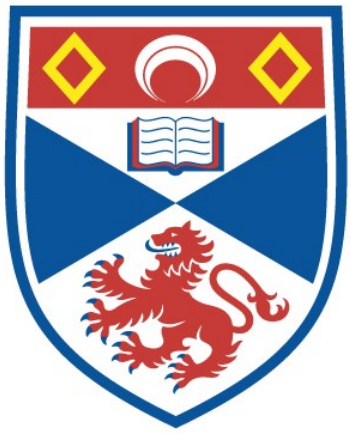
COMPUTATIONAL ANALYSIS OF TISSUE IMAGES IN CANCER DIAGNOSIS AND PROGNOSIS: MACHINE LEARNING-BASED METHODS FOR THE NEXT GENERATION OF COMPUTATIONAL PATHOLOGY

Neofytos Dimitriou

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews



2023

Computational analysis of tissue images in cancer diagnosis and prognosis: machine learning-based methods for the next generation of computational pathology

Neofytos Dimitriou

University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of

Doctor of Philosophy (PhD)

at the University of St Andrews

April 2022

Candidate's declaration

I, Neofytos Dimitriou, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 24,787 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I confirm that any appendices included in my thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

I was admitted as a research student at the University of St Andrews in September 2017.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date 30/01/2023 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree. I confirm that any appendices included in the thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

Date 30/01/2023 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Neofytos Dimitriou, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date 30/01/2023 Signature of candidate

Date 30/01/2023 Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Neofytos Dimitriou, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date 30/01/2023 Signature of candidate

Keywords: histology, histopathology, radiomics, quantitative histomorphometry, deep learning, whole slide imaging, digital pathology, gigapixel images, tissue heterogeneity, artefacts, image analysis, precision medicine, TNM, multiplex immunofluorescence, immunohistochemistry, hematoxylin, eosin, metastases detection, CAD, outcome prediction, survival analysis, ensemble.

Abstract

The focus of this work is to develop machine learning systems capable of tissue image analysis in the context of cancer diagnosis and prognosis. Such a system can not only identify new prognostic markers, but can also serve as a standalone clinical prediction rule, the premise being that its non–linear, multivariate nature may be capable of identifying and employing complex patterns that collectively provide accurate cancer diagnosis and prognosis, better than the clinical gold standard. The task, however, is very challenging because of the extremely high resolution of the images, highly heterogeneous microenvironment, multiple sources of noise and artifacts, and low–granularity of ground truth.

A starting point of related work which tackles the same task is the extraction of handcrafted features. I investigate the application of machine learning for prognosis using handcrafted features, and develop prognostic machine learning models that demonstrate better performances than baselines based on clinically employed prognostic systems, in two separate cohorts of colorectal and muscle–invasive bladder cancer patients. Moreover, analysis of the proposed methods provides insight behind the prognostic nature of characteristics within the microenvironment, not yet included in the clinical systems.

The emergence of deep learning has enabled analysis with images directly. Given the laborious, expensive, and human bias inducing nature of designing and building pipelines for handcrafted feature extraction, I investigate the application of deep learning on tissue images directly. In particular, I propose a framework that allows the training of models directly from exhaustively-tiled whole slide images with only patient–level ground truth, and demonstrate its effectiveness on colorectal cancer prognosis.

In my final work, I introduce a new type of CNN–based method, called Magnifying Networks, for gigapixel image analysis that does not require whole slide images to be patch–based preprocessed. Instead, MagNets dynamically extract patches from the tissue image based on the best magnification level, field–of–view, and location according to an optimizing task, and not based on generic, predefined or static ways. My results on the publicly available Camelyon16 and Camelyon17 datasets demonstrate the effectiveness of MagNets, as well as the proposed optimization framework, on the task of whole slide image classification. MagNets process far fewer patches from each slide than any of the existing end-to-end approaches (10

to 300 times fewer).

Acknowledgements

My deepest appreciation goes to my partner Rafaella for her constant support and care throughout my PhD journey. I also owe a debt of gratitude to my dear friends Christos, Andri, Soti, Marios, and Kristi for making my time at St Andrews enjoyable and memorable.

I would also like to express my heartfelt gratitude to my parents, Egli and Yiannakis. My mother has always exemplified the importance of perseverance and attention to detail, and her teachings have been invaluable in helping me navigate the challenges of life. My father has always been a constant presence in my life, offering his listening ear and unconditional support, and nurturing my curiosity from a young age. It is thanks to both of them that I am able to present this thesis with excitement and optimism for the future.

I would also like to express my sincere gratitude to Professor Tom Kelsey and Dr. Peter Bankhead, the internal and external examiners of my viva. Their feedback, which was both thorough and insightful, played a crucial role in improving my thesis. I am thankful to both of them for their time and expertise. I would also like to extend my gratitude to Professor David Harrison for his invaluable guidance and support throughout my interdisciplinary PhD journey. His expertise and willingness to answer my questions played a crucial role in the successful completion of my thesis.

Finally, I am deeply thankful to my supervisor and friend Oggie, to whom this thesis is dedicated. He has played a pivotal role in my academic journey and career, and his trust in my abilities, coupled with his wealth of experience, knowledge, enthusiasm, and optimism, has been crucial in helping me to thrive and succeed in both. He has always been a voice of reason and a steadfast source of support and guidance, especially during those times when the rest of the world seemed unhelpful or indifferent. I am grateful to him for all that he has done for me, and I look forward to continuing our work and friendship for many years to come.

This work was supported by the EPSRC (grant number 1950036) and the University of St Andrews (School of Computer Science).

Contents

Chapter 1

Introduction

1.1 Motivation, objectives, and thesis outline

Over the past decade, the field of digital pathology, which encompasses the processes of digitizing glass–mounted tissue specimens and analysing tissue images, has experienced exciting growth and has flourished into a promising avenue for computational pathology. More recently, we have witnessed the development and adoption of advanced visualization techniques (e.g. multiplex immunofluorescence), and specialized hardware, (e.g. whole slide scanners) that allow for more data to be captured from each tissue slide. In parallel, machine learning–based systems have dominated most areas of computer vision, and more broadly, image analysis, demonstrating unprecedented predictive capabilities across a wide spectrum of real world applications. With increasingly more hospitals committing to fully–digital workflows, new avenues for computational exploration of individualized disease tissue have emerged. If machine learning–based systems can leverage histopathological data to effectively address clinical tasks, computational systems for tissue image analysis will become an attractive alternative for a complete spectrum of activities in clinical pathology [80], and will inevitably transform the practice of pathology.

The objective of my thesis is to demonstrate that indeed, there exists a set of machine learning methods with nonlinear cognition capabilities that can leverage histopathological data and provide guidance for critical clinical questions. Due to the data–driven optimization of these systems, one of the premises of my work is that the underlying models will be able to identify patterns of clinical significance that have gone unnoticed to date, or have been too complex to allow for a standardized reporting protocol to be produced and used by pathologists. In the world of cancer diagnosis and prognosis, the proposed methods can help improve our understanding of the disease, and, if clinically adopted, in a very real sense, save human lives. The other premise of my work is that histopathological data contain information (object–based, or subvisual [16]) that is valuable to the quest for accurate cancer diagnosis and prognosis [96, 35, 209].

A cancer diagnosis is about identifying cancerous cells either within the primary site (cancer's origin), or in secondary sites for potential detection of metastasis. Cancer prognosis, on the other hand, is about predicting whether the patient will succumb to the disease within a specific timeline. Precise diagnosis and prognosis are very important for the identification of the most appropriate treatment, as well as to the subsequent clinical management of cancer patients. However, current practices and systems are unable to provide accurate diagnosis and prognosis for all cancer patients alike (see Section 1.1.1).

Chapters 3, 4, and 5 are concerned with the prognosis of colorectal and muscle-invasive bladder cancer patients, and Chapter 6 with the metastasis detection of breast cancer patients. An alternative separation is that in Chapters 3 and 4, the machine learning systems are trained using domain-specific expertise in the form of handcrafted features, whereas Chapters 5 and 6 are concerned with systems trained end–to–end with histopathological images. This separation is discussed more thoroughly in Chapter 2 with related work. In Chapter 7, the main achievements of this thesis are presented, and future research directions are discussed.

1.1.1 Clinical Applications

Cancer prognosis

For most types of cancer that form solid tumours, the tumour–node–metastasis (TNM) staging system is used to determine cancer staging and, together with a list of other factors [60], patient prognosis. "T" refers to the depth of local invasion, and "N" and "M" to the presence or lack of node and distant metastasis, respectively. An accurate prognosis forms the basis for determining the best treatment approach to the cancer patient [60].

TNM, however, serves certain groups of patients unfavourably as they exhibit higher survivability variance. Quantitative image analysis of digitized tissue from these patients constitutes a promising approach in the endeavour of cancer research (and of pathologists) to identify new prognostic markers (independent of TNM). However, a more important potential of image analysis–based systems, over and above the identification of prognostic markers in isolation, is that they can serve as standalone clinical prediction rules; the premise being that their non–linear, multivariate nature may be capable of identifying and employing complex prognostic patterns that collectively provide accurate cancer prognosis, better than the clinical gold standard. With an accurate patient prognosis, it becomes more feasible to identify the most effective treatment for each patient. In my work, I developed prognostic models for colorectal cancer (CRC), as well as muscle–invasive bladder cancer (MIBC) patients.

Colorectal cancer CRC is the third most common cancer worldwide and the leading cause of death among gastrointestinal tumours [67, 123]. Annually, there are 1.4 million new cases and more than half a million of

deaths around the globe [67]. A typical CRC diagnosis requires the evaluation of histopathological slides from a biopsy or resected specimen by a pathologist [133, 27, 68]. Subsequent to a positive diagnosis, prognosis is assessed based on the TNM staging system [60]. The TNM stage is considered by far one of the best predictors of CRC survival [68] and as a consequence, statistics specific to the stage primarily guide therapy. However, stages that exhibit higher variability in patient survival encounter greater uncertainty in therapy planning. The risk of disease–specific death within 5 years for stage II CRC patients is estimated to be 20%, and rises to 35% given a 10 year window [51, 150]. Nevertheless, there are no definite criteria for selecting which, if any, stage II patients should undergo adjuvant chemotherapy with different trials reaching inconsistent conclusions [10, 132].

Muscle-invasive bladder cancer Urothelial cancer of the bladder (bladder cancer) is one of the most prevalent cancers worldwide with approximately 430,000 new diagnoses each year [177]. High morbidity and mortality rates as well as high socioeconomic burden make bladder cancer a debilitating and often fatal disease [106, 114]. Even though the majority of bladder cancer patients are diagnosed with non-muscle-invasive bladder cancer, recurrence and progression of the disease may lead to MIBC [72]. Approximately 25% of newly diagnosed patients have MIBC. MIBC is an aggressive form of bladder cancer in which the cancerous cells have penetrated the neighbouring muscle tissue. Half of MIBC patients succumb to the disease within five years of the diagnosis. The high degree of uncertainty poses a huge psychological burden to these patients. To decrease the mortality rates, patients with a high risk of disease-specific death need to be identified more precisely, thereby allowing for better patient management and new treatments to be tested in the high–risk group.

Whole slide image analysis

For many clinical tasks, including cancer diagnosis and prognosis of solid tumours, tissue analysis constitutes the current gold standard [1]. Tissue analysis can be performed by a pathologist either under the microscope or, nowadays, digitally on whole slide images (WSIs). A WSI is a high resolution image created following the scanning of a tissue glass slide. WSIs are typically stored in a multi-resolution pyramid structure in the .tiff format. Image files contain multiple down-sampled versions of the original image. Each image in the pyramid is stored as a series of tiles, to facilitate rapid retrieval of subregions of the image. WSIs of tissue samples constitute a data–rich source that were previously only accessible to pathologists through a microscope [2]. In particular, prior to whole slide imaging, tissue image analysis was limited by the need to select fields of view upon which image analysis would be performed [80]. In addition, whole slide imaging has enabled the possibility of providing pathology services to locations with no, or limited, on–site pathology support [80], not to mention that it

can remove the cost, time, and risk factors associated with shipping glass slides [80]. Effective, efficient, and interpretable methods for WSI analysis are therefore becoming increasingly sought after.

I present two novel methodologies for WSI classification in Chapters 5 and 6, with the former addressing CRC prognosis (described above), and the latter, breast cancer (BC) metastases detection from nearby lymph node tissue images.

Breast cancer BC is the most common cancer worldwide ($>$ 2 million new cases every year), with more than half a million deaths in 2020 alone [188]. Sentinel axillary lymph node metastasis is typically the first manifestation of BC spreading [179]. Therefore, in routine clinical practice, lymph nodes are surgically removed, and subsequently examined by a pathologist. Histopathological analysis is, however, tedious and time–consuming with the reported overall concordance rate amongst pathologists being a mere 75% [62].

1.2 Challenges

There are many challenges impeding the successful application of machine learning systems in digital pathology. Some of the most important ones are listed below.

Extremely high–resolution images The comprehensive digital rendering of an entire glass slide, visible at resolutions of 0.5μm or lower pixel size, typically translates to a multi–gigabyte image with billions of pixels ($\approx$ 50GB of uncompressed data, just from the highest magnification level of each image alone). Analysis of these images therefore presents a novel challenge to quantitative image analysis, as the images cannot be loaded, in their entirety, into memory. The extremely high–resolution nature of these images gives rise to what is known as the "where" problem [57]; a problem that needs to be addressed first for image analysis to become computationally feasible. Methods addressing this problem need to employ a dimensionality reduction, or preprocessing strategy, manual or automatic, that enables for a set of images with less than a million pixels (often called image patches, or regions of interest (ROIs)) to be extracted from each gigapixel image. These image patches are then amenable to conventional image analysis methods. In other words, prior to any visual understanding optimization ("what" problem), the spatial distribution of relevant information from gigapixel images needs to be approximated ("where" problem). Neither the "where" nor the "what" problems are non–trivial to solve because of the complexity of the underlying disease process. This is discussed next.

Heterogeneous saliency Histopathological images, i.e. tissue images with a potential pathology, in contrast to many other image types (e.g.

radiology, cytology, natural images, etc.), display startling morphological, and spatial heterogeneity. For a tissue containing a tumour, this is unsurprising considering the inherent heterogeneity of cancer [71].

Cancerous tumours are not simply a growth of homogeneous cells but rather a heterogeneous mixture of cellular populations [42]. These cancer cell populations differ in terms of cellular morphology, gene expression, tumorigenic, angiogenic, proliferative, immunogenic, and metastatic potential, and genetic make–up [148, 143, 193, 42]. Similarly, the rest of the microenvironment, i.e. immune infiltrate, stromal compartment, and extracellular matrix, can be equally, if not more, heterogeneous. As a consequence, collectively, the dynamic topology of the tumour and its microenvironment, varies significantly both between (inter–tumour), and within the same (intra–tumour) lesions [42]. The heterogeneity of the underlying cell populations of a tumour, and its microenvironment, can be uniquely captured based on the morphology of tissue slides [16]. Considering that fitness is context–dependent [42], i.e. the survival and proliferation of cancer is microenvironment–dependent, understanding the entirety of the tumour–immune microenvironment in its highly heterogeneous state is vital, and in line with the goals of precision medicine, i.e. tailor interventions based on individual characteristics of patients [35].

The problem of tissue image understanding, i.e. in our case that is the finding of optimal histopathological features (handcrafted or unsupervised) for a clinical task, is referred to as the "what" problem [57] for the rest of this thesis.

Multiple sources of noise Artifacts can be introduced throughout the tissue sample preparation workflow as well as during the scanning process [116, 58]. These are generally unrelated to the underlying tissue biology, yet can occlude, or alter large parts of the final image. Common artifacts emerge from wrinkles and folding of tissue slices, dust, uneven tissue thickness (that can cause blurriness), and colour markings [116]. Moreover, imaging artifacts can appear from uneven illumination, focusing, image tiling, and fluorescence deposits and bleed–through [58, 35]. Finally, perhaps the most common artifact comes in the form of tissue colour variation (see Figure 1.1), and can appear as a consequence of different staining conditions, including that of using staining reagents from different manufacturers, scanning conditions, etc. [116, 58, 35]. The above artifacts, both their existence, and visual depiction, can vary significantly both within a single cohort, as well as across different cohorts. Examples of common artifacts are shown in Figure 1.2.

Low–granularity labels for high–level clinical tasks Low–level clinical tasks are ones that address issues like region identification (e.g. necrosis), and object detection and enumeration (e.g. cellular subtypes, mitosis, etc.) [103, 15]. These tasks, although often mundane and time–intensive,

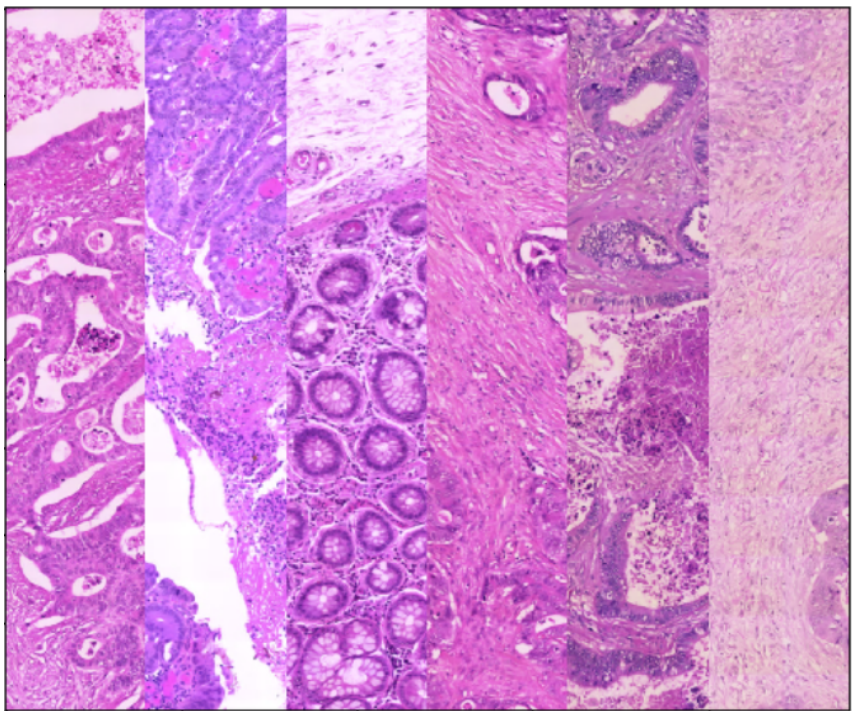


Figure 1.1: Examples of slides from different patients that exhibit different colour profiles.

are not difficult for the pathologist. High–level clinical problems, on the other hand, such as patient prognosis or predicting response to treatment, pose a much more challenging task to the pathologist [103, 15].

Ground truth annotations for tissue images typically involve the delineation of object boundaries and ROIs by one or more pathologists [103]. This level of annotation precision enabled the early success of image–based methods in digital pathology [57]. However, these annotations require large amounts of time and effort from the pathologists for a task that is otherwise rather mundane and cumbersome. For example, Janowczyk and Madabhushi [104] reported that in total more than 40 hours were spent to annotate 12,000 nuclei; a small fraction of all the nuclei present in all images. As a result, in practice, annotations are rarely pixel–level precise, and when they are, they are usually created at lower magnification levels which often results in numerous false positives and negatives [104]. Moreover, for some clinical tasks such as survival analysis, pixel–level precision ground truth does not exist. Instead, only low–granularity, e.g. patient–level rather than pixel–level [57], annotations can be provided. The different levels of annotations are visualized and explained in Figure 1.3. Therefore, tissue–based methods for addressing high–level clinical tasks are faced one the one hand with the challenging nature of the task, and on the other, with the paucity of high–level precise annotations which further limits the applicability of conventional techniques.

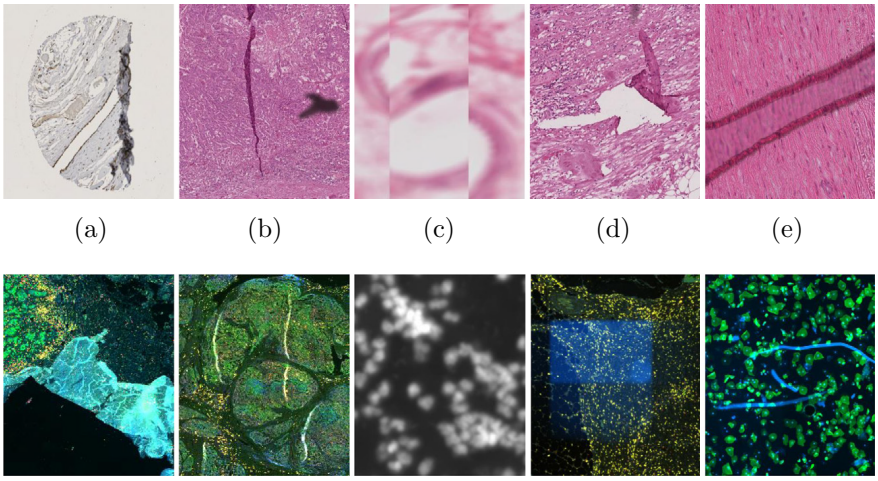


Figure 1.2: (a–e) Examples of artifacts from brightfield image capture. (a) Folded over section of a tissue. (b) Cutting artifact (dark pink line) and piece of dust (gray). (c) Out of focus and incorrectly stitched image. (d) Tear and fold. (e) Out-of-focus section bordered by a foreign object. (f–j) Examples of artifacts from fluorescence image capture. (f) A tear in a tissue section resulting in nonspecific fluorescence. (g) Cutting artifact (bright lines). (h) Out–of–focus nuclei. (i) Illumination artifact resulting in large blue squares. (j) Autofluorescence from hair in a urine cytology sample (blue lines).

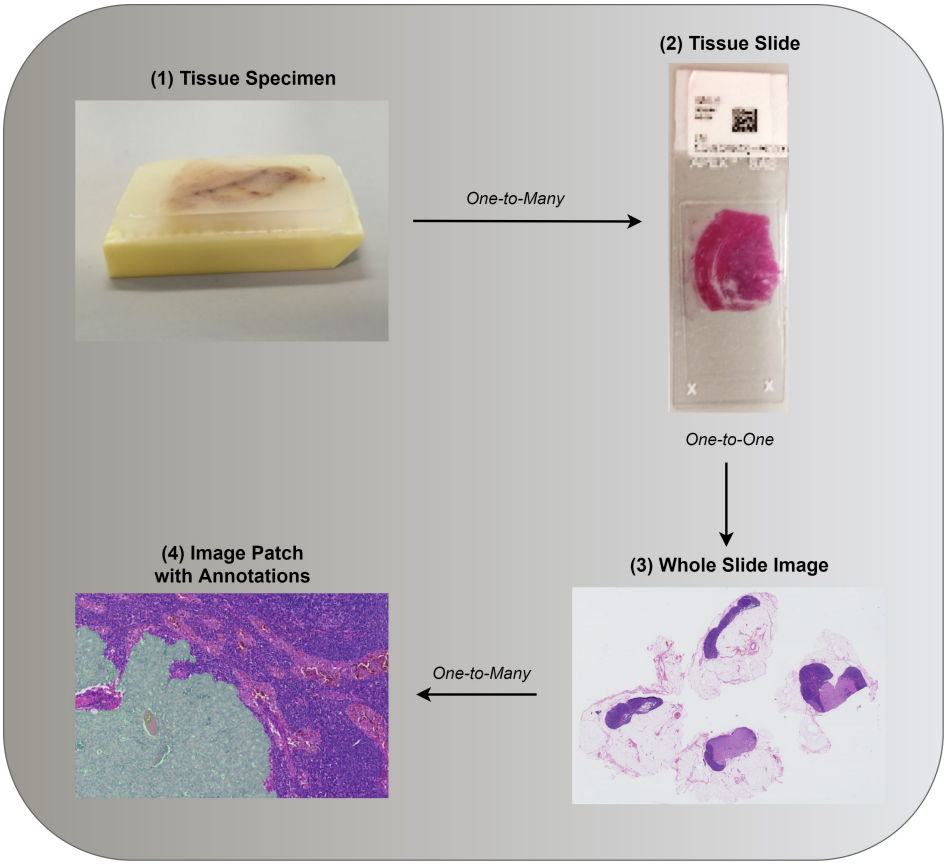


Figure 1.3: (1): Tissue specimen is often investigated as a potential predictor of patient diagnosis, prognosis, or other *patient-level* information. (2 - 3): Both in clinical practice and research, in the interest of time, a single tissue slide, or its digital counterpart, is often assessed. Annotations associated with a single tissue section can be provided such as whether a malignancy is present (*slide-level*). (4): Consequent to the gigapixel size of WSIs, image analysis requires further image reduction. Patches are often extracted based on *pixel-level* annotations from hand-delineated regions of interest. Information associated with individual patches is referred to as *patch-level*. Images (3) and (4) were taken from the public data set of Camelyon17 [8].

1.3 Contributions

Stage II colorectal cancer patients My methodology in Chapter 3 was shown to outperform significantly the clinical gold standard for stage II CRC prognosis based on two time cutoffs (5-year and 10-year). Specifically, the proposed method achieved AUROC of over 77% and 94% for 5 and 10-year prognoses respectively, compared to pT stage, which stratifies patients with the AUROC of approximately 62% both for 5- and 10-year prognosis, and the differentiation, which achieves the corresponding AUROC of approximately 62% and 65%, respectively. Moreover, the interpretability of the proposed method is assessed allowing clinicians to gain new insight by identifying prognostically the most salient features. In particular, my experiments demonstrated that a diverse set of characteristics of the entire microenvironment have a prognostic value.

Muscle–invasive bladder cancer patients Based on standardized quantification of tumour-immune features across whole slide images, and in conjunction with clinical information, in Chapter 4, I develop an ensemble machine learning model that correctly classifies 71.4% of the patients who succumb to MIBC, i.e. the patients at higher risk, using a 5-year cutoff. This is significantly higher than the 28.6% of the TNM staging system. Post-hoc analysis of the model reveals clinically relevant, immunological features for MIBC prognosis. The results of this work suggest that the characterization of a broad immune cell population (e.g. lymphocytes and macrophages), as well as their spatial organization in relation to cancer cells, enables a better estimation of 5-year survival compared to the TNM staging system in MIBC patients which, in turn, provides further biological insights. Most of my findings based on whole slide immunofluorescence images, and machine learning techniques are novel for MIBC, and also corroborate the existing literature on other types of cancer.

Automatically Inferred Phenotypes from whole slide images A novel deep learning–based framework for the prediction of CRC outcome from whole slide images is introduced. This framework can be optimized with low–granularity labels in three steps, and with an arbitrary number of image patches from each WSI. Unsupervised learning is first employed to categorise image patches into clusters based on (a) colour heterogeneity and (b) morphological appearance. Then, the discriminative nature of these clusters based on training patient-level performance is tested. To get cluster-level predictions and ultimately predict patient prognosis, the patch-level predictions are aggregated from discriminative clusters. Using a real-world data set, I demonstrate the effectiveness of the method and present a detailed analysis of its different elements which corroborate its ability to extract and learn salient and discriminative content.

End–to–end weakly supervised learning from whole slide images In Chapter 6, I propose a new type of neural network called Magnifying Network (MagNet), and outline an optimization protocol that enables end-to-end learning from whole slide images with low–granularity labels. MagNets provide a new way of solving both the "where" and "what" problems of gigapixel image analysis in an end–to–end fashion. The results on the publicly available Camelyon16 and Camelyon17 datasets corroborate the effectiveness and efficiency of MagNets and the proposed optimization framework for WSI classification. Importantly, MagNets provide process far fewer patches from each slide than any of the existing approaches (10 to 300 times less).

1.4 Publications

Chapter 3 was published in npj Digital Medicine journal in 2018 [58]. Chapter 4 was published in the journal Cancers in 2021 [79]. An early version of Chapter 5 was presented at the conference of BICOB in 2019 [217], and Chapter 6 is under peer-review (preprint available at arXiv [56]). I have also written a mini–review [57] and a book chapter [35] in 2019 and 2020, each providing a comprehensive overview of the current state of tissue image analysis in the context of computational pathology.

Chapter 2

Tissue slide analysis – Related Work

In this Chapter, I discuss the current methods and practices for tissue slide analysis in the context of the "where" and "what" problems. The former refers to the fact that the spatial distribution of salient information within a tissue slide is unknown [192]. Hence, solutions to the "where" problem either attempt to approximate the aforesaid spatial distribution (e.g. the attention models that are discussed in the following sections), or transform the problem based on certain assumptions (e.g. exhaustive tiling with weak supervision - assumes that the salient information available at the whole slide level is recognizable at the patch-level [192]). The "what" problem refers to the fact that the nature of salient information is unknown [192]. This could be addressed with the identification of visual patterns that are salient to the task at hand. An example would be the learning of a visual representation of cancer cell morphology by a CNN as means of classifying tumour vs. non-tumour regions.

The majority of tissue slides are captured using brightfield illumination, such as for slides stained with clinically routine haematoxylin and eosin (H&E). H&E is still the most important and commonly used histochemical staining method for studying and diagnosing tissue diseases in histopathology. However, imaging of H&E stained FFPE tissue has limitations including the inability to quantify the complex cellular states as well as to identify distinct cell populations in the tumour-immune microenvironment. With the advent of whole slide imaging and the increasing adoption of digital pathology in the clinic [79], multiplex methodologies have the potential to provide significantly more information about the underlying tumour-immune microenvironment than single-marker (i.e. single label immunohistochemistry) and conventional histochemical staining based methodologies [157]. The labelling of multiple cell types and their protein expression can be observed with multiplexed immunofluorescence (IF) which provides valuable information in cancer research and particularly in immunooncology [208]. WSI examples using multiplex IF and H&E

are shown in Chapters 3 and 4, and Chapters 5 and 6, respectively.

The clinical gold standard of histopathological analysis is first introduced, followed by the two most–widely employed types of histopathological analysis based on machine learning techniques.

2.1 Clinical gold standard – manual tissue analysis

In routine clinical practice, pathologists analyse tissue slides either under a microscope, or digitally, and visually recognize, semi–quantify and integrate multiple morphological features from the tissue in the context of the underlying disease [16]. Instrumental to the efficient analysis of the tissue slide, i.e. the "where" problem, is the practice of multi–scale tissue exploration, via microscope or software, that allows pathologists to explore the tissue dynamically while accumulating scale–specific information naturally and seamlessly [80].

With extensive, continual training, pathologists are able to rapidly extract morphological patterns associated with predefined criteria and features (to solve the "what" problem) [16]. These patterns can be object–based in that they can be described using cytologic terms, such as nucleus and cell, and can point to relationships with other objects, as well as to the adjacent tissue (benign or with invasion), glands, and other [80, 2]. When fatigue is not a factor, pathologists rarely fall for artifacts or staining variations [35]. Again, a consequence of their extensive training, but also, of the brilliancy of human vision and brain. The pathologists conclude upon the diagnosis, prognosis, and treatment plan based on (semi–quantitative) features, clinical guidelines, existing systems, and of course, their intuition, built upon years of experience and expertise [1].

Limitations Pathologists, like all humans, are fallible to human error [16, 121]. An extensive list of common cognitive and visual errors, in the context of tissue analysis, is provided in the work of Aeffner et al. [1]. It is therefore unsurprising, that inter– and intra–observer variability in reporting has been a long–standing issue in manual tissue analysis for both cancer diagnosis and prognosis [43, 62, 47]. Examples are seen on a wide range of clinical tasks, including counting of specific cell types, mitotic cells, and biomarker expression [43, 1, 16].

Moreover, the AJCC staging system follows a "tumour autonomous" paradigm, in that only characteristics associated to tumour cells are considered as prognostic factors [191]. However, with the emergence of immuno-oncology, an increasing number of studies has shown the critical role of the immune contexture (i.e. the spatial organization, density, and functional state of immune cell populations within the tumour microenvironment) on patient survivability, suggesting that it could be a valuable determinant of patient prognosis [73, 69]. Yet, despite consistent demonstration of the

prognostic significance of immune–related features, methodological shortcomings of quantifying these features in a standardized manner obstruct their current adoption in clinical practice [58].

Finally, histopathological reporting often requires the conversion of established pre–defined features, otherwise quantifiable across a continuum, into categorical ones based on population–level thresholds [35]; a practice which eliminates information that could otherwise be crucial for personalizing therapy plans [180].

2.2 Machine learning with handcrafted features

Machine learning–based methods for automated, or semi–automated, tissue analysis involve an initial extraction of so–called handcrafted features from each tissue image. As a result, the "what" and "where" problems are tackled within the feature extraction process.

2.2.1 Feature extraction

The "where" problem As we have established, conventional image analysis on the original tissue image is not computationally feasible. Therefore, for each set of handcrafted features, representative patches need to be identified. This process can be manual (e.g. a pathologist identifies the ROIs), semi–automatic (with the help of a computational model, yet pre– or post– manual processing still required), or fully–automatic (typically based on pre–trained segmentation models). Therefore, the "where" problem appears in the context of low–levels tasks, such as the identification of tumour budding, mitotic cells, invasive front, etc. for which high–granularity ground truth, e.g. pixel–level annotations, from pathologists is often provided [74, 122, 220, 57]. Ultimately, these methods return a set of image patches, amenable to conventional image analysis, so that handcrafted features can be extracted from them.

The "what" problem Feature extraction, in our context, refers to the engineering and extraction of explicitly defined features, also known as handcrafted features, from raw images. The key premise is appealing and intuitively sensible: to use human expertise to identify what elementary, low–level elements of an image are salient in that they capture important discriminative information on the one hand, and are yet compact enough to facilitate computational efficiency and reliable learning from available data. Handcrafted features can be general purpose descriptor–based features (domain–agnostic), or can instead exploit domain–specific knowledge, that is, human expert knowledge of pathology (knowledge–driven).

Knowledge–driven feature extraction The first applications of image analysis in digital pathology were the quantification of histopathological or immunohistochemical (i.e. based on protein expression) features with

known pathological significance. Object of interest counts (e.g. tumour and immune cells, etc.), their morphology quantifiers (e.g. size, eccentricity), or spatial statistics are some of the widely used ones [16, 15]. The appeal of these is twofold. Firstly, by their very nature they can be reasonably expected to exhibit saliency in the context of pathology slide analysis. Moreover, in most cases, they can be accurately measured from images: different staining and imaging modalities or biomarkers can be used to highlight the targets of interest, and image processing or computer vision algorithms (such as flood fill algorithms [146], morphological operators [107], blob detectors [211], etc.) can be used for quantitative feature extraction.

Domain–agnostic feature extraction Having demonstrated success on a wide variety of images of so-called natural scenes [152], local appearance descriptors originally proposed for more day-to-day applications of computer vision (e.g. location recognition, synthetic panoramic image generation, object localization, etc.) have been adopted first and applied on a diverse range of pathology image types [3, 211]. Popular and widely used examples include local binary patterns [64], scale invariant feature transform [135], and histogram of oriented gradients [5] based descriptors, for the quantification of subvisual textural heterogeneity measurements [124, 136, 203, 206, 16]. The value of domain–agnostic features is often not known a priori and is difficult, if not impossible, to sort and analyse by eye. Machine learning can be applied to data sets from digital pathology in order to understand the optimal features that allow clinical decision-making in the field of personalized medicine.

2.2.2 Short fat data

Subsequent to feature extraction on, typically, a small number of tissue slides, the output can be very high–dimensional with a large number of features per tissue image. We refer to this problem as the "short fat data" problem [50]. Frameworks for the development of machine learning solutions on "short fat data" often involve feature selection, model selection, algorithm selection, and model optimization so that a good fit between model and training data is identified [26]. However, machine learning algorithms are prone to overfitting, i.e. in finding and using patterns which arise from noise in the data. Patterns of noise, by definition, do not extend beyond the specific data set. Testing the performance of a trained model on unseen data constitutes the mainstay in evaluating the generalizability of a machine learning model, and therefore that it has not overfitted, both during (validation set) and at the end (testing set) of model development.

Feature selection It is very common for humans to generate large sets of features with the hopes that a subset might actually carry information about the problem at hand. Unfortunately, this is usually detrimental due to a phenomenon known as the "curse of dimensionality", first intro-

duced by Bellman [13], which refers to the fact that generalizing a particular problem correctly becomes exponentially harder as dimensionality grows [46]. The intuition behind this is straight-forward; any training set given even a moderate number of dimensions will only cover a minuscule fraction of the input space and as such the higher the number of dimensions the harder the learning process. Feature selection is the process of choosing the most relevant features according to a quantifiable metric, and forms the most common type of dimensionality reduction approach in computational pathology [202, 78]. The goal is to remove redundant, irrelevant and perhaps detrimental features from the original feature set.

Model selection and algorithm selection Both model selection and algorithm selection attempt to collectively maximize the predictive performance of the final machine learning model. Traditionally, model selection is the process in which a machine learning algorithm is configured. Most algorithms come with a number of configuration variables, commonly referred to as hyperparameters. Even though common hyperparameter configurations can be employed, it has been observed that hyperparameter tuning for a specific task can be the key between chance and state-of-the-art models [18]. Since manual tuning can be time consuming and counter-intuitive in high dimensional spaces, most ML methodologies adopt automated hyperparameter tuning.

One of the most popular approaches to model selection constitutes grid search where each hyperparameter is given a predefined list of values, and the best hyperparameter configuration is selected after evaluating all the combinations. For example, given hyperparameters A and B, and lists $V_A = [1, 10, 100]$ and $V_B = [0.1, 0.5]$, the following combinations are evaluated under grid search: $\forall(A, B) \in [(1, 0.1),\ (1, 0.5),\ (10, 0.1),\ (10, 0.5),\ (100, 0.1),\ (100, 0.5)]$. However, as shown by Bergstra and Bengio [17], random sampling provides a better tuning strategy. In random search, a number of hyperparameter configurations are evaluated by sampling from predefined hyperparameter distributions and densities. For example, given hyperparameters A and B with $D_A \sim N(0, 1)$ and $V_B = [0.1, 0.5]$, where D_A is a standard Gaussian distribution, the following five combinations could have been sampled and evaluated under random search: $\forall(A, B) \in [(-0.12, 0.1),\ (-0.14, 0.5),\ (-0.94, 0.5),(0.44, 0.1),\ (-1.3, 0.5)]$.

Not every machine learning algorithm will perform equally well on different problems and different data. In addition, there is no theoretical ranking suggesting that one algorithm is better than another [207]. Hence, similar to hyperparameter tuning, algorithm selection is yet another meta-optimization task that needs to be performed for maximizing predictive performance [171].

2.2.3 High–level clinical tasks

Machine learning methods can leverage "short fat data" to address high–level clinical tasks with low–granularity labels. Despite the aforementioned high dimensionality of the data, machine learning methods have already demonstrated their impressive capacity in a range of high–level clinical tasks [216, 140, 206, 96]. For example, Whitney et al. [206] developed predictive machine learning models that by leveraging handcrafted, histomorphometric features from H&E stained WSIs, they were able to predict the ODx risk category (a prognostic gene-expression panel) of breast cancer patients with high precision. However, in line with the objectives of my thesis, machine learning–based methods that leverage histopathological images for survival analysis are of more relevance. These are discussed next.

Cancer prognosis Survival analysis is broadly defined as the analysis of data that involve the time to the occurrence of an event of interest. In the quest for cancer prognosis, the event of interest is the death of an individual due to the disease. It differs from other types of statistical analysis as survival data can be censored, that is, the event of interest is only partially observed for some individuals. Given time points A and B that define the observation period of a study, and time points C and D that define the time a subject is at risk, Figure 2.1 illustrates the different types of censoring.

