Topology and Attention in Computational Pathology

by

Talha Qaiser

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Computer Science

April 2019

THE UNIVERSITY OF
WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

All gratitude to the Almighty who gave me strength and will power to conclude this work.

To Prof. Nasir M. Rajpoot, my PhD supervisor and mentor. I am truly grateful to you for offering me this opportunity and enduring support over the years. This thesis would not become possible without his kind feedback, motivation, and guidance. Under his guidance, I learnt a lot of academic and professional skills which are required to be a competent researcher. Thank you for this amazing experience. To Prof. David B. A. Epstein for his valuable feedback on my academic writing and his extremely useful suggestions on my persistent homology work. I am grateful to my external and internal examiners, Prof. Stephen McKenna and Prof. Hakan Ferhatosmanoglu for their valuable time and insightful comments on my thesis.

I would like to acknowledged the support from the Department of Computer Science, University of Warwick and University Hospital Coventry Warwickshire. I am also thankful to our academic collaborators Prof. Kazuaki Nakane, Prof. Paul Murray, Dr. Yee Wah Tsang, Prof. David Snead, Prof. Mohammad Ilyas, Dr. Abhik Mukherjee, Dr. Matthew Pugh, Dr. Sandra Margielewska, Dr. Robert Hollows, Prof. Naoya Sakamoto and Dr. Daiki Taniyama for their precious time and valuable discussions.

To all my friends and colleagues at Tissue Image Anayltics (TIA) lab Dr. Korsuk Sirinukunwattana, Dr. Violeta Kovacheva, Dr. Nicholas Trahearn, Dr. Shan-e-Ahmed Raza, Dr. Najah Alsubaie, Dr. Guannan Li, Dr. Mike T. Song, Mary Shapcott, Ruqayya Awan, Simon Graham, Muhammad Shaban, Navid Alemi Koohbanani, Jevgenij Gamper, Dr. Ali Khurram, Peter Byfield, Anna Lisowska,

Dr. Katherine Hewitt, Dr. Sajid Javed, Dr. Ksenija Benes, Dr. Moazam Fraz, Dr. Ayesha Azam, Hammam Alghamdi, and John Pocock. It was a great pleasure to have you and working with you. I have shared a lot of unforgettable memories with you all.

Finally, I would like to thank my parents, my younger sisters and brother for their continuous moral support along the way to complete my PhD thesis.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. I declare that, except where acknowledged, the material presented in this thesis is my own work, and has not been previously submitted for obtaining an academic degree.

Talha Qaiser
April 29, 2019

Abstract

Histopathology serves as the *gold standard* in the process of cancer diagnosis and unravelling the disease heterogeneity. In routine practice, a trained histopathologist performs visual examination of tissue glass slides under the microscope. The objective of the visual examination is to observe the morphological appearance of tissue sections, analyse the density of tumour rich areas, spatial arrangement, and architecture of different types of cells. However, careful visual examination of tissue slides is a demanding task especially when workloads are high, and the subjective nature of the histological grading inevitably leads to inter- and even intra-observer variability. Attaining high accuracy and objective quantification of tissue specimens in cancer diagnosis are some of the ongoing challenges in modern histopathology. With the recent advent of digital pathology, tissue glass slides can now be scanned with digital slides scanners to produce whole slide images (WSIs). A WSI contains a high-resolution pixel representation of tissue slide, stored in a pyramidal structure and typically containing 10^{10} pixels. Automated algorithms are generally based on the concepts of digital image analysis which can analyse WSIs to improve the precision and reproducibility in cancer diagnostics. The reliability of the results of an algorithm can be objectively measured and improved against an objective standard.

In this thesis, we focus on developing automated methods for quantitative assessment of histology WSIs with the aim of improving the precision and reproducibility of cancer diagnosis. More specifically, the designed automated computational pathology algorithms are based on deep learning models in conjunction with algebraic topology and visual attention mechanisms. To the best of our knowledge, the applicability of attention and topology based methods have not been explored in the domain of computational pathology. In this regard, we propose an algorithm for computing persistent homology profiles (topological features) and propose two variants for effective and reliable tumour segmentation of colorectal cancer WSIs. We show that incorporation of deep features along with topological features improves the overall performance for tumour segmentation.

