DEEP LEARNING FOR BIOMARKER AND OUTCOME PREDICTION IN CANCER

Dmitrii Bychkov, MSc (Tech.)

Institute for Molecular Medicine Finland – FIMM
Helsinki Institute of Life Science – HiLIFE
Faculty of Medicine
Doctoral Programme in Biomedicine
University of Helsinki
Finland

Academic dissertation

To be publicly discussed, with the permission of
the Faculty of Medicine of the University of Helsinki,
in Biomedicum Helsinki 1, Lecture Hall 3, Haartmaninkatu 8, Helsinki,
on February 18th, 2022 at 14:00.

Helsinki 2022

Supervised by

Docent, Johan Lundin, MD, PhD
Institute for Molecular Medicine Finland (FIMM)
Faculty of Medicine, University of Helsinki
Helsinki, Finland

Docent, Nina Linder, MD, PhD
Institute for Molecular Medicine Finland (FIMM)
Faculty of Medicine, University of Helsinki
Helsinki, Finland

Reviewed by

Associate Professor, Esa Rahtu, PhD
Tampere University of Technology
Tampere, Finland

Associate Professor, Mattias Rantalainen, PhD
Karolinska Institutet
Stockholm, Sweden

Opponent

Associate Professor, Lee A Cooper, PhD
Feinberg School of Medicine
Northwestern University
Chicago Illinois, USA

Custos

Docent Nina Linder, MD, PhD
Institute for Molecular Medicine Finland (FIMM)
Faculty of Medicine, University of Helsinki
Helsinki, Finland

The Faulty of Medicine uses the Urkund system (plagiarism recognition) to examine all doctoral dissertations.

Dissertationes Scholae Doctoralis Ad Sanitatem Investigandam
Universitatis Helsinkiensis (13/2022)

ISBN 978-951-51-7889-3 (Print)
ISBN 978-951-51-7890-9 (Online)
ISSN 2342-3161 (Print)
ISSN 2342-317X (Online)
`http://ethesis.helsinki.fi`
Unigrafia Oy, Helsinki 2022

To my family and friends

Abstract

Machine learning in the form of deep learning (DL) has recently transformed how computer vision tasks are solved in numerous domains, including image-based medical diagnostics. DL-based methods have the potential to enable more precise quantitative characterisation of cancer tissue specimens routinely analysed in clinical pathology laboratories for diagnostic purposes. Computer-assisted tissue analysis within pathology is not restricted to the quantification and classification of specific tissue entities. DL allows to directly address clinically relevant questions related to the prediction of cancer outcome and efficacy of cancer treatment.

This thesis focused on the following crucial research question: is it possible to predict cancer outcome, biomarker status, and treatment efficacy directly from the tissue morphology using DL without any special stains or molecular methods? To address this question, we utilised digitised hematoxylin-eosin-stained (H&E) tissue specimens from two common types of solid tumours – breast and colorectal cancer. Tissue specimens and corresponding clinical data were retrieved from retrospective patient series collected in Finland. First, a DL-based algorithm was developed to extract prognostic information for patients diagnosed with colorectal cancer, using digitised H&E images only. Computational analysis of tumour tissue samples with DL demonstrated a superhuman performance and surpassed a consensus of three expert pathologists in predicting five-year colorectal cancer-specific outcomes. Then, outcome prediction was studied in two independent breast cancer patient series. Particularly, generalisation of the trained algorithms to previously unseen patients from an independent series was examined on the large whole-slide tumour specimens. In breast cancer outcome prediction, we investigated a multitask learning approach by combining outcome and biomarker-supervised learning. Our experiments in breast and colorectal cancer show that tissue morphological features learned by the DL models supervised by patient outcome provided prognostic information independent of established prognostic factors such as histological grade, tumour size and lymph nodes status. Additionally, the accuracy of DL-based predictors was compared to other prognostic characteristics evaluated by pathologists in breast cancer, including mitotic count, nuclear pleomorphism, tubules formation, tumour necrosis and tumour-infiltrating lymphocytes. We further assessed if molecular biomarkers such as hormone receptor status and *ERBB2* gene amplification can be predicted from H&E-stained tissue samples obtained at the time of diagnosis from patients with breast cancer and showed that molecular alterations are reflected in the basic tissue morphology and can be captured with DL. Finally, we studied how morphological features of breast cancer can be linked to molecularly targeted treatment response. The results showed that *ERBB2*-associated morphology extracted with DL correlated with the efficacy of adjuvant anti-*ERBB2* treatment and can contribute to treatment-predictive information in breast cancer.

Taken together, this thesis shows the potential utility of DL in tissue-based characterisation of cancer for prediction of cancer outcome, tumour molecular status and efficacy of molecularly targeted treatments. DL-based analysis of the basic tissue morphology can provide significant predictive information and be combined with clinicopathological and molecular data to improve the accuracy of cancer diagnostics.

Tiivistelmä

Koneoppiminen syväoppimisen (SO) muodossa on muuttanut, miten tietokonenäön tehtävät ratkaistaan monilla toimialueilla, kuten lääketieteellisessä kuvantamisdiagnostiikkassa. SO-perusteiset menetelmät mahdollistavat tarkemman kvantitatiivisen karakterisoinnin syöpäkasvainnäytteistä, jotka rutiinisti analysoidaan kliinisen patologian laboratorioissa diagnosointia varten. Tietokoneavusteinen kudosanalyysi ei rajoitu ainoastaan tiettyjen kudosentiteettien määrittämiseen ja luokitteluun. SO:n avulla voidaan suoraan tutkia syövän ennustetta ja syöpähoitojen vastetta.

Tämä väitöskirja keskittyi tärkeään tutkimuskysymykseen: onko syövän ennuste, biomarkkerien status ja hoidon tehokkuus mahdollista ennustaa SO:lla suoraan kudosmorfologiasta ilman erillisiä värjäyksiä tai molekyylibiologisia testejä? Vastataksemme tähän kysymykseen käytimme digitaalisia hematoksyliini-eosiini (H&E)-värjättyjä kudosnäytteitä kahdesta tavallisesta kiinteästä kasvaimesta, rinta- ja paksusuolensyövästä. Kudosnäytteet ja niihin liittyvät kliiniset tiedot saatiin Suomessa kerätystä retrospektiivisestä potilassarjasta. Ensimmäiseksi kehitimme SO-algoritmin, jolla poimimme prognostisen tiedon paksusuolensyöpäpotilaista käyttäen ainoastaan digitalisoituja H&E-värjäyksiä. Kudosnäytteistä SO:lla tehty laskennallinen analyysi osoitti ihmisasiantuntijaa parempaa suorituskykyä ja ylitti kolmen patologian asiantuntijan antaman yksimielisen viiden vuoden ennusteen syövän lopputulemasta. Seuraavaksi lopputuleman ennustamista tutkittiin kahdessa erillisessä rintasyöpäpotilassarjassa. Erityisesti tutkimme koulutetun algoritmin kykyä yleistää syöpäkudosten kokoleikkeistä, jotka olivat peräisin erillisestä algoritmille aiemmin tuntemattomasta potilassarjasta. Rintasyövän ennusteen suhteen tutkimme ”multitask learning”-lähestymistapaa yhdistämällä eloonjäämis- ja biomarkkeri-valvotun oppimisen. Tutkimuksemme rinta- ja paksusuolensyövän osalta osoittavat, että SO-mallien avulla, jotka ovat opetettu potilaan eloonjäämisen mukaan, voidaan kudosmorfologian perusteella saada ennuste, joka on rippumaton aiemmin saatavilla olevista ennustetekijöistä, kuten histologisesta luokittelusta, kasvaimen koosta ja imusolmukkeiden statuksesta. Lisäksi SO-perusteisten ennusteiden tarkkuutta rintasyövässä verrattiin patologien arvioimiin syovän, kuten mitoosien lukumäärä, tuman pleomorfismiin, tubulusten tiehyeiden erilaistumisasteeseen, kasvaimen nekroosiin ja kasvaimen infiltroiviin lymfosyytteihin. Tutkimme myös, voiko rintasyöpäpotilailta syöpädiagnosoinnin yhteydessä saaduista H&E-värjätyistä kudosnäytteistä ennustaa molekulaarisia biomarkkereita, kuten hormonireseptoristatusta ja *ERBB2*-geenin monistumista. Tutkimuksemme osoitti, että molekulaariset muutokset löytyvät myös kudosmorfologiasta ja ne voi tunnistaa SO:n avulla. Lopuksi tutkimme, miten rintasyövän morfologiset piirteet voidaan yhdistää hoitovasteeseen. Tutkimuksemme osoitti, että SO:n tunnistama *ERBB2*-positiivisen kasvaimen morfologia korreloi anti-*ERBB2*-liitännäishoitojen tehokkuuden kanssa ja SO:ta voi käyttää ennustamaan rintasyövän lääkevastetta.

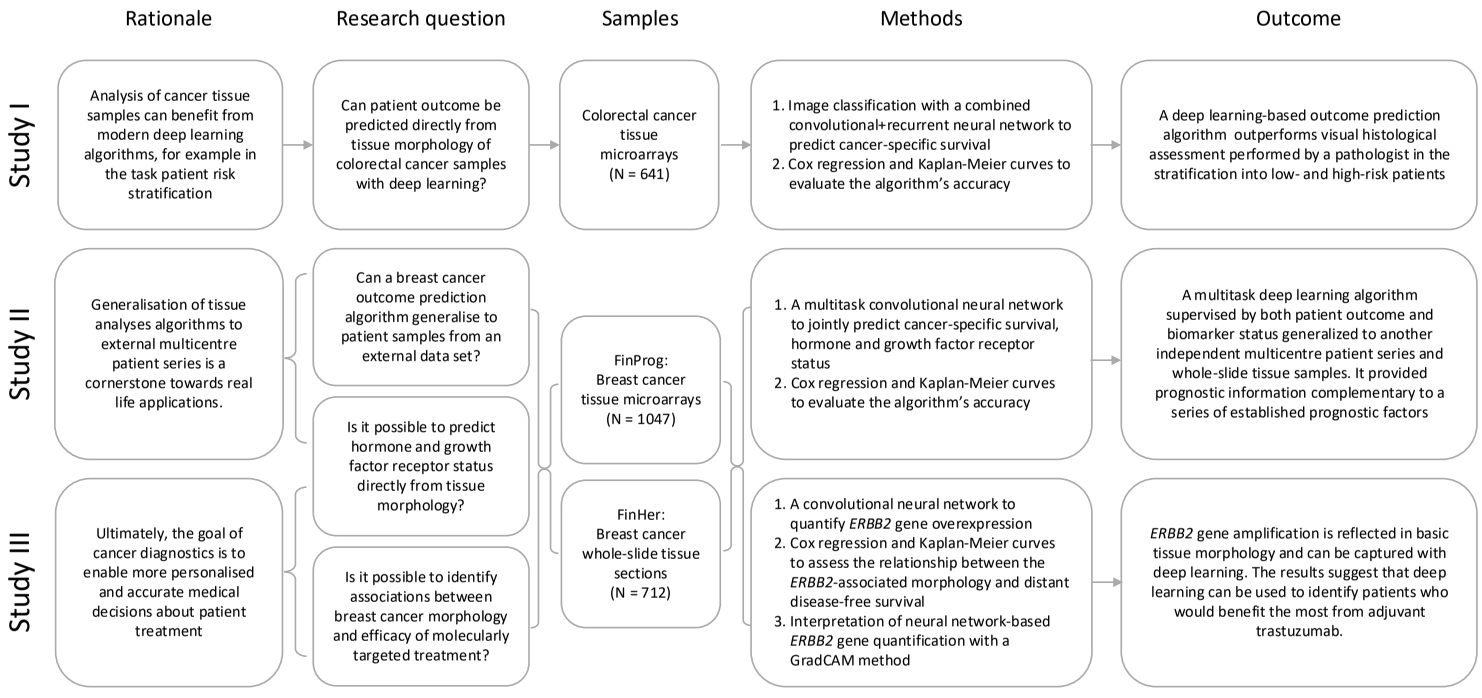
Tämän väitöskirjatyön tulokset osoittavat, että SO:n syöpäkudoksen karakterisointi voi olla hyödyllinen syövän ennusteen arvioinnissa sekä, molekulaarisen statuksen ja lääkevasteen ennustamisessa. SO-perusteinen kudosmorfologinen analyysi voi antaa merkittävää tietoa syövän ennusteesta ja se voidaan yhdistää kliiniseen patologiaan ja molekulaariseen informaatioon tarkemman syöpädiagnosoinnin mahdollistamiseksi.

Contents

Abbreviations

AP	Average precision
AUC	Area under the receiver operating characteristic curve
BCSS	Breast cancer-specific survival
CI	Confidence interval
CISH	Chromogenic *in situ* hybridisation
CNN	Convolutional neural network
DDFS	Distant disease-free survival
DL	Deep learning
ECW	Enhanced compressed wavelet
ER	Estrogen receptor
ERBB2	Erb-B2 receptor tyrosine kinase 2 gene
ERM	Empirical risk minimisation
FFPE	Formalin-fixed paraffin-embedded
GPU	Graphics processing unit
H&E	Hematoxylin and eosin
HER2	Human epidermal growth factor receptor 2
HR	Hazard ratio
IFV	Improved Fisher vector
IHC	Immunohistochemistry
LSTM	Long short-term memory
MAP	Maximum a posteriori probability estimate
ML	Machine learning
PH	Proportional hazards
PR	Progesterone receptor
ROC	Receiver operating characteristic curve
SVM	Support vector machine
TILs	Tumour-infiltrating lymphocytes
TMA	Tissue microarray
WS	Whole-slide
WSI	Whole-slide image

THESIS OVERVIEW



LIST OF ORIGINAL PUBLICATIONS

Publication I **Bychkov D**, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C, Walliander M, Lundin M, Haglund C & Lundin J "Deep learning based tissue analysis predicts outcome in colorectal cancer" *Scientific Reports* 8, 3395 (2018).

Publication II **Bychkov D**, Joensuu H, Nordling S, Tiulpin A, Kücükel H, Lundin M, Sihto H, Isola J, Lehtimäki T, Kellokumpu-Lehtinen PK, von Smitten K, Lundin J & Linder N "Outcome and Biomarker Supervised Deep Learning for Survival Prediction in Two Multicenter Breast Cancer Cohorts" *Journal of Pathology Informatics* 13:9 (2022).

Publication III **Bychkov D**, Linder N, Tiulpin A, Kücükel H, Lundin M, Nordling S, Sihto H, Isola J, Lehtimäki T, Kellokumpu-Lehtinen PK, von Smitten K, Joensuu H & Lundin J "Deep Learning Identifies Morphological Features in Breast Cancer Predictive of Cancer *ERBB2* Status and Trastuzumab Treatment Efficacy" *Scientific Reports* 11, 4037 (2021).

The publications are referred to in the text by their roman numerals. The original publications are reprinted with the permission of their copyright holders.

1 Introduction

Advances in deep learning (DL) techniques start to play an increasingly important role in healthcare in general, and image-based medical diagnostics in particular [1, 2, 3, 4, 5]. Computational image analysis empowered by DL has been used to address several tissue-based diagnostic tasks performed on digitised hematoxylin-and-eosin (H&E) tissue specimens. Examples include detection of cancerous tissue on whole-slide images (WSIs) [6, 7, 8, 9], histological grading of tumours [10, 11], quantification of tissue entities such as mitotic cells [12, 13], necrosis [14] and tumour-infiltrating lymphocytes (TILs) [15, 16, 17]. Addressing these tasks with supervised DL requires a significant amount of image annotations that are typically performed manually by human experts, e.g. pathologists. This approach is prone to potential human bias and does not allow to extract information that is not readily discernible by a human eye [18].

Recent studies go beyond expert-guided algorithms for the analysis of H&E tumour samples [18]. For example, genetic alterations in cancer can cause phenotypic changes in tumours and their surrounding tissue that can be captured with DL [19, 20, 21]. This approach eliminates potential human biases introduced through manual data annotations, thus reducing human labour work and, more importantly, allowing to discover how tumour morphology reflects molecular perturbations in cancer.

In the same way, as DL can be trained to predict genetic changes based on tissue morphology, DL algorithms have been applied to predict disease outcome, biomarker status and response to treatment. Studies have shown how DL trained with digitised H&E tissue specimens as input can predict outcome in patients with brain [22], breast [23], colorectal [24, 25], and other cancers [26, 27]. Moreover, existing molecular biomarkers can also be captured with DL [18]. For example, estrogen (ER) and progesterone (PR) receptor status in breast cancer are well-established therapy-specific prognostic biomarkers that predict whether a patient is likely to respond to hormone therapy. Similarly, breast cancer patients with *ERBB2* gene amplification are more likely to respond to anti-*ERBB2* targeted therapy [28, 29]. While prognostic biomarkers aid in patient stratification according to their risk of disease progression or death, predictive biomarkers also provide information on how the patients should be treated and the effect of the therapeutic intervention [30, 31]. Thus, with DL it could be possible to develop novel tissue-based prognostic and predictive biomarkers.

In this thesis, we study the use of DL algorithms for biomarker and outcome prediction directly from tissue morphology of breast and colorectal tumours. Predictions of disease outcome and the efficacy of molecularly targeted treatments are essential for the decision-making process, management and counselling of patients with cancer. We studied whether DL algorithms trained to predict the therapy-specific prognostic biomarker *ERBB2* based on morphology only, also can predict the efficacy of anti-*ERBB2* targeted therapy with trastuzumab in patients with breast cancer. We demonstrate how DL can complement expert-based visual tissue analysis to provide independent prognostic information in breast and colorectal cancer.

2 Review of the literature

2.1 Deep learning in computer vision

DL has during the last decade emerged as the state-of-the-art method within machine learning (ML) and artificial intelligence (AI) [1]. Recent DL methods have dramatically improved the the accuracy and efficiency of pattern recognition and representation learning in various domains, including self-driving cars [32], machine translation [33], finance [34], arts [35], and healthcare [36]. DL have demonstrated particular success in computer vision tasks related to image classification [9], object detection [12, 16] and segmentation [37, 17], and shown great potential in image-based medical diagnostics [4].