There are different statistical approaches to survival analysis. Likelihood-based multivariate analysis using Cox regression is one of the most employed methods for survival analysis, primarily due to its already wide adoption and recognition in the medical community, but also interpretability, and ability to handle censored data directly. However, Cox regression assumes that censoring is noninformative or in other words that there is no correlation between censoring and the event of interest [125], an assumption that is rarely tested in most of the existing literature [66]. Moreover, the underlying proportional-hazard assumption i.e. that the covariate effect remains constant over time, is more often than not violated with high–dimensional data, despite being a crucial prerequisite to to the method [178, 66]. The application of traditional machine learning algorithms for cancer prognosis therefore presents an attractive alternative for survival analysis of high–dimensional data, both for their less strict nature with data assumptions, and for their capacity to detect useful higher dimensional, nonlinear effects between the handcrafted features [23, 26]. Even though a body of work exists in the application of machine learning for cancer prognosis [202, 23], doing so with the use of whole tissue slide images is far scarcer, and indeed two of the publications of this thesis (Chapters 3 and 4) were amongst the first ones to do so. Instead, a larger body of work exists around image analysis (for prognosis) on a different type of histopathological images, called tissue microarray (TMA)

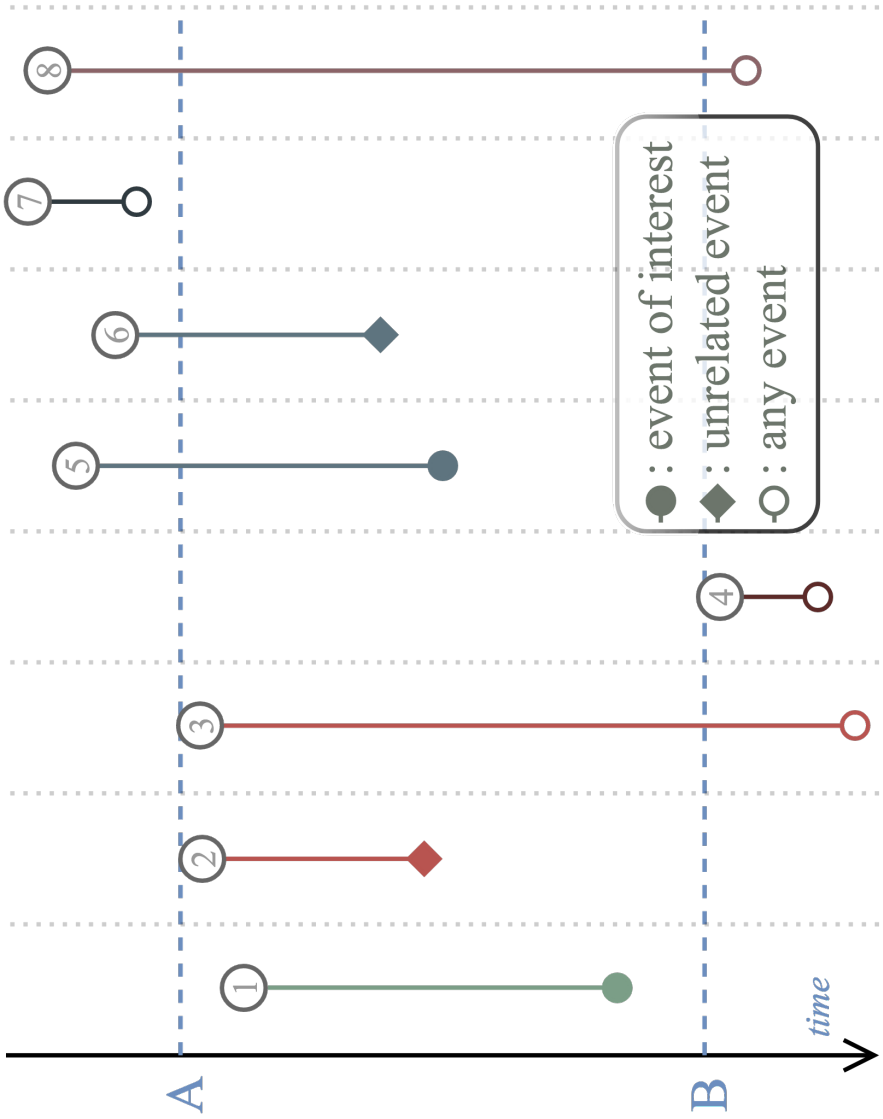


Figure 2.1: An illustration of different types of data censoring. Case 1 suffers from no censoring as surveillance is possible throughout and the event of interest occurs. Cases 2 and 3 are both *right-censored* but for different reasons; the monitoring of case 2 is disrupted due to an unrelated event whereas case 3 experiences some event only after the observation period $A - B$ is completed. Case 4 is *completely right-censored* and constitutes an edge case where the subject becomes at risk and experiences an event only after $A - B$. Case 5 is *left-censored* since the subject becomes at risk before the observation period begins. Cases 6 and 8 are simultaneously left- and right-censored, also known as *doubly censored*, and case 7 is *completely left-censored*.

images [12, 140]. Each TMA is a collection of cylindrical tissue samples taken from a single tissue specimen. TMA-based analysis, however, is different than WSI analysis (millions versus billions of pixels), and is therefore beyond the scope of this thesis.

2.3 Deep learning

In the previous decade most approaches focused on finding ways to explicitly extract features from images for models subsequently to employ [135, 53]. Therefore, feature extraction and model development were two distinct, independent stages that were performed sequentially, and where the former was based on human intuition of what constitutes a good feature. Automating this process through the use of convolutional neural networks (CNNs) has been shown to result in more discriminative features tailored for the problem at hand [61, 6, 166, 199]. This is one of the reasons behind the success of deep learning, and more broadly, neural network based learning, as feature extraction became a learning process, fundamentally intertwined with the learning of model parameters. The analysis of multi-gigabyte tissue images with deep learning is hampered by the challenges I introduced in Chapter 1, namely the "where" and "what" problems, as well as the multiple sources of noise, and the varying levels of label granularity.

2.3.1 The "where" problem

Practitioners need to come up with ways for either approximating the spatial distribution of the signal from gigapixel images, or performing some form of dimensionality reduction on the WSIs themselves.

Grid–like patch-based processing (or exhaustive tiling) has been the main method in the literature, yet for many clinical tasks such an approach can be inadequate, time and money consuming, and difficult to scale. Exhaustive tiling represents a "brute force" approach to the "where" problem by extracting image patches across the WSI at a predefined resolution, typically in a non-overlapping grid fashion, thereby imposing an *a priori* belief on the best properties of the extracted patches (magnification, field of view, location, etc.).

Patch extraction

Strongly supervised One way of identifying and extracting relevant information from gigapixel images relies on the use of annotations from domain–specific experts. For example, annotated slides with careful outlining of each lesion would provide a good approximation for the spatial distribution of salient regions ("where" problem) needed for metastasis detection [8]. There exists a relatively large body of work that follows this paradigm [131, 201, 127, 117, 113, 222, 120, 187]. Most of these approaches extract the ROIs from a single magnification level, e.g. the largest available

at 20× or 40×. A few, such as the approach of Sui et al. [187], instead extract patches from annotated areas at multiple magnification levels.

The fully–supervised nature of these approaches, however, limits their applicability to many clinical tasks for which annotations to this extent is either extremely laborious and expensive, or simply infeasible (e.g. cancer prognosis).

Weakly supervised In most cases ground truth labeling is done on the level of WSIs (slide–level) as opposed to individual patches or pixel–level. In the absence of higher granularity of labelling, the literature is divided into three main ways of tackling the "where" problem. The most prominent approach is to tile the entirety of a WSI, only perhaps excluding patches that do not meet certain image–based criteria based on computational methods (e.g. otsu, entropy, HSV colour space transformation, etc.) [162, 126, 194, 137, 55, 181, 192, 39, 94, 192]. The second approach involves random sampling from a grid–like patch population [90, 48]. One of the primary limitations of methodologies that use either of the above two approaches emerges as consequence of analysing a large input image by means of independent analysis of smaller patches. In particular, such approaches are inherently unable to capture information distributed over scales greater than the patch size. Explicitly modeling the spatial correlations between the patches has been proposed as a potential solution [127, 117, 219]. A few recent works have also employed instance–level self–supervision, under the multi–instance learning paradigm, to mitigate for the highly unbalanced nature of tiling (i.e. having only a small number of instances in a bag that are representative of the WSI) [126, 55].

Patch selection

Attention BenTaeib and Hamarneh employed a recurrent visual attention network that finds sub-regions of interest within images of $5,000 \times 5,000$ pixels [14]. Notably, these high–resolution images were tiled from fixed magnification scales. It is interesting to highlight that the processing of higher resolution patches came as a consequence of not employing any type of upsampling within the method. On the other hand, Qaiser and Rajpoot [167] used an attention network on images with $1,024 \times 1,024$ pixels, and at 2.5× magnification scale, to identify, extract, and process patches from higher, predefined magnification scales (10× or 20×). This approach, along with many others [170, 142, 205], attempts to solve a more complicated optimization problem, one that is non–differentiable (typically with reinforcement learning or variational methods [108]). More recent work has turned to differentiable alternatives [108, 170, 221, 52].

Nested attention Unfortunately, none of the above approaches can be employed on gigapixel images directly [118], and, instead, static, predefined preprocessing is still required. In parallel to our work in Chapter 6, Kong et al. [118] were the first to introduce the concept of nested attention, and by extending the attention module introduced by Katharopoulos

and Fleuret [108], proposed a two-layer hierarchical attention model that enables end-to-end training of deep learning models from WSIs.

2.3.2 The "what" problem

Although other computer vision approaches exist, convolutional neural network (CNN)–based methodologies have emerged as the most effective and popular choice as a way of automatically learning image features rather than handcrafting them [61, 6]. A CNN typically excels with image sizes of less than one million pixels [182, 91, 189, 218, 95]. Although there have been a few recent works that explored the use of higher resolution images (e.g. up to 8192×8192 [161]), the current state of the hardware cannot enable CNN–based learning directly from images with billions of pixels. Therefore, most methodologies address the "what" problem in the context of image patches (from WSIs), and not in the context of WSIs. A large body of work has also investigated better solutions to the "what" problem by incorporating contrastive loss [126], task-specific self-supervision [120], or using better pretrained networks [55].

2.3.3 Multiple sources of noise

Through training, the human brain can become adept at ignoring artifacts and staining variability, and honing in the visual information necessary for an accurate diagnosis. To facilitate an analogous outcome in deep learning models, there are generally two approaches that can be followed. The first involves explicit removal of artifacts (e.g., using image filters), as well the normalization of color variability [141]. In contrast, the second approach takes on a less direct strategy, augmenting data with often synthetically generated data which captures a representative variability in artifacts and staining, making their learning an integral part of the training process. Both approaches have been employed with some success to correct the variation from batch effect or from archived clinical samples from different clinics [29] though this finding has not been universal [131].

2.3.4 High–level clinical tasks

Despite the challenging nature of WSI analysis, when patch level labels are available, patch sampling coupled with hard negative mining can train deep learning models that in many cases match and even surpass the accuracy of pathologists on a number of high–level clinical tasks [166, 61, 8]. However, in many cases, sufficient ground truth cannot be attained either because it is very laborious and expensive, or simply because it is infeasible. In addition, this level of supervision may be limiting the potential of deep learning models as the models can only be as good as the annotations provided. Therefore, methods able to build deep learning solutions based on low–granularity labels are more desirable.

To work with slide and patient level labels, the current approaches focus on the *where* problem, or in other words, on approximating the spatial distribution of saliency. There are generally two approaches towards the *where* problem. The first uses a type of meta-learning, where in order to optimise the *where* problem, the *what* problem has to first be optimised. The second approach attempts to optimise both *what* and *where* problems simultaneously in an end-to-end setting. This is done by either forwarding a set of patches through a CNN and attending on a few or by localising and attending to a single patch at each time step.

Cancer prognosis Most early methods for deep learning survival analysis employ a Cox Proportional Hazards (CPH) layer on top of a CNN [196]. By using a non–linear model, such as a CNN, non–linear relationships between the features and the risk of death can be captured. However, in addition to certain limitations imposed by the nature of optimization that takes place within deep neural networks (e.g. batch based estimation of hazard ratios), the CPH layer is constrained by the nature of the method which assumes that the effect of the input features remains constant over time [196]. A new body of work has therefore explored different approaches, including the use of different loss functions [147], and reformulating survival analysis into a multitask (based on different time cutoffs) binary classification problem [70, 195]. More recently, various groups have explored the application of deep learning on WSIs for end-to-end prognosis using limited annotations (further discussion is provided in Chapter 5) [109, 183, 210].

Chapter 3

Stage II colorectal cancer prognosis - handcrafted features

3.1 Problem formulation

In present clinical practice, the main prognostic factors for colorectal cancer (CRC) comprise: (T) depth of tumour penetration through bowel wall, (N) presence or absence of nodal involvement, and (M) presence or absence of distant metastases. These form the basis of the five stage TNM staging system [28]. Stage 0 is least severe, with all the lesions restricted to the mucosa and the lamina propria. Local excision or simple polypectomy with clear margins is the most common treatment option. In Stage I, cancer may have grown into the muscularis mucosa or into the muscularis propria but has not spread deeper into the colon muscle wall, to nearby lymph nodes or other distant sites. Because CRC at this stage is still localized, it also has a high cure rate with wide surgical resection and anastomosis. Stage II characterizes CRC that has spread to or beyond the serosa and may have grown into nearby tissue or organs, but not to the lymph nodes and has not metastasised. Surgical resection is again the standard treatment, however high-risk patients may be offered chemotherapy. Stage III is characterized by lymph node involvement and the standard treatments are wide surgical resection and anastomosis, and adjuvant chemotherapy. Stage IV disease is characterized by metastatic disease. The treatment of CRC at this stage largely depends on the sites of metastatic disease.

In this chapter, I consider the problem of patient prognosis in the context of stage II CRC. Stage II patients do not experience nodal (N) or distant (M) metastasis of their cancer and so only the depth of local invasion (T) is reported under TNM staging (as either T3 or T4). The risk of disease–specific death within 5 years for stage II CRC patients is estimated to be 20%, and rises to 35% given a 10 year window [51, 150]. Nevertheless, there are no definite criteria for selecting which stage II patients (high-risk

patients) should undergo adjuvant chemotherapy [10, 132]. It is therefore imperative that better prognostic models for stage II CRC patients are developed to better aid clinical guidance, reduce the survivability variance, and ameliorate treatment research.

The aim of this work is to investigate whether machine learning models can leverage information from histopathological features of each patient to better assess their risk of disease–specific death when compared to the current clinical gold standard. The prognostic value of multiple features that are quantified across a continuum at patient–level, is difficult and laborious to sort and analyse by hand. Hence, machine learning is applied in order to identify and understand the optimal features that allow patient stratification. In particular, the present work makes the following contributions:

- A novel, principled framework for data–driven machine learning based personalized prognosis of stage II CRC cancer outcomes.
- Prognostic models that perform significantly better than the current clinical gold standard based on internal validation.
- Clinical insights into the disease, and empirical evidence behind the prognostic value of novel histopathological features.

The process of feature extraction from the patient images was performed prior to this work by collaborators, and was reported in previously published work [36, 37]. The automated segmentation of objects of interest involved manual examination of all segmented objects by domain-specific experts to ensure high quality segmentation. In addition, Definiens proprietary software (Tissue Studio® and Developer XD™) was used, and as such, reproducibility using open source software cannot be guaranteed.

3.2 Methods

3.2.1 Cohort

My cohort consisted of 180 Scottish patients who had been diagnosed with CRC and who underwent surgical resection, with a minimum follow-up of 11.5 years. Patients that succumbed within five days of the surgery were excluded to ensure that surgical complications did not contribute to the cause of death, as were the patients that received adjuvant chemotherapy due to potential effects on survival [154]. In total, seven patients from the initial cohort were removed. Table 3.1 summarizes the key clinical and demographic characteristics of the remaining cohort (173 patients). In the cohort there were 86 males and 87 females. Six patients were diagnosed with T1 stage, 7 with T2, 122 with T3, 37 with T4, and for 1 person, the T stage could not be determined. Most patients (163) experienced no nodal involvement (N0). There were 8 patients diagnosed with N1, and

Table 3.1: Summary of patient cohort statistics.

Number of patients		173
Age at surgery (years)		
	Range	62.5 ± 33.5
	Median	67
Gender		
	Male	86 (50%)
	Female	87 (50%)
T Stage		
	TX	1 (1%)
	T1	6 (3%)
	T2	7 (4%)
	T3	122 (71%)
	T4	37 (21%)
N Stage		
	N0	163 (94%)
	N1	8 (5%)
	N2	1 (1%)
	N3	1 (1%)
M Stage		
	MX	9 (5%)
	M0	161 (93%)
	M1	3 (2%)
Site		
	Rectum	56 (32%)
	Colon	117 (68%)
Differentiation		
	Undetermined	3 (2%)
	Poor	25 (14%)
	Moderate	138 (80%)
	Good	7 (4%)

1 patient for each of N2 and N3. Similarly, most patients were free of metastasis (161), with only 3 having being diagnosed with M1, and for 9 patients the M stage was not determined. The tumour site of origin was the colon and rectum for 117 and 56 of the patients respectively. Finally, the differentiation was poor, moderate, and good for 25, 138, and 7 patients, and for 3, differentiation could not be determined.

The use of tissue samples was approved by the East of Scotland Research Ethics Service (13/ES/0126). Further ethical clearance was not required as the acquired data was anonymized.

3.2.2 Feature extraction – handcrafted features

The digitization of the tissue samples, and subsequent quantification and extraction of histological features is described in the works of Caie et al. [36, 37]. For completeness, both processes are briefly described hereunder.

Tissue samples were prepared for multiplex IF with pan cytokeratin (panCK) and D2–40 antibodies, along with DAPI stain. These allowed for the detection of epithelial cells, lymphatic vessels, and cell nuclei. Detecting epithelial cells within lymphatic vessels (known as lymphatic vessel invasion) is what motivated the inclusion of the D2–40 marker, since it has been associated with nodal metastasis and poor prognosis in multiple carcinomas.

The invasive front was manually identified through the panCK channel of each WSI as captured at 4× magnification level. Fifteen evenly spaced high–resolution (20× magnification) images were captured across the invasive front of each sample. Regions of interest (ROIs) (including stroma, tumour glands, invasive tumour subpopulations, lymphatic vasculature, and cell nuclei) were detected and segmented from each imported image using Definiens AG image analysis software package. Each ROI was described by a collection of morphometric, spatial, and fluorescence related characteristics resulting in 123 histopathological features. A comprehensive list of the features is provided in Appendix A. Tumour buds are PanCK$^+$ objects with 5 or less associated nuclei, whereas poorly differentiated clusters (PDCs) have more than 5 PanCK$^+$ associated nuclei. PanCK$^+$ objects with no associated nuclei are small objects (below $50\mu m^2$) often depicting necrosis.

For each patient, pathological and demographic data was collected as well. The former set comprises the level of differentiation, site of primary tumour, and the corresponding disease stage, and the latter the patient's age, gender, survival status, and time until either death, or the end of study. Except for the last two survival features, which were combined to provide the dependent variable of this work, the remaining features were only used for the analysis of experimental results, and not for the actual learning and prediction.

3.2.3 Survival analysis

Herein, the event of interest was the death of an individual due to CRC. In the cohort, some patients were right-censored either because the end of study was reached and the event of interest did not occur, or because the patients succumbed to a cause other than CRC (abbreviated as OTD–censoring) [125]. The problem at hand was formalized as a binary, supervised classification task, whereby the prediction was that of a good or bad prognosis, i.e. survived or not, respectively. Commonly used statistics of prognosis (5–year and 10–year) were investigated [215]. The addition of the 10–year survival cutoff was motivated by my interest in investigating

whether more long term prognosis is possible, and if so, whether features associated with long term prognosis could be identified. Patients that succumbed to CRC within the designated cutoff were denoted as patients with a bad prognosis whereas those that survived were denoted as patients with a good prognosis. Inevitably, patients that died to an unrelated cause prior to the prognostic cutoff, i.e. they were part of the OTD–censored data, had to be excluded. In particular, there was a 17% patient exclusion for the 5–years prognostic cutoff, and a 32% for the 10–years, reducing the patient cohorts to 143 and 117 respectively. It is worth mentioning that removing these patients does not introduce bias since time to censoring was random, i.e. OTD–censoring was not known *a priori*. A consequence of this approach is that survival analysis was turned into a binary classification problem. Furthermore, due to removing censoring, traditional ML models were readily employable.

3.2.4 Data preparation

I followed the standard approach to algorithm training and evaluation, by splitting the cohort into non-overlapping training, validation, and test subsets. In particular, data was first randomly (with stratification) split into two, 70% and 30%, the latter being the test subset. Using tenfold cross-validation, the former, large subset was in each iteration of the process further randomly split into training and validation subsets. It is worth noting that, given the key aim of the present work, while the evaluation corpus contain only stage II patients, patients of different stages were included in the training data set (see Table 3.1). My hypothesis was that in spite of not being the target population for the prediction, useful pathological patterns could be learnt from this data too, allowing a degree of interpolation to take place. Stratified sampling was employed in order to maintain the prognosis distribution of each cohort as a means of countering the imbalanced nature of the data, and thus avoid class under-representation [115]. Lastly, features were normalized to zero mean and unity variance.

The training set for 5-year prognosis has 99 patients (78/21 patients with good/bad prognosis). For 10-year prognosis, the training set has 81 patients (54/27 patients with good/bad prognosis). The testing sets for 5 and 10-year prognosis have 44 (37/7), and 36 (25/11) patients respectively. In the testing set for 5-year prognosis, all patients are stage II with 34 having a T3 stage, and 10 having a T4 stage. In the training set for 5-year prognosis, there are 9 Stage IV patients, 1 Stage III patient, 76 Stage II patients (60 with T3 and 16 with T4), and 13 Stage I patients (6 with T1 and 7 with T2). In the testing set for 10-year prognosis, all patients are stage II with 34 having a T3 stage, and 10 having a T4 stage. In the training set for 10-year prognosis, there are 9 Stage IV patients, 1 Stage III patient, 58 Stage II patients (48 with T3 and 10 with T4), and 13 Stage I patients (6 with T1 and 7 with T2).

3.2.5 Baseline classifiers and performance assessment

I adopted several well-understood baseline classifiers, with different underlying assumptions (explicit or implicit) and mathematical underpinnings. In particular, I compared classifiers based on support vector machines (SVMs) (with *LSVM* translating to a linear kernel SVM, and *RSVM* to an SVM with radial basis function as kernel), random forests (RFs), k-nearest neighbours (KNN), naïve bayes (NB), and logistic regression (LR) [35]. Elastic cox (proportional hazards) regression (CPH) was also implemented for the purpose of comparative analysis with the ML-based models and prognosis binarization.

In an effort to capture performance adequately on a highly imbalanced data set, the area under the receiver operating characteristic curve (AUROC), as opposed to accuracy [128], is adopted as the primary performance measure. This is also in line with related work in literature, i.e. associated with survival analysis, wherein the so-called concordance index (C–index) or C–statistic is employed with the latter being a generalization of AUROC over regression problems with censored data [138]. Given the aforementioned preprocessing steps, i.e. censoring is eliminated, and regression is turned into binary classification, AUROC is equivalent to C–index. It is worth mentioning that recent work has reported that AUROC may in fact not be the best metric when working with imbalanced data sets [40]. For the sake of consistency with related work and ease of comparative analysis, I also report specificity and sensitivity, and accuracy for the final models.

3.2.6 Model selection

The capability of a model to represent information, as well the efficiency of its learning is governed by a number of parameters. These parameters, referred to as hyperparameters, need to be set prior to training. However, finding the optimal or close to optimal set of hyperparameter values is challenging. The commonly used and probably the simplest approach, in the form of a grid search has limited applicability due to its intractability for complex models. A random search over predefined ranges of hyperparameters often produces better results while being computationally less demanding [17]. However, both techniques are naïve as they do not take into account historical patterns.

Sequential model based global optimization (SMBO) techniques adopt a more sophisticated approach, approximating the often computationally expensive fitness function with a simpler surrogate [19]. Different SMBO approaches optimize different criteria which then guide the surrogate of the fitness function. The one adopted herein is tree-structured Parzen estimator (TPE), which optimizes the so-called "expected improvement". Conceptually, TPE initially behaves like a random search, subsequently refining the search so that hyperparameter values associated with poor performance are not re-visited [19, 98]. This process is guided probabilistically, using

suitable densities or distributions associated with the type of each hyperparameter. Those used in the present work are summarized in Table 3.2. Finally, as the loss function (of TPE) I used the negated AUROC (or concordance index for CPH) resulting from tenfold cross-validation, averaged over 20 independent runs, and ran the optimization for 500 iterations.

3.2.7 Feature selection

To mitigate for the so-called curse of dimensionality, which becomes of increasing concern with "short fat data" (see Chapter 2 for the definition), I examined the use of dimensionality reduction in the context of the problem at hand [86, 44]. In particular, motivated by their successful use in the existing literature [85], I employed sequential floating forward selection (SFFS) and sequential floating backward selection (SFBS) [85, 165, 102]. SFFS and SFBS constitute more sophisticated methods than the strictly monotonic sequential algorithms known as sequential forward selection (SFS) and sequential backward selection (SBS). Given an empty feature set, SFS sequentially adds the feature that improves a specified metric the most until the desire number of features is reached. SBS works in the same manner except that it begins with the whole feature set and proceeds by removing features. The main downside of both is that they do not reconsider the value of each added (SFS), or removed (SBS), feature, down the line – this is known as the "nesting effect". SFFS and SFBS attempt to address this by allowing the recursive removal or addition, respectively, of features at each step of the process. The implementation is provided in Appendix A.

3.3 Results

3.3.1 Full feature set based prognosis

Each baseline classifier's hyperparameter values are learnt by maximizing the corresponding average AUROC on the validation data corpus. Table 3.3 summarizes the results. The average AUROC across all classifiers was 0.89 both for 5 and 10-year prognosis. One-way analysis of variance (ANOVA) and Tukey's honest significance difference test (THSD) showed no statistical significance between classifiers for 10-year prognosis. The only statistically significant difference is that between NB and LR based approaches for 5-year prognosis (ANOVA p-value< 0.01, THSD p-value< 0.003).

To demonstrate the importance of model selection, I also compared the performance of all classifiers using hyperparameter values which were learnt as described in the previous section, and with the *a priori* set hyperparameters values based on the scikit–learn library [159]. As expected, using the latter approach a drop in the average AUROC was observed both for 5 and 10-year prognosis, to respectively 0.82 (approximately 8.0% drop) and 0.85 (approximately 4.5% drop). The results are visualized in Figure 3.1.

Table 3.2: The search space of each classifier based on the distributions over its hyperparameters (n.b. F denotes feature count; for biased categorical distributions, tuples (p_s, v) designate the sampling probability and the value assigned). C is a hyperparameter for regularization (smaller values specify stronger regularization), K is the number of neighbours in KNN, and P is the power parameter for the Minkowski metric. For CPH, the estimated set of alphas is retrieved based on an initial fit of the model on the training set, i.e. differs between different training subsets.

Classifier	Hyperparameter	Distribution	Values
LSVM	C	Log-uniform	$[\ln(1e{-}5),\ \ln(1e2)]$
	Class weight	Categorical	Balanced or none
RSVM	C	Log-uniform	$[\ln(1e{-}5),\ \ln(1e2)]$
	Gamma	Log–uniform	$[\ln(1e{-}3),\ \ln(1e3)]$
	Class weight	Categorical	Balanced or none
LR	Type of penalty	Categorical	L1 or L2
	C	Log–uniform	$[\ln(1e{-}5),\ \ln(1e2)]$
	Class weight	Categorical	Balanced or none
RF	Number of trees	Log–uniform integer	[10, 1000]
	Criterion	Categorical	Gini or entropy
	Maximum features	Biased categorical	[(0.2, $\sqrt{F}$), (0.1, $\ln(F)$), (0.1, F), (0.6, U[0, F])]
	Maximum depth	Biased categorical	[(0.1, 2), (0.1, 3), (0.1, 4), (0.7, None)]
	Bootstrap	Categorical	True or False
	Class weight	Categorical	Balanced or none
KNN	K	Log–uniform integer	[1, 50]
	Weights	Categorical	Uniform or Euclidean distance
	Metric	Categorical	Balanced or none
	P	Categorical	Balanced or none
CPH	C	Uniform	[0, 1]
	alpha	Categorical	Proposed categories

Table 3.3: Average AUROC and standard deviation (for n = 200) of trained classifiers on the training set using 20–times repeated 10–fold cross-validation.

	5-year	10-year
LSVM	0.89 ± 0.12	0.89 ± 0.13
RSVM	0.89 ± 0.13	0.89 ± 0.12
LR	0.91 ± 0.12	0.90 ± 0.13
RF	0.89 ± 0.13	0.89 ± 0.12
KNN	0.88 ± 0.12	0.89 ± 0.13
NB	0.86 ± 0.14	0.88 ± 0.12

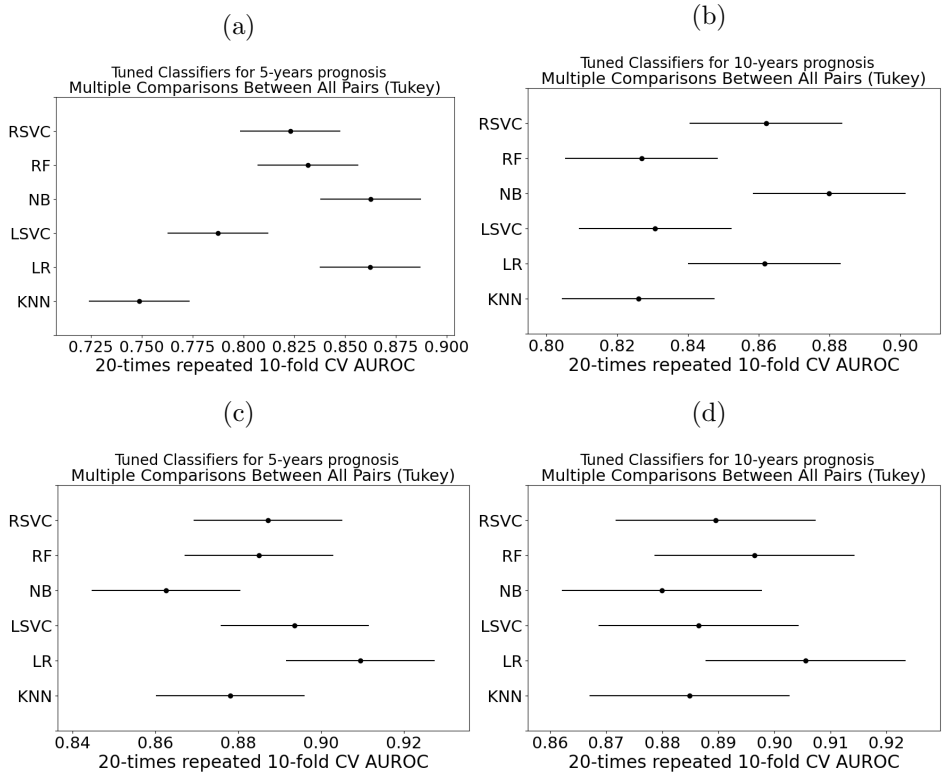


Figure 3.1: Tukey's significance difference test. No hyperparameter learning was employed in the experiments corresponding to the plots (a) and (b), in contrast to (c) and (d).

3.3.2 Reduced feature sets

Feature selection

The evaluation of each subset of features was performed by 10–fold cross validation on the training data. To reduce outcome variability caused by stochastic effects I adapt the method proposed by Dune et al.[59]. In particular, I performed SFFS and SFBS 40 times using different random partitions for the cross validation, each time retaining the feature subset that achieved the best performance. The optimal feature subset, i.e. the one corresponding to the highest 10–fold cross validation, from all the runs, were then aggregated (see Figures 3.2 and 3.3).

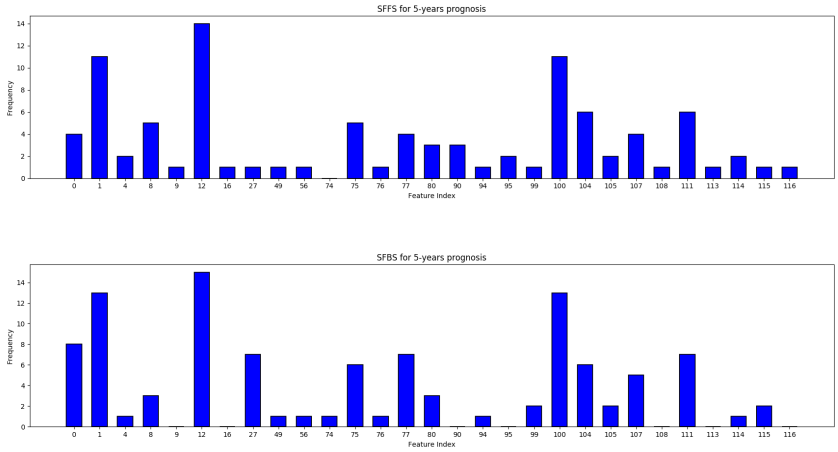


Figure 3.2: Frequency of occurrence of each feature from the 20 runs of SFFS and SFBS each for 5-year prognosis. Only features with at least one occurrence are shown for clarity.

Feature subsets from SFFS and SFBS were combined and sorted based on the frequency of occurrence. Starting with an empty set, features were added in incremental fashion based on their ranking using 20 times repeated 10-fold cross validation. The subset of features that achieved the highest averaged AUROC was selected for each prognostic term, see Table 3.4.

Experiments

I followed the same approach to classifier training, model selection, and evaluation as in the previous section (Section 3.3.1). The sole difference is that instead of the full feature set, for this set of experiments a reduced set of selected features (as described in the previous section) was used.

As expected, a significant improvement in performance is observed already at the coarsest level of analysis, with the average AUROC across classifiers reaching 0.94, both for 5 and 10-year prognosis. In line with

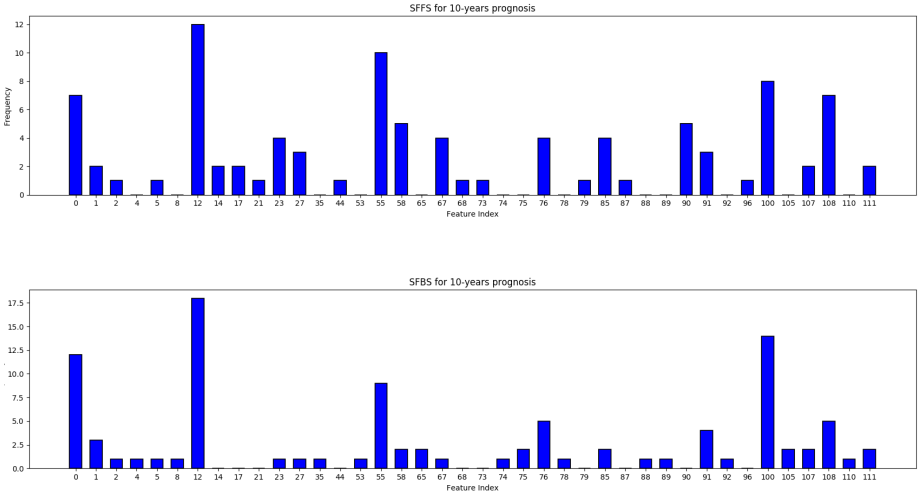


Figure 3.3: Frequency of occurrence of each feature from the 20 runs of SFFS and SFBS each for 10-year prognosis. Only features with at least one occurrence are shown for clarity.

Table 3.4: Features of significance to both prognosis terms, and those which were specific to a particular term; six features were used for both 5 and 10-year prognosis.

	#	Features
Unique to 5-year prognosis	4	Nuclei in tumour mean DAPI intensity, number of CK objects with no associated nuclei, average DAPI intensity (tumour area)
Unique to 10-year prognosis	3	Nuclei in tumour mean D240 intensity, mean compactness of tumour glands, number of PDCs
Common to both prognoses	3	Nuclei in tumour bud mean DAPI intensity, tumour gland relative area (%), sum area of vessels
CK pancytokeratin, *PDCs* poorly differentiated clusters		

my previous findings, no statistically significant difference was observed between different classifiers, except for the inferiority of RFs for 10-year prognosis (ANOVA p-value< 0.0001, THSD p-value< 0.01). Just as in the previous set of experiments, my data driven approach to hyperparameter selection was found to effect a statistically significant improvement over their being set *a priori*; see Figure 3.4.

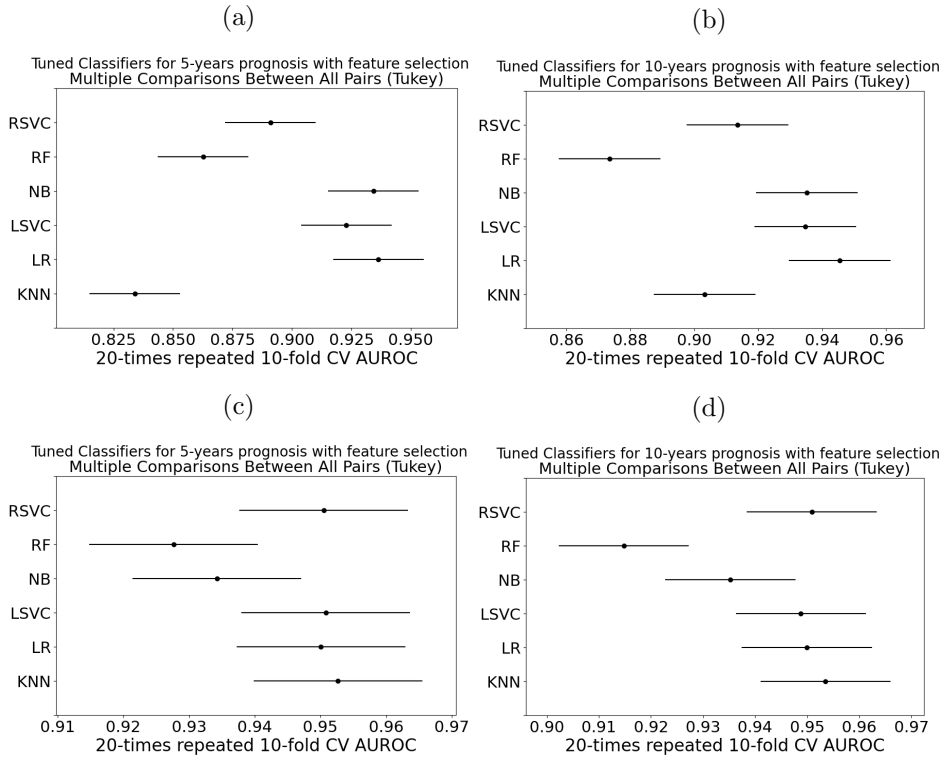


Figure 3.4: Tukey's significance difference test. No hyperparameter learning was employed in the experiments corresponding to the plots **a** and **b**, in contrast to **c** and **d**.

Table 3.5: Average AUROC and standard deviation (for n = 200) of each trained classifier using only features selected by SFFS and SFBS. The experiments were performed by 20 times repeating 10–fold cross–validation.

	5-year	10-year
LSVM	0.95 ± 0.08	0.95 ± 0.08
RSVM	0.95 ± 0.08	0.95 ± 0.08
LR	0.95 ± 0.08	0.95 ± 0.08
RF	0.93 ± 0.11	0.92 ± 0.10
KNN	0.95 ± 0.08	0.95 ± 0.07
NB	0.93 ± 0.10	0.94 ± 0.09

3.3.3 Final evaluation – internal validation

It can be readily seen that classifiers trained on the subset of features selected by SFFS and SFBS performed better, as illustrated in Tables 3.3 and 3.5. Though simple, the best performing classifier was found to be KNN based classifier (with the Minkowski distance metric) both for 5-year ($k = 36$) and 10-year prognosis ($k = 28$), as measured by the 20-times repeated 10-fold cross validation AUROC. The threshold was defined as the cutoff at which the sum of true positives and true negatives maximizes. Based on the training data set, the threshold was set to 0.222 for 5-year prognosis and 0.250 for 10-year prognosis. It is interesting to observe that the KNN based classifier performed poorly when the default hyperparameter values (of scikit–learn) were used. This observation further highlights the importance of properly tuning machine learning models to the task and data at hand [21].