We then present the first-ever systematic study (contest) for scoring the human epidermal growth factor receptor 2 (HER2) biomarker on breast cancer histology WSIs. Further, we devise a reinforcement learning based attention mechanism for HER2 scoring that sequentially identifies and analyses the diagnostically relevant regions within a given image, mimicking the histopathologist who would not usually analyse every part of the slide at the highest magnification. We demonstrate the proposed model outperforms other methods participated in our systematic study, most of them were using state-of-the-art deep convolutional networks. Finally, we propose a multi-task learning framework for simultaneous cell detection and classification, which we named as Hydra-Net. We then compute an image based biomarker which we refer as digital proximity signature (DPS), to predict overall survival in diffuse large B-cell lymphoma (DLBCL) patients. Our results suggest that patients with high collagen-tumour proximity are likely to experience better overall survival.

List of Publications

- **Talha Qaiser**, Yee-Wah Tsang, Daiki Taniyama, Naoya Sakamoto, Kazuaki Nakane, David Epstein, and Nasir Rajpoot. Fast and accurate Tumor segmentation of histology images using persistent homology and deep convolutional features. *Medical Image Analysis* (2019).

- **Talha Qaiser**, and Nasir M. Rajpoot. Learning Where to See: A Novel Attention Model for Automated Immunohistochemical Scoring. *IEEE Transactions on Medical Imaging* (2019).

- **Talha Qaiser**, Matthew Pugh, Sandra Margielewska, Robert Hollows, Paul Murray, and Nasir M. Rajpoot. Tumor-Collagen Digital Proximity Signature Indicates Prognostic Significance in Diffuse Large B-Cell Lymphoma. *Journal of Clinical Oncology*, ASCO Abstract (2019).

- **Talha Qaiser**, Abhik Mukherjee, Chaitanya Reddy Pb, Sai D. Munugoti, Vamsi Tallam, Tomi Pitkaho, Taina Lehtimki, Thomas Naughton, Matt Berseth, Anbal Pedraza, Ramakrishnan Mukundan, Matthew Smith, Abhir Bhalerao, Erik Rodner, Marcel Simon, Joachim Denzler, Chao-Hui Huang, Gloria Bueno, David Snead, Ian O Ellis, Mohammad Ilyas, and Nasir M. Rajpoot. HER 2 challenge contest: a detailed assessment of automated HER 2 scoring algorithms in whole slide images of breast cancer tissues. *Histopathology* 72, no. 2 (2018): 227-238.

- **Talha Qaiser**, Matthew Pugh, Sandra Margielewska, Robert Hollows, Paul Murray, and Nasir M. Rajpoot. Digital Tumor-Collagen Proximity Signature

Predicts Survival in Diffuse Large B-Cell Lymphoma. (Accepted in *ECDP* 2019)

- **Talha Qaiser**, Yee-Wah Tsang, David Epstein, and Nasir Rajpoot. Tumor segmentation in whole slide images using persistent homology and deep convolutional features. In Annual Conference on *Medical Image Understanding and Analysis*, (2017): pp. 320-329. Springer, Cham.
- **Talha Qaiser**, Korsuk Sirinukunwattana, Kazuaki Nakane, Yee-Wah Tsang, David Epstein, and Nasir Rajpoot. Persistent homology for fast Tumor segmentation in whole slide histology images. *Procedia Computer Science* 90 (2016): 119-124. (*Best Paper Award*)

Abbreviations

BC	Breast Cancer
BP	Back Propagation
CAC	Colorectal Adenocarcinoma
CAD	Computer-Assisted Diagnosis
CNN	Convolution Neural Network
CRC	Colorectal Cancer
CSV	Comma Separated Values
DAB	Diaminobenzidine
DCNN	Deconvolutional Neural Network
DCIS	Ductal Carcinoma in Situ
DLBCL	Diffuse Large B-cell Lymphoma
DPS	Digital Proximity Signature
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
FISH	Fluorescence In Situ Hybridisation
GEP	Gastroenteropancreatic Neuroendocrine Tumours

Glo1	Glyoxalase-1
GT	Ground Truth
H&E	Haematoxylin and Eosin
HER2	Human Epidermal growth factor Receptor 2
IARC	International Agency for Research on Cancer
IDC	Invasive Ductal Carcinoma
IHC	Immunohistochemical
IML	Interactive Machine Learning
IoR	Inhibition of Return
IQR	Interquartile Range
KLD	Kullback-Leibler Divergence
KM	Kaplan-Meier
***k*-NN**	*k*-Nearest Neighbours
LBP	Local Binary Patterns
LSTM	Long Short Term Memory
MEP	Multi-stage Ensembling Predictor
MTL	Multi-Task Learning
NCI	National Cancer Institute
NEQAS	National External Quality Assessment Scheme
OS	Overall Survival
PathSoc	Pathological Society of Great Britain & Ireland

