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

ARTIFICIAL INTELLIGENCE FOR BREAST CANCER PRECISION PATHOLOGY

Yinxi Wang

Karolinska Institutet

Stockholm 2022

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2022

ISBN 978-91-8016-845-8
Cover illustration: Yinxi Wang

Artificial intelligence for breast cancer precision pathology

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Yinxi Wang

The thesis will be defended in public at lecture hall Atrium, Nobels väg 12 B, Karolinska Institutet, Solna, at 1:00 PM on December 2nd, 2022

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

Dr. Johan Lindberg
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Opponent:
Professor Aristotelis Tsirigos
New York University School of Medicine
Department of Pathology

Examination Board:
Professor Johan Lundin
Karolinska Institutet
Department of Department of Global Public Health

Professor Johan Trygg
Umeå University
Department of Chemistry

Dr. Barbro Linderholm
University of Gothenburg
Department of Oncology

To my beloved family

致我亲爱的家人们

ABSTRACT

Breast cancer is the most common cancer type in women globally but is associated with a continuous decline in mortality rates. The improved prognosis can be partially attributed to effective treatments developed for subgroups of patients. However, nowadays, it remains challenging to optimise treatment plans for each individual. To improve disease outcome and to decrease the burden associated with unnecessary treatment and adverse drug effects, the current thesis aimed to develop artificial intelligence based tools to improve individualised medicine for breast cancer patients.

In **study I**, we developed a deep learning based model (DeepGrade) to stratify patients that were associated with intermediate risks. The model was optimised with haematoxylin and eosin (HE) stained whole slide images (WSIs) with grade 1 and 3 tumours and applied to stratify grade 2 tumours into grade 1-like (DG2-low) and grade 3-like (DG2-high) subgroups. The efficacy of the DeepGrade model was validated using recurrence free survival where the dichotomised groups exhibited an adjusted hazard ratio (HR) of 2.94 (95% confidence interval [CI] 1.24-6.97, $P = 0.015$). The observation was further confirmed in the external test cohort with an adjusted HR of 1.91 (95% CI: 1.11-3.29, $P = 0.019$).

In **study II**, we investigated whether deep learning models were capable of predicting gene expression levels using the morphological patterns from tumours. We optimised convolutional neural networks (CNNs) to predict mRNA expression for 17,695 genes using HE stained WSIs from the training set. An initial evaluation on the validation set showed that a significant correlation between the RNA-seq measurements and model predictions was observed for 52.75% of the genes. The models were further tested in the internal and external test sets. Besides, we compared the model's efficacy in predicting RNA-seq based proliferation scores. Lastly, the ability of capturing spatial gene expression variations for the optimised CNNs was evaluated and confirmed using spatial transcriptomics profiling.

In **study III**, we investigated the relationship between intra-tumour gene expression heterogeneity and patient survival outcomes. Deep learning models optimised from **study II** were applied to generate spatial gene expression predictions for the PAM50 gene panel. A set of 11 texture based features and one slide average gene expression feature per gene were extracted as input to train a Cox proportional hazards regression model with elastic net regularisation to predict patient risk of recurrence. Through nested cross-validation, the model dichotomised the training cohort into low and high risk groups with an adjusted HR of 2.1 (95% CI: 1.30-3.30, $P = 0.002$). The model was further validated on two external cohorts.

In **study IV**, we investigated the agreement between the Stratipath Breast, which is the modified, commercialised DeepGrade model developed in **study I**, and the Prosigna® test. Both tests sought to stratify patients with distinct prognosis. The outputs from Stratipath Breast comprise a risk score and a two-level risk stratification whereas the outputs from Prosigna® include the risk of recurrence score and a three-tier risk stratification. By comparing the number of patients assigned to ‘low’ or ‘high’ risk groups, we found an overall moderate agreement

(76.09%) between the two tests. Besides, the risk scores by two tests also revealed a good correlation (Spearman's rho = 0.59, P = 1.16E-08). In addition, a good correlation was observed between the risk score from each test and the Ki67 index. The comparison was also carried out in the subgroup of patients with grade 2 tumours where similar but slightly dropped correlations were found.

LIST OF SCIENTIFIC PAPERS

* Equal contribution.

I. **Wang Y**, Acs B, Robertson S, Liu B, Solorzano L, Wählby C, Hartman J, Rantalainen M. Improved breast cancer histological grading using deep learning. Annals of Oncology. 2022 Jan 1;33(1):89-98.

II. **Wang Y***, Kartasalo K*, Weitz P, Ács B, Valkonen M, Larsson C, Ruusuvuori P, Hartman J, Rantalainen M. Predicting Molecular Phenotypes from Histopathology Images: A Transcriptome-Wide Expression–Morphology Analysis in Breast Cancer. Cancer Research. 2021 Oct 1;81(19):5115-26.

III. **Wang Y**, Ali MA, Humphreys K, Hartman J, Rantalainen M. Transcriptional intra-tumour heterogeneity predicted by deep learning in routine breast histopathology slides provides independent prognostic information. *Manuscript*.

IV. **Wang Y***, Robertson S*, Karlsson E, Rantalainen M, Hartman J. Evaluation of the concordance in breast cancer risk stratification between a commercialised deep learning tool and the Prosigna® test. *Manuscript*.

CONTENTS

LIST OF ABBREVIATIONS

AIs	aromatase inhibitors
ASCO	American Society of Clinical Oncology
AUC	area under the receiver operating characteristic curve
BC	breast carcinomas
BH	Benjamini-Hochberg
CE	cross-entropy
Clinseq	Clinical Sequencing of Cancer in Sweden
CNB	core needle biopsy
CNN	convolutional neural network
CT	computed tomography
DCIS	ductal carcinoma *in situ*
DG	DeepGrade
EMO	expression-morphology
ER	estrogen receptor
ESMO	European Society for Medical Oncology
FDR	false discovery rate
FFPE	formalin-fixed paraffin-embedded
FNAB	fine-needle aspiration biopsy
FWER	family-wise error rate
GDPR	General Data Protection Regulation
HE	haematoxylin and eosin
HER2	human epidermal growth factor receptor 2
HR	hazard ratio
HSV	hue, saturation, and value
IHC	immunohistochemistry
INCA	Information Network for Cancer Care
ISH	*in situ* hybridization
KM	Kaplan-Meier
LASSO	least absolute shrinkage and selection operator
LCIS	lobular carcinoma *in situ*
LME	linear mixed effect

MLP	multi-layer perceptron
MRI	magnetic resonance imaging
MSE	mean square error
NHG	Nottingham histologic grade
NKBC	national quality registry for breast cancer
OD	optical density
OOD	out of distribution
PARP	Poly(ADP-ribose) polymerase
PD-L1	programmed death-ligand 1
PR	progesterone receptor
RGB	red, green and blue
ROC	receiver operating characteristic
ROR	risk of recurrence
RT-PCR	reverse transcription polymerase chain reaction
SCAN-B	Sweden Cancerome Analysis Network – Breast
SGD	stochastic gradient descent
ST	spatial transcriptomics
SVD	singular value decomposition
TCGA	The Cancer Genome Atlas
TDLU	terminal duct lobular unit
TNBC	triple-negative breast cancer
WSI	whole slide image

1 INTRODUCTION

The breast is an organ with a glandular structure. There are three major components inside a female breast, namely adipose tissue, connective tissue and glandular tissue. The glandular structure is also known as the mammary gland. Within the mammary gland, there are on average 15-20 lobes, each divides into branches of smaller lobules that comprise of clusters of ductules (also known as alveolar buds in organs undergoing differentiation or acinus in morphologically mature organs (1)) and intralobular terminal ducts. Together the lobules and extralobular terminal ducts form the major functional unit called 'terminal duct lobular unit' (TDLU) that produces milk during lactation.

Breast cancer, or breast carcinoma (BC) is a malignant disease in the breast that mainly arises from the TDLU. Carcinomas originating from the ducts are called ductal carcinomas whereas those originating from the lobules are named as lobular carcinomas. Breast cancer is considered as a heterogeneous disease that can be grouped from various levels based on molecular profiles, morphological patterns or degrees of progression. To date, pathological examination is the gold standard for diagnosing breast cancer and for guiding treatment plans. However, the assessment has relatively large inter- and intra-observer variations (2–4), making it difficult to provide optimal clinical decision support. In addition, the heterogeneous nature of breast cancer can result in treatment resistance, causing long term recurrence and metastases (5)(6). Therefore, improving the accuracy of cancer diagnosis and deepening our knowledge in cancer subtyping is vital to ultimately tailor treatment regimens for each patient.

Deep learning is a subfield of machine learning that has evolved rapidly within recent years. A strength of deep learning models is the ability to automatically learn representations that are associated with learning objectives (7). This advantage has enabled the possibility of analysing large volumes of image data without the need for manual feature engineering. In the healthcare domain, a considerable number of studies have been carried out to facilitate disease diagnosis and precision medicine by analysing medical data such as digitised histopathology WSIs with deep learning models.

In the field of breast pathology, molecular profiling is available to assist with characterising tumours, estimating patient prognosis or predicting treatment efficacy. With the advent of deep neural networks, novel opportunities are now unveiled to strengthen precision medicine; in addition, these cost-effective solutions exhibit strong potential to reduce the workload for healthcare professionals and to relieve health economic burden globally.

2 LITERATURE REVIEW

2.1 BREAST CANCER

2.1.1 Breast cancer epidemiology

Female breast cancer is the most common cancer type globally, with over 2.2 million new diagnoses, accounting for 24.5% of all new cancers in 2020 in women (8). Besides, it has the highest age-standardised incidence rate of 47.8 per 100,000 person-years among cancers from all sites. In Sweden, an increasing trend in breast cancer incidence has been observed since the 1970s, with age adjusted rate ranging from 200.6 to 211.6 per 100,000 person-years within the year 2017-2019; the number dropped significantly in the year 2020, with age adjusted incidence rate 195.2 per 100,000 person-years and 7570 newly diagnosed cases, which could be partially explained by temporarily ceased screening during the COVID-19 pandemic (9). In contrast to this high incidence rate is a gradually declining mortality rate, with an age-standardised mortality rate of 13.6 per 100,000 person-years worldwide in 2020，accounting for the second largest cause of cancer death aside from lung cancer. In Sweden, the mortality rate slightly declined over the past decades, and the estimated age-standardised mortality rate was 26.9 per 100,000 person-years in 2020. Possible explanations behind the improved patient prognosis include the implementation of screening programs and recent achievements in targeted therapies.

The risk factors associated with BC can be divided into genetic factors and non-genetic factors. Genetic factors include germline mutations (i.e. hereditary BC characterised by deleterious mutations in anti-oncogenes *BRCA1* or *BRCA2*), african-american ethnicity, and dense breast whereas the most well known non-genetic causes include hormonal risk factors such as low parity, late age at first birth, early menarche or late menopause (10) and environmental risk factors, such as alcohol consumption, tobacco smoking and a high-fat diet.

2.1.2 Breast cancer screening and diagnosis

Between the 1880s and 1990s, the evolution of imaging techniques prompted the wide adoption of mammography screening that contributed to earlier diagnosis of breast cancer and an effectively declined mortality rate ranging from 21% to 39% (11,12).

Since 1997, the nationwide mammography screening for breast cancer was fully introduced and the National Board of Health and Welfare of Sweden recommends women between 40 and 74 to participate in screening. In addition to mammography, magnetic resonance imaging (MRI) also serves as an adjunctive screening method.

Suspected breast cancer needs to be confirmed pathologically with biopsies. The tissue sampling is typically performed by fine-needle aspiration biopsy (FNAB) cytology and core needle biopsy (CNB) technique (13).

FNAB is fast and minimally invasive but has limitations such as the inability in separating carcinoma *in situ* with invasive carcinoma, or providing morphological information that is important in planning preoperative treatment.

CNB has the advantage of preserving morphological patterns. The tissue is typically formalin fixed and paraffin embedded (FFPE), subsequently sectioned and stained with HE or immunohistochemistry (IHC). Staining helps to reveal information such as tumour grade or the expression of biomarkers through pathological examination. A good grasp of tumour characteristics helps to predict potential benefits of certain therapeutic choices. CNB is therefore mandatory when neoadjuvant therapy is applicable in order to assist with treatment selections.

2.1.3 Histopathological examination of breast cancer

Once the cancer is removed during surgery, surgical resected specimens are subsequently sliced across the frontal or sagittal plane. A proportion of the cut slices are prepared into fresh frozen materials and are typically used in molecular analysis. Other slices are processed to FFPE tissue sections for in-depth pathological assessment.

FFPE can preserve tissue structures for years, allowing the materials to be archived or used in retrospective research studies over the long term. After FFPE fixation, tissue slices are placed on a glass, and typically stained with two major types of methods for different diagnostic purposes. For the purpose of evaluating morphological patterns and tumour aggressiveness, HE staining is applied. Haematoxylin stains cell nuclei into blue-purple colour, whereas eosin stains extracellular matrix as well as cytoplasm into pink colour. Distinct colours allow clear separation of microscopic features that reveals essential information such as tumour's growth patterns or grade. Another purpose is to study the expression level of key proteins, which are biomarkers that reveal crucial characteristics of tumours and serve as important indicators for therapeutic strategies (more detail is introduced in **Section 2.1.7**). IHC staining is used for this objective. Antibodies in the solution can bind specifically to targeting antigens (the protein under analysis) within the tissue, a kind of reporter molecule is then recruited or bound to the antibody that is either labelled with a chromogenic substrate or fluorescent dye and yields respective colours. Whether or to what extent does a tumour exhibit certain types of biomarkers can thus be assessed by quantifying the abundance of positively stained cells.

Histopathological examination is considered the gold standard for cancer diagnosis, and tumour characteristics obtained via pathological assessments are pivotal in planning treatment and subsequent management options.

2.1.4 Breast cancer subtypes

BC are malignancies developed from epithelial cells. Due to its heterogeneous nature, BC can be divided into subtypes from various angles. For instance, when the growth of cancer cells are restricted in milk ducts or TDLUs, the lesion is called ductal carcinoma *in situ* (DCIS) or lobular carcinoma *in situ* (LCIS); in contrast, when the proliferation grows beyond the ductal

wall and starts to penetrate normal surrounding tissue, the tumour is considered invasive carcinoma; correspondingly, depending on the location where the cancer originates, there are invasive ductal carcinomas and invasive lobular carcinomas; molecular techniques such as gene expression profiling provides insights into the intrinsic differences between BC, and identified several molecular subtypes with distinct clinical behaviours and prognosis; clinically, some of these subtypes can be determined by IHC surrogates (14). In addition, standardised grading (**Section 2.1.5**) and staging systems (**Section 2.1.6**) have also been proposed to divide BC into subgroups with different degrees of aggressiveness and progression.

2.1.5 Breast cancer grading

Two separate grading systems are adopted to evaluate the degree of differentiation for DCIS and invasive breast cancer individually. Both require pathological examination of tumour morphological patterns from HE stained FFPE specimens. According to the European guidelines, DCIS are graded as low, medium or high using an integration of nuclear and structural features. Although widely implemented, the system lacks consistency in defining each grade, resulting in a moderate reproducibility for DCIS grading and hence highlighting a need to refine the current system with unified standards (15).

Since cancer *in situ* does not necessarily predict malignancy and therefore typically remains untreated, the grading of invasive cancer plays a more central role in understanding tumour progression and establishing treatment plans. Existing recommendations suggest grading BC with the Nottingham (Elston-Ellis) (16) modification of the Scarff-Bloom-Richardson grading system.

The grading criteria include three levels of morphological changes: mitotic count, nuclear pleomorphism and tubular formation. Each subcomponent is to be scored from 1 to 3 (**Table 1**). Mitotic count is defined as the number of mitosis within 10 high power fields. In general, the scoring shall be performed in the most proliferative regions demonstrated by highest mitotic density such as the invasive front of a tumour. In case of a heterogeneous tumour where mitotic figures exhibit a distinct regional variation, carefully identified hotspots are especially important and it is recommended to score in areas with least differentiation (17).

The evaluation of nuclear pleomorphism is also recommended in peripheral regions or less differentiated areas, and the scoring is based on the extent of nuclear atypia by comparing the size, shape, vesicularity and the presence of nucleoli with normal breast epithelium. This subcomponent often exhibits less intra-observer variability compared to mitosis counts (18).

The examination of tubular formation is performed by assessing the entire region of a WSI with a low power field. Tubular structures are defined by clear central lumens with polarised surrounding cancer cells. The scoring is based on the proportion of identified structures composing a tumour.

The overall histological grade (Nottingham histologic grade, NHG) is calculated by summing up scores from the above three components, which also ranges from grade 1 to 3 (**Figure 1**). It

is a well-established independent prognostic factor in previous studies (19),(20),(21). Tumours with lower grade are typically associated with a lower risk of disease recurrence and are therefore often treated conservatively; in contrast, tumours with higher grade grow more aggressively, and may be benefited from adjuvant chemotherapy.

The estimated ratio for these three grades are 2:3:5 (16), however, huge discrepancies have been observed with grade 2 tumours of up to 62% (22), implying noticeable inconsistencies among observers (23),(24), which is more common on judgements between NHG 1 and 2 tumours (25). It is worth noting that although the histological grades are assigned as discrete values, the corresponding morphological changes are associated with a continuous spectrum of differentiation, hence, cancers with grade 1 or 3 lesions normally exhibit more homogeneous growth patterns whereas grade 2 tumours often pertain to heterogeneous characteristics that can be reflected from both molecular profiles (26) and morphological features (27). A comprehensive and in-depth understanding of diagnosis warrants appropriate treatment, it is therefore more challenging to select optimal treatment regimens for patients with grade 2 tumours. Aside from the substantial variance in tumour appearance, the subjectivity in manual assessments further poses challenges towards a good adherence to agreed protocols. Taken together, these phenomena emphasise the need in developing quantitative methods that increase the reproducibility of BC grading.

Table 1. Breast cancer grading criteria by subcomponents.

Score	Tubular formation	Nuclear poleomorphism	mitotic count [a] (field diameter 0.40 mm)
1	Tubular structures can be found in >75% of the entire tumour	Nuclei appears with <1.5 times larger in size compared with normal epithelium, only minor variation in shape can be observed	<=4
2	Tubular structures can be found in 10–75% of the entire tumour	Nuclei is 1.5–2 times larger than normal epithelium, with visible vesicular, nucleoli, and moderate variation in shape	5 to 9
3	Tubular structures can be found in <10% of the entire tumour	Vesicular nuclei is > 2 times larger in size, with prominent nucleoli and distinct variation in shape	>=10

a. The threshold varies with different choices of field diameter

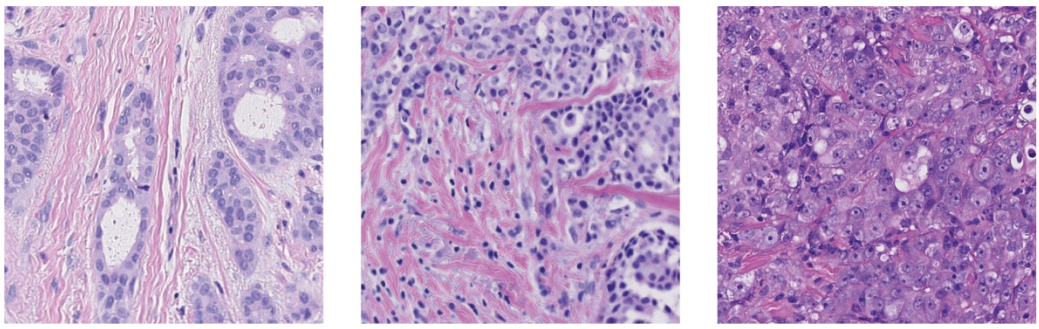

Figure 1. Examples of grade 1, 2 and 3 breast tumours. Left, NHG 1 tumour with clear tubular structures. The polarised cells surrounding each lumen are preserved with regular shapes and sizes with non-discernable nucleoli. Middle, NHG 2 tumour. The normal lumens are not observable, tumour cells start to grow in varied shapes and sizes, nucleoli become discernible. Right, NHG 3 tumour. Tubular structure is missing, and tumour cells exhibit bizarre forms. Images were taken from regions of HE stained FFPE tissue sections, under the 20X magnification.

2.1.6 The TNM staging system

The TNM staging system evaluates the growth and spread behaviour of tumours, and consists of three aspects: tumour size (T), lymph node metastasis (N) and distance metastasis (M).

Tumour size is measured by the longest diameter of the entire tumour. It has been indicated by many studies as a powerful predictor for patient prognosis (28),(29). However, studies showed that for the basal-like subtype, tumour size no longer carries significant prognostic value (30), especially with respect to long-term survival (31).

The number of axillary lymph node metastasis is another important factor relating to patient outcome. With effective detection of early stage BC, total axillary dissection is no longer favoured as the large proportion of patients tend to have clear axillary lymph nodes (32); this has urged the need for alternative examination of the sentinel lymph nodes. Sentinel lymph nodes are the first to receive lymphatic fluid from the cancer site, thus, cancer cells are likely to be detected first from these nodes (33) and in theory, negative findings in sentinel lymph nodes predict negative axillary lymph nodes. It was also confirmed from a study that the number of cancer cells in the sentinel lymph node is in high concordance (96.8%) with the status of axillary lymph nodes (34). Hence, nowadays, a primary assessment of the sentinel lymph nodes is recommended, sparing the unnecessary axillary lymph node dissection and accompanying side effects (35). In case of positive findings that suggest cancer spread, axillary lymph node dissection is then performed to allow for exhaustive assessment (36). The number of positive nodes is also highly correlated with tumour size (37), but the relationship diminishes in basal-like tumours (30), indicating evident dissimilarities between these subtypes.