2.1.1 Introduction to artificial neural networks

Essentially, DL is a reincarnation of artificial neural networks – a broad family of ML models composed of simple computational units called neurons. The basic idea behind artificial neuron dates back to 1958 when a Perceptron was first conceived as a simplified mathematical model of how neurons function in the human brain [38]. Mathematically, an artificial neuron is a scalar product of two vectors or a weighted sum of inputs, followed by an activation function:

$$a = f\left(w^T x + b\right). \tag{1}$$

Here, $x \in \mathbb{R}^n$ is an n-dimensional input vector and w is a set of weights such that each x_i is associated with its weight w_i, and $f(\cdot)$ is an activation function. Popular activation functions have been *sigmoid* and *hyperbolic tangent* nonlinearities. The former maps neuron outputs to $[0,1]$ range and has a natural probabilistic interpretation. The latter scales the output values to $[-1,1]$. Both functions, though, may lead to a problem called vanishing gradients and make deep networks difficult to train [39]. To surpass the problem, other activation functions have been proposed, including the currently popular rectified linear unit (ReLU) [40].

A neuron in equation (1) can either take raw data values or outputs of other neurons as inputs, suggesting that neurons can be organised in an interconnected structure and constitute artificial neural networks. Importantly, neural connections have to be organised in an acyclic manner. Typically, neurons are arranged in layers of three types: *input* layers, *hidden* layers and *output* layers. Networks with at least one hidden layer are often referred to as the Multi-Layer Perceptron (MLP). As

the number of hidden layers grows, the networks become deep and consequently give rise to DL [1]. Vanilla MLP, also known as a fully-connected network - is the simplest and most generic architecture yet not optimal in, e.g. computer vision applications. A variety of network architectures have been designed to efficiently handle different data types, such as image data or data that exhibit temporal dynamic behaviour [41, 42].

2.1.2 Network architectures for computer vision

In computer vision, *convolutional neural networks* (CNNs) [41, 43] – a special type of feed-forward neural networks that efficiently handle the grid-like structure of images have a central role. As the name suggests, the CNNs are based on the operation called convolution or cross-correlation. This operation performs pattern matching through the multiplication of inputs at each spatial location with a kernel – a three-dimensional matrix of weights [44]. Important properties of the CNNs are sparse connectivity and parameter sharing [44]. These properties significantly reduce the number of model parameters, resulting in improved computational efficacy and reduced memory requirements compared to the fully-connected architecture [44]. Typically, CNNs represent a feature pyramid, where blocks of layers learn intermediate feature representations. Many variants of the original CNN architecture have been proposed in recent years. Popular examples include, AlexNet [45], VGG [46], ResNets [47], Inception [48].

2.1.3 Supervised learning

Many computer vision tasks, including image classification, pixel-level segmentation or bounding-box object detection, can be formalised, for example either as classification or regression and solved in a supervised fashion. In supervised learning [44], a dataset $\boldsymbol{D}$ is represented by pairs $\{(x^{(i)}, y^{(i)})\}_{i=0}^{N-1}$, where each data point $x^{(i)}$ has a corresponding target or *label* $y^{(i)}$, and N is a the size of the dataset drawn from a joint distribution $p(x,y)$. A DL algorithm, e.g. a neural network with a fixed structure is defined as a parametric function $f(x;\theta)$ that provides a mapping between observations $x^{(i)}$ and corresponding labels $y^{(i)}$. The goal of learning is to find an optimal set of parameters θ of the model by minimising an objective function $J(\theta)$:

$$J(\theta) = \mathbb{E}_{(x,y)\sim p(x,y)}\left[L(f(x;\theta),y)\right] \rightarrow \min_{\theta}, \tag{2}$$

where $L(\cdot)$ is the distance between model predictions and labels – a measure of the

algorithm's performance. Since the underlying distribution $p(x,y)$ of the data is typically unknown, the expectation $\mathbb{E}$ is calculated across an *empirical* distribution of the observed data $\hat{p}(x,y)$ that we call a *training set*:

$$\mathbb{E}_{(x,y)\sim\hat{p}(x,y)}\left[L(f(x;\theta),y)\right] = \frac{1}{N}\sum_{i=0}^{N-1} L(f(x^{(i)};\theta),y^{(i)}). \tag{3}$$

Minimising the average training error is known as *empirical risk minimisation* (ERM) [44]. Other approaches to estimate θ exist, e.g. a maximum a posteriori probability estimate (MAP) and maximum likelihood estimate (MLE) [49].

The choice of distance measure $L(\cdot)$ depends on the task. For classification problems, *cross-entropy* is a standard function. A binary version of the cross-entropy (BCE) loss for model $f(x;\theta)$ that outputs predictions $\hat{y} \in [0,1]$ looks as follows:

$$L_{BCE}(f(x;\theta),y) = -\frac{1}{N}\sum_{i=0}^{N-1} y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i). \tag{4}$$

Extensions of the original cross-entropy were proposed, for example a *focal loss* was introduced to address class imbalance problem for CNN-based object detectors [50]. When $f(x;\theta)$ solves a regression problem such that $\hat{y} \in \mathbb{R}$, *mean squared error* (MSE) loss is a standard choice:

$$L_{MSE}(f(x;\theta),y) = -\frac{1}{N}\sum_{i=0}^{N-1} (y_i-\hat{y}_i)^2. \tag{5}$$

As the loss function is defined, the training procedure aims to minimise the loss by iteratively updating the parameters of the model. The gradient of the loss function with respect to the parameters of the model defines the best direction along which the parameters should be changed:

$$\nabla_\theta L(f(x;\theta),y) = \frac{\partial L}{\partial \theta} = \left[\frac{\partial L}{\partial \theta_1}, \frac{\partial L}{\partial \theta_2}, ..., \frac{\partial L}{\partial \theta_n}\right]. \tag{6}$$

Calculating first-order partial derivatives of the loss with respect to θ is done using the *backpropagation* algorithm [51] or the chain rule [49]. Once the gradients can be computed, the Gradient Descent algorithm is applied by repeatedly calculating the gradient and performing a parameter update until a specific stop criterion is met. In practical applications with large-scale datasets, which is often the case in computer vision, computing the loss function on the entire training set becomes

problematic. To overcome this challenge, the loss function can be approximated on batches - smaller portions of training data. This is referred to as mini-batch or stochastic mini-batch gradient descent (SGD). The algorithm for implementing mini-batch SGD is given below:

Algorithm 1: Mini-batch Gradient Descent Algorithm

Requre: $\boldsymbol{D}$ - training data;
$\theta \leftarrow$ initialise model parameters;
$\eta \leftarrow$ initialise learning rate;
while *stop criterion not met* **do**
  $\{(x^{(batch)}, y^{(batch)})\} \leftarrow$ sample a mini-batch of m pairs from $\boldsymbol{D}$;
  Compute outputs: $\hat{y}^{(batch)} \leftarrow f(x^{(batch)}; \theta)$;
  Compute gradient: $\hat{g} \leftarrow \frac{1}{m} \nabla_\theta L(\hat{y}^{(batch)}, y^{(batch)})$;
  Apply update: $\theta \leftarrow \theta - \eta \hat{g}$
end

The result of the training procedure heavily depends on a hyperparameter η called *learning rate*. Typically, the learning rate is initialised with a small positive value. e.g. 10^{-3}, which defines the size of a step that SGD takes along the gradient (downhill) at each iteration. Original SGD have seen many modifications meant to improve the speed of convergence [52]. Some recent and popular versions of SGD include Adagrad [53], Adadelta [54] and Adam [55]. Most of them leverage the idea of the adaptive learning rate for the individual parameters, which is claimed to improve the speed and convergence of the models.

2.1.4 Regularisation and model selection

DL models are often overparameterised, which can lead to *overfitting* - an undesired effect when a model demonstrates high accuracy on a training set but fails to generalise to a new, unseen set of data. Model generalisation is often assessed by splitting the dataset at hands into three non-overlapping parts:

- Training set – used to estimate model parameters θ;
- Validation set – used for hyperparameter (e.g. number of hidden layers, learning rate) tuning and for early stopping;
- Test set – used to obtain an unbiased estimate of model performance on unseen data.

Often, **cross-validation** is used for hyperparameter tuning and model selection [56, 57]. Particularly, the dataset is split into K equally-sized non-overlapping

folds. Then, $K-1$ folds are used for training, and the k_{th} fold is used for validation. The process is repeated K times, shuffling the folds in such a way that a different validation set is used at each iteration.

A common approach used for decades to improve model generalisation is to add a regularisation or parameter **penalty term** $R(\theta)$ to the equation (3), such that ERM parameter estimates $\hat{\theta}$ are obtained by solving equation (7):

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{i=0}^{N-1} L(f(x^{(i)};\theta), y^{(i)}) + \gamma R(\theta), \tag{7}$$

where $\gamma \in [0,\infty]$ is a hyperparameter that defines the strengths of regularisation. Two popular choices of $R(\theta)$ in neural networks and other ML models are *lasso regression* [58] and *ridge regression* [59]. The former is know as $L1$ regularisation and takes the form of $\sum|\theta|$, the later is called $L2$ and defined as $\sum\theta^2$. Intuitively, $L1$ leads to sparse θ during training, whereas $L2$ enforces diffuse values of θ by pushing model parameters towards 0. A combination of both is called *elastic net* [60]. It is possible to demonstrate that using $L1$ in equation (7) is equivalent to MAP estimate for θ with a Laplace prior [61], whereas $L2$ becomes equivalent to Gaussian prior [62].

In a supervised setting, **multitask learning** [63] has been proven effective to yield better generalisation of the models. The "multitasking" is achieved by introducing additional output nodes to the network to predict different but related targets, i.e. solving several tasks simultaneously. The tasks still share common inputs and hidden layers, hence imposing additional constraints on the parameters of the model [64].

More recently, a **dropout** technique was introduced to address regularisation specifically in deep neural networks [65]. It keeps individual neurons inactive during training with some probability p (a hyperparameter) and can effectively complement $L1$ and $L2$.

Image augmentation [66] methods have an important role in computer vision. These methods are particularly effective when data labelling for supervised training is laborious, time-consuming and expensive. In that situation, the training set can be significantly enlarged by augmenting existing labelled data and creating new artificial observations. Image transformations often used to generate augmented data include rotations, flipping, shears, adding noise and performing contrast and brightness perturbations [67].

2.2 Preparation and visual examination of tissue specimens for research and diagnostic purposes

Visual microscopic assessment of tissue morphology, supplemented by molecular methods, such as immunohistochemistry and *in situ* hybridization, are standard methods in cancer diagnostics and research [68]. The cancer tissue morphology, protein and gene expression pattern, characteristics of the tumour microenvironment are associated with the aggressiveness of the disease, risk of recurrence and efficacy of specific treatments [69]. Below we briefly summarise the main steps for tissue preparation, staining and visual examination used for diagnostic and research purposes.

2.2.1 Sample preparation and staining

After a tissue sample is obtained from the patient during surgery or as a biopsy, it requires preparation to prevent tissue degradation and preserve morphological structures, and molecular composition [70]. First, the tissue is immersed and fixed in a formaldehyde (formalin) solution, then dehydrated using ethanol and finally embedded into paraffin [70]. The resulting formalin-fixed paraffin-embedded (FFPE) tissue blocks are then typically cut into 3-5 μm thick sections with a microtome and mounted onto microscopic glass slides for diagnostic and research purposes, and the remaining tissue blocks can be archived for later use [70].

Pathology laboratories typically prepare one or more tissue blocks per tissue specimen, and multiple slides are cut from each block, which subsequently are subject to different stainings. A technique called tissue microarrays (TMA) was developed to gather tissue from multiple specimens into a single tissue block [71]. That is achieved by punching 0.6 - 1 mm cylinder-shaped tissue cores from different FFPE blocks and transferring those to a densely and precisely arrayed recipient block. As a result, up to 100-150 individual patient samples can be represented on the same slide cut from a TMA block and analysed simultaneously [72].

Tissue sections are transparent and almost colourless; thus, they need to be stained or dyed to make tissue components visible. Various staining procedures have been developed to increase contrast, visualise the morphological structures and highlight specific tissue entities [73]. The most common staining protocol is based on a combination of hematoxylin and eosin (H&E) dyes [68]. Hematoxylin stains the DNA of the cell nuclei dark blue, whereas eosin stains the cytoplasm and extracellular components pink. Other tissue structures take up a combination of those colours with different shades, hues, and textures [68]. Another commonly

used staining technique is immunohistochemistry (IHC) [74]. In contrast to the H&E, IHC relies on the specific interaction between an antigen and an antibody directed against it. The antibody-antigen interaction is visualised using chromogenic or fluorescent methods. The IHC method is beneficial in identifying and localising specific proteins of interest, individual cells and cell populations [74].

2.2.2 Visual tissue examination in cancer diagnostics

Pathologists examine tissue samples under a microscope to assess morphological characteristics of tumour tissue, including the exact size of lesions, patterns of growth, surgical margins, histological type and grade, the presence and quantity of specific cell types, as well as expression of proteins and genes through molecular methods. Specifically, the histological grade is a measure of the degree of tumour differentiation. Tumours that closely resemble normal tissue with regards to morphological features are considered well differentiated, and tumours that have lost their resemblance with healthy tissue are considered poorly differentiated. Poorly differentiated are more aggressive and likely to metastasise [75]. Grading systems differ depending on the cancer type, e.g., a three-tier grading system in breast cancer is based on a semiquantitative assessment of the following morphological features: percentage of tubule formation, the degree of nuclear pleomorphism, and a mitotic count within a defined tissue area defined by fields-of-view analysed by microscopy [75]. A four-tier standard used for grading colorectal cancer is defined by the degree of glandular structures formation and the least differentiated tumour areas [76, 77]. Tumour size and type, lymph node status and presence of distant metastases constitute the stage of the disease [78]. The TNM classification is the most widely used cancer staging system [78]. Other tissue features that can be assessed by pathologists include the presence of tissue necrosis, immune cell infiltration, cancer-induced angiogenesis, and various stromal features associated with cancer [69].

Molecular pathology is an essential component of cancer diagnostics that supplements the examination of tissue morphological features [79]. It is typically performed through IHC to evaluate specific proteins that serve as prognostic and therapy-specific biomarkers. Prognostic biomarkers provide information about patient outcome regardless of treatment, whereas therapy-specific prognostic biomarkers give information on the likelihood that a patient will respond to a certain therapeutic intervention [30]. Molecular characterisation of tumours allows to better understand underlying molecular pathways that drive the disease and tailor the therapy for individual cancer patients. For example, hormone receptor status as assessed by IHC analysis of estrogen receptor (ER) and progesterone receptor (PR)

status is routinely determined in the diagnosis of breast cancer [28]. Overexpression of HER2 protein – human epidermal growth factor receptor 2, encoded by amplified *ERBB2* gene, is an established prognostic and therapy-specific biomarker. Amplified *ERBB2* gene is a molecular alteration that can be targeted with anti-*ERBB2* therapy, such as trastuzumab [29]. Other biomarkers that complement the diagnosis of breast cancer include proliferative activity of tumours evaluated through IHC analysis of the Ki67 protein [80]. Ki67 is expressed in dividing cells and serves as a prognostic biomarker [80, 81] in the management of various malignancies.

2.3 Deep learning for cancer tissue analysis

2.3.1 Applications in cancer diagnostics

Deep learning-based approaches within digital pathology and cancer research have already shown promising results in a substantial series of image-based quantification and classification tasks and have the potential to address issues related to reproducibility and throughput in tissue examination within pathology [82]. The development of microscopy scanners, capable of digitising entire tissue specimens and produce so-called whole-slide images (WSIs), allow both visualisation and image analysis of the digital samples for diagnostic and research purposes [83]. Digitisation of large retrospective series of tissue samples combined with comprehensive clinical information, such as information on outcome and treatment [84, 85, 86] provide opportunities for mining knowledge about the disease, improve cancer diagnostics and support clinical decision-making [4].

Conventional tasks in tissue examination within pathology that have been successfully tackled with DL methods include counting mitosis [87, 13], quantifying tumour-infiltrating immune cells [15, 16, 17], assessing the grade of tumour differentiation [88, 89], segmenting specific tissue entities such as glands [90, 91], stroma [92], vascular structures and other tissue components [93]. Quantification of these tissue entities often serves as intermediate steps to address "higher level" biological and clinical questions such as response to treatment, need for surgical intervention, and disease outcome prediction. With DL, it has become possible to go beyond expert-supervised tasks in tissue analysis [94, 5] and directly predict clinical endpoints, such as disease outcome [23] and response to treatment [18]. Recent studies demonstrate that using digitised tumour samples stained for basic morphology, it is possible to extract information not readily discernible by the human eye. For example, molecular and genomic alterations can be reflected in the tissue morphology and identified with DL algorithms [95].

2.3.2 Outcome prediction

Early studies that addressed cancer outcome prediction directly from the tumour morphology as revealed by H&E staining relied on conventional image feature extraction followed by a machine learning classifier such as Support Vector Machine (SVM) [96] or logistic regression. Examples include survival prediction in the lung [97, 26, 98] and breast cancer [99]. With the advances in DL methods, researchers have started to adapt convolutional neural networks (CNNs) for feature extraction as a powerful alternative to hand-crafted features.

CNNs have been trained using tissue entity labels made by experts and supervised learning approaches to quantify tissue entities that are known to be predictive of survival. For example, the tumour-stroma ratio has been shown to be an independent prognostic factor in various solid tumours, including breast, colorectal and lung cancer [100]. A study on colorectal cancer used CNNs to classify tumour and stroma regions in TMAs and demonstrated that a high stroma to tumour ratio was associated with a worse prognosis than a low stroma ratio in patients with rectal cancer [101]. Similarly, in a study on outcome prediction in colorectal cancer, CNNs were used to segment cancer-associated stroma, tumour epithelium, and lymphocytes in WSIs, followed by survival analysis [25]. The results showed that a "deep stroma score" was an independent prognostic factor for patient overall survival with a hazard ratio (HR) of 1.63. In prostate cancer, a pre-trained CNN (Inception-V3) [102] was used for patch-level Gleason pattern [103] segmentation which was then used in risk stratification for disease progression [104].

