compared to the external test set, with slightly higher decreases for attention-based models. For the AF model, the decrease in median Spearman correlation is 0.029, for the AP model 0.037, for the MF model 0.012 and for the MP model 0.012. In the spatial transcriptomics analysis, differences both in Spearman correlations and proportions of variance explained between the four models were marginal.

The simulation experiments of this study indicate that the performance improvements of attention-based MIL seen in classification might also translate to regression objectives to some degree. However, bulk gene expression prediction might not meet the conditions under which a MIL approach is beneficial. There appears to be no benefit of attention models for gene expression prediction in this study. Potentially, the prediction of gene expression is currently not limited through the choice of modelling approach but through data availability. Furthermore, it appears like the decrease in performance between internal and external data is slightly higher for the attention models, probably due to the increased complexity of the modelling approach. Attention-based MIL models therefore appear worthwhile to benchmark also for tasks outside of classification, however, they need to be carefully evaluated, both with regards to their assumptions and generalizability.

6.3 Study III & IV

Study III and Study IV are closely intertwined. In Study III, we conducted the AutomatiC Registration Of Breast cAncer Tissue (ACROBAT) challenge that was held in conjunction with the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2023 conference in Singapore. The purpose of the challenge was to establish the current state-of-the-art in the registration of differently stained WSIs that originate from slides from routine clinical workflows. Study IV describes the data set that we published to facilitate the ACROBAT challenge.

The ACROBAT data set consists of 1,153 female primary breast cancer patients that are part of the SöS study. 750 patients were randomly sampled for the training set and 100 patients for the validation set. The 303 test set cases were chosen as a subset of the SöS study that was previously selected for a different research project such that semantic annotations for different tissue types are available. Cases were excluded and randomly replaced by a new case if visual assessment of the WSIs revealed that the diagnostic slide contains two sections from the same tissue, as it would be unclear which of these sections should be the

target for registration. For each case in the training set, there is one WSI of H&E-stained tissue, along one to four WSIs with IHC-stained tissue with the four routine diagnostic stains ER, PGR, HER2 and KI67. The training set contains 3,406 WSIs. For cases in the validation and test set, there is one WSI of H&E-stained tissue and one corresponding randomly selected WSI of IHC-stained tissue. The validation set contains 200 WSIs, while the test set contains 606 WSIs. All WSIs were anonymised and image pyramid levels of 10X and lower magnifications were published. Alongside the WSIs, we anonymised clinical information such as year of diagnosis, grade, and biomarker statuses for the test set, as well as semantic annotations. Semantic annotations include the classes normal tissue, invasive cancer (IC), artefact, DCIS, LCIS and non-malignant changes (NMC). This data was not published alongside the WSIs. An overview of the cases, WSIs, stains, antibodies and scanners is available in Table 1.

	Training	Validation	Test	Total
cases	750	100	303	1503
slides	3406	200	606	4212
stain/antibodies				
H&E	750 (22%)	100 (50%)	303 (50%)	1153 (27.4%)
ER	732 (21.5%)	29 (14.5%)	84 (13.9%)	845 (20.1%)
KI67	732 (21.5%)	29 (14.5%)	82 (13.5%)	843 (20%)
PGR	728 (21.4%)	28 (14%)	81 (13.4%)	837 (19.9%)
HER2	464 (13.6%)	14 (7%)	56 (9.2%)	534 (12.7%)
scanners				
C13220	559 (16.4%)	38 (19%)	205 (33.8%)	802 (19%)
C12000-02	884 (26%)	46 (23%)	203 (33.5%)	1133 (26.9%)
C12000-22	1963 (57.6%)	116 (58%)	198 (32.7%)	2277 (54.1%)

Table 1. Overview of cases, WSIs, stains, antibodies, and scanner models in the ACROBAT training, validation, and test set.

All cases in the validation and test set were randomly assigned to 13 annotators, who placed pairs of landmarks in matched H&E and IHC WSIs, as shown in Figure 10. Annotators were asked to place 50 landmark pairs for each case. While there was only one phase of annotations in the validation set, there were two in the test set. Landmarks from the first annotation round in the H&E WSIs were randomly shifted by up to 115 µm and a second annotator was asked to move the landmark

in the H&E WSI to the correct location, based on the landmark position in the IHC WSI from the first annotation phase. In total, 35,760 landmark pairs were placed in the validation set and during the first and second annotation phase. This results in 5020 landmark pairs with one annotator each for performance evaluation in the validation set and 13,130 pairs with two annotators each for the test set after excluding landmarks with an annotator disagreement of more than 115 µm.

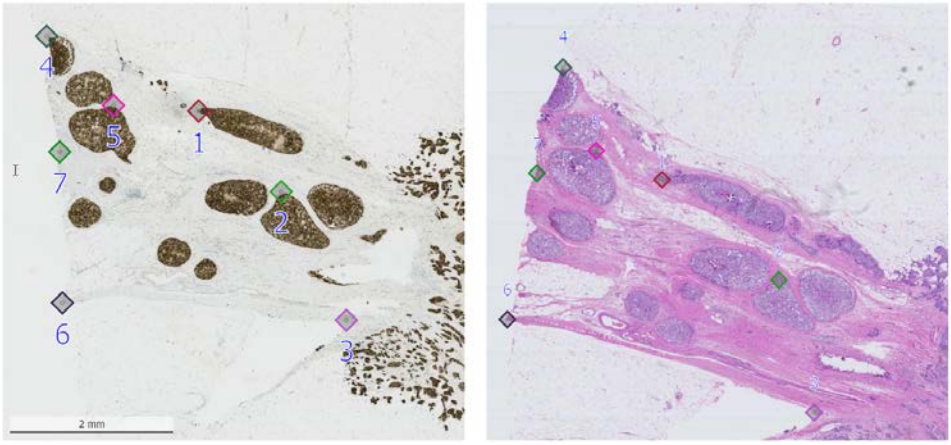

Figure 10. Example of pairs of landmarks in corresponding WSIs of H&E and IHC-stained tissue sections.

The ACROBAT challenge was conducted between the 1st of April 2022 and the 26th of August 2022, with an associated workshop at the MICCAI 2023 conference during which the test set leaderboard was announced publicly. Both for the validation and test set, we published landmarks that were placed in the IHC WSIs. Participants were then asked to register these landmarks to the corresponding H&E WSIs and submit registered landmark coordinates in micrometres. Participants had the opportunity to submit validation set landmarks on the ACROBAT challenge website (https://acrobat.grand-challenge.org/) to receive quantitative feedback on their registration performance. This submission system is intended to remain available indefinitely. Eight teams submitted all necessary materials to qualify to be ranked in the test set leaderboard.

The primary ranking metric was chosen as the median 90th percentile of target-registration-errors (TREs). The TRE is computed by obtaining the mean distance in micrometres between the registered landmark position and the positions set by the first and second annotator. The 90th percentiles are then computed across all landmarks within each WSI. Besides the median 90th percentile, we investigated the ranking for the 90th percentile of 90th percentiles, the mean 90th percentile, the median and mean TREs across all landmarks without

aggregating, and the mean reduction of TREs in percent between unregistered and registered landmark positions. For reference, we computed the corresponding metrics for the distance between first and second annotators (DBAs).