Distant metastasis is the third component of the TNM system. M0 denotes no observed distant metastasis in other organs, BC with this stage accounts for more than 95% of cancer cases in different populations (38,39),(40); M1 indicates the presence of metastatic breast cancer and is an indication of worse prognosis (41).

2.1.7 Breast cancer biomarkers

The expression level of several proteins is also required to characterise breast tumours. These biomarkers include estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and Ki67. The assessment is carried out mainly with IHC staining, and sometimes in conjunction with *in situ* hybridization (ISH) to confirm the abundance of HER2.

ER consists of two types, ER-α,and ER-β, both are nuclear transcription factors. The major type that is expressed in mammary glands is ER-α, encoded by *ESR1*. ER-α acts to regulate cell proliferation and differentiation and typically has an abundant expression in early stage breast cancers, whereas ER-β exerts an anti-proliferative effect with only low or no expression in both early stage and advanced breast cancers (42). Therefore, the expression of ER-α is a vital indicator for the applicability of endocrine therapy. Clinical evaluation of ER status relies on IHC staining and counting the percentage of positively stained nuclei among cells within the invasive cancer regions. If no cancer cells are immunoreactive, it is considered an ER negative tumour, otherwise positive. However, the cutoff in determining ER positivity varies across sites, and a tumour with 1%-10% positive cancer cells is typically considered as ER low positive (43). Due to limited understanding towards biological behaviours and prognosis of ER low positive tumours, the planning of personalised therapy for this subgroup often requires additional information such as the intensity level and the existence of other biomarkers. In comparison, the cutoff of 10% for defining ER positivity from Swedish guidelines is less ambiguous and also commonly chosen in research (44),(45). ER positive cases account for up to 84% of all BCs (46), and this subgroup is typically associated with better prognosis compared with ER negative groups for women diagnosed after age 40 (47).

PR is encoded by the gene *PGR*. In BC, PR has been recognised to regulate expression of target genes that mediate cell proliferation (48) but whether the existence of PR is predictive to treatment outcomes remains unclear (49)(50). In Sweden, PR status is also determined via IHC staining with the same threshold of 10% (51).

HER2 is encoded by *ERBB2*, and its expression is primarily evaluated by IHC. In cases where no immunoreactive cells or more than 10% of the epithelial cells exhibit incomplete membrane staining, the tumour is regarded as HER2 negative; in comparison, a tumour is defined as HER2 positive if more than 10% of cells are completely and intensively stained positive. For borderline cases with only weak to moderate stains in >10% of cells or with =<10% of cells with complete staining, the tumour is further examined with the ISH test. HER2 positive cases account for 9%-16% of all BC and a positive biomarker status is an indicator of worse prognosis in node-positive patients (4,52).

Tumours devoid of ER, PR and HER2 are referred to as triple-negative breast cancer (TNBC). The subgroup constitutes around 10-20% breast cancers (53) and exhibits worse prognosis.

Ki67 is a crucial biomarker reflecting cell proliferation activity and is encoded by the *MKI67* gene. It is examined by IHC assay and reported as the proportion of immunoreactive cells with

a range between 0 to 100%. Due to recognised discrepancies in preparation and calibration procedures, there are no globally applied cutoffs for Ki67, instead, the threshold is defined individually across sites and regions. Apart from the lack of unified cutoffs, there's no consensus regarding how the Ki67 shall be quantified either. In Sweden, the guideline used to suggest scoring Ki67 by counting 200 cells within a hotspot region, but the latest update suggests to perform global scoring (54). The insufficient consensus on scoring protocols and considerable inter-observer variability (55) have largely restricted the clinical utility of Ki67. As a consequence, although it has been reported that the Ki67 index correlates with tumour aggressiveness and predicts patient prognosis (56), no consensus is reached regarding its clinical value.

2.1.8 Breast cancer treatment

A careful and comprehensive examination of tumour grade, stage, subtype and the status of biomarkers make up the basic workflow in clinical decision making.

For primary early-stage tumours, surgical removal is typically the main type of treatment and is performed with either breast-conserving surgery or mastectomy.

Postoperative adjuvant radiotherapy is another standard treatment modality for primary tumours to kill cancer cells in the remaining cancer region by targeting the cells' DNA. It is also an effective palliative therapy for metastatic breast tumours (57).

After surgery, systemic treatment such as chemotherapy is often prescribed to eradicate remaining cancer cells, facilitating prolonged survival. The current guideline suggests that the chemotherapy shall be administered for HER2-positive, triple-negative and luminal B-like tumours (58). Due to its unpleasant short-term side effects such as causing hair loss, vomiting, diarrhoea and peripheral neuropathy, as well as the considerable long-term side effects that increase the risk of heart failure, mental dysfunction or leukaemia (57), the administration of chemotherapy shall be motivated with efficacy outweighing the accompanied risks. It is therefore also of special importance to seek for less toxic alternatives, such as targeted therapies for subgroups of BC. In addition, novel molecular or imaging diagnostic tools that are capable of assisting with de-escalation or escalation of chemotherapy are also desirable elements constituting personalised modern BC management.

Endocrine therapy is an alternative systemic therapy for hormone receptor positive patients. The mainstream treatments consist of two subtypes, one is Tamoxifen that functions as an ER antagonist to block the effect of estrogen on ER positive cells; The other is aromatase inhibitors (AIs) that function by inhibiting the synthesis of estrogen. Co-administration of adjuvant chemotherapy and Tamoxifen is also used in treating ER positive patients, to further reduce the risk of treatment resistance (59).

Targeted therapy is another class of treatments that works by targeting specific molecules or pathways. For instance, Trastuzumab, which targets the HER2-protein, is applied to treat HER2 positive patients. Poly(ADP-ribose) polymerase (PARP) inhibitors that induce apoptosis

among cancer cells with *BRCA1* or *BRCA2* mutation (60) are currently approved to treat BRCA mutated breast cancers (61). Furthermore, immune checkpoint inhibitors such as the inhibitor for the programmed death-ligand 1 (PD-L1), atezolizumab, can bind to PD-L1 and inhibit its interaction with PD-1 and B7-1, enhancing the T cell-induced cytotoxicity (62,63). The concurrent administration of nab-paclitaxel and atezolizumab has also demonstrated favourable survival outcomes (63) and has been used for treating PD-L1 positive TNBC patients (64).

Preoperative systemic therapy, regardless of the type of the choices that are covered above, is referred to as neoadjuvant therapy, and is recommended in the presence of advanced tumours without metastasis or inoperable inflammatory tumours (65) to shrink the tumour size. The pathological response to neoadjuvant therapy not only enables less invasive surgical strategies but also serves as a critical indicator on how tumours react to therapeutic agents, hence, providing a unique prognostic value to further guide subsequent treatment decisions.

2.1.9 Prognostic markers

Modern techniques have assisted with the development of novel prognostic factors that facilitate refined BC subtyping and individualised clinical decision making. One widely adopted assay is the Prosigna® test (NanoString Technologies, Seattle, USA) which was developed based on the expression profiling of 50 genes (PAM50)(66) and works to classify BC into four intrinsic subtypes, namely Luminal-A, Luminal-B, HER2-enriched and basal-like. In Sweden, the test is recommended to assess the recurrence risk for patients that are postmenopausal, with ER-positive, HER2-negative and node-negative breast cancer, in an attempt to determine the applicability of adjuvant chemotherapy (67). Oncotype DX® (Exact Sciences Corp., Madison, USA) is a 21-gene signature assay that also computes a risk of recurrence score, besides, the test provides insights to potential benefit of adjuvant chemotherapy (68). Aside from global classification models, improved stratification of NHG 2 tumours have received special attention due to the noticeable heterogeneous morphology and biological behaviour among cancers in this subgroup. Using RNA sequencing, a molecular tool developed by Wang et.al integrated 34 genes and dichotomised NHG 2 tumours into two levels with distinct recurrence free survival (69). Albeit a plethora of subtyping criteria that have been introduced in practice, it is widely acknowledged that the development of malignancy as a continuous and heterogeneous process poses a challenge to effective personalised medicine, highlighting the need for novel approaches that improve tumour stratification.

2.1.10 Spatial heterogeneity

Spatial intra-tumour heterogeneity is a reflection of cell-subpopulations that carry distinct molecular phenotypes. The dynamic yet closely interlinked gene expression can be regulated at different levels and stages by events such as genetic mutation, epigenetic changes, alterations in mRNA synthesis, processing and translation. These factors give rise to clonal heterogeneity and variations in expression pattern within a tumour. Such spatial intra-tumor heterogeneity

has been regarded as one major reason for treatment resistance (70) and a prognostic factor in various types of cancers (71)(72)(73).

Spatial transcriptomics (ST) accounts for a major genre of techniques for quantifying gene expression while preserving spatial information. There are three ways to recover spatial coordinates, by using ligate oligonucleotide barcodes that are attached to RNA molecules, by *in situ* hybridization or by *in situ* sequencing (74). Leveraging these techniques enables an in-depth understanding for depicting the evolutionary pattern of tumour cells. It provides insights into how tumour subpopulations started from genomic subclones within the *in situ* cancer obtain invasiveness and grow into neighbouring regions (75). In addition, the spatial expression proximity can also be exploited to study gene-gene or cell-cell interactions, making it possible to picture a transcriptome-wide network of cellular communications (76). Enriched information acquired with spatial constraints effectively facilitates research into the tumour microenvironment (77)(78).

Deep learning models have also been used to predict molecular phenotypes from medical image data such as HE stained WSIs (79,80)(81)(82), constituting a novel alternative to spatial transcriptomics (83)(84). Although intra-tumour heterogeneity and its consequence in patient prognosis remains to be an intricate problem, modern technologies have brought about a wealth of technical solutions that foster novel research opportunities in this domain.

2.2 COMPUTATIONAL PATHOLOGY

The past twenty years have witnessed a rapid evolution of scanning technology that enabled the speedy digitisation of WSIs with a very high resolution. Besides, the accompanying development in image archiving and communication systems further strengthen the image management process (85). Together, these advances in computer and electronic technology have prepared modern pathology to be gradually migrated from traditional microscope-based examination to digital-based practice where pathologists can perform the assessment in front of the computer. The digitisation facilitates efficient remote image transfer and review not only for diagnosis but also for research or educational purposes. Moreover, it spawned a novel research field: computational pathology.

The definition of computational pathology is not necessarily confined with computer-aided pathologic diagnostic tools, rather, it includes the entire framework that comprises experimental design, image analysis and statistical and computational modelling to address scientific problems that can be answered with pathological data (86). WSIs, either stained with, for example, IHC or HE serve as the main image modality that is analysed in computational pathology, and the analyses often include image processing, segmentation, object detection or identification, and prediction of histopathology characteristics as well as clinical outcomes. A common goal shared among research in this field can be summarised as to deliver decision support for pathologists and other clinicians with respect to, for example, pathological diagnosis or patient risk stratification.

The quantitative modelling approaches have the favourable property to not only potentially reduce inter- or intra-observation variabilities, but also to extract clinically relevant information beyond current routine pathology. Moreover, with the abundance of multi-'omics' data and other imaging modalities such as computed tomography (CT) and MRI, it is believed that computational pathology can serve to bridge the information gap, providing a more comprehensive understanding to individual tumours for improved precision medicine (87).

2.3 MACHINE LEARNING

While the term 'artificial intelligence' covers a broad concept of disciplines where human intelligence can be formalised and learned by machines, 'machine learning' as one of its subfields focuses mainly on training models with training data sets and generating accurate predictions on unseen test data (88). Many research topics in the medical image analysis domain fall within the scope of supervised learning, where the training data comprises images that serve as input together with their labels as ground truth. The model optimisation process therefore aims to learn the model parameters from the input feature representations and minimise the prediction error in comparison with the ground truth targets.

Before the explosion of scientific interest in deep learning, traditional machine learning techniques that utilise hand-crafted features dominated the domain of medical image analysis (89)(90). Later, the introduction of deep learning models, in particular CNNs, aided in overcoming the necessity for manual feature selection and has yielded superior performance in various image analysis contexts (91).

2.3.1 Artificial neural network

The development of the perceptron marks a milestone in the history of deep learning. It was mainly inspired by two findings, one was the McCulloch-Pitts Neuron (92) in mimicking the biological activation of a neuron that triggers an impulse to connected neurons when a given stimulus surpasses a threshold; the other was the theory in learning, proposed by Donald Hebb (93), which stated that the connection between neurons increases if one continuously participates in activating the other.

A perceptron equation with a vector of N inputs (**x**) takes the following form:

$$y = f\left(\sum_{j=1}^{N} w_i x_i + \theta\right) \qquad (2\text{-}1)$$

Where **w** is the vector of coefficients. For a binary classification problem, if the sum of the linear combinations of inputs with coefficients has a higher value than the threshold, θ, the function $f(\cdot)$ outputs 1, otherwise 0. The basic structure can only be applied for linear separable scenarios, which largely limited its use in advanced and complex requirements. The problem was solved later with the introduction of the multi-layer perceptron (MLP) (94).

The MLP shares the basic structure of the perceptron but with more layers of neurons inserted between the input and output layer, as shown in **Figure 2**.

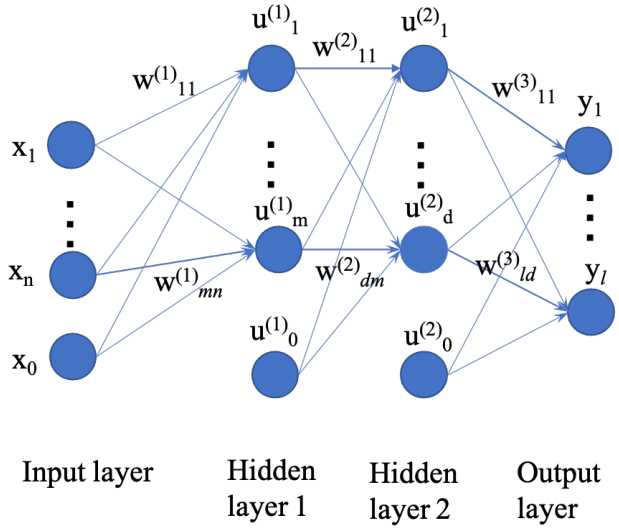

Figure 2. The architecture of a multi-layer perceptron with two hidden layers.

The model is a 3-layer neural network as there are three layers of weights that are adaptable during training and it contains two hidden layers.

Linear combination from the first layer is computed with the following formula where x_1, …, x_n are the inputs. Note that to illustrate the bias term, $x_0 = 1$ was added to the figure. The learnable parameters w_{ji} are often referred to as weights with w_{j0} as biases, where $j = 1 \ldots m$:

$$a_j^{(1)} = \sum_{i=1}^{N} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \qquad (2\text{-}2)$$

Next, $a^{(1)}_j$ go through an activation function $f(\cdot)$ to form the output of the j^{th} neuron in the hidden layer *1*, $u^{(1)}_j$, where $f(\cdot)$ is designed to be non-linear and differentiable:

$$u_j^{(1)} = f\left(a_j^{(1)}\right) \qquad (2\text{-}3)$$

In the same manner, $u^{(1)}_j$ contributes to compute the input for hidden layer 2. The output neurons from the hidden layer 2 provide an input $a^{(3)}_l$ for the last layer, where l denotes the number of neurons (classes) in the output layer.

Finally, an activation function is applied to each output in generating the final prediction such that:

$$y_l = \sigma\left(a_l^{(3)}\right) \qquad (2\text{-}4)$$

For regression objectives, the activation function is an identity function so that $y_l = a_l$, whereas for binary classification objectives, the activation function is either the sigmoid (2-5) or the softmax function (2-6) for binary or multiclass problem, respectively:

$$\sigma(a) = \frac{1}{1 + \exp(-a)} \quad (2\text{-}5)$$

$$\sigma(\boldsymbol{a})_i = \frac{\exp(a_i)}{\sum_{j=1}^{L} \exp(a_j)} \quad (2\text{-}6)$$

A deep neural network often contains several hidden layers, and by combining the above components together, the model can be expressed with the formula below and is regarded as the forward propagation of input information:

$$\hat{y}_l(\mathbf{x}, \mathbf{w}) = \sigma(\sum_d w_{ld}^{(L)} f(\sum_m w_{dm}^{(L-1)} f(\ldots f(\sum_i w_{ji}^{(1)} xi + \mathrm{w}_{j0}^{(1)}) + \ldots) + \mathrm{w}_{d0}^{(L-1)}) + \mathrm{w}_{l0}^{(L)}) \quad (2\text{-}7)$$

2.3.2 Error function

To compare model predictions with the ground truth, an error function (also known as loss function, or cost function) is designed to suit specific needs from different learning objectives. For regression models, given a set of N inputs $\mathbf{X}=\{x_1 \ldots x_n\}$ and associated continuous values $\mathbf{Y}=\{y_1 \ldots y_n\}$ with model prediction $\hat{\mathbf{Y}}$, the commonly used mean square error (MSE) loss can be used:

$$L_{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad (2\text{-}8)$$

Whereas for binary classification tasks, a cross-entropy (CE) loss is used where $y_i \in \{0, 1\}$, denoting the class label and p_i is the model prediction:

$$L_{CE} = -\sum_{i=1}^{N} \{y_i \log(p_i) + (1 - y_i) log(1 - p_i)\} \quad (2\text{-}9)$$

The function can be written in a generalised form that also applies to multiclass problems and here y_{ci} is a vector with length l, only the entry corresponding to the specific class label takes the value 1 and the rest are 0s:

$$L_{CE} = -\sum_{i=1}^{N} \sum_{c=1}^{L} y_{ci} \log(pc_i) \quad (2\text{-}10)$$

2.3.3 Gradient descent and backpropagation

For ANNs, training a model implies the procedure of feeding batches of inputs into the model and updating the weights in an iterative manner by minimising the errors within each batch. The most efficient way to achieve this is to take a small step towards the direction where the error decreases at its fastest speed, and this is accomplished by updating the weights to the opposite of the local gradient (derivative) with regard to the loss function: $-\nabla L(\mathbf{w})$, such that:

$$\mathbf{w}_{new} = \mathbf{w}_{old} - \eta \nabla L(\mathbf{w}_{old}) \quad (2\text{-}11)$$

Here, the η (*learning rate*) controls the strength of the updates and the optimisation strategy is called *gradient descent*.

The gradient of the error message is computed efficiently backward to each layer through a technique called *backpropagation* (95). Let L_n be the error associated with the n^{th} input. Suppose the aim is to update weight $w^{(2)}_{dm}$, given L_n, we hence need to compute the derivative of error function with respect to $w^{(2)}_{dm}$:

$$\partial L_n/(\partial w^{(2)}_{dm}) \qquad (2\text{-}12)$$

To process is illustrated in **Figure 3**:

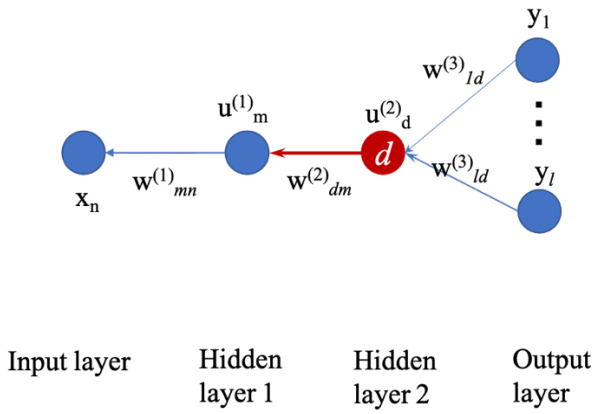

Figure 3. Error backpropagation process for a subset of the network in **Figure 2**.

It can be seen from the forward pass that the $w^{(2)}_{dm}$ contributes to the error through the calculation of $a^{(2)}_{nd}$. For simplicity, the subscription n is omitted from the equation. Based on the chain rule, the derivative can be decomposed into the following term:

$$\frac{\partial L_n}{\partial w^{(2)}_{dm}} = \frac{\partial L_n}{\partial a^{(2)}_d} \frac{\partial a^{(2)}_d}{\partial w^{(2)}_{dm}} \qquad (2\text{-}13)$$

And by defining:

$$\delta^{(2)}_d = \frac{\partial L_n}{\partial a^{(2)}_d} \qquad (2\text{-}14)$$

Equation *2-13* can be re-written as:

$$\frac{\partial L_n}{\partial w^{(2)}_{dm}} = \delta^{(2)}_d u^{(1)}_m \qquad (2\text{-}15)$$

In this way, the gradient with regard to a given weight can be expressed as the product of the error signal at the output end of that weight (indicated by the red neuron in **Figure 3**) and the input signal that connects to the other side of the weight (96).