Kaplan-Meier (KM) survival curves were employed to visualize the difference in survivability between the predicted prognosis groups, and, for object quantification, the log-rank test was used (summary tables A.1 and A.2 are also provided in the appendix). For 5-year prognosis, my KNN based approach achieved the AUROC of 0.77, effecting a good separation patients into high and low risk (p-value< 0.02). On 10-year prognosis, the classifier demonstrated performance which can be described as nothing short of outstanding, achieving AUROC of 0.94, and even better separation between high and low risk patients (log-rank test p-value< 0.0001). The sensitivity of 42.9%, specificity of 89.2%, and accuracy of 81.8% were achieved for 5-year prognosis, and the sensitivity of 100%, specificity of 84%, and accuracy of 88.9%, for 10-year prognosis. In comparison, the differentiation (poor/moderate vs. good), which is considered a prognostic factor independent of TNM, and T stage discrimination (T3 vs. T4) results are summarized in Figures 3.5, 3.6, and 3.7, as well as in Table 3.6.

To compare my method against CPH, I followed the same approach to hyperparameter tuning as with the machine learning models. Moreover, (i) the patients in the testing sets for 5 and 10–year prognosis are kept the same, (ii) both the full feature set as well as the reduced one (for each prognostic cutoff) are used, and (iii) I train CPH models with and without prognosis binarization, where in the latter case, I include the previously removed censored patients in the training set (the training sets size increases from 99 patients to 129, and from 81 to 137 for 5 and 10-year prognosis respectively). The results are shown in Table 3.7. For 5-year prognosis, CPH models performs better without feature selection, whereas for 10-year prognosis, no difference is observed. When including censored patients in the training set, surprisingly, no substantial gain in performance is observed, and in fact in most cases, there is a drop. The best CPH model for each prognostic cutoff is evaluated in the testing set. It can be readily seen that the KNN models performed better than the CPH models for both prognostic cutoffs.

Table 3.6: Summary of low vs. high risk patient separation results.

	Differentiation (5/10-year)	T stage (5/10-year)	KNN (5/10-year)
Specificity	0.95/0.88	0.82/0.84	0.89/0.84
Sensitivity	0.39/0.36	0.43/0.46	0.43/1.00
Accuracy	0.84/0.72	0.75/0.72	0.82/0.89
AUROC	0.62/0.62	0.62/0.65	0.77/0.94

Table 3.7: Comparison between the proposed machine learning method and CPH with and without feature selection for both 5-year and 10-year prognoses. A "-" means that there was no binarization and that censored data were included in the training set. *FS* stands for feature selection.

Method	FS	5-year	10-year	Train set	Test set
KNN		✓		88%	-
KNN	✓	✓		95%	77%
CPH	✓	✓		72%	-
CPH	✓	-		71%	-
CPH		✓		81%	-
CPH		-		83%	71%
KNN			✓	89%	-
KNN	✓		✓	95%	94%
CPH	✓		✓	84%	71%
CPH	✓		-	82%	-
CPH			✓	84%	70%
CPH			-	82%	-

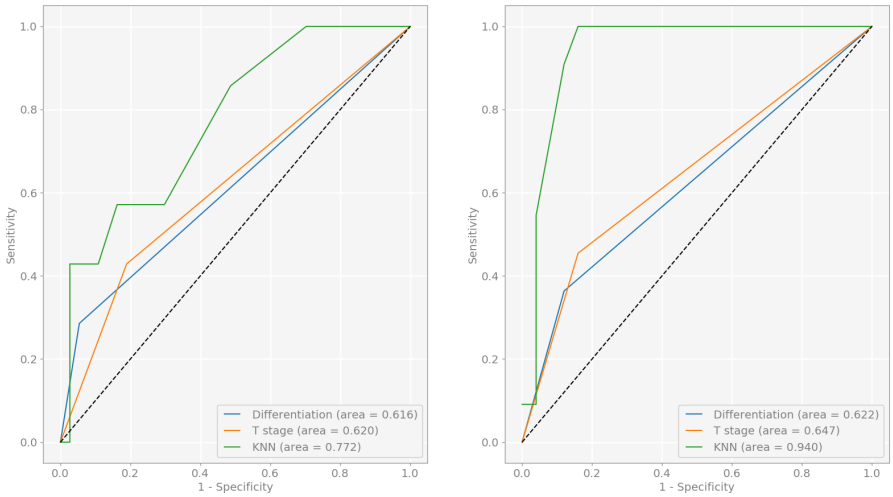


Figure 3.5: ROC curves for the two prognostic terms of interest.

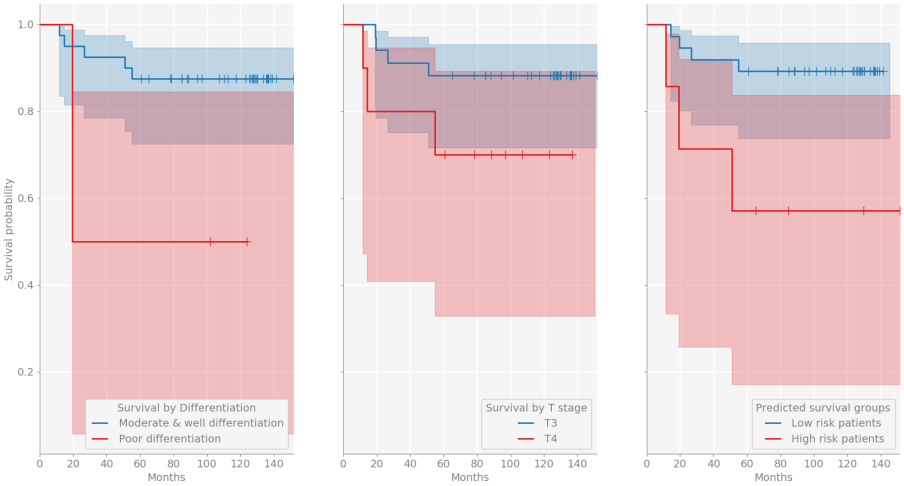


Figure 3.6: KM curves for 5-year prognosis.

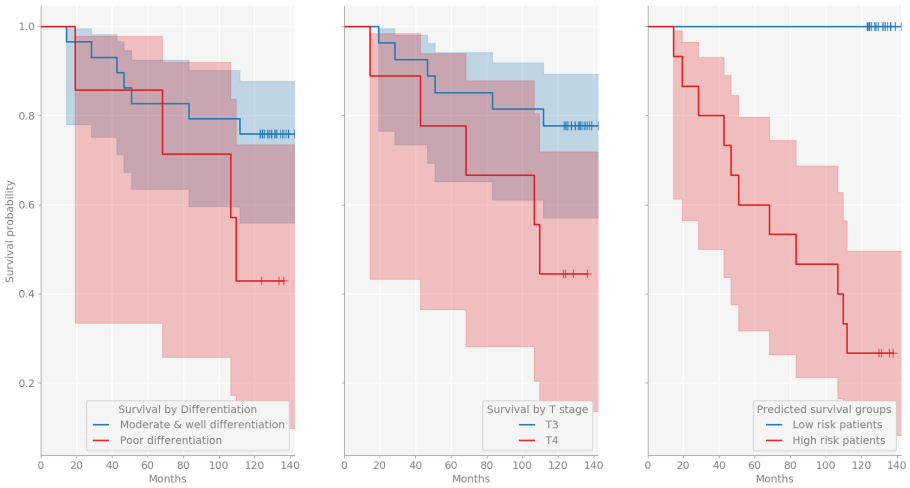


Figure 3.7: KM curves for 10-year prognosis.

3.4 Discussion

Colorectal cancer is a highly heterogeneous disease which limits the prognostic accuracy of the simplistic TNM staging system or singular features such as tumour budding [93], or lymphatic vessel invasion and density [33]. Prior work on the use of automated image analysis and machine learning, and other types of cancer has focused on parameters solely from tumour cells [204, 198]. However, the evidence from an increasing number of studies suggests that the tumour microenvironment is just as informative [186, 38, 92, 100] which motivated us to use information not only from tumour nuclei but also from numerous hierarchical features such as texture, morphology, fluorescence intensity, and spatial relationships across the entire tumour–immune microenvironment.

Hence, I introduced a novel and carefully crafted machine learning based framework capable of personalized prediction of survival for stage II CRC patients. My methodology was shown to outperform significantly the current gold standard in the form of TNM staging. Specifically, an AUROC of over 77% and 94% for 5 and 10-year prognosis respectively was achieved, compared to the clinical gold-standard of T stage, which stratifies patients with the AUROC of approximately 62% both for 5 and 10-year prognosis, and the differentiation, which achieves the corresponding AUROC of approximately 62% and 65% respectively. When compared to CPH, the most widely employed method for multivariate survival analysis in the literature, my models performed better consistently across various setups. When evaluated on the testing set, there was a 6% difference in the AUROC for 5-year prognosis, and a 13% difference for 10-year prognosis. Finally, high interpretability of the proposed approach was demonstrated (see Table 3.4), allowing clinicians to gain new insight by identifying prognostically the most salient features.

Confirming findings from prior empirical research as well as one of the premises of the present work, my experiments demonstrated that a diverse set of characteristics of the entire microenvironment have a prognostic value. This explains the outstanding performance of my method and the major improvement on the current state of the art which focuses on a single aspect thereof (usually tumour cells). DAPI intensity within the nuclei of tumour buds was consistently found to carry the greatest prognostic weight, which too agrees with previous empirical findings [24, 139]. Furthermore, this feature was highly correlated with tumour bud nuclei morphometry whereby features linked to larger and more irregular shaped nuclei were associated with poorer prognosis. Tumour gland nuclear morphometry, also found to be of major prognostic importance, has also been identified in the past [149, 63]. Other selected features included known histopathological features such as the number of PDCs [11], the number and area of lymphatic vessels [33], and the shape and area of tumour glands [99, 169].

It is interesting to observe and comment on my finding that certain

features were specifically associated with a particular prognostic term; a hypothesis that is relatively under explored in the literature, yet, if proven right, can have significant implications [212]. Having looked at this in detail, I found high correlation between these features and survival outcomes, suggesting that the features are not specific to set survival times *per se* but are rather associated with poorer outcomes. For example, the number of small pan cytokeratin positive objects with no associated nuclei was found to be an important feature for 5-year survival. On the other hand, the number of PDCs was found to be an important prognostic feature for 10-year survival. Nevertheless, both were highly correlated with the number of tumour buds.

The proposed method is likely far from clinical adoption primarily due to the semi-automated nature of the feature extraction approach. Feature robustness against differences in staining and sectioning would need to be assessed in new cohorts. Moreover, further validation would be needed for the machine learning models to ensure that they are not confounded by any subtle differences in scanning, staining, etc. It is likely that a more refined methodology might be needed for feature extraction, one that involves deep learning. As future work, the problem could be reformulated into a single multitask binary classification problem (based on different time cutoffs) rather than two independent problems [70, 195].

3.5 Implementation details

The machine learning framework was implemented using the following packages in Python: Pandas [145], Numpy [89], scikit-learn [159], matplotlib [97], SciPy [200], hyperopt [18], and lifelines [54]. The feature selection algorithms SFFS and SFS were implemented as shown in Appendix A (SFBS and SBS were omitted for brevity). Other key code sections are also provided in the appendix. The CPH models were implemented using scikit–survival [164].

Chapter 4

Muscle-invasive bladder cancer prognosis - handcrafted features

4.1 Problem formulation

Like in the previous chapter, the primary goal of this work is to develop prognostic models that leverage patient–level data (herein, WSIs and clinical records) in order to identify those patients that are at risk of disease–specific death. In particular, the problem of MIBC prognosis is addressed. As described in Chapter 1, patients with MIBC face a rather poor prognosis, with half succumbing to the disease within 5 years. To decrease mortality rate, patients with a high risk of disease-specific death need to be identified more precisely, thereby allowing for better patient management and new treatments to be tested in the high-risk group.

MIBC prognosis is a fitting problem for the machine learning framework I proposed in the previous chapter, given its challenging nature [175, 81]. Based on the same framework (with minor modifications) as in the previous chapter, I develop machine learning based prognostic models for 5–year prognosis with different combinations of image, clinical, and spatial features. In particular, the contribution of this chapter is threefold:

- This is the first work where machine learning is applied on histopathological features that were derived from across entire MIBC tissue sections with multiple fluorescence immune markers and across both the tumour core and the invasive front of whole slide immunofluorescence images.

- Using image, spatial, and clinical features, the proposed machine learning methodology improves the accuracy of 5–year prognosis for MIBC patients by a large margin when compared to the current gold standard, TNM.

- My findings reinforce the importance of the immune contexture in cancer prognosis.

The processes of feature extraction from the patient WSIs, including that of cell nuclei detection, epithelial cell segmentation, and cell classification, were developed and executed by our collaborators in Definiens GmbH. The robustness of these methods (including segmentation) has been tested in previous works (e.g. the work by Brie et al. [30]) and is therefore not examined in this chapter.

4.2 Methods

4.2.1 Cohort

Tissue specimens from patients with MIBC who underwent radical cystectomy at Royal Infirmary and Western General Hospital in Edinburgh between the years 2006 to 2013 were collated into a cohort. Patients were excluded from this study either due to incomplete clinical records, extensive tissue section artefacts, or data censoring. The final study cohort was comprised of 78 patients. Archived formalin-fixed paraffin-embedded (FFPE) tissue blocks presenting the deepest invasion of cancer were selected for each patient based on both macroscopic and microscopic examination of haematoxylin and eosin (H&E) stained slides by a pathologist and a research scientist. The corresponding unstained tissue sections were collected from the NHS Lothian NRS BioResource Research Tissue Bank, conforming to protocols approved under the ethical status granted by the East of Scotland Research Ethics Service (Ethical Approval Ref: 10/S1402/33) and with written informed consent from all the patients. All experiments were performed in accordance with the relevant guidelines and regulations.

Clinicopathological data that included age, sex and TNM stage status along with survival data was retrieved from available electronic medical records. Patients were followed up for a total time of 113 months with a median survival time of about 24 months. Median age of the patients was 68 years (range 29–87 years) with 43 males and 35 females. According to the TNM staging system guidelines [4], my cohort consists of 17 stage II, 29 stage IIIA, 5 stage IIIB, and 27 stage IV patients. Twenty seven patients had distant metastasis at time of surgery. No positive lymph nodes were found in 57 patients and 1–2 lymph nodes contained tumour cells in 21 patients. Of the 78 patients, 53 patients died due to bladder cancer. The clinicopathological characteristics of the cohort are summarised in Table 4.1.

In order to maintain the anonymity of the patient information, the samples were de-identified prior to conducting this study.

Characteristics	**Summary**
MIBC patients	$N = 78$
Median survival (range)	19 (1-113) months
Age	66 ± 11 years
Gender	55% Male; 45% Female
TNM stage	
II	17 (22%)
IIIA	29 (37%)
IIIB	5 (6%)
IV	27 (35%)
Tumour (T)	
T2	18 (23%)
T3	39 (50%)
T4	21 (27%)
Node (N)	
N0	57 (73%)
N1	13 (17%)
N2	8 (10%)
Metastasis (M)	
M0	51 (65%)
M1	27 (35%)

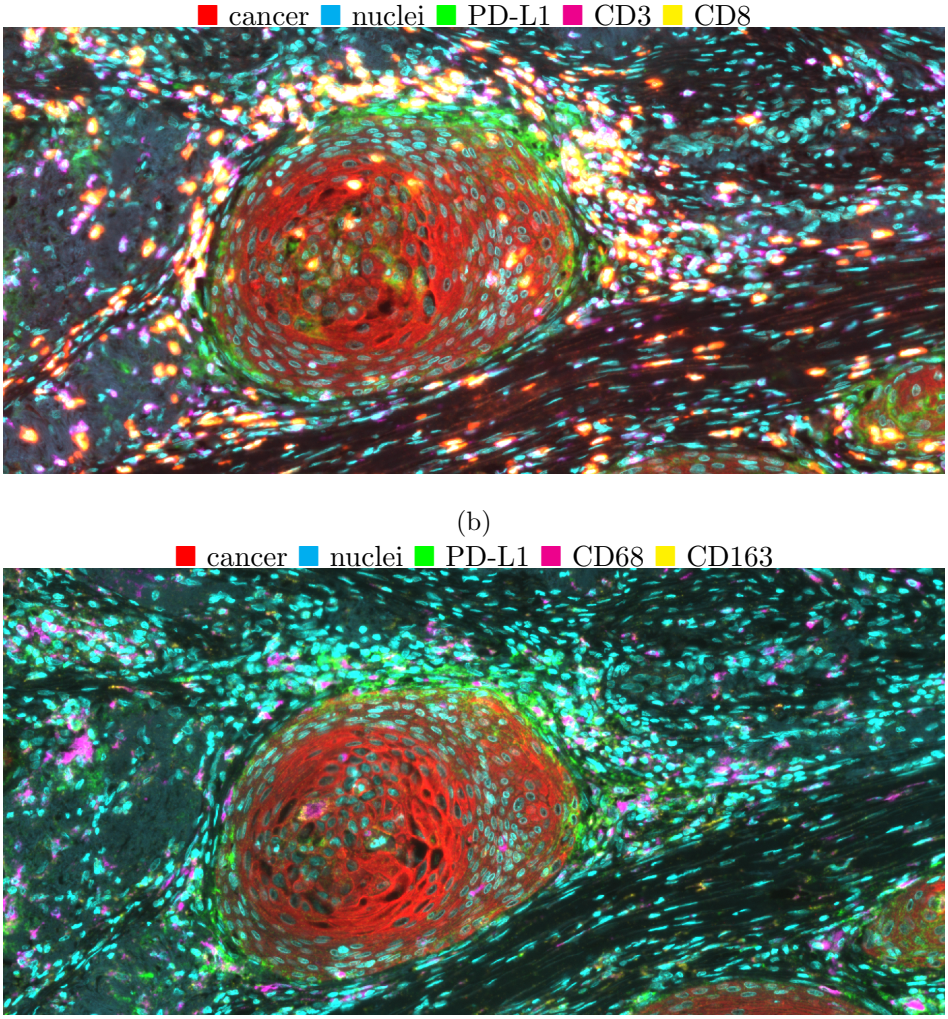
Table 4.1: Patient cohort characteristics.

4.2.2 Feature extraction – handcrafted features

A different member of the team implemented the processes that were needed for feature extraction from immunofluorescence WSIs. Hereunder, a brief overview is given. A more in–depth discussion is provided for the spatial feature extraction process since I was involved in its implementation (including the Ripley's K and L functions).

For each patient, immunofluorescence was performed on two serial $3\mu m$ thick sections of FFPE tissue sections. Primary antibodies against PanCK, CD3, CD8, CD68, CD163 and PD–L1 were used to label urothelial cells, general T-cells, cytotoxic T-cells, M1/M2 (total) macrophages, M2 macrophages and immune checkpoint ligand PD-L1, respectively. Nuclei were counter-stained with Hoechst. The multiplex immuno–labeled tissue slides were scanned at $20\times$ magnification and digitized into whole slide fluorescence images using a Carl Zeiss AxioScan.Z1 scanner (Zeiss, Göttingen, Germany). Examples are shown in Figure 4.1.

(a)



(b)

Figure 4.1: Examples of (a) TILs (*Nuclei: Cyan, Cancer: Red , PD-L1: Green, CD3: Purple, CD8: Yellow*) and (b) TAMs (*Nuclei: Cyan, Cancer: Red , PD-L1: Green, CD68: Purple, CD163: Yellow*) visualized using multiplexed immunofluorescence.

Detection of cell nuclei For nucleus detection, the methodology described by Brieu and Schmidt [32] was used. The methodology is comprised of four distinct stages: (a) training a classification RF on manually annotated foreground and background regions, defined as regions with or without cell nucleus respectively [31], (b) training a regression RF to generate proximity maps using coordinates from manual annotations of cell nuclei, (c) training a regression RF to generate surface area maps using manual annotations [32], and (d) localizing nuclei centers based on the proximity and surface area maps that were generated in (b) and (c) [32]. A proximity map encloses the distance to the closest nucleus center for each pixel of the input image. A surface area model provides a mask of the initial image wherein each pixel is either zero, if it is not a part of a nucleus, or a positive real number, if it is part of a nucleus. The positive real number is the area of the corresponding nucleus. Data augmentation with varying scale, rotation, as well as intensity for both PanCK and Hoechst IF channels was implemented in all stages.

Segmentation of epithelial cells for the identification of tumour buds For the quantification of tumours buds (TBs), segmentation of epithelial cells was required. The CNN–RF methodology described by Brieu et al. [29] and extended in the work of Brieu et al. [30] was adopted. Briefly, a convolutional neural network was trained on an annotated data set of epithelium and non-epithelium images. The Hoechst, PanCK, CD3 and CD8 IF channels were normalized following the approach described by Brieu et al. [29]. Once trained, the convolutional neural network produced a coarse segmentation mask of epithelium regions. The predicted epithelium probability layer was used together with the original immunofluorescence channels as input to a RF. Finer-grained segmentation masks were produced by the RF, enabling a more accurate segmentation of the epithelium. As detailed in previous work [30], the output of the CNN–RF is ensembled with the output of a semantic segmentation network [176] to generate the final epithelium segmentation results. Tumour buds (TBs) were classified as epithelium objects containing one to four nuclei [30].

Cell classification For cell classification, given normalized immunofluorescence channels (normalized as described by Brieu et al. [29]), each cell nuclei is defined as the center of a 11×11 pixels region and the mean normalized intensity of the region is computed for each immunofluorescence marker (CD3, CD8, CD68, CD163, PD-L1 and PanCK). Cells are classified as positive or negative for a given IF marker if the corresponding mean intensity is above or below a determined threshold (the threshold for all the IF markers was set to $32/256 = 0.125$ based on optimization on a small subset of the training set), respectively.

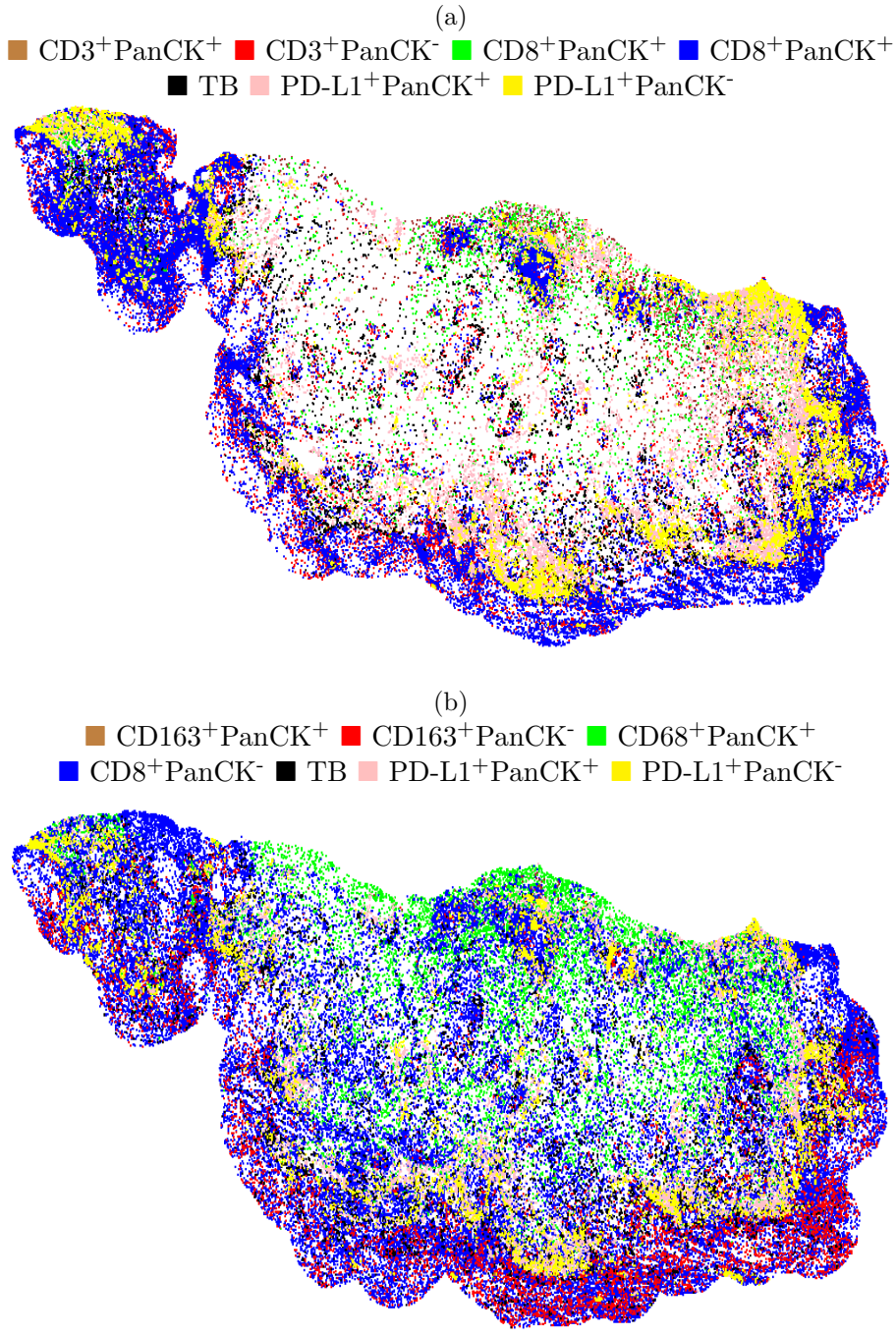


Figure 4.2: Nuclei localization of (a) TILs (*$CD3^+PanCK^+$: Brown, $CD3^+PanCK^-$: Red, $CD8^+PanCK^+$: Green, $CD8^+PanCK^-$: Blue, TB: black, $PD\text{-}L1^+PanCK^+$: Pink, $PD\text{-}L1^+PanCK^-$: Yellow*) and (b) TAMs (*$CD163^+PanCK^+$: Brown, $CD163^+PanCK^-$: Red, $CD68^+PanCK^+$: Green, $CD68^+PanCK^-$: Blue, TB: Black, $PD\text{-}L1^+PanCK^+$: Pink, $PD\text{-}L1^+PanCK^-$: Yellow*) across the WSI.

Fully automated feature extraction The entire FFPE tissue section of each MIBC patient was digitized into a WSI, encompassing both muscle-invasive urothelial carcinoma as well as adjacent benign tissue. Multiplex immunofluorescence enabled the detection of tumour infiltrating lymphocytes (TILs) (general CD3 and cytotoxic CD8 T-cells), tumour-associated macrophages (TAMs) (total CD68 macrophages and M2 CD163 macrophages), PD-L1$^+$ cells, cell nuclei (Hoechst), and epithelial cancer cells (PanCK) including TBs across the WSI of each patient. Machine learning-based image analysis (as described in the previous paragraphs) allowed for the exhaustive localization of each cell (across the core and invasive front), subsequently classified depending on its immunofluorescence signal as either a: 1) TB (panCK$^+$ i.e. marked as epithelial cancer cell), 2) M1 macrophage (CD163$^-$CD68$^+$), 3) M2 macrophage (CD163$^+$), 4) total macrophage (CD163$^+$CD68$^+$), 5) general T cell (CD3$^+$CD8$^-$), 6) cytotoxic T cell (CD3$^+$CD8$^+$), or 7) PD-L1$^+$ cell. Other than PD-L1, the rest of the categories are mutually exclusive to each other. Based on the above seven classes, a total of 186 quantitative features were extracted from the tumour core and invasive front of each WSI including the number and density of different cell types, the total size of tumour areas, as well as the pairwise spatial distributions between immune and cancer cells. The tumour core is defined as the main tumour mass and the invasive front as the border of the tumour core with a width of $1000\mu m$ ($500\mu m$ inside and $500\mu m$ outside of the border defining the invasive frontin and frontout, respectively) as shown in Figure 4.8. Feature extraction was performed using Definiens Tissue Phenomics® software (Definiens AG, Munich, Germany) [88, 25, 7]. A comprehensive list of the features is provided in the Appendix.

Spatial statistics The point coordinates of cell nuclei and immune checkpoint ligand PD-L1 expression were localized across the WSIs as shown in Figure 4.2. The Ripley's K function [174] was adopted for investigating how TBs, PD-L1, and the different populations of immune cells are distributed around each other. Given two populations X and Y, Ripley's K function estimates the density of Y within a circle of radius r around points X. As illustrated in Figure 4.3, assuming a Poisson distribution, the Ripley's K function can identify whether a population Y is dispersed, randomly distributed, or clustered around another population X. The K function is given as:

$$K_{xy}(r) = \frac{1}{\lambda_y}\mathbb{E}[\text{number of points y within a distance r around a point x}] \tag{4.1}$$

where $\mathbb{E}[\cdot]$ encloses all of the points of type y within a distance r of a randomly selected point of type x and λ_y is the number of points y per unit area in the region of interest. Theoretically, if the point pattern of points Y around X follows complete spatial randomness, also known as

a homogeneous Poisson process, the value of K function is πr^2. The L function [22] is a modification of equation 4.1, so that the expected output value is r, i.e. :

$$L_{xy}(r) = \sqrt{\frac{K_{xy}(r)}{\pi}} \tag{4.2}$$

This enables a more intuitive interpretation of the function's output value in relation to r. The L function was calculated for TIMs, TAMs, and PD-L1 surrounding TBs as well as PD-L1 surrounding TAMs and TILs for a series of increasing distances r where $r \in \{20, 50, 100, 150, 200, 250\}$ μm. While some approaches calculate the area under the curve of the L function against different r values [41], I provided the pairwise spatial distributions between PD-L1, TBs, and the immune populations directly to the ML classifiers as distinct features.

4.2.3 Binary survival analysis

As with the previous chapter, patient survivability was binarized based on a specific time cutoff. In particular, given the highly aggressive nature of MIBC, only the 5-year prognostic cutoff was investigated [58, 112, 153]. Patients that succumbed to MIBC within 5 years were denoted as patients with a bad prognosis whereas those that survived the 5 year cutoff were denoted as patients with a good prognosis. Inevitably, patients that died to an unrelated cause prior to the prognostic cutoff, i.e. they were part of the OTD–censored data (see the previous chapter for more information on censoring) had to be excluded (19% patient exclusion, i.e. from 96 to 78).

4.2.4 Model selection, algorithm selection, and performance evaluation

Both model selection and algorithm selection attempt to collectively maximize the predictive performance of the final ML model. However, ML algorithms are prone to overfitting, i.e. in finding and using patterns which arise from noise in the data. Such noisy patterns do not generally extend beyond the specific data set since noise is typically random. With both a small data set and a complicated model, the likelihood of overfitting increases. Testing the performance of a trained ML model on unseen data constitutes the mainstay in evaluating the generalizability of a ML model, and therefore, in identifying whether a model has overfitted. As such, a subset of the initial cohort was kept aside as the testing data set. In particular, using stratified random sampling, two subsets were created, the training set with 75% of the initial data (58 patients), and the testing set with 25% (20 patients). The testing set was only used at the performance evaluation stage to avoid introducing bias to the generalization performance estimate.

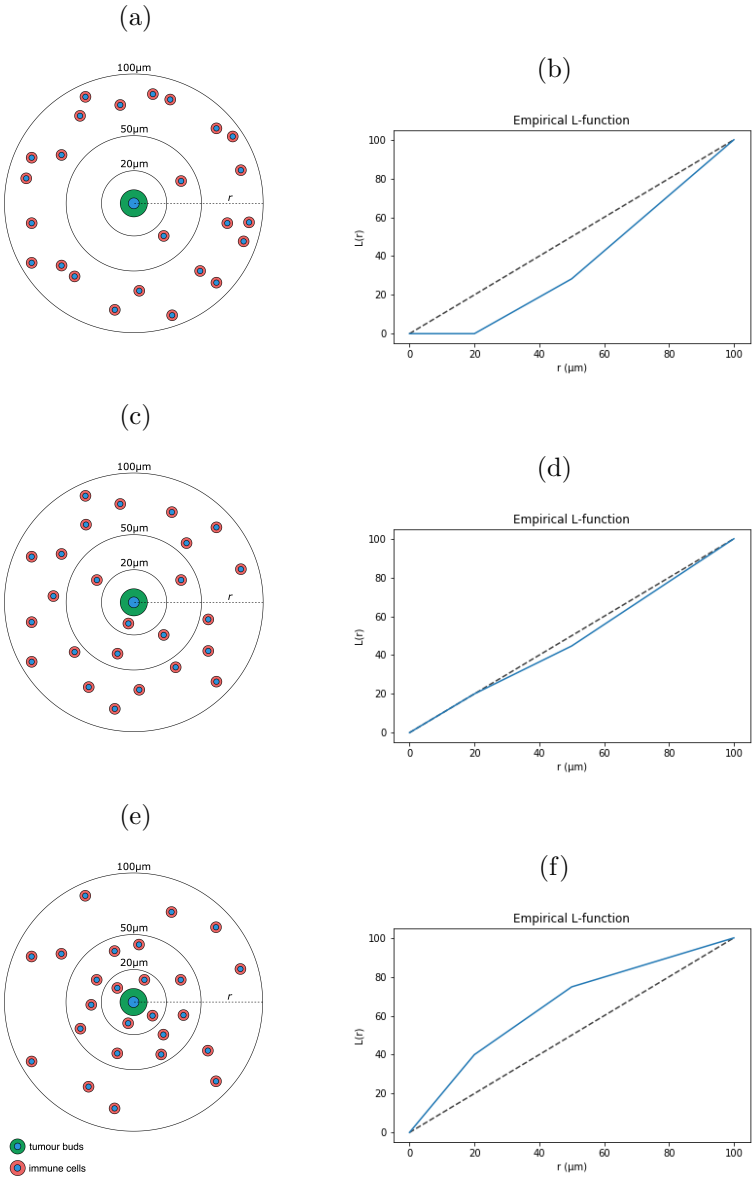


Figure 4.3: Schematic representation of different immune cell distributions from the nuclear centre of a tumour bud (a, c, e), and their corresponding L function values at different radii (b, d, f). The immune cell population is either (a–b) dispersed, (c–d) randomly distributed, or (d–e) clustered around the TB.

Classifier	Hyperparameter	Distribution	Values
LSVM	C	Log-uniform	[ln (1e−5) , ln (1e2)]
	Class weight	Categorical	[Balanced. None]
RSVM	C	Log-uniform	[ln (1e−5) , ln (1e2)]
	Gamma	Log–uniform	[ln (1e−3) , ln (1e3)]
	Class weight	Categorical	[Balanced. None]
LR	Type of pernalty	Categorical	[L1, L2, Elastic net, None]
	C	Log–uniform	[ln (1e−5) , ln (1e2)]
	L1 ratio	Uniform	[0, 1]
	Class weight	Categorical	[Balanced. None]
DT	Criterion	Categorical	[Gini, Entropy]
	Maximum features	Uniform integer	[1, max_features]
	Maximum depth	Categorical	[2, 3, 4, None]
	Class weight	Categorical	[Balanced. None]
RF	Number of trees	Log–uniform integer	[10, 1000]
	Criterion	Categorical	[Gini, Entropy]
	Maximum features	Categorical	[1, max_features]
	Maximum depth	Categorical	[2, 3, 4, None]
	Bootstrap	Categorical	[True, False]
	Class weight	Categorical	[Balanced, None]
KNN	K	Log–uniform integer	[1, 50]
	Metric	Categorical	[Balanced, None]
	P	Uniform integer	[1, 6]

Table 4.2: The search space of each classifier based on predefined distributions and densities over its hyperparameters.

Finally, for hyperparameter tuning, the same approach as the previous chapter is followed. In particular, 200 hyperparameter configurations were randomly sampled and evaluated for each machine learning algorithm using TPE. The distributions and densities used are shown in Table 4.2. Furthermore, each machine learning algorithm was first tuned using 5–fold cross validation and then compared against each other using 2–fold cross validation. This nested cross validation translates to optimizing the hyperparameters of each ML algorithm twice, and then measuring their performance on the corresponding evaluation folds (see Figure 4.4). Subsequently, the ML algorithm which performed better than the rest across two different training and validation folds is selected. It is important to highlight how in most cases each ML algorithm is evaluated based on two different hyperparameter configurations. Nevertheless, once the ML algorithm has been selected, yet another hyperparameter tuning phase is implemented to find an optimal hyperparameter configuration based on the whole training data set.

4.2.5 Baseline classifiers and performance assessment

Five machine learning algorithms with different theoretical underpinnings were selected to investigate whether the extracted features could predict 5 year survivability in MIBC patients; decision tree (DT), RF, SVM, LR, and KNN. The optimizing metric throughout experimentation was the area under the receiver operating characteristic (AUROC). At the final evaluation phase, classification accuracy, sensitivity, specificity, F1 score, and hazard ratios were also computed for ease of comparative analysis. It is important to highlight that in this work a true positive is defined as a correct prediction for a patient with bad prognosis, whereas a true negative is a correct prediction for a patient with good prognosis. In order to compute the aforementioned metrics, optimal threshold values were automatically selected at the final stage based on the training set performance. Hazard ratios and the associated confidence intervals were calculated using univariate Cox regression.

4.2.6 Stratified sampling

In order to avoid sampling subsets with different class distributions (classes are based on survival with a 5 year cutoff) than the original cohort, stratified sampling was used. Intuitively, when sampling from a data set with stratification, proportionally many patients from each class are sampled. As an example, given a data set with 75 patients of class C_1 and 25 patients of class C_2, a 20% sample would contain 15 patients from class C_1 and 5 patients from class C_2. In my case, the classes C_1, C_2 refer to patients with good and bad prognosis.

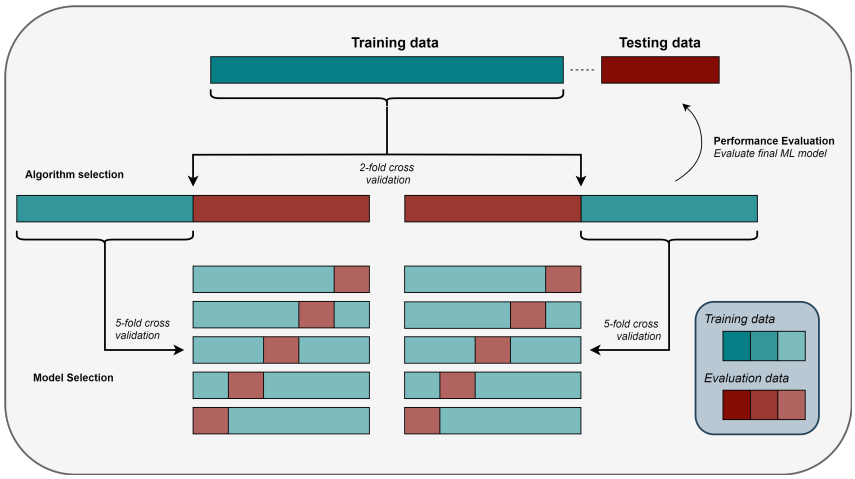


Figure 4.4: Pictorial representation of nested cross validation with an independent testing set. (A) Performance Evaluation: The best ML algorithm (selected by the outer cross validation, see *B*) was trained on the training data set and subsequently evaluated on the testing data set. (B) Algorithm Selection: Each ML algorithm (with hyperparameters tuned based on the inner cross validation, see *C*) was trained and tested on the corresponding training and evaluation folds respectively. The best ML algorithm was selected based on the average performance of both evaluation folds. (C) Model Selection: ML models with randomly sampled hyperparameter configurations were trained and tested based on a fivefold cross validation. The best hyperparameter configuration for each ML algorithm was selected based on the performance on all five evaluation folds.