Besides an evaluation of algorithm performances, we also investigated covariates that impact algorithm TREs and the DBA using LME models. We used the $log10$-transformed TREs or DBA as the endogenous variable. As random effects, we included the combination of first and second annotator, as well as the WSI ID. As fixed effects, we included the antibody of the IHC WSI with ER as the reference, the semantic segmentation of a landmark with normal tissue as the reference category, the distance in millimetres between a landmark and the centre of tissue mass, the slide age in years, the NHG for landmarks within IC regions and the interaction of biomarker status and antibody for landmarks within IC regions. Based on the availability of this additional information, we included 11,465 test set landmarks into the LME analysis.

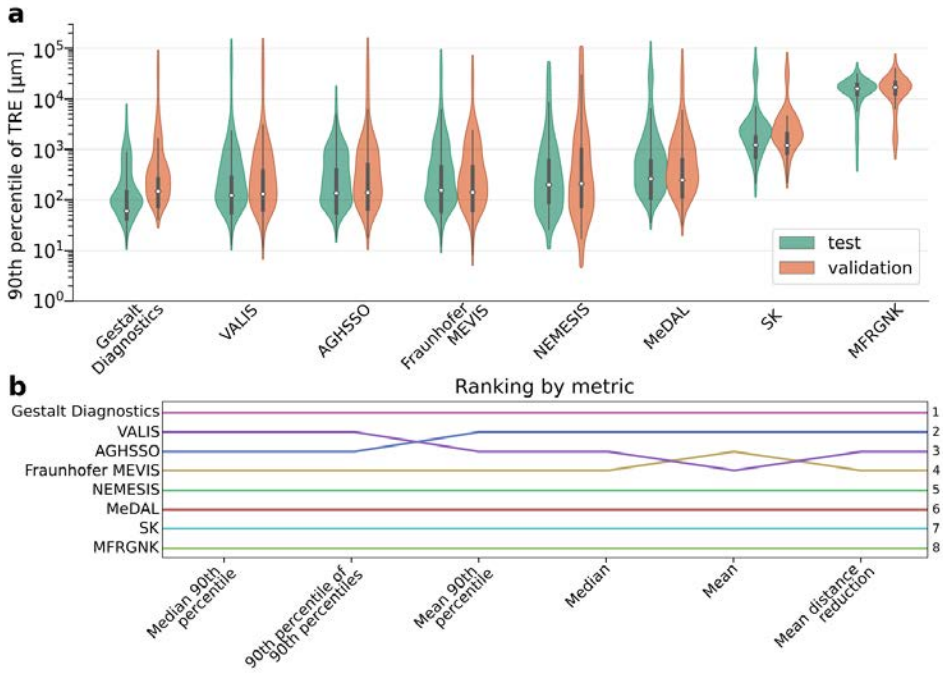

Figure 11. Distributions of 90th percentiles and rankings for different metrics. a) shows the distributions of 90th percentiles of TREs in the validation and test set as violin plots. b) displays the rank of different teams for each computed metric.

The highest score in the challenge was achieved by the team Gestalt Diagnostics, with a median 90th percentile in the test set of 60.1 [55.8, 68.6] µm. The median 90th percentile by the following teams is ca. twice as high, with 123.3 [98.5, 144.1]

µm for VALIS, 137.6 [120.3, 176.7] µm for AGHSSO and 155.3 [123.1, 184.7] µm for Fraunhofer MEVIS. Lower ranked methods have a median 90th percentile in the range of three or more times the score of Gestalt Diagnostics. Figure 11a) shows the distributions of 90th percentiles of TREs in the validation and test set as violin plots. Testing for differences in distributions with Mann-Whitney U rank tests implies that only for Gestalt Diagnostics, there is a difference between the distributions for validation and test set based on BH-adjusted p-values. Figure 11b) shows the ranking for all computed metrics, which are mostly stable across metrics, with some deviations for VALIS.

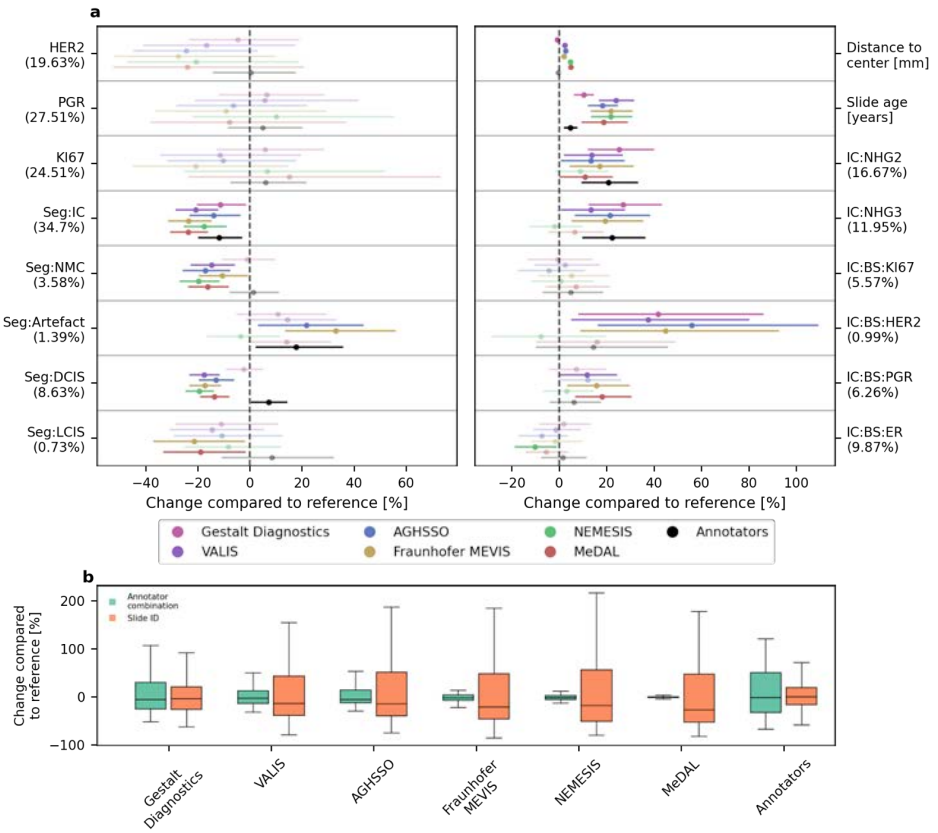

Figure 12. Fixed effect coefficients and conditional means of random effects of the LME model analysis, both for the TREs of the six teams with the highest ranking, as well as the DBA. a) displays the change in percent for a unit increase in the fixed effect, including 95% confidence intervals. If the confidence interval contains zero, the respective marker is transparent. Units are indicated for continuous fixed effects, whereas the percentage next to categorical fixed effects refers to the percentage of landmarks in the category. Effects with the prefix Seg refer to the semantic annotation of landmarks. b) displays boxplots that describe the distributions of percentage changes of conditional means of the random effects. The boxes indicate the lower to upper quartile of the distributions, while whiskers extend to 1.5 times the interquartile range or the maximum or minimum value. Outliers are not shown.