PCMS	Percentage with Complete Membrane Staining
PD-1	Programmed Death 1
PHP	Persistent Homology Profile
POMDP	Partially Observable Markov Decision Process
RBF	Radial Basis Function
ResNet	Residual Neural Network
RF	Random Forest
RNN	Recurrent Neural Network
ROIs	Regions Of Interest
SVM	Support Vector Machine
TMA	Tissue Micro Arrays
WSI	Whole Slide Image
WSISA	Whole Slide Histopathological Images Survival Analysis

Chapter 1

Introduction

Cancer is an ensemble of related diseases originating from uncontrolled proliferation of cells and regarded as a major challenge in modern medicine [1]. In 2018, nearly 17 million new cancer cases were diagnosed and 9.5 million cancer related deaths occurred, as reported by the International Agency for Research on Cancer (IARC) [2]. In males, the most common types of cancer are lung, prostate, and colorectal cancer, whereas in females breast, colorectal, and lung cancer are the most common cancers [3]. Carcinogenesis in humans is linked with progressive genetic abnormalities, which result in the transformation of normal cells to tumour cells. Generally, the proliferation of healthy cells is intricately regulated by genes, whereas in tumour cells, a variety of mutations lead to uncontrolled cell growth [4]. Broadly, cancer can be categorised into benign tumour (non-invasive) and malignant tumour (able to metastasise, invasive). In metastasis, tumour cells spread from the primary tumour (origin) to nearby tissues, lymph nodes or organs, and initiate tumour development in other parts of the body.

Histology and cytology serve as the backbone in the process of cancer diagnosis and understanding disease heterogeneity. The diagnostic process starts with extraction of a tissue sample from a biopsy, which is generally an organ-specific process. Some common ways of collecting the tissue sample include: needle biopsy, endoscopy, and surgery. The next step is to preserve the gross tissue specimen by freezing or paraffin embedding. A microtome is normally used to section the tissue into thin slices, typically of 5 microns in thickness. Sectioned tissue regions are then mounted on to glass slides, and stained with special chemical markers to highlight different tissue structures. One of the most commonly used stains is Haematoxylin and Eosin (H&E). Haematoxylin binds with nucleic acids (DNA, RNA) and dyes dense nuclei as dark blue or violet, whereas Eosin dyes cytoplasmic substance as pink, including proteins, nutrients and muscles (connective) tissues. Immunohistochemical (IHC) staining is another commonly used approach, where chemical biomarkers identify the over-expression of a specific protein. The routine IHC staining for breast cancer patients includes oestrogen and progesterone (ER/PR) hormone receptors, Ki67 for tumour cell proliferation, and a membranous marker for Human epidermal growth factor receptor 2 (HER2). After staining, the sections are visually examined by an expert pathologist under the optical microscope to determine if the specimen is malignant or benign.

1.1 Tumour Morphology and Histological Analysis

The objective of visual examination is to determine the histological grade of cancer which is normally done by observing the morphological appearance of tissue sections, quantifying the density of tumour rich areas, analysing spatial arrangement and structure of tumour cells [5]. Histological grading varies for different types of cancer, but generally, tumours are graded into 3 categories low-grade, moderate, and high-grade tumours [6]. Grading information is later used for selecting treatment options and predictive analysis.

In this section, we provide an overview to the histology of different cancer types presented in this thesis.

1.1.1 Colorectal Cancer

Colorectal cancer (CRC), also known as colon or bowel cancer, originates in the colon or the rectum, and occurs due to the abnormal growth pattern of cells. CRC is the third most commonly diagnosed cancer in males and the second most in females [7], with an estimated 1.4 million cases and 693,000 deaths occurring in 2012 [8]. According to the National Cancer Institute (NCI), around 4.3% of the human population will be diagnosed with CRC during their lifetime, based on 2012-2014 data[1]. CRC results from excessive growth of malignant (cancer) cells in the colon or the rectum. The most common form of CRC is adenocarcinoma (found in up to 95% of CRC cases) which develops in epithelial glandular cells — these are the glands that secrete mucus, which lubricates the colorectal region.