4.2.7 Feature space and feature selection

In order to capture multiple aspects of the disease, features from both clinical reports and whole slide immunofluorescence images were quantified. Herein, the number and density of PD-L1 positive and negative immune cell populations, as well as of TBs from the WSIs are labelled as "image features". The pairwise spatial distributions between immune cells and TBs are termed "spatial features". And lastly, clinicopathological features such as age, gender, and TNM stage are termed "clinical features". Altogether 201 features were quantified – 126 image, 60 spatial, and 15 clinical features. To investigate whether smaller feature spaces result in better ML models, I ran the same ML workflow over different feature sets. In particular, my experiments were based on the following 7 feature sets: (i) image, (ii) spatial, (iii) clinical, (iv) image and spatial, (v) image and clinical, (vi) spatial and clinical, (vii) image, spatial, and clinical.

4.3 Results

	Training set		Testing set	
Evaluation metrics	**Ensemble**	**TNM**	**Ensemble**	**TNM**
AUROC	**98.3**	71.6	**89.3**	64.3
Accuracy	**94.8**	65.5	**80.0**	50.0
Specificity	**89.5**	**89.5**	**100.0**	**100.0**
Sensitivity	**97.4**	53.8	**71.4**	28.6
F1 score	**96.2**	67.7	**83.3**	44.0
Hazard ratio	**45.9** (**6.2**, **341.1**)*	4.4 (2.3, 8.6)*	**32.5** (**3.9**, **270.3**)*	3.3 (1.0, 11.0)*

Table 4.3: Comparison between the ensemble model and TNM staging. * 95% Confidence Interval.

		Prognosis				Prognosis		
		Bad	Good	Total		Bad	Good	Total
Ensemble	Bad	38	2	40	Bad	10	0	10
	Good	1	17	18	Good	4	6	10
	Total	39	19	58	Total	14	6	20
		Bad	Good	Total		Bad	Good	Total
TNM	Bad	21	2	23	Bad	4	0	4
	Good	18	17	35	Good	10	6	16
	Total	39	19	58	Total	14	6	20

Table 4.4: The confusion matrices produced by the proposed ensemble model and TNM staging on the training set (left) and testing set (right).

F	**LR**	**KNN**	**LSVM**	**RSVM**	**DT**	**RF**
I	69.8 ± 13.3	61.8 ± 5.7	$\mathbf{** 72.8 \pm 0.3}$	60.8 ± 8.2	63.3 ± 3.3	57.8 ± 3.2
C	$\mathbf{58.1 \pm 0.1}$	59.8 ± 7.2	55.9 ± 5.9	48.6 ± 5.6	50.1 ± 2.4	56.6 ± 9.1
S	46.8 ± 3.3	46.2 ± 4.7	42.7 ± 1.8	38.0 ± 0.0	$\mathbf{49.2 \pm 3.3}$	43.7 ± 13.0
{I, C}	62.0 ± 0.6	55.3 ± 4.4	56.5 ± 8.5	50.1 ± 11.5	$\mathbf{** 68.8 \pm 0.8}$	70.4 ± 7.9
{I, S}	$\mathbf{** 70.2 \pm 14.7}$	59.9 ± 2.4	49.0 ± 2.5	58.6 ± 11.9	51.0 ± 1.5	64.1 ± 7.6
{C, S}	44.3 ± 9.8	49.4 ± 3.2	$\mathbf{57.9 \pm 2.4}$	47.1 ± 4.4	44.8 ± 2.7	46.6 ± 15.4
{I, C, S}	56.3 ± 0.2	55.0 ± 4.0	60.2 ± 8.2	61.4 ± 4.6	66.6 ± 9.1	$\mathbf{** 67.3 \pm 5.8}$

Table 4.5: Results for algorithm selection from the nested cross validation on the training set with AUROC as the performance metric. For each feature space, the best ML classifier is indicated in bold. Amongst the best classifiers of each feature space (in bold), my ensemble model uses those with a marked difference in performance (**). *F*, I, C, and S are abbreviations for feature space, set of image, clinical, and spatial feature, respectively.

4.3.1 Proposed ensemble model

For each of the tested feature sets, the classifier with the highest average AUROC on nested cross–validation was selected. In case of similar average AUROC between two ML classifiers using the same feature set, I selected the one exhibiting the least variance. The results are shown in Table 4.5. Since multiple classifiers exhibited competitive performance, yet used different underlying feature sets, instead of employing a single classifier, I combined the best ones into an ensemble model. In particular, my ensemble model consists of a linear support vector machine (LSVM) that uses image features (72.8 ± 0.3 AUROC), a DT that uses image and clinical features (68.8 ± 0.8 AUROC), a LR that uses image and spatial features (70.2 ± 14.7 AUROC), and a RF that uses all features (67.3 ± 5.8 AUROC). Following hyperparameter tuning for each one of the selected classifiers on the whole training set, without cross validation, the ensemble model was evaluated on the independent testing set achieving 89.3% AUROC and a highly significant separation of patients into low and high risk groups (p value$= 7e-06$). Patients were classified as having a bad prognosis by the ensemble model if two or more of the submodels predicted a bad prognosis.

On a sidenote, although reversing the predictions of the RSVM model with spatial features, which achieved the worst results (38.0 ± 0 AUROC) across all feature combinations and algorithms, would have been an interesting experiment to pursue, further validation would be needed to ensure that I hadn't in the meantime introduce bias in model selection.

4.3.2 Pessimistic bias

The large difference between the generalization estimates of algorithm selection and performance evaluation (see Table 4.5 and Table 4.3) can be mostly attributed to pessimistic bias [171]. Given an already small data set, withholding half of the training data set for evaluation, due to twofold cross validation, increases the chance that a ML model will underfit, i.e. its maximum representation capacity will not be reached [171]. Therefore, the generalization estimate from performance evaluation (Table 4.3) is more reliable since the whole training set was used.

4.3.3 Comparing against TNM staging

In order to compare against the gold standard in clinical practice, TNM, patients had to be stratified into low and high risk groups. Based on a pairwise log-rank test comparison in the training data set, stage II and III patients were considered as the low risk group whereas stage IV patients were considered as the high risk group. The Kaplan-Meier and ROC curves of TNM staging and the proposed ensemble model on the testing set are shown in Figure 4.5. To allow further comparative analysis, Kaplan-Meier curves of other clinicopathological features, such as age (threshold optimized in training set) and gender, are shown in Figure 4.6. In addition,

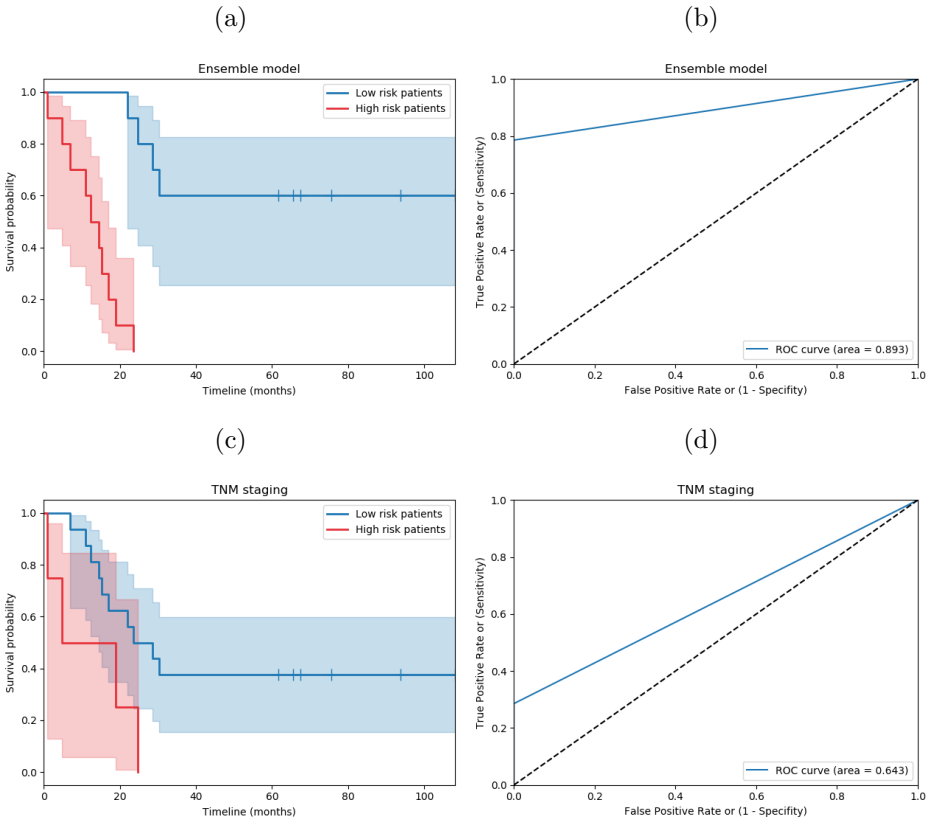


Figure 4.5: Kaplan-Meier and ROC curves on the testing set for the ensemble model and TNM. Separation was significant based on the (a) ensemble model (p value $= 7e-06$, $N_{LowRisk} = 10$ & $N_{HighRisk} = 10$) and (c) TNM (p value $= 0.04$, $N_{LowRisk} = 16$ & $N_{HighRisk} = 4$).

the Kaplan-Meier and ROC curves of each submodel of the ensemble model are shown in Figure 4.7.

4.3.4 Post-hoc analysis of features

For each classifier of the ensemble model, post-hoc analysis was conducted to reveal the features guiding survivability prediction. The feature considered at each node of a DT is readily interpretable. For the LR, its coefficients determine the importance as well as the positive or negative effect of each feature on patient prognosis. Mean decrease in Gini index was calculated for each feature of the RF based on the underlying decision trees [134]. Finally, since the selected SVM had a linear kernel, feature ranking coefficients were readily available [87]. A threshold was set to filter out features with low feature importance. In particular, the threshold was set to two times the mean importance of all features for the DT, LR, and RF, whereas two times the median importance was used for the LSVM.

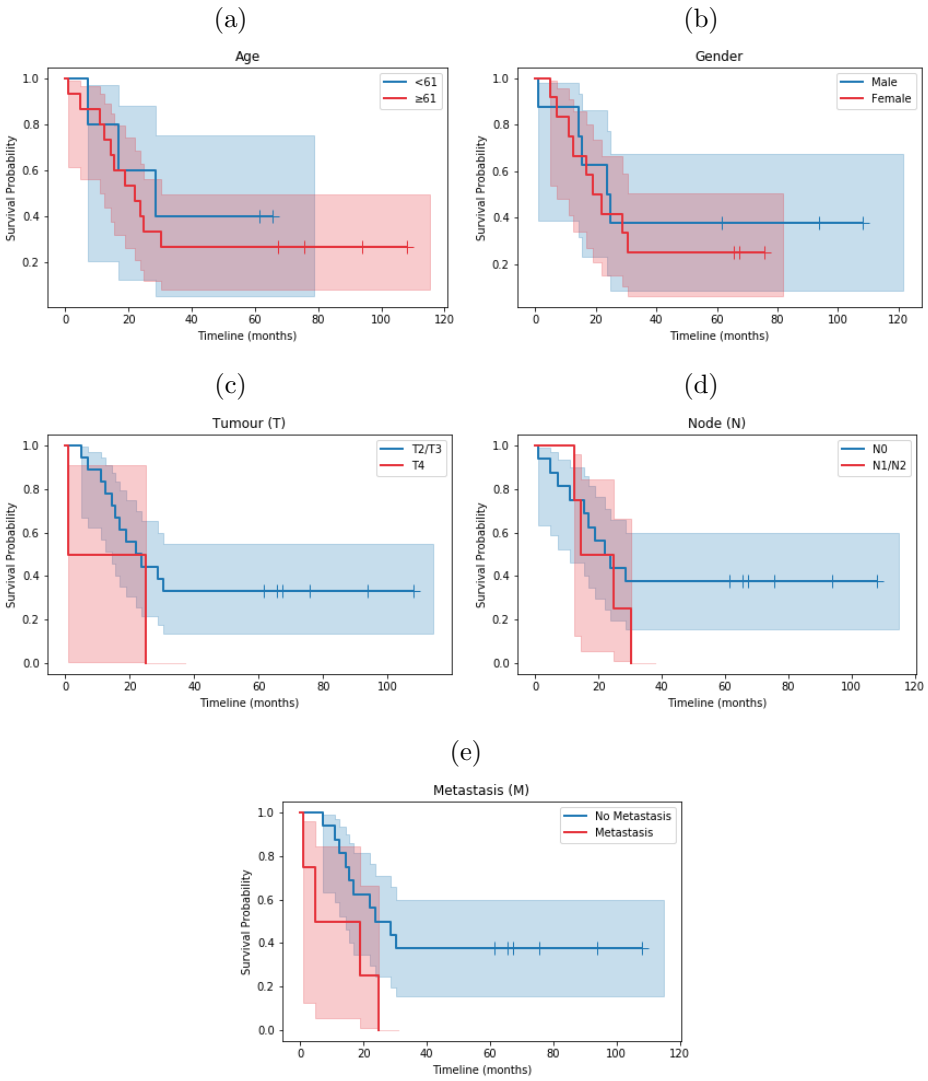
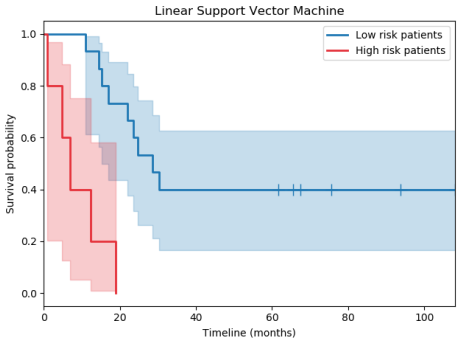
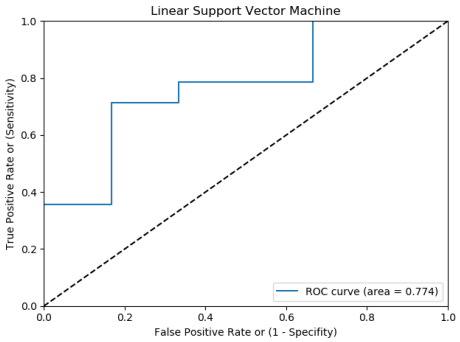


Figure 4.6: Kaplan-Meier curves of (a) Age (p value = 0.57, $N_{<61} = 5$ & $N_{\geq 61} = 15$), (b) Gender (p value = 0.62, $N_{Female} = 12$ & $N_{Male} = 8$), (c) Tumour (p value = 0.25, $N_{T2/T3} = 10$ & $N_{T4} = 2$), (d) Node (p value = 0.36, $N_{N0} = 16$ & $N_{N1/N2} = 4$), and (e) Metastasis (p value = 0.04, $N_{No} = 16$ & $N_{Yes} = 4$).

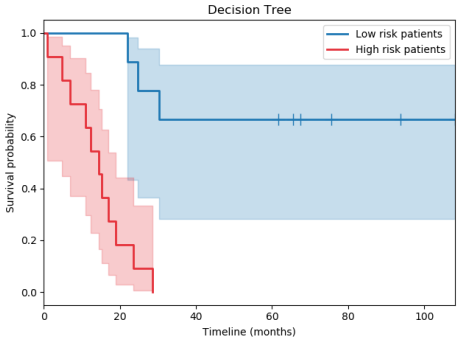
(a)


Linear Support Vector Machine
Low risk patients
High risk patients
Survival probability
Timeline (months)

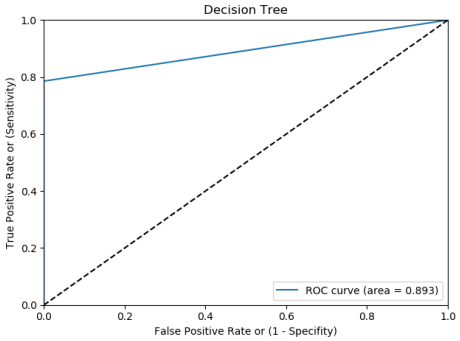
(b)


Linear Support Vector Machine
True Positive Rate or (Sensitivity)
False Positive Rate or (1 - Specifity)
ROC curve (area = 0.774)

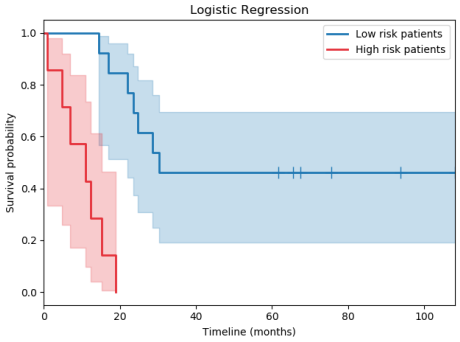
(c)


Decision Tree
Low risk patients
High risk patients
Survival probability
Timeline (months)

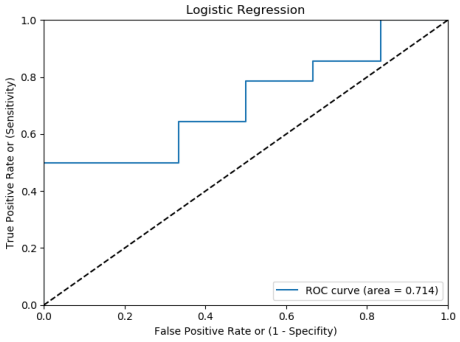
(d)


Decision Tree
True Positive Rate or (Sensitivity)
False Positive Rate or (1 - Specifity)
ROC curve (area = 0.893)

(e)


Logistic Regression
Low risk patients
High risk patients
Survival probability
Timeline (months)

(f)


Logistic Regression
True Positive Rate or (Sensitivity)
False Positive Rate or (1 - Specifity)
ROC curve (area = 0.714)

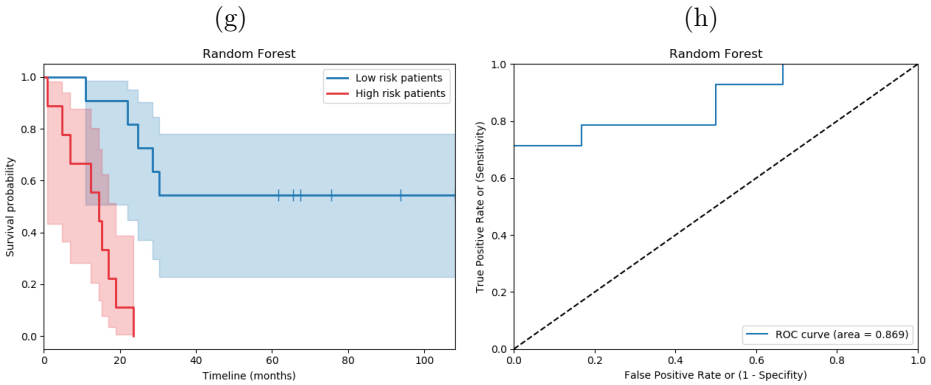


Figure 4.7: Kaplan-Meier and ROC curves on the testing set for each of the ML classifiers used in the ensemble model. (a–b) Image features (p value $= 6e-05$, $N_{LowRisk} = 15$ & $N_{HighRisk} = 5$), (c–d) image and clinical features (p value $= 1e-04$, $N_{LowRisk} = 9$ & $N_{HighRisk} = 11$), (e–f) image and spatial features (p value $= 5e-05$, $N_{LowRisk} = 13$ & $N_{HighRisk} = 7$), (g–h) image, clinical, and spatial features (p value $= 8e-06$, $N_{LowRisk} = 11$ & $N_{HighRisk} = 9$).

There were 8, 10, 25, and 16 important features for DT, LR, RF, LSVM, respectively which are listed in Tables 4.6 and 4.7.

Classifiers	**Bad Prognosis**	**Good Prognosis**
LR	Density of TB frontin/core	Density of $CD8^+$ frontin/frontout/core Density of $CD3^+$ frontout/core Density of $CD68^+$ frontin/frontout L(TB, $CD3^+$, 20)
LSVM	Density of TB frontin/core	Density of $CD8^+$ frontin/frontout/core Density of $CD3^+$ frontin/frontout/core Density of $CD68^+$ frontin/frontout/core Number of $CD3^+$ frontin Number of $PDL1^+CD163^-CD68^+$ frontout Number of $PDL1^+CD163^+CD68^+$ frontout

Table 4.6: The features that contribute to a good and bad prognosis according to the LR and the LSVM. L(x,y,r): the L function value of y in respect to x for distance r.

For the LR and LSVM submodels, high density of TBs in both the invasive frontin and tumour core is highlighted as an indicator of bad prognosis. On the contrary, high density of $CD8^+$, $CD3^+$ and $CD68^+$ cells is consistently identified as a marker of good prognosis. In addition, high number of $CD3^+$, $CD68^+PD\text{-}L1^+$ and $CD163^+PD\text{-}L1^+$ cells in the invasive front as well as the presence of $CD3^+$ cells within a distance of $20\mu m$ from TBs

Classifiers	**Important Features**
DT	Number of PD-L1$^+$ frontout
	Density of CD163$^+$ frontout
	Density of PD-L1$^+$CK$^+$ core
	Number of TB frontout
	Density of CD3$^+$ frontout
	Number of CD68$^+$ frontout
	Density of CD68$^+$ frontin/frontout
RF	Number of CD68$^+$ frontout
	Density of CD68$^+$ frontin/frontout/core
	Number of CD3$^+$ core
	Density of CD3$^+$ frontout
	Number of CD8$^+$ frontout/core
	Density of CD8$^+$ frontout/core
	Number of TB core
	Density of PD-L1$^+$ frontout
	Density of PD-L1$^+$CK$^+$ core
	Density of PD-L1$^+$CK$^-$ frontin/frontout
	Number of CD163$^+$CD68$^+$ frontout
	Number of NucleiCK$^+$ frontin
	TNM IIIA
	TNM IV
	L(TB, CD3$^+$, 20)
	L(TB, CD8$^+$, 20)
	L(TB, PD-L1$^+$, 20)
	L(TB, CD8$^+$, 50)
	L(CD163$^+$, PD-L1$^+$, 150)

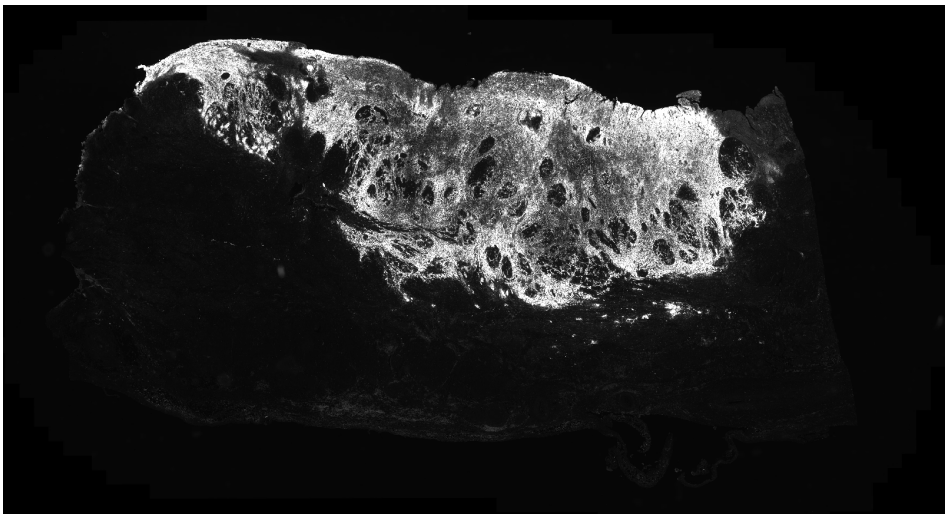
Table 4.7: The most important features for estimating patient prognosis by the DT and the RF. L(x,y,r): the L function value of y in respect to x for distance r.

are associated with good prognosis.

For the DT submodel, low density of CD68$^+$, high PD-L1$^+$ expression, and high number of TBs (all in frontout) lead to bad prognosis, whereas, given a low density of CD68$^+$ and PD-L1$^+$ expression in frontout, prognosis depends on the number of CD68$^+$ in frontin. Finally, the majority of the patients with good prognosis had high CD68$^+$ in frontout, nonzero PD-L1$^+$ expression in core, low CD163$^+$ in frontout, and high CD3$^+$ in frontout.

Similar to the previous submodels, TBs and CD68$^+$ cells were the most important predictors of 5 year prognosis for RF. In addition, RF employed more spatial features than any of the other submodels including but not limited to PD-L1$^+$ expression within a distance of $20\mu m$ from TBs and $150\mu m$ from M2 macrophages as well as the presence of CD3$^+$ and CD8$^+$ cells within a distance range of $20 - 50\mu m$ from TB.

(a)



(b)

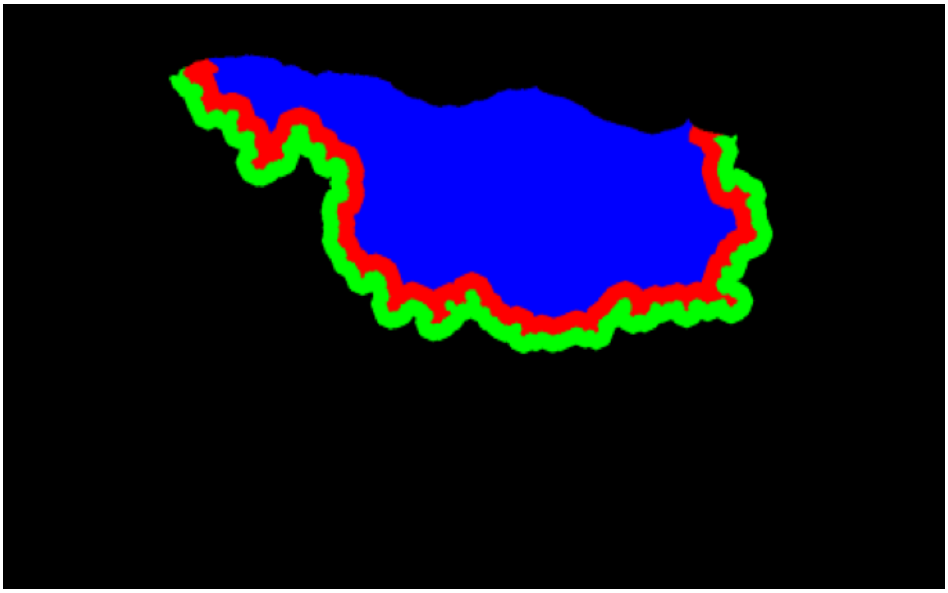


Figure 4.8: (a) Whole slide immunofluorescence image based on the PanCK channel, (b) Segmentation of the corresponding tissue (a) into tumour core (*blue*), invasive frontin (*red*) and frontout (*green*) using the PanCK channel.

4.4 Discussion

In the last decade, advances in the rapidly growing field of tumour immunology have provided further insights into the dynamic nature of the multifaceted immune response throughout the various stages of cancer initiation, evasion, and progression. Concomitantly, multiple research groups have successfully leveraged this new knowledge to improve cancer prognosis, thereby providing evidence for the clinical relevance of immuno-oncology [156]. Accurate patient prognosis is crucial for improving the survival rates of cancer patients since it is a prerequisite to delivering the most effective treatment for each patient. In fact, multiple papers have shown that the quantitative characterization of the tumour-immune microenvironment components, including TILs, TAMs, and immune checkpoints, can yield information of prognostic relevance [83, 214, 111]. Particularly, tumour cells surrounded by a large number of prominent intra-tumoural and peri-tumoural TILs and M1 macrophages have been related to better prognosis in several types of cancer [144], whereas a high content of M2 macrophages and TBs has been associated with poorer outcome [214]. In addition, related research has reported the significance of PD-L1 expression on tumour tissues as an independent poor prognostic factor [213]. In this paper, I have investigated for the first time the prognostic relevance of immune system biomarkers, TILs, TAMs, TBs, and PD-L1, across whole slide immunofluorescence images of MIBC patients.

H&E is still the most important and commonly used histochemical staining method for studying and diagnosing tissue diseases in histopathology. However, imaging of H&E stained FFPE tissue has limitations including the inability to quantify the complex cellular states as well as to identify distinct cell populations in the tumour-immune microenvironment. With the advent of whole slide imaging and the increasing adoption of digital pathology in the clinic [34], multiplex methodologies have the potential to provide significantly more information about the underlying tumour-immune microenvironment than single-marker (i.e. single label immunohistochemistry) and conventional histochemical staining based methodologies [157]. The development of single protein-based biomarkers to explain patient-level behaviour is hindered by the vast signalling network mediating the heterotypic cell-cell crosstalk between cancer, stromal and immune cells. Instead, with multiplexed methodologies, various proteins can be simultaneously captured on a single tissue sample, encapsulating the tumour-immune architecture from the cellular level down to the subcellular, ultimately providing more information about the microenvironment.

In the proposed approach, multiplexed immunofluorescence was used to visualize TBs, general and cytotoxic T-cells, M1, M2, and total macrophages, and their co-expression of immune checkpoint ligand PD-L1 in order to quantify their numbers and densities, as well as their pairwise spatial distributions across defined areas (tumour core, invasive frontin and frontout) within a WSI. Over the last decade, multiple studies have investigated the

topographical distribution of the immune cells within the tumour microenvironment [110, 119]. It is known that tumour-infiltrating immune cells are scattered in the tumour core and the invasive front whereas their density in each tumour region is correlated with patient outcome [75]. Furthermore, the analysis of multiple tumour regions (tumour core and invasive front) was shown to improve the prediction accuracy of patient survival compared to single-region analysis [76, 75]. In addition, Immunoscore, a classification system based on the quantification of two lymphocyte populations (CD3 and CD8) within the tumour core and the invasive front of tumour, has been shown to have a prognostic significance superior to that of the TNM staging system in patients with colorectal carcinoma [77, 76]. The image, spatial, and clinical features contain a multitude of information about the state of the disease and collectively portray a more holistic view of each patient pathophysiology. I hypothesized that these features can predict the aggressiveness of MIBC, and therefore suggest whether a patient should be considered low or high risk of disease-specific death.

ML contains a plethora of classifiers that have been employed with success in multiple instances, including diagnosis, segmentation, prognosis, and even therapy planning [140]. In the proposed methodology, survival analysis is turned into a binary classification problem, thus enabling traditional ML algorithms and workflows to be readily employable. In addition, to counter the possibility of overfitting due to having a small data set yet high dimensional feature space, nested cross validation with a separate testing set was adopted. The proposed ensemble model significantly surpasses under all metrics – AUROC, Accuracy, Specificity, F1 score, Hazard ratio – (89.3%/80.0%/71.4%/83.3%/32.5) the gold standard, TNM staging (64.3%/50.0%/28.6%/44.0%/3.3), as summarized in Table 4.3. The confusion matrices are shown in Table 4.4. The ensemble model consists of a LSVM that uses image features, a DT that uses image and clinical features, a LR that uses image and spatial features, and a RF that uses all features. It is interesting to note that the DT performed almost on par with the ensemble model on the testing set. It only underperformed on the log rank test. It even achieved a higher accuracy than the ensemble model and correctly classified one more instance of bad prognosis than the ensemble model. However, the correctly classified instance became the last to succumb in the high risk group (the effect can be observed by comparing the last patient to succumb to the disease within the high risk group between DT in Figure 4.7 and the ensemble model in Figure 4.5), extending the confidence intervals of the high risk group, and thus decreasing the separation of the two groups. Further validation would be needed in order to test the hypothesis that the DT and the ensemble model perform equally (since this was observed on the testing set, rather than the validation set), and that perhaps the spatial features as well as the complexity of an ensemble model are both unnecessary.

The results of this work suggest that the characterization of a broad

immune cell population enables a better estimation of survival compared to TNM staging system in MIBC patients which in turn provides further biological insights. Most of the findings based on whole slide immunofluorescence images are novel for MIBC, and also corroborate with existing literature on other types of cancer [84, 173, 111, 9, 45, 151]. In particular, I found that high content of TBs in the invasive frontin, frontout, and tumour core as well as low number of $CD68^+$ cells and high PD-L1 expression in the invasive frontout are indicators of bad prognosis [84, 30, 173, 111]. High density of $CD8^+$, $CD3^+$, and $CD68^+$ cells in the invasive frontin, frontout, and tumour core was associated with good prognosis by the proposed models [9, 45]. In addition, high number of $CD3^+$ and $CD68^+$ cells as well as high number of $CD3^+$ cells clustered within a distance of $20\mu m$ from TBs were linked to good prognosis [151]. Finally, I found that high density of $CD163^+$ cells without PD-L1 expression in frontout is associated with bad prognosis by the DT submodel, whereas the LSVM submodel employed high number of $CD163^+$PD-$L1^+$ cells in frontout as an indication of good prognosis. Number and density, although often found to be highly correlated, capture different types of information, and often turn out to be both beneficial with an additive benefit when both included. However, density is fundamentally a more standardised type of feature to measure since it generalizes over variable size tissue by definition, i.e. measured as pixels of object of interest divided by total number of pixels of the tissue image (or region of interest). Whether density is more likely to generalize to new data sets (perhaps ones created in different institutions), or whether the exclusion of number as a type of feature could result in more robust classifiers, remains to be explored in future work as more empirical data would be needed.

In summary, I have demonstrated that ML classifiers using image and spatial features from WSIs combined with clinical features from medical records can separate MIBC patients into low and high risk groups for 5 year prognosis. The present approach outperforms the current clinical staging system, TNM, reinforcing the importance of standardized quantification of immunological features across WSIs, as well as the adoption of ML into the clinic. Moreover, my findings show that investigating features from the tumour-immune microenviroment in relation to survival can provide further insights for histopathological studies, thereby contributing to better ways for predicting survivability and enabling better quality of care.

4.5 Implementation details

The machine learning framework was implemented using the following packages in Python: Pandas [145], Numpy [89], scikit-learn [159], matplotlib [97], SciPy [200], hyperopt [18], and lifelines [54]. The spatial feature extraction process was implemented in R. Code sections of the methodology can be found in Appendix A.

Chapter 5

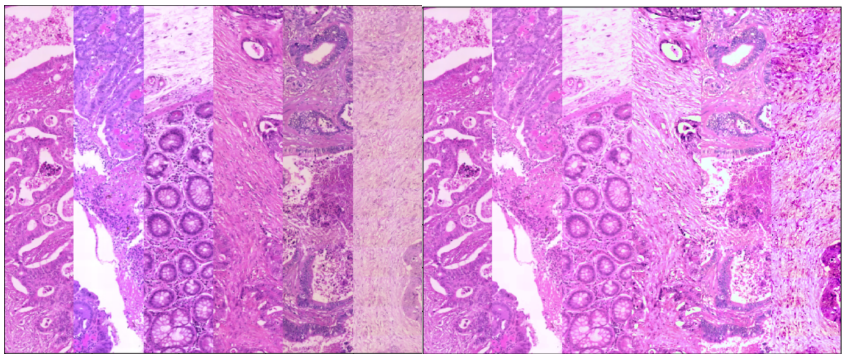
Stage I and II colorectal cancer prognosis - billions of pixels

5.1 Problem formulation

In this chapter, I consider the problem of CRC prognosis and propose a deep learning based method, trained on unannotated whole slide histopathology images, that predicts the disease–specific survivability of CRC patients. The whole slide histopathological images survival analysis (WSISA) framework, introduced by Zhu et al. [223], was one of the earliest and most promising WSI–based methods for high–level clinical task optimization with low–granularity labels. The framework described herein extends WSISA in the following ways:

- A more sophisticated approach to reducing the dimensionality of image patches prior to clustering is introduced that is based on (i) colour heterogeneity and (ii) morphological heterogeneity.
- An effective way to aggregate patch–level predictions to slide-level to patient-level (*two-step* aggregation) using machine learning techniques is proposed.

One of the principal contributions of this work is a deep learning–based system that is able to extract and learn salient, discriminative, and clinically meaningful content from a real–world data set (CRC WSIs) with low–granularity labels (patient prognosis). A subset of stage I and II CRC patients succumbs to the disease within 5 years of the diagnosis (20% of stage II and 9% of stage I). By identifying those patients that are at higher risk, treatment options, normally given to patients diagnosed with a higher stage, can be considered. Note that in this chapter I extend the methodology presented in my published work [217], and conduct further experiments on a larger cohort of patients (from 34 to 110) and WSIs (from 34 to 209).



(a) (b)

Figure 5.1: Chromatic normalization examples (left & right: original & normalized tiled strips).

5.2 Methods

5.2.1 Cohort

In this work, I utilize digitized whole slide images of archived diagnostic histopathological tissue sections stained with haematoxylin and eosin (H&E); see examples in Figure 5.1.

The WSIs are from patients operated in NHS Lothian hospitals between the years of 2002 and 2007. In particular, my data set comprises WSIs of tissue sections from archived FFPE tissue blocks of CRC stage I and stage II patients who underwent surgical resection. For each patient, there is either one or (typically) two WSIs associated with them. When there is only one WSI available, it is always of tumour tissue. However, when two WSIs are available, one of the two may contain normal tissue (more details were not available). This work was conducted in accordance with the declaration of Helsinki and no patient identifiable information was provided to the researchers. Ethical approval was obtained after review by the NHS Lothian NRS BioResource, REC approved Research Tissue Bank (REC approval ref: 15/ES/0094), granted by East of Scotland Research Ethics Service.

Apart from the WSIs, each patient data sample is accompanied by follow-up information, including date of death, and whether this patient dies of CRC. All CRC patients were either stage I or stage II at the time of surgery. For almost half the patients, there is no other clinical information available. For a subset of the patients (114 out of the initial 223) for which incomplete clinical records were available, about 75% of the patients were stage I, and 25% were stage II. Beyond the stage, however, no other information (representative of the final data set) could be extracted from the records. The original slides were stained using H&E at the time of treatment, and were scanned using a ZEISS Axio scan Z1 (Zeiss, Oberkochen,

DE) whole slide scanner with a 40× objective. The scale of a single pixel represents 0.111μm × 0.111μm of the actual size. The digital camera used was a Hitachi HVF2025SCL with an exposure time of 200μs.

The entire data set contains 223 patients with a total of 343 WSIs. Unfortunately, the files of many of these WSIs were corrupted (85 patients had to be removed). The smallest image is 1GB and the largest one is 13GB, with an average size of the WSIs being 8GB, which is approximately 300,000 pixels by 200,000 pixels. The bit depth is 24 with 3 channels. The original data was in the CZI format. To handle WSIs in the CZI format, the Python libraries bioformats [129] and javabridge (available at https://github.com/LeeKamentsky/python-javabridge) were used.

5.2.2 Binary survival analysis

Similar to the previous chapters, patient survivability was binarized based on a 5-year prognostic cutoff. Patients that succumbed to CRC within 5 years were denoted as patients with a bad prognosis whereas those that survived the 5 year cutoff were denoted as patients with a good prognosis. Inevitably, patients that died to an unrelated cause prior to the prognostic cutoff, i.e. they were part of the OTD–censored data, had to be excluded (28 patients removed).

5.2.3 Data preparation

Following the binarization of prognosis, and removal of corrupted WSI files, the portion of the data set that was usable for this work contained 110 patients with a total of 209 WSIs. A subset of the patients was kept aside as the testing data set. In particular, using stratified random sampling, two subsets were created, the training set with 75% of the initial data (82 patients, 154 WSIs), and the testing set with 25% (28 patients, 55 WSIs). There were 51 patients (94 WSIs) with good prognosis, and 31 patients (60 WSIs) with bad prognosis in the training set. In the testing set, there were 17 patients (33 WSIs) with good prognosis, and 11 patients (22 WSIs) with bad prognosis.