The results of the LME model analysis for the six highest ranked teams and the DBA in black are shown in Figure 12. Based on the coefficients from the LME analysis in Figure 12a), it appears that the antibody of a stain does not impact algorithm performances. A possible exception might be HER2, for which the point estimates of the fixed effects for all teams are below zero. The semantic annotations of tissue regions appear to generally be negatively associated with the TRE compared to the reference category of normal tissue, with the exception of artefacts that might only exist in one image of an image pair. A higher NHG in IC regions is associated with higher TREs. Positive biomarker statuses in IC regions are associated with higher TREs for HER2 and potentially with a weaker association for PGR. One of the main drivers of TREs appears to be the continuous slide age. At a slide age of four years, all teams except for Gestalt Diagnostics have an increase of TRE of ca. 100%. The other continuous fixed effect, the distance of a landmark to the tissue centre of mass, is also associated with higher TREs except for Gestalt Diagnostics. This is the only case in which a statistically significant coefficient for a team has a different sign from the coefficients for the other teams. Based on the LME analysis, we found fewer associations for the fixed effects of the DBA compared to TREs, and effect sizes appear to be smaller. The distributions of changes in percent for conditional means of the random effects are displayed in Figure 12b). The interquartile range for the annotator combination is highest for the DBA, followed by Gestalt Diagnostic and decreasing with worse performance. In contrast, the interquartile range for the WSI ID is lowest for annotators, followed by Gestalt Diagnostics and approximately increasing with worse performance.

The ACROBAT challenge establishes the current state-of-the-art in WSI registration. Mean registration errors in the order of magnitude of 100 µm and below might already approach the limit of possible performance for non-consecutive sections. While improvements in robustness are likely still desirable, errors this low enable a wide range of research and clinical applications. Furthermore, the challenge has generated novel insights into what affects WSI registration performances and can therefore guide future methods development. The difference between validation and test set distribution of performance metrics and the LME analysis indicate that for the highest ranked team, Gestalt Diagnostics, the DBA might be too high to measure further improvements in the method.

6.4 Study V

In Study V, we investigated whether cancer detection models that were optimised with registered annotations are inferior to cancer detection models that were optimised with annotations that were directly generated for the target WSIs. This has the potential to increase the usefulness of existing annotations and reduce the need for new annotations, which can be costly to generate. Furthermore, we evaluated the correlation between cancer detection performances and KI67 scores.

The data set that we used in this study is a subset of the SöS study. It includes WSIs from 272 breast cancer cases. For each case, there is one H&E and one KI67 WSI, with 544 WSIs in total. Annotations of invasive cancer regions are available for all WSIs. Annotations that were generated for the H&E WSIs were registered to the KI67 WSIs with the WSI registration algorithm proposed in [84], which corresponds to team AGHSSO in Study III. The data set was split into a development set for model optimisation and hyperparameter tuning including 218 cases, and a test set that includes 54 cases. The split was stratified based on low or high KI67 expression, split on the median KI67 score. We then optimised ResNet18 CNN models that predict whether a tile contains invasive cancer. One ensemble of models was optimised with the annotations that were generated directly in the KI67 WSIs. Another ensemble of models was optimised with the registered annotations.

	AUROC	Dice	Jaccard	Accuracy	Specificity	Sensitivity	Precision
manual annot.	0.974 [0.964, 0.982]	0.816 [0.768, 0.858]	0.718 [0.662, 0.771]	0.919 [0.899, 0.936]	0.921 [0.898, 0.94]	0.915 [0.882, 0.94]	0.78 [0.72, 0.835]
registered annot.	0.974 [0.965, 0.982]	0.813 [0.765, 0.858]	0.716 [0.657, 0.769]	0.921 [0.9, 0.939]	0.931 [0.908, 0.951]	0.888 [0.851, 0.917]	0.798 [0.737, 0.853]
BH-adj. p-value	0.962	0.879	0.885	0.891	0.006	<0.001	0.017

Table 2. Performance metrics for cancer detection models trained with manual annotations that were generated directly for the KI67 WSIs and for models trained with registered annotations.

All model performances were computed with annotations that were generated directly in the KI67 WSIs. We computed AUROC, Dice score, Jaccard index, accuracy, specificity, sensitivity/recall, and precision for the tiles within each WSI. This yields one value for each performance metric for each WSI. We then compared the distributions of performance metrics between the two modelling approaches with paired Wilcoxon signed rank tests. Performance metrics, including bootstrapped 95% confidence intervals and BH-adjusted p-values from the Wilcoxon tests are available in Table 2. The models differ in calibration, as visible from the significant differences for specificity, sensitivity, and precision, but not in performance, as is apparent from AUROC, Dice score, Jaccard index and accuracy.

	AUROC	Dice	Jaccard	Accuracy	Specificity	Sensitivity	Precision
manual label-registered label	0.954 [0.904, 0.983]	0.948 [0.907, 0.974]	0.948 [0.907, 0.974]	0.882 [0.784, 0.947]	0.879 [0.765, 0.95]	0.918 [0.857, 0.963]	0.964 [0.938, 0.981]
KI67-score - manual label	0.175 [-0.115, 0.452]	0.267 [-0.03, 0.538]	0.269 [-0.026, 0.542]	-0.029 [-0.348, 0.285]	-0.226 [-0.538, 0.109]	0.389 [0.138, 0.605]	0.149 [-0.161, 0.441]
KI67-score-registered label	0.221 [-0.06, 0.482]	0.286 [0.001, 0.547]	0.286 [0.006, 0.543]	-0.066 [-0.369, 0.246]	-0.257 [-0.547, 0.068]	0.477 [0.23, 0.681]	0.124 [-0.174, 0.417]

Table 3. Spearman correlations between performance metrics of the two modelling approaches, as well as between the performance metrics and the KI67 scores of the respective WSIs.

Furthermore, we investigated the correlations between model performance metrics and performance metrics and KI67 scores. The results are available in Table 3. Performance metrics between models are highly correlated. Regarding the correlation between performance metrics and the KI67-score, only the confidence interval of the sensitivity/recall does not include 0, indicating a higher sensitivity for WSIs with a higher KI67-score.

This study indicates that cancer detection models for different IHC stains can be trained with registered H&E annotations without a decrease in performance compared to generating new annotations for each stain. Compared to directly using registered annotations to identify cancer regions in IHC WSIs, this can be necessary if there are no (consecutive) H&E sections available. Most semantic annotations in histopathology currently exist for H&E WSIs. This study provides some evidence that the usefulness of these annotations can be substantially increased through WSI registration. However, validation with external data and further IHC stains may be necessary to draw firm conclusions.

7 Discussion

Since I began working on this thesis in May 2019, the field of computational pathology has made significant advances. There are, for example, several studies that thoroughly investigate AI systems that automate Gleason grading of prostate biopsies, some of which were conducted with the participation of Karolinska Institutet [85]–[87]. Further examples that go beyond what pathologists can visually assess include the prediction of novel prognostic markers [88]–[90] and genetic alterations [91], [92]. I hope that this thesis will make a modest contribution towards the advancement of computational pathology in research and translational applications. The studies comprising this thesis focus on the development, application, and evaluation of methods for gene expression prediction from WSIs, as well as WSI registration. The assessment of gene expression from WSIs is an application that can generally not be performed by human pathologists. WSI registration can be fully automated or based on manual input, either of rotation and translation or the selection of matching landmarks, where the registration precision increases to some degree with the number of landmarks placed. However, highly accurate non-rigid registrations require sophisticated algorithms and can generally not be performed by pathologists at scale. It can therefore be argued that this thesis focuses on methods that expand the capabilities of pathologists, rather than automating routine tasks.