In routine diagnostic process of CRC, the pathologist analyses tissue sections on glass slides under the microscope to observe morphological features and the variability of nuclear morphology. The morphology of epithelial glandular cells is one of the most important features used in clinical practice for the grading of

[1] https://seer.cancer.gov/statfacts/html/colorect.html

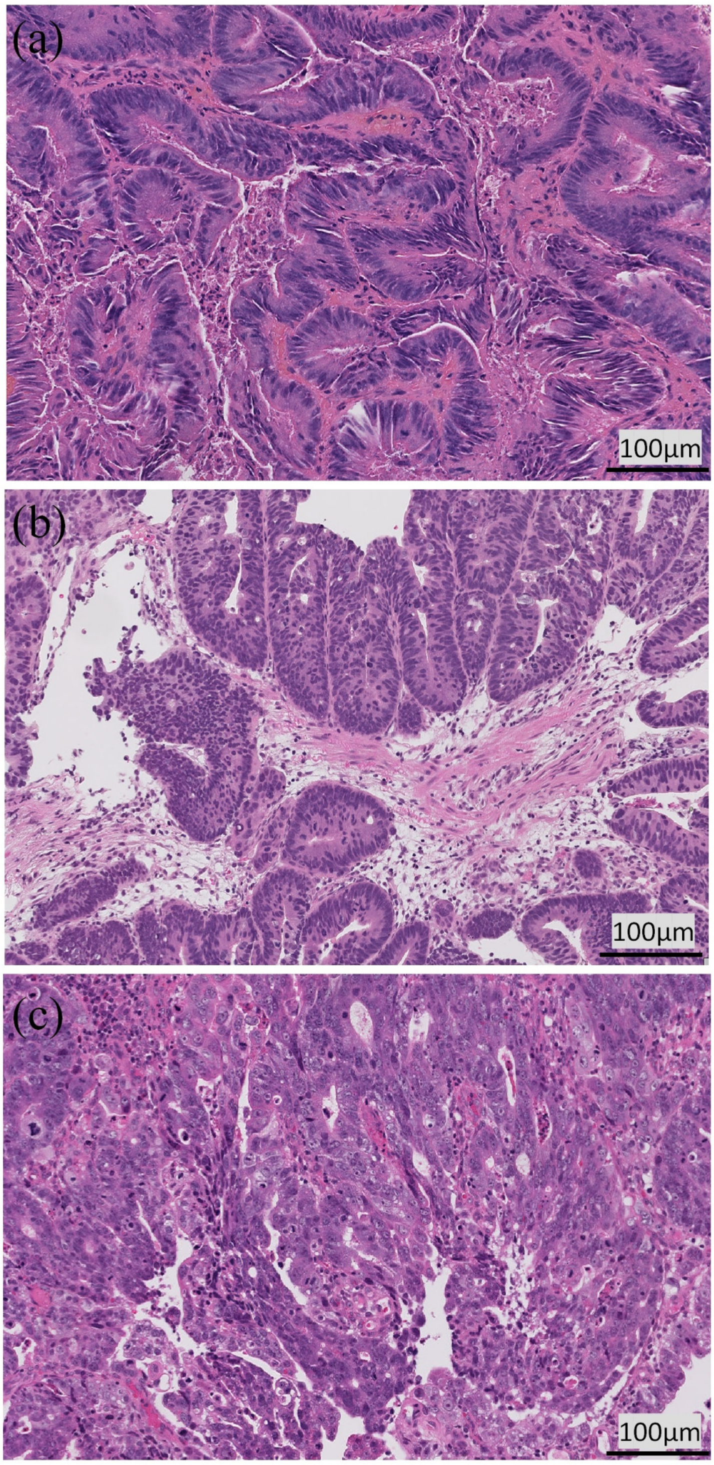


Figure 1.1: Visual fields showing different histology grades for colorectal adenocarcinoma cancer. (a) Well differentiated, (b) moderately differentiated, and (c) poorly differentiated.