Chromatic normalization Histopathological tissue sections or WSIs are often examined individually by pathologists, who mainly focus on relative colour and pattern differences within a single tissue section. It is rare to compare directly different slides in order to make a diagnosis; each slide is examined to identify particular spatial or pattern characteristics. However, in the application of quantitative analysis and medical statistics for diagnosis and prognosis, different overall absolute colour value can have a detrimental effect especially when the slide count is low (as in my case).

The variation in terms of colour distribution is ultimately the difference in the amount of light absorbed; Figure 5.1 shows examples of slides from

different patients. As can be seen, the colour profile exhibits great inter–patient variability. While it is true that the use of greyscale would address this problem, it also effects a loss of valuable histopathological information. I instead apply the Reinhard normalization; a colour normalization method with low computational overhead and memory requirements [172] that, despite its simplicity, has been show to yield good results in histopathology images [155]. The implementation of Reinhard normilization was adopted from https://github.com/Peter554/StainTools. Code is provided in Appendix C.

Patch extraction Patches are extracted after the entire WSI has been down-sampled and normalized. In particular, tiles of size 224×224 pixels are extracted from the 1/16 resolution image (n.b. $40\times$ magnification level was used in acquisition at 0.111μm/pixel). To construct an end-to-end pipeline, I approach the problem of relevant patch selection through the use of fully automatic clustering which does not assume or require application of human prior knowledge.

Data augmentation To prevent the model from over–fitting, I apply heavy data augmentation during training. This design choice was inspired by the recent work of Pohjonen et al. [163], and others [65], who proposed a heavy augmentation strategy as a way of improving the generalization performance of neural networks against distributional shifts (e.g. stainings with different colour profiles). Specifically, the transformations used in this work can be categorised into three clusters:

- Trivial augmentations that do not alter the content of the image, e.g. 90° rotations, vertical and horizontal reflections.
- Non-trivial augmentations that alter the content of the image, e.g. Gaussian blurring, elastic deformation, etc.
- Colour augmentations, e.g. randomly changing the brightness, contrast, etc. of the image.

During training, a transformation from each one of the clusters is randomly sampled and used to augment each image. Examples for all transformations in isolation and combined are shown in Figure 5.2. The code at https://github.com/gatsby2016/Augmentation-PyTorch-Transforms was used for the implementation of the transformations in PyTorch. The parameters of each transformation are included in Appendix C.

5.2.4 Patch clustering

A major problem for patch–level based classification approaches is that there is no ground truth label for each individual patch. In order to overcome this issue, I broadly consider a patch either to be (sufficiently)

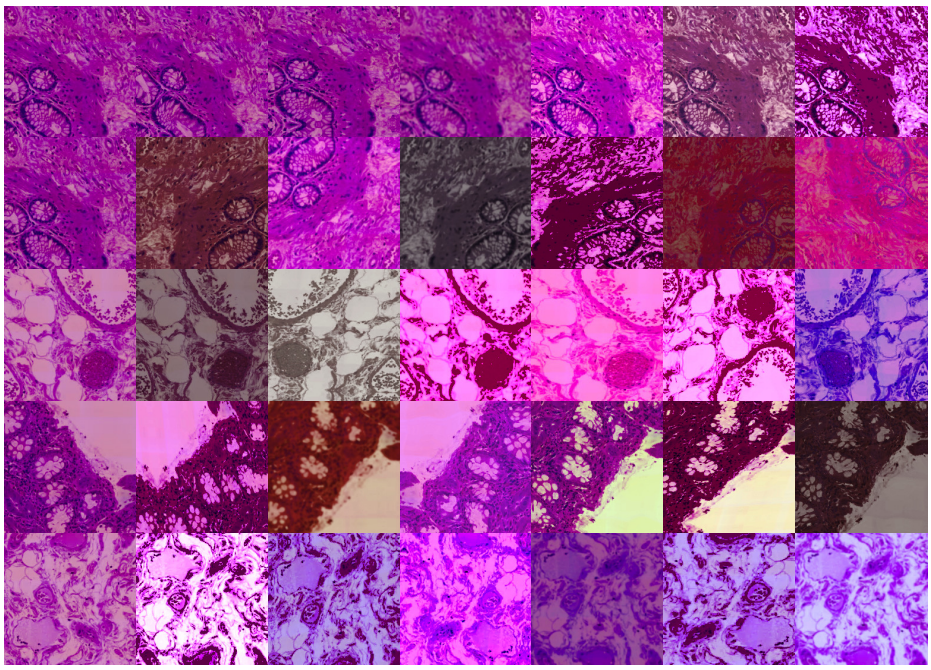


Figure 5.2: Data augmentation examples in isolation (first row; in column order (i) no augmentation, (ii) elastic transformation, (iii) affine transformation with mirror padding, (iv) Gaussian blurring, (v) brightness and contrast adjusted, (vi) brightness, contrast, hue and saturation adjusted, (vii) adjusted in the HED space) and randomly combined (remaining rows with the first column showing the image prior to any augmentation).

discriminative or not. This alone does not get one much further as it is very difficult to extract the discriminative subset of patches without expert knowledge and intensive human labour. Therefore, to obtain a collection of discriminative patches, an unsupervised learning method is used to cluster similar patches into several groups. In particular, I apply the k–means algorithm to group the patches from a subset of WSIs in the training set. To increase the robustness of the result to the random initialization of parameters I perform multiple clusterings using different random starting parameters, and adopt the one associated with the lowest loss, thereby avoiding sub-optimal local minima. In this work I adopt two clustering approaches (implementations provided in Appendix C), described next.

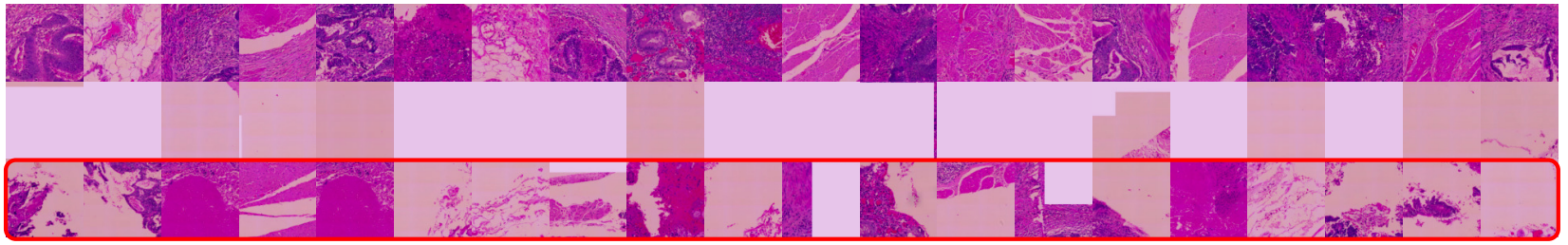


Figure 5.3: Examples from each cluster ($k = 3$) based on information density clustering. The third cluster (framed in red) was found to be the most discriminative.

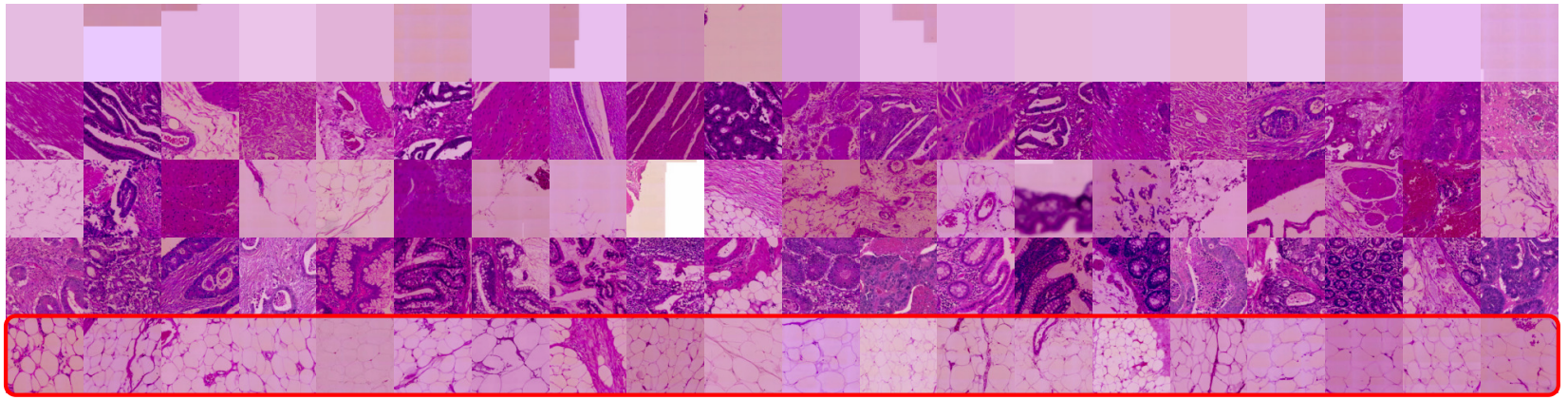


Figure 5.4: Examples from each cluster ($k = 5$) based on phenotype based clustering. The fifth cluster (framed in red) was found to be the most discriminative.

Figure 5.5: Examples from each cluster ($k = 10$) based on phenotype based clustering. The tenth cluster (framed in red) was found to be the most discriminative.

Table 5.1: Summary of data flow and transformation at different stages of the proposed algorithm employing phenotype clustering based patch selection.

Operation	Input dimension
Pre–trained VGG16_bn	$224 \times 224 \times 3$
Global average pooling	$7 \times 7 \times 512$
Dimension reduction (PCA)	512
k–means clustering	50

Information density clustering Information density is a simple but efficient way to group the extracted patches. In particular, since peripheral patches tend to contain large uniform areas, they are suited for compression by the DEFLATE algorithm used by the PNG image format.

The information ratio is defined here as the inverse of data compression ratio, $IR = \frac{1}{CR} = \frac{S_c}{S_u}$ where S_u is the bit size of an uncompressed image. For a 224×224 pixel RGB 8-bit image, S_u is $150,528$ bytes, and S_c is the size of the corresponding losslessly compressed PNG file. Examples are shown in Figure 5.3. Similar to my previous work (see Figure 5.6), most patches fall into one of the three clusters (CR–1 = 0.1, 0.4, 0.7). However, it is important to note that this does not necessarily imply that they are more pertinent for prognosis i.e. the ultimate task. For example, the spatial arrangement of immune and cancer cells in peripheral regions around the tumour is known to be informative in this regard.

Phenotype clustering I also developed a new phenotype clustering approach. The motivation behind phenotype clustering stems from the observation that the extracted patches exhibit significant heterogeneity; see Figure 5.6(b). Because it is computationally expensive to perform clustering in the original $150,528$ dimensional space ($224 \times 224 \times 3$), herein (instead of performing simple down-sampling) I used an ImageNet pre-trained CNN (*VGG16_bn* from PyTorch's torchvision.models subpackage) to generate phenotypes and then principal component analysis for dimensionality reduction. A summary of the process is shown in Table 1 and visual examples in Figures 5.4 and 5.5.

5.2.5 Patch–level CNN prognosis

CNN based classifiers are trained with patches from different clusters and used to determine which clusters are discriminative. The well-known InceptionNet-v3 network is employed as it outperformed other CNN architectures (most ResNets, DenseNets, and VGG-like architectures from the torchvision.models package) in preliminary experiments based on one-epoch training, and evaluation at the patient level (both using the training set).

I optimize the CNN at patch level and regard the ground truth la-

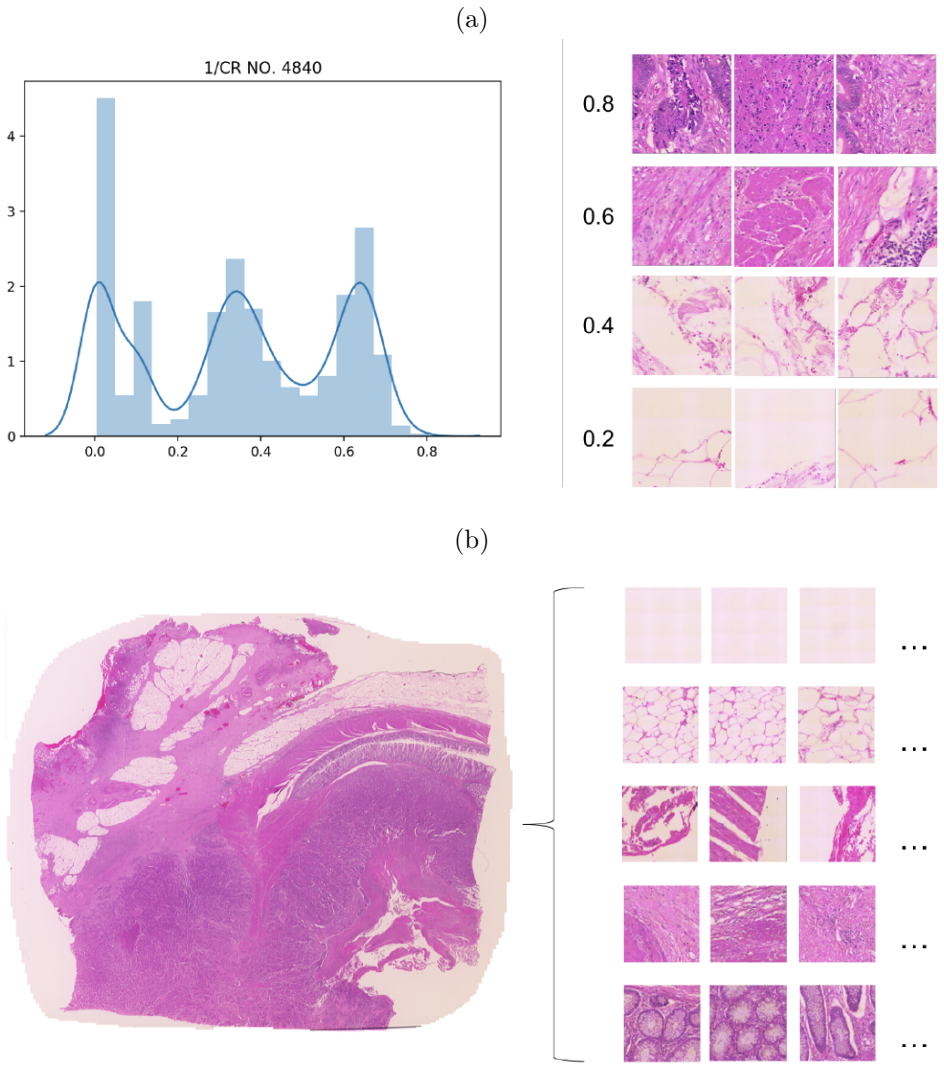


Figure 5.6: (a) Distribution of patch memberships (left) across different information ratio based clusters and the corresponding visual examples (right). (b) An example of a WSI (left) and sample patches (right) from the different phenotype based clusters inferred automatically.

bel of the patient, i.e. 5-year prognosis, as its label. The fact that not all patches contain discriminative features (for survival prediction) motivated my choice to train a network for each cluster separately and independently. During optimization, at every epoch, I also calculate the patient-level statistics on the training set using the aggregation methods that are described in the next section. The clusters which correspond to the networks that have the highest patient-level accuracy on the training set for each clustering technique (ID, PH5 and PH10) are inferred to be discriminative. Note that for most of the clusters, convergence was either not possible, or the patient-level training accuracy was worse than random guessing.

The CNNs are trained using Stochastic Gradient Descent (SGD) for 4 epochs. A batch size of 256 is employed. The initial learning rate is set to 0.02 (based on the learning range test described in the work of Smith and Topin [184]; implementation provided in Appendix C) and is decayed using a cosine annealing scheduler. A momentum of 0.9 and a weight decay of 0.0001 are used. The InceptioNet-v3 networks are initialized using pretrained networks on ImageNet. Two fully connected layers are placed on top of the CNN (a hidden layer with 512 neurons and the output layer with 2 neurons). The first layer is randomly initialized, whereas the weights of the second layer are set to zero, and the bias to $log(n_{gp}/n_{bp})$ where n_{gp} and n_bp are the number of good and bad prognosis patients, respectively, in the training set.

5.2.6 Aggregation of predictions

After the patch levels predictions (of a specific cluster) are made for the WSIs of a patient, these are aggregated into patient level predictions. Since each WSI has different numbers of tiles from a given cluster, and each patient has one or two WSIs, patch level predictions are represented by normalized histograms, thus effecting a homogeneous representation.

With the normalized histograms for every patients as input, an SVM classifier is trained to learn the cluster level outcome. Hyperparameter tuning with grid–search is employed (kernel: "linear" or "rbf", C: [0.001, 0.01, 0.1, 1, 10], and for rbf, gamma: [0.01, 0.003, 0.001, 0.0003, 0.0001, 0.00003, 0.00001]) with the loss function being the negated accuracy resulting from tenfold cross-validation, averaged over 30 independent runs. Majority voting is also implemented as a baseline.

5.3 Results and Discussion

In the past few years (following the publication of the early version of this work [217]) various groups have explored the application of deep learning on WSIs for end-to-end colorectal cancer prognosis [109, 183, 210]. Kather et al. [109] employed patch-level supervised learning and trained CNNs using

Table 5.2: Summary of results; prefixes ID and Ph refer to respectively information density and phenotype based clustering, followed by the corresponding number of clusters. The train set is divided into training (75%) T and validation (25%) V subsets for patch-level evaluation, yet for the patient-level performance, the entirety of the train set is used for the training, tuning, and evaluation of the SVM.

	Patch–level		Patient–level		Patient–level	
	T	V	Train set		Test set	
	Acc	Acc	Acc	F1	Acc	F1
ID3–CNN–Vote	0.68	0.64	0.62	0.01	0.61	0.00
ID3–CNN–SVM			0.58	0.56	0.64	**0.62**
Ph5–CNN–Vote	0.75	0.55	0.73	0.52	0.74	**0.63**
Ph5–CNN–SVM			0.70	0.64	0.67	0.61
Ph10–CNN–Vote	0.78	0.54	0.75	0.57	0.68	0.47
Ph10–CNN–SVM			0.80	0.73	0.75	**0.70**

pixel-level annotations of nine different tissue classes; (i) adipose tissue, (ii) debris, (iii) lymphocytes, (iv) muscle, (v) cancer-associated stroma, (vi) colorectal adenocarcinoma epithelium, and (vii) mucus, (viii) normal colon mucosa, and (ix) background. A "deep stroma" score calculated from tissue decomposition of the first five of these classes was found to be an independent prognostic factor of overall survival. Skrede et al. [183] turned to both weak supervision (using the label of the slide for each of its constituent tiles) and multi-instance learning for training CNNs, and had them evaluated in large, independent patient populations - both CNNs were found to be strong predictors of cancer–specific survival. Notably, they also used a segmentation network (trained with pixel-level annotations) as a preprocessing strategy for identifying areas within a given WSI with high tumour content [183]. Similarly, Wulczyn et al. [210] used a segmentation network that classifies a given a region as tumour vs. normal. Tumour regions from each WSI were then used to train CNNs in a weakly supervised fashion. Finally, human-interpretable histologic features were generated using a deep-learning-based image-similarity model [210]. The feature described as "poorly differentiated tumor cell clusters adjacent to adipose tissue" was found to be the most prognostic in isolation. In contrary to the above works, I approach the problem of relevant patch selection (defined as the "where" problem in previous chapters) through the use of fully automatic clustering which does not assume or require application of human prior knowledge.

I start my analysis by examining the overall prediction results and, in particular, the effect that different clustering and aggregation techniques, and their parameters have. For comparison, in addition to the SVM based aggregation described in the previous section, I also present results for majority voting based aggregation. A summary is provided in Table 5.2.

There are several important observations that are readily apparent from the table. Firstly, for all approaches, a better patch-level accuracy on the validation set does not correspond to better generalization performance at the patient-level. The availability of two WSIs for most patients, with one of them potentially containing no tumour regions, increases the label noise when translated from the patient-level down to the patch-level. Therefore, patch-level accuracy did not guide the selection of the most dicriminative clusters (as it did in my previous work [217]). Instead, at each epoch, the network is evaluated at patient-level by first training and tuning an SVM and then using it to get patient-level predictions on the entirety of the training set. Training set patient-level accuracy (highest between the SVM and majority voting aggregation techniques) was therefore the guiding metric for the selection of the most discriminative cluster of each clustering technique. Secondly, phenotype based clustering with larger k performed best at the patient-level in both training and testing sets. Interestingly, the manner of decision fusion played an important role in two out of the three clustering techniques. SVM based fusion dramatically improves algorithms with both information density and large k phenotype based clustering, achieving (64%/62%) and (75%/70%) accuracy and F1 score respectively. For the phenotype based clustering with $k = 5$, the manner of decision fusion (majority vote vs. SVM based) had a lesser effect on performance.

Finally, I sought additional insight and examined the best cluster from each method (image samples from all clusters are shown in Figures 5.3, 5.4, and 5.5). What I found was that the semantics of some of the clusters were easier to interpret than others. For example, the discriminative cluster of the phenotype based clustering with $k = 5$ (bottom row of Figure 5.4) could be easily interpreted as adipose tissue with clusters of tumour and lymphocyte cells appearing in some of the images. On the other hand, the most discriminative cluster of the large phenotype based clustering approach, and the one that performed the best, was harder to summarize into a single category (bottom row of Figure 5.5). Cancer-associated stroma, colorectal adenocarcinoma epithelieum, and muscle is seen in most of the sampled examples (for reference I used the examples shown in the works of [109, 210]).

The earlier version of this work was one of the first to address one of the most challenging problems in the emerging sphere of digital pathology – that of using images not previously annotated by a pathologist to develop algorithms that can be applied automatically to generate diagnostic and prognostic information from WSIs. Almost all current applications of CNN for WSI CRC prognosis require careful annotation of tissue images by a qualified pathologist (examples covered in the first paragraph), and this is a rate limiting step. The novel algorithm I introduced addresses the overwhelming amount of data by automatic, unsupervised discriminative patch selection and inference of prognosis on the level of the patient us-

ing decision fusion based on SVMs. On a real-world corpus my phenotype based clustering employed in conjunction with the aforementioned techniques achieved promising performance both in terms of overall accuracy and F1 score.

5.4 Implementation details

The framework described in this chapter was implemented using the following packages in Python: PyTorch [158], Numpy [89], Pandas [145], matplotlib [97], Pillow [49], and scikit-learn [159]. To handle WSIs in the CZI format, the Python libraries bioformats [129] and javabridge (available at https://github.com/LeeKamentsky/python-javabridge) were used. The implementation of Reinhard normilization was adopted from https://github.com/Peter554/StainTools. The code at https://github.com/gatsby2016/Augmentation-PyTorch-Transforms was used for the implementation of the data augmentation in PyTorch. The learning rate range test was implemented using the package at https://github.com/davidtvs/pytorch-lr-finder. Key code sections are provided in Appendix C.

Chapter 6

Breast cancer metastasis detection - billions of pixels

6.1 Problem formulation

Like the previous chapter, I consider the use of DL on images with billions of pixels. Most of the published methods preprocess these high–resolution images into a set of smaller image patches, thereby imposing an *a priori* belief on the best properties of the extracted patches (magnification, field of view, location, etc.). As an alternative to gigapixel image classification, herein, I introduce a new family of neural networks, henceforth referred to as Magnifying Networks (MagNets).

MagNets learn to use an attention based mechanism to decide on a coarse to fine basis the regions of the gigapixel image that need to be analysed at an increasingly fine scale. Incidentally, this is conceptually similar to a pathologist's knowledge and attention based use of magnification with a brightfield miscroscope. A microscope has multiple magnification settings that enable the user to view a specimen at different scales. Starting at the lowest magnification setting, the entire specimen can be observed. As the magnification is increased, finer detail is accessed, while at the same time, a smaller part of the specimen is displayed. During a visual examination, the clinician finds areas of interest at lower magnification levels and then examines them further at higher and higher magnification levels, accruing in the process information from all magnification levels that collectively enable a clinical decision to be made. Similarly, a MagNet starts at the lowest magnification level and recursively identifies, magnifies, and analyses areas of interest with more fine-grain detail (see Figure 6.1).

While remaining within the realm of weakly supervised learning, a MagNet nests spatial transformer modules [101] (with a differentiable upsampling mechanism) such that patches from higher magnifications are recursively extracted based on the assessment of lower magnification images. Depending on the amount of magnification at the current magnifying layer, a version of the WSI at a higher resolution can be accessed by the

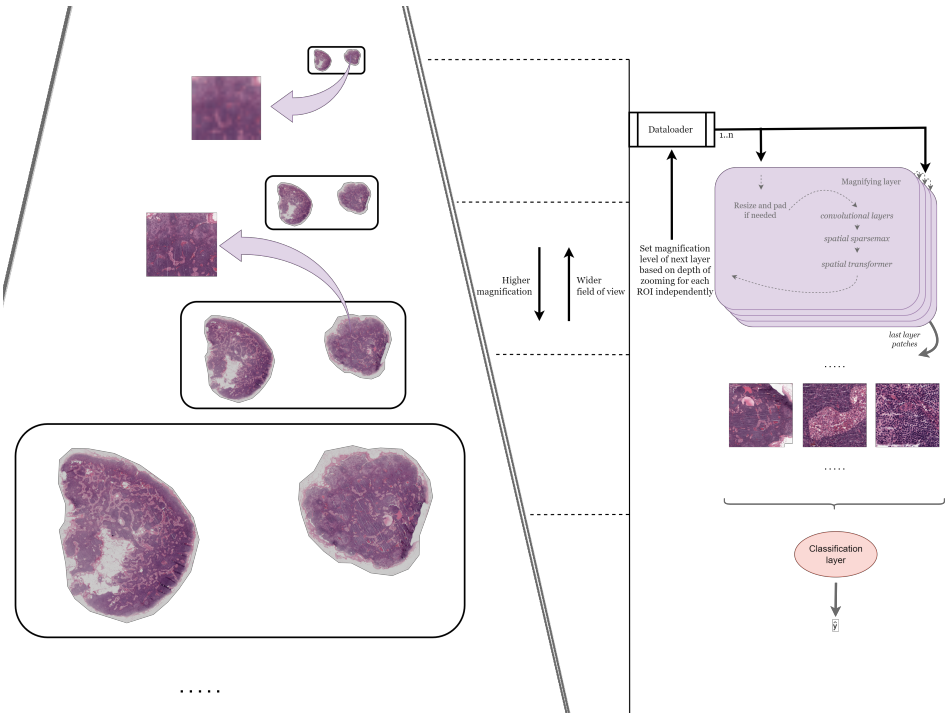


Figure 6.1: An Illustration of the architecture of MagNet. The depicted model consists of four magnifying layers and a classification layer. For each ROI of each magnifying layer, the right level of image resolution is set based on the level of magnification so far. Note that the ROIs of the last layer can span across different magnification levels, and with varying levels of fidelity, thereby providing information across multiple resolutions, and multiple fields of views.

subsequent layer, as illustrated in Figure 6.1.

MagNets provide a novel way of solving both the "where" (i.e. the identification of informative patches within a WSI) and "what" (i.e. visual understanding of an individual patch) problems of gigapixel image analysis in an end-to-end fashion [57]. Importantly, as I show in my experiments, my models can be optimized without the need for extra supervision for the "where" problem (e.g. boundary boxes). I conduct experiments by benchmarking MagNets on the Camelyon data sets. In particular, 3– and 4–layer MagNets are trained and evaluated on the task of WSI classification [57] using the publicly available data set of the Camelyon challenge [8]. MagNets offer an attractive alternative to gigapixel image analysis, especially in the context of digital pathology, as they come with innate transparency (embedded hard attention), no preprocessing requirements (i.e. end-to-end training capability with gigapixel images), and an ability to perform both localization and classification tasks with no additional information (only slide-level information is used). My contributions are:

- In the context of WSI classification of metastases, I propose the possibility of identifying and magnifying ROIs starting from a very low resolution downsampled version of the WSI (3 channels, 56×56 pixels), and experimentally show that recursively identifying and magnifying ROIs allows for the extraction of informative areas across magnification levels, without having to preprocess billions of pixels.
- Without leaving the weakly supervised paradigm, I explore nested attention using the spatial transformer module for gigapixel image analysis.
- To the best of my knowledge, this is the first work that automatically finds, and fuses information from multiple *learnt* magnifications on WSIs. The proposed method is able to exploit rich contextual and salient features, overcoming the typical problem of patch–based processing that poorly capture the information that is distributed beyond the patch size.

6.2 Methods

6.2.1 Cohort

The Camelyon data sets contain WSIs from surgically resected lymph nodes of breast cancer patients. These WSIs were independently curated across multiple hospitals [61, 130]. Camelyon16 includes images from 238 normal and 160 cancerous tissue sections whereas the publicly available portion of Camelyon17 has a total of 500 WSIs (318 normal, 182 cancerous) grouped into artificial patients [8]. In addition, in the case of metastasis, metadata is available as to the extent of the metastasis (macrometastasis, micrometastasis, or isolated tumour cells (ITC)). Since only a few cases contain the much more difficult ITC type of metastasis (36 cases, i.e. $\approx 4\%$ of all cases), it is unlikely that they are sufficiently representative of the ITC class. Therefore, they are excluded from the training data set.

I follow the protocol described in the Camelyon competition website, and in addition set aside 25% of Camelyon17 as a testing set [187]) (73 WSIs with metastases, 36 with ITC, 17 with micro-, and 20 with macro-, and 88 WSIs of normal tissue). I shuffle the remaining WSIs from Camelyon17 with the Camelyon16 WSIs, and train on the 80% and validate the better models from the remaining 20%. The best MagNets (based on the validation set) are retrained on both the training and validation data, and evaluated on the testing set. Since ITC cases were excluded and the models were trained for WSI classification, rather than patient–level pN prediction, the MagNet models were not evaluated on the privately held testing set.

The pixel-level annotations that are available for some of the WSIs of Camelyon are not used in my work, other than to gain insights in post-processing, e.g. as in Figures 6.6 and 6.7. During training, I only use

the binary slide-level label that indicates the presence, or lack thereof, of cancerous cells within the gigapixel image.

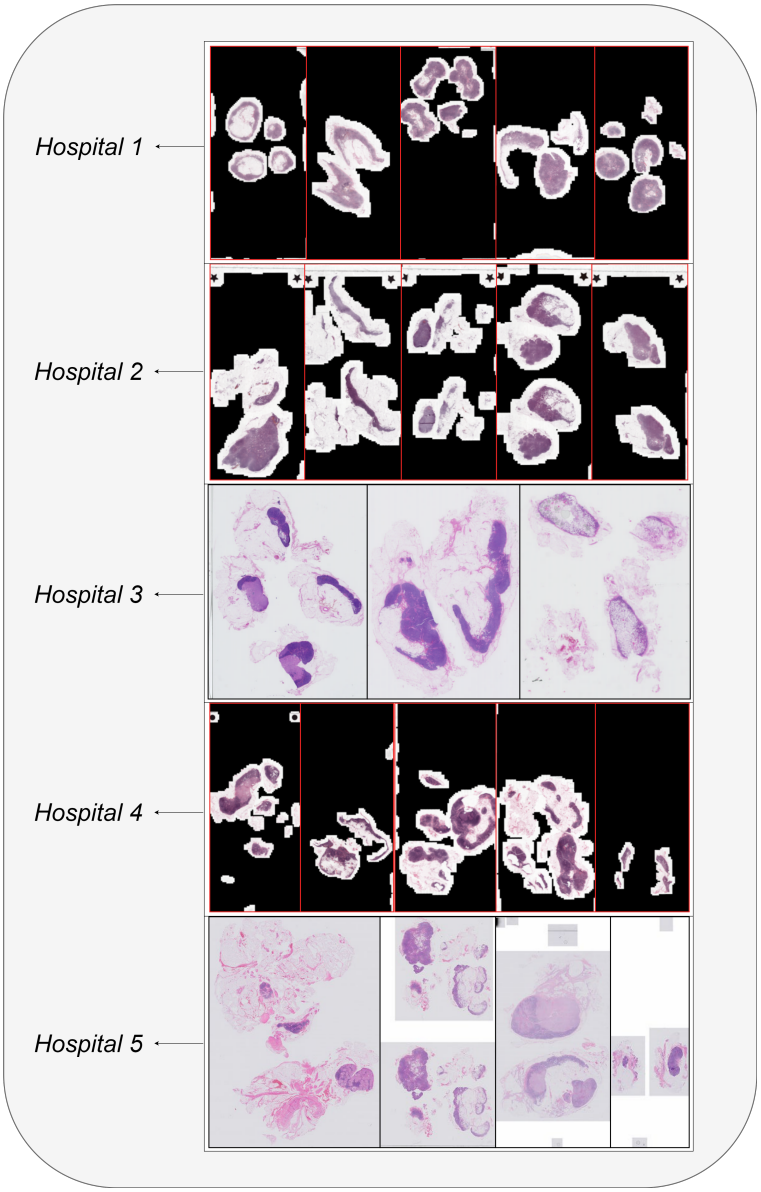


Figure 6.2: Randomly sampled WSIs from each hospital.

6.2.2 Data preparation

A significant advantage of MagNet over the methods described in the existing literature is that no preprocessing is required for the input WSI. Therefore, no data preparation stage was needed.

6.2.3 Magnifying networks

A MagNet consists of N magnifying layers followed by a classification layer. The magnifying layers are responsible for identifying the information relevant to the task at hand within a WSI (in the form of image patches), whereas the classification layer is concerned with the visual understanding of the extracted information.

Magnifying layer

Consider a single gigapixel image I_0 that will pass through a MagNet.

Resizing and padding As I subsequently employ convolutional layers expecting 56×56 pixel images, input I, either as a single input image (e.g. I_0) or a set of images, is resized to $56 \times h_i$ or $w_i \times 56$, based on bilinear interpolation, with the smaller side, h_i or w_i, then symmetrically padded (new pixels are black to match the filter, see Section 6.3) so that $h_i = 56$ or $w_i = 56$ accordingly. For the purpose of up-sampling (explained shortly in detail, see the "Sampling" paragraph), a larger version I' (112×112 pixels) is also generated using the same protocol. Note that although preliminary experiments were conducted using larger images as input to the magnifying layers (I and I' with 112×112 and 224×224 resolutions respectively), single GPU training of multiple, stacked magnifying layers was not possible with these resolutions.

Convolutional layers. Salient regions in each image patch vary significantly in size. This comes as a consequence of the varying levels of metastasis, but also of the lack of standardization in WSI digitization across different institutes and scanners (see Figure 6.2). Therefore, the right kernel size for the convolutional operations varies depending on I. Hence, I stack convolution layers with different kernel sizes similarly to InceptionNet–v3 [190].

Let *Conv2D* be a $n \times n$ convolution layer (with padding set to 1), followed by Batch Normalization and a ReLU nonlinearity. *MaxPool* is a max pooling operation with a 3×3 kernel and padding. I define a "Branch" as the simultaneous forward pass of the input through five layers where a *layer* sequentially applies a number of *Conv2D* and *MaxPool* operations. In particular, the five layers are:

- 1×1 *Conv2D*
- 1×1 *Conv2D* $\rightsquigarrow$ 3×3 *Conv2D*
- 1×1 *Conv2D* $\rightsquigarrow$ 3×3 *Conv2D* $\rightsquigarrow$ 3×3 *Conv2D*
- 1×1 *Conv2D* $\rightsquigarrow$ 3×3 *Conv2D* $\rightsquigarrow$ 3×3 *Conv2D* $\rightsquigarrow$ 3×3 *Conv2D*
- *MaxPool* $\rightsquigarrow$ 1×1 *Conv2D*

The outputs of all of the layers above are concatenated into a single tensor. Since padding is employed, the output has the same height and width as the input. MagNets use patch and layer–specific "Branches", e.g. a 2-layer MagNet with two patches extracted at each magnifying layer has six of these layers (two at the first layer, and four at the second).

The first "Branch" takes an input with 3 features (e.g. an image), and outputs a tensor with 15 features, i.e. each of the five layers outputs tensors with 3 features. The second "Branch" takes the output of the first "Branch" and outputs a tensor with 40 features (8 from each of the five layers). In the third "Branch", given the input of the second "Branch", 1 feature is extracted from each of the five layers, and the concatenated output is forwarded through a 1×1 *Conv2D* that returns a tensor with 1 feature. Given that the input images in my experiments had a 56×56 pixel resolution, the output from the third "Branch" is a tensor with $56 \times 56 \times 1$ dimensions. An implementation of "Conv2D" and "Branch", as described above, is provided in Appendix D.

Spatial transformer. A spatial transformer network (STN) is used to transform hard attention based cropping into a differentiable process. An STN consists of three parts; a localization network, a grid generator, and a sampler [101].

The *localization network* in the literature is typically a fully connected or a recurrent neural network [185] that receives an input from a CNN, and its role is to output a spatial transformation of the coordinate space of the original image [170]. However, due to their high demand for GPU VRAM, owing to their large number of parameters, both options are impractical for employment within MagNets. Instead, MagNets utilize a spatial sparsemax at the last convolutional layer whose output can be used to infer the spatial transformation (hard attention based cropping) parameters (s, t_x, t_y) directly. In particular, the dimensions of the output of the last convolutional layer are the same as the input image, i.e. 56×56 pixels. Following the application of the spatial sparsemax operation, the output can be thought of as a probability mass function with the expected $L1$ norm translating to the scaling parameter (s), and the translation parameters (t_x, t_y) obtained by expected value over indices of the x-axis and y-axis respectively.

Given the transformation parameters s for isotropic scaling and t_x, t_y for translation along each axis, I further constrain the parameters as follows:

$$s = max(s, 0.05) \tag{6.1}$$

$$t_x = tanh(t_x) \tag{6.2}$$

$$t_y = tanh(t_y) \tag{6.3}$$

with θ of spatial (affine) transformation A_θ:

$$\theta = \begin{bmatrix} s & 0 & t_x \\ 0 & s & t_y \end{bmatrix} \tag{6.4}$$

The *tanh* constraint on the translation parameters implicitly forces the network to favour centre extraction, whereas the minimum bound imposed on the scaling helped ensure that I do not get vanishing gradients for some STs during the early stages of training. An implementation of the above technique is provided in Figure D.3.

The *grid generator* then creates the desired grid by multiplying θ with a 56×56 pixel meshgrid. Finally, an image can be interpolated onto the grid using a *sampler*.

Sampler A sampler takes a set of sampling points along with an image, and applies a differentiable sampling kernel (e.g. bilinear) to produce the sampled image. Bilinear interpolation is a poor choice for a sampling kernel for my work as shown in the empirical analysis below (performed on the training set), and also supported by the literature, e.g. it performs poorly under severe scale changes [105], with poor gradient propagation. Wei et al. [105] proposed an alternative sampler, Linearized Multi-Sampling, whose gradients are resilient to the amount of scaling. I use the original implementation of this sampler provided by Wei et al. [105].

The empirical analysis of bilinear interpolation and Linearized Multi-Sampling was performed on a subset of the training set, with similar results observed given any WSI. In particular, given a *downsampled* WSI, I iteratively move part of the image to a specific location (locations defined by a grid as shown in Figure 6.3), and use the ST to get the gradients based on an $L2$ loss that will allow the part to be moved back to its original place. I observe that the quality of the gradients suffers when the image is downsampled and bilinear sampling is employed. On the other hand, the method that was introduced by Wei et al. demonstrates high quality gradient estimates even in extreme cases of downsampling as shown in Figure 6.3.

Sampling This is the part that makes each layer "magnifying". MagNet applies the transformation A_θ on I' instead of I, thereby allowing the output to contain information (finer-grain) that was not present in I. An example of a magnifying layer that outputs two patches is shown in Figure 6.4. By stacking multiple magnification layers together, MagNets are able to retrieve information from increasingly higher magnification levels.