The modelling approach developed in Study I appears promising, with modest improvements in prediction performance but substantial improvements in computational efficiency, both for training and prediction. We aimed at a fair evaluation against several plausible baseline models, but potential comparative benefits for performance need to be further validated. In Study II, we investigated several modelling approaches for gene expression prediction with the objective to evaluate attention-based MIL models in this context. However, no relevant differences between modelling approaches became apparent. It is unclear whether this study is limited through data availability and if differences would become apparent if more training data were available. For now, it appears like the simplest modelling approach is preferable as there are no performance differences.

There are several studies that evaluate the potential of AI models for gene expression prediction [65], [66], [93], [94]. It appears therefore well established that there is a learnable association between morphology in WSIs and gene

expression of individual genes. Generally, the prediction performances in these studies are similar. This could indicate that these studies are limited through the same factors. Limits on the performance could either be imposed by current modelling approaches or the availability of training data. While all studies deploy slightly different modelling approaches, all of these are based on CNNs and often ImageNet weights. Future research should investigate whether performance improvements are possible with additional data. It is also possible that the association between morphology and gene expression does not allow for higher prediction performances. Furthermore, all these studies rely on TCGA data to some degree, if not entirely. It is therefore essential that findings are further validated with additional data sets. While prediction performances generally do not appear high enough to replace molecular assays, current performances might already suffice to detect patients that could benefit from further molecular profiling. Furthermore, gene expression prediction from routine diagnostic slides has the potential to be a cost-effective means to facilitate large-scale epidemiological studies based on archived materials.

The spatial transcriptomics analysis of gene expression predictions in Study II and [65] is based on a manual registration between sequenced slides and diagnostic H&E WSIs. Generally however, the main applications of WSI registration currently revolve around 3D reconstruction [95], [96], the transfer of annotations or segmentations between WSIs as in Study V and [97], [98], stain-guided learning [99]–[101], virtual staining [102]–[105], and the analysis of multiplex stained histology [106], [107]. These are promising avenues of research that can deliver both new biomedical insights, as well as tools for clinicians and researchers. However, the state-of-the-art of WSI registration on data from routine clinical workflows was unclear prior to the ACROBAT challenge. While the ANHIR WSI registration challenge [108] laid important groundwork for the evaluation of WSI registration methods, the training and test data were not sufficiently separated to independently assess performance. The ANHIR data set consists of WSIs with a higher variety of stains and organs than the ACROBAT data set. During internal evaluations of some of the ANHIR methods, we realised that despite very high reported robustness in the ANHIR challenge, methods failed on a significant proportion of data from routine workflows, potentially due to lower quality of tissue materials and more artefacts. I therefore believe that the ACROBAT challenge is a highly useful point of reference to move the field of WSI registration forward, which will in turn enable further research studies that require WSI

registration. Research community challenges have the advantage of independently evaluating methods through a third party, in this case me and my collaborators Masi Valkonen and Leslie Solorzano. While the evaluation of registration performance through landmarks has its limitations, e.g. only estimating performances in the location of landmarks, which might in turn be a biased selection that favours landmarks that are easy to recognize by humans, we believe to have performed one of the most thorough investigations of WSI registration to date. Furthermore, we could elucidate factors that impact algorithms with the LME analysis. The submission system for the validation data is fully automated and will remain open as a resource for the community without a planned expiration date. Validation with the test data will require contacting us to avoid that a high number of submissions on the test set invalidates its independence.

Besides the challenge itself, I hope that the ACROBAT data set will prove a valuable resource for the research community. To the best of my knowledge, the ACROBAT data set is the currently largest publicly available WSI data set that has multiple stains available for sections from the same tumour. While WSIs are only available at 10X and it was not possible to share clinical information, I believe that this data set can still be useful for many avenues of research outside of WSI registration. These include the developments of digital staining and stain transfer methods, stain-guided learning, tissue segmentation and classification, as well as artefact detection and unsupervised pre-training. Many research groups do not have the resources to digitise large numbers of slides and can only conduct research with publicly available data. Current incentives in academia do not always encourage data sharing, and I therefore hope that this data set can aid in advancing the field of computational pathology.

The final study of this thesis, Study V, is an application example of WSI registration. I intended to showcase how current WSI registration methods can be used to answer research questions in computational pathology. I believe that besides the performance of WSI registration methods, their increasing ease of use will lead to a broader proliferation of these methods for research that requires fusing information from multiple WSIs. Therefore, I expect that the proportion of studies deploying WSI registration will increase during the coming years.

However, despite the rapid advances in computational pathology in research studies in recent years, the impact on clinical practice is as of yet limited [45], [109]. To this day, relatively few regulatory approved applications, e.g. compared

to radiology, are available. Reis-Filho and Kather recently identified five obstacles that currently prevent the broad application of computational pathology tools into clinical routines [109]. First, there can be cultural resistance to new technologies in healthcare systems if the current approaches appear to work sufficiently well. Second, differences in tissue cutting, fixation and staining procedures can lead to diverging algorithm performances for different laboratories, which might then necessitate laboratory-specific validation or calibration. Third, validation of computational pathology biomarkers needs to be as rigorous as for e.g. genetic biomarkers, but current studies often lack external validation or sufficiently complete reporting of results to assess limitations. Fourth, there are rapidly evolving regulatory requirements, which can make it difficult for companies to bring their products to market. Fifth, there is a high initial cost for scanners when transitioning to digital pathology workflows, while the benefit in cost or patient outcomes is yet unclear. Despite these limitations, it is expected that computational pathology will have a significant impact on clinical routines during the next decades.

As Reis-Filho and Kather point out, rigorous validation in external data is essential. Future work might therefore be necessary to gather further evidence for the findings of this thesis and many current studies in the field. However, the incentives in academia might not favour such validation studies but rather reward novelty. Nevertheless, I believe that e.g. further validating the findings from the ACROBAT challenge with data from other institutions and organs is a worthwhile endeavour. We are currently attempting to lay the groundwork for this with the ACROBAT 2023 challenge, which adds undisclosed IHC stains and IHC-IHC image pairs to the test set, albeit still from the SöS study. Computational pathology methods also have the potential to expand global access to precision diagnostics. However, rigorous validation in the respective populations will be crucial. It is therefore essential to begin data collection for these validation studies as soon as possible to guarantee patient safety globally, particularly in resource-constrained environments where patients may be less protected through appropriate regulatory frameworks. Besides a lack of external validation, another current limitation of the field could be data set sizes. In order to fully unlock the potential of AI to identify patterns that human experts may not yet have discovered, AI models need to be trained with labels that are not based on human expertise, such as outcome data. Current data set sizes, particularly in breast cancer where many patients are cured, may still be insufficient to train models directly with outcome

data such as recurrences due to small numbers of events. Furthermore, there is a lack of standardisation both for image acquisition and formats, as well as pathology reports. This lack of standardisation makes data curation and preparation for large-scale studies very laborious. Without standardisation of pathology reports, it can be difficult to obtain consistent information across data sources. The DICOM standard may alleviate divergent imaging formats. Structured and standardised reporting is an active area of research that can improve the completeness of information in pathology reports [110]. This may in turn aid computational pathology through consistently providing necessary information for training and labels. While these are limitations that are unlikely to be solved through methods development, there are also remaining challenges that could be overcome through technological improvements. One such limitation is the generalisation of methods across different laboratories, where differences in sample preparation and image acquisition procedures can pose substantial challenges. This does not only affect algorithm performances, but also calibration, which is necessary for the deployment of methods. Research into image normalisation and domain generalisation techniques will therefore be crucial. I did not actively pursue work in this area in terms of this thesis, but believe that research in this field is essential for the clinical translation of computational pathology methods.