Another group of studies used CNN-derived features to perform outcome prediction with no direct domain expertise involved. One of the first purely outcome-supervised methods used a pre-trained VGG-16 [46] architecture to extract image features from colorectal cancer TMA samples [24]. A recurrent architecture was then trained on VGG feature vectors to jointly aggregate patch-level information and predict five-year cancer-specific survival [24]. A related approach for TMA-based breast cancer outcome prediction combined pre-trained VGG-16 feature extraction, feature pooling with Improved Fisher Vector encoding and classification with SVM [23]. A similar study on breast cancer also combined VGG-16 features with SVM to predict the risk of recurrence based on H&E-stained TMA samples [105]. Fully end-to-end (without intermediate steps) outcome-supervised methods have been studied for survival prediction in patients with brain [22, 106], lung cancer [106], mesotheliomas [107], colorectal cancer [108], and across ten different cancer types [27] based on data from The Cancer Genome Atlas (TCGA) [86]. Moreover, the study on TCGA Low-Grade Glioma (LGG) and Glioblastoma (GBM) cohorts integrated information on genomic biomarkers with morphology-derived features to improve the prognostic accuracy of DL models [22].

Some unsupervised approaches for tissue image subtyping, i.e. clustering, based on visual similarity, were used on digitised tissue samples of cholangiocarcinoma [109], and various cancer types [110] from the TCGA archive. Histological tissue subtypes/clusters were then evaluated using Cox survival regression. A comprehensive study across 28 TCGA cancer types also utilised transfer learning with the Inception architecture to predict patients' survival [111].

2.3.3 Biomarker and treatment response prediction

Several studies have investigated the prediction of molecular biomarkers in breast cancer directly from H&E samples. Prediction of ER status was approached with DL [105] and using nuclear morphometric pattern analysis [112] on breast cancer TMA samples. Later, morphological properties of breast cancer were analysed to find associations with ER, PR, Ki67, and HER2 expression in TMA tissue specimens [113]. Recent studies expanded receptor status prediction in breast cancer by analysing WSIs of tissue sections [114, 115, 116, 95]. The performance of the DL algorithms used for biomarker prediction in breast cancer has varied depending on the size of the datasets used for training and validation.

In gastrointestinal cancers, microsatellite instability was predicted directly from WSIs of tissue specimens [19]. Moreover, it was demonstrated that messenger RNA expression can be predicted from WSIs of H&E samples [81], and across 28 different tumour types, including breast and colorectal cancers [21].

With DL, it becomes possible to identify morphological features of tumour tissue that could be predictive of a positive response to both chemotherapy [117, 118] and targeted therapy [18]. Most of the cancer therapies are effective only in a subset of patients; thus, a more accurate segregation of responders from non-responders can help to minimise side effects of the treatments [18]. One way to predict treatment response from the H&E tumour tissue samples is through already known molecular biomarkers, e.g. hormone-receptor status and *ERBB2* gene amplification in breast cancer patients [116]. Alternatively, treatment response can be predicted directly from the basic morphology images, without intermediate identifications of predefined molecular biomarkers [18].

Since treatment response information is typically not readily available, few studies explored the feasibility of tissue-based therapy-specific prognostic biomarkers. Prediction of response to neoadjuvant chemotherapy using histological tissue images has been investigated in breast cancer patients [117, 118]. Pathological complete response prediction with DL achieved a ROC AUC of 0.847 on a validation set of 117 breast cancer patients [118]. Prediction of response to immunotherapy has been addressed in patients with malignant melanoma and non-small cell lung cancer directly from H&E-stained tissue samples [119, 120].

Similarly to outcome prediction, prediction of treatment response can lead to the identification of novel tissue-based biomarkers and have an impact on clinical decision-making [18].

2.4 Methodological aspects of survival modelling

Statistical methods that model expected time before a particular event of interest occurs, as a function of covariates, are called *Survival Models* [121]. In the context of cancer outcome prediction, examples of such events could be local or regional disease recurrence, development of distant metastasis or death. Each patient is treated as an individual observation followed for a specific duration of time. We refer to this time as a follow-up time, during which the event may or may not occur. Time-to-event data present challenges that stem from incomplete observations, known as *censored* observations [122]. *Censoring* refers to the situation when an individual is lost to follow-up for whatever reason, or situations when information, whether the event of interest occurred or not, is not available [122]. Established and widely used methods that deal with censored data include the non-parametric Kaplan-Meier estimator [123], and Cox survival regression [124]. The Cox regression is a linear model with a restrictive assumption that the effect of the covariates does not change over time, i.e. is independent of time. Therefore, the original Cox method is frequently referred to as a proportional hazards regression. Violating the assumption of proportional hazards (PH) may lead to faulty conclusions, though it is rarely checked for in practical applications [125, 126, 127]. Variations of the original Cox method have been proposed with relaxed assumptions [128, 129], as well as alternative approaches such as the accelerated failure time model [130]. However, these approaches are less frequently used.

Extensions of some traditional machine learning algorithms such as survival SVM [131], and Random Survival Forest [132] have been developed to handle censored data. None of the methods described above are designed to work with raw image data and require preliminary image feature extraction. This is where convolutional neural networks have prominent advantages and researchers have started to combine feature extraction with CNNs with established statistical models to perform image-based survival modelling. The most relevant studies in the field of digital pathology [105, 24, 104, 101] have been cited in the previous chapters.

More advanced neural network-based approaches have adopted the original Cox *Partial Likelihood* for end-to-end modelling of time-to-event data. This method was first described as a DeepSurv neural network [133] and based on the work that dates back to 1995 [134]. In the following paragraphs, we will take a closer look at this method.

Formally, in survival analysis each observation i can be expressed as a triple $(x_i, t_i.\delta_i)$, where x_i is a set of covariates, $t_i > 0$ is the follow-up time or time-to-event, and $\delta_i \in 0, 1$ is the binary event indicator or censoring status. The Cox PH

model estimates the risk or a probability of the event to occur at time t, given that the individual has survived until that time. This is called a hazard function, and in the Cox framework, it is expressed as follows:

$$\lambda(t,x) = \lambda_0(t)exp(\beta^T x). \tag{8}$$

The first term of the hazard function in equation (8) is called the baseline hazard and depends only on time. The baseline hazard remains unspecified; thus, no particular survival time distribution is assumed in the model. That is one of the reasons why the Cox PH model has been commonly used. The second term depends only on the covariates x_i but not time. That fact defines the proportional hazard assumption, i.e. the effects of covariates on survival are constant over time; β is a vector of regression coefficients.

It is essential to briefly explain the estimation of the Cox PH model as it directly affects how the model can be adapted for image-based time-to-event modelling with DL. The full maximum likelihood requires to specify the baseline hazard $\lambda_0(t)$. In 1972 a partial likelihood [124] that depends only on the parameters of the model and does not depend on the underlying hazard function $\lambda_0(t)$ was proposed:

$$L(\beta,x) = \prod_{i:\delta_i=1} \left(\frac{e^{h_i}}{\sum_{j:t_j \geq t_i} e^{h_j}} \right), \tag{9}$$

where $h_i = \beta^T x_i$ - predicted risk for individual i; The corresponding log partial likelihood is:

$$l(\beta,x) = \sum_{i:\delta_i=1} \left(h_i - log \sum_{j:t_j \geq t_i} e^{h_j} \right). \tag{10}$$

Minimising the log-likelihood in equation (10) turns the outcome prediction task into a ranking problem. Several studies have used this method for direct outcome prediction from tumour tissue morphology [110, 22, 107, 135].

The likelihood function in equation (10) only considers the subjects that experienced the event by the end of follow-up. Censored observations are included only in the risk set, i.e. in the denominator of the equation (9). Importantly, equation (9) is valid only for continuous-time survival data, which is not the case in most practical applications where multiple events may occur at the same (discrete) time, i.e. resulting in ties. The most common approaches to handle tied data are Breslow [136] and Efron [137] approximations to the discrete likelihood and the exact

method that considers all possible orders of events that occurred at the same time. The exact approach becomes computationally intensive as the number of ties grows. Efron's method is considered to give a better approximation to the original partial likelihood [138]:

$$L(\beta,x) = \prod_{i:\delta_i=1} \left[\frac{\prod_{j\in H_i} e^{h_j}}{\prod_{l=0}^{m-1} \left(\sum_{j:t_j\geq t_i} e^{h_j} - \frac{l}{m} \sum_{j\in H_i} e^{h_j} \right)} \right], \tag{11}$$

where H_i is a set of individuals that failed at time i, and m is a number of ties; The corresponding log likelihood looks as follows:

$$l(\beta,x) = \sum_{i:\delta_i=1} \left(\sum_{j\in H_i} h_j - \sum_{l=0}^{m-1} \log \Big(\sum_{j:t_j\geq t_i} e^{h_j} - (l/m) \sum_{j\in H_i} e^{h_j} \Big) \right). \tag{12}$$

Still, most practical applications within cancer research and pathology have relied on the Breslow approximation or original likelihood, ignoring ties. A non-vision-based study on patient-specific kidney graft survival analysis with DL-adapted Efron's method [139].

Cox Partial Likelihood adaptation in DL has its limitations, e.g. lack of hazard proportionality [127]. A *Complete Hazard Ranking* (Guan Rank) method was proposed to address some of the limitations [140]. The Guan Rank algorithm assigns hazard ranks to all observations in the training set, including the censored ones. That provides a complete set of labels for subjects and allows to train arbitrary regression algorithms, e.g. a DL model in an end-to-end fashion [140].

3 Aims of the study

The overall aim of this doctoral thesis was to investigate whether patient outcome and tumour biomarker status can be predicted from cancer tissue morphology by deep learning applied to digitised H&E-stained tumour specimens.

Specifically, the aims were:

1. To assess whether patient outcome can be predicted from tissue morphology of breast and colorectal cancer samples using outcome and biomarker supervised deep learning

2. To study if *ERBB2* gene amplification can be predicted directly from tissue morphology in breast cancer

3. To assess if the efficacy of a molecularly targeted treatment in breast cancer can be predicted based on tissue morphological features learned through biomarker supervised deep learning

4 Materials and methods

4.1 Patient series and samples

4.1.1 Colorectal cancer tissue microarray series (I)

In Study I, we investigated whether patient outcome can be predicted based on digitised H&E-stained tissue samples of primary colorectal cancer using an outcome supervised DL approach. The samples originated from a retrospective cohort of 641 consecutive patients diagnosed with colorectal cancer. All patients underwent primary tumour resection at the Helsinki University Central Hospital in 1989-1998 [141]. Tissue cores were punched from the most representative areas of the original formalin-fixed and paraffin-embedded (FFPE) tumour blocks, i.e. typically from the least differentiated parts of the tumour and assembled into tumour TMA blocks. Then, the TMA blocks were cut into four-micrometre thick sections, stained for basic morphology (H&E) and digitised with a WS scanner.

Patient survival data, i.e. follow-up time and disease outcome were available for each of the patients. This information was obtained from the Finnish Population and Register Centre and Statistics Finland. Clinicopathological characteristics related to the patients were extracted from pathology reports and included histological grade and Dukes' stage of the disease. Additionally, each TMA spot, representing a patient's tumour, was classified by three pathologists low-risk and high-risk groups. The experts were blind to patient outcome and were guided purely by the morphology of each TMA sample. A consensus score defined by a majority vote among the three experts was derived and referred to as a Visual Risk Score. The Visual Risk Scoring was performed to allow direct comparison of expert-based and DL-based patient outcome prediction (Table 1).

A total of thirty-nine patients were excluded from the analysis. Twenty-four patient samples were detached or had no tumour tissue. Fifteen patients were excluded due to misdiagnosis or postoperative death.

Ethical approvals were obtained from The Hospital District of Helsinki and Uusimaa (Dnro HUS 226/E6/06, extension TMK02 §66 17.4.2013) and the National Supervisory Authority for Welfare and Health (Valvira Dnro 10041/06.01.03.01/2012). Written informed consent was not required because patient consent could not be obtained since the study was retrospective and the number of specimens was extensive.

Table 1: Clinicipathological characteristics of the colorectal cancer patients.

	Included Patients	All patients
No. of patients	420	641
Age at diagnosis (years)		
< 50	53 (12.6%)	77 (12%)
50 - 64	123 (29.3%)	189 (29.5%)
65 - 74	145 (34.5%)	216 (33.7%)
$\geqslant$ 75	99 (23.6%)	159 (24.8%)
Average	65.4	65.9
Gender		
Male	227 (54%)	340 (53%)
Female	193 (46%)	301 (47%)
Stage		
Dukes' A	51 (12.1%)	93 (14.5%)
Dukes' B	141 (33.6%)	231 (36%)
Dukes' C	114 (27.1%)	166 (25.9%)
Dukes' D	114 (27.1%)	149 (23.2%)
NA	0 (0%)	2 (0.3%)
Histological grade (WS)		
Low (I-II)	285 (67.9%)	439 (68.4%)
High (III-IV)	135 (32.1%)	200 (32.5%)
NA	0 (0%)	2 (0.3%)
Visual Risk (TMA)		
Low	173 (41.2%)	-
High	225 (53.6%)	-
NA	22 (5.2%)	-

4.1.2 Breast cancer FinProg tissue microarray series (II, III)

In Study II and III, breast cancer H&E-stained FFPE TMA tumour samples were collected from the FinProg original series [85] and from the FinProg validation series [142]. The original FinProg series is a nationwide cohort that consists of 93% of all breast cancer cases diagnosed in 1991 and 1992 within five selected geographical regions in Finland [143]. The FinProg validation cohort included 565 women treated for breast cancer at the Departments of Surgery and Oncology, Helsinki University Hospital in 1987-1990. The combined FinProg and FinProg validation set included 2,313 patients. A total of 1,047 patient samples were available for analysis after the exclusions (Table 2) (described in Study II, supplementary figure 1).

Table 2: Clinicipathological characteristics of the FinProg patients.

	Training and tuning N = 693		Test set patients N = 354		Included patients N = 1047		All patients N = 1299	
	N	%	N	%	N	%	N	%
Histological grade (WS)								
I	98	*14.1*	68	*19.2*	166	*15.9*	226	*17*
II	244	*35.2*	127	*35.9*	371	*35.4*	450	*35*
III	168	*24.2*	68	*19.2*	236	*22.5*	273	*21*
NA	183	*26.4*	91	*25.7*	274	*26.2*	350	*27*
ERBB2 status (CISH)								
Negative	557	*80.4*	288	*81.4*	845	*80.7*	944	*73*
Positive	136	*19.6*	66	*18.6*	202	*19.3*	216	*17*
Na							139	*10*
Estrogen receptor								
Negative	221	*31.9*	111	*31.4*	332	*31.7*	364	*28*
Positive	472	*68.1*	243	*68.6*	715	*68.3*	812	*63*
NA							123	*9*
Progesterone receptor								
Negative	326	*47.0*	163	*46.0*	489	*46.7*	539	*41*
Positive	367	*53.0*	191	*54.0*	558	*53.3*	638	*49*
NA							122	*9*
Cancer-specific survival								
Censored	483	*69.7*	254	*71.8*	737	*70.4*	979	*75*
Uncensored	210	*30.3*	100	*28.2*	310	*29.6*	205	*16*

Patient clinical information, tumour characteristics and patient outcome were retrieved from the hospital records, the Finnish Cancer Registry and Statistics Finland. Those included histological grade, tumour size, stage, axillary lymph

node status, estrogen (ER) and progesterone (PR) receptor expression, *ERBB2* gene amplification and patient survival (Table 2). Additionally, the following characteristics regarding the patient tumours were assessed by a pathologist via visual examination of the digitized tumour TMA samples:

- Degree of immune cell infiltration: Low / High
- Pleomorphism: Minimal / Moderate / Marked
- Mitotic events per 1 high-power field: <1 / 1 / >1
- Tubule formation: <10% / 10 - 75% / >75%
- Necrosis: absent / present

Similar to the colorectal cohort described above, a Visual Risk Score (low vs. high) was assigned to each patient based on the visual assessment of the corresponding H&E-stained TMA samples. Follow-up time with disease-specific outcome was available for each patient.

ER and PR receptor expression was determined for each tumour sample with immunohistochemistry [85]. Quantification of *ERBB2* gene amplification was performed by chromogenic *in situ* hybridisation (CISH). Ethical approvals were obtained from The Hospital District of Helsinki and Uusimaa (Dnro 94/13/03/02/2012) and the National Supervisory Authority for Welfare and Health (Valvira Dnro 7717/06.01.03.01/2015).

4.1.3 Breast cancer FinHer whole-slide image series (II, III)

Studies II and III also comprised 712 H&E-stained FFPE WS tumour sections from the FinHer trial (ISRCTN76560285) [144]. The trial included 1,010 women with primary breast cancer that had undergone breast cancer surgery [84]. Expression of ER, PR and *ERBB2* was determined with IHC and *ERBB2* gene amplification was confirmed with CISH (Table 3). Patients with *ERBB2*-positive cancer (N=232) were randomly assigned either to receive or not to receive adjuvant anti-*ERBB2* treatment (trastuzumab; Herceptin). The study was approved by the institutional review board (HUS 177/13/03/02/2011). Patients' written informed consent was acquired for further research to be carried out on in their tissue material.

Table 3: Clinicipathological characteristics of the FinHer patients.

	Included patients N = 712		All patients N = 1009	
	N	%	N	%
Histological grade (WS)				
I	95	*13.3*	150	*14.9*
II	276	*38.8*	397	*39.3*
III	303	*42.6*	414	*41.0*
NA	38	*5.3*	48	*4.8*
ERBB2 status (CISH)				
Negative	548	*77.0*	776	*76.9*
Positive	164	*23.0*	233	*23.1*
Estrogen receptor				
Negative	211	*29.6*	280	*27.8*
Positive	501	*70.4*	729	*72.2*
Distant disease-free survival				
Censored	593	*83.3*	846	*83.8*
Uncensored	119	*16.7*	163	*16.2*

4.2 Digitisation of samples

Digitisation of tissue samples was performed with a WS image (WSI) scanner (Pannoramic 250 FLASH, 3DHISTECH Ltd., Budapest, Hungary). The scanner was equipped with a 20x objective, numerical aperture 0.80. Acquired images were initially stored in a MIRAX WSI file format (3DHISTECH Ltd.). MIRAX files were further converted to an Enhanced Compressed Wavelet (ECW) file format (developed by Earth Resource Mapping, currently owned by Hexagon, Stockholm, Sweden). The ECW files were compressed with the ratio 1:10 and uploaded to the WS management server (Aiforia Technologies OY, Helsinki, Finland) for long-term storage, and visual exploration. Pixel size of the resulting images represented an area of $0.22\mu m$ x $0.22\mu m$.