The magnification level from which I' is extracted is set dynamically. In particular, given h_0, w_0 as the height and width of a WSI (at the highest magnification level, i.e. pyramid level 0), and h_c, w_c as the height and width of a requested ROI (based on the affine transformation of the STN), the magnification level m is calculated as follows:

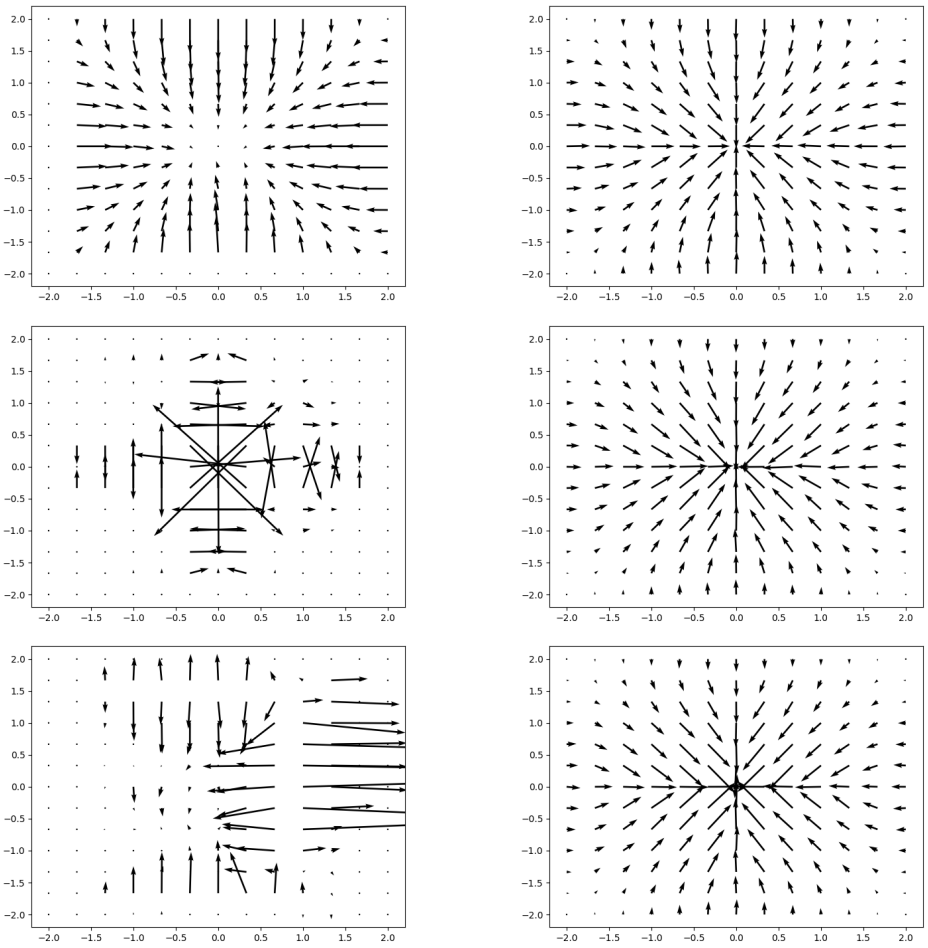


Figure 6.3: Gradient analysis of using bilinear sampling (Left) versus the Linearized Multi–Sampling approach [105]. In order from top to bottom, the image is not downsampled, downsampled by a factor of 4, and downsampled by 8.

$$
\begin{aligned}
R_h &= \left\lfloor log_2 \left(\frac{h_o}{h_c} \right) \right\rfloor, \\
R_w &= \left\lfloor log_2 \left(\frac{w_o}{w_c} \right) \right\rfloor, \\
R &= max(R_h, R_w), \\
m &= max(m_{max} - R, 0),
\end{aligned}
$$

where m_{max} is the total number of magnification levels of the WSI. For example, given a WSI with $50,000 \times 100,000$ pixels, and 9 magnification levels, access to a specific magnification level depends on the requested area (width $\times$ height) as follows:

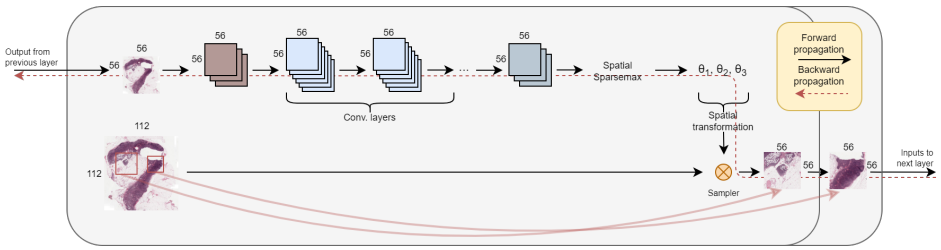


Figure 6.4: An illustration of a single magnifying layer that outputs two patches. The convolutional layers are independent between the two patches. The red squares illustrate the affine transformation based on the outputted thetas. Note that if this was the last magnifying layer, the image size of the patches would have been 224×224.

Width		Height	WSI resolution	level
$\geq 25,000$	and	$\geq 50,000$	171×391 pixels	8
$\geq 12,500$	or	$\geq 25,000$	391×782 pixels	7
$\geq 6,250$	or	$\geq 12,500$	$782 \times 1,563$ pixels	6
$\geq 3,125$	or	$\geq 6,250$	$1,563 \times 3,125$ pixels	5
$\geq 1,563$	or	$\geq 3,125$	$3,125 \times 6,250$ pixels	4
≥ 782	or	$\geq 1,563$	$6,250 \times 12,500$ pixels	3
≥ 391	or	≥ 782	$12,500 \times 25,000$ pixels	2
≥ 171	or	≥ 391	$25,000 \times 50,000$ pixels	1
< 171	and	< 391	$50,000 \times 100,000$ pixels	0

The above assumes that the spatial resolution of the WSI is halved at each subsequent level, which is indeed the case for the Camelyon data set.

Classification layer

The images of the last magnifying layer are sampled using a grid with 224×224 pixel resolution (instead of 56×56 pixels). These images are forwarded through an ImageNet–pretrained CNN (InceptionNet-v3) that outputs a feature map into a Gated Recurrent Unit (GRU) network. The output of the GRU is passed through a FCNN (two layers with 512 and 256 hidden neurons respectively) to output a slide–level $\hat{y}$ estimate on whether the given WSI contains cancer or not.

Auxiliary classifiers

A form of both self–supervision and weak–supervision is introduced by using two auxiliary classifiers. These are ImageNet–pretrained ResNets–18 networks [91] that output a slide-level prediction using the extracted images from magnifying layer 1 and layer 3 respectively. Cross-entropy is used between the slide-level labels and the ResNet-18 outputs in a weakly–supervised fashion. In addition, the paradoxical loss was also employed

as a form of self–supervision [142]. The premise of the paradoxical loss is that information presented at layer 3 images should provide an equally good, or better, prediction than that from layer 1. Under this assumption, instances where the opposite is observed are viewed as "undesirable and paradoxical" [142]. The paradoxical loss over M inputs is computed as follows:

$$L_1 = \frac{1}{M}\sum_{i=0}^{M} max(P_1 - P_3, 0),$$

where P_1, P_3 is the estimated probabilities of identifying the true class label (slide–level) by patches from layer 1 and 3 respectively.

Configurations

The network consists of L magnifying layers each of which can access increasingly higher magnification scales as determined dynamically from the degree of zooming (i.e. s) thus far. At each layer l, P_l number of patches are extracted (ROIs).

A consequence of the recurrent nature of MagNets is that an exponential number of patches are extracted and analysed from a single gigapixel image, if more than 1 patch is extracted per layer. In particular, given a constant P across the layers:

$$\text{Total patches extracted in layer } l = \begin{cases} l, & \text{if } P = 1. \\ P^l, & \text{otherwise.} \end{cases}$$

I find that a combination of $[2, 3]$ for P_l, (i.e. 2 ROIs are extracted in some magnifying layers, whereas in others, 3 ROIs are extracted) provides a balance between a sufficient rate of expansion (breadth), while allowing for up to 4-layer MagNets (depth) to be trained on a GPU with 24 GB of VRAM.

6.3 Evaluation

Data augmentation For any given image (both during training and inference), I apply a filter that sets grey image pixels (including the degenerate form of grey that is white) black. In particular, these are pixels for which the corresponding red, green, and blue channel values differ from each other by less than 15 (scale $0 \ldots 255$). This filter removes background and the various scanning artefacts (smudges, etc.) which are most strongly visible in otherwise nearly uniform regions of the slide. I employ neither colour normalization nor random colour perturbation [131]. I find the latter to be ineffective (based on the validation set), whereas the former is avoided since it would add a significant computational overhead to WSI analysis. Synthetic data augmentation I perform during training involves horizontal and vertical mirroring, and rotation by 90, 180, and 270 degrees.

Training The final networks were trained using the Adam optimizer for 200 epochs. However, during both hyperparameter tuning and the ablation experiments, the models were only trained for 20 epochs. A batch size of 16 and 8 was employed for MagNet networks with 3 and 4 layers respectively. The initial learning rate was set to $3 \times 10^{-}5$ and was decayed using a cosine annealing scheduler. Both ResNet-18 and InceptioNet-v3 networks are initialized using pretrained networks on ImageNet. The ST convolutional layers are randomly initialized.

"Frozen" patch Some WSIs have already been preprocessed so that they only contain regions with tissue, whereas others depict all of the tissue slide. This diversity comes as a consequence of the differences in the clinical pipelines leading to the creation of WSIs, e.g. due to different scanning profiles (see Figure 6.2). In order to mitigate for the above variance, I freeze the 1st patch of the 2nd layer so that it always attends to the whole input image. This allows for the image to catch-up in quality in the cases where a large amount of zooming was required at the first magnifying layer, i.e. when the WSI shows the whole tissue slide (this can be observed in all of the examples in Figure 6.6, as indicated by the vertical, orange arrow).

Loss functions I employ the paradoxical loss function (L_1) as a form of self–supervision for the convolutional layers in the weakly–supervised STNs. In addition, cross-entropy is used between the slide-level labels and both the last output of the GRU (L_2) as well as the ResNet-18 outputs(L_3). The final loss function is computed as the sum of L_1, L_2, and L_3.

Table 6.1: The results of MagNets on the testing set against baselines and existing competitive methods. Given X/Y percentages, X corresponds to AUROC, and Y to accuracy.

	# patches	Macro (%)	Micro (%)	ITC (%)	Macro (%) & Micro (%)	All (%)
Mean RGB Baseline [192]	-	59/-	57/-	-	58/-	-
DSMIL-LC [126]	> 5000	-	-	-	92/90	-
HAS [118]	> 5000	-	-	-	-/83	-
3-layer MagNet	28	95/88	71/78	57/69	84/77	71/64
4-layer MagNet	64	91/89	76/83	63/70	84/81	75/66

Table 6.2: Ablation experiments for different components of a 3-layer MagNet using three random seeds, and random training-validation splits. AUROC decreases when a smaller number of patches is used, and when L_2, L_3, or "Frozen" patch are omitted.

# patches	L_2	L_3	Frozen patch	AUROC [%]
3, 2, 3	✓	✓	✓	68.8
2, 2, 2	✓	✓	✓	63.1
2, 3, 2	✓	✓	✓	66.1
3, 2, 3	✓		✓	67.3
3, 2, 3		✓	✓	68.0
3, 2, 3			✓	66.9
3, 2, 3	✓	✓		62.4
3, 2, 3				62.9

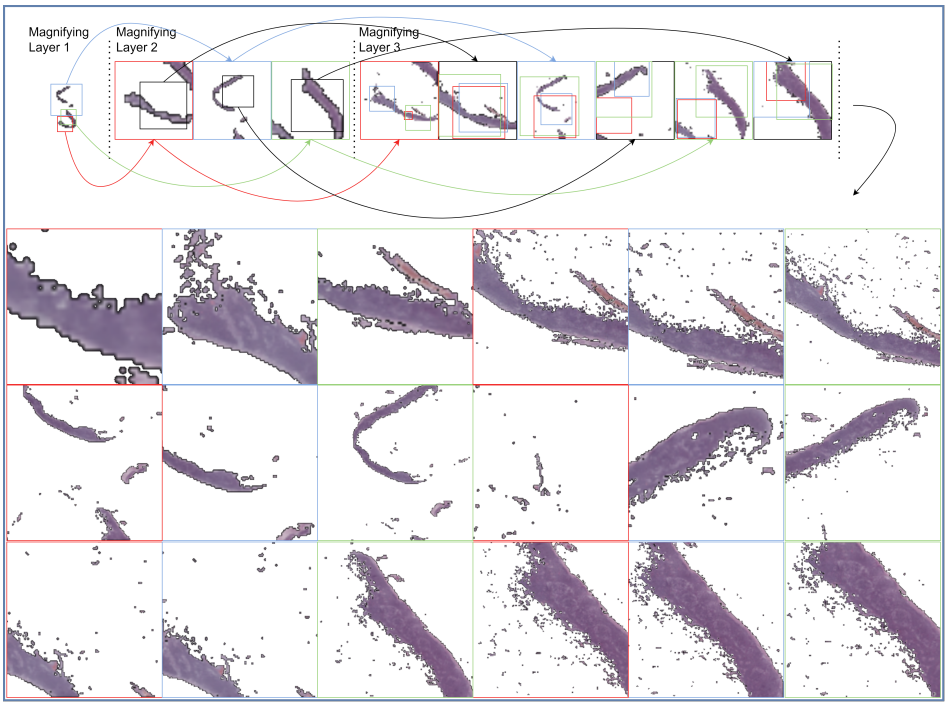


Figure 6.5: A visualization of a forward pass of a WSI with micro metastasis (from the testing set) through a 3–layer MagNet model. The background of the images is shown in white for visualization purposes. The 18 image patches that are extracted at the last magnifying layer are passed forward through the classification layer.

6.4 Results & discussion

To evaluate the proposed method, using the optimization framework described in the previous section I trained 3-layer and 4-layer MagNets on

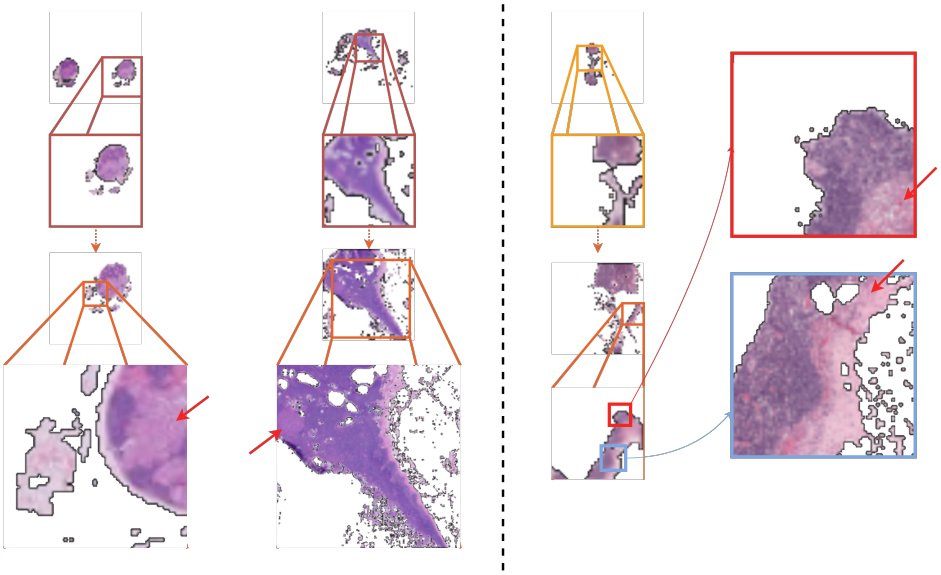


Figure 6.6: Cancerous regions of macro, and micro–metastasis as identified by the 3–layer MagNet model (on the left), and of micro–metastasis as identified by the 4–layer MagNet (on the right). The pointing red arrows show cancer regions based on the annotations provided by the pathologists at the highest magnification scale.

the task of cancer metastasis detection from WSIs. A summary of the results is presented in Table 6.1 which shows the AUROC – the standard evaluation metric used in related literature [201, 192, 61, 130] – and the accuracy (threshold was set to 0.5). The inclusion of loss functions L_1 and L_3, the "frozen" patch, as well as the specific MagNet configuration (i.e. patches per layer) was supported by the outcomes of the ablation studies in Table 6.2. In parallel to my work, Kong et al. [118] were the first to introduce the concept of nested attention, and by extending the attention module introduced by Katharopoulos and Fleuret [108], proposed a two-layer hierarchical attention model that enables end-to-end training of deep learning models from WSIs. Although conceptually similar, MagNets further extend the idea of nested attention by allowing an arbitrary number of attention layers (called magnification layers herein), and by not enforcing any *a priori* properties on the selected patches. For ease of comparative analysis, we include a baseline encoder as reported by Tellez et al. [192] based on average colour intensity (termed RGB baseline), as well as the dual-stream multiple instance learning network (DSMIL) [126] and the two-stage hierarchical attention sampling method (HAS) [118] as evaluated on micro- and macro-metastases (Camelyon16). DSMIL constitutes one of the most competitive methods in the weakly supervised paradigm, but involves extensive preprocessing steps; namely, the extraction of millions of patches

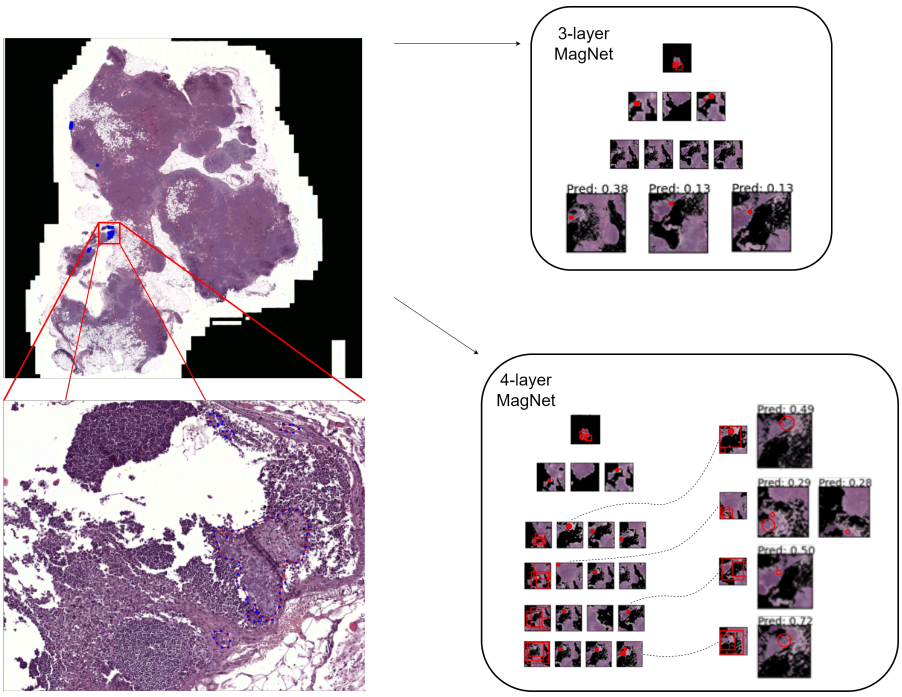


Figure 6.7: A WSI from Hospital 3 wherein the 4-layer MagNet correctly identified a cancerous region, whereas the 3-layer MagNet produced a false negative.

at different magnification scales [126]. On the other hand, HAS has no preprocessing steps. Nevertheless, in contrary to MagNet, it requires a large number of patches to be dynamically extracted from each attention layer (50–100), with each layer specific to a predefined magnification scale (as selected prior to training), and each patch loci predefined in a grid-like fashion. Although I make no argument against this design, I hypothesise that MagNets are able to solve the "where" problem more efficiently (as indicated by the much smaller number of patches explored and magnified) due to the lack of such constraints. The 4-layer MagNet almost meets the accuracy of the best model of HAS (81% vs 83%) on classifying tumour vs normal WSI with 54% sensitivity, 94% specificity and 65% F1 score.

MagNets exhibit robust and effective exploration capabilities, namely attending to image content in an attention driven manner, exploring slides at varying magnification levels best suited to the task at hand and learning how to fuse relevant information both within the same WSI region and across different regions and magnification levels. In addition, the classifier (in the form of InceptionNet) demonstrates an excellent ability to distinguish normal from cancerous tissue across irrespectively of the magnification scale. Examples corroborating this are shown in Figure 6.6. Interestingly, both MagNets perform well on ITC cases despite the lack of

Table 6.3: The results of my MagNet models on WSI subsets of the testing set (the percentages correspond to AUROC) sorted by their hospital of origin.

3-layer MagNet	Macro- & Micro-	Macro-	Micro-	All
Hospital 1	89%	95%	80%	73%
Hospital 2	60%	77%	**54%**	58%
Hospital 3	96%	97%	**95%**	91%
Hospital 4	87%	98%	81%	68%
Hospital 5	-	92%	-	79%
4-layer MagNet	Macro-	Micro-	Macro- & Micro-	All
Hospital 1	88%	97%	84%	76%
Hospital 2	85%	77%	**85%**	74%
Hospital 3	84%	97%	**58%**	78%
Hospital 4	75%	93%	65%	68%
Hospital 5	-	92%	-	79%

ITC examples in the training set.

I re-evaluate the final MagNet models (both the 3-layer and the 4-layer) against the *testing set* of WSIs but this time I take into consideration the different hospitals they WSIs came from. There are (15,11,3,4), (13,7,6,2), (19,2,2,4), (14,8,6,3), (16,8,0,7) of normal tissue, ITC, micro- and macro-metastases cases, respectively, for hospitals 1 to 5 (see examples from the different hospitals in Figure 6.2). A summary of the results is provided in Table 6.3. I compute the ranking capabilities (AUROC) of MagNets between normal and tumour cases scanned from the same hospital for all five hospitals independently. No or minimal discrepancy is observed between the performance of the models for Hospitals 1, 4, and 5. However, the 4-layer MagNet performs better than the 3-layer MagNet on micro metastasis cases from hospital 2 (54% vs 85%), and for cases from hospital 3, the opposite is observed, i.e. 3-layer MagNet performs better (95% vs 58%). For Hospital 2, with 6 micro metastasis cases, the false negatives (i.e. classification of a WSI with cancer as being normal) from the 3-layer MagNet was the source of the discrepancy. The extra magnifying layer of the 4-layer MagNet provided higher resolution images that, for the above cases, was needed for the cancer to appear in the patches. In Figure 6.7 we show an example of a WSI which is incorrectly classified as negative by a 3-layer MagNet, but correctly as positive by a 4-layer MagNet, together with explanatory visualizations corresponding to the two networks. For Hospital 3, the discrepancy came down to the decision of one case with micro metastasis (hospital 3 only had 2 micro metastasis cases). The 4-layer MagNet misses the part of the WSI that had cancer from the very first magnifying layer, whereas the 3-layer MagNet correctly classifies it (this

example is shown in Figure 6.5).

I also investigate the scaling that is typically learnt by the 4-layer MagNet across the different hospitals and different cases (normal vs different types of mestastases). No major difference is observed between layers 2, 3 and 4 with the average scale learnt being 0.5. Nevertheless, the standard deviation ranges significantly from 0.1 to 0.3. For layer 1, a mean scale difference is observed between the different hospitals, with the average and standard deviation being 0.21 ± 0.13, 0.31 ± 0.15, 0.32 ± 0.22, 0.25 ± 0.13, and 0.32 ± 0.22 for hospitals 1 to 5 in order. Since there are no universal scanning settings (see Figure 6.2, e.g. hospitals 1, 2 and 4 scanned the whole tissue slide whereas hospitals 3 and 5 applied a form of cropping in most cases), the shift in the mean scale seems to be the model's approach to generalizing across different hospitals.

MagNets do not require WSIs to be patch–based preprocessed. Instead, the network starting from the lowest magnification level, can dynamically explore a WSI in continually higher magnification levels, as seen fit by MagNet and in the visual context of the specific WSI. The premise being that the patches with the best magnification level, field–of–view, and location according to the optimizing task will be dynamically extracted. Indeed for the task of breast cancer metastasis detection, my MagNet models performed extremely well given that they only had to process from 28 to 64 image patches per WSI, far less than any of the existing approaches (10 to 300 times less) [126].

6.4.1 Limitations

The experimental results we present in this work concern the use of MagNets as effective predictors for whole slide image analysis. However, for completeness and future use in mind, we have also investigated them in terms of efficiency. Traditionally, a PyTorch dataloader is accessed once per mini–batch. In contrast, because of the magnifying nature of MagNet, during a single forward pass through a MagNet, the WSIs in the mini–batch would need to be accessed multiple times to progressively retrieve smaller regions of the original WSI at a higher magnification level. As a result, I had to implement a stateful Dataset class with the "__get_item__" changed so that it returns the right patches based on the dynamically changing metadata. As can be expected, the addition above introduces a time overhead, a consequence of not using the PyTorch dataloaders in the traditional, heavily optimized, way.

This overhead can be an impediment to training on images with much larger than, say, 56×56 pixel resolution (during the magnifying layers), or training on much larger data sets, given that my MagNet models on average needed 30 minutes per training epoch. Therefore, in cases where higher resolution is needed to be able to see where to zoom, training MagNets may be very computationally expensive. However, it is worth mentioning that the above overhead does not constitute an issue for MagNet's clinical

adoption. It takes on average less than 5 minutes to get predictions for the whole testing set, i.e. ≈ 2 seconds per slide when using a mini–batch of 16.

The Camelyon data set provides a fitting optimization challenge for MagNets, that of breast cancer metastases detection, considering the varying granularity that needs to be assessed in having to predict macro–metastases, micro–metastases, and isolated tumour cells (ITC) from WSIs. However, it could be argued that the configurations of MagNet I explore herein are not adequate for clinical adoption. In particular, due to the unconstrained and non–exhaustive nature of exploration, a MagNet could miss a region containing a metastasis early on, thus producing a false negative. With clinical adoption in mind, a more exhaustive exploration would be required by perhaps increasing the number of patches that are extracted at each magnifying layer. Moreover, although the Camelyon data sets are useful and indeed the most appropriate public corpora for the evaluation of MagNets on the task of gigapixel image analysis, it is important to appreciate that they were collected for a very specific set of analytical tasks, namely the localization of tumour regions and the holistic classification of WSIs as cancerous or not. However, the above results cannot guarantee the same efficacy of MagNets when applied on a problem where "identification of patches to zoom-in" is not as clear-cut. Moreover, since the hospital origin of each WSI was not considered while creating the testing set, it is possible that generalization of the trained models may not extend beyond these 5 hospitals.

6.5 Conclusions

In this work, I introduced the MagNet – a neural network consisting of fully-connected, convolutional, and recurrent layers that employ STs in a novel manner so as to facilitate attention and data driven recurrent exploration and, ultimately, end-to-end learning over gigapixel images. The built-in hard attention mechanism of MagNets makes them well suited for clinical use. In particular, the explanations generated by MagNets are visually intuitive, e.g. as shown in Figure 6.6), for a domain-specific expert to interpret (as they visually depict a subset of the original WSI) and can be generated on the go without any additional overhead. Moreover, MagNets can be optimized without extra supervision (e.g. bounding boxes) for the task at hand. This is of high significance since for most clinical tasks, collecting ground truth data required for a higher degree of supervision is either extremely laborious and expensive, or simply not possible, e.g. in the case of patient prognosis.

6.6 Implementation details

MagNets and the optimization framework were implemented using the following packages in Python: PyTorch [158], OpenSlide [82], Scikit-

Image [197], Numpy [89], matplotlib [97], Pillow [49], and scikit-learn [159]. The sparsemax implementation was adopted from https://github.com/deep-spin/entmax [160]. The attention layer was inspired by the implementation at https://github.com/TolgaOk/Differentiable-Hard-Attention-Module. Key code sections are provided in Appendix D.

Chapter 7

Conclusions

Pathology has embraced the era of digitization, with volumes of digitized tissue being archived across the world and emerging research demonstrating its potential use in the sphere of image analysis and computational pathology. Tissue images present novel challenges to image analysis, namely, extremely high–dimensionality, heterogeneous saliency, and multiple sources of artifacts and noise. For supervised approaches, a family of learning techniques that has revolutionized both image analysis, and the learning–from–data paradigm more in general, tissue image analysis for a clinical task can be difficult to apply, given the paucity of ground truth labels, and often low precision. Despite these challenges, a large body of work exploits these images (e.g. by identifying and extracting patterns and morphological information), and applies machine learning for a handful of clinical tasks. From the perspective of histopathology, this is unsurprising, given the known association between histopathology and a wide spectrum of clinical outcomes. From the perspective of the machine learning engineer, this is also unsurprising, given the successful application of machine learning on seemingly every branch of science, and known superiority over traditional statistical methods when it comes to high dimensional data.

The first achievement of my thesis is the successful application of learning–from–data, i.e. machine learning, methods for cancer prognosis. Cancer prognosis based on histopathology is the gold standard in the clinic for solid tumours. Yet, the censored nature of data associated with survival analysis precludes the use of traditional machine learning models. The framework I introduced in Chapter 3, and later reused in Chapter 4, offers a straightforward, yet rigorous and principled method for developing prognostic machine learning systems using short fat data. Unlike much of the existing prognostic models based on machine learning and Cox regression, I take the needed measurements to account for overfitting, and internally validate the final model of each work. Moreover, I highlight the importance of model and algorithm selection, and propose approaches to each. On the clinical side, based on feature selection strategies and techniques, as well as the innate interpretability of my models, I unveil clinically use-

ful information behind the underlying prognostic modelling. Indeed (as per my first premise), the underlying models identify prognostic patterns that would have been too hard to obtain otherwise, both due to the high dimensional nature of the data, and the highly–correlated covariates. A future modification of my framework in Chapters 3 and Chapters 4 would be the use of multitask survival analysis, which is a generalization over my current approach in that binary classifiers trained on multiple cutoffs are combined so that time–to–event is not lost completely, while still using as much censored data in each cutoff problem as possible [70, 195].

The second achievement of my thesis is the successful application of deep learning on whole slide images with minimum ground truth. It is important to differentiate from my works above wherein, although minimum ground truth is used in the optimization of the prognostic models, the feature extraction process involved training models for low–level tasks, such as nuclei identification and tissue segmentation, based on high precision annotations from domain experts. Instead, deep learning methods are able to learn directly from the high–dimensional ($> 100,000$ dimensions) space of images, i.e. the "what" problem can be solved end–to–end. However, WSIs are not your typical images, as they often reside in a space of billions of dimensions, far exceeding the capabilities of the current state of hardware (not to mention the scalability issues that will emerge in the optimization of the network) for end–to–end analysis. Therefore, a fundamental aspect of tissue image analysis is the approach taken to address the "where" problem. A solution to this problem is non–trivial due to the underlying complexity of the disease. The heterogeneity of cancerous cells from the same tumour, between tumours or different sites of a patient, and between patients of the same stage is indeed staggering. Fields like molecular immuno–oncology, a field that aims to provide better therapies based on cell–level molecular analysis, are coming in terms with the realization that a single (or even a dozen) cell(s) cannot plausibly represent the state of the disease. Besides, the microenvironment surrounding tumour cells is equally important. For example, tumour cells surrounded by a large number of prominent intra-tumoural and peri-tumoural TILs and M1 macrophages have been related to better prognosis in several types of cancer [144], whereas a high content of M2 macrophages and TBs has been associated with poorer outcome [214]. In Chapter 5, I construct a multi–step pipeline that is able to address both the "where" and "what" problems effectively, and with low–granularity labels (patient–level).

In particular, in an attempt to introduce a way to filter out unnecessary parts of the WSI, while remaining within the weakly supervised paradigm, and requiring no manual input, clustering is explored given exhaustively tiled WSIs. However, clustering approaches do not work well with images due to their high–dimensionality. Two types of image descriptors are introduced; one to capture the morphological heterogeneity, and one to capture the colour heterogeneity of the image patches. A CNN

is trained with each cluster, and based on the training performance at patient–level, clusters with relevant information are annotated as discriminative. Since the initial WSI was partitioned, in order to produce a decision at the patient–level, patch–level decisions need to be aggregated. I reformulate the problem into a binary classification problem (a method I also adopted in Chapters 3 and 4), thus enabling the use of machine learning algorithms. By accumulating the patch–level CNN output of the filtered (i.e. ones that were in discriminative clusters) patches of a WSI into a normalized histogram, a SVM could be trained directly with the patient–level labels. Following internal validation (on a hold–out testing set), I demonstrate that the proposed method is able to extract and learn salient, discriminative, and clinically meaningful content from a real–world data set. There was one fundamental limitation to this method, however, which I sought to improve in my work in Chapter 6. In particular, the initial step of exhaustive tiling assumes that information is only available at a single magnification level. Moreover, there is also a loss of spatial information effected by the separation of patches and the arbitrariness of the loci of patch boundaries with respect to the imaged physical tissue.

In Chapter 6, I introduced a new type of CNN–based method, called MagNet, for gigapixel image analysis that does not require WSIs to be patch–based preprocessed. Instead, the network starting from the lowest magnification level, can dynamically explore a WSI in continually higher magnification levels, as seen fit by MagNet and in the visual context of the specific WSI. The premise was that in this way, the patches would be dynamically extracted with the best magnification level, field–of–view, and location according to the optimizing task, and not based on generic, predefined or static ways. Indeed for the task of breast cancer metastasis detection, my MagNet models performed extremely well given that they only had to process from 28 to 64 image patches per WSI, far less than any of the existing approaches (10 to 300 times less). However, the recurrent application of attention layers, especially ones which are partly differentiable and weakly supervised, is by no means an easy process to optimize. In fact, the quality of gradients that pass through billinear sampling was shown to decrease catastrophically when using increasingly more downsampled versions of an image. Therefore, an optimization framework is also proposed, under which multi–layer MagNets can be successfully optimized. Given the innate transparency of the hard attention in MagNets, no patch–based preprocessing requirements, and ability to perform both localization and classification tasks with low–granularity labels, MagNets offer an attractive alternative for gigapixel image analysis. Further validation of the method is paramound; the effectiveness of MagNets can be explored on different WSI data sets, and under different clinical settings. I am particularly interested in exploring the ability of MagNets to provide patient prognosis next, a clinical problem that was the main focus of my first three works. In particular, the cohort of Chapter 5 would provide

a challenging next direction for MagNets. Since the patients can have a couple of WSIs, an aggregation step would be needed for fusing the slide-level decision into patient-level. Another interesting direction would be the optimization of MagNets on the task of mutation prediction [168].

Appendix A

Appendix A

This is the appendix of Chapter 3.

Table A.1: Summary table of survival for the groups of the final KNN model on the 5–year testing cohort.

Low risk patients								
Months	0	20	40	60	80	100	120	140
At risk	37	35	34	33	30	25	20	1
Censored	0	0	0	0	3	8	13	32
Events	0	2	3	4	4	4	4	4
High risk patients								
Months	0	20	40	60	80	100	120	140
At risk	7	5	5	4	3	2	2	1
Censored	0	0	0	0	1	2	2	3
Events	0	2	2	3	3	3	3	3

Table A.2: Summary table of survival for the groups of the final KNN model on the 10–year testing cohort.

Low risk patients								
Months	0	20	40	60	80	100	120	140
At risk	21	21	21	21	21	21	21	1
Censored	0	0	0	0	0	0	0	20
Events	0	0	0	0	0	0	0	0
High risk patients								
Months	0	20	40	60	80	100	120	140
At risk	15	13	12	9	8	7	4	0
Censored	0	0	0	0	0	0	0	4
Events	0	2	3	6	7	8	11	11

Table A.3: A list of the 123 image analysis-based features imported into the machine learning workflow for Chapter 3. *ck* = pancytokeratin, *pdc* = poorly differetiated clusters, *lvi* = lymphatic vessel invasion.