8 Conclusions

In this thesis, I focused on the development, application, and evaluation of methods for gene expression prediction from WSIs, as well as WSI registration methods. While the gene expression prediction methods evaluated in Study II ultimately appear to not improve prediction performance, the method proposed in Study I offers a computationally highly efficient approach for end-to-end training that may improve prediction performance. The ACROBAT challenge and its associated data set could significantly advance the narrow field of WSI registration, as it is currently the most thorough comparison of WSI registration methods. The data set is the to date largest publicly available data set with WSIs of multiple stains from the same tumour, which may enable many avenues of research in the field, which is still limited through data availability. In Study V, we could demonstrate how WSI registration can facilitate the investigation of novel research questions. The main limitation of the studies is the lack of validation in external data. Nevertheless, I believe that the findings of this thesis will prove useful for the computational pathology research community. While there are many interesting directions for future work, I believe that particularly the rigorous validation of existing studies is currently the most pressing. Furthermore, the investigation of methods that improve generalisation across data sources is crucial to facilitate the translation of research findings into tools that can benefit patients.

9 Acknowledgements

I would like to express my sincerest gratitude to everyone who contributed to this thesis and who supported me throughout my doctoral studies.

First and foremost, I would like to thank my supervisor **Mattias Rantalainen**. You are brilliant, kind, hard-working and generous and I appreciate deeply how much I have learned from you throughout the years, both academically and personally. Your attention to detail and foresight continue to amaze me.

I would also like to thank my co-supervisor **Johan Hartman**. Your deep expertise in pathology and ambition to make a difference with your work has enabled much of my research and your infectious positivity has been an inspiration.

To my co-supervisor **Martin Eklund**, thank you for your support. Academia is better for having someone with your clarity of thought and integrity contribute to it.

My co-supervisor **Henrik Grönberg** for helping with your expertise in and the discussions around prostate cancer.

Furthermore, I would like to thank **Anne Martel** for serving as my opponent and **Gustaf Edgren**, **Pernilla Wikström** and **Claes Lundström** for serving as the examination board for my dissertation. **Mark Clements** for chairing my dissertation seminar and your delightful presence.

My friends and colleagues from the Predictive Medicine Group, **Duong Tran, Ariane Buckenmeyer, Abhinav Sharma, Bojing Liu, Constance Boissin, Leslie Solorzano, Yanbo Feng, Yujie Xiang.** Thank you for your essential help and support in my research studies. It would not have been possible without you, you made my doctoral studies so much more enjoyable!

I would also like to thank my colleagues from Stratipath. **Stephanie Robertson, Sandy Kang Lövgren, Kajsa Ledesma Eriksson, Johnson Ho, Yinxi Wang, Binbin Su, Emma Jansson, Fredrik Wetterhall, Jennie Ousbäck, Lars Lengquist, Annica Jämtén Ericsson** and **Emelie Karlsson.** You are hard-working, competent, kind, and a pleasure to work with. You always had my back during the last years.

Furthermore, I would like to thank my friends and colleagues from Martin Eklund's group for their contributions and support. **Nita Mulliqi, Kimmo Kartasalo, Xiaoyi Ji, Henrik Olsson, Peter Ström, Kelvin Szolnoky, Matteo Titus**, **Alessio Crippa, Andrea Discacciati**. Your amazing work has been an inspiration!

Members of the ABCAP research consortium contributed to this thesis, particularly the ACROBAT challenge, and I would like to express my gratitude to **Masi Valkonen, Umair Khan, Circe Carr, Aino Kuusela** and **Pekka Ruusuvuori** from the University of Turku, Finland. **Sandra Sinius Pouplier, Dusan Rasic** and **Anne-Vibeke Laenkholm** from Zealand University Hospital, Denmark. **Sonja Koivukoski** and **Leena Latonen** from the University of Eastern Finland. Without your firm support, the ACROBAT challenge could not have succeeded. Thank you for all the hard work and the patience! Finally, **Balazs Acs**, for your impressive determination and competence!

My colleagues from MEB, who have made these years so much fun! **Shuang Hao, Emilio Ugalde Morales, Andreas Jangmo, Nikolaos Skourlis, Enoch Yi-Tung Chen, Tyra Lagerberg, Hilda Björk Daníelsdóttir, Qian Yang, Marica Leone, Ruyue Zhang, Alexander Ploner, Erin Gabriel** and everybody who contributes to the MEB spirit. **Alessandra Nanni** and **Anna Berglund**, without whom I could not have mastered the administrative duties.

My friends. **Tobias Visser**, **Philipp Rosin, Aylin Sevis, Heather Iriye, Xavier Job, Victor Ahlén, Eleni Frountzou, Xinhe Mao, Maryam Hami** for your unconditional friendship and support. **Jonas Lantto** and **Felix Bangel** for always being wrong about everything. **Pascal Rütten** and **Carlos Kraft** for inspiring me with your drive. **Wanda Christ** for your kindness and positivity. **Juliette Rivard, Felix Vaisfeld, Sebastian Bruckmann, Miquel Martí, Pavel Karpashevich, Ines & Maya Alsheh Ali, Jordi Riera, Adría Cruz, Bastiaan Hezemans, Sebastian Köster, Oliver Mendoza, Larissa Anthofer, Joshua Fineberg, Francesca Stirbu, Viktor Wohlfahrt, Mattis Steuer, André de Oliveira Gomes, Michael Baumgartner, Rosana Hinojosa** for all the living, dancing, adventures, and shenanigans. **Jet Termorshuizen, Tom Niessen, Nick Judd, Tobias Grelsson** for mostly letting me down softly. **Lea Petermann** for keeping Humorle in check. **Yannick Stolpmann, Lilly Kohaus, Velten Stille, Daniel Tillmann, Vitus Appel, Martin Kramer, Christoph Haarburger** for making my time in Aachen so much better.

Silke & Fritz Joussen, Friedrich Thiemann and **Barbara Höcker** for your kindness, support and patience during all these years.

Laura Olivera Nieto for your brilliance, determination, and confidence. I have learned so much from you. Thank you for all the support during my doctoral studies, you have my deepest gratitude.

Last but not least, I would like to thank my family **Moritz, Hannah, Birgit, Josef, Johanna,** and my late grandparents **Guido, Elisabeth,** and **Josef**. You have given me more than I can ever express here, and I could not have done this without you.