4.3 Computer vision methods

4.3.1 Image pre-processing and augmentation

In Study I, a grid of tiles was generated from each of the digitised TMA images with an average size of 3500 x 3500 pixels. The size of each tile was 224 x 224 pixels and determined by the input size of the VGG-16 [145] convolutional neural network (CNN) used as a feature extractor. A total of 256 tiles (16 x 16 grid) overlapped by 15 pixels were extracted from each TMA image. Extracted tiles were then colour-channel-normalized according to the mean pixel values calculated on the ImageNet [146] training set: (Red: 123.68; Green: 116.779; Blue: 103.939). No image augmentation was applied in Study I.

In Study II and III, we utilized the same strategy for image pre-processing and augmentation on the FinProg and the FinHer tissue images. The size of individual TMA spot image was 3500 x 3500 pixels on average. When training on the FinProg TMA spot images, square crops were extracted at a random location. One crop of size 950 x 950 pixels per TMA spot was extracted at each training iteration, i.e. at each epoch the networks were supplied with a different set of crops that originated from various locations of the individual TMA spots included in the training set. A batch of 16 random crops constituted one training iteration for outcome prediction in Study II and biomarker prediction in Study III. Thus, in both studies, we used input tensors of size [950, 950, 3, 16] (height, width, colour channels, batch size). All input tensors were normalized with mean and standard deviation, as estimated on the training-set images: mean - (Red: 0.8198558, Green: 0.78990823, Blue: 0.91205645), standard deviation - (Red: 0.1421396, Green: 0.15343277, Blue: 0.07634846).

After image normalisation, we performed on-the-fly training image augmentations using SOLT data augmentation library (`https://github.com/MIPT-Oulu/solt`) [147] with the following parameters:

- random scaling with 0.5 probability and 0.3 scale range
- random rotation with 0.5 probability and $\pm$ 90-degree range
- random shear with 0.5 probability and 0.2 shear range
- random gamma correction with 0.5 probability and 0.3 gamma range

A centre crop of size 2100 x 2100 pixels was extracted from the FinProg TMAs at validation and test phases. No image augmentation was applied at validation and testing.

For evaluations on the FinHer patient series in Study II and Study III, the WS images first passed a quality check. We used the HistoQC [148] quality control

tool with default parameters to exclude image regions with artefacts such as out of focus and blurry areas, tissue folding, pen markings, bubbles, and coverslip edges. Non-overlapping tiles of 950 x 950 pixels with a step size of 950 pixels were extracted from the qualified regions of the WS tissue images. Since the FinHer tissue material was used solely for model evaluation, no augmentations were applied to the extracted tiles.

4.3.2 Outcome prediction for colorectal cancer

For the colorectal cancer outcome prediction task, we utilised a transfer learning approach. We used the VGG-16 convolutional neural network [145] trained on the ImageNet dataset [146] and, without additional fine-tuning, extracted intermediate activations from the second last fully-connected layer. The VGG-16 activations were extracted from each tile individually and, in total, from 256 tiles from each TMA spot image. Individual tiles were propagated through the VGG-16 network to generate 4096-bin feature vectors. These feature vectors served as input to the downstream classifiers trained to predict a five-year disease-specific outcome. Specifically, a linear Support Vector Machine (SVM), a gaussian Naive Bayes classifier and a Logistic Regression were trained in a binary classification setting to predict whether a patient survived five years after the initial diagnosis or died of cancer. These traditional machine learning algorithms were used to establish a baseline classification performance and were then compared with a proposed DL approach.

The DL method proposed consisted of a three-layer one-dimensional Long Short-Term Memory (LSTM) [42] recurrent architecture that was trained to perform the same binary classification task. The LSTM network had 264, 128 and 64 memory cells at the first, second and third hidden layers respectively. A hyperbolic tangent non-linearity was used for all hidden units. The architecture was trained with a binary cross-entropy loss function and Adadelta [54] - an adaptive learning rate optimiser using default parameters (learning rate 1.0, decay 0.0). The LSTM network performed image summarisation and classification jointly as one task.

The LSTM architecture training was performed in three-fold cross-validation with 220 training samples, 60 validation samples, and 140 samples in the test set at each fold. Regularisation techniques included Elastic Net [60] (l1: 0.005; l2: 0.005) applied at each hidden layer of the LSTM. The input and the last hidden layers adopted dropout method [65] at training phase.

In contrast to the LSTM, the other (baseline) algorithms used for classification could not integrate information from multiple image tiles, thus required additional steps to

aggregate features extracted from individual tiles. To this end, we adopted Improved Fisher Vectors (IFV) encoding [149] and produced global descriptors of TMA spots. This transformation resulted in 2048-dimensional IFV representations of TMA images and served as input to the traditional (not DL) classifiers. Finally, these classifiers we trained on the same three-fold cross-validation splits as described above.

4.3.3 Outcome and biomarker prediction in breast cancer

For breast cancer outcome and biomarker prediction tasks, we also utilised a transfer learning approach. Our DL model was built around a ResNet [47] convolutional neural network backbone, pre-trained with ImageNet [146]. In our network architecture, convolutional feature maps from the ResNet-34 backbone were globally average pooled [150] to produce a fixed-size feature vector, making the pipeline flexible with regards to the input image size. The pooled feature vectors were then passed through two types of fully connected branches, where one branch predicted disease outcome, and the other branches performed biomarker status classification. Thereby, in addition to predicting the primary endpoint, i.e. breast cancer-specific survival, the network jointly learned to predict the ER and *ERBB2* status of the tumour samples. Predicting multiple endpoints simultaneously is referred to as multi-task learning [63].

The outcome prediction problem was turned into a regression task with the mean squared error loss (MSE) using a GuanRank transformation [140]. The GuanRank is a non-parametric ranking-based technique that transforms time-to-event data into a linear space of hazard ranks representing breast cancer-specific survival for each patient. Focal loss (alpha=0.25, gamma=2) [50] was used to optimise the biomarker status prediction branches in solving the binary classification task. The focal and MSE losses were combined for end-to-end training of the entire multi-task architecture.

The outcome and biomarker prediction DL pipeline was trained on the FinProg TMA images with five-fold cross-validation. Stratified sampling in cross-validation preserved target class distributions within each fold. During the first three epochs, only the weights of fully-connected branches were optimised. Then, the last three convolutional layers of the ResNet backbone were released and trained for 100 more epochs together with the fully connected layers.

For training, we used Adam [55] – an adaptive learning rate optimisation algorithm with an initial learning rate set to 1e-4. The learning rate was decreased by a factor of 10 at epoch number 10 and number 50. The L2 regularisation term was added to

the loss function with a weight decay parameter set to 1e-3. A dropout layer [65] (p = 0.3) was introduced before the fully connected branches. The convolutional stack of our architecture was fine-tuned from the ImageNet pre-trained weights, whereas the fully connected blocks were initialised with random weights.

Evaluation of the models trained in cross-validation was performed on the FinProg test-set TMA samples and the FinHer WS samples. The FinHer tumour samples were not used for training, thus served as an independent test-set patient series.

4.3.4 Activation maps for biomarker prediction

We applied two different approaches to understand what drives network decisions in predicting *ERBB2* gene amplification from the H&E tumour samples. First, we gathered network predictions for the *ERBB2* gene amplification and overlaid those on top of the corresponding locations of the WS tissue sections from the FinHer dataset (H&E-*ERBB2* score maps). The score maps allow studying the distribution of *ERBB2*-associated morphologies within and across the tissue samples. Then, we applied a Gradient-weighted Class Activation Mapping technique - Grad-CAM [151]. The Grad-CAM identified tissue image features that were most informative to determine the *ERBB2* gene amplification status. The results from both approaches were visualised as heatmaps.

4.4 Performance evaluation and statistical analysis

Statistical analysis of time-to-event data included a non-parametric Kaplan-Meier method [123] to estimate and visualise survival profiles of patient subgroups. Time-to-event was time from the initial diagnosis to cancer-specific death or the development of distant metastasis, i.e. distant disease-free survival (DDFS). Statistical significance of the difference observed between patient survival profiles was evaluated with the logrank test [152]. To quantify the effect size between patient subgroups, we reported hazard ratios assessed with univariate and multivariate Cox Proportional Hazards regression [124]. For the sake of interpretability, we split the patients into low-risk and high-risk subgroups, using the median risk score value as a threshold. The agreement between the predicted risk scores and the observed follow-up time and censoring status was estimated with the concordance index (C-index) [153]. We also used the area under the receiver operating characteristic curve (ROC AUC) [154] to compare binary classifiers' accuracy. The Venkatraman permutation test [155] was used to compare ROC AUCs.

In the biomarker status prediction, we used ROC and Precision-Recall [156] curves

to summarise the accuracy of binary classifiers. Average precision (AP) was calculated from the Precision-Recall. Corresponding confidence intervals for ROC AUCs and APs were estimated using a stratified bootstrapping technique in 2,000 iterations. Logistic regression was used to test the independence of DL-derived biomarker predictors from other clinical covariates in the multivariate analysis.

5 Results

5.1 Colorectal cancer outcome prediction (I)

Using the logrank test, we first identified that the Digital Risk Score (deep learning) (p-value <0.001), the Visual Risk Score (p-value 0.00016), histological grade assessed on the whole-slides (p-value 0.00016), age group (p-value 0.012) and Dukes' stage (p-value <0.001) were statistically significant predictors of the disease-specific survival. Corresponding survival curves were calculated using the Kaplan-Meier method (Figure 1). Regarding the ROC AUC, the Digital Risk Score predicted disease-specific outcome with an average AUC = 0.69; the Visual Risk Score demonstrated an AUC = 0.58, whereas histological grade had an AUC=0.57. According to the Venkatraman's test, the ROC AUC demonstrated by the DL approach in predicting five-year disease-specific outcome was significantly higher than that of the Visual Risk Score (p-value 0.025) and histological grade (p-value 0.003). Thus, among the tissue-level predictors, the Digital Risk Score was significantly better in discriminating low- and high-risk patients regarding colorectal cancer outcome at five years after initial diagnosis.

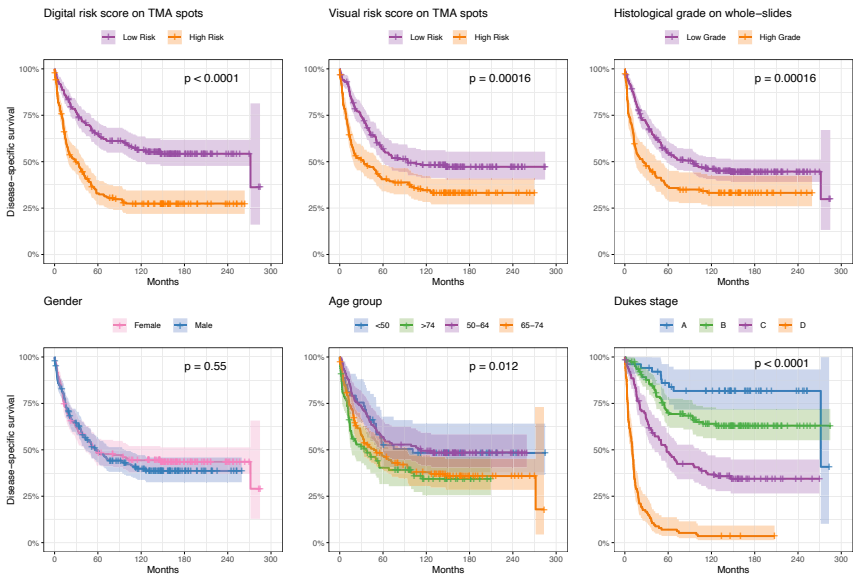


Figure 1: Kaplan-Meier survival curves based on different prognostic factors of colorectal patient outcome.

To estimate hazard ratios (HR) associated with each prognostic factor, we applied a semi-parametric Cox proportional hazards (PH) survival regression. In a univariate

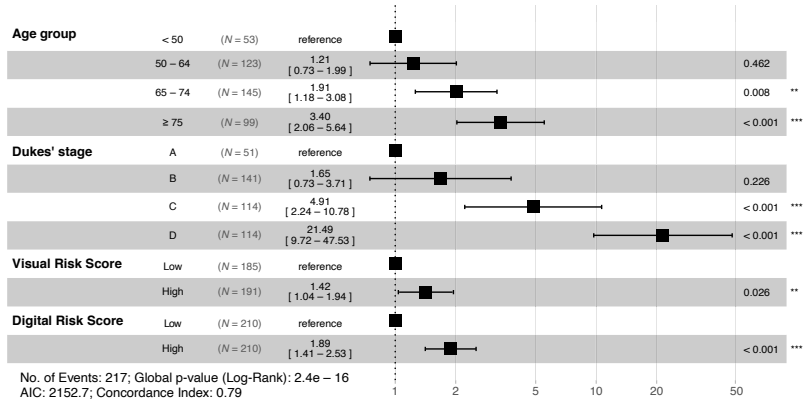


Figure 2: Multivariate Cox proportional-hazards analysis in the studied series of patients with colorectal cancer. The Digital Risk Score remained as an independent predictor of patient outcome when adjusted for other prognostic factors.

Cox PH analysis, the Digital Risk Score reached an HR of 2.3, CI 95% 1.79-3.03; the Visual Risk Score reached an HR of 1.67, CI 95% 1.28-2.19; histological grade demonstrated an HR of 1.65, CI 95% 1.30-2.15; Dukes' stage D had an HR of 20.29, CI 95% 10.44-39.44, and age at diagnosis using a cut-off at 65 years demonstrated an HR of 1.42 CI 95% 0.93-2.19. In a multivariate Cox PH analysis, gender was excluded as it was not a statistically significant predictor of survival in the univariate model. Goodness-of-fit test with the χ^2 statistics identified that histological grade violated the proportional hazards assumption (p-value = 0.014). Therefore, the final multivariate Cox PH model was stratified by histological grade (Figure 2). The Digital Risk Score remained as an independent predictor of the disease-specific survival with an HR of 1.89, CI 95% 1.41-2.53, p-value <0.001, when adjusted for other predictors.

The algorithm's performance was also evaluated on 181 hold-out patients and demonstrated comparable accuracy to that in cross-validation. In a univariate Cox PH analysis, it reached an HR of 2.11, CI 95% 1.20 - 3.73, logrank p-value 0.008, and ROC AUC 0.67 (Figure 3). In a multivariate analysis, the Digital Risk Score remained an independent predictor of disease-specific survival with an HR of 2.06, CI 95% 1.15 - 3.70, when adjusted for age group and histological grade (Figure 4). In the hold-out set of 181 patients, histological grade did not violate the proportional hazards assumption; therefore, it was included in the multivariate model.

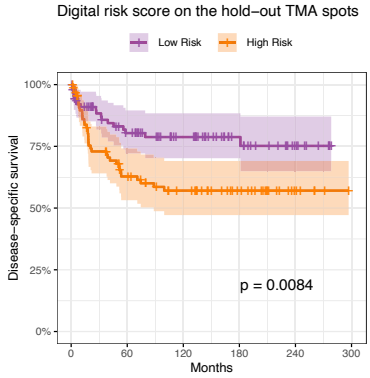


Figure 3: Kaplan-Meier survival curves for the 181 hold-out patients, stratified by high and low risk score as predicted by the deep learning algorithm (the Digital Risk Score).

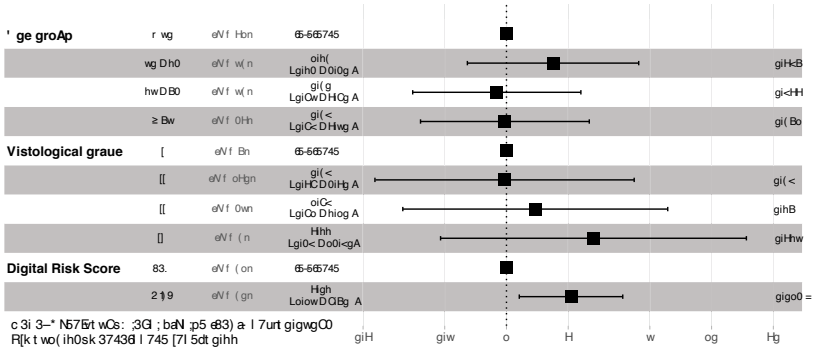


Figure 4: Multivariate Cox proportional-hazards analysis in the hold-out set of patients with colorectal cancer.

5.2 Breast cancer outcome prediction (II)

The "solo" and multitask models were separately evaluated on a test set of 354 hold-out patients from the FinProg series. In the FinProg test set, the "solo" models achieved an HR of 1.70, CI 95% 1.10 – 2.60 in a univariate Cox PH regression, p-value 0.009 and a c-index of 0.57 4. The multitask models reached an HR of 2.00, CI 95% 1.30 – 3.00, p-value < 0.001, and a c-index of 0.59 4. These results were compared to several tissue characteristics visually assessed by a pathologist on the same set of 354 TMA samples from the FinProg series. Those tissue characteristics included the subcomponents of histological grade: mitotic count, nuclear pleomorphism, tubule formation; other tissue characteristics included

tumour necrosis, TILs, and the Visual Risk Score.

Results of the univariate Cox PH analysis for each predictor are summarized in Table 4. The TMA-based grade was established from the histological grade subcomponents and reached an HR of 3.00 CI 95% 1.50 – 6.10, p-value 0.002, and a c-index of 0.60. The original histological grading assessed on WS (WS grade) by pathologists at the time of diagnosis demonstrated an HR of 4.00 CI 95% 2.00 – 8.30, p-value <0.001 and a c-index of 0.64. The presence of necrotic tissue was associated with an HR of 5.00 CI 95% 2.40 – 10.00, p-value <0.001, whereas a higher number of TILs was not a statistically significant predictor of survival in a univariate Cox PH regression. The Visual Risk Score reached an HR of 1.80 CI 95% 1.20 – 2.70, p-value 0.004 and a c-index of 0.58 (Table 4).