Number of vessels (sum)
Sum area of vessels (sum)
Lymphatic vessel density (mean)
Mean intensity of d240 in vessels (mean)
Number of tumour buds (sum)
Sum area of tumour buds (sum)
Tumour bud density (mean)
Number of tumour buds containing partial nuclei (sum)
Nuclei in tumour bud_asymmetry (mean)
Nuclei in tumour bud_border index (mean)
Nuclei in tumour bud_meanck intensity (mean)
Nuclei in tumour bud_meand240 intensity (mean)
Nuclei in tumour bud_meandapi intensity (mean)
Nuclei in tumour bud_mean_area (mean)
Nuclei in tumour bud_mean_ellipticity (mean)
Nuclei in tumour bud_mean_ratio of ck intensity (mean)
Nuclei in tumour bud_mean_ratio of d240 intesnity (mean)
Nuclei in tumour bud_mean_ratio of dapi intensity (mean)
Nuclei in tumour bud_mean_roundness (mean)
Nuclei in tumour bud_mean_shape index (mean)
Nuclei in tumour bud_mean_borderlength (mean)
Nuclei in tumour bud_mean_circularity (mean)
Nuclei in tumour bud_mean_compactness (mean)
Nuclei in tumour bud_mean_density (mean)
Nuclei in tumour bud_mean_length (mean)
Nuclei in tumour bud_mean_lengthwidth (mean)
Nuclei in tumour bud_mean_pixel texture of ck intensity (mean)
Nuclei in tumour bud_mean_pixel texture of d240 intensity (mean)
Nuclei in tumour bud_mean_pixel texture of dapi intensity (mean)
Nuclei in tumour bud_mean_width (mean)
Nuclei in stroma_asymmetry (mean)
Nuclei in stroma_border index (mean)
Nuclei in stroma_meanck intensity (mean)
Nuclei in stroma_meand240 intensity (mean)
Nuclei in stroma_meandapi intensity (mean)
Nuclei in stroma_mean_area (mean)
Nuclei in stroma_mean_ellipticity (mean)
Nuclei in stroma_mean_ratio of ck intensity (mean)
Nuclei in stroma_mean_ratio of d240 intensity (mean)
Nuclei in stroma_mean_ratio of dapi intensity (mean)
Nuclei in stroma_mean_roundness (mean)
Nuclei in stroma_mean_shape index (mean)

Nuclei in stroma_mean_borderlength (mean)
Nuclei in stroma_mean_circularity (mean)
Nuclei in stroma_mean_compactness (mean)
Nuclei in stroma_mean_density (mean)
Nuclei in stroma_mean_length (mean)
Nuclei in stroma_mean_lengthwidth (mean)
Nuclei in stroma_mean_pixel texture of ck intensity (mean)
Nuclei in stroma_mean_pixel texture of d240 intensity (mean)
Nuclei in stroma_mean_pixel texture of dapi intensity (mean)
Nuclei in stroma_mean_width (mean)
Nuclei in tumour_asymmetry (mean)
Nuclei in tumour_border index (mean)
Nuclei in tumour_meanck intensity (mean)
Nuclei in tumour_meand240 intensity (mean)
Nuclei in tumour_meandapi intensity (mean)
Nuclei in tumour_mean_area (mean)
Nuclei in tumour_mean_ellipticity (mean)
Nuclei in tumour_mean_ratio of ck intensity (mean)
Nuclei in tumour_mean_ratio of d240 intensity (mean)
Nuclei in tumour_mean_ratio of dapi intensity (mean)
Nuclei in tumour_mean_roundness (mean)
Nuclei in tumour_mean_shape index (mean)
Nuclei in tumour_mean_borderlength (mean)
Nuclei in tumour_mean_circularity (mean)
Nuclei in tumour_mean_compactness (mean)
Nuclei in tumour_mean_density (mean)
Nuclei in tumour_mean_length (mean)
Nuclei in tumour_mean_lengthwidth (mean)
Nuclei in tumour_mean_pixel texture of ck intensity (mean)
Nuclei in tumour_mean_pixe texture of d240 intensity (mean)
Nuclei in tumour_mean_pixel texture of dapi intensity (mean)
Nuclei in tumour_mean_width (mean)
Tumour bud with ≤ 2 nuclei (sum)
Tumour bud with $>= 3$ nuclei (sum)
Number of pdcs (sum)
Number of ck objects with no associated nuclei (sum)
Area of tumour bud with only partial nuclei (sum)
Area of pdcs (sum)
Area of ck objects with no associated nuclei (sum)
Mean ck intensity in tumour buds (mean)
Mean ck intensity in tumour bud with debris nuclei (mean)
Mean ck intensity of ck objects with no associated nuclei (mean)
Mean ck intensity of pdcs (mean)
Area of tumour bud invading a vessel (sum)
Area of pdc invading a vessel (sum)

Number of bordering tumour bud and vessel (sum)
Number of bordering pdcs and vessels (sum)
Number of lvi events by tumour buds (sum)
Number of lvi events by pdcs (sum)
Number of vessels bordering a tumour bud lvi event (sum)
Area of vessel border to a tumour bud lvi event (sum)
Area of vessel border to a pdc event (sum)
No. of vessels bordered to tumour glands (sum)
Area of vessels bordering to tumour glands (sum)
Area of tumour glands invading vessels(sum)
Number of tumour gland lvi events (sum)
Necrosis relative area (%) (mean)
Stroma relative area (%) (mean)
Tumour gland relative area (%) (mean)
Average ck intensity (stroma area) (mean)
Average ck intensity (tumour area) (mean)
Average dapi intensity (stroma area) (mean)
Average dapi intensity (tumour area) (mean)
Mean border index of tumour glands (mean)
Mean area (µm^2) of tumour glands (mean)
Mean roundness of tumour glands (mean)
Mean compactness of tumour glands (mean)
Mean pixel texture of ck in tumour glands (mean)
Mean pixel texture of d240 in tumour glands (mean)
Mean pixel texture of dapi in tumour glands (mean)
Mean border to stroma (µm) of tumour glands (mean)
Mean length/width of tumour glands (mean)
Mean length (µm) of tumour glands (mean)
Mean rectangular fit of tumour glands (mean)
Mean rel. border to stroma of tumour glands (mean)
Mean elliptic fit of tumour glands (mean)
Mean asymmetry of tumour glands (mean)
Mean width (µm) of tumour glands (mean)
Mean circularity of tumour glands (mean)
Mean ellipticity of tumour glands (mean)
Mean perimeter of tumour glands (mean)

```
def SFFS(X, y, cls, cv_folds, n_repeats, rs, delta):
    """
    X: Data with (M, F) dimensions
    y: Labels (M) dimensions
    cls: Classifier
    cv_folds: Number of folds for cross-validation
    n_repeats: Number of repeats of cross-validation
    rs: Random seed for reproducibility
    delta: Maximum number of features to use
    """
    feature_scores = {} # To store best score for selected features
                        # to delta features) along with their indices
    feature_subset = [] # Indices of features currently in use
    pass_step_2 = True
    k = 0 # Number of features in use
    while k < delta:
        # Step 1
        feature_index_FS, mean_FS, std_FS = \
            SFS(X, y, feature_subset, cls, cv_folds, n_repeats, rs)
        k += 1 # Increment number of selected features
        feature_subset.append(feature_index_FS) # Add index of new feature
        feature_scores[k] = (mean_FS, std_FS, feature_subset[:])
        pass_step_2 = False
        # Step 2
        if k > 1:
            feature_index_BS, score_BS, std_BS = \
                SBS(X, y, feature_subset, cls, cv_folds, n_repeats, rs)
            # To recursively remove more features (Step 3):
            # (a) feature selected by SBS != feature selected by SFS above
            # (b) score of new combination of features surpassed previous
            if not (
                feature_subset[feature_index_BS] == feature_index_FS \
                or score_BS <= feature_scores[k - 1][0]
                ):
                del feature_subset[feature_index_BS]
                feature_scores[k - 1] = (score_BS, std_BS, feature_subset[:])
                k -= 1 # Decrement number of selected features
                pass_step_2 = True
        # Step 3
        while k > 2 and pass_step_2:
            feature_index_BS, score_BS, std_BS = \
                SBS(X, y, feature_subset, cls, cv_folds, n_repeats, rs)
            # If new score for this number of selected features is worse
            # than previous, then stop recursively removing features
            if score_BS <= feature_scores[k-1][0]:
                break
            else:
                del feature_subset[feature_index_BS]
                feature_scores[k - 1] = (score_BS, std_BS, feature_subset[:])
                k -= 1
    return feature_scores
```

Figure A.1: Python code for SFFS.

```
sklearn.model_selection import *
def SFS(X, y, feature_subset, cls, cv_folds, n_repeats, rs):
    """
    X: Data with (M, F) dimensions
    y: Labels (M) dimensions
    feature_subset: Subset of features to use
    cls: Classifier
    cv_folds: Number of folds for cross-validation
    n_repeats: Number of repeats of cross-validation
    rs: Random seed for reproducibility
    """
    baseline_roc_mean = 0
    baseline_roc_sem = 0
    to_add_feature = None
    # Get the scores of adding each remaining feature independently
    for feature_index in range(len(X.columns)):
        # Make sure selected feature has not already been selected
        if feature_index in feature_subset:
            continue
        else:
            # Use repeated stratified cross validation
            shuffle = RepeatedStratifiedKFold(cv_folds, n_repeats, rs)
            # Train and evaluate given classifier only on the features selected
            # so far plus the new feature to be considered.
            score = cross_val_score(
                cls, X.iloc[:, feature_subset + [feature_index]], y,
                cv=shuffle, scoring='roc_auc', n_jobs=1
            )
            if (score.mean() > baseline_roc_mean):
                to_add_feature = feature_index
                baseline_roc_mean = score.mean()
                baseline_roc_sem = scipy.stats.sem(score)
    return (to_add_feature, baseline_roc_mean, baseline_roc_sem)
```

Figure A.2: Python code for SFS.

```
def filter_n_years_prognosis(data, n):
    # Remove patients that died before n years due to a non CRC-related cause.
    data = data[(data["SurvivalMonths"] > n * 12) | (data["EventOccurred"] == 1)]
    # Change response variable to 0 if patient didn't die due to CRC in n years.
    data["EventOccurred"][data["SurvivalMonths"] > n * 12] = 0
    return data
```

Figure A.3: The function for binarising prognosis for a given pandas dataframe *data* and a specified cutoff n.

```
from sklearn.model_selection import *
def objective(params):
    local_X = X.copy()
    cls = params["classifier"]
    shuffle = RepeatedStratifiedKFold(n_folds, n_repeats, rs)
    roc_auc_score = cross_val_score(cls, local_X, y, cv=shuffle,
                                scoring='roc_auc', n_jobs=1)
    return {
        'loss': 1 - numpy.array(roc_auc_score).mean(),
        'status': STATUS_OK
    }
```

Figure A.4: Objective function for hyperparameter tuning using Hyperopt [20].

Appendix B

Appendix B

This is the appendix of Chapter 4.

Table B.1: A list of the image features in Chapter 4.

area_Core_SerialSection2
area_FrontIn_SerialSection2
area_FrontOut_SerialSection2
number_of_CD68p_Core_SerialSection2
number_of_CD68p_FrontIn_SerialSection2
number_of_CD68p_FrontOut_SerialSection2
number_of_CD163p_Core_SerialSection2
number_of_CD163p_FrontIn_SerialSection2
number_of_CD163p_FrontOut_SerialSection2
number_of_CD163nCD68p_Core_SerialSection2
number_of_CD163nCD68p_FrontIn_SerialSection2
number_of_CD163nCD68p_FrontOut_SerialSection2
number_of_CD163pCD68p_Core_SerialSection2
number_of_CD163pCD68p_FrontIn_SerialSection2
number_of_CD163pCD68p_FrontOut_SerialSection2
number_of_PDL1pCK_Core_SerialSection2
number_of_PDL1pCK_FrontIn_SerialSection2
number_of_PDL1pCK_FrontOut_SerialSection2
number_of_PDL1pNonCK_Core_SerialSection2
number_of_PDL1pNonCK_FrontIn_SerialSection2
number_of_PDL1pNonCK_FrontOut_SerialSection2
number_of_PDL1p_Core_SerialSection2
number_of_PDL1p_FrontIn_SerialSection2
number_of_PDL1p_FrontOut_SerialSection2
number_of_PDL1pCD163nCD68p_Core_SerialSection2
number_of_PDL1pCD163nCD68p_FrontIn_SerialSection2
number_of_PDL1pCD163nCD68p_FrontOut_SerialSection2
number_of_PDL1pCD163pCD68p_Core_SerialSection2
number_of_PDL1pCD163pCD68p_FrontIn_SerialSection2

number_of_PDL1pCD163pCD68p_FrontOut_SerialSection2
number_of_PDL1pCD163p_Core_SerialSection2
number_of_PDL1pCD163p_FrontIn_SerialSection2
number_of_PDL1pCD163p_FrontOut_SerialSection2
number_of_PDL1pCD68p_Core_SerialSection2
number_of_PDL1pCD68p_FrontIn_SerialSection2
number_of_PDL1pCD68p_FrontOut_SerialSection2
Density_of_CD163p_Core_SerialSection2
Density_of_CD163p_FrontIn_SerialSection2
Density_of_CD163p_FrontOut_SerialSection2
Density_of_CD68p_Core_SerialSection2
Density_of_CD68p_FrontIn_SerialSection2
Density_of_CD68p_FrontOut_SerialSection2
number_of_TBsmall_Core_SerialSection2
number_of_TBsmall_FrontIn_SerialSection2
number_of_TBsmall_FrontOut_SerialSection2
Density_of_TBsmall_Core_SerialSection2
Density_of_TBsmall_FrontIn_SerialSection2
Density_of_TBsmall_FrontOut_SerialSection2
number_of_NucCK_Core_SerialSection2
number_of_NucCK_FrontIn_SerialSection2
number_of_NucCK_FrontOut_SerialSection2
number_of_NucNonCK_Core_SerialSection2
number_of_NucNonCK_FrontIn_SerialSection2
number_of_NucNonCK_FrontOut_SerialSection2
number_of_Nuc_Core_SerialSection2
number_of_Nuc_FrontIn_SerialSection2
number_of_Nuc_FrontOut_SerialSection2
Density_of_PDL1p_Core_SerialSection2
Density_of_PDL1p_FrontIn_SerialSection2
Density_of_PDL1p_FrontOut_SerialSection2
Density_of_PDL1pCK_Core_SerialSection2
Density_of_PDL1pCK_FrontIn_SerialSection2
Density_of_PDL1pCK_FrontOut_SerialSection2
Density_of_PDL1pNonCK_Core_SerialSection2
Density_of_PDL1pNonCK_FrontIn_SerialSection2
Density_of_PDL1pNonCK_FrontOut_SerialSection2
area_Core_SerialSection1
area_FrontIn_SerialSection1
area_FrontOut_SerialSection1
number_of_CD3p_Core_SerialSection1
number_of_CD3p_FrontIn_SerialSection1
number_of_CD3p_FrontOut_SerialSection1
number_of_CD8p_Core_SerialSection1
number_of_CD8p_FrontIn_SerialSection1

number_of_CD8p_FrontOut_SerialSection1
number_of_PDL1p_Core_SerialSection1
number_of_PDL1p_FrontIn_SerialSection1
number_of_PDL1p_FrontOut_SerialSection1
number_of_TBsmall_Core_SerialSection1
number_of_TBsmall_FrontIn_SerialSection1
number_of_TBsmall_FrontOut_SerialSection1
number_of_CD3pPDL1p_Core_SerialSection1
number_of_CD3pPDL1p_FrontIn_SerialSection1
number_of_CD3pPDL1p_FrontOut_SerialSection1
number_of_CD8pPDL1p_Core_SerialSection1
number_of_CD8pPDL1p_FrontIn_SerialSection1
number_of_CD8pPDL1p_FrontOut_SerialSection1
number_of_PDL1pCK_Core_SerialSection1
number_of_PDL1pCK_FrontIn_SerialSection1
number_of_PDL1pCK_FrontOut_SerialSection1
number_of_PDL1pNonCK_Core_SerialSection1
number_of_PDL1pNonCK_FrontIn_SerialSection1
number_of_PDL1pNonCK_FrontOut_SerialSection1
Density_of_CD3p_Core_SerialSection1
Density_of_CD3p_FrontIn_SerialSection1
Density_of_CD3p_FrontOut_SerialSection1
Density_of_CD8p_Core_SerialSection1
Density_of_CD8p_FrontIn_SerialSection1
Density_of_CD8p_FrontOut_SerialSection1
Density_of_TBsmall_Core_SerialSection1
Density_of_TBsmall_FrontIn_SerialSection1
Density_of_TBsmall_FrontOut_SerialSection1
number_of_NucCK_Core_SerialSection1
number_of_NucCK_FrontIn_SerialSection1
number_of_NucCK_FrontOut_SerialSection1
number_of_NucNonCK_Core_SerialSection1
number_of_NucNonCK_FrontIn_SerialSection1
number_of_NucNonCK_FrontOut_SerialSection1
number_of_Nuc_Core_SerialSection1
number_of_Nuc_FrontIn_SerialSection1
number_of_Nuc_FrontOut_SerialSection1
number_of_CD3pCD8nPDL1p_Core_SerialSection1
number_of_CD3pCD8nPDL1p_FrontIn_SerialSection1
number_of_CD3pCD8nPDL1p_FrontOut_SerialSection1
number_of_CD3pCD8n_Core_SerialSection1
number_of_CD3pCD8n_FrontIn_SerialSection1
number_of_CD3pCD8n_FrontOut_SerialSection1
Density_of_PDL1p_Core_SerialSection1
Density_of_PDL1p_FrontIn_SerialSection1

Density_of_PDL1p_FrontOut_SerialSection1
Density_of_PDL1pCK_Core_SerialSection1
Density_of_PDL1pCK_FrontIn_SerialSection1
Density_of_PDL1pCK_FrontOut_SerialSection1
Density_of_PDL1pNonCK_Core_SerialSection1
Density_of_PDL1pNonCK_FrontIn_SerialSection1
Density_of_PDL1pNonCK_FrontOut_SerialSection1

Table B.2: A list of the spatial features in Chapter 4.

L_function_r20_TB_CD68
L_function_r20_TB_CD163
L_function_r20_TB_PDL1_SerialSection2
L_function_r50_TB_CD68
L_function_r50_TB_CD163
L_function_r50_TB_PDL1_SerialSection2
L_function_r100_TB_CD68
L_function_r100_TB_CD163
L_function_r100_TB_PDL1_SerialSection2
L_function_r150_TB_CD68
L_function_r150_TB_CD163
L_function_r150_TB_PDL1_SerialSection2
L_function_r200_TB_CD68
L_function_r200_TB_CD163
L_function_r200_TB_PDL1_SerialSection2
L_function_r250_TB_CD68
L_function_r250_TB_CD163
L_function_r250_TB_PDL1_SerialSection2
L_function_r20_CD68_PDL1
L_function_r20_CD163_PDL1
L_function_r50_CD68_PDL1
L_function_r50_CD163_PDL1
L_function_r100_CD68_PDL1
L_function_r100_CD163_PDL1
L_function_r150_CD68_PDL1
L_function_r150_CD163_PDL1
L_function_r200_CD68_PDL1
L_function_r200_CD163_PDL1
L_function_r250_CD68_PDL1
L_function_r250_CD163_PDL1
L_function_r20_TB_CD3
L_function_r20_TB_CD8
L_function_r20_TB_PDL1_SerialSection1
L_function_r50_TB_CD3
L_function_r50_TB_CD8

L_function_r50_TB_PDL1_SerialSection1
L_function_r100_TB_CD3
L_function_r100_TB_CD8
L_function_r100_TB_PDL1_SerialSection1
L_function_r150_TB_CD3
L_function_r150_TB_CD8
L_function_r150_TB_PDL1_SerialSection1
L_function_r200_TB_CD3
L_function_r200_TB_CD8
L_function_r200_TB_PDL1_SerialSection1
L_function_r250_TB_CD3
L_function_r250_TB_CD8
L_function_r250_TB_PDL1_SerialSection1
L_function_r20_CD3_PDL1
L_function_r20_CD8_PDL1
L_function_r50_CD3_PDL1
L_function_r50_CD8_PDL1
L_function_r100_CD3_PDL1
L_function_r100_CD8_PDL1
L_function_r150_CD3_PDL1
L_function_r150_CD8_PDL1
L_function_r200_CD3_PDL1
L_function_r200_CD8_PDL1
L_function_r250_CD3_PDL1
L_function_r250_CD8_PDL1

Table B.3: A list of the clinical features in Chapter 4..

Age
Gender_Female
Gender_Male
T_T2
T_T3
T_T4
N_N0
N_N1
N_N2
M_No
M_Yes
TNM_II
TNM_IIIA
TNM_IIIB
TNM_IV

Appendix C

Appendix C

This is the appendix of Chapter 5.

```
# One random trivial transformation
myTransforms.RandomChoice([myTransforms.RandomHorizontalFlip(p=1),
                           myTransforms.RandomVerticalFlip(p=1),
                           myTransforms.AutoRandomRotation()]),
# One random non-trivial transformation
myTransforms.RandomChoice([myTransforms.RandomElastic(alpha=2, sigma=0.06),
                           myTransforms.RandomAffineCV2(alpha=0.1),
                           myTransforms.RandomGaussBlur(radius=[0.5, 1.5])]),
# One colour transformation
myTransforms.ColorJitter(brightness=(0.65, 1.35), contrast=(0.5, 1.5)),
myTransforms.RandomChoice([myTransforms.ColorJitter(saturation=(0, 2), hue=0.3),
                           myTransforms.HEDJitter(theta=0.05)])
```

Figure C.1: Data augmentation during training using "myTransforms" as implemented at https://github.com/gatsby2016/Augmentation-PyTorch-Transforms.

```
 # Normalize staining using a target WSI
reader = bioformats.ImageReader(target_path)
reader.rdr.setId(target_path)
# Accessing the 4th layer (1/16 of the original resolution)
reader.rdr.setSeries(downsampling_plane=4)
WSI_target = reader.read(z=0, rescale=False)

normalizer = staintools.ReinhardColorNormalizer()
normalizer.fit(WSI_target)

reader = bioformats.ImageReader(input_path)
reader.rdr.setId(input_path)
# Accessing the 4th layer (1/16 of the original)
reader.rdr.setSeries(downsampling_plane=4)
WSI_input = reader.read(z=0, rescale=False)

WSI_input_norm = normalizer.transform(whole_image)
```

Figure C.2: Code for normalizing a WSI using staintools' Reinhard normalization.

```
from sklearn.cluster import KMeans
# Different random seeds are tested - lowest loss selected
rs = 0
# Sample 20% of the training cohort of patients (stratified)
_X_train_subset = X_train.groupby(
    y_train["CRC death"], group_keys=False
).apply(
    lambda x: x.sample(
        frac=0.2,
        random_state=rs
    )
)
# Get the WSIs of these patients
train_imgs_subset = [
    img for _list in
        X_train_subset["FilePath"].tolist() for img in _list
]
# For each WSI, iterate over the image patches, and get inverse of
# data compression rate (see main text for definition)
for _path in train_imgs_subset:
    _path = _path.split(".")[0] + "_normalized"
    _cur_patches = glob.glob(_path + "/*.png", recursive=True)
    for _cur_patch in _cur_patches:
        _train_subset_sizes.append(os.path.getsize(_cur_patch)/150528.0)
random.shuffle(_train_subset_sizes)

kmeans = KMeans(init="k-means++", n_clusters=3, n_init=10, random_state=rs)
kmeans.fit(np.array(_train_subset_sizes).reshape(-1, 1)) # Training
... kmeans.predict(np.array(_cur_img_patches).reshape(-1, 1)) # Clustering
# For all patches of each WSI, assign their clustering labels into spreadsheets.
```

Figure C.3: Implementation of the information density clustering approach.

```
from sklearn.cluster import KMeans
from torchvision.models import vgg16_bn
from sklearn.decomposition import PCA

pretrained_CNN = vgg16_bn(pretrained=True).cuda()
pca = PCA(n_components=50)
# Different random seeds are tested - lowest loss selected
rs = 0
# Sample 20% of the training cohort of patients (stratified)
X_train_subset = X_train.groupby(
    y_train["CRC death"], group_keys=False
).apply(
    lambda x: x.sample(
        frac=0.2,
        random_state=rs
    )
)
# Get the WSIs of these patients
train_imgs_subset = [
    img for _list in
        X_train_subset["FilePath"].tolist() for img in _list
]
# Iterate over all image patches and extract features using pretrained_CNN
# Further reduce the dimensionality using PCA
 all_img_patches = pca.fit_transform(all_img_patches)
kmeans1 = KMeans(init="k-means++", n_clusters=5, n_init=10, random_state=0)
kmeans2 = KMeans(init="k-means++", n_clusters=10, n_init=10, random_state=0)
# Training
kmeans1.fit(all_img_patches)
kmeans2.fit(all_img_patches)

# PH_5 Clustering
... kmeans1.predict(np.array(_cur_img_patches).reshape(-1, 1))
# PH_10 Clustering
... kmeans2.predict(np.array(_cur_img_patches).reshape(-1, 1))
# For all patches of each WSI, assign their clustering labels into spreadsheets.
```

Figure C.4: Implementation of the phenotype clustering approach.

```
from torch_lr_finder import LRFinder
optimizer = torch.optim.SGD(model.parameters(), lr=0.00001, weight_decay=0)
lr_finder = LRFinder(model, optimizer, criterion, device="cuda")
lr_finder.range_test(trainloader, end_lr=1, num_iter=100)
```

Figure C.5: Implementation of the learning rate range test using the package at https://github.com/davidtvs/pytorch-lr-finder.

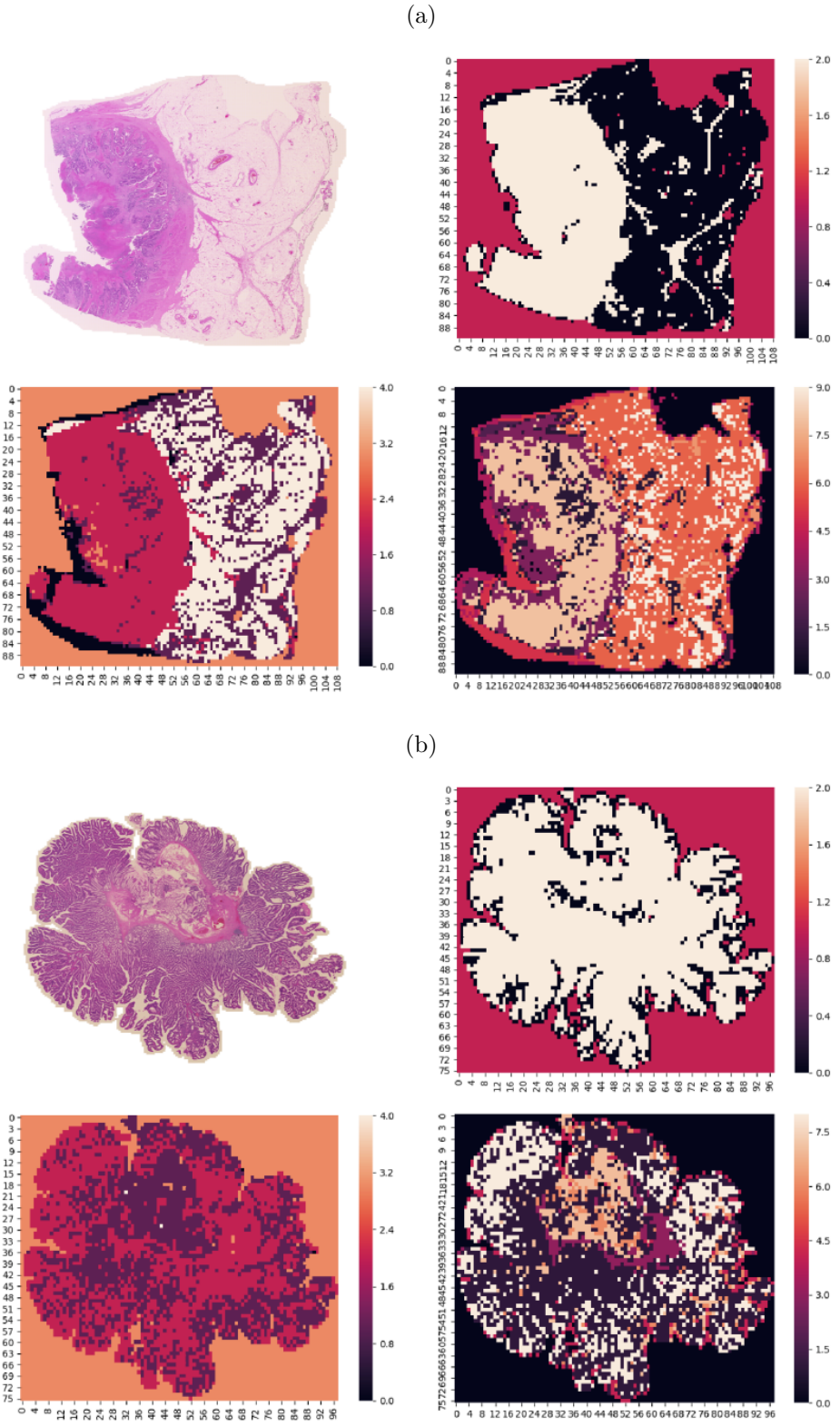


Figure C.6: Examples of clustering results from the early version of my work. Each square block of four images comprises the original WSI (top left), and the corresponding information density (top right) and phenotype based (k = 5 and k = 10 respectively bottom left and bottom right) clustering labels, colour coded.

Appendix D

Appendix D

This is the appendix of Chapter 6.

```
import torch
import torch.nn as nn
import torch.nn.functional as F

class BasicConv2d(nn.Module):
    def __init__(self, in_channels, out_channels,
                 last_layer=False, **kwargs):
        super(BasicConv2d, self).__init__()
        self.conv = nn.Conv2d(
            in_channels, out_channels, bias=False, **kwargs
        )
        if not last_layer:
            self.bn2 = nn.BatchNorm2d(out_channels, eps=0.00001)
        self.last_layer=last_layer

    def forward(self, x):
        x = self.conv(x)
        if not self.last_layer:
            x = self.bn2(x)
        return F.relu(x, inplace=True)
```

Figure D.1: PyTorch implementation of "Conv2D" as described in Chapter 6.

```
class Branch2d(nn.Module):
    def __init__(self, in_channels, features,
                 out_channels, **kwargs):
        super(Branch2d, self).__init__()
        self.layer1 = BasicConv2d(
            in_channels, features, kernel_size=1, **kwargs)
        self.layer2 = nn.Sequential(
            BasicConv2d(in_channels, max(int(3/4 * features), 1),
                        kernel_size=1, **kwargs),
            BasicConv2d(max(int(3/4 * features), 1), features,
                        kernel_size=3, padding=1, **kwargs)
        )
        self.layer3 = nn.Sequential(
            BasicConv2d(in_channels, max(int(3/4 * features), 1),
                        kernel_size=1, **kwargs),
            BasicConv2d(max(int(3/4 * features), 1), features,
                        kernel_size=3, padding=1, **kwargs),
            BasicConv2d(features, features,
                        kernel_size=3, padding=1, **kwargs)
        )
        self.layer4 = nn.Sequential(
            nn.MaxPool2d(kernel_size=3, stride=1,
                         padding=1, ceil_mode=True),
            BasicConv2d(in_channels, features,
                        kernel_size=1, **kwargs)
        )
        self.layer5 = nn.Sequential(
           BasicConv2d(in_channels, max(int(3 / 4 * features), 1),
                       kernel_size=1, **kwargs),
           BasicConv2d(max(int(3 / 4 * features), 1), features,
                       kernel_size=3, padding=1, **kwargs),
           BasicConv2d(features, features,
                       kernel_size=3, padding=1, **kwargs),
           BasicConv2d(features, features,
                       kernel_size=3, padding=1, **kwargs)
        )
        self.downsample = None
        if out_channels!=-1:
            self.downsample = BasicConv2d(features * 5, 1, kernel_size=1)

    def forward(self, x):
        layer1 = self.layer1(x)
        ...
        layer5 = self.layer5(x)
        outputs = [layer1, layer2, layer3, layer4, layer5]
        outputs = torch.cat(outputs, 1)
        return self.downsample(outputs) if self.downsample else outputs
```

Figure D.2: PyTorch implementation of a "Branch" as described in Chapter 6.

```
from torch.autograd import Variable
from entmax import entmax15 # sparsemax variant

class STN(nn.Module):
    ...
    def transform_feature_map(self, feature_map):
        B, C, Y, X = feature_map.size()
        # evenly spaced X numbers from -1 to 1 across Y axis
        grid_x = Variable(torch.as_tensor(np.linspace(-1, 1, X))
        # evenly spaced Y numbers from -1 to 1 across X axis
        grid_y = Variable(torch.as_tensor(np.linspace(-1, 1, Y))
        # Pass feature map through an entmax activation (sparsemax variant)
        sparsemax_map = entmax15(
            feature_map.view(B, C, -1), dim=-1
        ).view(B, C, Y, X)
        # Compute the translation affine parameters
        mean_x = (F.tanh(sparsemax_map.sum(-2)) * grid_x).sum(-1)
        mean_y = (F.tanh(sparsemax_map.sum(-1)) * grid_y).sum(-1)

        difference_x = (grid_x - torch.unsqueeze(mean_x, -1))
            .view(B, C, 1, X)
        difference_y = (grid_y - torch.unsqueeze(mean_y, -1))
            .view(B, C, Y, 1)
        # Compute the scale affine parameter by obtaining the expected L1 norm
        scale = = (
                (torch.abs(difference_x) + torch.abs(difference_y))
                * sparsemax_map
        ).sum(-1).sum(-1)
        # Avoid "out of bound" from having a scale that exceeds
        # the boundaries of the image
        scale = torch.min(
            scale,
            1 - (
                torch.max(torch.abs(mean_x), torch.abs(mean_y))
            )
        )
        # Constrain the scale between 0.05 and 1.0
        scale = torch.min(
            torch.max(scale, torch.Tensor([0.05]).cuda()),
            torch.Tensor([1.0]).cuda()
        )
        return mean_x, mean_y, scale # return affine parameters
    ...
```

Figure D.3: PyTorch implementation of inferring the affine parameters from a given feature map for the STN.

Bibliography

[1] F. Aeffner, K. Wilson, N. T. Martin, J. C. Black, C. L. L. Hendriks, B. Bolon, D. G. Rudmann, R. Gianani, S. R. Koegler, J. Krueger, and G. D. Young. The gold standard paradox in digital image analysis: Manual versus automated scoring as ground truth. *Archives of Pathology and Laboratory Medicine*, 141(9):1267–1275, May 2017. 11, 20

[2] F. Aeffner, M. Zarella, N. Buchbinder, M. Bui, M. Goodman, D. Hartman, G. Lujan, M. Molani, A. Parwani, K. Lillard, O. Turner, V. Vemuri, A. Yuil-Valdes, and D. Bowman. Introduction to digital image analysis in whole-slide imaging: A white paper from the digital pathology association. *Journal of Pathology Informatics*, 10(1):9, 2019. 11, 20

[3] T. J. Alhindi, S. Kalra, K. H. Ng, A. Afrin, and H. R. Tizhoosh. Comparing lbp, hog and deep features for classification of histopathology images. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018. 22

[4] American Joint Committee on Cancer. AJCC - Cancer Staging Manual. https://cancerstaging.org/references-tools/deskreferences/Pages/default.aspx. 50

[5] O. Arandjelović. Object matching using boundary descriptors. *In Proc. British Machine Vision Conference*, 2012. DOI: 10.5244/C.26.85. 22

[6] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, G. Fernandez, J. Zeineh, M. Kohl, C. Walz, F. Ludwig, S. Braunewell, M. Baust, Q. Dang Vu, M. Nguyen Nhat To, and P. Aguiar. Bach: Grand challenge on breast cancer histology images. *arXiv:1808.04277*, 2018. 26, 28

[7] M. Athelogou, G. Schmidt, A. Schäpe, M. Baatz, and G. Binnig. Cognition network technology–a novel multimodal image analysis technique for automatic identification and quantification of biological image contents. In *Imaging Cellular and Molecular Biological Functions*, pages 407–422. Springer, 2007. 55

[8] P. Bandi, O. Geessink, Q. Manson, M. van Dijk, M. Balkenhol, M. Hermsen, E. B. Bejnordi, B. Lee, K. Paeng, A. Zhong, Q. Li, G. F. Zanjani, S. Zinger, K. Fukuta, D. Komura, V. Ovtcharov, S. Cheng, S. Zeng, J. Thagaard, B. A. Dahl, H. Lin, H. Chen, L. Jacobsson, M. Hedlund, M. Cetin, E. Halici, H. Jackson, R. Chen, F. Both, J. Franke, H. Kusters-Vandevelde, W. Vreuls, P. Bult, B. van Ginneken, J. van der Laak, and G. Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019. 16, 26, 28, 90, 91

[9] T. A. Barnes and E. Amir. Hype or hope: the prognostic value of infiltrating immune cells in cancer. *British Journal of Cancer*, 117(4):451, 2017. 73

[10] C. Barone. Adjuvant chemotherapy of colon cancer current strategies. *European Journal of Cancer Supplements*, 6(14):60–63, October 2008. DOI: 10.1016/j.ejcsup.2008.06.024. 11, 32

[11] V. Barresi, L. R. Bonetti, A. Ieni, R. A. Caruso, and G. Tuccari. Poorly differentiated clusters: clinical impact in colorectal cancer. *Clinical Colorectal Cancer*, 16(1):9–15, March 2017. DOI: 10.1016/j.clcc.2016.06.002. 46

[12] A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. van de Vijver, R. B. West, M. van de Rijn, and D. Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine*, 3(108), Nov. 2011. 26

[13] R. Bellman. Adaptive control processes: A guided tour. (A RAND Corporation Research Study). Princeton, N. J.: Princeton University Press, XVI, 255 p. (1961)., 1961. 23

[14] A. BenTaieb and G. Hamarneh. Predicting cancer with a recurrent visual attention model for histopathology images. In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention*, pages 129–137. Springer International Publishing, 2018. 27

[15] K. Bera, I. Katz, and A. Madabhushi. Reimagining t staging through artificial intelligence and machine learning image processing approaches in digital pathology. *JCO Clinical Cancer Informatics*, (4):1039–1050, 2020. 13, 14, 22

[16] K. Bera, K. A. Schalper, D. L. Rimm, V. Velcheti, and A. Madabhushi. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology*, 16(11):703–715, 2019. 9, 13, 20, 22

[17] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, February 2012. 23, 36

[18] J. Bergstra, D. Yamins, and D. D. Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, volume 28, pages I–115–I–123, 2013. 23, 47, 73

[19] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2546–2554. Curran Associates, Inc., 2011. 36

[20] J. S. Bergstra, D. Yamins, and D. D. Cox. Hyperopt: a python library for optimizing the hyperparameters of machine learning algorithms., 2013. 119

[21] J. S. Bergstra, D. Yamins, and D. D. Cox. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 115–123, Atlanta, Georgia, USA, June 2013. PMLR. 43

[22] J. Besag. Contribution to the discussion on dr ripley's paper. *Journal of the Royal Statistical Society*, 39:193–195, 1977. 56

[23] H. P. Bhambhvani, A. Zamora, E. Shkolyar, K. Prado, D. R. Greenberg, A. M. Kasman, J. Liao, S. Shah, S. Srinivas, E. C. Skinner, and J. B. Shah. Development of robust artificial neural networks for prediction of 5-year survival in bladder cancer. *Urologic Oncology*, 39(3):193.e7–193.e12, 2021. 24

[24] A. Bhangu, G. Wood, A. Mirnezami, A. Darzi, P. Tekkis, and R. Goldin. Epithelial mesenchymal transition in colorectal cancer: seminal role in promoting disease progression and resistance to neoadjuvant therapy. *Surgical Oncology*, 21(4):316–323, December 2012. DOI: 10.1016/j.suronc.2012.08.003. 46

[25] G. Binnig, R. Huss, and G. Schmidt. *Tissue phenomics: Profiling cancer patients for treatment decisions*. CRC Press, 2018. 55

[26] S. Borhani, R. Borhani, and A. Kajdacsy-Balla. Artificial intelligence: A promising frontier in bladder cancer diagnosis and outcome prediction. *Critical Reviews in Oncology/Hematology*, 171:103601, 2022. 22, 24

[27] H. Brenner, M. Kloor, and C. P. Pox. Colorectal cancer. *The Lancet*, 383(9927):1490–1502, November 2013. DOI: 10.1016/S0140-6736(13)61649-9. 11

[28] J. D. Brierley, M. K. Gospodarowicz, and C. Wittekind, editors. *TNM classification of malignant tumours*. John Wiley & Sons, Nashville, TN, 8 edition, Dec. 2016. 31

[29] N. Brieu, C. G. Gavriel, D. J. Harrison, P. D. Caie, and G. Schmidt. Context-based interpolation of coarse deep learning prediction maps for the segmentation of fine structures in immunofluorescence images. In *Medical Imaging 2018: Digital Pathology*, volume 10581, page 105810P. International Society for Optics and Photonics, 2018. 28, 53

[30] N. Brieu, C. G. Gavriel, I. P. Nearchou, D. J. Harrison, G. Schmidt, and P. D. Caie. Automated tumour budding quantification by machine learning augments tnm staging in muscle-invasive bladder cancer prognosis. *Scientific Reports*, 9(1):5174, 2019. 50, 53, 73

[31] N. Brieu, O. Pauly, J. Zimmermann, G. Binnig, and G. Schmidt. Slide-specific models for segmentation of differently stained digital histopathology whole slide images. In *Medical Imaging 2016: Image Processing*, volume 9784, page 978410, 2016. 53

[32] N. Brieu and G. Schmidt. Learning size adaptive local maxima selection for robust nuclei detection in histopathology images. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 937–941, 2017. 53

[33] C. Cacchi, D. V. Arnholdt, H. Jahnig, M. Anthuber, A. Probst, D. V. Oruzio, and B. Markl. Clinical significance of lymph vessel density in t3 colorectal carcinoma. *International Journal of Colorectal Disease*, 27(6):721–726, January 2012. DOI: 10.1007/s00384-011-1373-7. 46

[34] P. Caie, N. Dimitriou, I. Nearchou, O. Arandjelovic, and D. Harrison. Artificial intelligence driving automated pathology: icaird and beyond. In *VIRCHOWS ARCHIV*, volume 475, pages S60–S60, 2019. 71

[35] P. D. Caie, N. Dimitriou, and O. Arandjelović. Chapter 8 - precision medicine in digital pathology via image analysis and machine learning. In S. Cohen, editor, *Artificial Intelligence and Deep Learning in Pathology*, pages 149–173. Elsevier, 2021. 9, 13, 18, 20, 21, 36

[36] P. D. Caie, A. K. Turnbull, S. M. Farrington, A. Oniscu, and D. J. Harrison. Quantification of tumour budding, lymphatic vessel density and invasion through image analysis in colorectal cancer. *Journal of Translational Medicine*, 12(1):156, June 2014. DOI: 10.1186/1479-5876-12-156. 32, 34

[37] P. D. Caie, Y. Zhou, A. K. Turnbull, A. Oniscu, and D. J. Harrison. Novel histopathologic feature identified through image analysis augments stage II colorectal cancer clinical reporting. *Oncotarget*, 7(28):44381–44394, July 2016. 32, 34

[38] A. Calon, E. Lonardo, A. Berenguer-Llergo, E. Espinet, X. Hernando-Momblona, M. Iglesias, M. Sevillano, S. Palomo-Ponce, D. V. F. Tauriello, D. Byrom, C. Cortina, C. Morral, C. Barceló, S. Tosi, A. Riera, C. S. Attolini, D. Rossell, E. Sancho, and E. Batlle. Stromal gene

expression defines poor-prognosis subtypes in colorectal cancer. *Nature Genetics*, 47(4):320–329, February 2015. DOI: 10.1038/ng.3225. 46

[39] G. Campanella, W. K. V. Silva, and J. T. Fuchs. Terabyte-scale deep multiple instance learning for classification and localization in pathology. *arXiv:1805.06983*, 2018. 27

[40] A. M. Carrington, P. W. Fieguth, H. Qazi, A. Holzinger, H. H. Chen, F. Mayr, and D. G. Manuel. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC Medical Informatics and Decision Making*, 20(1), Jan. 2020. 36

[41] J. L. Carstens, P. C. De Sampaio, D. Yang, S. Barua, H. Wang, A. Rao, J. P. Allison, V. S. LeBleu, and R. Kalluri. Spatial computation of intratumoral t cells correlates with survival of patients with pancreatic cancer. *Nature Communications*, 8:15095, 2017. 56

[42] J. Cassidy and A. Bruna. Tumor heterogeneity. In R. Uthamanthil and P. Tinkey, editors, *Patient Derived Tumor Xenograft Models*, pages 37–55. Academic Press, 2017. 13

[43] I. Chandler and R. S. Houlston. Interobserver agreement in grading of colorectal cancers—findings from a nationwide web-based survey of histopathologists. *Histopathology*, 52(4):494–499, 2008. 20

[44] G. Chandrashekar and F. Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, January 2014. DOI: 10.1016/j.compeleceng.2013.11.024. 37

[45] N. Chaput, M. Svrcek, A. Aupérin, C. Locher, F. Drusch, D. Malka, J. Taïeb, D. Goéré, M. Ducreux, and V. Boige. Tumour-infiltrating cd68+ and cd57+ cells predict patient outcome in stage ii–iii colorectal cancer. *British Journal of Cancer*, 109(4):1013–1022, 2013. 73

[46] L. Chen. Curse of dimensionality. In L. LIU and M. T. ÖZSU, editors, *Encyclopedia of Database Systems*, pages 545–546. Springer US, 2009. 23

[47] A. C. Chi, N. Katabi, H.-S. Chen, and Y.-S. L. Cheng. Interobserver variation among pathologists in evaluating perineural invasion for oral squamous cell carcinoma. *Head and Neck Pathology*, 10(4):451–464, 2016. 20

[48] P. Chikontwe, M. Kim, S. J. Nam, H. Go, and S. H. Park. Multiple instance learning with center embeddings for histopathology classification. In A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 519–528, Cham, 2020. Springer International Publishing. 27

[49] A. Clark. Pillow (pil fork) documentation, 2015. 87, 107

[50] B. Clarke and J.-H. Chu. Generic feature selection with short fat data. *Journal of the Indian Society of Agriculture Statistics*, 68(2):145–162, 2014. 22

[51] C. C. Compton. Optimal pathologic staging: defining stage II disease. *Clinical Cancer Research*, 13(22):6862s–6870s, November 2007. DOI: 10.1158/1078-0432.ccr-07-1398. 11, 31

[52] J.-B. Cordonnier, A. Mahendran, A. Dosovitskiy, D. Weissenborn, J. Uszkoreit, and T. Unterthiner. Differentiable patch selection for image recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. 27

[53] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005. 26

[54] C. Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40):1317, 2019. 47, 73

[55] O. Dehaene, A. Camara, O. Moindrot, A. de Lavergne, and P. Courtiol. Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology. *arXiv*, Dec. 2020. 27, 28

[56] N. Dimitriou and O. Arandjelovic. Magnifying Networks for Images with Billions of Pixels. *arXiv e-prints*, 2021. 18

[57] N. Dimitriou, O. Arandjelović, and P. D. Caie. Deep learning for whole slide image analysis: An overview. *Frontiers in Medicine*, 6:264, 2019. 12, 13, 14, 18, 21, 90

[58] N. Dimitriou, O. Arandjelović, D. J. Harrison, and P. D. Caie. A principled machine learning framework improves accuracy of stage II colorectal cancer prognosis. *npj Digital Medicine*, 2018. 13, 18, 21, 56

[59] K. Dunne, P. Cunningham, and F. Azuaje. Solutions to instability problems with sequential wrapper-based approaches to feature selection. Technical report, Journal of Machine Learning Research, 2002. 40

[60] S. B. Edge and C. C. Compton. The american joint committee on cancer: the 7th edition of the ajcc cancer staging manual and the future of tnm. *Annals of Surgical Oncology*, 17(6):1471–1474, 2010. DOI: 10.1245/s10434-010-0985-4. 10, 11

[61] B. B. Ehteshami, M. Veta, P. Johannes van Diest, and et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017. 26, 28, 91, 102

[62] J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. A. Tosteson, H. D. Nelson, M. S. Pepe, K. H. Allison, S. J.