Finally, I would also like to thank whomever I inevitably forgot here for their forbearance.

10 References

[1] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, no. 1, pp. 57–70, Jan. 2000.

[2] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, no. 5, pp. 646–674, Mar. 2011.

[3] E. J. Odes *et al.*, "Earliest hominin cancer: 1.7-million-year-old osteosarcoma from Swartkrans Cave, South Africa," *S. Afr. J. Sci.*, vol. 112, no. 7/8, p. 5, Jul. 2016.

[4] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, May 2021.

[5] F. Cardoso *et al.*, "Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up," *Ann. Oncol.*, vol. 30, no. 8, pp. 1194–1220, Aug. 2019.

[6] M. J. Thun, M. S. Linet, J. R. Cerhan, C. A. Haiman, and D. Schottenfeld, *Cancer Epidemiology and Prevention*. Oxford University Press, 2017.

[7] A. Kwong, "Is smoking a risk factor of breast cancer?," *Nov. Approaches Cancer Study*, vol. 2, no. 3, Apr. 2019.

[8] "Standardiserat vårdförlopp bröstcancer." https://kunskapsbanken.cancercentrum.se/diagnoser/brostcancer/vardforlopp/ (accessed Aug. 21, 2023).

[9] G. Cserni and A. Sejben, "Grading Ductal Carcinoma In Situ (DCIS) of the Breast – What's Wrong with It?," *Pathol. Oncol. Res.*, vol. 26, no. 2, pp. 665–671, Apr. 2020.

[10] C. W. Elston and I. O. Ellis, "Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up," *Histopathology*, vol. 19, no. 5, pp. 403–410, Nov. 1991.

[11] B. Acs *et al.*, "Variability in Breast Cancer Biomarker Assessment and the Effect on Oncological Treatment Decisions: A Nationwide 5-Year Population-Based Study," *Cancers*, vol. 13, no. 5, Mar. 2021.

[12] X. Chen, Y. Yuan, Z. Gu, and K. Shen, *Accuracy of estrogen receptor, progesterone receptor, and HER2 status between core needle and open excision*

biopsy in breast cancer: a meta-analysis. Centre for Reviews and Dissemination (UK), 2012.

[13] I. Paterni, C. Granchi, J. A. Katzenellenbogen, and F. Minutolo, "Estrogen receptors alpha (ERα) and beta (ERβ): subtype-selective ligands and clinical potential," *Steroids*, vol. 90, pp. 13–29, Nov. 2014.

[14] K. H. Allison *et al.*, "Estrogen and Progesterone Receptor Testing in Breast Cancer: ASCO/CAP Guideline Update," *J. Clin. Oncol.*, vol. 38, no. 12, pp. 1346–1366, Apr. 2020.

[15] K. B. Horwitz and C. A. Sartorius, "90 YEARS OF PROGESTERONE: Progesterone and progesterone receptors in breast cancer: past, present, future," *J. Mol. Endocrinol.*, vol. 65, no. 1, pp. T49–T63, Jul. 2020.

[16] Z. Mitri, T. Constantine, and R. O'Regan, "The HER2 Receptor in Breast Cancer: Pathophysiology, Clinical Use, and New Advances in Therapy," *Chemother. Res. Pract.*, vol. 2012, p. 743193, Dec. 2012.

[17] S. C. Y. Leung *et al.*, "Analytical validation of a standardised scoring protocol for Ki67 immunohistochemistry on breast cancer excision whole sections: an international multicentre collaboration," *Histopathology*, vol. 75, no. 2, pp. 225–235, Aug. 2019.

[18] M. Dowsett *et al.*, "Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group," *J. Natl. Cancer Inst.*, vol. 103, no. 22, pp. 1656–1664, Nov. 2011.

[19] T. O. Nielsen *et al.*, "Assessment of Ki67 in Breast Cancer: Updated Recommendations from the International Ki67 in Breast Cancer Working Group," *J. Natl. Cancer Inst.*, vol. 113, no. 7, pp. 808–819, 2021.

[20] "Nationellt vårdprogram bröstcancer." https://kunskapsbanken.cancercentrum.se/diagnoser/brostcancer/vardprogram/sammanfattning/ (accessed Aug. 22, 2023).

[21] J. S. Parker *et al.*, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *J. Clin. Oncol.*, vol. 27, no. 8, pp. 1160–1167, Mar. 2009.

[22] S. Paik *et al.*, "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer," *N. Engl. J. Med.*, vol. 351, no. 27, pp. 2817–2826, Dec. 2004.

[23] B. Demir Cetinkaya and C. Biray Avci, "Molecular perspective on targeted therapy in breast cancer: a review of current status," *Med. Oncol.*, vol. 39, no. 10, p. 149, Jul. 2022.

[24] C. K. Zhou *et al.*, "Prostate cancer incidence in 43 populations worldwide: An analysis of time trends overall and by age group," *Int. J. Cancer*, vol. 138, no. 6, pp. 1388–1400, Mar. 2016.

[25] G. Gandaglia *et al.*, "Epidemiology and Prevention of Prostate Cancer," *Eur Urol Oncol*, vol. 4, no. 6, pp. 877–892, Dec. 2021.

[26] O. Bratt *et al.*, "The Swedish national guidelines on prostate cancer, part 1: early detection, diagnostics, staging, patient support and primary management of non-metastatic disease," *Scand. J. Urol.*, vol. 56, no. 4, pp. 265–273, Aug. 2022.

[27] H. Grönberg *et al.*, "Prostate cancer screening in men aged 50-69 years (STHLM3): a prospective population-based diagnostic study," *Lancet Oncol.*, vol. 16, no. 16, pp. 1667–1676, Dec. 2015.

[28] P. Ström, T. Nordström, M. Aly, L. Egevad, H. Grönberg, and M. Eklund, "The Stockholm-3 Model for Prostate Cancer Detection: Algorithm Update, Biomarker Contribution, and Reflex Test Potential," *Eur. Urol.*, vol. 74, no. 2, pp. 204–210, Aug. 2018.

[29] D. F. Gleason, "Classification of prostatic carcinomas," *Cancer Chemother. Rep.*, vol. 50, no. 3, pp. 125–128, Mar. 1966.

[30] J. I. Epstein, M. B. Amin, V. E. Reuter, and P. A. Humphrey, "Contemporary Gleason Grading of Prostatic Carcinoma: An Update With Discussion on Practical Issues to Implement the 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma," *Am. J. Surg. Pathol.*, vol. 41, no. 4, pp. e1–e7, Apr. 2017.

[31] J. I. Epstein, W. C. Allsbrook Jr, M. B. Amin, L. L. Egevad, and ISUP Grading Committee, "The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma," *Am. J. Surg. Pathol.*, vol. 29, no. 9, pp. 1228–1242, Sep. 2005.

[32] L. Egevad *et al.*, "Standardization of Gleason grading among 337 European pathologists," *Histopathology*, vol. 62, no. 2, pp. 247–256, Jan. 2013.

[33] J. I. Epstein *et al.*, "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.*, vol. 40, no. 2, pp. 244–252, Feb. 2016.