We then evaluated how the DL-based outcome prediction can complement visual tissue assessment by a pathologist. To this end, we separately combined the "solo" and multitask models with visual TMA-based histological grading in the Cox PH analysis. The multivariate Cox PH regression showed that the multitask model was an independent predictor of breast cancer-specific survival when adjusted for the visual TMA-based histological grade with an HR of 1.70 CI 95% 1.10 – 2.70, a p-value of 0.017 and a c-index of 0.63. A similar c-index (0.63) was observed when the multitask model was combined with the Visual Risk Score. The "solo" model was not a statistically significant predictor of breast cancer-specific survival when adjusted for TMA-based histological grade. Nor was the "solo" model statistically independent when adjusted for the Visual Risk core.

Table 4: Univariate Cox Proportional Hazards analysis of tissue characteristics assessed on tissue microarrays (TMA) within the FinProg test set. Association of the variables with breast cancer-specific survival is reported as effect size (hazard ratio, HR) and a concordance index (c-index). Prognostic performance of the "solo" and the multitask models is compared to tissue characteristics assessed by a pathologist, including histological grade assessed on whole-slide tissue sections (WS).

Variable:	N	HR	CI 95%	p-val	c-index
Mitotic count (TMA)					
Low	256	reference			
Moderate	43	1.50	0.88 - 2.70	0.132	0.57
High	31	2.00	1.10 - 3.60	0.023*	
Pleomorphism (TMA)					
Minimal	45	reference			
Moderate	193	1.90	0.86 - 4.20	0.11	0.59
Marked	92	3.00	1.34 - 6.70	0.008**	
Tubule formation (TMA)					
High	49	reference			0.54
Low	281	2.20	1.10 - 4.60	0.029*	
Histological Grade (TMA)					
I	74	reference			
II	194	2.10	1.10 - 3.80	0.022*	0.60
III	62	3.0	1.50 - 6.10	0.002**	
Histological Grade (WS)					
I	64	reference			
II	119	2.70	1.30 - 5.30	0.006**	0.64
III	61	4.00	2.00 - 8.30	<0.001***	
Tumour necrosis (TMA)					
Absent	320	reference			0.54
Present	11	5.00	2.40 - 10.00	<0.001***	
Tumour-infiltrating lymphocytes (TMA)					
Low	289	reference			0.54
High	50	1.60	0.94 - 2.60	0.083	
Visual Risk (TMA)					
Low	213	reference			0.58
High	114	1.80	1.20 - 2.70	0.004**	
Axillary lymph node status					
Negative	200	reference			0.62
Positive	128	2.40	1.60 - 3.60	<0.001***	
Tumour size (cm)	336	1.50	1.30 - 1.70	<0.001***	0.71
"Solo" Model (TMA)					
Low Risk	177	reference			
High Risk	177	1.70	1.10 - 2.60	0.009**	0.57
Multitask Model (TMA)					
Low Risk	177	reference			0.59
High Risk	177	2.00	1.30 - 3.00	<0.001***	

Then, survival analysis was expanded by combining TMA-based histological grade, necrosis and TILs as covariates in the same multivariate Cox PH regression together with the DL-based predictors. The multitask model remained independent of other covariates in predicting breast cancer-specific survival with an HR of 1.70 CI 95% 1.06 – 2.70, p-value 0.029 and a c-index of 0.66.

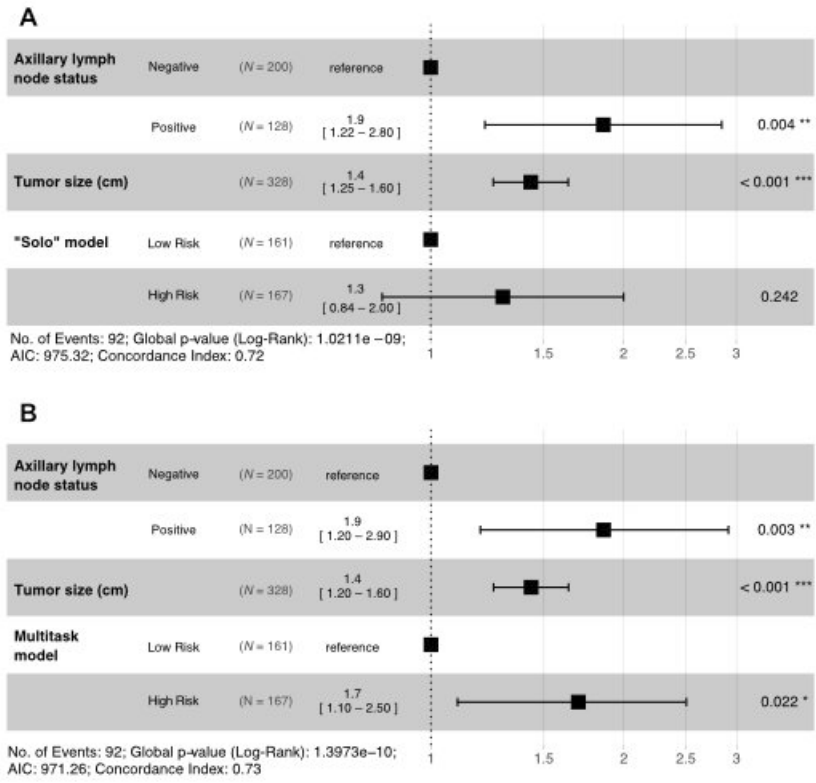


Figure 5: Multivariate Cox Proportional Hazards analysis of deep learning models together with prognostic factors related to the extent of the disease in breast cancer, i.e. spread of the cancer to axillary lymph nodes and size of the primary tumour in the FinProg test set. The results show that multitask training (B) was an independent predictor of survival as compared to outcome-supervised training only (A).

Conventional histological grading of the WS tissue samples was available for the FinProg patient's tumours. Thus, we evaluated the prognostic value of the outcome-supervised DL combined with WS histological grade. The multitask CNN remained independent of WS histological grade, whereas the "solo" model was not a significant predictor of breast cancer-specific survival. The compound multitask model adjusted for WS histological grade had a c-index of 0.66. Tumour size and axillary lymph node status were also included in the multivariate Cox PH regression where the DL predictor reached an HR of 1.70 CI 95% 1.10 – 2.50, p-value 0.022 and a c-index of 0.73 (Figure 5)

Finally, the DL models trained on the TMA samples from the FinProg patient series were applied to WSIs of whole tissue sections from the independent FinHer patient series. Univariate Cox PH regression showed that both multitask, and "solo" models were statistically significant predictors of distant disease-free survival (DDFS) in patients from the FinHer series. The "solo" model reached an HR of 1.80, CI 95% 1.30 – 2.70, a p-value of 0.002 and a c-index of 0.57. The multitask model achieved an HR of 1.70, CI 95% 1.20 – 2.60, p-value 0.003 and a c-index 0.57. We then evaluated both models in a multivariate Cox PH regression adjusted for the WS histological grade and observed that both models were statistically significant predictors of survival, independent of histological grade on the whole-slide level (Table 5). The "solo" model reached an HR of 1.70, CI 95% 1.10 – 2.50, a p-value of 0.009 and a c-index of 0.60 in a multivariate Cox PH analysis, whereas the multitask model reached an HR of 1.509, CI 95% 1.00 – 2.30), a p-value of 0.033 and a c-index of 0.59 (Table 5).

Table 5: Multivariate Cox Proportional Hazards regression of deep learning-based outcome predictions adjusted for tumour histological grade on the independent FinHer patient series.

		"Solo" Model (WS)			Multitask Model (WS)		
Variable:	N	HR	CI 95%	p-val	HR	CI 95%	p-val
Predicted Risk							
Low	337	reference			reference		
High	337	1.70	1.10-2.50	0.009	1.50	1.00-2.30	0.033
Histological Grade							
Low (I & II)	371	reference			reference		
High (III)	303	1.60	1.10-2.30	0.022	1.50	1.00-2.20	0.037
		c-index 0.60			c-index 0.59		

5.3 Biomarker prediction in breast cancer (III + unpublished)

5.3.1 Receptor status prediction

Deep CNN models were separately trained to predict ER and PR status, and *ERBB2* gene amplification directly from the H&E-stained FinProg TMA samples. The models were then evaluated on two datasets: the FinProg holdout patients (N=354) and the FinHer patient cohort (N=712) (Figure 6).

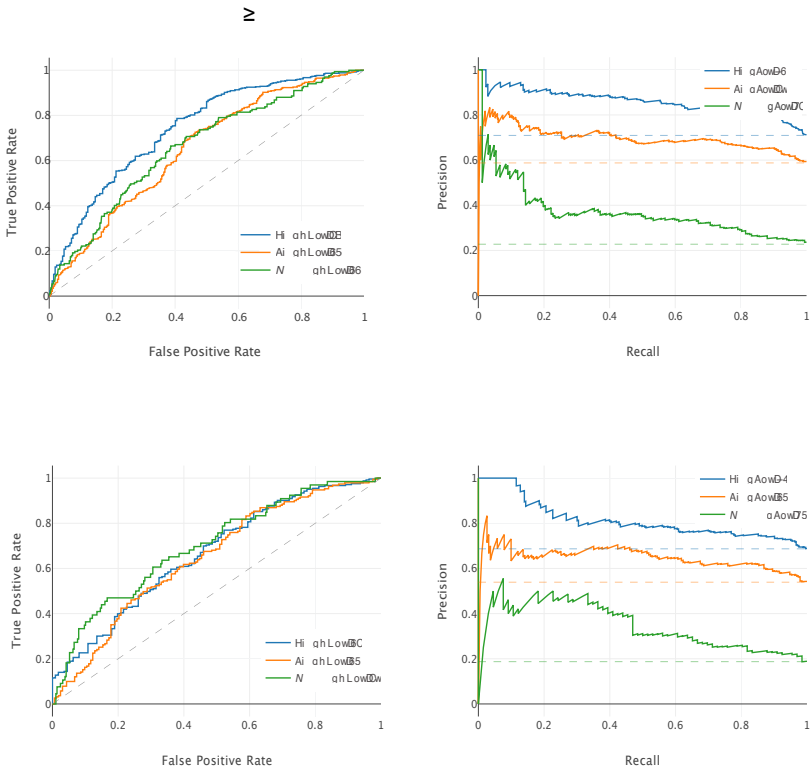


Figure 6: Receiver operating characteristic (ROC) and Precision-Recall analyses on 712 FinHer patients (A) and on 354 FinProg holdout patients (B). Dashed lines depict the performance of random classifier an the baseline precision on the ROC and PRec plots accordingly.

Receiver Operating Characteristic (ROC) and Precision-Recall curves for each of the three binary predictors are shown (Figure 6). The area under the ROC curve of each predictor is compared to that of a random classifier - AUC 0.5. The average precision (AP) of each classifier is compared to a baseline precision - a fraction of the positive observations in the test set.

On the FinHer independent test set we observed that ER was predicted with an AUC of 0.74 (CI 95% 0.70 – 0.78) and AP of 0.86 (CI 95% 0.83 – 0.89) with a baseline AP of 0.71; PR was predicted with AUC of 0.65 (CI 95% 0.61 – 0.69) and AP of 0.70 (CI 95% 0.67 – 0.73) with baseline AP 0.59; and *ERBB2* was predicted with an AUC of 0.66 (CI 95% 0.62 – 0.71) and AP of 0.37 (CI 95% 0.33 – 0.44) with a baseline AP of 0.23.

5.3.2 *ERBB2*-linked morphology and prediction of trastuzumab treatment efficacy

In the FinHer trial, chromogenic *in situ* hybridisation (CISH) *ERBB2*-positive patients were randomised to receive or not to receive adjuvant trastuzumab (Herceptin) as a part of the clinical trial. Survival analysis was performed separately in both subgroups to identify whether *ERBB2*-associated morphology was predictive of DDFS. Kaplan-Meier curves and corresponding hazard ratios are shown (Figure 7). DL-derived *ERBB2* score (H&E *ERBB2* score) was dichotomized into *high* and *low* categories reflected by two curves on the Kaplan-Meier plots (Figure 7). The graphs show that FinHer patients with a high H&E *ERBB2* score and who received trastuzumab had a more favourable outcome than patients with a low score (HR, 0.37; CI 95%, 0.15-0.93; p-value 0.034). Among CISH *ERBB2*-positive patients that did not receive trastuzumab, a high H&E *ERBB2* score indicated a less favourable DDFS (HR, 2.03; CI 95%, 0.69-5.94; p-value 0.20) as compared to a low score (Figure 7).

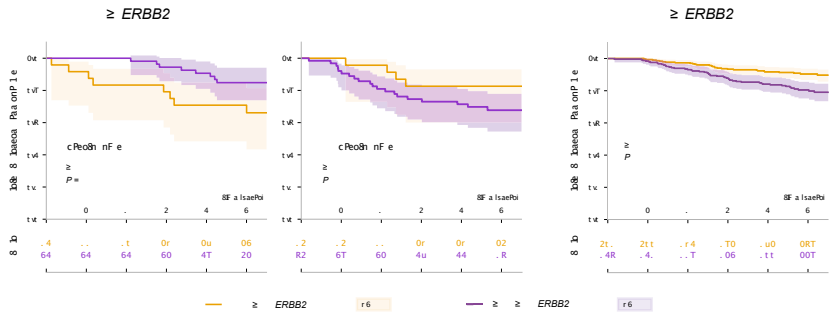


Figure 7: Morphology-based H&E-*ERBB2* score and distant disease-free survival in the FinHer trial series. (A) Evaluation of H&E-*ERBB2* scores and distant disease-free survival (DDFS) in patients with *ERBB2*-positive breast cancer as determined by chromogenic in situ hybridization (CISH). (B) CISH *ERBB2*-negative patients in the FinHer series were stratified by the H&E-*ERBB2* score. None of the CISH *ERBB2*-negative patients received adjuvant trastuzumab.

5.3.3 Activation maps for *ERBB2* gene amplification

We used explainable-AI techniques to provide explainability behind the H&E-based *ERBB2* predictions. First, we overlayed a heatmap of *ERBB2* scores on top of the original H&E input images and referred to them as the score maps. Then, we applied the Grad-CAM method to identify which areas of the H&E images were most relevant in predicting the *ERBB2* amplification.

The resulting score maps present an overview of how the *ERBB2* gene associated morphologies predicted by the DL model are spatially distributed across the WS H&E tissue sample. The analysis of the H&E-*ERBB2* score maps and corresponding Grad-CAM heatmaps showed significant heterogeneity of the *ERBB2*-associated morphologies discovered by the DL algorithm. The Grad-CAM activation maps indicated that tumour epithelium and in situ carcinoma components were most predictive of *ERBB2* gene amplification. Individual epithelial cells and fibroblasts in the stromal regions also appeared informative in predicting the *ERBB2* gene amplification (Figure 8).

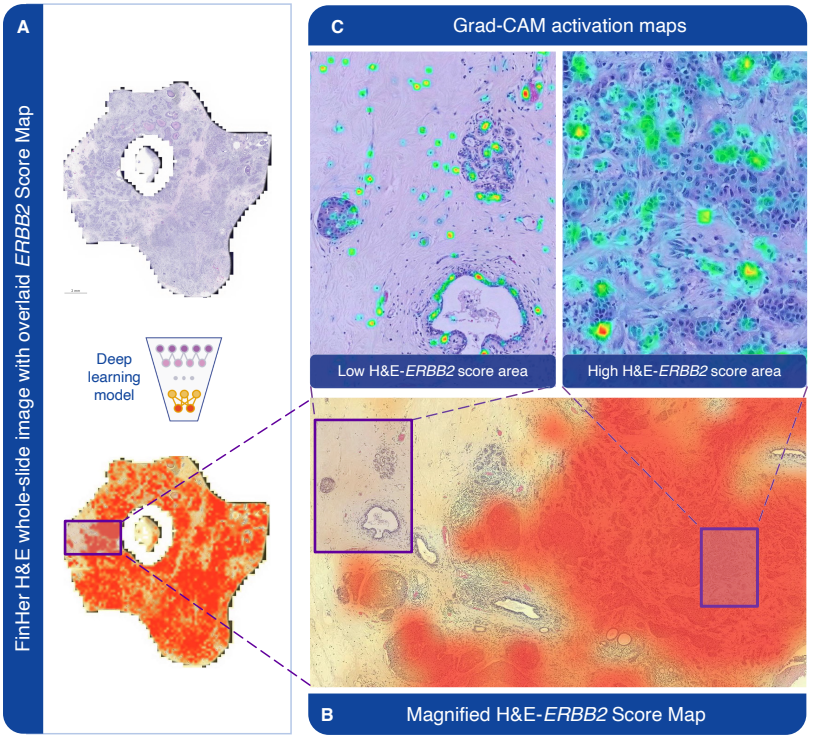


Figure 8: A hematoxylin-eosin (H&E)-based *ERBB2* score map and Grad-CAM activation maps for an individual whole-slide (WS) FinHer patient sample. (A) H&E-*ERBB2* score map as predicted by the deep learning (DL) algorithm. The score map was overlaid as a heatmap on top of the H&E-stained WS tissue sample, representing variable levels of the DL-derived *ERBB2* score. (B) Magnified image of the corresponding score map, representing high H&E-*ERBB2* scores with red. (C) Grad-CAM activation maps of the regions predicted to have a low H&E-*ERBB2* score (left) and a high H&E-*ERBB2* score (right). The sample presented is *ERBB2*-positive by chromogenic in situ hybridization and had a high overall H&E-*ERBB2* score.