colorectal adenocarcinoma (CAC) [9]. CAC tumours can be broadly categorised as low grade or high grade. A low-grade tumour refers to glandular cancer cells with a well-differentiated morphological appearance and tumour, with cancer cells that are poorly differentiated, or undifferentiated are known as a high-grade tumour. In addition to low and high-grade tumours, there also exists moderately differentiated tumour regions. Generally, the histological grading for CAC is estimated by subjective assessment of glandular formation [10]. CAC tissue slides with 95% of glandular structures are categorised as well differentiated tumours. A grade of moderately differentiated is assigned to cases where glandular structures lie between 50% and 95%. Poorly differentiated and undifferentiated grades are those where the glandular formulation lies within 5-50% and $< 5\%$, respectively. Figure 1.1 shows the visual fields of CAC with different histological grades. As tumour transforms from low to high grade, tumour epithelial cells exhibit atypical characteristic. For example, nuclei become relatively large, with heterogeneous chromatin texture and irregularities in their shape and size. Due to uncontrolled cell division, for moderately and poorly differentiated grades, the structure of individual glands is difficult to discern. Prior to cancer grading or quantitative analysis, identification of tumour-rich areas is considered to be one of the primary tasks for a pathologist and for a computer-aided diagnosis system.

1.1.2 Breast Cancer

Breast cancer (BC) is the most commonly diagnosed cancer among women and the second leading cause of cancer related deaths worldwide [11]. According to Cancer Research UK, the risk for women being diagnosed with BC is one in eight in the United Kingdom, and approximately 11,600 women died from BC in 2012 [12]. Broadly, BC can be categorised into two groups, based on tumour origin and morphology: ductal carcinoma in situ (DCIS) and invasive ductal carcinoma (IDC). DCIS is a non-invasive carcinoma and the most common form of BC (found in up to 80% of diagnosed BC cases [13]) where the growth of malignant cells is confined only to breast ducts. IDC is an advanced stage of invasive carcinoma, where malignant epithelial cells infiltrate through the ductal wall and invade into the regions of the nearby tissue. The Nottingham Histologic Grading system is the most widely accepted histologic grading system [14], which classifies tumour into 3 grades based on the morphological appearance of malignant epithelial cells.

In routine diagnostic practice of BC, tumour tissue sections are stained with H&E and then examined under the optical microscope for morphological assessment and grading. In addition, tissues are stained with IHC to evaluate biomarker