[34] J. T. Bishoff *et al.*, "Prognostic utility of the cell cycle progression score generated from biopsy in men treated with prostatectomy," *J. Urol.*, vol. 192, no. 2, pp. 409–414, Aug. 2014.

[35] M. R. Cooperberg *et al.*, "Validation of a cell-cycle progression gene panel to improve risk stratification in a contemporary prostatectomy cohort," *J. Clin. Oncol.*, vol. 31, no. 11, pp. 1428–1434, Apr. 2013.

[36] J. Cuzick *et al.*, "Prognostic value of a cell cycle progression signature for prostate cancer death in a conservatively managed needle biopsy cohort," *Br. J. Cancer*, vol. 106, no. 6, pp. 1095–1099, Mar. 2012.

[37] N. Erho *et al.*, "Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy," *PLoS One*, vol. 8, no. 6, p. e66855, Jun. 2013.

[38] M. Marrone, A. L. Potosky, D. Penson, and A. N. Freedman, "A 22 Gene-expression Assay, Decipher® (GenomeDx Biosciences) to Predict Five-year Risk of Metastatic Prostate Cancer in Men Treated with Radical Prostatectomy," *PLoS Curr.*, vol. 7, Nov. 2015.

[39] P. L. Nguyen *et al.*, "Ability of a Genomic Classifier to Predict Metastasis and Prostate Cancer-specific Mortality after Radiation or Surgery based on Needle Biopsy Specimens," *Eur. Urol.*, vol. 72, no. 5, pp. 845–852, Nov. 2017.

[40] J. Cullen *et al.*, "A Biopsy-based 17-gene Genomic Prostate Score Predicts Recurrence After Radical Prostatectomy and Adverse Surgical Pathology in a Racially Diverse Population of Men with Clinically Low- and Intermediate-risk Prostate Cancer," *Eur. Urol.*, vol. 68, no. 1, pp. 123–131, Jul. 2015.

[41] G. Eure *et al.*, "Use of a 17-Gene Prognostic Assay in Contemporary Urologic Practice: Results of an Interim Analysis in an Observational Cohort," *Urology*, vol. 107, pp. 67–75, Sep. 2017.

[42] E. A. Klein *et al.*, "A 17-gene assay to predict prostate cancer aggressiveness in the context of Gleason grade heterogeneity, tumor multifocality, and biopsy undersampling," *Eur. Urol.*, vol. 66, no. 3, pp. 550–560, Sep. 2014.

[43] D. Knezevic *et al.*, "Analytical validation of the Oncotype DX prostate cancer assay - a clinical RT-PCR assay optimized for prostate needle biopsies," *BMC Genomics*, vol. 14, p. 690, Oct. 2013.

[44] S. K. Van Den Eeden *et al.*, "A Biopsy-based 17-gene Genomic Prostate Score as a Predictor of Metastases and Prostate Cancer Death in Surgically Treated Men with Clinically Localized Disease," *Eur. Urol.*, vol. 73, no. 1, pp. 129–138, Jan. 2018.

[45] L. Pantanowitz, A. Sharma, A. B. Carter, T. Kurc, A. Sussman, and J. Saltz, "Twenty Years of Digital Pathology: An Overview of the Road Travelled, What is on the Horizon, and the Emergence of Vendor-Neutral Archives," *J. Pathol. Inform.*, vol. 9, p. 40, Nov. 2018.

[46] C. Chang, B. Moon, A. Acharya, C. Shock, A. Sussman, and J. Saltz, "Titan: a high-performance remote-sensing database," in *Proceedings 13th International Conference on Data Engineering*, Apr. 1997, pp. 375–384.

[47] A. Goode, B. Gilbert, J. Harkes, D. Jukic, and M. Satyanarayanan, "OpenSlide: A vendor-neutral software foundation for digital pathology," *J. Pathol. Inform.*, vol. 4, p. 27, Sep. 2013.

[48] R. Singh, L. Chubb, L. Pantanowitz, and A. Parwani, "Standardization in digital pathology: Supplement 145 of the DICOM standards," *J. Pathol. Inform.*, vol. 2, p. 23, May 2011.

[49] A. Saco, J. Ramírez, N. Rakislova, A. Mira, and J. Ordi, "Validation of Whole-Slide Imaging for Histolopathogical Diagnosis: Current State," *Pathobiology*, vol. 83, no. 2–3, pp. 89–98, Apr. 2016.

[50] S. Mukhopadhyay *et al.*, "Whole Slide Imaging Versus Microscopy for Primary Diagnosis in Surgical Pathology: A Multicenter Blinded Randomized Noninferiority Study of 1992 Cases (Pivotal Study)," *Am. J. Surg. Pathol.*, vol. 42, no. 1, pp. 39–52, Jan. 2018.

[51] L. A. Cooper, E. G. Demicco, J. H. Saltz, R. T. Powell, A. Rao, and A. J. Lazar, "PanCancer insights from The Cancer Genome Atlas: the pathologist's perspective," *J. Pathol.*, vol. 244, no. 5, pp. 512–524, Apr. 2018.

[52] J. Borovec, A. Munoz-Barrutia, and J. Kybic, "Benchmarking of Image Registration Methods for Differently Stained Histological Slides," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct. 2018, pp. 3368–3372.

[53] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychol. Rev.*, vol. 65, no. 6, pp. 386–408, Nov. 1958.

[54] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[55] K. Fukushima, "Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, 1980.

[56] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[57] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," vol. 37, pp. 448–456, 07--09 Jul 2015.

[58] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[59] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans. Syst. Man Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[60] P. Bándi, M. Balkenhol, B. van Ginneken, J. van der Laak, and G. Litjens, "Resolution-agnostic tissue segmentation in whole-slide histopathology images with convolutional neural networks," *PeerJ*, vol. 7, p. e8242, Dec. 2019.

[61] M. Macenko *et al.*, "A method for normalizing histology slides for quantitative analysis," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, Jun. 2009, pp. 1107–1110.

[62] C. Szegedy *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2015, pp. 1–9.

[63] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818–2826.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778.

[65] Y. Wang *et al.*, "Predicting Molecular Phenotypes from Histopathology Images: A Transcriptome-Wide Expression-Morphology Analysis in Breast Cancer," *Cancer Res.*, vol. 81, no. 19, pp. 5115–5126, Oct. 2021.

[66] P. Weitz *et al.*, "Transcriptome-wide prediction of prostate cancer gene expression from histopathology images using co-expression-based convolutional neural networks," *Bioinformatics*, vol. 38, no. 13, pp. 3462–3469, Jun. 2022.

[67] M. Ilse, J. Tomczak, and M. Welling, "Attention-based Deep Multiple Instance Learning," vol. 80, pp. 2127–2136, 2018.

[68] M. Y. Lu *et al.*, "AI-based pathology predicts origins for cancers of unknown primary," *Nature*, May 2021.

[69] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nat Biomed Eng*, Mar. 2021.

[70] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer US.

[71] D. R. Cox, "Regression Models and Life-Tables," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 34, no. 2, pp. 187–220, 1972.

[72] R. A. Fisher, "XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance," *Earth Environ. Sci. Trans. R. Soc. Edinb.*, vol. 52, no. 2, pp. 399–433, Jan. 1919.

[73] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software, Articles*, vol. 67, no. 1, pp. 1–48, 2015.