6 Discussion

Recent advances in machine learning applied to digital image analysis, particularly DL-based methods [1], enable training of algorithms for classification and quantification tasks in the analysis of tissue specimens with the aim of improving the accuracy of tumour profiling and cancer diagnostics. Such tasks include counting cells [87, 16, 13], assessing the grade of tumour differentiation [88, 89], and segmentation of specific tissue entities, such as epithelium and stroma [90, 157]. These tissue characteristics are often assessed with the aim to achieve better patient stratification, predict patient outcomes, and in decision-making related to patient treatment. However, training DL algorithms for these tasks with supervised machine learning typically requires extensive and laborious image annotations by human experts. In this thesis, we evaluated if it is possible to bypass the expert-supervised tissue characterisation and, instead, train the algorithms to directly predict clinically relevant endpoints such as disease outcome or response to treatment. Access to digitised large-scale retrospective patient cohorts with outcome data, improved computational capacity based on massively parallel graphics processing units (GPUs), and novel DL methods inspired us to evaluate and study outcome-supervised learning in colorectal and breast cancer.

Analysis of tumour specimens through conventional microscopy is prone to subjectiveness and results in intra- and inter-observer variability [158, 159, 160]. A global shortage of experts and an increasing number of cancer cases lead to a delayed diagnosis for cancer patients [161, 162, 163]. There are opportunities to address these challenges through digitisation of pathology and application of novel DL-based tools.

The goal of this thesis was to explore the use of DL in image-based tissue diagnostics. Specifically, we investigated whether patient outcomes and molecular biomarkers can be predicted directly from digitised tissue samples stained with H&E to visualise basic tissue morphology. We evaluated whether patient outcome can be predicted directly from H&E-stained digitised specimens of colorectal tumours. While earlier attempts to predict disease outcome in breast cancer [99] used conventional computer vision techniques for image feature extraction, we were among the first to apply DL approaches in breast cancer outcome prediction. We used the VGG-16 [164] architecture pre-trained on the ImageNet dataset [45] and performed transfer learning by tuning the CNN to predict the five-year disease-specific outcome in patients with colorectal cancer.

Another novelty proposed in our study was the adaptation of a recurrent LSTM architecture [42] to summarise spatial patch-level information across TMA samples.

Spatial feature aggregation using the LSTM was compared to established methods where patch-level features were first summarised using Improved Fisher Vector encoding [149] and then passed to traditional ML classifiers: support vector machine, naive bayes and logistic regression. We observed that the LSTM approach surpassed the accuracy of traditional classifiers in terms of hazard ratio and ROC AUC. Importantly, the LSTM adaptation facilitated the interpretability of our method. Inspired by the work on explainable memory design in LSTM units [165], we visualised how a combined CNN+LSTM architecture segregates tissue patches, e.g. tumour epithelial cells, from non-informative areas of the TMA spots, such as image background. Our DL approach demonstrated a super-human performance compared to a consensus of three pathologists in TMA-based outcome prediction in colorectal cancer. The DL-based outcome predictor was independent of histological grade, age at diagnosis, and stage of the disease.

We then extended our initial experiments on colorectal cancer in several aspects and to another type of cancer. Instead of using a single patient series, we incorporated two independent multicentre breast cancer patient series in the analysis. Building partly upon previous studies on breast cancer [23], we also included evaluation of the generalisation of the algorithms by using one patient series (FinProg) for training and another (FinHer) for external validation. Analysis of the WS tissue sections from the FinHer cohort suggested that algorithms trained on TMAs generalise to the highly heterogeneous tissue morphologies intrinsic to the large WS tissue areas. The multitask algorithm trained on TMAs remained a statistically significant predictor of patient outcome and independent of other established prognostic factors, e.g. histological grade, tumour size and axillary lymph node status. Statistical independence from other prognostic factors suggests that DL can extract prognostic features that complement features extracted through visual examination of tissue samples. Thus, these DL-based prognostic features can be combined with expert-derived features to reach higher accuracy in patient outcome prediction. We additionally combined the DL-based outcome predictor with several tissue entities assessed by a pathologist on the FinProg TMA samples. Those entities included components of histological grade, i.e. mitotic count, nuclear pleomorphism, and tubule formation [166]. We conclude that the DL algorithm provides statistically significant and independent prognostic information and complement expert knowledge, resulting in the improved overall accuracy of patient outcome prediction.

The multitask DL algorithm was trained to jointly predict breast cancer outcome and commonly used molecular biomarkers in breast cancer diagnostics, namely ER status and *ERBB2* gene amplification directly from H&E-stained tumour specimens. As both biomarkers reflect patient prognosis, we hypothesised that

auxiliary supervision by the ER and *ERBB2* status could facilitate neural network training and eventually lead to improved accuracy and better generalisation of the algorithms. In contrast to the purely outcome-supervised training, the multitask approach enabled the extraction of image features that remained independent of the expert-derived features, as described in the previous paragraph. Our observations suggest that information about the molecular alterations related to the ER status and *ERBB2* gene overexpression are encoded in the morphology of breast tumours, and a machine learning-based approach allows to extract valuable prognostic and predictive information, which is not readily discernible by a human eye alone on H&E-stained samples.

The ability to predict the expression of specific proteins and amplification status of genes directly from tissue morphology has recently introduced novel possibilities for H&E tissue-based diagnostics [113]. In addition, it was shown that RNA expression can be predicted from WS H&E samples across 28 different tumour types, including breast and colorectal cancers [21]. In the current thesis, we validated that the ER and PR status and *ERBB2* amplification can be predicted directly from the H&E-stained samples. Similar recent studies on molecular biomarkers in breast cancer quantified directly from basic morphology images support our findings by reporting comparable accuracy [113, 114, 115, 95]. The assessment of ER, PR and *ERBB2* status is critical in clinical decision-making on molecularly target therapies. For example, overexpression and/or amplification of the *ERBB2* gene is detected in about 20% of breast cancers, and can be targeted by an anti-*ERBB2* therapy with trastuzumab. We went beyond the H&E-based prediction of *ERBB2* gene amplification and explored whether morphological features associated with *ERBB2* gene amplification could be directly linked to the efficacy of trastuzumab therapy. Our H&E based predictor of *ERBB2* amplification also identified patient subgroups associated with higher or lower efficacy of anti-*ERBB2* treatment, based on the tissue morphology and independent of the *ERBB2* status determined by the CISH. Thereby, *ERBB2*-linked morphology may contain therapy-predictive information to complement established CISH-based molecular profiling. Thus, we hypothesize that identification of DL-based morphological features predictive of the molecular status also can predict the efficacy of a molecularly targeted treatment and should be further studied as biomarkers for clinical decision-making in breast cancer.

It is crucial to generate visual clues of what morphological features the DL-based algorithms have learned in order to achieve more explainable outputs of the analyses. We present attempts to achieve this in the form of activation maps. We adapted both DL score maps and an established method called Grad-CAM [151], which show that the algorithms specifically focused on the malignant epithelial cells and paid less attention to stromal regions. Interpretability and explainability of the DL algorithms

are highly important, especially in the context of clinical decision-making and future studies have to address this even more rigorously.

In the colorectal cancer outcome prediction study, the main limitation was that we analysed a single-centre patient series with a relatively limited number of patients. The tumour samples were represented by small TMA core sections manually selected from the original whole tissue block. Analysis of the WS tumour sections and tissue material from independent patient series is essential to evaluate the generalisation of the DL models [108].

Our DL model was trained to predict a binary endpoint at the five-year cut-off from the initial diagnosis. Dichotomising the patients into survivors and non-survivors simplifies statistical learning, but at the same time, disregards the follow-up time distribution of the patient series. This simplification could lead to limited prognostic accuracy of the DL models. We addressed some of these limitations in the breast cancer outcome prediction studies. Specifically, we incorporated analysis of the WS sections and used an independent patient series to evaluate the generalisation of the trained algorithm. We also utilised the GuanRank method [140] to incorporate the entire spectrum of time-to-event data into the learning procedure.

In breast cancer outcome prediction, cancer-specific survival was used to train the algorithm on the FinProg data, whereas distant disease-free survival was used as an endpoint for evaluation on the FinHer series. Although a strong correlation has been shown between disease-free and overall survival in studies on early breast cancer [167], the strength of correlation between breast cancer-specific survival and distant disease-free survival remains to be established. The tissue samples used in our breast cancer study were centrally scanned using the same instrument. Thereby, we could not evaluate how well the trained DL algorithms generalise to image data obtained with different instruments.

In our studies on breast cancer, we did not have access to consecutive tissue sections with *ERBB2* gene amplification results. For example, consecutive slides or multiplexed analysis of the same H&E tissue section with CISH, or *ERBB2* protein expression assessed with immunohistochemistry, or in situ sequencing could assess if the DL-learned *ERBB2*-associated and treatment efficacy predictive features in the H&E morphology originate from the specific cells that exhibit the gene amplification or represent some more general tissue features. Overall accuracy, especially the specificity of the *ERBB2* predictions, could be improved by incorporating more training samples from external datasets. The heterogeneity of the DL-based H&E *ERBB2* features and their effect on survival should also be analysed in WS tissue section and include more patients. To confirm our findings, analyses of external patient series treated with adjuvant or neoadjuvant

anti-*ERBB2*-targeted therapy are also needed. In future research, the prognostic accuracy and generalisation of the DL models can be further improved by training DL algorithms on datasets that cover an even larger spectrum of variations of tissue morphologies, including training on WSIs of whole tissue sections. Quantifying established prognostic features, such as mitotic count, tubule formation, nuclear pleomorphism, and TILs using machine learning algorithms instead of visual assessment could further improve the accuracy, consistency and reproducibility of outcome prediction. A combination of computationally quantified prognostic features with features learned through end-to-end outcome supervised learning should be addressed in future studies.

Taken together, we demonstrate how tissue-based DL can aid in the exploration and identification of biomarkers predictive of outcome and treatment efficacy in cancer. We showed that the machine learning-based approach allows extracting valuable information directly from the basic morphology of tumours that is not readily discernible by a human eye alone. Computer-assisted tissue analysis can complement the conventional microscopic and molecular analysis of tumour tissue rather than replace these methods. Future studies should further explore these novel imaging biomarkers to achieve improved diagnostic and prognostic accuracy and stratification of cancer patients according to the expected efficacy of targeted treatments. Finally, a combination of clinicopathological, molecular, genomic and tissue-based features assessed in large-scale datasets can provide improved methods for precision medicine and help to better understand the underlying biological mechanisms of cancer.

7 Conclusions

In this doctoral thesis, we developed and evaluated deep learning computer-vision approaches for outcome and biomarker prediction directly based on tissue morphology in digitised tumour samples. The main conclusions of our work are as follows:

1. Cancer patient outcome prediction is feasible with deep learning applied to conventional hematoxylin-eosin-stained tumour samples.

2. Deep learning can identify novel tissue features predictive of protein expression and gene amplification directly from tumour morphology.

3. Established prognostic factors and tumour features extracted with deep learning approaches can complement each other and lead to more accurate outcome prediction and interpretable tumour tissue analysis.

Acknowledgements

Over the course of my doctoral studies, I was privileged to meet and work with amazing people. So I am taking these last pages of the book to express my warmest appreciation to those who guided, supported, motivated me, and simply was always nearby.

First and foremost, I am sincerely grateful to my supervisors Johan Lundin and Nina Linder for their guidance, encouragement, constructive criticism and friendly advice during these years. I would like to thank you for always giving me the freedom to explore and for the opportunities you provided beyond my thesis project.

I express my gratitude to Lee Cooper from Northwestern University for accepting the invitation to take the role of the opponent. I also would like to thank the pre-examiners of the dissertation, Esa Rahtu from Tampere University of Technology and Mattias Rantalainen from Karolinska Institutet for their valuable and thorough comments on my dissertation.

This work was conducted at the Institute for Molecular Medicine Finland (FIMM), University of Helsinki, and I would like to acknowledge Olli Kallioniemi, Jaakko Kaprio and Mark Daly who have acted as directors of FIMM and provided an exceptional environment for academic work. I also acknowledge the financial support from Biomedicum Foundation, Orion Research Foundation, and K. Albin Johanssons stiftelse. In addition, our studies have received funding from Sigrid Jusélius Foundation and iCAN- Digital Precision Cancer Medicine Flagship.

The work presented in this thesis would not have been possible without a big group of collaborators I have had the pleasure to work with. I want to thank Clare Verrill, Stig Nordling, Heikki Joensuu, Aleksei Tiulpin, Panu Kovanen, Harri Sihto, Jorma Isola, Tiina Lehtimäki, Pirkko-Liisa Kellokumpu-Lehtinen, and Karl von Smitten for successful collaboration.

During these years, I have been fortunate to be surrounded by outstanding, talented and extremely supportive colleagues in our research group – Hakan, Oscar, Antti, Micke, Sebastian, Klaus and Otto. Thank you for your support, scientific and unscientific discussions, lunches together and time outside the lab. I have learned a lot from each of you. I also want to express my gratitude to our ex-group members – Riku and Margarita. I have been lucky to meet you and to work together with you at the beginning of my PhD.

I would also like to thank all the amazing people I met at FIMM and in Biomedicum – Annabrita, Katja, Teijo, Lassi, Sami, Heikki, Lea, Shabbeer, Tanya, James, Denis,

Maria, Evgeny, Sergey, Emmy, Tero, Gretchen, and many others who created a family-like atmosphere and working environment at FIMM during these years.

Special thanks go to my friends who have been exposed to all my ups and downs along the way – Bulat, Chiara, Andrew and Ilida. Thank you for your support and friendship. Of course, none of this would have been possible without my friends who have been with me since the dawn of time – Dimas F, Ivan V, Sergey J, Maxim S, Alexey and Olia K, Alex G. I am very grateful to have you all in my life!

Finally, I would like to deeply thank my parents Svetlana and Viktor. Thank you for your support and care, not only during my PhD but throughout my entire life. Most of all, my warmest gratitude goes to Keti for the endless love, patience, encouragement, and positive energy you gave me throughout this PhD journey. Thank you!

Dmitrii B.
Espoo, January 2022

References

Page(s)

[1] LeCun, Y, Bengio, Y, & Hinton, G. (2015) Deep learning. *Nature* **521**, 436–444. 1, 3, 4, 39

[2] Hajjar, A. E & Rey, J.-F. (2020) Artificial intelligence in gastrointestinal endoscopy: general overview. *Chinese Medical Journal* **133**, 326–334. 1

[3] Solomou, A, Apostolos, A, & Ntoulias, N. (2020) Artificial intelligence in magnetic resonance imaging: A feasible practice? *Journal of Medical Imaging and Radiation Sciences* **51**, 501–502. 1

[4] Acs, B, Rantalainen, M, & Hartman, J. (2020) Artificial intelligence as the next step towards precision pathology. *Journal of Internal Medicine* **288**, 62–81. 1, 3, 11

[5] Cui, M & Zhang, D. Y. (2021) Artificial intelligence and computational pathology. *Laboratory Investigation* **101**, 412–422. 1, 11

[6] Bejnordi, B. E, Veta, M, van Diest, P. J, van Ginneken, B, Karssemeijer, N, Litjens, G, van der Laak, J. A. W. M, Hermsen, M, Manson, Q. F, Balkenhol, M, Geessink, O, Stathonikos, N, van Dijk, M. C, Bult, P, Beca, F, Beck, A. H, Wang, D, Khosla, A, Gargeya, R, Irshad, H, Zhong, A, Dou, Q, Li, Q, Chen, H, Lin, H.-J, Heng, P.-A, Haß, C, Bruni, E, Wong, Q, Halici, U, Ümit Öner, M, Cetin-Atalay, R, Berseth, M, Khvatkov, V, Vylegzhanin, A, Kraus, O, Shaban, M, Rajpoot, N, Awan, R, Sirinukunwattana, K, Qaiser, T, Tsang, Y.-W, Tellez, D, Annuscheit, J, Hufnagl, P, Valkonen, M, Kartasalo, K, Latonen, L, Ruusuvuori, P, Liimatainen, K, Albarqouni, S, Mungal, B, George, A, Demirci, S, Navab, N, Watanabe, S, Seno, S, Takenaka, Y, Matsuda, H, Phoulady, H. A, Kovalev, V, Kalinovsky, A, Liauchuk, V, Bueno, G, Fernandez-Carrobles, M. M, Serrano, I, Deniz, O, Racoceanu, D, & and, R. V. (2017) Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199. 1

[7] Halicek, M, Shahedi, M, Little, J. V, Chen, A. Y, Myers, L. L, Sumer, B. D, & Fei, B. (2019) Head and neck cancer detection in digitized whole-slide histology using convolutional neural networks. *Scientific Reports* **9**. 1

[8] Pham, H. H. N, Futakuchi, M, Bychkov, A, Furukawa, T, Kuroda, K, & Fukuoka, J. (2019) Detection of lung cancer lymph node metastases from whole-slide histopathologic images using a two-step deep learning approach. *The American Journal of Pathology* **189**, 2428–2439. 1

[9] Hekler, A, Utikal, J. S, Enk, A. H, Solass, W, Schmitt, M, Klode, J, Schadendorf, D, Sondermann, W, Franklin, C, Bestvater, F, Flaig, M. J, Krahl, D, von Kalle, C, Fröhling, S, & Brinker, T. J. (2019) Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *European Journal of Cancer* **118**, 91 – 96. 1, 3

[10] Wei, J. W, Tafe, L. J, Linnik, Y. A, Vaickus, L. J, Tomita, N, & Hassanpour, S. (2019) Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Scientific Reports* **9**. 1

[11] Iizuka, O, Kanavati, F, Kato, K, Rambeau, M, Arihiro, K, & Tsuneki, M. (2020) Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Scientific Reports* **10**. 1

[12] Balkenhol, M. C. A, Tellez, D, Vreuls, W, Clahsen, P. C, Pinckaers, H, Ciompi, F, Bult, P, & van der Laak, J. A. W. M. (2019) Deep learning assisted mitotic counting for breast cancer. *Laboratory Investigation* **99**, 1596–1606. 1, 3

[13] Mahmood, T, Arsalan, M, Owais, M, Lee, M. B, & Park, K. R. (2020) Artificial intelligence-based mitosis detection in breast cancer histopathology images using faster r-CNN and deep CNNs. *Journal of Clinical Medicine* **9**, 749. 1, 11, 39

[14] Arunachalam, H. B, Mishra, R, Daescu, O, Cederberg, K, Rakheja, D, Sengupta, A, Leonard, D, Hallac, R, & Leavey, P. (2019) Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models. *PLOS ONE* **14**, e0210706. 1