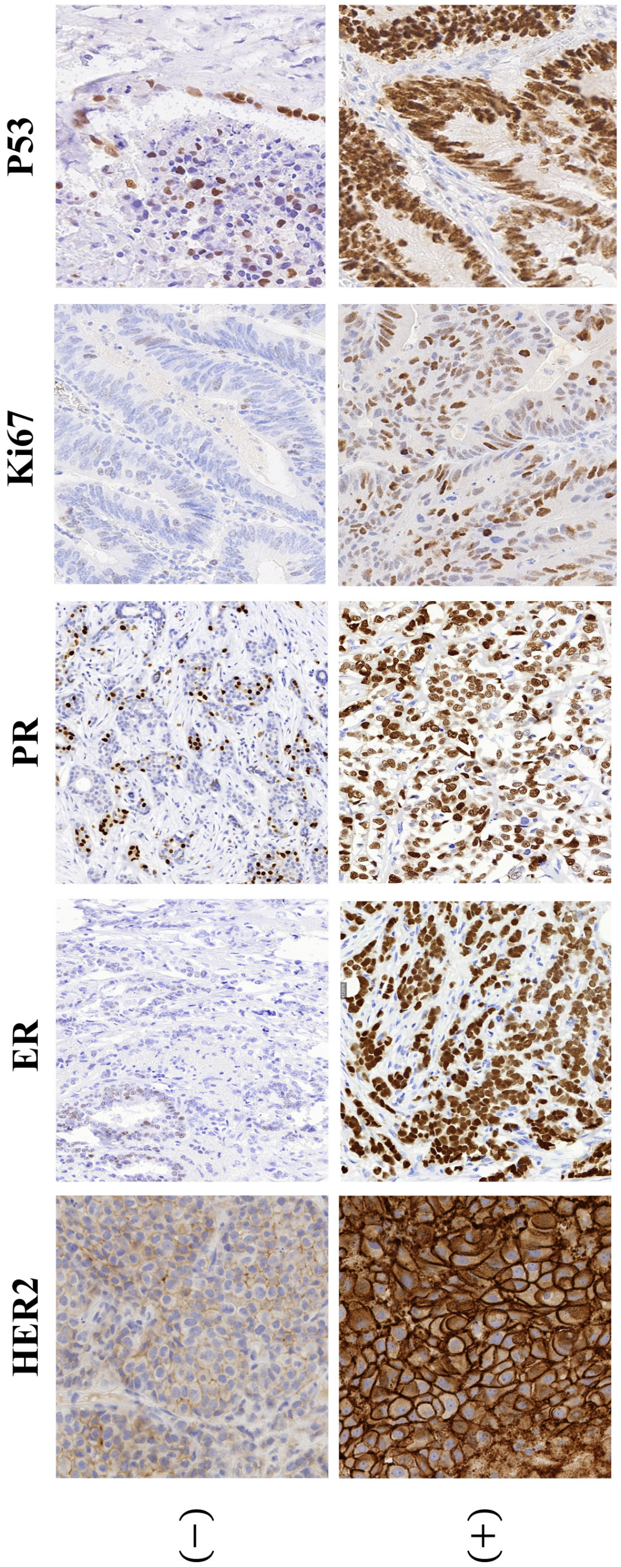


Figure 1.2: Examples of commonly used immunohistochemical stained visual fields with negative (first column) and positive (second column) response.

expression for prognostic and predictive purposes. The most commonly used IHC stains are HER2, ER, PR, Ki67, and p53, as shown in Figure 1.2 with positive and negative visual fields. More specifically ER, PR and HER2 are considered as an integral part for routine BC diagnosis. As opposed to other commonly used IHC stains, HER2 is a membranous stain (as shown in Figure 1.3) and the over amplification of HER2 is estimated by the percentage of cells with complete membrane staining. HER2 mainly influences the growth of malignant epithelial cells located in invasive tumour regions. In HER2+ (positive) patients, tumour cells experience uncontrolled growth as compared to HER2- (negative) cases. Invasive breast carcinomas cases are recommended for HER2 testing and the overexpression of HER2 protein is identified in nearly 20-30% of cases[15]. Recent studies have reported the HER2 status as a predictive and prognostic factor which is associated with poor prognosis, lower survival, and high recurrence [16]. Precise quantification of HER2 overexpression is also crucial for ensuring that HER2+ patients receive appropriate anti-HER2 hormonal therapies [17].

1.1.3 Diffuse Large B-cell Lymphoma

Lymphomas are malignancies derived from lymphocytes (a type of white blood cell) and are broadly categorised into B-cell and T-cell lymphomas, reflecting the origin of the respected cell. B-cell malignancies are further categorised into low grade and high-grade lymphomas. More specifically, diffuse large B-cell lymphoma (DLBCL) is an aggressive (rapidly-growing) malignancy, affecting the growth of B-type lymphocytes that are responsible for antibody production. DLBCL is the most common high-grade lymphoma in Western populations [18], affecting individuals with a median age of 70 years at the time of diagnosis [19].

Routine pathological assessment is performed by a hematopathologist who examines the tissue samples under the microscope dissected from the affected lymph nodes. The objective of the visual examination is to assess the nodal architecture and categorise tissue samples in four stages as recommended by the Ann Arbor classification system [20] (Table 1.1). Patients with stage I or II are considered to have an early stage DLBCL and are normally treated with either complete or a brief course of chemotherapy. DLBCL patients with stage III or IV are considered to have an advanced stage disease. The most common route for the treatment of high-stage DLBCL is a combination of chemotherapy and monoclonal rituximab.

The introduction of modern chemotherapeutic regimens which include rituximab has led to improved survival for DLBCL patients [21]. Despite these advances, approximately 40% of patients will not show a lasting response to therapy and