Schnit, F. P. OMalley, and D. L. Weaver. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA*, 313(11):1122, Mar. 2015. 12, 20

[63] H. G. Eynard, E. A. Soria, E. Cuestas, R. A. Rovasio, and A. R. Eynard. Assessment of colorectal cancer prognosis through nuclear morphometry. *Journal of Surgical Research*, 154(2):345–348, June 2009. DOI: 10.1016/j.jss.2008.06.022. 46

[64] J. Fan and O. Arandjelović. Employing domain specific discriminative information to address inherent limitations of the LBP descriptor in face recognition. *In Proc. IEEE International Joint Conference on Neural Networks*, pages 3766–3772, 2018. 22

[65] K. Faryna, J. van der Laak, and G. Litjens. Tailoring automated data augmentation to h&e-stained histopathology. In *Medical Imaging with Deep Learning*. PMLR, 2021. 78

[66] Q. Feng, M. T. May, S. Ingle, M. Lu, Z. Yang, and J. Tang. Prognostic models for predicting overall survival in patients with primary gastric cancer: A systematic review. *Biomed Res. Int.*, 2019:5634598, 2019. 24

[67] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International Journal of Cancer*, 136(5):E359–E386, 2015. 10, 11

[68] M. Fleming, S. Ravula, S. F. Tatishchev, and H. L. Wang. Colorectal carcinoma: pathologic aspects. *Journal of Gastrointestinal Oncology*, 3(3):153–173, September 2012. DOI: 10.3978/j.issn.2078-6891.2012.030. 11

[69] F. Foerster, M. Hess, A. Gerhold-Ay, J. U. Marquardt, D. Becker, P. R. Galle, D. Schuppan, H. Binder, and E. Bockamp. The immune contexture of hepatocellular carcinoma predicts clinical outcome. *Scientific reports*, 8(1):5351, 2018. 20

[70] S. Fotso. Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework. *arXiv e-prints*, 2018. 29, 47, 110

[71] E. J. Fox and L. A. Loeb. One cell at a time. *Nature*, 512(7513):143–144, 2014. 13

[72] M. Frantzi, K. E. Van Kessel, E. C. Zwarthoff, M. Marquez, M. Rava, N. Malats, A. S. Merseburger, I. Katafigiotis, K. Stravodimos, W. Mullen, et al. Development and validation of urine-based peptide biomarker panels for detecting bladder cancer in a multi-center study. *Clinical Cancer Research*, 22(16):4077–4086, 2016. 11

[73] W. H. Fridman, L. Zitvogel, C. Sautès-Fridman, and G. Kroemer. The immune contexture in cancer prognosis and treatment. *Nature reviews Clinical oncology*, 14(12):717, 2017. 20

[74] T. J. Fuchs, P. J. Wild, H. Moch, and J. M. Buhmann. Computational pathology analysis of tissue microarrays predicts survival of renal clear cell carcinoma patients. *Med. Image Comput. Comput. Assist. Interv.*, 11(Pt 2):1–8, 2008. 21

[75] J. Galon, A. Costes, F. Sanchez-Cabo, A. Kirilovsky, B. Mlecnik, C. Lagorce-Pagès, M. Tosolini, M. Camus, A. Berger, P. Wind, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science*, 313(5795):1960–1964, 2006. 72

[76] J. Galon, B. Mlecnik, G. Bindea, H. K. Angell, A. Berger, C. Lagorce, A. Lugli, I. Zlobec, A. Hartmann, C. Bifulco, et al. Towards the introduction of the 'immunoscore'in the classification of malignant tumours. *The Journal of pathology*, 232(2):199–209, 2014. 72

[77] J. Galon, F. Pagès, F. M. Marincola, H. K. Angell, M. Thurin, A. Lugli, I. Zlobec, A. Berger, C. Bifulco, G. Botti, et al. Cancer classification using the immunoscore: a worldwide task force. *Journal of Translational Medicine*, 10(1):205, 2012. 72

[78] S. S. Garapati, L. Hadjiiski, K. H. Cha, H.-P. Chan, E. M. Caoili, R. H. Cohan, A. Weizer, A. Alva, C. Paramagul, J. Wei, and C. Zhou. Urinary bladder cancer staging in CT urography using machine learning. *Medical Physics*, 44(11):5814–5823, 2017. 23

[79] C. G. Gavriel, N. Dimitriou, N. Brieu, I. P. Nearchou, O. Arandjelović, G. Schmidt, D. J. Harrison, and P. D. Caie. Assessment of immunological features in muscle-invasive bladder cancer prognosis using ensemble learning. *Cancers*, 13(7):1624, 2021. 18, 19

[80] F. Ghaznavi, A. Evans, A. Madabhushi, and M. Feldman. Digital imaging in pathology: Whole-slide imaging and beyond. *Annual Review of Pathology: Mechanisms of Disease*, 8(1):331–359, 2013. 9, 11, 12, 20

[81] A. P. Glaser, D. Fantini, A. Shilatifard, E. M. Schaeffer, and J. J. Meeks. The evolving genomic landscape of urothelial carcinoma. *Nature Reviews Urology*, 14(4):215, 2017. 49

[82] A. Goode, B. Gilbert, J. Harkes, D. Jukic, and M. Satyanarayanan. OpenSlide: A vendor-neutral software foundation for digital pathology. *Journal of Pathology Informatics*, 4(1):27, Jan. 2013. 106

[83] M. J. Gooden, G. H. de Bock, N. Leffers, T. Daemen, and H. W. Nijman. The prognostic influence of tumour-infiltrating lymphocytes in cancer: a systematic review with meta-analysis. *British Journal of Cancer*, 105(1):93, 2011. 71

[84] F. Gujam, D. McMillan, Z. Mohammed, J. Edwards, and J. Going. The relationship between tumour budding, the tumour microenvironment and survival in patients with invasive ductal breast cancer. *British Journal of Cancer*, 113(7):1066, 2015. 73

[85] M. N. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener. Histopathological image analysis: a review. *IEEE Reviews in Biomedical Engineering*, 2:147–171, 2009. DOI: 10.1109/RBME.2009.2034865. 37

[86] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, March 2003. 37

[87] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422, Jan 2002. 65

[88] N. Harder, M. Athelogou, H. Hessel, N. Brieu, M. Yigitsoy, J. Zimmermann, M. Baatz, A. Buchner, C. G. Stief, T. Kirchner, et al. Tissue phenomics for prognostic biomarker discovery in low-and intermediate-risk prostate cancer. *Scientific Reports*, 8(1):4470, 2018. 55

[89] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020. 47, 73, 87, 107

[90] N. Hashimoto, D. Fukushima, R. Koga, Y. Takagi, K. Ko, K. Kohno, M. Nakaguro, S. Nakamura, H. Hontani, and I. Takeuchi. Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 27

[91] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 28, 97

[92] A. Heindl, S. Nawaz, and Y. Yuan. Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology. *Laboratory Investigation*, 95(4):377–384, January 2015. DOI: 10.1038/labinvest.2014.155. 46

[93] M. Horcic, V. H. Koelzer, E. Karamitopoulou, L. Terracciano, G. Puppa, I. Zlobec, and A. Lugli. Tumor budding score based on 10 high-power fields is a promising basis for a standardized prognostic scoring system in stage ii colorectal cancer. *Human Pathology*, 44(5):697–705, May 2013. DOI: 10.1016/j.humpath.2012.07.026. 46

[94] L. Hou, D. Samaras, M. T. Kurc, Y. Gao, E. J. Davis, and H. J. Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016. 27

[95] G. Huang, Z. Liu, and Q. K. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. 28

[96] S. Huang, J. Yang, S. Fong, and Q. Zhao. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Letters*, 471:61–71, 2020. 9, 24

[97] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. 47, 73, 87, 107

[98] F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *Proceedings of the 5th International Conference on Learning and Intelligent Optimization*, pages 507–523. Springer-Verlag, 2011. DOI: 10.1007/978-3-642-25566-3_40. 36

[99] S. O. Hynes, H. G. Coleman, P. J. Kelly, S. Irwin, R. F. O'Neill, R. T. Gray, C. McGready, P. D. Dunne, S. McQuaid, J. A. James, M. Salto-Tellez, and M. B. Loughrey. Back to the future: routine morphological assessment of the tumour microenvironment is prognostic in stage II/III colon cancer in a large population-based study. *Histopathology*, 71(1):12–26, April 2017. DOI: 10.1111/his.13181. 46

[100] C. Isella, A. Terrasi, S. E. Bellomo, C. Petti, G. Galatola, A. Muratore, A. Mellano, R. Senetta, A. Cassenti, C. Sonetto, L. Inghirami, G. andTrusolino, Z. Fekete, M. Ridder, P. Cassoni, G. Storme, A. Bertotti, and E. Medico. Stromal contribution to the colorectal cancer transcriptome. *Nature Genetics*, 47(4):312–319, February 2015. DOI: 10.1038/ng.3224. 46

[101] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. *CoRR*, abs/1506.02025, 2015. 89, 94

[102] A. Jain and D. Zongker. Feature selection: evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, February 1997. DOI: 10.1109/34.574797. 37

[103] A. Janowczyk and A. Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7(1):29, 2016. 13, 14

[104] A. Janowczyk and A. Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7(1):29, 2016. 14

[105] W. Jiang, W. Sun, A. Tagliasacchi, E. Trulls, and K. M. Yi. Linearized multi-sampling for differentiable image transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 95, 96

[106] A. M. Kamat, N. M. Hahn, J. A. Efstathiou, S. P. Lerner, P.-U. Malmström, W. Choi, C. C. Guo, Y. Lotan, and W. Kassouf. Bladder cancer. *The Lancet*, 388(10061):2796–2810, 2016. 11

[107] Y. Karunakar and A. Kuwadekar. An unparagoned application for red blood cell counting using marker controlled watershed algorithm for android mobile. In *2011 Fifth International Conference on Next Generation Mobile Applications, Services and Technologies*, pages 100–104. IEEE, 2011. 22

[108] A. Katharopoulos and F. Fleuret. Processing megapixel images with deep attention-sampling models. In *ICML*, 2019. 27, 28, 102

[109] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber, L. Jansen, C. C. Reyes-Aldasoro, I. Zörnig, D. Jäger, H. Brenner, J. Chang-Claude, M. Hoffmeister, and N. Halama. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16(1):1–22, 01 2019. 29, 84, 86

[110] J. N. Kather, M. Suarez-Carmona, P. Charoentong, C.-A. Weis, D. Hirsch, P. Bankhead, M. Horning, D. Ferber, I. Kel, E. Herpel, et al. Topography of cancer-associated immune cells in human solid tumors. *Elife*, 7:e36967, 2018. 72

[111] M. D. Keller, C. Neppl, Y. Irmak, S. R. Hall, R. A. Schmid, R. Langer, and S. Berezowska. Adverse prognostic value of pd-l1 expression in primary resected pulmonary squamous cell carcinomas and paired mediastinal lymph node metastases. *Modern Pathology*, 31(1):101, 2018. 71, 73

[112] N. Kemi, M. Eskuri, and J. H. Kauppila. Tumour-stroma ratio and 5-year mortality in gastric adenocarcinoma: a systematic review and meta-analysis. *Scientific Reports*, 9(1):1–6, 2019. 56

[113] M. Khened, A. Kori, H. Rajkumar, G. Krishnamurthi, and B. Srinivasan. A generalized deep learning framework for whole-slide image segmentation and analysis. *Sci. Rep.*, 11(1):11579, June 2021. 26

[114] M. A. Knowles and C. D. Hurst. Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity. *Nature Reviews Cancer*, 15(1):25–41, 2015. 11

[115] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. 35

[116] D. Komura and S. Ishikawa. Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018. 13

[117] B. Kong, X. Wang, Z. Li, Q. Song, and S. Zhang. Cancer metastasis detection via spatially structured deep network. In M. Niethammer, M. Styner, S. Aylward, H. Zhu, I. Oguz, P.-T. Yap, and D. Shen, editors, *Information Processing in Medical Imaging*, pages 236–248. Springer International Publishing, 2017. 26, 27

[118] F. Kong and R. Henao. Efficient classification of very large images with tiny objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2384–2394, June 2022. 27, 100, 102

[119] L. König, F. D. Mairinger, O. Hoffmann, A.-K. Bittner, K. W. Schmid, R. Kimmig, S. Kasimir-Bauer, and A. Bankfalvi. Dissimilar patterns of tumor-infiltrating immune cells at the invasive tumor front and tumor center are associated with response to neoadjuvant chemotherapy in primary breast cancer. *BMC Cancer*, 19(1):120, 2019. 72

[120] N. A. Koohbanani, B. Unnikrishnan, S. A. Khurram, P. Krishnaswamy, and N. Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging*, 40(10):2845–2856, 2021. 26, 28

[121] K. Kourou, K. P. Exarchos, C. Papaloukas, P. Sakaloglou, T. Exarchos, and D. I. Fotiadis. Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. *Computational and Structural Biotechnology Journal*, 19:5546–5555, 2021. 20

[122] M. Kruk, J. Kurek, S. Osowski, R. Koktysz, B. Swiderski, and T. Markiewicz. Ensemble of classifiers and wavelet transformation for improved recognition of fuhrman grading in clear-cell renal carcinoma. *Biocybernetics and Biomedical Engineering*, 37(3):357–364, 2017. 21

[123] C. Langner and N. Schneider. Prognostic stratification of colorectal cancer patients: current perspectives. *Cancer Management and Research*, page 291, July 2014. DOI: 10.2147/cmar.s38827. 10

[124] G. Lee, R. Veltri, G. Zhu, S. Ali, J. Epstein, and A. Madabhushi. Nuclear shape and architecture in benign fields predict biochemical recurrence in prostate cancer patients following radical prostatectomy: Preliminary findings. *European Urology Focus*, 3(4–5):457–466, 2017. 22

[125] K. Leung, M. R. Elashoff, and A. A. Afifi. Censoring issues in survival analysis. *Annual Review of Public Health*, 18(1):83–104, May 1997. 24, 34

[126] B. Li, Y. Li, and K. W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14318–14328, June 2021. 27, 28, 100, 102, 103, 105

[127] Y. Li and W. Ping. Cancer metastasis detection with neural conditional random field. *arXiv:1806.07064*, 2018. 26, 27

[128] C. X. Ling, J. Huang, and H. Zhang. Auc: A statistically consistent and more discriminating measure than accuracy. In *Proceedings*

of the 18th International Joint Conference on Artificial Intelligence, IJCAI'03, pages 519–524, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc. 36

[129] M. Linkert, C. T. Rueden, C. Allan, J.-M. Burel, W. Moore, A. Patterson, B. Loranger, J. Moore, C. Neves, D. MacDonald, A. Tarkowska, C. Sticco, E. Hill, M. Rossner, K. W. Eliceiri, and J. R. Swedlow. Metadata matters: access to image data in the real world. *Journal of Cell Biology*, 189(5):777–782, May 2010. 77, 87

[130] G. Litjens, P. Bandi, B. Ehteshami Bejnordi, O. Geessink, M. Balkenhol, P. Bult, A. Halilovic, M. Hermsen, R. van de Loo, R. Vogels, F. Q. Manson, N. Stathonikos, A. Baidoshvili, P. van Diest, C. Wauters, M. van Dijk, and J. van der Laak. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience*, 7(6), May 2018. 91, 102

[131] Y. Liu, K. Gadepalli, M. Norouzi, E. G. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, Q. P. Nelson, S. G. Corrado, D. J. Hipp, L. Peng, and C. M. Stumpe. Detecting cancer metastases on gigapixel pathology images. *CoRR*, abs/1703.02442, 2017. 26, 28, 98

[132] L. Lombardi, F. Morelli, S. Cinieri, D. Santini, N. Silvestris, N. Fazio, L. Orlando, G. Tonini, G. Colucci, and E. Maiello. Adjuvant colon cancer chemotherapy: where we are and where we'll go. *Cancer Treatment Reviews*, 36:S34–S41, November 2010. DOI: 10.1016/s0305-7372(10)70018-9. 11, 32

[133] M. B. Loughrey, P. Quirke, and N. A. Shepherd. Dataset for colorectal cancer histopathology reports. *The Royal College of Pathologists*, 343, 2014. 11

[134] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts. Understanding variable importances in forests of randomized trees. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 431–439. Curran Associates, Inc., 2013. 65

[135] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2003. 22, 26

[136] C. Lu, D. Romo-Bucheli, X. Wang, A. Janowczyk, S. Ganesan, H. Gilmore, D. Rimm, and A. Madabhushi. Nuclear shape and orientation features from h&e images predict survival in early-stage estrogen receptor-positive breast cancers. *Laboratory Investigation*, 98(11):1438–1448, 2018. 22

[137] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, June 2021. 27

[138] M. Luck, T. Sylvain, H. Cardinal, A. Lodi, and Y. Bengio. Deep learning for patient-specific kidney graft survival analysis. *ArXiv*, abs/1705.10245, 2017. 36

[139] A. Lugli, E. Karamitopoulou, and I. Zlobec. Tumour budding: a promising parameter in colorectal cancer. *British Journal of Cancer*, 106(11):1713–1717, April 2012. DOI: 10.1038/bjc.2012.127. 46

[140] A. Madabhushi and G. Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016. 24, 26, 72

[141] D. R. Magee, D. E. Treanor, D. Crellin, M. Shires, K. J. E. Smith, K. Mohee, and P. Quirke. Colour normalisation in digital histopathology images. 2009. 28

[142] S. Maksoud, K. Zhao, P. Hobson, A. Jennings, and B. C. Lovell. Sos: Selective objective switch for rapid immunofluorescence whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 27, 98

[143] A. Marusyk and K. Polyak. Tumor heterogeneity: Causes and consequences. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1805(1):105–117, 2010. 13

[144] Y. Masugi, T. Abe, A. Ueno, Y. Fujii-Nishimura, H. Ojima, Y. Endo, Y. Fujita, M. Kitago, M. Shinoda, Y. Kitagawa, et al. Characterization of spatial distribution of tumor-infiltrating cd8+ t cells refines their prognostic utility for pancreatic cancer survival. *Modern Pathology*, page 1, 2019. 71, 110

[145] W. McKinney. Data structures for statistical computing in python. In S. van der Walt and J. Millman, editors, *Proceedings of the 9th Python in Science Conference*, Proceedings of the Python in Science Conference, pages 56–61. SciPy, 2010. 47, 73, 87

[146] N. Mehta, A. Raja'S, and V. Chaudhary. Content based sub-image retrieval system for high resolution pathology images using salient interest points. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3719–3722. IEEE, 2009. 22

[147] A. Meier, K. Nekolla, L. C. Hewitt, S. Earle, T. Yoshikawa, T. Oshima, Y. Miyagi, R. Huss, G. Schmidt, and H. I. Grabsch. Hypothesis-free deep survival learning applied to the tumour microenvironment in gastric cancer. *The Journal of Pathology: Clinical Research*, 6(4):273–282, 2020. 29

[148] F. Michor and K. Polyak. The origins and implications of intratumor heterogeneity: Fig. 1. *Cancer Prevention Research*, 3(11):1361–1364, 2010. 13

[149] Y. Nakashima, T. Yao, M. Hirahashi, S. Aishima, Y. Kakeji, Y. Maehara, and M. Tsuneyoshi. Nuclear atypia grading score is a useful prognostic factor in papillary gastric adenocarcinoma. *Histopathology*, 59(5):841–849, November 2011. DOI: 10.1111/j.1365-2559.2011.04035.x. 46

[150] R. Nauta, D. M. Stablein, and D. Holyoke. Survival of patients with stage b2 colon carcinoma. *Archives of Surgery*, 124(2):180, February 1989. DOI: 10.1001/archsurg.1989.01410020050008. 11, 31

[151] I. P. Nearchou, K. Lillard, C. G. Gavriel, H. Ueno, D. J. Harrison, and P. D. Caie. Automated analysis of lymphocytic infiltration, tumor budding, and their spatial relationship improves prognostic accuracy in colorectal cancer. *Cancer Immunology Research*, 7(4):609–620, 2019. 73

[152] M. Niemeyer and O. Arandjelović. Automatic semantic labelling of images by their content using non-parametric Bayesian machine learning and image search using synthetically generated image collages. *In Proc. IEEE International Conference on Data Science and Advanced Analytics*, pages 160–168, 2018. 22

[153] A. Noon, P. Albertsen, F. Thomas, D. Rosario, and J. Catto. Competing mortality in patients diagnosed with bladder cancer: evidence of undertreatment in the elderly and female patients. *British Journal of Cancer*, 108(7):1534–1540, 2013. 56

[154] M. O'Neil and I. Damjanov. Histopathology of colorectal cancer after neoadjuvant chemoradiation therapy. *The Open Pathology Journal*, 3(2):91–98, September 2009. DOI: 10.2174/1874375700903020091. 32

[155] S. Otálora, N. Marini, D. Podareanu, R. Hekster, D. Tellez, J. Van Der Laak, H. Müller, and M. Atzori. stainlib: a python library for augmentation and normalization of histopathology h&e images. *bioRxiv*, 2022. 78

[156] F. Pagès, B. Mlecnik, F. Marliot, G. Bindea, F.-S. Ou, C. Bifulco, A. Lugli, I. Zlobec, T. T. Rau, M. D. Berger, et al. International validation of the consensus immunoscore for the classification of colon cancer: a prognostic and accuracy study. *The Lancet*, 391(10135):2128–2139, 2018. 71

[157] E. R. Parra, A. Francisco-Cruz, and I. I. Wistuba. State-of-the-art of profiling immune contexture in the era of multiplexed staining and digital analysis to study paraffin tumor tissues. *Cancers*, 11(2):247, 2019. 19, 71

[158] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in*

Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019. 87, 106

[159] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 37, 47, 73, 87, 107

[160] B. Peters, V. Niculae, and A. F. Martins. Sparse sequence-to-sequence models. In *Proc. ACL*, 2019. 107

[161] H. Pinckaers, B. van Ginneken, and G. Litjens. Streaming convolutional neural networks for end-to-end learning with multi-megapixel images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1581–1590, 2022. 28

[162] A. Pirovano, H. Heuberger, S. Berlemont, S. Ladjal, and I. Bloch. Automatic feature selection for improved interpretability on whole slide imaging. *Machine Learning and Knowledge Extraction*, 3(1):243–262, 2021. 27

[163] J. Pohjonen, C. Stürenberg, A. Föhr, A. Rannikko, T. Mirtti, and E. Pitkänen. Exposing and addressing the fragility of neural networks in digital pathology, 2022. 78

[164] S. Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. 47

[165] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, November 1994. DOI: 10.1016/0167-8655(94)90127-9. 37

[166] T. Qaiser, A. Mukherjee, C. Reddy PB, D. S. Munugoti, V. Tallam, T. Pitkäaho, T. Lehtimäki, T. Naughton, M. Berseth, A. Pedraza, R. Mukundan, M. Smith, A. Bhalerao, E. Rodner, M. Simon, J. Denzler, C. Huang, G. Bueno, D. Snead, O. I. Ellis, I. Ilyas, and N. Rajpoot. HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues. *Histopathology*, 72(2):227–238, 2017. 26, 28

[167] T. Qaiser and M. N. Rajpoot. Learning where to see: A novel attention model for automated immunohistochemical scoring. *arXiv*, 2019. 27

[168] H. Qu, M. Zhou, Z. Yan, H. Wang, V. K. Rustgi, S. Zhang, O. Gevaert, and D. N. Metaxas. Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning. *npj Precision Oncology*, 5(1):87, Sept. 2021. 112

[169] R. Rajaganeshan, R. Prasad, P. J. Guillou, C. R. Chalmers, N. Scott, R. Sarkar, G. Poston, and D. G. Jayne. The influence of invasive

growth pattern and microvessel density on prognosis in colorectal cancer and colorectal liver metastases. *British Journal of Cancer*, 96(7):1112–1117, March 2007. DOI: 10.1038/sj.bjc.6603677. 46

[170] J. Ramapuram, M. Diephuis, R. Webb, and A. Kalousis. Variational saccading: Efficient inference for large resolution images. In *BMVC*, 2019. 27, 94

[171] S. Raschka. Model evaluation, model selection, and algorithm selection in machine learning, 2018. 23, 64

[172] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001. 78

[173] S. Riihijärvi, I. Fiskvik, M. Taskinen, H. Vajavaara, M. Tikkala, O. Yri, M.-L. Karjalainen-Lindsberg, J. Delabie, E. Smeland, H. Holte, et al. Prognostic influence of macrophages in patients with diffuse large b-cell lymphoma: a correlative study from a nordic phase ii trial. *Haematologica*, 100(2):238–245, 2015. 73

[174] B. D. Ripley. Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):172–192, 1977. 55

[175] A. G. Robertson, J. Kim, H. Al-Ahmadie, J. Bellmunt, G. Guo, A. D. Cherniack, T. Hinoue, P. W. Laird, K. A. Hoadley, R. Akbani, et al. Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell*, 171(3):540–556, 2017. 49

[176] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 53

[177] O. Sanli, J. Dobruch, M. A. Knowles, M. Burger, M. Alemozaffar, M. E. Nielsen, and Y. Lotan. Bladder cancer. *Nature Reviews Disease Primers*, 3:17022, 2017. 11

[178] M. Schemper. Cox analysis of survival data with non-proportional hazard functions. *Statistician*, 41(4):455, 1992. 24

[179] R. S. Schwartz and J. K. Erban. Timing of metastasis in breast cancer. *New England Journal of Medicine*, 376(25):2486–2488, 2017. 12

[180] A. Shaban-Nejad, M. Michalowski, and D. L. Buckeridge. Health intelligence: how artificial intelligence transforms population and personalized health. 1(1):53, 2018. 21

[181] Y. Sharma, A. Shrivastava, L. Ehsan, C. A. Moskaluk, S. Syed, and D. E. Brown. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. In *MIDL*, 2021. 27

[182] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 28

[183] O.-J. Skrede, S. D. Raedt, A. Kleppe, T. S. Hveem, K. Liestøl, J. Maddison, H. A. Askautrud, M. Pradhan, J. A. Nesheim, F. Albregtsen, I. N. Farstad, E. Domingo, D. N. Church, A. Nesbakken, N. A. Shepherd, I. Tomlinson, R. Kerr, M. Novelli, D. J. Kerr, and H. E. Danielsen. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *The Lancet*, 395(10221):350–360, Feb. 2020. 29, 84, 85

[184] L. N. Smith and N. Topin. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. *arXiv e-prints*, page arXiv:1708.07120, Aug. 2017. 84

[185] K. S. Sønderby, K. C. Sønderby, L. Maaløe, and O. Winther. Recurrent spatial transformer networks. *CoRR*, abs/1509.05329, 2015. 94

[186] T. Sugai, N. Yamada, M. Eizuka, R. Sugimoto, N. Uesugi, M. Osakabe, K. Ishida, K. Otsuka, A. Sasaki, and T. Matsumoto. Vascular invasion and stromal s100a4 expression at the invasive front of colorectal cancer are novel determinants and tumor prognostic markers. *Journal of Cancer*, 8(9):1552–1561, 2017. DOI: 10.7150/jca.18685. 46

[187] D. Sui, W. Liu, J. Chen, C. Zhao, X. Ma, M. Guo, and Z. Tian. A pyramid architecture-based deep learning framework for breast cancer detection. *Biomed Res. Int.*, 2021:2567202, Oct. 2021. 26, 27, 91

[188] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021. 12

[189] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, E. R. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. 28

[190] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision, 2015. 93

[191] J. M. Taube, J. Galon, L. M. Sholl, S. J. Rodig, T. R. Cottrel, N. A. Giraldo, A. S. Baras, S. S. Patel, R. A. Anders, D. L. Rimm, and A. Cimino-Mathews. Implications of the tumor immune microenvironment for staging and therapeutics. *Modern Pathology*, 31(2):214–234, 2017. 20

[192] D. Tellez, G. Litjens, J. van der Laak, and F. Ciompi. Neural image compression for gigapixel histopathology image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):567–578, 2021. 19, 27, 100, 102

[193] M. Tellez-Gabriel, B. Ory, F. Lamoureux, M.-F. Heymann, and D. Heymann. Tumour Heterogeneity: The Key Advantages of Single-Cell Analysis. *International Journal of Molecular Sciences*, 17:2142 – 2142, 2017. 13

[194] H. Tokunaga, Y. Teramoto, A. Yoshizawa, and R. Bise. Adaptive weighting multi-field-of-view cnn for semantic segmentation in pathology. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12589–12598, 2019. 27

[195] L. A. Vale-Silva and K. Rohr. MultiSurv: Long-term cancer survival prediction using multimodal deep learning. *bioRxiv*, 2020. 29, 47, 110

[196] L. A. Vale-Silva and K. Rohr. Long-term cancer survival prediction using multimodal deep learning. *Scientific Reports*, 11(1), June 2021. 29

[197] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014. 107

[198] M. E. Vandenberghe, M. L. J. Scott, P. W. Scorer, M. Söderberg, D. Balcerzak, and C. Barker. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Scientific Reports*, 7:45938, April 2017. DOI: 10.1038/srep45938. 46

[199] M. Veta, J. Y. Heng, N. Stathonikos, E. B. Bejnordi, F. Beca, T. Wollmann, K. Rohr, A. M. Shah, D. Wang, M. Rousson, M. Hedlund, D. Tellez, F. Ciompi, E. Zerhouni, D. Lanyi, M. Viana, V. Kovalev, V. Liauchuk, A. H. Phoulady, T. Qaiser, S. Graham, N. Rajpoot, E. Sjöblom, J. Molin, K. Paeng, S. Hwang, S. Park, Z. Jia, I. E. Chang, Y. Xu, H. A. Beck, J. P. van Diest, and P. J. Pluim. Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. *Medical Image Analysis*, 54:111–121, 2019. 26

[200] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. 47, 73

[201] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and H. A. Beck. Deep learning for identifying metastatic breast cancer. *arXiv:1606.05718*, 2016. 26, 102

[202] G. Wang, K.-M. Lam, Z. Deng, and K.-S. Choi. Prediction of mortality after radical cystectomy for bladder cancer by machine learning

techniques. *Computers in Biology and Medicine*, 63:124–132, 2015. 23, 24

[203] X. Wang, C. Barrera, P. Velu, K. Bera, P. Prasanna, M. Khunger, A. Khunger, V. Velcheti, and A. Madabhushi. Computer extracted features of cancer nuclei from h&e stained tissues of tumor predicts response to nivolumab in non-small cell lung cancer. *Journal of Clinical Oncology*, 36(15_suppl):12061–12061, 2018. 22

[204] X. Wang, A. Janowczyk, Y. Zhou, R. Thawani, P. Fu, K. Schalper, V. Velcheti, and A. Madabhushi. Prediction of recurrence in early stage non-small cell lung cancer using computer extracted nuclear features from digital h&e images. *Scientific Reports*, 7(1), October 2017. DOI: 10.1038/s41598-017-13773-7. 46

[205] Y. Wang, K. Lv, R. Huang, S. Song, L. Yang, and G. Huang. Glance and focus: a dynamic approach to reducing spatial redundancy in image classification. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2432–2444. Curran Associates, Inc., 2020. 27

[206] J. Whitney, G. Corredor, A. Janowczyk, S. Ganesan, S. Doyle, J. Tomaszewski, M. Feldman, H. Gilmore, and A. Madabhushi. Quantitative nuclear histomorphometry predicts oncotype DX risk categories for early stage ER+ breast cancer. *BMC Cancer*, 18(1), 2018. 22, 24

[207] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *Transactions on Evolutionary Computation*, 1(1):67–82, 1997. 23

[208] F. P. Wong, W. Wei, W. J. Smithy, B. Acs, I. M. Toki, R. K. Blenman, D. Zelterman, M. H. Kluger, and L. D. Rimm. Multiplex quantitative analysis of tumor-infiltrating lymphocytes and immunotherapy outcome in metastatic melanoma. *Clinical Cancer Research*, 25(8):2442–2449, 2019. 19

[209] Y. Wu, M. Cheng, S. Huang, Z. Pei, Y. Zuo, J. Liu, K. Yang, Q. Zhu, J. Zhang, H. Hong, D. Zhang, K. Huang, L. Cheng, and W. Shao. Recent advances of deep learning for computational histopathology: Principles and applications. *Cancers*, 14(5), 2022. 9

[210] E. Wulczyn, D. F. Steiner, M. Moran, M. Plass, R. Reihs, F. Tan, I. Flament-Auvigne, T. Brown, P. Regitnig, P.-H. C. Chen, N. Hegde, A. Sadhwani, R. MacDonald, B. Ayalew, G. S. Corrado, L. H. Peng, D. Tse, H. Müller, Z. Xu, Y. Liu, M. C. Stumpe, K. Zatloukal, and C. H. Mermel. Interpretable survival prediction for colorectal cancer using deep learning. *npj Digital Medicine*, 4(1), Apr. 2021. 29, 84, 85, 86

[211] F. Xing and L. Yang. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. *IEEE Reviews in Biomedical Engineering*, 9:234–263, 2016. 22

[212] J. Xu, Z. Sun, H. Ju, E. Xie, Y. Mu, J. Xu, and S. Pan. Construction of novel prognostic nomogram for mucinous and signet ring cell colorectal cancer patients with a survival longer than 5 years. *International Journal of General Medicine*, Volume 15:2549–2573, Mar. 2022. 47

[213] S. Xue, G. Song, and J. Yu. The prognostic significance of pd-l1 expression in patients with glioma: a meta-analysis. *Scientific Reports*, 7(1):4231, 2017. 71

[214] T. Yagi, Y. Baba, K. Okadome, Y. Kiyozumi, Y. Hiyoshi, T. Ishimoto, M. Iwatsuki, Y. Miyamoto, N. Yoshida, M. Watanabe, et al. Tumour-associated macrophages are associated with poor prognosis and programmed death ligand 1 expression in oesophageal cancer. *European Journal of Cancer*, 111:38–49, 2019. 71, 110

[215] C.-N. Yu, R. Greiner, H.-C. Lin, and V. Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. 34

[216] K.-H. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Ré, D. L. Rubin, and M. Snyder. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications*, 7:12474, August 2016. DOI: 10.1038/ncomms12474. 24

[217] X. Yue, N. Dimitriou, and O. Arandjelović. Colorectal cancer outcome prediction from H&E whole slide images using machine learning and automatically inferred phenotype profiles. *BICOB*, pages 139–149, 2019. 18, 75, 84, 86

[218] S. Zagoruyko and N. Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016. 28

[219] G. F. Zanjani, S. Zinger, and H. N. P. de With. Cancer detection in histopathology whole-slide images using conditional random fields on deep embedded spaces. In *Medical Imaging 2018: Digital Pathology*. SPIE, mar 2018. 27

[220] M. D. Zarella, C. Yeoh, D. E. Breen, and F. U. Garcia. An alternative reference space for H&E color normalization. *PLoS One*, 12(3):e0174489, 2017. 21

[221] J. Zhang, K. Ma, J. V. Arnam, R. Gupta, J. Saltz, M. Vakalopoulou, and D. Samaras. A joint spatial and magnification based attention framework for large scale histopathology classification. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2021. 27

[222] Y. Zhao, F. Yang, Y. Fang, H. Liu, N. Zhou, J. Zhang, J. Sun, S. Yang, B. Menze, X. Fan, and J. Yao. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 26

[223] X. Zhu, J. Yao, F. Zhu, and J. Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6855–6863, July 2017. 75