[15] Turkki, R, Linder, N, Kovanen, P, Pellinen, T, & Lundin, J. (2016) Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples. *Journal of Pathology Informatics* **7**, 38. 1, 11

[16] Linder, N, Taylor, J. C, Colling, R, Pell, R, Alveyn, E, Joseph, J, Protheroe, A, Lundin, M, Lundin, J, & Verrill, C. (2018) Deep learning for detecting tumour-infiltrating lymphocytes in testicular germ cell tumours. *Journal of Clinical Pathology* **72**, 157–164. 1, 3, 11, 39

[17] Stenman, S, Bychkov, D, Kucukel, H, Linder, N, Haglund, C, Arola, J, & Lundin, J. (2021) Antibody supervised training of a deep learning based algorithm for leukocyte segmentation in papillary thyroid carcinoma. *IEEE Journal of Biomedical and Health Informatics* **25**, 422–428. 1, 3, 11

[18] Echle, A, Rindtorff, N. T, Brinker, T. J, Luedde, T, Pearson, A. T, & Kather, J. N. (2020) Deep learning in cancer pathology: a new generation of clinical biomarkers. *British Journal of Cancer* **124**, 686–696. 1, 11, 13

[19] Kather, J. N, Pearson, A. T, Halama, N, Jäger, D, Krause, J, Loosen, S. H, Marx, A, Boor, P, Tacke, F, Neumann, U. P, Grabsch, H. I, Yoshikawa, T, Brenner, H, Chang-Claude, J, Hoffmeister, M, Trautwein, C, & Luedde, T. (2019) Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine* **25**, 1054–1056. 1, 13

[20] He, B, Bergenstråhle, L, Stenbeck, L, Abid, A, Andersson, A, Borg, Å, Maaskola, J, Lundeberg, J, & Zou, J. (2020) Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering* **4**, 827–834. 1

[21] Schmauch, B, Romagnoni, A, Pronier, E, Saillard, C, Maillé, P, Calderaro, J, Kamoun, A, Sefta, M, Toldo, S, Zaslavskiy, M, Clozel, T, Moarii, M, Courtiol, P, & Wainrib, G. (2020) A deep learning model to predict RNA-seq expression of tumours from whole slide images. *Nature Communications* **11**. 1, 13, 41

[22] Mobadersany, P, Yousefi, S, Amgad, M, Gutman, D. A, Barnholtz-Sloan, J. S, Vega, J. E. V, Brat, D. J, & Cooper, L. A. D. (2018) Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences* **115**, E2970–E2979. 1, 12, 15

[23] Turkki, R, Byckhov, D, Lundin, M, Isola, J, Nordling, S, Kovanen, P. E, Verrill, C, von Smitten, K, Joensuu, H, Lundin, J, & Linder, N. (2019) Breast cancer outcome prediction with tumour tissue images and machine learning. *Breast Cancer Research and Treatment* **177**, 41–52. 1, 11, 12, 40

[24] Bychkov, D, Linder, N, Turkki, R, Nordling, S, Kovanen, P. E, Verrill, C, Walliander, M, Lundin, M, Haglund, C, & Lundin, J. (2018) Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific Reports* **8**. 1, 12, 14

[25] Kather, J. N, Krisam, J, Charoentong, P, Luedde, T, Herpel, E, Weis, C.-A, Gaiser, T, Marx, A, Valous, N. A, Ferber, D, Jansen, L, Reyes-Aldasoro, C. C, Zörnig, I, Jäger, D, Brenner, H, Chang-Claude, J, Hoffmeister, M, & Halama, N. (2019) Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine* **16**, e1002730. 1, 12

[26] Yu, K.-H, Zhang, C, Berry, G. J, Altman, R. B, Ré, C, Rubin, D. L, & Snyder, M. (2016) Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications* **7**. 1, 11

[27] Wulczyn, E, Steiner, D. F, Xu, Z, Sadhwani, A, Wang, H, Flament-Auvigne, I, Mermel, C. H, Chen, P.-H. C, Liu, Y, & Stumpe, M. C. (2020) Deep learning-based survival prediction for multiple cancer types using histopathology images. *PLOS ONE* **15**, e0233678. 1, 12

[28] Nicolini, A, Ferrari, P, & Duffy, M. J. (2018) Prognostic and predictive biomarkers in breast cancer: Past, present and future. *Seminars in Cancer Biology* **52**, 56–73. 1, 10

[29] Hayes, D. F. (2019) HER2 and breast cancer — a phenomenal success story. *New England Journal of Medicine* **381**, 1284–1286. 1, 10

[30] Oldenhuis, C, Oosting, S, Gietema, J, & de Vries, E. (2008) Prognostic versus predictive value of biomarkers in oncology. *European Journal of Cancer* **44**, 946–953. 1, 9

[31] Chang, J. Y & Ladame, S. (2020) in *Bioengineering Innovative Solutions for Cancer*. (Elsevier), pp. 3–21. 1

[32] Biggi, G & Stilgoe, J. (2021) Artificial intelligence in self-driving cars research and innovation: A scientometric and bibliometric analysis. *SSRN Electronic Journal*. 3

[33] Stahlberg, F. (2020) Neural machine translation: A review. *Journal of Artificial Intelligence Research* **69**, 343–418. 3

3 [34] Cao, L. (2020) AI in finance: A review. *SSRN Electronic Journal*.

3 [35] Mazzone, M & Elgammal, A. (2019) Art, creativity, and the potential of artificial intelligence. *Arts* **8**, 26.

3 [36] Chen, M & Decary, M. (2019) Artificial intelligence in healthcare: An essential guide for health leaders. *Healthcare Management Forum* **33**, 10–18.

3 [37] Ronneberger, O, Fischer, P, & Brox, T. (2015) U-net: Convolutional networks for biomedical image segmentation. *CoRR* **abs/1505.04597**.

3 [38] Rosenblatt, F. (1958) The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* **65**, 386–408.

3 [39] Glorot, X & Bengio, Y. (2010) *Understanding the difficulty of training deep feedforward neural networks.*, JMLR Proceedings eds. Teh, Y. W & Titterington, D. M. (JMLR.org), Vol. 9, pp. 249–256.

3 [40] Glorot, X, Bordes, A, & Bengio, Y. (2011) *Deep Sparse Rectifier Neural Networks*, Proceedings of Machine Learning Research eds. Gordon, G, Dunson, D, & Dudík, M. (PMLR, Fort Lauderdale, FL, USA), Vol. 15, pp. 315–323.

4 [41] Cun, Y. L, Jackel, L, Boser, B, Denker, J, Graf, H, Guyon, I, Henderson, D, Howard, R, & Hubbard, W. (1989) Handwritten digit recognition: applications of neural network chips and automatic learning. *IEEE Communications Magazine* **27**, 41–46.

4, 24, 39 [42] Hochreiter, S & Schmidhuber, J. (1997) Long short-term memory. *Neural Computation* **9**, 1735–1780.

4 [43] LeCun, Y, Kavukcuoglu, K, & Farabet, C. (2010) *Convolutional networks and applications in vision*. (IEEE).

4, 5 [44] Goodfellow, I, Bengio, Y, & Courville, A. (2016) *Deep Learning*. (MIT Press). `http://www.deeplearningbook.org`.

4, 39 [45] Krizhevsky, A, Sutskever, I, & Hinton, G. E. (2012) *ImageNet Classification with Deep Convolutional Neural Networks* eds. Pereira, F, Burges, C. J. C, Bottou, L, & Weinberger, K. Q. (Curran Associates, Inc.), Vol. 25.

4, 12 [46] Simonyan, K & Zisserman, A. (2015) *Very Deep Convolutional Networks for Large-Scale Image Recognition* eds. Bengio, Y & LeCun, Y.

4, 25 [47] He, K, Zhang, X, Ren, S, & Sun, J. (2016) *Deep Residual Learning for Image Recognition*. pp. 770–778.

4 [48] Szegedy, C, Wei Liu, Yangqing Jia, Sermanet, P, Reed, S, Anguelov, D, Erhan, D, Vanhoucke, V, & Rabinovich, A. (2015) *Going deeper with convolutions*. pp. 1–9.

5 [49] Bishop, C. M. (2006) *Pattern Recognition and Machine Learning*. (Springer).

5, 25 [50] Lin, T.-Y, Goyal, P, Girshick, R, He, K, & Dollar, P. (2017) *Focal Loss for Dense Object Detection*.

5 [51] Rumelhart, D. E, Hinton, G. E, & Williams, R. J. (1986) Learning representations by back-propagating errors. *Nature* **323**, 533–536.

[52] Nocedal, J & Wright, S. J. (2006) *Numerical Optimization*. (Springer, New York, NY, USA), second edition. 6

[53] Duchi, J, Hazan, E, & Singer, Y. (2011) Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* **12**, 2121–2159. 6

[54] Zeiler, M. D. (2012) Adadelta: An adaptive learning rate method. 6, 24

[55] Kingma, D. P & Ba, J. (2017) Adam: A method for stochastic optimization. 6, 25

[56] Hastie, T, Tibshirani, R, & Friedman, J. (2009) *The elements of statistical learning: data mining, inference and prediction*. (Springer), 2 edition. 6

[57] Murphy, K. P. (2012) *Machine Learning: A Probabilistic Perspective*. (MIT Press). 6

[58] Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267–288. 7

[59] Hoerl, A. E & Kennard, R. W. (2000) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **42**, 80–86. 7

[60] Zou, H & Hastie, T. (2005) Addendum: Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 768–768. 7, 24

[61] Kabán, A. (2007) On bayesian classification with laplace priors. *Pattern Recognition Letters* **28**, 1271–1282. 7

[62] Cousineau, D & Hélie, S. (2013) Improving maximum likelihood estimation using prior probabilities: A tutorial on maximum a posteriori estimation and an examination of the weibull distribution. *Tutorials in Quantitative Methods for Psychology* **9**, 61–71. 7

[63] Caruana, R. A. (1993) *Multitask Connectionist Learning*. pp. 372–379. 7, 25

[64] Baxter, J. (1995) *Learning Internal Representations*. (ACM Press), pp. 311–320. 7

[65] Srivastava, N, Hinton, G, Krizhevsky, A, Sutskever, I, & Salakhutdinov, R. (2014) Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15**, 1929–1958. 7, 24, 26

[66] Perez, L & Wang, J. (2017) The effectiveness of data augmentation in image classification using deep learning. 7

[67] Shorten, C & Khoshgoftaar, T. M. (2019) A survey on image data augmentation for deep learning. *Journal of Big Data* **6**. 7

[68] Chan, J. K. C. (2014) The wonderful colors of the hematoxylin–eosin stain in diagnostic surgical pathology. *International Journal of Surgical Pathology* **22**, 12–32. 8

[69] Hanahan, D & Weinberg, R. A. (2011) Hallmarks of cancer: The next generation. *Cell* **144**, 646–674. 8, 9

[70] Junqueira, L. C. (2010) *Junqueiras Basic Histology: Text and Atlas*. (McGraw-Hill Education - Europe). 8

[71] Kononen, J, Bubendorf, L, Kallionimeni, A, Bärlund, M, Schraml, P, Leighton, S, Torhorst, J, Mihatsch, M. J, Sauter, G, & Kallionimeni, O.-P. (1998) Tissue microarrays for high- 8 throughput molecular profiling of tumor specimens. *Nature Medicine* **4**, 844–847.

[72] Kallioniemi, O.-P. (2001) Tissue microarray technology for high-throughput molecular 8 profiling of cancer. *Human Molecular Genetics* **10**, 657–662.

8 [73] Weinberg, R. A. (2013) *The Biology of Cancer*. (W.W. Norton & Company).

[74] Coons, A. H, Creech, H. J, & Jones, R. N. (1941) Immunological properties of an antibody 9 containing a fluorescent group. *Experimental Biology and Medicine* **47**, 200–202.

[75] Elston, C & Ellis, I. (1991) pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. 9 *Histopathology* **19**, 403–410.

[76] Compton, C. C. (2002) Pathologic prognostic factors in the recurrence of rectal cancer. 9 *Clinical Colorectal Cancer* **2**, 149–160.

[77] Hamilton, S, Vogelstein, B, Kudo, S, Riboli, E, Nakamura, S, Hainaut, P, Rubio, C, Sobin, L, Fogt, F, Winawer, S, Goldgar, D, & Jass, J. (2000) *World Health Organization Classification of Tumours: Pathology and Genetics of Tumours of the Digestive System*. (IARC Press), pp. 9 105–119.

[78] Sobin, L, Gospodarowicz, M, & Wittekind, C. (2009) *TNM classification of malignant* 9 *tumours*. (Chichester, West Sussex, UK ; Hoboken, NJ : Wiley-Blackwell, 2010.).

9 [79] Hamilton, S. R. (2012) Molecular pathology. *Molecular Oncology* **6**, 177–181.

[80] Albarracin, C & Dhamne, S. (2014) Ki67 as a biomarker of prognosis and prediction: Is it 10 ready for use in routine pathology practice? *Current Breast Cancer Reports* **6**, 260–266.

[81] Wang, Y, Kartasalo, K, Valkonen, M, Larsson, C, Ruusuvuori, P, Hartman, J, & Rantalainen, M. (2020) Predicting molecular phenotypes from histopathology images: a transcriptome- 10, 13 wide expression-morphology analysis in breast cancer.

[82] Bera, K, Schalper, K. A, Rimm, D. L, Velcheti, V, & Madabhushi, A. (2019) Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nature* 11 *Reviews Clinical Oncology* **16**, 703–715.

[83] Holzinger, A, Malle, B, Kieseberg, P, Roth, P. M, Müller, H, Reihs, R, & Zatloukal, K. (2017) *Machine Learning and Knowledge Extraction in Digital Pathology Needs an Integrative Approach* eds. Holzinger, A, Goebel, R, Ferri, M, & Palade, V. (Springer International 11 Publishing, Cham), pp. 13–50.

[84] Joensuu, H, Bono, P, Kataja, V, Alanko, T, Kokko, R, Asola, R, Utriainen, T, Turpeenniemi-Hujanen, T, Jyrkkiö, S, Möykkynen, K, Helle, L, Ingalsuo, S, Pajunen, M, Huusko, M, Salminen, T, Auvinen, P, Leinonen, H, Leinonen, M, Isola, J, & Kellokumpu-Lehtinen, P.-L. (2009) Fluorouracil, epirubicin, and cyclophosphamide with either docetaxel or vinorelbine, with or without trastuzumab, as adjuvant treatments of breast cancer: Final results of the 11, 21 finher trial. *Journal of Clinical Oncology* **27**, 5685–5692.

[85] Joensuu, H, Isola, J, Lundin, M, Salminen, T, Holli, K, Kataja, V, Pylkkänen, L, Turpeenniemi-Hujanen, T, von Smitten, K, & Lundin, J. (2003) Amplification of erbb2 and erbb2 expression are superior to estrogen receptor status as risk factors for distant recurrence in pt1n0m0 breast cancer. *Clinical Cancer Research* **9**, 923–930. 11, 20, 21

[86] Cooper, L. A, Demicco, E. G, Saltz, J. H, Powell, R. T, Rao, A, & Lazar, A. J. (2018) PanCancer insights from the cancer genome atlas: the pathologist's perspective. *The Journal of Pathology* **244**, 512–524. 11, 12

[87] Roux, L, Racoceanu, D, Loménie, N, Kulikova, M, Irshad, H, Klossa, J, Capron, F, Genestie, C, Naour, G, & Gurcan, M. (2013) Mitosis detection in breast cancer histological images an ICPR 2012 contest. *Journal of Pathology Informatics* **4**, 8. 11, 39

[88] Kallen, H, Molin, J, Heyden, A, Lundstrom, C, & Astrom, K. (2016) *Towards grading gleason score using generically trained deep convolutional neural networks*. (IEEE). 11, 39

[89] Ström, P, Kartasalo, K, Olsson, H, Solorzano, L, Delahunt, B, Berney, D. M, Bostwick, D. G, Evans, A. J, Grignon, D. J, Humphrey, P. A, Iczkowski, K. A, Kench, J. G, Kristiansen, G, van der Kwast, T. H, Leite, K. R. M, McKenney, J. K, Oxley, J, Pan, C.-C, Samaratunga, H, Srigley, J. R, Takahashi, H, Tsuzuki, T, Varma, M, Zhou, M, Lindberg, J, Lindskog, C, Ruusuvuori, P, Wählby, C, Grönberg, H, Rantalainen, M, Egevad, L, & Eklund, M. (2020) Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet Oncology* **21**, 222–232. 11, 39

[90] Sirinukunwattana, K, Snead, D. R. J, & Rajpoot, N. M. (2015) A stochastic polygons model for glandular structures in colon histology images. *IEEE Transactions on Medical Imaging* **34**, 2366–2378. 11, 39

[91] Sirinukunwattana, K, Pluim, J. P, Chen, H, Qi, X, Heng, P.-A, Guo, Y. B, Wang, L. Y, Matuszewski, B. J, Bruni, E, Sanchez, U, Böhm, A, Ronneberger, O, Cheikh, B. B, Racoceanu, D, Kainz, P, Pfeiffer, M, Urschler, M, Snead, D. R, & Rajpoot, N. M. (2017) Gland segmentation in colon histology images: The glas challenge contest. *Medical Image Analysis* **35**, 489–502. 11

[92] Linder, N, Konsti, J, Turkki, R, Rahtu, E, Lundin, M, Nordling, S, Haglund, C, Ahonen, T, Pietikäinen, M, & Lundin, J. (2012) Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. *Diagnostic Pathology* **7**, 22. 11

[93] Chan, L, Hosseini, M. S, Rowsell, C, Plataniotis, K. N, & Damaskinos, S. (2019) *HistoSegNet: Semantic Segmentation of Histological Tissue Type in Whole Slide Images*. 11

[94] Shah, P, Kendall, F, Khozin, S, Goosen, R, Hu, J, Laramie, J, Ringel, M, & Schork, N. (2019) Artificial intelligence and machine learning in clinical development: a translational perspective. *npj Digital Medicine* **2**. 11

[95] Gamble, P, Jaroensri, R, Wang, H, Tan, F, Moran, M, Brown, T, Flament-Auvigne, I, Rakha, E. A, Toss, M, Dabbs, D. J, Regitnig, P, Olson, N, Wren, J. H, Robinson, C, Corrado, G. S, Peng, L. H, Liu, Y, Mermel, C. H, Steiner, D. F, & Chen, P.-H. C. (2021) Determining breast cancer biomarker status and associated morphological features using deep learning. 11, 13, 41 *Communications Medicine* **1**.