The error $\delta^{(2)}_d$ can be calculated using the chain rule again, by summing up the partial derivatives computed from all units that accept inputs from the current unit d during the forward pass:

$$\delta_d^{(2)} = \sum_l \frac{\partial L_n}{\partial a_l^{(3)}} \frac{\partial a_l^{(3)}}{\partial a_d^{(2)}} \qquad (2\text{-}16)$$

Rearranging the formula with equations *2-2*, *2-3* and *2-16*, the $\delta^{(2)}{}_d$ can be formulated as:

$$\delta_d^{(2)} = f'\left(a_d^{(2)}\right) \sum_l w_{ld}^{(3)} \delta_l^{(3)} \qquad (2\text{-}17)$$

And the error signal for the output layer can be calculated with:

$$\delta_l^{(3)} = \frac{\partial L_n}{\partial \hat{y}_l} \frac{\partial \hat{y}_l}{\partial a_l^{(3)}} \qquad (2\text{-}18)$$

Where $\hat{y}_l$ is the model prediction of n^{th} input.

Once the errors are calculated for each neuron, the weights can thus be updated using the gradient descent technique introduced above.

2.3.4 Stochastic gradient descent

In practice, updating the weights with the gradients that are calculated over the entire dataset is often not feasible, due to the large memory consumption and long processing time. Instead, computing the gradients with randomly sampled small batches (mini-batch) of data is a preferable choice, this is the technique called 'stochastic gradient descent' (SGD). With SGD, the model is optimised against the loss in an iterative process, performing the updates based on every mini-batch, enabling a local optima to be reached.

2.3.5 Introduction to CNN

For histopathological images, the arrangement of pixels in 2D directions contains crucial information of how tumour forms and grows. Therefore, it is desirable to retain the spatial relationships when learning representations from the images; in comparison, the orientation does not carry important information. The convolution operation enables processing of both centering and neighbouring pixels, which is suitable for this scenario, hence has gained great attention in computational pathology. A basic CNN mainly consists of the following units: convolution layer, pooling layer, fully-connected layer and output layer (97).

2.3.5.1 Convolutional layer

A convolutional layer is typically referred to as a kernel, it is a matrix with randomly initialised 'weights', these weights will be constantly adjusted during training. For images, the kernel typically has 2 dimensions (i.e. 5 × 5), and for histopathological images with 3 colour channels (red, green and blue), the kernel will have an additional dimension (i.e. 5 × 5 × 3) so that the kernel and image match in shape during matrix multiplication. The convolution procedure starts by sliding the kernel across an image. At each sliding location, the element-wise product between kernel and input signal is computed and summed up, forming the output of the current receptive field. The same procedure is performed repeatedly until the kernel scans through the

entire image. In general, a sliding stride is primarily defined to determine the amount of shift in pixels between each convolution operation.

The kernel is also referred to as a filter for its ability in returning higher value when the feature within a receptive field correlates with the shape that the set of kernel weights stands for. For instance, a kernel with larger weights diagonally has the capacity in detecting slash patterns (\ ). Hence, in comparison to input devoid of any pattern, it tends to generate higher value through the convolution step when the input also contains a slash pattern.

2.3.5.2 Activation layer

The output within each receptive field is subsequently passed through an activation layer, for example a rectified linear unit (ReLU) (98). The layer functions by mapping the input with a simple rule: ReLU = max(0, x) where x is the input of the activation function. The activation layer serves to gain non-linearity, therefore expanding the capacity of CNN models to learn non-linear patterns. Outputs from this layer are concatenated with their spatial coordinates and regarded as feature maps, as they reflect the mapping of where a desired feature is detected in the previous layer.

2.3.5.3 Pooling layer

A pooling layer refers to the step where output feature maps are further reduced in size, retaining only one value within each prespecified window size (i.e. 2 × 2). The value is computed based on different pooling functions, and is typically the maximum or average of pixel values within the receptive field for max pooling or average pooling, respectively. This step improves the model's tolerance towards small transitions in input signals and reduces the required parameters by shrinking the output features; together, it renders the model a simpler and more robust property.

2.3.5.4 The deep structure

Before the output layer, a CNN model is configured by repeatedly stacking the three aforementioned layers in a sequential manner.

Information revealed from an image possesses a hierarchical structure: the intricate shapes are constructed on top of basic elements. Accordingly, simpler features learned by earlier layers become inputs for later layers, and are in turn combined through convolution to facilitate the detection of more complex features. As the information passes by more layers, the kernels gradually acquire the power to recognise objects with an increased complexity.

Models with deeper configurations can possess more capacity for learning complicated patterns, i.e. morphological features in the case of histopathology, especially when the dataset has few classes but is rich in image data, and hence may be more suitable for modelling histopathological features (99)(100). On the other hand, high similarities can be found between representations generated from consecutive intermediate layers, and the observations become more profound with deeper or wider networks, suggesting that some layers can be pruned

without sacrificing model performance (101). But this phenomenon can be attenuated with large training data, hence, its influence on histopathological datasets with millions of inputs remains unknown.

2.3.5.5 Fully connected layer

Towards the final layers of the CNN, a set of features with various degrees of complexity have been extracted. To link the information with training objectives, an integration of learned features is required. The fully connected layer is designed to provide an extensive combination of all features from the convolutional part of the model. Therefore, one or more fully connected layers are typically inserted between the last convolutional unit and the output layer.

2.3.6 Applications of deep CNN models in the medical domain

The past decade has witnessed a soaring trend in the application of CNNs for histopathological analysis in a large variety of prediction tasks (82)(102), some have demonstrated promising outcomes. Campanella, G. et al designed a CNN based framework for the prediction of prostate cancer and skin basal cell carcinoma on HE stained images that achieved outstanding performance (AUC higher than 0.98). The model is expected to help preclude over 75% slides that do not carry valuable information for cancer diagnosis without diminishing the sensitivity for cancer detection on patient level (103). Another study showed that CNN can both detect prostate cancer with core needle biopsies with an AUC of 0.997 and perform Gleason grading with performance comparable to specialists (104).

Apart from performing routine diagnostic tasks, the strong ability to learn features from data has rendered CNNs the potential to distinguish refined tumour characteristics that are not currently achievable through pathological assessment. One application scenario is to predict disease outcome and improve patient risk stratification (105). For example, using features extracted from HE stained TMA tissues by CNN models, Bychkov et al. stratified colorectal cancer patients into two risk groups that exhibited distinct survival outcomes (105); alternatively, pathological data can be combined with genetic and clinical data, accounting for more relevant biological mechanisms to facilitate the prediction of patient prognosis (106)(107).

The detection of somatic mutations with lung cancer WSIs by Coudray et al uncovered an innovative finding that CNNs can potentially link morphological appearance to their underlying molecular changes. This discovery soon received increased attention and fueled similar research in other cancer types such as liver cancer (108), colon cancer (109) and breast cancer (110). Two studies successfully predict genetic alterations in a pan-cancer manner, indicating a shared association between genetic and morphology phenotypes across different cancer types and can be captured by deep learning models (111),(112). Apart from mutations, prediction of gene expression and molecular subtypes are also achievable (111),(112),(79),(84).

3 RESEARCH AIMS

The overall aim of the doctoral thesis was to develop and validate deep learning models for the purpose of improved prognostic risk stratification of breast cancer patients, and for prediction of molecular characteristics of breast tumours, using digital histopathological WSIs.

The specific aims for each study are listed below:

- **Study I** aimed to improve risk stratification of grade 2 tumours by training deep learning models on morphological patterns in histopathology images of grade 1 and 3 tumours.
- **Study II** aimed to develop and validate deep learning models for the ability to predict tumour average level mRNA expression of individual genes, and to predict intra-tumour spatial gene expression values.
- **Study III** extends the results from **study II** by quantifying multiple metrics of intra-tumour gene expression heterogeneity and assessing the prognostic value through survival analysis.
- **Study IV** aimed to compare the agreements in prognostic risk stratification between Stratipath Breast, a product developed for clinical use based on the concept proposed in **study I**, and the Prosigna® gene signature assay, a product for clinical use based on molecular profiling.

4 MATERIALS AND METHODS

4.1 DATA SOURCES

Patients and study materials used in the current thesis were collected from four studies. For each enrolled patient, demographic information was extracted from health records and the HE stained tissue slides were retained, digitised into WSIs. The information of data sources is summarised below:

The Clinseq study

The Clinseq study (Clinical Sequencing of Cancer in Sweden) consists of female participants that were included in two cohorts: Libro-1 and Karma. Libro-1 (113) recruited patients retrospectively in 2009 who were younger than 80 years old by the age of diagnosis, and received surgery between 2001 to 2008 at the Karolinska University Hospital. Karma (114) enrolled patients who were diagnosed with breast cancer at the South General Hospital in 2012. Clinical characteristics and follow-up information were retrieved from the Stockholm-Gotland Regional Breast Cancer quality register and the Information Network for Cancer Care (INCA) (115) respectively. INCA includes detailed records for patients with primary invasive or *in situ* breast cancer from 2007 to 2018, whereas the Stockholm-Gotland Regional Breast Cancer quality register contains historical data for INCA up to 2007. RNA-seq data were extracted according to the protocols outlined in the previous study (69) and were originally retrieved with fresh frozen tumour tissues from surgery excision. HE stained FFPE slides were digitised with a combination of Hamamatsu XR and S360 scanners. The Clinseq cohort was used extensively as the training and internal test sets for **study I-III**.

The Cancer Genome Atlas (TCGA) breast cancer study

The study comprises patients that were diagnosed with invasive breast cancer and underwent surgical resection. No adjuvant therapy was applied prior to surgery. The demographic and follow-up information was retrieved from previous publications while the histological grading and individual subcomponent scores were manually extracted from the pathological reports. The RNA-seq data were acquired from http://cancergenome.nih.gov/ in June 2014, under the approval from the TCGA data access committee (dbGAP project ID 5621). The digitised diagnostic slides were downloaded from the official portal https://portal.gdc.cancer.gov. The slides were scanned by Aperio scanners. The TCGA dataset served as the training and internal test sets for **study I-III**.

The SöS-BC-1 study

The SöS-breast cancer batch 1 (SöS-BC-1) cohort contains retrospectively enrolled patients that were diagnosed of breast cancer at the Stockholm South General Hospital between April 2012 to October 2014, and between October 2015 and May 2018. All clinical information was retrieved from the national quality registry for breast cancer (NKBC) (116). NKBC is a reconstructed INCA register established in 2019, with detailed records regarding patient

demographics, tumour characteristics, treatments and follow-up information. The slides were collected and digitised in-house using a Hamamatsu Nanozoomer XR scanner. In **study I**, SöS-BC-1 was assigned as the training and internal test set for the development and evaluation of the DeepGrade model in distinguishing NHG 1 and 3 tumours. In **study III**, it acted as an external test set to validate the prognostic significance of the proposed risk stratification model.

The SCAN-B-Lund study

The SCAN-B-Lund cohort (Sweden Cancerome Analysis Network - Breast) prospectively enrolled patients that were diagnosed of primary invasive breast cancer in Lund from 2010 to 2019 (117). Basic tumour characteristics such as biomarker status, tumour grade, treatment and follow-up information were retrieved from INCA. The slides were scanned on a Hamamatsu Nanozoomer XR scanner. The cohort was used as an external test set to evaluate the prognostic significance of the proposed risk stratification models for **study I** and **III**. A subset of patients with the RNA-seq data available from the NCBI Gene Expression Omnibus (Accession Nos. GSE81538) formed the external test set (ABiM cohort) for **study II** (118).

The Prosigna cohort

The Prosigna cohort consists of patients that were diagnosed of invasive breast cancer from March 2020 to March 2022 at the Karolinska University Hospital or Södersjukhuset, and also subjected to molecular diagnostics using the Prosigna® test. Participants shared the characteristics of being postmenopausal with ER positive, HER2 negative and node negative tumours. Clinical information regarding crucial tumour characteristics were extracted from the clinical records; a pair of FFPE sections originating from surgical specimens were either analysed with the Prosigna® test at Karolinska University Hospital or stained with HE and underwent the Stratipath breast test at Karolinska Institutet. The cohort composed the study material for **study IV**.

4.2 WHOLE SLIDE IMAGES

The HE stained FFPE slides from Swedish cohorts were scanned locally with Hammamastu Nanozoomer digital scanners of model XR or S360 at 40X, whereas slides from TCGA were downloaded from the digital portal as previously described.

The digitised slides are referred to as WSIs, each consisting of billions of pixels with several gigabytes in size. To facilitate data visualisation at different magnification levels, in the scanner native output image files, a set of lower resolution images are computed sequentially and stacked on top of the full resolution originals, forming an image pyramid, with the lowest resolution on top. This structure facilitates a smooth loading procedure as it is often necessary to examine the image under various magnification levels, such as during pathological examination. The image with the highest resolution forms the bottom of the image pyramid and the size is reduced sequentially forming the subsequent levels (**Figure 4A**).

Images are matrices of pixels. A pixel is composed of red, green and blue (RGB) colour channels with each digit ranging from 0 to 255. In this format, an image is analysed by the computer vision model as 3D tensors of pixel values with the dimension of $N_{rows} \times N_{columns} \times 3$ as is shown in **Figure 4B**. It is worth noting that there is a variation for representing an image across different programming libraries, hence, the order of these three components can differ.

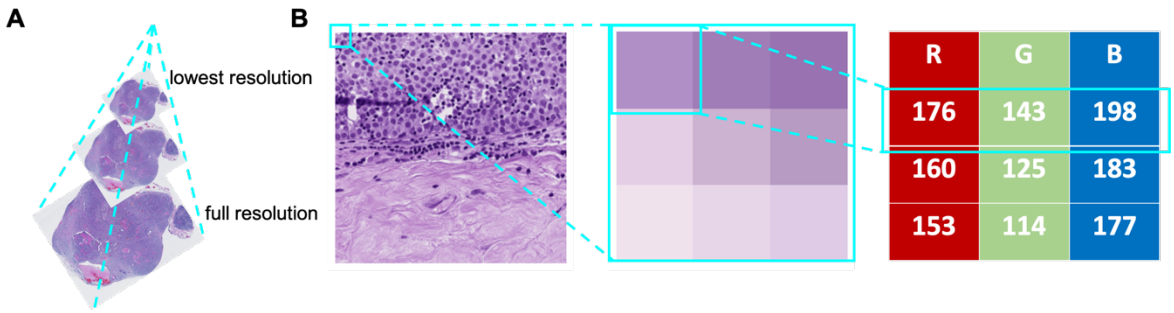

Figure 4. Demonstration of image pyramid and digitised HE stained WSIs.

4.3 IMAGE PREPROCESSING

As an initial step, a standardised preprocessing pipeline was applied to each WSI. The entire workflow was summarised in **Figure 5** and consisted of tissue region identification, tiling, quality control, colour normalisation and tumour region segmentation. Tissue regions were extracted from downsampled WSIs from the image pyramid. For slides scanned in-house with the Hamamatsu scanners, a downsampling factor 32 was used consistently across all slides; For TCGA slides that were scanned with Aperio scanners of multiple prototypes, when the same downsampling factor was not available, a lower resolution image with the compression level closest to 32 was used. The extracted low resolution image was transformed to HSV (*hue*, *saturation*, and *value*) colour space where a mask was generated with Otsu's thresholding (119) on the *saturation* channel to detect tissue components; In order to filter out pen marks, another binary mask was generated to exclude pixels with a *hue* value less than 0.75; The tissue region was retained with logical intersection of the two masks, followed by removing small holes or noises that have a radius of up to 10 pixels.

Next, we sequentially tiled the image at full resolution into smaller image patches of size 1196 × 1196, followed by a 2X downsampling, so that the final output has the shape of 598 × 598 pixels and represents the magnification of 20X. This granularity was chosen to ensure that the images are composed of tubular structures while details such as cellular patterns can be retained. Another reason to scale tiles to 20X resolution was to mitigate the blurry artefacts that were prominent under full resolution. Slides from the SöS-BC-1 cohort were tiled with 25% overlap to balance the total number of tiles, while 50% overlap was chosen for the other cohorts.

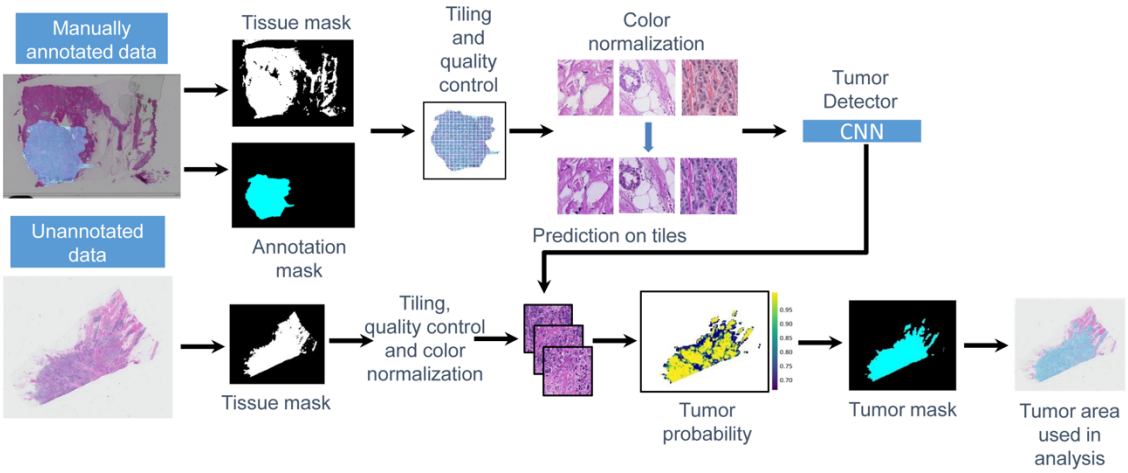

Figure 5. Illustration of the preprocessing workflow. Adapted from Wang *et al*. 2022 (120), with permission from Elsevier [author's material].

4.4 QUALITY CONTROL

Although downsampling from 40X to 20X can significantly reduce blurry artefacts, tissues containing subregions that are out of focus remain problematic; furthermore, images composed of adipose tissue and devoid of epithelial cells provide limited value for prediction objectives, hence, a quality control measure was adopted to remove blurry images by computing the variance of pixel values by convolving with a Laplacian operator (121). It generates the second order derivative of a given image, enabling the detection of edges with fast transition in grayscale intensity. In-focus images can be dominated by sharp edges with a high value in the variance whereas the out-of-focus images are characterised by a slow transition in grey values that is associated with a low variance.

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \qquad (4\text{-}1)$$

The calculated focus measurement was thus compared with an empirically chosen threshold of 500, and tiles with a value lower than the threshold were considered as blurred and were excluded from subsequent analysis.

4.5 COLOUR NORMALISATION

Tissues possess a diverse capacity in absorbing stains and such intrinsic difference is exploited in assessing lesions. However, slides prepared under divergent protocols, such as scanned by different scanners, typically exhibit noticeable colour variations that could confound model predictions. To alleviate this, we employed and modified the colour normalisation following Macenko's method (122).

The basic idea is to firstly transfer images from RGB colour space into optical density (OD) space (123), making the colour components linearly separable as in (4.2), where I_c denotes the RGB colour vector. Secondly, a slide level stain vector was estimated with Macenko's method. Thirdly, pixel level concentration coefficients were computed using the slide level stain vector

and the pixel level OD values (124). Lastly, the normalised OD value was obtained by multiplying a reference stain vector and the pixel level concentration value.

$$\boldsymbol{OD} = (-1 * log(\boldsymbol{I_C}/255) \quad (4\text{-}2)$$

In practice, it is advisable to adjust the brightness of each incoming image before the colour normalisation step. This was achieved with a slide level luminosity reference. In brief, 100 randomly sampled tiles from one WSI were concatenated and transformed from the RGB to CIELAB colour space. The 95th percentile of the L^* channel value was chosen as the slide reference. Using this reference, the pixel values within each tile were adjusted by either linearly scaled between 0 to 255 according to the reference, or setting to 255 if the value exceeded the given reference. Lastly, we transformed the image back to RGB profiles for colour normalisation.

The next step was to estimate stain vectors. Two types of stain vectors were needed. One was the reference vector as a standard to normalise all images to, which was estimated with 3000 random gathered tiles within the pool of all available WSIs in the training set; the other was the slide level stain vector for each WSI and was computed with 100 randomly sampled tiles from the WSI under study. To estimate a stain vector, luminosity adjusted image tiles in OD space were first randomly sampled and concatenated. Secondly, singular value decomposition (SVD) was performed using the concatenated images in the previous step. Next, each pixel was projected onto the plane that was spanned by the two vectors corresponding to the two largest singular values. After this, the vectors that have 99th percentile or 1st percentile among all angles with respect to the first vector were regarded as the *H* and *E* vectors, respectively. The percentiles were chosen empirically to allow for more robust estimation.

4.6 CANCER DETECTION MODEL

For slides from the Clinseq cohort, the tumour region was annotated by experienced pathologists. For slides from the rest of the cohorts, tumour masks were generated from a deep learning based tumour detection model. The model was trained with tiles from the Clinseq cohort where the annotations were used as ground truth labels.

The binary cancer detection model was trained with WSIs from the Clinseq data. Patients were split into training (N = 149), validation (N = 36) and test (N = 47) sets. Within the training set, 80% were used to train the model and 20% as the tuning set to monitor model performance. An Inception V3 model (125) with 'ImageNet' (126) pretrained weights was optimised with the Adam (127) optimiser. The learning rate was set to 1e-4, with $\beta1 = 0.9$, $\beta2 = 0.999$, ε = None, decay = 0. Each mini-batch contained 24 cancer tiles and 24 non-cancer tiles. The model was evaluated on the tuning set after every 50 iterations. Model predictions were compared with ground truth annotations with the cross-entropy loss function. We employed early stopping with a patience of 10 and the criteria of non-improvement was considered as having less than 0.003 change in the tuning set loss. Data augmentations with random flips and 90° rotations were applied to the training samples.