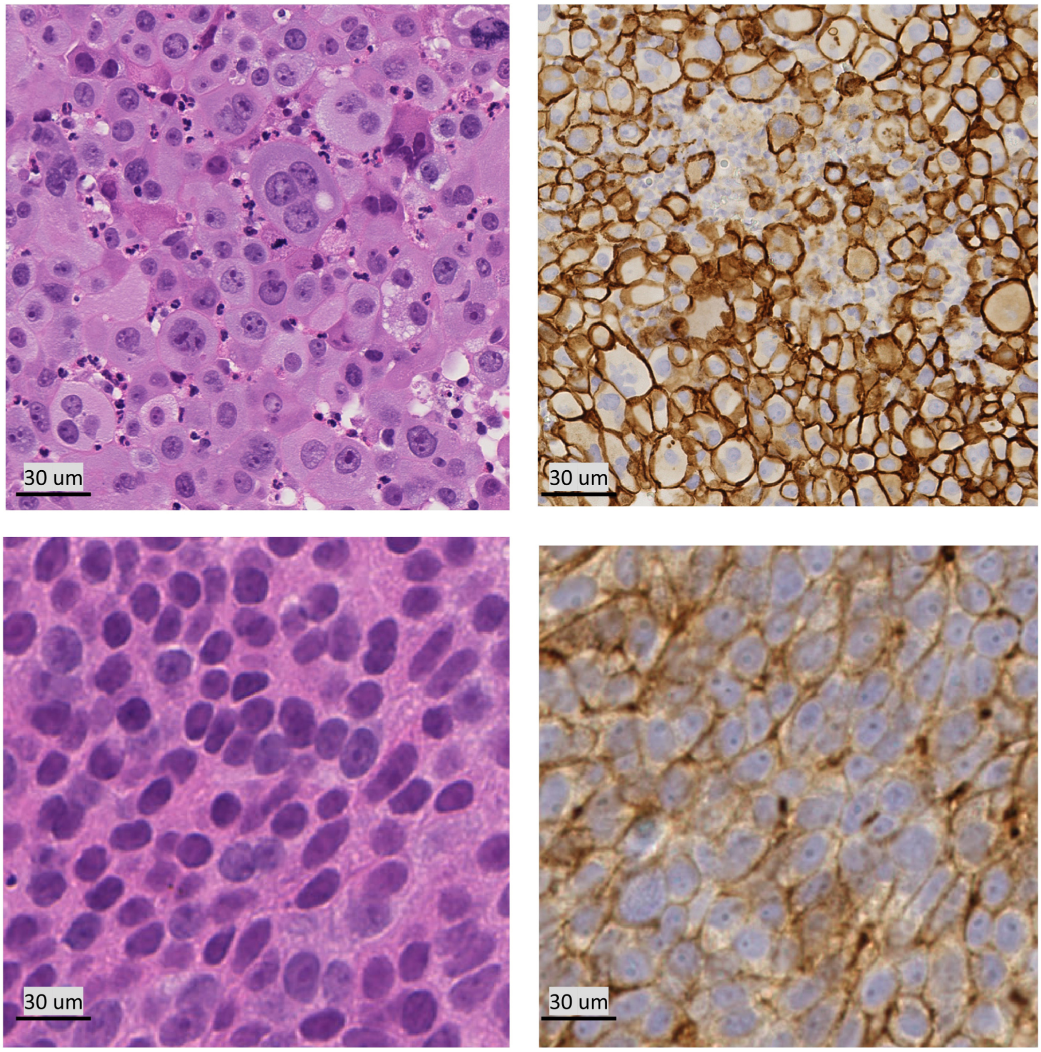


Figure 1.3: Example of two visual fields (*left*) H&E stain and (*right*) HER2 stain.

Table 1.1: Ann Arbor staging of diffuse large B-cell lymphoma [20]

Stage	**Staging description**
I	A single organ or site contains tumour (around a single lymph node)
II	At least two organs or sites in lymphatic regions of same side of diaphragm contain tumour
III	Lymphatic regions (including organs and lymph nodes) contain tumour on both side of the diaphragm
IV	Diffuse or disseminated association of one of multiple extralymphatic organs (like liver, lung nodules, bone marrow)