11 [96] Cortes, C & Vapnik, V. (1995) Support-vector networks. *Machine learning* **20**, 273–297.

[97] Wang, H, Xing, F, Su, H, Stromberg, A, & Yang, L. (2014) Novel image markers for non-small cell lung cancer classification and survival prediction. *BMC Bioinformatics* **15**, 11 310.

[98] Luo, X, Zang, X, Yang, L, Huang, J, Liang, F, Rodriguez-Canales, J, Wistuba, I. I, Gazdar, A, Xie, Y, & Xiao, G. (2017) Comprehensive computational pathological image analysis 11 predicts lung cancer prognosis. *Journal of Thoracic Oncology* **12**, 501–509.

[99] Beck, A. H, Sangoi, A. R, Leung, S, Marinelli, R. J, Nielsen, T. O, van de Vijver, M. J, West, R. B, van de Rijn, M, & Koller, D. (2011) Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational* 11, 39 *Medicine* **3**, 108ra113–108ra113.

[100] Wu, J, Liang, C, Chen, M, & Su, W. (2016) Association between tumor-stroma ratio and prognosis in solid tumor patients: a systematic review and meta-analysis. *Oncotarget* **7**, 12 68954–68965.

[101] Geessink, O. G. F, Baidoshvili, A, Klaase, J. M, Bejnordi, B. E, Litjens, G. J. S, van Pelt, G. W, Mesker, W. E, Nagtegaal, I. D, Ciompi, F, & van der Laak, J. A. W. M. (2019) Computer aided quantification of intratumoral stroma yields an independent prognosticator 12, 14 in rectal cancer. *Cellular Oncology* **42**, 331–341.

[102] Szegedy, C, Vanhoucke, V, Ioffe, S, Shlens, J, & Wojna, Z. (2016) *Rethinking the Inception* 12 *Architecture for Computer Vision*. (IEEE).

[103] Gleason, D. F & and, G. T. M. (1974) Prediction of prognosis for prostatic adenocarcinoma 12 by combined histological grading and clinical staging. *Journal of Urology* **111**, 58–64.

[104] Nagpal, K, Foote, D, Liu, Y, Chen, P.-H. C, Wulczyn, E, Tan, F, Olson, N, Smith, J. L, Mohtashamian, A, Wren, J. H, Corrado, G. S, MacDonald, R, Peng, L. H, Amin, M. B, Evans, A. J, Sangoi, A. R, Mermel, C. H, Hipp, J. D, & Stumpe, M. C. (2019) Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. 12, 14 *npj Digital Medicine* **2**.

[105] Couture, H. D, Williams, L. A, Geradts, J, Nyante, S. J, Butler, E. N, Marron, J. S, Perou, C. M, Troester, M. A, & Niethammer, M. (2018) Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *npj Breast* 12, 13, 14 *Cancer* **4**.

[106] Tang, B, Li, A, Li, B, & Wang, M. (2019) CapSurv: Capsule network for survival analysis 12 with whole slide pathological images. *IEEE Access* **7**, 26022–26030.

[107] Courtiol, P, Maussion, C, Moarii, M, Pronier, E, Pilcer, S, Sefta, M, Manceron, P, Toldo, S, Zaslavskiy, M, Stang, N. L, Girard, N, Elemento, O, Nicholson, A. G, Blay, J.-Y, Galateau-Sallé, F, Wainrib, G, & Clozel, T. (2019) Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature Medicine* **25**, 1519–1525. 12, 15

[108] Skrede, O.-J, Raedt, S. D, Kleppe, A, Hveem, T. S, Liestøl, K, Maddison, J, Askautrud, H. A, Pradhan, M, Nesheim, J. A, Albregtsen, F, Farstad, I. N, Domingo, E, Church, D. N, Nesbakken, A, Shepherd, N. A, Tomlinson, I, Kerr, R, Novelli, M, Kerr, D. J, & Danielsen, H. E. (2020) Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *The Lancet* **395**, 350–360. 12, 42

[109] Muhammad, H, Sigel, C. S, Campanella, G, Boerner, T, Pak, L. M, Büttner, S, IJzermans, J. N. M, Koerkamp, B. G, Doukas, M, Jarnagin, W. R, Simpson, A. L, & Fuchs, T. J. (2019) in *Lecture Notes in Computer Science*. (Springer International Publishing), pp. 604–612. 12

[110] Zhu, X, Yao, J, Zhu, F, & Huang, J. (2017) *WSISA: Making Survival Prediction From Whole Slide Histopathological Images*. 12, 15

[111] Fu, Y, Jung, A. W, Torne, R. V, Gonzalez, S, Vöhringer, H, Shmatko, A, Yates, L. R, Jimenez-Linan, M, Moore, L, & Gerstung, M. (2020) Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer* **1**, 800–810. 12

[112] Rawat, R. R, Ruderman, D, Macklin, P, Rimm, D. L, & Agus, D. B. (2018) Correlating nuclear morphometric patterns with estrogen receptor status in breast cancer pathologic specimens. *npj Breast Cancer* **4**. 13

[113] Shamai, G, Binenbaum, Y, Slossberg, R, Duek, I, Gil, Z, & Kimmel, R. (2019) Artificial intelligence algorithms to assess hormonal status from tissue microarrays in patients with breast cancer. *JAMA Network Open* **2**, e197700. 13, 41

[114] Rawat, R. R, Ortega, I, Roy, P, Sha, F, Shibata, D, Ruderman, D, & Agus, D. B. (2020) Deep learned tissue "fingerprints" classify breast cancers by ER/PR/her2 status from h&e images. *Scientific Reports* **10**. 13, 41

[115] Naik, N, Madani, A, Esteva, A, Keskar, N. S, Press, M. F, Ruderman, D, Agus, D. B, & Socher, R. (2020) Deep learning-enabled breast cancer hormonal receptor status determination from base-level h&e stains. *Nature Communications* **11**. 13, 41

[116] Bychkov, D, Linder, N, Tiulpin, A, Kücükel, H, Lundin, M, Nordling, S, Sihto, H, Isola, J, Lehtimäki, T, Kellokumpu-Lehtinen, P.-L, von Smitten, K, Joensuu, H, & Lundin, J. (2021) Deep learning identifies morphological features in breast cancer predictive of cancer ERBB2 status and trastuzumab treatment efficacy. *Scientific Reports* **11**. 13

[117] Dodington, D. W, Lagree, A, Tabbarah, S, Mohebpour, M, Sadeghi-Naini, A, Tran, W. T, & Lu, F.-I. (2021) Analysis of tumor nuclear features using artificial intelligence to predict response to neoadjuvant chemotherapy in high-risk breast cancer patients. *Breast Cancer Research and Treatment* **186**, 379–389. 13

[118] Li, F, Yang, Y, Wei, Y, He, P, Chen, J, Zheng, Z, & Bu, H. (2021) Deep learning-based predictive biomarker of pathological complete response to neoadjuvant chemotherapy from histological images in breast cancer. *Journal of Translational Medicine* **19**. 13

[119] Harder, N, Schönmeyer, R, Nekolla, K, Meier, A, Brieu, N, Vanegas, C, Madonna, G, Capone, M, Botti, G, Ascierto, P. A, & Schmidt, G. (2019) Automatic discovery of image-based signatures for ipilimumab response prediction in malignant melanoma. *Scientific Reports* **9**. 13

[120] Madabhushi, A, Wang, X, Barrera, C, & Velcheti, V. (2019) Predicting response to immunotherapy using computer extracted features of cancer nuclei from hematoxylin and eosin (h&e) stained images of non-small cell lung cancer (nsclc). *United States Patent Application 20190259154*. 13

[121] Lee, E. T & Go, O. T. (1997) SURVIVAL ANALYSIS IN PUBLIC HEALTH RESEARCH. *Annual Review of Public Health* **18**, 105–134. 14

[122] Schober, P & Vetter, T. R. (2018) Survival analysis and interpretation of time-to-event data. *Anesthesia & Analgesia* **127**, 792–798. 14

[123] Kaplan, E. L & Meier, P. (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481. 14, 26

[124] Cox, D. R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 187–202. 14, 15, 26

[125] Bellera, C. A, MacGrogan, G, Debled, M, de Lara, C. T, Brouste, V, & Mathoulin-Pélissier, S. (2010) Variables with time-varying effects and the cox model: Some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Medical Research Methodology* **10**. 14

[126] Xue, X, Xie, X, Gunter, M, Rohan, T. E, Wassertheil-Smoller, S, Ho, G. Y, Cirillo, D, Yu, H, & Strickler, H. D. (2013) Testing the proportional hazards assumption in case-cohort analysis. *BMC Medical Research Methodology* **13**. 14

[127] Babińska, M, Chudek, J, Chełmecka, E, Janik, M, Klimek, K, & Owczarek, A. (2015) Limitations of cox proportional hazards analysis in mortality prediction of patients with acute coronary syndrome. *Studies in Logic, Grammar and Rhetoric* **43**, 33–48. 14, 16

[128] Platt, R. W. (2004) A proportional hazards model with time-dependent covariates and time-varying effects for analysis of fetal and infant death. *American Journal of Epidemiology* **160**, 199–206. 14

[129] Buchholz, A & Sauerbrei, W. (2011) Comparison of procedures to assess non-linear and time-varying effects in multivariable models for survival data. *Biometrical Journal* **53**, 308–331. 14

[130] Swindell, W. R. (2009) Accelerated failure time models provide a useful statistical framework for aging research. *Experimental Gerontology* **44**, 190–200. 14

[131] Van Belle, V, Pelckmans, K, Suykens, J, & Van Huffel, S. (2008) *Survival SVM : a Practical scalable algorithm*. pp. 89–94. 14

[132] Ishwaran, H, Kogalur, U. B, Blackstone, E. H, & Lauer, M. S. (2008) Random survival forests. *The Annals of Applied Statistics* **2**. 14

[133] Katzman, J. L, Shaham, U, Cloninger, A, Bates, J, Jiang, T, & Kluger, Y. (2018) DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology* **18**. 14

[134] Faraggi, D & Simon, R. (1995) A neural network model for survival data. *Statistics in Medicine* **14**, 73–82. 14

[135] Muhammad, H, Xie, C, Sigel, C. S, Doukas, M, Alpert, L, & Fuchs, T. J. (2021) Epic-survival: End-to-end part inferred clustering for survival analysis, featuring prognostic stratification boosting. 15

[136] Breslow, N. (1974) Covariance analysis of censored survival data. *Biometrics* **30**, 89–99. 15

[137] Efron, B. (1977) The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association* **72**, 557–565. 15

[138] Hertz-Picciotto, I & Rockhill, B. (1997) Validity and efficiency of approximation methods for tied survival times in cox regression. *Biometrics* **53**, 1151–1156. 16

[139] Luck, M, Sylvain, T, Cardinal, H, Lodi, A, & Bengio, Y. (2017) Deep learning for patient-specific kidney graft survival analysis. *ArXiv* **abs/1705.10245**. 16

[140] Huang, Z, Zhang, H, Boss, J, Goutman, S. A, Mukherjee, B, Dinov, I. D, & and, Y. G. (2017) Complete hazard ranking to analyze right-censored data: An ALS survival study. *PLOS Computational Biology* **13**, e1005887. 16, 25, 42

[141] Linder, N, Martelin, E, Lundin, M, Louhimo, J, Nordling, S, Haglund, C, & Lundin, J. (2009) Xanthine oxidoreductase – clinical significance in colorectal cancer and in vitro expression of the protein in human colon cancer cells. *European Journal of Cancer* **45**, 648–655. 18

[142] Lundin, J. (2003) Infopoints: A web-based system for individualised survival estimation in breast cancer. *BMJ* **326**, 29–29. 20

[143] Joensuu, H, Lehtimäki, T, Holli, K, Elomaa, L, Turpeenniemi-Hujanen, T, Kataja, V, Anttila, A, Lundin, M, Isola, J, & Lundin, J. (2004) Risk for Distant Recurrence of Breast Cancer Detected by Mammography Screening or Other Methods. *JAMA* **292**, 1064–1073. 20

[144] Joensuu, H, Kellokumpu-Lehtinen, P.-L, Bono, P, Alanko, T, Kataja, V, Asola, R, Utriainen, T, Kokko, R, Hemminki, A, Tarkkanen, M, Turpeenniemi-Hujanen, T, Jyrkkiö, S, Flander, M, Helle, L, Ingalsuo, S, Johansson, K, Jääskeläinen, A.-S, Pajunen, M, Rauhala, M, Kaleva-Kerola, J, Salminen, T, Leinonen, M, Elomaa, I, & Isola, J. (2006) Adjuvant docetaxel or vinorelbine with or without trastuzumab for breast cancer. *New England Journal of Medicine* **354**, 809–820. 21

[145] Simonyan, K & Zisserman, A. (2015) *Very Deep Convolutional Networks for Large-Scale Image Recognition* eds. Bengio, Y & LeCun, Y. 23, 24

[146] Russakovsky, O, Deng, J, Su, H, Krause, J, Satheesh, S, Ma, S, Huang, Z, Karpathy, A, Khosla, A, Bernstein, M, Berg, A, & Fei-Fei, L. (2015) Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* **115**, 211–252. Publisher Copyright: © 2015, Springer Science+Business Media New York. Copyright: Copyright 2015 Elsevier B.V., All rights reserved. 23, 24, 25

[147] Tiulpin, A, Panfilov, E, & Tiulpin. (2020) Mipt-oulu/solt: Improved api, speed, bugfix a big release. 23

[148] Janowczyk, A, Zuo, R, Gilmore, H, Feldman, M, & Madabhushi, A. (2019) Histoqc: An open-source quality control tool for digital pathology slides. *JCO clinical cancer informatics* **3**, 1–7. 23

[149] Perronnin, F, Sánchez, J, & Mensink, T. (2010) in *Computer Vision – ECCV 2010*. (Springer Berlin Heidelberg), pp. 143–156. 25, 40

[150] Lin, M, Chen, Q, & Yan, S. (2014) Network in network. *CoRR* **abs/1312.4400**. 25

[151] Selvaraju, R. R, Cogswell, M, Das, A, Vedantam, R, Parikh, D, & Batra, D. (2019) Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* **128**, 336–359. 26, 41

[152] Bland, J. M & Altman, D. G. (2004) The logrank test. *BMJ* **328**, 1073. 26

[153] Brentnall, A. R & Cuzick, J. (2018) Use of the concordance index for predictors of censored survival data. *Statistical Methods in Medical Research* **27**, 2359–2373. 26

[154] Bradley, A. P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* **30**, 1145–1159. 26

[155] Venkatraman, E. S. (2000) A permutation test to compare receiver operating characteristic curves. *Biometrics* **56**, 1134–1138. 26

[156] Saito, T & Rehmsmeier, M. (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE* **10**, e0118432. 26

[157] Hassanpour, S, Korbar, B, Olofson, A, Miraflor, A, Nicka, C, Suriawinata, M, Torresani, L, & Suriawinata, A. (2017) Deep learning for classification of colorectal polyps on whole-slide images. *Journal of Pathology Informatics* **8**, 30. 39

[158] Hamilton, P. W, van Diest, P. J, Williams, R, & Gallagher, A. G. (2009) Do we see what we think we see? the complexities of morphological assessment. *The Journal of Pathology* **218**, 285–291. 39

[159] Polley, M.-Y. C, Leung, S. C. Y, McShane, L. M, Gao, D, Hugh, J. C, Mastropasqua, M. G, Viale, G, Zabaglo, L. A, Penault-Llorca, F, Bartlett, J. M, Gown, A. M, Symmans, W. F, Piper, T, Mehl, E, Enos, R. A, Hayes, D. F, Dowsett, M, & Nielsen, T. O. (2013) An international ki67 reproducibility study. *JNCI: Journal of the National Cancer Institute* **105**, 1897–1906. 39

[160] Maguire, A. (2014) Controversies in the pathological assessment of colorectal cancer. *World Journal of Gastroenterology* **20**, 9850. 39

[161] Robboy, S. J, Weintraub, S, Horvath, A. E, Jensen, B. W, Alexander, C. B, Fody, E. P, Crawford, J. M, Clark, J. R, Cantor-Weinberg, J, Joshi, M. G, Cohen, M. B, Prystowsky, M. B, Bean, S. M, Gupta, S, Powell, S. Z, Speights, V. O, Gross, D. J, & Black-Schaffer, W. S. (2013) Pathologist workforce in the united states: I. development of a predictive model to examine factors influencing supply. *Archives of Pathology & Laboratory Medicine* **137**, 1723–1732. 39

[162] Wilson, M. L, Fleming, K. A, Kuti, M. A, Looi, L. M, Lago, N, & Ru, K. (2018) Access to pathology and laboratory medicine services: a crucial gap. *The Lancet* **391**, 1927–1938. 39

[163] Kleinert, S & Horton, R. (2018) Pathology and laboratory medicine: the cinderella of health systems. *The Lancet* **391**, 1872–1873. 39

[164] Simonyan, K & Zisserman, A. (2015) *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 39

[165] Karpathy, A, Johnson, J, & Li, F. (2015) Visualizing and understanding recurrent networks. *CoRR* **abs/1506.02078**. 40

[166] Elston, C. W & Ellis, I. O. (2002) 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* **41**, 154–161. 40

[167] Saad, E. D, Squifflet, P, Burzykowski, T, Quinaux, E, Delaloge, S, Mavroudis, D, Perez, E, Piccart-Gebhart, M, Schneider, B. P, Slamon, D, Wolmark, N, & Buyse, M. (2019) Disease-free survival as a surrogate for overall survival in patients with HER2-positive, early breast cancer in trials of adjuvant trastuzumab for up to 1 year: a systematic review and meta-analysis. *The Lancet Oncology* **20**, 361–370. 42