Model performance was evaluated with the Receiver Operating Characteristic (ROC) curves and a threshold to dichotomise output probabilities was chosen using the validation set with Youden's index (128) on the ROC curve.

For cohorts without invasive cancer annotations, the trained model was applied to each tile and by using the same cut-off threshold, a tile was predicted as cancer or non-cancer. To mimic manual annotation with a coarse boundary, post-processing was performed by first mapping tiles with their binary predicted labels back to the original location, and then using morphological closing and opening operations to remove holes or objects with a size less than 405 pixels. The output was a binary tumour mask that was eventually employed to label image tiles. We retained the tiles within the region labelled as 'tumour' by the model for subsequent analysis.

4.7 APPLICATION OF MACHINE LEARNING MODELS

4.7.1 Training deep CNN models with a classification objective

In **study I**, an ensemble of 20 deep learning models with the Inception V3 architecture were trained to classify NHG 1 and NHG 3 cases. The training (N = 674), tuning set (N = 170) and internal test sets (N = 351) comprised a combination of WSIs from the Clinseq, TCGA and SöS-BC-1 cohort. The models were initiated with the weights pre-optimised using the 'ImageNet' dataset. Models were trained with the SGD optimiser and a learning rate of 1e-3. A mini-batch with balanced sampling of 32 tiles from two classes were used and the model performance was evaluated on the tuning set for every 250 iterations. The learning rate was set to reduce by 50% if the model did not improve for 10 epochs. The same loss function, early stopping criteria and data augmentation were applied as depicted in **section 4.6**. In the test sets, the predicted probability of NHG 3 for each tile was averaged across all 20 predictions and the slide level probability was obtained using the upper quantile across all tile level predictions within that slide. ROC curves were generated to evaluate model performance.

4.7.2 Training deep CNN models with a regression objective

In **study II**, tiled images were used as input to optimise the Inception V3 models with slide level gene expression as the outcomes. The regression models were configured by switching the last layer with one neuron followed by a linear activation. Models were initiated with the 'ImageNet' weights and were trained for each individual gene separately. Adam optimiser was used together with a learning rate of 1e-6 and default parameters including $\beta1 = 0.9$, $\beta2 = 0.999$, ε = None, decay = 0. Mean squared error was used as the loss function.

Clinseq and TCGA datasets were merged and split on patient level to form the training (N = 558), tuning (N = 139), validation (N = 122) and internal test sets (N = 172).

The models were trained on the training data with a mini-batch of 32 tiles, and evaluated on the tuning set every 150 iterations (partial epoch). Early stopping was employed with the

patience set to 80 partial epochs and the minimum change in loss set to 0.003. Data augmentation was applied as previously described.

4.7.3 Regularisation

Regularisation is often applied in an attempt to prevent overfitting in many machine learning scenarios. Depending on whether a regularisation term is explicitly added to the optimisation function, the technique can be further divided as explicit or implicit regularisation.

4.7.3.1 implicit regularisation

Early stopping (129) techniques in optimising a CNN model is an implicit regularisation, the method stops the training process when model performance plateaus with a predefined criteria on the tuning set, thus, avoiding overfitting on the training data.

Batch normalisation (130) is added before the activation layer with an aim to normalise input values so that each mini-batch has 0 mean and unit standard deviation. The normalised mini-batches then pass through a scale and shift operation with two learnable parameters respectively, and eventually forming the input for the activation layer. Although study demonstrated that this regularisation only marginally improves generalisability, we can potentially benefit from a more stable learning curve and less computation time (131).

Dropout (132) works as another implicit regularisation technique by randomly setting 0 to the neurons within a layer with a probability of *p* during training time. Only the rest of the neurons contribute to computing the outputs and accept error signals from the backpropagation procedure for each iteration. Dropout is not used during inference time, hence, for testing the model, the weights associated to each neuron are scaled by a factor of (*1-p*), adjusting for additional activations compared with the training time.

4.7.3.2 explicit regularisation

In model optimisation, a regularisation term can be added to the loss function. It is also regarded as a penalty term since its presence enforces a constraint on learned parameters.

Given *N* samples with *d* predictors, a training loss can be defined with the following formula where $\Phi(\cdot)$ is a mapping function applied on each input x_n, $\boldsymbol{\theta}$ is the coefficient vector and $R(\boldsymbol{\theta})$ denotes the regularisation term:

$$L_{\text{train}}(\theta) = \sum_{n=1}^{N}\left(y_n - \sum_{p=1}^{d}\Phi_p\left(x_n\right)\theta_p\right)^2 - \lambda * R(\boldsymbol{\theta}) \qquad (4\text{-}3)$$

The two commonly used terms are L1-regularisation (Least Absolute Shrinkage and Selection Operator, LASSO) (133) where:

$$R(\theta) = \sum_{p=1}^{d} |\,\theta_p| \qquad (4\text{-}4)$$

And L2-regularisation (Ridge) (134) where:

$$R(\theta) = \sum_{p=1}^{d} \theta_P^2 \qquad (4\text{-}5)$$

L1-regularisation encourages sparse solutions where only a subset of features have a non-zero coefficient while L2-regularisation addresses the multicollinearity problems when working with high-dimensional data.

A linear combination of these two methods forms the so-called elastic net regularisation with the following formula (135):

$$\lambda \left(\frac{1-\alpha}{2} \sum_{p=1}^{d} \theta_P^2 + \alpha \sum_{p=1}^{d} |\theta_p| \right) \qquad (4\text{-}6)$$

Where an additional parameter α was introduced to control the ratio of L1-regularisation.

In **study III**, we applied elastic net penalty while optimising the partial likelihood function of a Cox proportional hazard model; more details regarding the Cox model are provided in **section 4.8.3.2**.

4.8 STATISTICAL ANALYSIS

4.8.1 Assessment of classification performance

In a classification context (**study I**), the model performance was primarily evaluated with the Area Under the ROC Curve (AUC). The curve serves a reflection on the model's overall capacity in separating positive and negative cases. During inference time, a set of probabilities are generated on all input data and by setting an output probability as the cut-off threshold, test data can be dichotomised into two predicted classes; Next, by comparing the predicted labels with the associated ground truths, a pair of true positive rate (sensitivity) and false positive rate (1-specificity) can be computed. Following the same manner, each threshold would contribute with a different pair of sensitivity and specificity. The ROC is then constructed by plotting all ranked measurements with y-axis being the sensitivity and x-axis as the (1-specificity). In case when a model is completely devoid of discriminative capacity, the ROC curve is a diagonal line as it always separates data with 50% sensitivity and specificity, the associated AUC is equal to 0.5. On the contrary, a best performing model with 100% sensitivity and specificity has an AUC of 1. In between, a larger AUC indicates better model performance.

4.8.2 Statistical association analysis

4.8.2.1 Spearman correlation

In **study II**, we studied the association between RNA-seq measured and model predicted gene expression levels without assuming an underlined linear dependency, hence, the Spearman's

rank correlation was chosen. Instead of measuring directly to what extent two variables change their values together, Spearman's method evaluates the correlation of ranks of the values.

For n pairs of samples from $\boldsymbol{x} = [x_1,...,x_n]^T$ $\boldsymbol{y} = [y_1,...,y_n]^T$, assuming $R(\cdot)$ being the rank of a variable and let d_i denotes the difference between the ranks of two observations ($d_i = R(x_i) - R(y_i)$), the coefficient rho (ρ) can be calculated as:

$$rho = 1 - \frac{6\Sigma\, \mathrm{d}_i^2}{n(n^2-1)} \qquad (4\text{-}7)$$

The coefficient ranges from -1 to 1, with a higher absolute value indicating a stronger monotonic association whereas 0 means no identified association. Meanwhile, the direction of the association is reflected by the positivity or negativity.

4.8.2.2 coefficient of determination

To evaluate how well the feature representations (i.e. input image data) that have been learned by the model predict outcomes such as gene expression, we calculated the proportion of variance explained (known as Coefficient of determination).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (4\text{-}8)$$

Where y_i stands for the ground truth for the i^{th} input, $\hat{y}_i$ is its model predicted value, and $\bar{y}$ is the mean among all input data. It can be seen from the formula that the better the model is, the closer the score is to 1; On the other hand, if a model constantly outputs the $\bar{y}$, the score then becomes zero. It is also not uncommon to have a negative R^2 when the model generates predictions that are seriously wrong.

4.8.2.3 Linear mixed effects model

If the data has a hierarchical structure, such as multiple measurements from the same patients, then, there is a dependency among observations. Linear mixed effects (LME) (136) models account for this by a model parametrised by fixed effects and random effects terms in the model. The model takes the following form:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon} \qquad (4\text{-}9)$$

Where $\boldsymbol{y}$ is the outcome variable, $\boldsymbol{X}$ is the design matrix of explanatory variables whose fixed effect coefficients are expressed as a vector $\boldsymbol{\beta}$; $\boldsymbol{Z}$ is the matrix for random effects whereas the random effect coefficients are expression with $\boldsymbol{u}$ which follows a normal distribution and the variance was estimated; finally, the $\boldsymbol{\varepsilon}$ denotes the error term.

For instance, in **study II**, we applied LME to examine the relationship between gene expression estimate from ST measurements and from model predictions, 22 tumours were included in the study with 12 regions of interest each, resulting in a total of 264 observations per gene. The outcome can then be expressed with a vector $\boldsymbol{y}$ with a shape of 264 × 1, and model predictions

as well as fixed intercept forms the design matrix $\boldsymbol{X}$ with a shape of 264 × 2; in addition, the tumour index was added to the model as the random effect term, therefore, $\boldsymbol{Z}$ had the shape of 264 × 22; lastly, the residual was also a 264 × 1 column vector. A p-value examines whether the null hypothesis of zero coefficient can be rejected.

4.8.2.4 Multiple testing

Statistical inference is accompanied with the risk of rejecting a true null hypothesis, which is known as the type I error. With the widely adopted threshold of 0.05, performing 1000 tests when the null hypothesis is true will still generate 50 significant results on average. It is therefore necessary to control for the inflation of false positives by correcting for multiple testing.

One way is to apply the Benjamini-Hochberg (BH) method (137) in controlling the false discovery rate, which is the expected proportion of committing a type I error. The procedure starts from ranking all p-values from each comparison, and then calculates the BH critical value per test with the equation below:

$$\text{Critical Value} = (k/N)*\alpha \qquad (4\text{-}10)$$

Where k is the rank for p_k, N is the total number of tests and α is the significant level defined by the user. Next, find the largest p-value that is smaller than the corresponding critical value, the null hypothesis is then rejected for the current test and all the other tests with rankings prior to this test. Equivalently, one can compute a FDR-adjusted p-value, and compare with the prespecified significant level, with the following function (138):

$$p_{\mathrm{BH}(i)} = \min\{\min j \geq i\{Np_j/j\}, 1\} \qquad (4\text{-}11)$$

While the FDR is the expected ratio of false discoveries, another similar concept named the family-wise error rate (FWER) depicts the probability of making at least one type I error. This couples with a more stringent controlling method – the Bonferroni correction (139). The method rejects a null hypothesis if $p < \alpha/N$, where α is the significant level and N is the total number of tests.

4.8.3 Survival analysis

Survival analysis refers to the type of statistical modelling of time-to-event data. It is widely used in medical research to investigate patient prognosis. An event is defined prior to the analysis, and is often chosen as recurrence of disease or death in the context of cancer studies. We performed survival analysis in **study I** and **study III**.

In both studies, the follow-up time was defined as the date of diagnosis with breast cancer to the date of experiencing an event or loss of follow-up. The event was defined as the recurrence, metastasis of breast cancer or death. A participant didn't experience an event before withdrawing from the study or by the end of the study was censored (right censoring).

4.8.3.1 Kaplan-Meier (KM) curve

The proportion of patients that remains to be event-free per interval at each event time can be depicted by the Kaplan-Meier curves. The y-axis of the curve is the probability of survival past time *t*, the x-axis is the time. A log-rank test is performed to assess whether there are statistical differences in survival between different groups of patients. The test is non-parametric, as it does not rely on any parametric assumptions. The test statistics is obtained by firstly comparing the observed number of events to the expected number of events at each event time and then sum the values up across the entire study period, for *k* groups of patients, the output follows approximately a chi-square distribution:

$$X^2 = \sum_{i=1}^{k} \frac{(\sum O_{it} - \sum E_{it})^2}{\sum E_{it}} \qquad (4\text{-}12)$$

4.8.3.2 Cox proportional hazards regression analysis

When evaluating the effect of multiple risk factors that simultaneously influence patients survival, we performed Cox proportional hazards regression. The model provides the estimation of HR and has the form:

$$h(t|\boldsymbol{X}) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k) \qquad (4\text{-}13)$$

Where h0(*t*) denotes the baseline hazard when all the variables are set to zero and h(*t*|**X**) is the hazard at time t. Thus, the HR between two values of a predictor X_1 can be calculated as:

$$HR = h_0(t)\exp(\beta_1 X_1 = a) \,/\, h_0(t)\exp(\beta_1 X_1 = b) = \exp(\beta_1(a-b)) \qquad (4\text{-}14)$$

It describes the increase or decrease in risk of experiencing the event, in the group under study compared with the reference group. The ratio is widely used to describe the prognostic value for risk factors.

The Cox model is regarded as semi-parametric for the fact that it does not make assumptions on the distribution of the baseline hazard, rather, it assumes proportional hazards over time, which entails a constant HR between subgroups of patients regardless of time. In violation of the proportional hazard assumption, stratified Cox regression was considered as an alternative approach. By stratifying upon the variable that violates the assumption, it models different baseline hazard functions for each stratum of the variable.

4.9 ETHICAL CONSIDERATIONS

The research projects shared a common focus in developing deep learning based approaches for improved precision medicine in breast cancer, and the models were optimised with HE stained FFPE slides towards prediction of histological grade, gene expression as well as patient prognosis. For this purpose, HE stained FFPE slides, clinical characteristics and follow-up information were collected to evaluate model efficacy. All participants provided signed information consent voluntarily, and no intervention was included in any of the studies. These

retrospective research studies therefore posed no additional harm that would lead to increased patients' disease burden.

Data that involves identifiable information are considered sensitive and were handled with special care to comply with the General Data Protection Regulation (GDPR) together with the Swedish Data Protection Act.

All data were stored and analysed on password protected systems at the Department of Medical Epidemiology and Biostatistics. When the computation needed to be performed in the computer clusters in Finland, both patient ID and gene names were ensured to be anonymised prior to data transfer, and the transfer process was also conducted in an encrypted manner. All access to data was restricted to study specific use.

5 RESULTS

5.1 STUDY I

The Nottingham histological grading system provides a standardised manner in assessing tumour aggressiveness and serves as an independent prognostic factor that can guide clinical decision making. However, as tumorigenesis denotes a gradual loss of normal function or organisation, the malignant change often appears with a continuous spectrum of morphological abnormality, making it hard to classify tumours into discrete stratum. This difficulty has led to the profound phenomenon of inter-assessor variability when grading tumours, especially for the intermediate class (NHG 2). The current study aims to train a binary classification model with NHG 1 and 3 tumours, and then use the model to dichotomise NHG 2 tumours into 2 groups with Grade 1-like or Grade 3-like appearances, respectively.

Clinseq, TCGA and SÖS-BC-1 were combined and split into training, test 1 and test 2 groups. An ensemble of 20 Inception V3 models (DeepGrade) were optimised separately using the training set and evaluated using test sets. The model achieved an AUC of 0.927, 0.937 and 0.919 on the combined test 1 and test 2 data per cohort respectively (**Figure 6A-C**), suggesting a good separation between the two grade groups which possess relatively extreme and homogenous morphological changes.

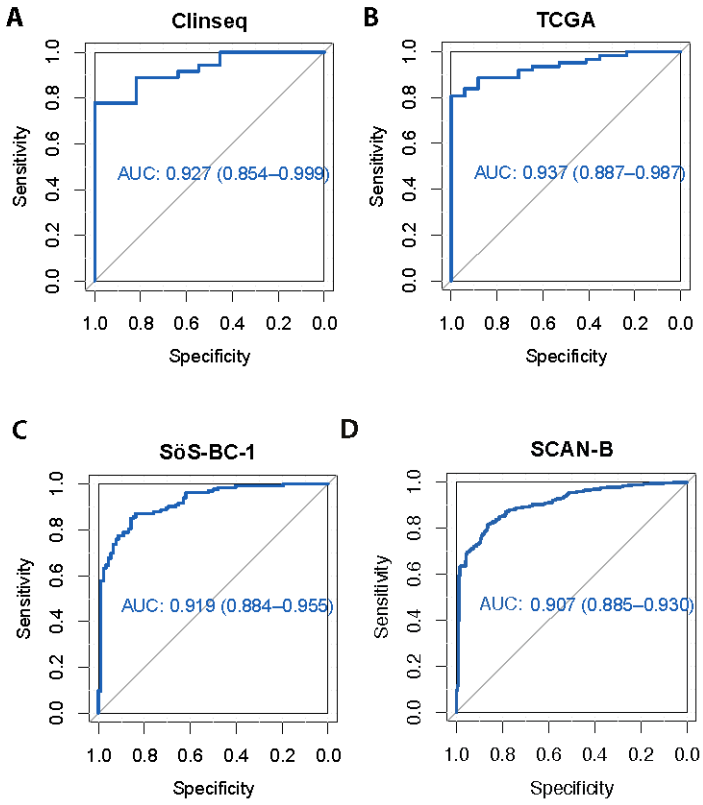

Figure 6. ROC curves for discriminating NHG 1 and NHG 3 tumours. **A.** Results on the Clinseq data. **B.** Results on the TCGA data. **C.** Results on the SÖS-BC-1 data. **D.** Results from the external test cohort SCAN-B-Lund. Modified from Wang *et al*. 2022 (120), with permission from Elsevier [author's material].

Next, we applied the DeepGrade model to stratify patients with NHG 2 tumours and obtained two groups representing NHG 2 with low risk (DG2-low, N = 242, 65.0%) and NHG 2 with high risk (DG2-high, N = 130, 35.0%) which showed a significant difference in recurrence free survival, reflected by KM curves (P = 0.0016, log-rank test, **Figure 7A**). A Cox regression analysis was also performed to study the HR while adjusting for other covariates including age, tumour size, lymph node status, ER status, HER2 status, and the results suggested that the DeepGrade risk stratification was an independent prognostic factor with a HR of 2.94 (95% CI: 1.24-6.97, P = 0.015, **Figure 7B**).

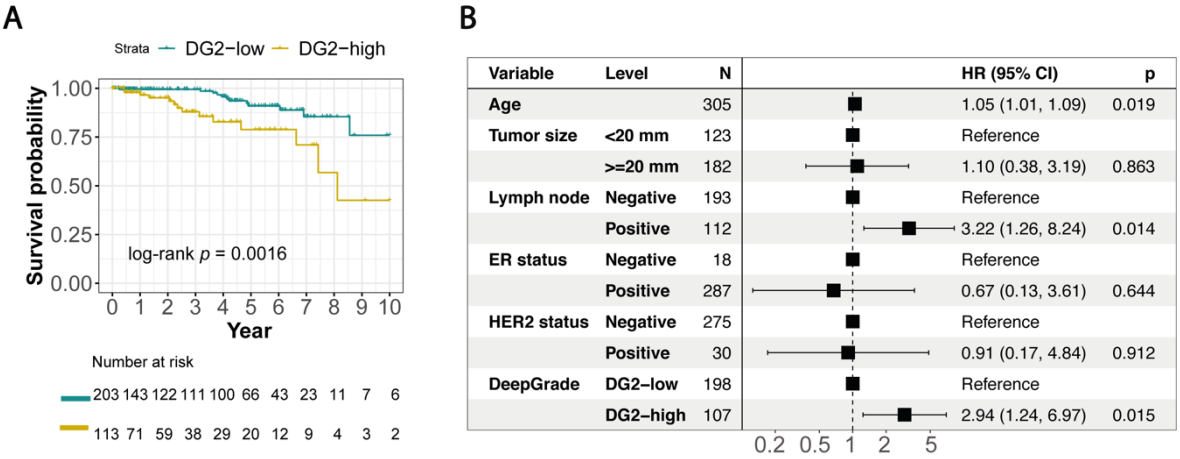

Figure 7. Risk stratification in the training cohort. **A.** Results from the Kaplan-Meier estimator. **B.** Results from the multivariable Cox regression analysis. Modified from Wang *et al*. 2022 (120), with permission from Elsevier [author's material].

The performance of the DeepGrade model on unseen data was validated using the SCAN-B-Lund cohort that involves 1,262 patients. For the classification of NHG 1 and NHG 3, the model yielded a slightly dropped performance with an AUC of 0.907 (**Figure 6D**).