[74] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[75] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Stat. Soc.*, 1995.

[76] Cancer Genome Atlas Research Network, "The Molecular Taxonomy of Primary Prostate Cancer," *Cell*, vol. 163, no. 4, pp. 1011–1025, Nov. 2015.

[77] Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, Oct. 2012.

[78] J. Holm *et al.*, "Risk factors and tumor characteristics of interval cancers by mammographic density," *J. Clin. Oncol.*, vol. 33, no. 9, pp. 1030–1037, Mar. 2015.

[79] M. Rantalainen *et al.*, "Sequencing-based breast cancer diagnostics as an alternative to routine biomarkers," *Sci. Rep.*, vol. 6, p. 38037, Nov. 2016.

[80] L. H. Saal *et al.*, "The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine," *Genome Med.*, vol. 7, no. 1, p. 20, Feb. 2015.

[81] L. Löfgren *et al.*, "Validation of data quality in the Swedish National Register for Breast Cancer," *BMC Public Health*, vol. 19, no. 1, p. 495, May 2019.

[82] G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 3146–3154, 2017.

[83] P. Weitz, Y. Wang, J. Hartman, and M. Rantalainen, "An investigation of attention mechanisms in histopathology whole-slide-image analysis for regression objectives," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, IEEE, Oct. 2021.

[84] M. Wodzinski and H. Müller, "DeepHistReg: Unsupervised Deep Learning Registration Framework for Differently Stained Histology Samples," *Comput. Methods Programs Biomed.*, vol. 198, p. 105799, Jan. 2021.

[85] P. Ström *et al.*, "Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study," *Lancet Oncol.*, vol. 21, no. 2, pp. 222–232, Feb. 2020.

[86] W. Bulten *et al.*, "Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study," *Lancet Oncol.*, vol. 21, no. 2, pp. 233–241, Feb. 2020.

[87] W. Bulten *et al.*, "Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge," *Nat. Med.*, vol. 28, no. 1, pp. 154–163, Jan. 2022.

[88] Y. Wang *et al.*, "Improved breast cancer histological grading using deep learning," *Ann. Oncol.*, vol. 33, no. 1, pp. 89–98, Jan. 2022.

[89] S. Foersch *et al.*, "Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer," *Nat. Med.*, vol. 29, no. 2, pp. 430–439, Feb. 2023.

[90] O.-J. Skrede *et al.*, "Deep learning for prediction of colorectal cancer outcome: a discovery and validation study," *Lancet*, vol. 395, no. 10221, pp. 350–360, Feb. 2020.

[91] J. N. Kather *et al.*, "Pan-cancer image-based detection of clinically actionable genetic alterations," *Nature Cancer*, vol. 1, no. 8, pp. 789–799, Aug. 2020.

[92] A. Binder *et al.*, "Morphological and molecular breast cancer profiling through explainable machine learning," *Nature Machine Intelligence*, vol. 3, no. 4, pp. 355–366, Mar. 2021.

[93] B. Schmauch *et al.*, "A deep learning model to predict RNA-Seq expression of tumours from whole slide images," *Nat. Commun.*, vol. 11, no. 1, p. 3877, Aug. 2020.

[94] Y. Fu *et al.*, "Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis," *Nature Cancer*, vol. 1, no. 8, pp. 800–810, Aug. 2020.

[95] K. Kartasalo, L. Latonen, J. Vihinen, T. Visakorpi, M. Nykter, and P. Ruusuvuori, "Comparative analysis of tissue reconstruction algorithms for 3D histology," *Bioinformatics*, vol. 34, no. 17, pp. 3013–3021, Sep. 2018.

[96] Y. Song, D. Treanor, A. J. Bulpitt, and D. R. Magee, "3D reconstruction of multiple stained histology images," *J. Pathol. Inform.*, vol. 4, no. Suppl, p. S7, Mar. 2013.

[97] Z. Huang *et al.*, "Artificial intelligence reveals features associated with breast cancer neoadjuvant chemotherapy responses from multi-stain histopathologic images," *NPJ Precis Oncol*, vol. 7, no. 1, p. 14, Jan. 2023.

[98] H. Duanmu *et al.*, "A spatial attention guided deep learning system for prediction of pathological complete response using breast cancer histopathology images," *Bioinformatics*, vol. 38, no. 19, pp. 4605–4612, Sep. 2022.

[99] A. Su *et al.*, "A deep learning model for molecular label transfer that enables cancer cell identification from histopathology images," *NPJ Precis Oncol*, vol. 6, no. 1, p. 14, Mar. 2022.

[100] R. Turkki, N. Linder, P. E. Kovanen, T. Pellinen, and J. Lundin, "Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples," *J. Pathol. Inform.*, vol. 7, p. 38, Sep. 2016.

[101] M. Valkonen *et al.*, "Cytokeratin-Supervised Deep Learning for Automatic Recognition of Epithelial Cells in Breast Cancers Stained for ER, PR, and Ki-67," *IEEE Trans. Med. Imaging*, vol. 39, no. 2, pp. 534–542, Feb. 2020.

[102] E. A. Burlingame *et al.*, "SHIFT: speedy histological-to-immunofluorescent translation of a tumor signature enabled by deep learning," *Sci. Rep.*, vol. 10, no. 1, p. 17507, Oct. 2020.

[103] H. Wieslander, A. Gupta, E. Bergman, E. Hallström, and P. J. Harrison, "Learning to see colours: Biologically relevant virtual staining for adipocyte cell images," *PLoS One*, vol. 16, no. 10, p. e0258546, Oct. 2021.

[104] K. de Haan *et al.*, "Deep learning-based transformation of H&E stained tissues into special stains," *Nat. Commun.*, vol. 12, no. 1, p. 4884, Aug. 2021.

[105] U. Khan, S. Koivukoski, M. Valkonen, L. Latonen, and P. Ruusuvuori, "The effect of neural network architecture on virtual H&E staining: Systematic assessment of histological feasibility," *PATTER*, vol. 0, no. 0, Apr. 2023.

[106] D. Schapiro *et al.*, "MCMICRO: a scalable, modular image-processing pipeline for multiplexed tissue imaging," *Nat. Methods*, vol. 19, no. 3, pp. 311–315, Mar. 2022.

[107] J.-R. Lin *et al.*, "Multiplexed 3D atlas of state transitions and immune interaction in colorectal cancer," *Cell*, vol. 186, no. 2, pp. 363–381.e19, Jan. 2023.

[108] J. Borovec *et al.*, "ANHIR: Automatic Non-Rigid Histological Image Registration Challenge," *IEEE Trans. Med. Imaging*, vol. 39, no. 10, pp. 3042–3052, Oct. 2020.

[109] J. S. Reis-Filho and J. N. Kather, "Overcoming the challenges to implementation of artificial intelligence in pathology," *J. Natl. Cancer Inst.*, vol. 115, no. 6, pp. 608–612, Jun. 2023.

[110] J. A. A. Snoek, I. D. Nagtegaal, S. Siesling, E. van den Broek, H. J. van Slooten, and N. Hugen, "The impact of standardized structured reporting of pathology reports for breast cancer care," *Breast*, vol. 66, pp. 178–182, Dec. 2022.