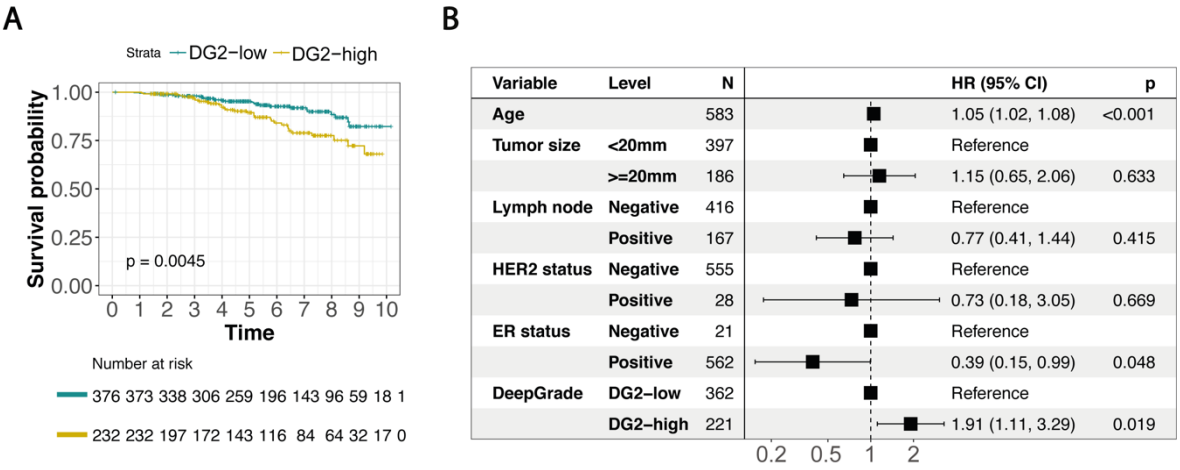

Figure 8. Validation on the external test cohort. Results displayed in the same manner as in **Figure 7.** Modified from Wang et al. 2022 (120), with permission from Elsevier [author's material].

Among the 608 NHG 2 patients, 376 (61.8%) were stratified as DG2-low, whereas 232 (38.2%) as DG2-high. A significant difference in survival probabilities were revealed from the KM curve (P =

0.0045, log-rank test, **Figure 8A**) and the adjusted HR between low and high risk groups was 1.91 (95% CI: 1.11-3.29, P = 0.019, **Figure 8B**). Modified from Wang et al. 2022 (120), with permission from Elsevier [author's material].

5.2 STUDY II

In **study II**, we employed deep learning based models (EMO) as a scalable approach to predict tumour average gene expression values for the whole transcriptome; In addition, the developed models were further applied to generate spatial gene expression estimations within each WSI, whose efficacy was validated through spatial transcriptomics.

The models were trained and initially evaluated using the Clinseq and TCGA cohorts. Prior to the analysis, patients were split into training, validation and internal test sets with 697, 122 and 172 patients each. The ABiM cohort (N = 350) was selected as an external test set. Inception V3 model was applied to predict gene expression from image tiles for a total of 17,695 genes. The trained models were first applied to the validation set and assessed by Spearman's correlation as well as the coefficient of determination (R^2_{pred}). In brief, regarding the monotonic association, 9,334 (52.75%) genes showed significant correlations with the RNA-seq measurements revealed by a FDR-adjusted $P < 0.05$ (Spearman's correlation). With respect to the proportion of variance explained, 1,026 (5.80%) genes had R^2_{pred} higher than 0.2, among which, 196 had a R^2_{pred} more than 0.3 and less than 0.4; Another 26 genes had a R^2_{pred} higher than 0.4 (**Figure 9A**).

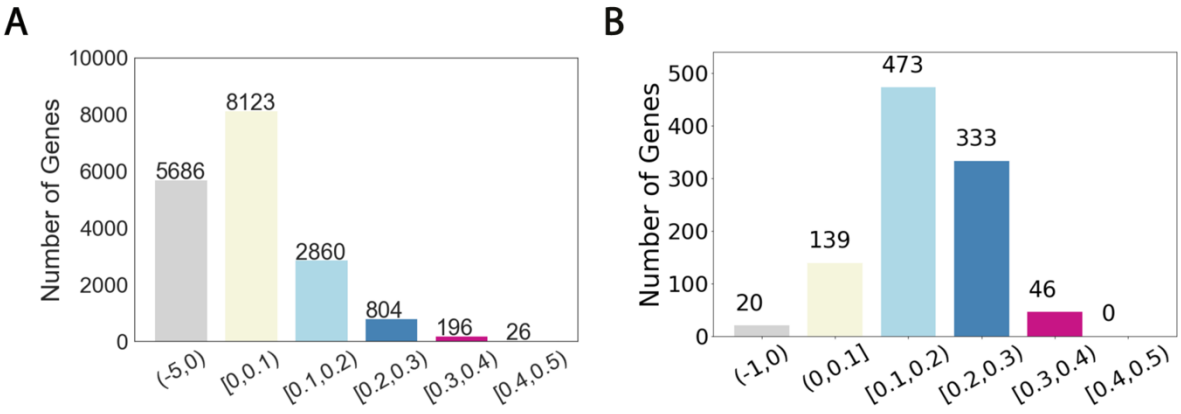

Figure 9. Barplot showing the number of genes within each coefficient of determination interval. **A**. Results in the training set. **B**. Results in the internal test set. Regenerated from Wang & Kartasalo et al. 2021 (140).

The models were filtered by the criteria of $R^2_{pred} > 0.2$ and FDR-adjusted $P < 0.001$ (Spearman's correlation) for final assessment in the internal and external test sets which resulted in a total of 1,011 models.

In the internal test set, 876 out of 1,011 (86.65%) genes can be successfully validated under the criteria of Bonferroni-adjusted P value <0.05 (Spearman's correlation). In the external test set, among the 995 available genes, 908 (91.26%) demonstrated significant relations.

To account for variations in scale, the coefficient of determination was only computed in the internal test set. 333 and 46 models had R^2_{pred} between [0.2, 0.3) and between [0.3, 0.4) intervals, respectively, constituting 479 models with $R^2_{pred} > 0.2$ in the internal test set (**Figure 9B**).

A group of 11 genes (*BIRC5*, *CCNB1*, *CDC20*, *CDCA1*, *CEP55*, *KNTC2*, *MKI67*, *PTTG1*, *RRM2*, *TYMS*, *UBE2C*) from the PAM50 gene panel was previously described as the proliferation signature, whose average expression estimates carry prognostic value in predicting patient survival outcomes (141,142). We therefore incorporated the panel to study the efficacy of predicting proliferation scores with EMO models compared with RNA-seq measurements. A significant and good correlation was revealed in each of the three datasets, with Spearman's ρ ranging from 0.55 to 0.67 (**Figure 10A-C**). A high proliferation score can be an indicator of more proliferative cell activity, as suggested by visual similarity between the heatmaps of predicted proliferation score over WSIs, and the IHC stained Ki67 levels (**Figure 10D**).

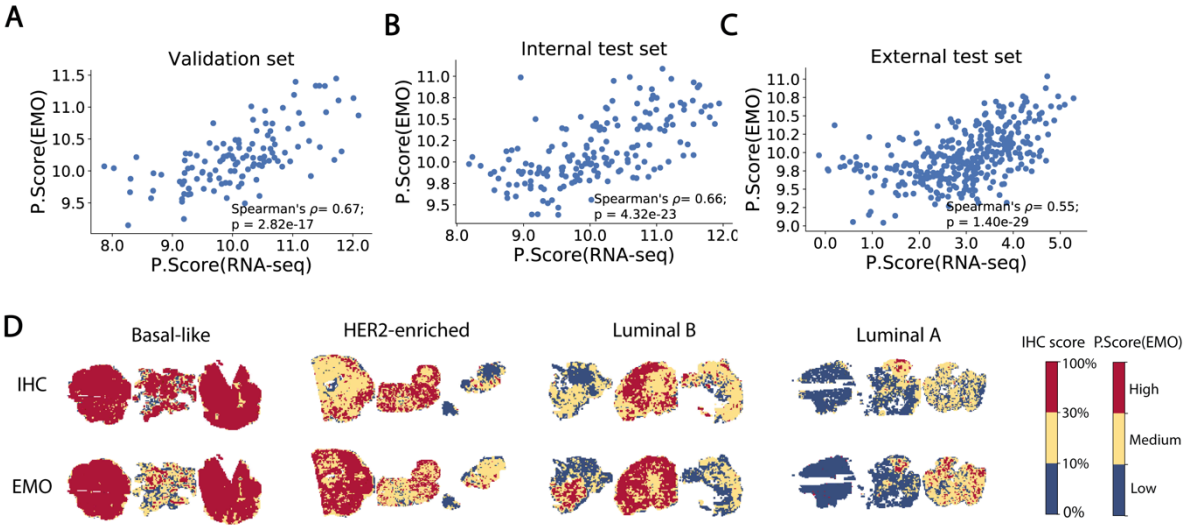

Figure 10. Validation of EMO predicted proliferation score. **A.** Correlation between EMO predictions and RNAseq measurements in the validation set. **B.** Same analysis as in **A**, with results from the internal test set. **C.** Same analysis as in A, with results from the external test set. **D.** Visual similarity between the IHC stained Ki67 abundance and EMO prediction. Three sample images were displayed with respect to each intrinsic molecular subtype. Modified from Wang & Kartasalo *et al*. 2021 (140).

Aside from visual inspection, the ability to predict spatial gene expression value was further validated stringently with ST technique on the Nanostring GeoMX DSP platform (NanoString Technologies). The analysis consisted of a panel of 76 genes, and was performed on 22 tumours. For each tumour, a pair of FFPE sections were retained, with 12 ROIs preselected on both sections. Next, the ST estimation per gene was carried out within individual regions of interest on one section, while the corresponding EMO model was applied to predict the gene expression in the same regions using the other HE stained section. In the end, for each gene, the intra-tumour expression prediction performance by EMO (EMO-spatial) was evaluated with a LME. The model predicted ST estimates by incorporating both EMO-spatial predictions

as fixed effect and slide ID as random effect. A significant association between ST and EMO-spatial was retained for 59 genes (77.63%) (FDR-adjusted $P < 0.05$, likelihood ratio test).

Taken together, the results supported our hypothesis that EMO models can serve as an alternative method in quantifying tumour average gene expression and spatial gene expression variability for a large number of genes.

5.3 STUDY III

Intra-tumour heterogeneity has been widely studied as a risk factor for the prognosis of breast cancer, it is therefore vital to identify patients with higher degree of heterogeneity with implication of treatment resistance. In the current study, we extracted texture based features from spatial gene expression patterns to model the association between intra-tumour gene expression heterogeneity and patient risk of recurrence.

Using a nested cross validation procedure, a Cox proportional hazards regression model with elastic net regularisation was optimised, including the hyperparameters, which enabled intrinsic variable selection and interpretation (non-zero coefficient) and calculation of a risk score per patient. By stratifying patients with the median of risk scores across the training cohort, we demonstrated that the proposed model enabled a separation of patients with distinct risk of recurrence (P = 7.5E-09, log-rank test, **Figure 11A**). When adjusted for age, tumour size, lymph node status, ER status, HER2 status and grade, the prognostic value remained to be significant with an adjusted HR of 2.1 (95% CI 1.3-3.30, P = 0.002, **Figure 11B**).

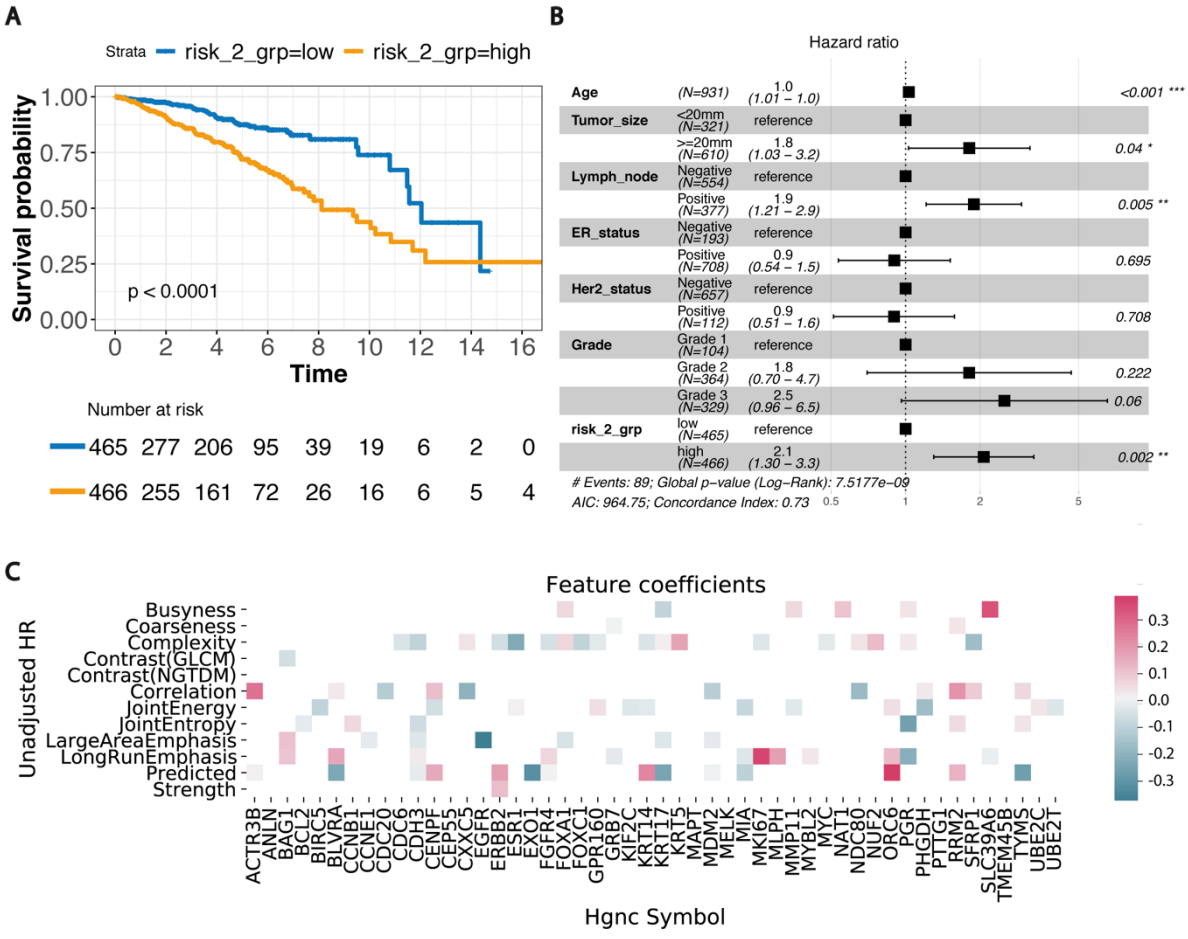

Figure 11. Summary of results in the training set. **A.** Distinct survival probabilities revealed by the Kaplan-Meier curves. **B.** Model risk stratification exhibited an independent prognostic value, via multivariable Cox proportional hazards regression analysis. **C.** Included features and associated coefficients by the optimised model.

Next, in an attempt to investigate the generalisability of the risk stratification model, the model was retrained with all training data and applied to two external test cohorts. A total of 90 features from 45 genes were included by the final model (**Figure 11C**). The predicted expression level of *ORC6* within a WSI was assigned the largest coefficient, indicating a strong predictive effect towards risk of recurrence. The contribution to hazard was not only related to genes but also to specific textures. For instance, the spatial expression pattern of *MKI67* (LongRunEmphasis) was an important factor for patient outcome, but not the expression level *MKI67*, according to the model; Similar importance was observed for the Busyness value of *SLC39A6*. On the other hand, the LongAreaEmphasis of *EGFR* together with the slide level prediction of *EXO1* as well as *TYMS* were also among the most important features but with protective effect towards patient prognosis.

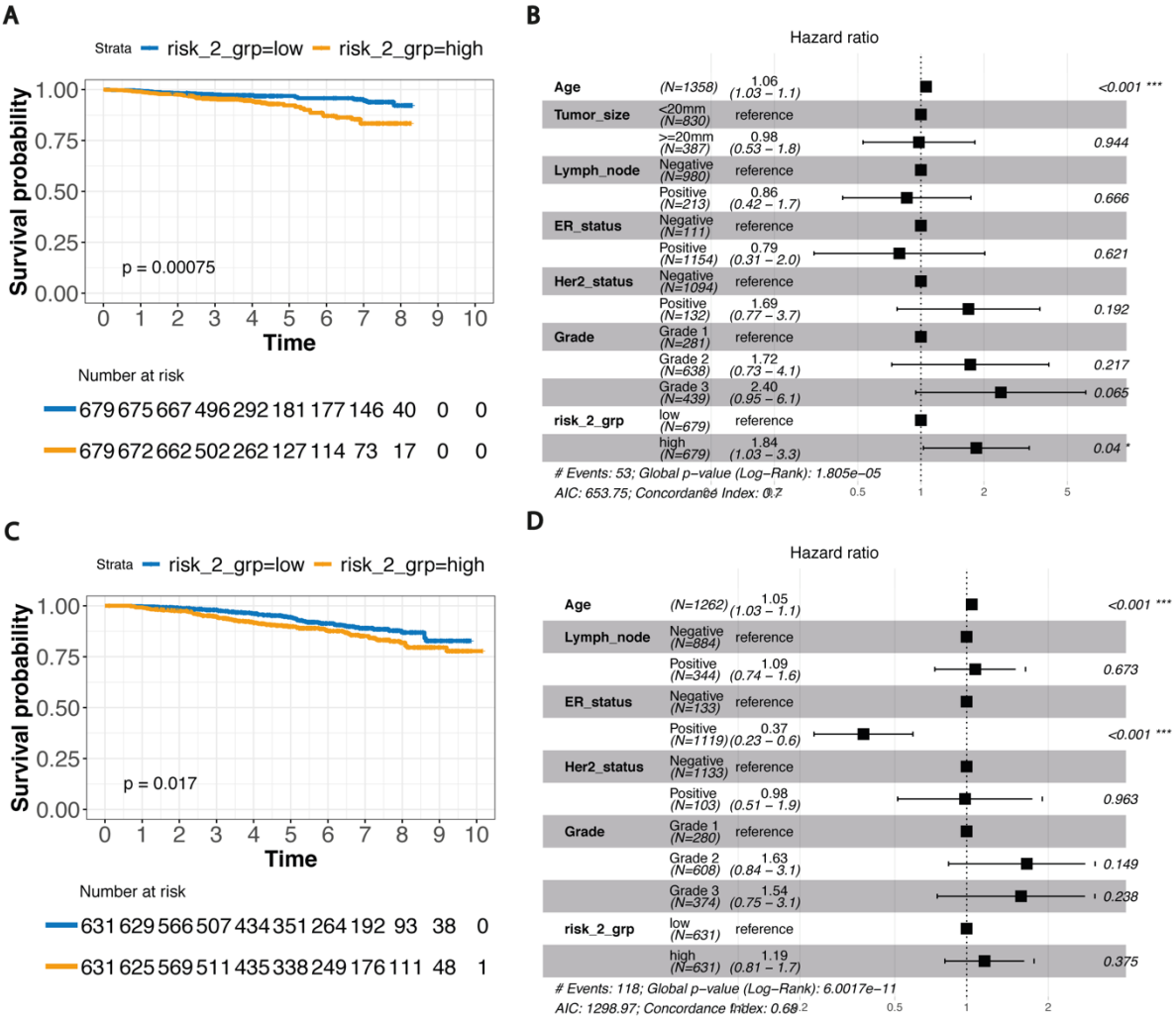

Figure 12. External validation of proposed risk stratification model. **A**. KM curves for SöS-BC-1 data. **B**. Multivariable Cox proportional hazards regression analysis for SöS-BC-1 data. **C**. Same analysis as in **A**, with SCAN-B-Lund data. **D**. Same analysis as in **B**, with SCAN-B-Lund data.

Results from the SöS-BC-1 dataset provided a strong support for the independent prognostic significance of the proposed model, where dichotomised risk groups demonstrated distinctive survival probabilities (P = 0.00075, log-rank test, **Figure 12A**) with an adjusted HR of 1.84 (95% CI: 1.03-3.3, P = 0.04, **Figure 12B**). In the SCAN-B-Lund cohort, the same trend was observed with KM curves (P = 0.017, log-rank test, **Figure 12C**), but the risk factor became insignificant, after adjusting for other covariates (HR: 1.19; 95% CI: 0.81-1.7, P = 0.375, **Figure 12D**). Nevertheless, the point estimate indicated an increased risk of experiencing disease recurrence in the high risk group.

5.4 STUDY IV

In **study IV**, we compared the agreement of two risk stratification tools, one is the Prosigna® test which was developed with integrated information of PAM50 gene panel and tumour size, and aims to provide intrinsic subtypes as well as recurrence risk predictions; The other is the Stratipath Breast model, which is a modified, commercialised product based on the DeepGrade model that was developed in **study I**.

A total of 122 primary breast cancer patients who shared the charactersitics of being postmenopausal, ER-positive, HER2-negative and node negative were included in the study, 88 patients had a valid test results generated from both tools. The tools reached an agreement on 20 and 15 cases, with the assigned risk classes of low and high, respectively (**Table 2**). This resulted in an agreement of 76.09% with a kappa of 0.51, between the low and high groups. Among the Prosigna® test low risk group, six patients were escalated as high risk by Stratipath Breast, while among the high risk group by Prosigna® test, five patients were de-escalated as low risk by Stratipath Breast.

Table 2. Number of patients assigned to each risk group by two tests.

		Prosigna risk group			
Stratipath Breast	**Low**	**Intermediate**	**High**	**Not available**	**All**
	N (%)	N (%)	N (%)	N (%)	N (%)
Low	20 (16.39)	24 (19.67)	5 (4.10)	1 (0.82)	50 (40.98)
High	6 (4.92)	18 (14.75)	15 (12.30)	1 (0.82)	40 (32.79)
Not available	11 (9.02)	17 (13.93)	4 (3.28)	0 (0.00)	32 (26.23)
All	37 (30.33)	59 (48.36)	24 (19.67)	2 (1.64)	122 (100.00)

The Spearman's coefficient of 0.59 between the risk of recurrence (ROR) score from the Prosigna® test and risk score from the Stratipath Breast indicated a good association (**Figure 13A**); the coefficient was 0.47 between Ki67-index and risk score from Stratipath Breast, which also suggested a good correlation (**Figure 13B**). A significantly different distribution of Ki67 index was observed between low and intermediate (P = 2.19E-04, Mann–Whitney test), as well as between low and high risk groups (P = 5.15E-04, Mann–Whitney test) by Prosigna® test but not between intermediate and high risk groups (P = 0.769, Mann–Whitney test, **Figure 13C**);

In comparison, Ki67 was significantly higher in the high risk group, compared with that in low risk group, by Stratipath Breast (P = 2.64E-04, Mann–Whitney test, **Figure 13D**).

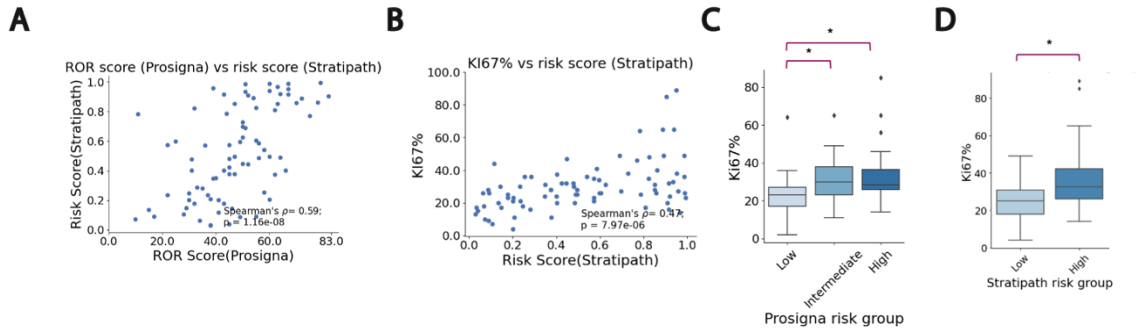

Figure 13. Comparison of risk scores and Ki67 index. **A.** The correlation between risk scores by Prosigna® and Stratipath Breast. **B.** The correlation between Ki67 index and Stratipath Breast risk scores. **C.** Distribution of Ki67 index by Prosigna® risk groups. **D.** Distribution of Ki67 index by Stratipath Breast risk groups.

As grade 2 tumours exhibit larger variation in terms of growth patterns, a risk stratification is therefore considered to be more valuable in this subgroup. Hence, we performed the same comparisons within the NHG2 subgroup. Results indicated a similar association with slight decreases both between two risk scores (Spearman's ρ = 0.45, **Figure 14A**), and between Ki67 index with Stratipath risk score (Spearman's ρ = 0.37, **Figure 14B**). The distribution of Ki67 index between low and intermediate risk group and between low and high risk groups for Prosigna® test continued to be significant, with p-value of 0.02044 and 0.04014 respectively; On the contrary, no difference was found between intermediate and high risk group (P = 0.953, Mann–Whitney test, **Figure 14C**). The distribution of Ki67 was similar between low and high risk groups identified by Stratiapth breast (P = 0.07, Mann–Whitney test, **Figure 14D**).

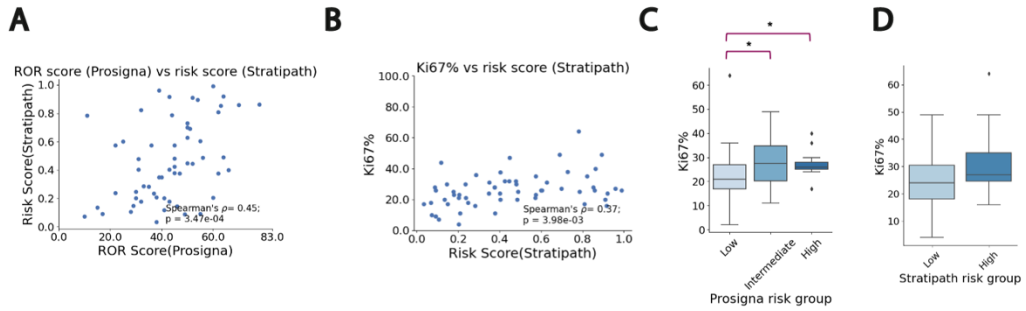

Figure 14. Same analysis as in **Figure 13**, but only included patients with NHG2 tumours.

6 DISCUSSION

Histopathological examination of HE stained FFPE slides has been playing a pivotal role in diagnosing cancer and designing therapeutic strategies. With the advances in target therapies, it is now possible for patients characterised with special subtypes of breast cancers to be spared from chemotherapy and the concomitant toxical side effects. An optimal management of breast cancer relies on an accurate and comprehensive evaluation of tumour characteristics.

The blooming field of machine learning techniques for medical image analysis has made encouraging achievements in the past decade not only in assisting with cancer diagnosis but also in discovering novel biomarkers for refined patient stratifications. The current thesis sits amid the wide variety of deep learning based applications on HE stained images, with the shared aim to improve precision medicine for breast cancer patients.

6.1 INTERPRETATIONS AND CLINICAL IMPLICATIONS

According to the presence of biomarkers, ER+ breast cancer is the largest subgroup, accounting for 84% of the patient population (4). To date, owing to the introduction of hormone therapy, a considerably improved disease outcome has been observed in this subgroup. However, partially due to intra-tumour heterogeneity that arise from the genetic or epigenetic changes, up to one third of ER+ cases can develop treatment resistance to endocrine therapy (143). These observations formulate two critical difficulties in achieving precision medicine, one relates to how to better identify lower risk patients that could be spared from cytotoxic chemotherapy; The other states the importance of discerning patients that are more prone to develop treatment resistance.

For the first aspect, an abundance of patient risk stratification approaches have been developed. OncotypeDX® and Prosigna® are two sequencing based assays that have been made available In Sweden for clinical use to assist with determining the additional benefit of chemotherapy for post-menopausal women with ER+ HER2- tumours without axillary node metastases. Both of the methods have been evaluated with large clinical trials (144) and are endorsed in the international guidelines including American Society of Clinical Oncology (ASCO) (145), and the European Society for Medical Oncology (ECMO) (146). For both neoadjuvant and post operative treatment, the use of chemotherapy in conjunction with either endocrine or anti-HER2 therapy have to be evaluated by cancer intrinsic subtypes and the chemotherapy can be spared for the majority of the cases for Luminal A-like tumours. The Swedish guideline (146) recommends the use of surrogate subtypes for determining luminal cancers which incorporate information including histological grade, IHC measured biomarkers and gene expression profiles. Based on IHC staining, ER+HER2- breast cancers shall be further divided into Luminal A-like and Luminal B-like subtypes, conditioning on tumour grade. Specifically, for the most indecisive grade 2 cases, Ki67 index shall be evaluated. Quantification of Ki67 has been a time consuming task and is associated with marked inter-observer variability that requires scrupulous calibration across sites. As a consequence, the cut-off to discretise Ki67 levels has been under continuous discussion (147)(148) and the current guideline suggests the

use of gene expression profiles to confirm the need of chemotherapy in Ki67 intermediate patients. In **study I**, we proposed a model (DeepGrade) to facilitate with restratification of grade 2 tumours, and its efficacy is validated with an external population based cohort. In addition, in **study IV**, we compared the risk assignments between the Prosigna® test and the commercialised DeepGrade model (Stratipath Breast). In both studies, the Ki67 distribution was not significantly different between two model stratified grade 2 risk groups, indicating that the model was capable of extracting morphological features unrelated to Ki67 expression. The results suggested that on one hand, our proposed model alleviates the influence of inconcordant Ki67 scoring when making the judgement, on the other hand, it provides additional information while predicting patient recurrence risk. Compared to multigene assays, another strength of the proposed model is its convenient application. Unlike genomic tests, the model requires no additional tissue sampling or processing step as it was developed solely with diagnostic HE slides, the effort to train technicians can then be spared. Furthermore, the model is a software that can be easily integrated into a digital pathological system, enabling a scalable and flexible deployment in clinical settings.

Treatment resistance for both endocrine therapy and HER2-target therapy can be ascribed to genetic or epigenetic changes that lead to lowered affinity to drug binding, altered signalling pathways and interfered immune response (143). Extensive studies have pointed out that inter- and intra-tumour heterogeneity play a key role behind the treatment failure (149). It is therefore of central importance to recognise heterogeneous tumours with strong invasive potential. In **study II**, we developed deep learning based models (EMO) to enable a cost-efficient measurement of the spatial resolved gene expression for the whole transcriptome and through various validation modalities, we demonstrated that the models can be applied to evaluate intra-tumour gene expression heterogeneity. We continued to design a machine learning based workflow for this objective in **study IV**, and assessed the value of intra-tumour heterogeneity measurements with patient survival outcome. Through external validation, we showed that the model stratification served as an independent prognostic factor in dichotomising ER+ and HER2-subgroups, suggesting a potential utility to assist with clinical decisions for the application of aggressive therapeutic choices.

6.2 METHODOLOGICAL CONSIDERATIONS

6.2.1 Generalisability

All studies included in the thesis involved training of machine learning models with an abundance of labelled data and providing predictions on samples that were not engaged in the training process. A common emphasis and difficulty in this regime is to ensure the generalisability of the optimised model when applied on unseen data. To address this, in all of the four studies, we employed the optimised model on external test datasets and we showed that in **study I**, the classification performance between grade 1 vs grade 3 remained high with an AUC of 0.907 (95%CI: 0.885-0.930) in the external test set SCAN-B-Lund; more importantly, the efficacy of proposed risk stratification model on grade 2 patients was confirmed in the SCAN-B-Lund cohort, with an adjusted HR of 1.91(95% CI: 1.11-3.29, P =

0.019). In **study II**, the validation of tumour average gene expression prediction performance was carefully conducted in a sequential manner that involves validation, internal and external test data; An orthogonal evaluation of the spatial gene expression variations was also carried out using the spatial transcriptomic technique to demonstrate that CNN models can predict molecular phenotypes with learned morphological patterns from WSIs. In **study III**, we leveraged the study outcome from **study II** and built a risk stratification model on top of intra-tumoural heterogeneity measurements, and the model was also tested in two external cohorts (SöS-BC-1 and SCAN-B-Lund) which both demonstrated superior survival rates in identified low risk groups; In **study IV**, the risk stratification outputs by the commercialised version of DeepGrade model (Stratipath Breast) were compared against the Prosigna® test, without model retraining, this comparison is a direct reflection on the performance of Stratipath Breast in real-world settings. A moderate agreement was found between the two tests, as in concordance with previous studies evaluating different multi-gene risk stratification assays (150)(151), the results suggested that Stratipath Breast as a more cost-effective alternative has the potential to be applied in assessing the applicability of chemotherapy for primary breast cancer patients.

6.2.2 Handling of domain shift and outliers

Deep learning models possess distinctive capacity in memorising feature representations from the training data which guarantees superb learning outcomes in the training set. However, this property may lead to catastrophic failures on unseen data that exhibit different appearances since the latter are likely to have been generated from a distinct origin. Having dissimilar distributions between source data and target data is defined as domain shift, this phenomenon poses considerable challenge towards the deployment of machine learning models in clinical settings. There are several common causes behind the domain shift that are particularly associated with the analysis of histopathological images, including the change of colour profiles, both during slide preparation and WSI digitalisation; the shift in intensity and contrast (152), the existence of disparity between ethnic groups (153)(154).

To address the above issues, it is necessary to train deep NNs with large datasets. However, it takes extraneous efforts to collect, label and process images in extending the datasets, hence, standardisation and augmentation techniques are typically considered during model training to suppress the dissimilarity in distributions between training and test sets and to increase data variations that the model sees. To achieve the above goals, we applied colour normalisation prior to training and data augmentation including rotation and flip during training in the current thesis.

Colour normalisation has been an important aspect in histopathological image analysis with abundance methods proposed ranging from histogram matching to deep learning based models. We employed Macenko's approach in the current thesis as the algorithm is straightforward and easy to compute. Because it is reasonable to assume a single stain matrix is shared within a whole slide image, we slightly modified the algorithm to compute the stain matrix on whole slide level instead of tile level. But the method has drawbacks such as abnormal colour profiles resulting from negative coefficients in the SVD computation. In comparison, spectral

decomposition proposed by Rabinovich and Agarwal (155) made use of non-matrix factorization (NMF) which has the strength to only output positive coefficients, but suffer from inconsistent solutions due to no closed-form solutions. To address this issue, Vahadane (156) developed a method with sparse NMF, by imposing sparseness that effectively reduced the solution space but is computationally expensive.

Furthermore, colour shift is not the only one source of variation, the use of scanners from different manufactures can also result in changes in contrast or brightness. It is therefore worthwhile considering directly minimising discrepancies between domains. Generative Adversarial Networks (GANs) (157) are an example for this objective, but training such a network often requires paired samples such as a slide being scanned by different scanners. An extension to this is the cycleGANs (158)(159) that overcome the limitation and have reached state-of-the-art performance, however, its applicability in clinical settings is still questionable since the algorithm is not transparent and often referred to as a 'blackbox'. More comprehensive comparisons can be found in the review paper (160). A study comparing different colour normalisation methods as a preprocessing step reached the conclusion that methods yielded comparable performance in subsequent analysis, and the choice of evaluation metrics could largely affect the conclusion in selecting a winning method. The results also indicated that, aside from choosing a colour normalisation approach in the preprocessing step, selecting an appropriate model architecture for solving subsequent learning objectives seems to play a more important role (161).

Effective detection of out of distribution (OOD) observations is expected to largely secure a stable performance in real world application. Several methods have been proposed for this purpose, including imposing a confidence measurement mechanism such as adding a separate confidence branch (162); enabling model uncertainty measurement (163); post hoc temperature scaling on the output softmax score that sought to maximise the separation between inliers and outliers (164). In addition, distance based methods targeting either the data or feature space have also been widely adopted such as measuring the Mahalanobis-distance to class centroids in the latent space (165) or comparing the residual distance to the hyperplane spanned by principal directions in the embedding space (166).

Working with histopathology images typically allows for less OOD observation as the inputs are typically confined to HE stained tissue sections. However, bad practice or mistakes in sample preparation can generate artefacts such as bubbles, glues, or abnormal staining artefacts that interfere with model performance. The tissue mask generation and quality control steps integrated in our preprocessing pipeline can largely reduce the influence of such problems, but in practice, we also observed that tissues stained with cytokeratins (Ck5/Ck8-18) by IHC staining revealed a redish colour and the quality control algorithms failed to detect this OOD sample. To increase the robustness of the model, it is worthwhile to incorporate the above mentioned methods in the future.

6.2.3 Model calibration

Another challenge pertains to the calibration of modern neural networks. In a classification setting where the output denotes the probability of belonging to a certain class, training towards a reduced cross entropy loss amounts to minimising the negative log likelihood, or equivalently, maximising the likelihood of the data. It is possible that even if the accuracy remains unchanged, the loss can be further decreased by imposing higher output probabilities. Therefore, it should be noted that overfitting doesn't necessarily appear with decreased accuracy, rather, it can also be associated with less accurate recovering of ground truth distributions. This observation was further supported by experiments that, on the test set, when the model started to overfit to loss, it exhibited a slightly increased classification accuracy at the same time (167).

To address the insufficient calibration of output probabilities, in **study I**, the cutoff for dichotomising NHG 1 and 3 was selected using the Youden's index on an AUC curve to maximise sensitivity and specificity. This method was chosen because the ultimate goal was to provide good restratification of NHG 2 cases with the binary classification model that had been trained against NHG 1 and 3 cases. Alternative methods have been proposed such as binning the predicted probabilities and replacing the outputs within each bin with the average number of positive samples (168), or a generalised form of binning with isotonic regression (169). Alternatively, it is also feasible to calibrate the output probabilities by passing CNN outputs through a sigmoid function, which is referred to as Platt scaling (170),(171).

6.2.4 The choice of modelling

In both **study I** and **II**, the context falls within the scope of weakly supervised learning, where all tiles within one WSI share a class label. During inference, the predictions were generated for each tile, and we obtained the slide level estimation by taking the upper quantile (**study I**) or the mean (**study II**) of tile level predictions.

The tile-to-slide mapping can be achieved with more sophisticated approaches. For instance, with an attention based multiple instance learning, the slide level scores are calculated with weighted average of tile level representations, where the weights are learned with an additional neural network (172). It is however unclear if the network guarantees superior performance (173). Besides, the training is computationally demanding, so studies typically chose to generate feature representations with pre-trained weights from a commonly used benchmarking dataset such as 'ImageNet'. Since the dataset only contains images from unrelated domains, it is then unclear if models learn enough clinically relevant representations.

In the current thesis, the CNNs were trained with randomly sampled tiles, where the spatial relationships across tiles were not considered. Previous studies suggested treating consecutive tiles as a temporal sequence, and training the model with a recurrence neural network (103) or the extended long short-term memory network (174)(105). Likewise, tiles can also be regarded as nodes and connected based on spatial coordinates or similarities. The learning process can then be formulated with a graph-based approach, such as graph convolutional network (175).

7 CONCLUSIONS

We developed deep learning models to stratify breast cancer patients with the aim of facilitating personalised treatment regimens. Models were optimised using HE stained WSIs that are routinely collected for diagnosis purposes, no additional tissue sampling was needed, providing a potential cost-efficient solution compared to molecular diagnosis.

In **study I**, we developed a deep learning based model to further stratify patients with NHG 2 tumours, the findings were validated in external cohorts. The methodology provides additional value that is expected to lower the over- or undertreatment rates for patients with intermediate risks.

In **study II**, we trained deep learning models to estimate gene expressions using HE stained WSIs for the whole transcriptome and verified that over half of the genes can be successfully predicted. We further validated that the model has the capacity to generate spatial gene expression predictions, allowing for large scale studies on intra-tumour heterogeneity with higher efficiency and lower cost.

In **study III**, we quantified intra-tumour gene expression heterogeneity with the models optimised in **study II**. In addition, we developed a Cox proportional hazards regression model to predict patient survival outcome with the integrated heterogeneity measurements. We demonstrated that the model stratification was an independent risk factor for breast cancer prognosis.

In **study IV**, we compared the risk stratification results between the Prosigna® test and the Stratipath Breast, the latter is a commercialised version of the risk stratification model that was developed in **study I**. A moderate agreement was observed, and the results indicated a need for further pathological examination with a larger study population to better compare the risk assessment tools.

8 POINTS OF PERSPECTIVE

As deep learning based models have demonstrated their utility in wide ranges of medical applications, there remains a great deal of unsolved questions that can be potentially investigated with a deep learning model. In addition, more studies are required to ensure the safety of adopting these models in real-world settings. There are several aspects that can be investigated:

From the technical perspective, to better learn both cellular and tubular morphologies that appear under different magnifications, a multi-scale modelling approach shall be adopted to generate more accurate replicates with respect to each subcomponent that constitute the final histological grading score. Training an end-to-end multi-scale model is a desirable approach. Alternatively, one can also imitate the clinical workflow by starting with a single-scale model for coarse slide evaluations and gradually employ other models that correspond to higher magnifications to predict on most suspicious, zoomed in regions. This can be achieved by imposing a model uncertainty measurement.

To ensure the model can be transferred on real-world dataset, it is preferable to integrate outlier detection and model calibration so that domain shift or batch effect can be effectively captured and controlled.

Besides, the model performance shall be further validated on larger cohorts while the efficacy shall be confirmed with randomised controlled trials.

From the clinical perspective, while the patient prognosis has been studied from different aspects, it is now attempting to investigate the potential benefit of adjuvant chemotherapy, adding more powerful evidence to guide clinical decision making.

Furthermore, there are several critical questions that are expected to be tackled with the assistance of computational pathology, such as to identify what characteristics are shared with those *in situ* cancers that are prone to gain invasiveness. Similarly, deep learning based image analysis can be exploited to study the epithelial-mesenchymal transition that relates to tumour progression and metastasis, the gene expression prediction models that have been developed in the current thesis can be employed for this purpose. Another future research area is to obtain a more comprehensive understanding about each individual tumour by effectively integrating radiological and histopathological image data.

9 ACKNOWLEDGEMENTS

While the memory of receiving this PhD offer from Mattias is still vivid, time flies and I've found myself reaching towards the end of this PhD journey. At this point, I cannot articulate how thankful I am to all of the wonderful people that I have met and collaborated with during the past four years, the thesis work would not have been possible without your support. I would particularly like to express my sincere gratefulness to:

My brilliant and kind supervisor **Mattias Rantalainen**, words are never enough to express my gratitude for you. Thank you for introducing me to computational pathology and entrusting me with so many important projects. Thank you for providing me with generous technical and pedagogical support with your broad and in-depth knowledge in machine learning and statistics and thank you for sparing no effort to patiently guide me on the road towards becoming an independent researcher. I have witnessed how much time and effort you have spent to guarantee us with optimal research resources and environment and I feel so fortunate to be able to work in your team.

My co-supervisor **Johan Hartman**, thank you for guiding me into the fancy breast pathology research area and the essential support in all of my studies. I'm incredibly grateful for your stepping up without hesitation to provide generous and invaluable help on the last study. Your enthusiasm, positivity and kindness has been so empowering and inspirational.

My co-supervisor **Martin Eklund**, thank you for involving me in the group activities and discussions when I started my very first journey in deep learning research at MEB. I'm sincerely thankful for such a solid and enjoyable starting point. I've been benefiting greatly from your guidance.

My co-supervisor **Johan Lindberg**, thank you for being such a strong support in the field of bioinformatics. Knowing that I could always turn to you for help, I've had the luxury of working in this unfamiliar domain with much more confidence.

My mentor **Fang Fang**, every time I talked to you, I could receive so much strength and positive energy. I feel so encouraged and inspired thinking of you and your warm support. You are my role model.

I would also like to thank **Aristotelis Tsirigos** for being my opponent, **Johan Lundin, Johan Trygg** and **Barbro Linderholm** for being members of my examination board. Thank you **Mark Clements** for kindly being the chairperson on my dissertation, and your warm-hearted advice and support.

Thank you **Jonas Ludvigsson** and **Alexander Ploner** for taking part in my half-time review and for your encouraging, valuable comments that help to shape the second half of my PhD study.

My collaborators and coauthors for the studies, **Kimmo Kartasalo**, **Stephanie Robertson**, **Balazs Acs**, **Masi Valkonen**, **Christer Larsson**, **Maya Alsheh Ali**, **Emelie Karlsson**, **Emmanouil Sifakis**, **Ioannis Zerdes**, **Dimitrios Salgkamis**, **Theodoros Foukakis**, **Nikolaos Tsiknakis**: the work could not have taken such great shape without your solid expertise. I'm also very grateful for your fast response and valuable inputs. Thank you for giving a high priority to our collaborations although with an already heavy workload on your table. It was a huge joy working with all of you. I would like to especially thank **Carolina Wählby,** for the insightful advice on the breast cancer grading project and for your great speech at many conferences that keeps influencing and inspiring me; **Pekka Ruusuvuori**, for your warm support, valuable advice and insightful opinions on my studies. The successful

collaboration under your supervision has been such a great joy and precious memory in my PhD journey; **Keith Humphreys**, for sharing your valuable inputs and kind advice on my project and my career plan. I'm so lucky to work with you.

My huge gratitude goes to members from Rantalainen's research group: **Bojing Liu**, for our lovely conversations and the laughter that always brings me energy and keeps me moving forward. **Constance Boissin**, you are such a multi-skill person, I enjoy discussing research and life with you so much. Your good taste not only in research but also in all aspects of life always brings sunshine to me. **Daniel Garcia León**, for sharing your talented ideas that never fail to impress and inspire me. **Leslie Solorzano Vargas**, for the extraordinary data visualisations that have always been an immense pleasure to appreciate. **Yanbo Feng**, for your strong expertise and always being there when I need your help. **Philippe Weitz** for those many great collaborations, and the help with proofreading the thesis. **Abhinav Sharma** for our fun discussions about research and life. **Nguyen Thuy Duong Tran** for your exquisite work and cheerful conversations.

Thank you **Paul Lichtenstein** for all of your kind help and quick response throughout this time, thank you for arranging the meeting during the pandemic, making sure I have good attachment to the intended timeline and study outcome.

Thank you **Alessandra Nanni**, for your rich experience and extraordinary valuable advice throughout my PhD study. Thank you for spending time and effort to patiently answer many tricky questions. I'm sincerely grateful for your support.

I would like to thank **Sara Hägg** for always giving me a big, supportive smile, for your numerous warm-hearted advice, for the many fun chats about work and life. I genuinely thank you for your extra support.

A huge appreciation goes to **Cristina Johnell**, for leading MEB through the challenging pandemic and your supportive encouragement in our many cheerful conversations.

Thank you, **Anne-Vibeke Lænkholm** and **Leena Latonen,** for the intriguing research presentations about breast cancer and computational image analysis. Your enthusiasm and expertise are so inspirational to me and thank you for your thoughtful advice on my studies. Thank you **Umair Khan, Sandra Kristiane Sinius Pouplier** and **Dusan Rasic** for the joyful conversations at the ABCAP meeting.

A big thanks to **Xinhe Mao**, **Xiaoyi Ji**, for taking part in my pre-dissertation seminar. Thank you for spending time and effort to help me better prepare my defence.

A special thanks to **Örjan Smedby** and **Rodrigo Moreno** for kindly inviting me to your seminars and warmly welcoming me in your group activities. I have learned a lot from your brilliant research team and I enjoyed the enthusiastic discussion so much with the talented researchers **Atif Mehmood**, **Jingru Fu** and **Fabian Sinzinger**.

I would like to thank **Jiangwei Sun**, for the treasured memories that I can never imagine without you inviting me to write the 'Big Data' book with you. I would also like to thank **Lin Li, Weiwei Bian**, **Xia Li** and **Zheng Ning** for the joyful, meaningful and unforgettable experience that we had together while writing the book.

Many thanks to **Chen Wang**, **Xinge Li**, **Ji Zhang, Muyi Yang, Xinsong Chen** and **Shuang Hao** for all the help with defence application and thoughtful career advice.

I want to also thank the amazing MEBers and my friends outside MEB: **Anders Forss**, **Bjorn Roelstraete**, **Peter Lind**, **Yun Du**, **Yufeng Chen, Lu Pan**, **Shengxin Liu**, **Yuanhang Yang**,

Erwei Zeng, **Cen Chen**, **Ge Bai**, **Xiaoying Kang**, **Le Zhang**, **Bowen Tang**, **Mengping Zhou**, **Honghui Yao, Nanbo Zhu**, **Chenxi Qin**, **Tong Gong**, **Shuyang Yao**, **Weiyao Yin**, **Jie Song**, **Henrik Olsson**, **Mattias Hammarström**, **Yuliya Leontyeva**, **Nita Mulliqi**, **Nurgul Batyrbekova**, **Yuqi Zhang**, **Yuying Li**, **Mujin Ye**, **Hilda Björk Danielsdottir**, **Jet Termorshuizen**, **Aleksandra Kanina** and **Miriam Martini** for our inspiring and cheerful chat about life and work.

A special thank you to **Panagiotis Papapetrou**, my journey in data science started from DAMIs, thank you for holding the incredible courses. Your enthusiasm in research and teaching has provided a profound influence to me.

A big thanks goes to **Isak Samsten**, for the wonderful data science course that has fundamentally shaped my approach to research. I've learned so much from you and I was so fortunate to have your strong recommendation letter for my doctoral application. Thank you for your warm support.

I would also like to thank my old friends from the HI program: **Hongxia Zhang**, **Sachiko Lim, Hamed Khodabakhshi** and **Hayat Ibrahim**, for being wonderful classmates, for the jokes, advice, support and sweet memories.

I wish to give a special thanks to the group of amazing people from the Stratipath team, **Fredrik Wetterhall**, **Annica Jämtén Ericsson**, **Lars Lengquist**, **Jennie Ousbäck**, and of course the fabulous R&D team **Johnson Ho**, **Sandy Kang Lövgren**, **Binbin Su** and **Kajsa Ledesma Eriksson**, for turning the research idea to real-world applications that bring true benefits to patients. Your enthusiasm, dedication and strong domain expertise has been such a powerful motivator to me.

A unique thank you goes to my childhood friends **Shuyang Liu**, **Bing Xie**, **Miaomiao Sun** and **Yue Jiang**, for your warm support and comfort and **Bo Wang** for the high dose of jokes.

最重要的是谢谢我的家人们。感谢爸爸，妈妈和姥姥给予我无条件的爱与支持，感谢您们耐心的安慰与鼓励，也感谢您们为我带来的快乐和温暖；感谢老公韵章给我的爱与帮助。你是我的榜样，也是我最亲密的朋友。

10 REFERENCES

1. Russo J, Russo IH. Development of the human breast [Internet]. Vol. 49, Maturitas. 2004. p. 2–15. Available from: http://dx.doi.org/10.1016/j.maturitas.2004.04.011

2. Dooijeweert C, Diest PJ, Willems SM, Kuijpers CCH, Wall E, Overbeek LIH, et al. Significant inter- and intra-laboratory variation in grading of invasive breast cancer: A nationwide study of 33,043 patients in the Netherlands [Internet]. Vol. 146, International Journal of Cancer. 2020. p. 769–80. Available from: http://dx.doi.org/10.1002/ijc.32330

3. Page DL. Interobserver Agreement and Reproducibility in Classification of Invasive Breast Carcinoma: An NCI Breast Cancer Family Registry Study [Internet]. Vol. 18, Breast Diseases: A Year Book Quarterly. 2007. p. 67. Available from: http://dx.doi.org/10.1016/s1043-321x(07)80050-6

4. Acs B, Fredriksson I, Rönnlund C, Hagerling C, Ehinger A, Kovács A, et al. Variability in Breast Cancer Biomarker Assessment and the Effect on Oncological Treatment Decisions: A Nationwide 5-Year Population-Based Study. Cancers [Internet]. 2021 Mar 9;13(5). Available from: http://dx.doi.org/10.3390/cancers13051166

5. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012 Jan 18;481(7381):306–13.

6. Polyak K, Shipitsin M, Campbell-Marrotta LL, Bloushtain-Qimron N, Park SY. Breast tumor heterogeneity: causes and consequences [Internet]. Vol. 11, Breast Cancer Research. 2009. Available from: http://dx.doi.org/10.1186/bcr2279

7. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May 28;521(7553):436–44.

8. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries [Internet]. Vol. 71, CA: A Cancer Journal for Clinicians. 2021. p. 209–49. Available from: http://dx.doi.org/10.3322/caac.21660

9. National Board of Health and Welfare (Socialstyrelsen). Statistics on Cancer Incidence 2020. Available online: https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/statistik/2021-12-7701.pdf (accessed on 31 August 2022).

10. Sun YS, Zhao Z, Yang ZN, Xu F, Lu HJ, Zhu ZY, et al. Risk Factors and Preventions of Breast Cancer. Int J Biol Sci. 2017 Nov 1;13(11):1387–97.

11. Nyström L, Andersson I, Bjurstam N, Frisell J, Nordenskjöld B, Rutqvist LE. Long-term effects of mammography screening: updated overview of the Swedish randomised trials. Lancet. 2002 Mar 16;359(9310):909–19.

12. Tabár L, Vitak B, Chen HH, Duffy SW, Yen MF, Chiang CF, et al. The Swedish Two-County Trial twenty years later. Updated mortality results and new insights from long-term follow-up. Radiol Clin North Am. 2000 Jul;38(4):625–51.

13. Field AS, Schmitt F, Vielh P. IAC Standardized Reporting of Breast Fine-Needle Aspiration Biopsy Cytology. Acta Cytol. 2017;61(1):3–6.

14. Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, Hu Z, et al. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. Clin Cancer Res. 2004 Aug 15;10(16):5367–74.

15. Cserni G, Sejben A. Grading Ductal Carcinoma In Situ (DCIS) of the Breast – What's Wrong with It? [Internet]. Vol. 26, Pathology & Oncology Research. 2020. p. 665–71. Available from: http://dx.doi.org/10.1007/s12253-019-00760-8

16. Elston CW, Ellis IO. 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. 1991 Nov;19(5):403–10.

17. Lanjewar S, Patil P, Fineberg S. Pathologic reporting practices for breast cancer specimens after neoadjuvant chemotherapy—a survey of pathologists in academic institutions across the United States [Internet]. Vol. 33, Modern Pathology. 2020. p. 91–8. Available from: http://dx.doi.org/10.1038/s41379-019-0326-5

18. Meyer JS, for the Cooperative Breast Cancer Tissue Resource, Alvarez C, Milikowski C, Olson N, Russo I, et al. Erratum: Breast carcinoma malignancy grading by Bloom–Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index [Internet]. Vol. 18, Modern Pathology. 2005. p. 1649–1649. Available from: http://dx.doi.org/10.1038/modpathol.3800479

19. Thomas JSJ, St J Thomas J, Kerr GR, Jack WJL, Campbell F, McKay L, et al. Histological grading of invasive breast carcinoma â□□ a simplification of existing methods in a large conservation series with long-term follow-up [Internet]. Vol. 55, Histopathology. 2009. p. 724–31. Available from: http://dx.doi.org/10.1111/j.1365-2559.2009.03429.x

20. Fitzgibbons PL, Page DL, Weaver D, Thor AD, Allred DC, Clark GM, et al. Prognostic factors in breast cancer. College of American Pathologists Consensus Statement 1999. Arch Pathol Lab Med. 2000 Jul;124(7):966–78.

21. Hayes DF, Isaacs C, Stearns V. Prognostic factors in breast cancer: current and new predictors of metastasis. J Mammary Gland Biol Neoplasia. 2001 Oct;6(4):375–92.

22. van Dooijeweert C, van Diest PJ, Ellis IO. Grading of invasive breast carcinoma: the way forward. Virchows Arch. 2022 Jan;480(1):33–43.

23. Ellis IO, Coleman D, Wells C, Kodikara S, Paish EM, Moss S, et al. Impact of a national external quality assessment scheme for breast pathology in the UK. J Clin Pathol. 2006 Feb;59(2):138–45.

24. Dalton LW, Pinder SE, Elston CE, Ellis IO, Page DL, Dupont WD, et al. Histologic grading of breast cancer: linkage of patient outcome with level of pathologist agreement. Mod Pathol. 2000 Jul;13(7):730–5.

25. van Dooijeweert C, van Diest PJ, Willems SM, Kuijpers CCHJ, van der Wall E, Overbeek LIH, et al. Significant inter- and intra-laboratory variation in grading of invasive breast cancer: A nationwide study of 33,043 patients in the Netherlands. Int J Cancer. 2020 Feb 1;146(3):769–80.

26. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst. 2006 Feb 15;98(4):262–72.

27. McCart Reed AE, Kutasovic JR, Lakhani SR, Simpson PT. Invasive lobular carcinoma of the breast: morphology, biomarkers and 'omics. Breast Cancer Res. 2015 Jan 30;17:12.

28. Rakha EA, El-Sayed ME, Lee AHS, Elston CW, Grainge MJ, Hodi Z, et al. Prognostic significance of Nottingham histologic grade in invasive breast carcinoma. J Clin Oncol. 2008 Jul 1;26(19):3153–8.

29. Saadatmand S, Bretveld R, Siesling S, Tilanus-Linthorst MMA. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173,797 patients. BMJ. 2015 Oct 6;351:h4901.

30. Foulkes WD, Grainge MJ, Rakha EA, Green AR, Ellis IO. Tumor size is an unreliable predictor of prognosis in basal-like breast cancers and does not correlate closely with lymph node status [Internet]. Vol. 117, Breast Cancer Research and Treatment. 2009. p. 199–204. Available from: http://dx.doi.org/10.1007/s10549-008-0102-6

31. Dent R, Hanna WM, Trudeau M, Rawlinson E, Sun P, Narod SA. Time to disease recurrence in basal-type breast cancers: effects of tumor size and lymph node status. Cancer. 2009 Nov 1;115(21):4917–23.

32. Veronesi U, Salvadori B, Luini A, Greco M, Saccozzi R, del Vecchio M, et al. Breast conservation is a safe method in patients with small cancer of the breast. Long-term results of three randomised trials on 1,973 patients. Eur J Cancer. 1995 Sep;31A(10):1574–9.

33. Veronesi U, Galimberti V, Zurrida S, Merson M, Greco M, Luini A. Prognostic significance of number and level of axillary node metastases in breast cancer. The Breast. 1993 Dec 1;2(4):224-8.

34. Hua G, Jégou H. Computer Vision – ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings. Springer; 2016. 919 p.

35. Swenson KK, Nissen MJ, Ceronsky C, Swenson L, Lee MW, Tuttle TM. Comparison of side effects between sentinel lymph node and axillary lymph node dissection for breast cancer. Ann Surg Oncol. 2002 Oct;9(8):745–53.

36. ASCO Guideline Recommendations for Sentinel Lymph Node Biopsy in Early-Stage Breast Cancer: Guideline Summary [Internet]. Vol. 1, Journal of Oncology Practice. 2005. p. 134–6. Available from: http://dx.doi.org/10.1200/jop.1.4.134

37. Carter CL, Allen C, Henson DE. Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. Cancer. 1989 Jan 1;63(1):181–7.

38. Gentil J, Dabakuyo TS, Ouedraogo S, Poillot ML, Dejardin O, Arveux P. For patients with breast cancer, geographic and social disparities are independent determinants of access to specialized surgeons. A eleven-year population-based multilevel analysis. BMC Cancer. 2012 Aug 13;12:351.

39. Moorman PG, Jones BA, Millikan RC, Hall IJ, Newman B. Race, anthropometric factors, and stage at diagnosis of breast cancer. Am J Epidemiol. 2001 Feb 1;153(3):284–91.

40. Wani SQ, Khan T, Wani SY, Koka AH, Arshad S, Rafiq L, et al. Clinicoepidemiological analysis of female breast cancer patients in Kashmir. J Cancer Res Ther. 2012 Jul;8(3):389–93.

41. Cardoso F, Fallowfield L, Costa A, Castiglione M, Senkus E. Locally recurrent or metastatic breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment

and follow-up [Internet]. Vol. 22, Annals of Oncology. 2011. p. vi25–30. Available from: http://dx.doi.org/10.1093/annonc/mdr372

42. Paterni I, Granchi C, Katzenellenbogen JA, Minutolo F. Estrogen receptors alpha (ERα) and beta (ERβ): subtype-selective ligands and clinical potential. Steroids. 2014 Nov;90:13–29.

43. Allison KH, Hammond MEH, Dowsett M, McKernin SE, Carey LA, Fitzgibbons PL, et al. Estrogen and Progesterone Receptor Testing in Breast Cancer: ASCO/CAP Guideline Update. J Clin Oncol. 2020 Apr 20;38(12):1346–66.

44. Hoefnagel LDC, Moelans CB, Meijer SL, van Slooten HJ, Wesseling P, Wesseling J, et al. Prognostic value of estrogen receptor α and progesterone receptor conversion in distant breast cancer metastases. Cancer. 2012 Oct 15;118(20):4929–35.

45. Bentzon N, Düring M, Rasmussen BB, Mouridsen H, Kroman N. Prognostic effect of estrogen receptor status across age in primary breast cancer. Int J Cancer. 2008 Mar 1;122(5):1089–94.

46. Allison KH, Hammond MEH, Dowsett M, McKernin SE, Carey LA, Fitzgibbons PL, et al. Estrogen and Progesterone Receptor Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Guideline Update [Internet]. Vol. 144, Archives of Pathology & Laboratory Medicine. 2020. p. 545–63. Available from: http://dx.doi.org/10.5858/arpa.2019-0904-sa

47. Sopik V, Sun P, Narod SA. The prognostic effect of estrogen receptor status differs for younger versus older breast cancer patients. Breast Cancer Res Treat. 2017 Sep;165(2):391–402.

48. Giulianelli S, Vaqué JP, Soldati R, Wargon V, Vanzulli SI, Martins R, et al. Estrogen receptor alpha mediates progestin-induced mammary tumor growth by interacting with progesterone receptors at the cyclin D1/MYC promoters. Cancer Res. 2012 May 1;72(9):2416–27.

49. Jalava P, Kuopio T, Huovinen R, Laine J, Collan Y. Immunohistochemical staining of estrogen and progesterone receptors: aspects for evaluating positivity and defining the cutpoints. Anticancer Res. 2005 May;25(3c):2535–42.

50. Viale G, Regan MM, Maiorano E, Mastropasqua MG, Dell'Orto P, Rasmussen BB, et al. Prognostic and Predictive Value of Centrally Reviewed Expression of Estrogen and Progesterone Receptors in a Randomized Trial Comparing Letrozole and Tamoxifen Adjuvant Therapy for Postmenopausal Early Breast Cancer: BIG 1-98 [Internet]. Vol. 25, Journal of Clinical Oncology. 2007. p. 3846–52. Available from: http://dx.doi.org/10.1200/jco.2007.11.9453

51. Cui X, Schiff R, Arpino G, Osborne CK, Lee AV. Biology of progesterone receptor loss in breast cancer and its implications for endocrine therapy. J Clin Oncol. 2005 Oct 20;23(30):7721–35.

52. Rakha EA, Pinder SE, Bartlett JMS, Ibrahim M, Starczynski J, Carder PJ, et al. Updated UK Recommendations for HER2 assessment in breast cancer. J Clin Pathol. 2015 Feb;68(2):93–9.

53. Badve S, Dabbs DJ, Schnitt SJ, Baehner FL, Decker T, Eusebi V, et al. Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. Mod Pathol. 2011 Feb;24(2):157–67.

54. Leung SCY, Nielsen TO, Zabaglo LA, Arun I, Badve SS, Bane AL, et al. Analytical validation of a standardised scoring protocol for Ki67 immunohistochemistry on breast cancer excision whole sections: an international multicentre collaboration. Histopathology. 2019 Aug;75(2):225–35.

55. Leung SCY, Nielsen TO, Zabaglo L, Arun I, Badve SS, Bane AL, et al. Analytical validation of a standardized scoring protocol for Ki67: phase 3 of an international multicenter collaboration. NPJ Breast Cancer. 2016 May 18;2:16014.

56. Denkert C, Loibl S, Müller BM, Eidtmann H, Schmitt WD, Eiermann W, et al. Ki67 levels as predictive and prognostic parameters in pretherapeutic breast cancer core biopsies: a translational investigation in the neoadjuvant GeparTrio trial. Ann Oncol. 2013 Nov;24(11):2786–93.

57. Souchon R, Wenz F, Sedlmayer F, Budach W, Dunst J, Feyer P, et al. DEGRO practice guidelines for palliative radiotherapy of metastatic breast cancer: bone metastases and metastatic spinal cord compression (MSCC). Strahlenther Onkol. 2009 Jul;185(7):417–24.

58. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, Zackrisson S, Senkus E. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of oncology. 2019 Aug 1;30(8):1194-220.

59. Tamoxifen After Adjuvant Chemotherapy for Premenopausal Women With Lymph Node-Positive Breast Cancer: International Breast Cancer Study Group Trial 13-93 [Internet]. Vol. 24, Journal of Clinical Oncology. 2006. p. 1332–41. Available from: http://dx.doi.org/10.1200/jco.2005.03.0783

60. Farmer H, McCabe N, Lord CJ, Tutt ANJ, Johnson DA, Richardson TB, et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature. 2005 Apr 14;434(7035):917–21.

61. Patel M, Nowsheen S, Maraboyina S, Xia F. The role of poly(ADP-ribose) polymerase inhibitors in the treatment of cancer and methods to overcome resistance: a review. Cell Biosci. 2020 Mar 11;10:35.

62. Herbst RS, Soria JC, Kowanetz M, Fine GD, Hamid O, Gordon MS, et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. Nature. 2014 Nov 27;515(7528):563–7.

63. Schmid P, Chui SY, Emens LA. Atezolizumab and Nab-Paclitaxel in Advanced Triple-Negative Breast Cancer. Reply. N Engl J Med. 2019 Mar 7;380(10):987–8.

64. Lyons TG. Targeted Therapies for Triple-Negative Breast Cancer [Internet]. Vol. 20, Current Treatment Options in Oncology. 2019. Available from: http://dx.doi.org/10.1007/s11864-019-0682-x

65. Liu SV, Melstrom L, Yao K, Russell CA, Sener SF. Neoadjuvant therapy for breast cancer [Internet]. Vol. 101, Journal of Surgical Oncology. 2010. p. 283–91. Available from: http://dx.doi.org/10.1002/jso.21446

66. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009 Mar 10;27(8):1160–7.

67. Harris LN, Ismaila N, McShane LM, Hayes DF. Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline Summary [Internet]. Vol. 12, Journal of Oncology Practice. 2016. p. 384–9. Available from: http://dx.doi.org/10.1200/jop.2016.010868

68. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004 Dec 30;351(27):2817–26.

69. Wang M, Klevebring D, Lindberg J, Czene K, Grönberg H, Rantalainen M. Determining breast cancer histological grade from RNA-sequencing data. Breast Cancer Res. 2016 May 10;18(1):48.

70. Pareja F, Marchiò C, Geyer FC, Weigelt B, Reis-Filho JS. Breast Cancer Heterogeneity: Roles in Tumorigenesis and Therapeutic Implications [Internet]. Vol. 9, Current Breast Cancer Reports. 2017. p. 34–44. Available from: http://dx.doi.org/10.1007/s12609-017-0233-z

71. Fan M, Xia P, Clarke R, Wang Y, Li L. Radiogenomic signatures reveal multiscale intratumour heterogeneity associated with biological functions and survival in breast cancer. Nat Commun. 2020 Sep 25;11(1):4861.

72. Gerlinger M, Catto JW, Orntoft TF, Real FX, Zwarthoff EC, Swanton C. Intratumour heterogeneity in urologic cancers: from molecular evidence to clinical implications. Eur Urol. 2015 Apr;67(4):729–37.

73. López-Carrasco A, Berbegall AP, Martín-Vañó S, Blanquer-Maceiras M, Castel V, Navarro S, et al. Intra-Tumour Genetic Heterogeneity and Prognosis in High-Risk Neuroblastoma. Cancers [Internet]. 2021 Oct 15;13(20). Available from: http://dx.doi.org/10.3390/cancers13205173

74. Liu B, Li Y, Zhang L. Analysis and Visualization of Spatial Transcriptomic Data. Front Genet. 2021;12:785290.

75. Casasent AK, Schalck A, Gao R, Sei E, Long A, Pangburn W, et al. Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. Cell. 2018 Jan 11;172(1-2):205–17.e12.

76. Cang Z, Nie Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. Nat Commun. 2020 Apr 29;11(1):2084.

77. Berglund E, Maaskola J, Schultz N, Friedrich S, Marklund M, Bergenstråhle J, et al. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. Nat Commun. 2018 Jun 20;9(1):2419.

78. Moncada R, Barkley D, Wagner F, Chiodin M, Devlin JC, Baron M, et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas [Internet]. Vol. 38, Nature Biotechnology. 2020. p. 333–42. Available from: http://dx.doi.org/10.1038/s41587-019-0392-8

79. Schmauch B, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nat Commun. 2020 Aug 3;11(1):3877.

80. Fu Y, Jung AW, Torne RV, Gonzalez S, Vöhringer H, Shmatko A, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat Cancer. 2020 Aug;1(8):800–10.

81. Website [Internet]. Available from: https://doi.org/10.48550/arXiv.2104.09310

82. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018 Oct;24(10):1559–67.

83. Wang Y, Kartasalo K, Weitz P, Ács B, Valkonen M, Larsson C, et al. Predicting Molecular Phenotypes from Histopathology Images: A Transcriptome-Wide Expression–Morphology Analysis in Breast Cancer [Internet]. Vol. 81, Cancer Research. 2021. p. 5115–26. Available from: http://dx.doi.org/10.1158/0008-5472.can-21-0482

84. He B, Bergenstråhle L, Stenbeck L, Abid A, Andersson A, Borg Å, et al. Integrating spatial gene expression and breast tumour morphology via deep learning. Nat Biomed Eng. 2020 Aug;4(8):827–34.

85. Nam S, Chong Y, Jung CK, Kwak TY, Lee JY, Park J, et al. Introduction to digital pathology and computer-aided pathology. J Pathol Transl Med. 2020 Mar;54(2):125–34.

86. Fuchs TJ, Buhmann JM. Computational pathology: challenges and promises for tissue analysis. Comput Med Imaging Graph. 2011 Oct;35(7-8):515–30.

87. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: Challenges and opportunities. Med Image Anal. 2016 Oct;33:170–5.

88. Dick S. Artificial intelligence.

89. Haralick RM, Shanmugam K, Dinstein I 'hak. Textural Features for Image Classification [Internet]. Vol. SMC-3, IEEE Transactions on Systems, Man, and Cybernetics. 1973. p. 610–21. Available from: http://dx.doi.org/10.1109/tsmc.1973.4309314

90. Komura D, Ishikawa S. Machine Learning Methods for Histopathological Image Analysis [Internet]. Vol. 16, Computational and Structural Biotechnology Journal. 2018. p. 34–42. Available from: http://dx.doi.org/10.1016/j.csbj.2018.01.001

91. Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep Learning in Medical Image Analysis. Adv Exp Med Biol. 2020;1213:3–21.

92. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. 1943. Bull Math Biol. 1990;52(1-2):99–115; discussion 73–97.

93. Huxley J. The Organization of Behavior: A Neuropsychological Theory. 1949.

94. Haykin S. Neural networks: a comprehensive foundation. Prentice Hall PTR; 1994.

95. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. nature. 1986 Oct;323(6088):533-6.

96. Bishop CM. Pattern Recognition and Machine Learning. Springer Verlag; 2006. 738 p.

97. LeCun Y, Kavukcuoglu K, Farabet C. Convolutional networks and applications in vision [Internet]. Proceedings of 2010 IEEE International Symposium on Circuits and Systems. 2010. Available from: http://dx.doi.org/10.1109/iscas.2010.5537907

98. Jun S. Bayesian Inference and Learning for Neural Networks and Deep Learning [Internet]. 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). 2020. Available from: http://dx.doi.org/10.1109/icaiic48513.2020.9065256

99. Spanhol FA, Oliveira LS, Petitjean C, Heutte L. Breast cancer histopathological image classification using Convolutional Neural Networks [Internet]. 2016 International Joint Conference on Neural Networks (IJCNN). 2016. Available from: http://dx.doi.org/10.1109/ijcnn.2016.7727519

100. Basha SHS, Shabbeer Basha SH, Dubey SR, Pulabaigari V, Mukherjee S. Impact of fully connected layers on performance of convolutional neural networks for image classification [Internet]. Vol. 378, Neurocomputing. 2020. p. 112–9. Available from: http://dx.doi.org/10.1016/j.neucom.2019.10.008

101. Nguyen T, Raghu M, Kornblith S. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. arXiv preprint arXiv:2010.15327. 2020 Oct 29.

102. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA. 2017 Dec 12;318(22):2199–210.

103. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019 Aug;25(8):1301–9.

104. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. Lancet Oncol. 2020 Feb;21(2):222–32.

105. Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C, et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci Rep. 2018 Feb 21;8(1):3395.

106. Skrede OJ, De Raedt S, Kleppe A, Hveem TS, Liestøl K, Maddison J, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. Lancet. 2020 Feb 1;395(10221):350–60.

107. Chen RJ, Lu MY, Weng WH, Chen TY, Williamson DFK, Manz T, et al. Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images [Internet]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021. Available from: http://dx.doi.org/10.1109/iccv48922.2021.00398

108. Chen M, Zhang B, Topatana W, Cao J, Zhu H, Juengpanich S, et al. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning [Internet]. Vol. 4, npj Precision Oncology. 2020. Available from: http://dx.doi.org/10.1038/s41698-020-0120-3

109. Bilal M, Raza SEA, Azam A, Graham S, Ilyas M, Cree IA, et al. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. Lancet Digit Health. 2021 Dec;3(12):e763–72.

110. Qu H, Zhou M, Yan Z, Wang H, Rustgi VK, Zhang S, et al. Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning. NPJ Precis Oncol. 2021 Sep 23;5(1):87.

111. Kather JN, Heij LR, Grabsch HI, Kooreman LFS, Loeffler C, Echle A, et al. Pan-cancer image-based detection of clinically actionable genetic alterations [Internet]. Available from: http://dx.doi.org/10.1101/833756

112. Fu Y, Jung AW, Torne RV, Gonzalez S, Vöhringer H, Shmatko A, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis [Internet]. Available from: http://dx.doi.org/10.1101/813543

113. Holm J, Humphreys K, Li J, Ploner A, Cheddad A, Eriksson M, et al. Risk factors and tumor characteristics of interval cancers by mammographic density. J Clin Oncol. 2015 Mar 20;33(9):1030–7.

114. Rantalainen M, Klevebring D, Lindberg J, Ivansson E, Rosin G, Kis L, et al. Sequencing-based breast cancer diagnostics as an alternative to routine biomarkers. Sci Rep. 2016 Nov 30;6:38037.

115. Emilsson L, Lindahl B, Köster M, Lambe M, Ludvigsson JF. Review of 103 Swedish Healthcare Quality Registries. J Intern Med. 2015 Jan;277(1):94–136.

116. Löfgren L, Eloranta S, Krawiec K, Asterkvist A, Lönnqvist C, Sandelin K, et al. Validation of data quality in the Swedish National Register for Breast Cancer. BMC Public Health. 2019 May 2;19(1):495.

117. Saal LH, Vallon-Christersson J, Häkkinen J, Hegardt C, Grabau D, Winter C, 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. 2015 Feb 2;7(1):20.

118. Brueffer C, Vallon-Christersson J, Grabau D, Ehinger A, Häkkinen J, Hegardt C, et al. Clinical Value of RNA Sequencing-Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network-Breast Initiative. JCO Precis Oncol [Internet]. 2018 Mar 9;2. Available from: http://dx.doi.org/10.1200/PO.17.00135

119. Otsu N. A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics. 1979 Jan;9(1):62-6.

120. Reprinted from Wang Y, Acs B, Robertson S, Liu B, Solorzano L, Wählby C, et al. Improved breast cancer histological grading using deep learning. Ann Oncol. 2022 Jan;33(1):89–98. Copyright Elsevier Reprinted from Publication title, Vol /edition number, Author(s), with permission from Elsevier [author's material]

121. Pech-Pacheco JL, Cristobal G, Chamorro-Martinez J, Fernandez-Valdivia J. Diatomautofocusing in brightfield microscopy: a comparative study [Internet]. Proceedings 15thInternational Conference on Pattern Recognition. ICPR-2000.

122. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, et al. Amethod for normalizing histology slides for quantitative analysis [Internet]. 2009 IEEEInternational Symposium on Biomedical Imaging: From Nano to Macro. 2009.Available from: http://dx.doi.org/10.1109/isbi.2009.5193250.

123. Fawcett J, Scott J. A rapid and precise method for the determination of urea. Journal of clinical pathology. 1960 Mar 1;13(2):156-9.

124. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. Anal Quant Cytol Histol. 2001 Aug;23(4):291–9.

125. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 2818-2826).

126. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015 Dec;115(3):211-52.

127. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014 Dec 22.

128. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32-5.

129. Raskutti G, Wainwright MJ, Yu B. Early stopping for non-parametric regression: An optimal data-dependent stopping rule [Internet]. 2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton). 2011. Available from: http://dx.doi.org/10.1109/allerton.2011.6120320

130. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. InInternational conference on machine learning 2015 Jun 1 (pp. 448-456). PMLR.

131. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning (still) requires rethinking generalization. Communications of the ACM. 2021 Feb 22;64(3):107-15.

132. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research. 2014 Jan 1;15(1):1929-58.

133. Tibshirani R. Regression Shrinkage and Selection Via the Lasso [Internet]. Vol. 58, Journal of the Royal Statistical Society: Series B (Methodological). 1996. p. 267–88. Available from: http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x

134. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970 Feb 1;12(1):55-67.

135. Zou H, Hastie T. Regularization and variable selection via the elastic net [Internet]. Vol. 67, Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005. p. 301–20. Available from: http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x

136. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, et al. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol. 2009 Mar;24(3):127–35.

137. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical andpowerful approach to multiple testing. J R Stat Soc Series B Methodol 1995;57:289–300.

138. Benjamini Y, Heller R, Yekutieli D. Selective inference in complex research [Internet]. Vol. 367, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2009. p. 4255–71. Available from: http://dx.doi.org/10.1098/rsta.2009.0127

139. Dunn OJ. Multiple comparison among means. J Am Stat Assoc 1961; 56: 52– 64

140. Reproduced from Wang Y, Kartasalo K, Weitz P, Ács B, Valkonen M, Larsson C, et al. Predicting Molecular Phenotypes from Histopathology Images: A Transcriptome-Wide Expression-Morphology Analysis in Breast Cancer. Cancer Res. 2021 Oct 1;81(19):5115–26. [author's material]

141. Martín M, Prat A, Rodríguez-Lescure A, Caballero R, Ebbert MTW, Munárriz B, et al. PAM50 proliferation score as a predictor of weekly paclitaxel benefit in breast cancer. Breast Cancer Res Treat. 2013 Apr;138(2):457–66.

142. Nielsen TO, Parker JS, Leung S, Voduc D, Ebbert M, Vickery T, et al. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. Clin Cancer Res. 2010 Nov 1;16(21):5222–32.

143. Tang Y, Wang Y, Kiani MF, Wang B. Classification, Treatment Strategy, and Associated Drug Resistance in Breast Cancer. Clin Breast Cancer. 2016 Oct;16(5):335–43.

144. Vieira AF, Schmitt F. An Update on Breast Cancer Multigene Prognostic Tests-Emergent Clinical Biomarkers. Front Med. 2018 Sep 4;5:248.

145. Raab R, Ismaila N, Andre F, Stearns V, Kalinsky K. Biomarkers for Adjuvant Endocrine and Chemotherapy in Early-Stage Breast Cancer: ASCO Guideline Update Q and A. JCO Oncol Pract. 2022 Sep;18(9):646–8.

146. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2019 Oct 1;30(10):1674.

147. Cheang MCU, Chia SK, Voduc D, Gao D, Leung S, Snider J, et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. J Natl Cancer Inst. 2009 May 20;101(10):736–50.

148. Coates AS, Winer EP, Goldhirsch A, Gelber RD, Gnant M, Piccart-Gebhart M, et al. Tailoring therapies--improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. Ann Oncol. 2015 Aug;26(8):1533–46.

149. Reinhardt F, Franken A, Fehm T, Neubauer H. Navigation through inter- and intratumoral heterogeneity of endocrine resistance mechanisms in breast cancer: A potential role for Liquid Biopsies? Tumour Biol. 2017 Nov;39(11):1010428317731511.

150. Abdelhakam DA, Hanna H, Nassar A. Oncotype DX and Prosigna in breast cancer patients: A comparison study. Cancer Treat Res Commun. 2021 Jan 7;26:100306.

151. Bartlett JMS, Bayani J, Marshall A, Dunn JA, Campbell A, Cunningham C, et al. Comparing Breast Cancer Multiparameter Tests in the OPTIMA Prelim Trial: No Test Is More Equal Than the Others. J Natl Cancer Inst [Internet]. 2016 Sep;108(9). Available from: http://dx.doi.org/10.1093/jnci/djw050

152. Stacke K, Eilertsen G, Unger J, Lundstrom C. Measuring Domain Shift for Deep Learning in Histopathology. IEEE J Biomed Health Inform. 2021 Feb;25(2):325–36.

153. Smotkin D, Nevadunsky NS, Harris K, Einstein MH, Yu Y, Goldberg GL. Histopathologic differences account for racial disparity in uterine cancer survival. Gynecol Oncol. 2012 Dec;127(3):616–9.

154. Huo D, Hu H, Rhie SK, Gamazon ER, Cherniack AD, Liu J, et al. Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. JAMA Oncol. 2017 Dec 1;3(12):1654–62.

155. Rabinovich A, Agarwal S, Laris C, Price J, Belongie S. Unsupervised color decomposition of histologically stained tissue samples. Advances in neural information processing systems. 2003;16.

156. Vahadane A, Peng T, Sethi A, Albarqouni S, Wang L, Baust M, et al. Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images. IEEE Trans Med Imaging. 2016 Aug;35(8):1962–71.

157. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. Communications of the ACM. 2020 Oct 22;63(11):139-44.

158. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision 2017 (pp. 2223-2232).

159. Shaban MT, Tarek Shaban M, Baur C, Navab N, Albarqouni S. Staingan: Stain Style Transfer for Digital Histological Images [Internet]. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). 2019. Available from: http://dx.doi.org/10.1109/isbi.2019.8759152

160. Roy S, Kumar Jain A, Lal S, Kini J. A study about color normalization methods for histopathology images. Micron. 2018 Nov;114:42–61.

161. Pontalba JT, Gwynne-Timothy T, David E, Jakate K, Androutsos D, Khademi A. Assessing the Impact of Color Normalization in Convolutional Neural Network-Based Nuclei Segmentation Frameworks. Front Bioeng Biotechnol. 2019 Nov 1;7:300.

162. DeVries T, Taylor GW. Learning confidence for out-of-distribution detection in neural networks. arXiv preprint arXiv:1802.04865. 2018 Feb 13.

163. Meinke A, Hein M. Towards neural networks that provably know when they don't know. arXiv preprint arXiv:1909.12180. 2019 Sep 26.

164. Liang S, Li Y, Srikant R. Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690. 2017 Jun 8.

165. Lee K, Lee K, Lee H, Shin J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems. 2018;31.

166. Sjögren R, Trygg J. Out-of-Distribution Example Detection in Deep Neural Networks using Distance to Modelled Embedding. arXiv preprint arXiv:2108.10673. 2021 Aug 24.

167. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. InInternational conference on machine learning 2017 Jul 17 (pp. 1321-1330). PMLR.

168. Zadrozny B, Elkan C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. InIcml 2001 Jun 28 (Vol. 1, pp. 609-616).

169. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. InProceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining 2002 Jul 23 (pp. 694-699).

170. Platt, John et al. Probabilistic outputs for support vectormachines and comparisons to regularized likelihoodmethods. Advances in large margin classifiers, 10(3):61–74, 1999.

171. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. InProceedings of the 22nd international conference on Machine learning 2005 Aug 7 (pp. 625-632).

172. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. InInternational conference on machine learning 2018 Jul 3 (pp. 2127-2136). PMLR.

173. Weitz P, Wang Y, Hartman J, Rantalainen M. An investigation of attention mechanisms in histopathology whole-slide-image analysis for regression objectives. InProceedings of the IEEE/CVF International Conference on Computer Vision 2021 (pp. 611-619).

174. Ren J, Karagoz K, Gatza M, Foran DJ, Qi X. Differentiation among prostate cancer patients with Gleason score of 7 using histopathology whole-slide image and genomic data. Proc SPIE Int Soc Opt Eng [Internet]. 2018 Feb;10579. Available from: http://dx.doi.org/10.1117/12.2293193

175. Chen RJ, Lu MY, Shaban M, Chen C, Chen TY, Williamson DF, Mahmood F. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2021 Sep 27 (pp. 339-349). Springer, Cham.