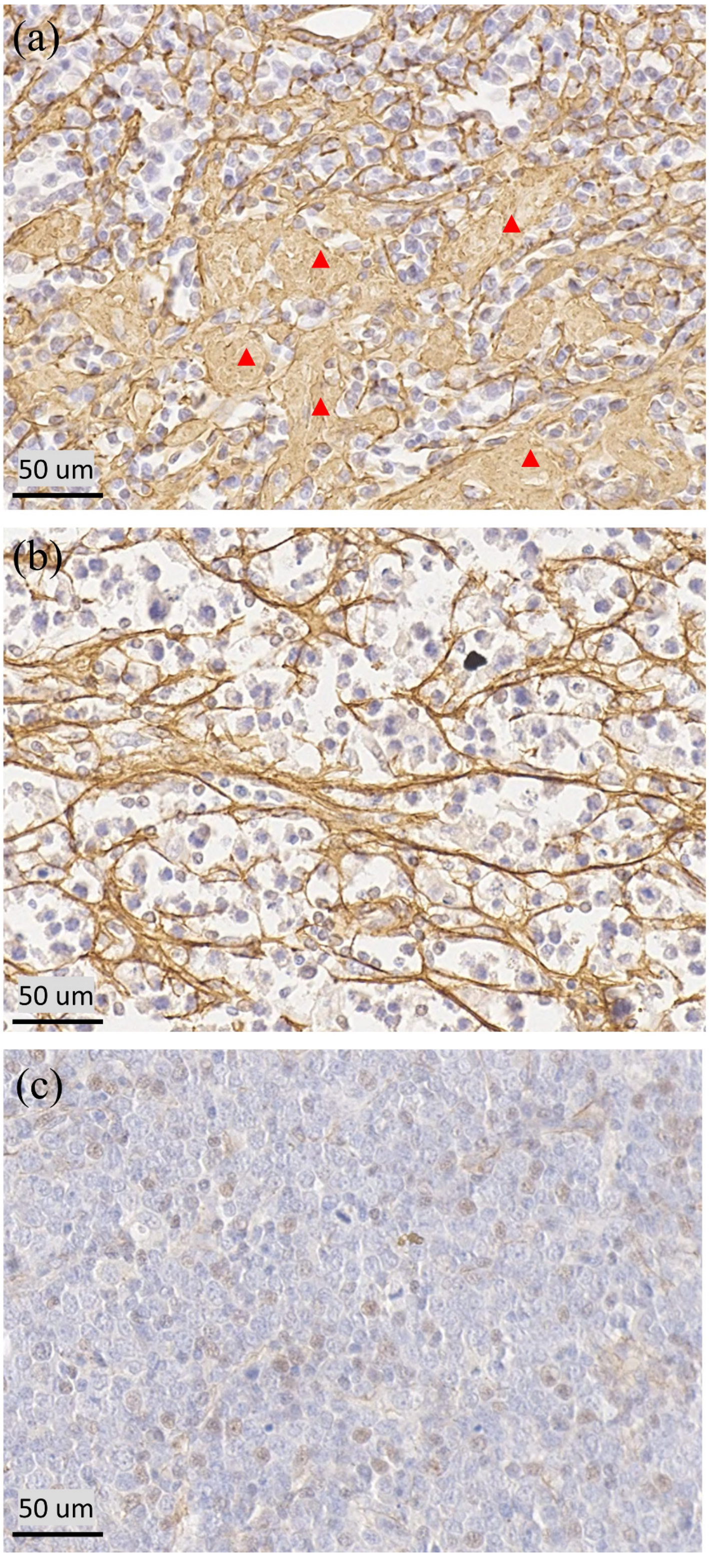


Figure 1.4: Example of three visual fields showing the proximity of collagen and tumour/normal cells with varying quantity of collagen fibre. (a) red triangle shows the abundance of collagen (b) chicken wire structure of collagen and (c) sparse representation of collagen.

will inevitably die of their disease [22]. This variability in treatment response in part reflects the biological heterogeneity of the disease. In recent years, significant progress has been made in unravelling this heterogeneity. Elucidation of the effects of cell-of-origin gene expression profile and most recently exome and transcriptome sequencing, have identified biological and clinically distinct subgroups of the disease [23, 24]. This has led to better prognostic stratification in routine clinical practice and offers the potential for better-targeted therapies in the coming years. Recent studies have also begun to investigate the role of the tumour and stromal microenvironment in DLBCL [25, 26]. An important component of the acellular stromal microenvironment is collagen (as shown in Figure 1.4) and previous studies have linked the presence of collagen with patient outcome [27], and staging of tumours [28].

1.2 Digital Pathology

Visual examination of tissue mounted on glass slides under the microscope to analyse the morphological features is the *gold standard* for cancer diagnosis [29]. However, careful visual examination of tissue slides is difficult when workloads are high, and the subjective nature of the task inevitably leads to inter- and even intra-observer variability in some areas of diagnosis and disease classification [30, 31]. Objective quantification of tissue specimens and attaining high accuracy in cancer diagnosis are some of the ongoing challenges in tissue histology [32, 33]. One of the obvious reasons is the subjective nature of cancer grading and tissue slide preparation. Previously, several modifications have been suggested to overcome subjectivity while grading tumours. The Bloom-Richardson grading system [14] and revised HER2 guidelines and recommendations [34] are two such examples. To some extent, these recommendations have served to improve standards in cancer grading but by and large, these recommendations lack quantitative precision and may still lead to inter- and intra-observer variability. In contrast, automated algorithms can analyse the data with reproducible results. The reliability of the results of an algorithm can be objectively measured (for example against a patient's subsequent clinical progress) and then improved against an objective standard. This is now possible with the advent of digital slides scanners, as hundreds of glass tissue slides can now be scanned in a single run of the scanner.

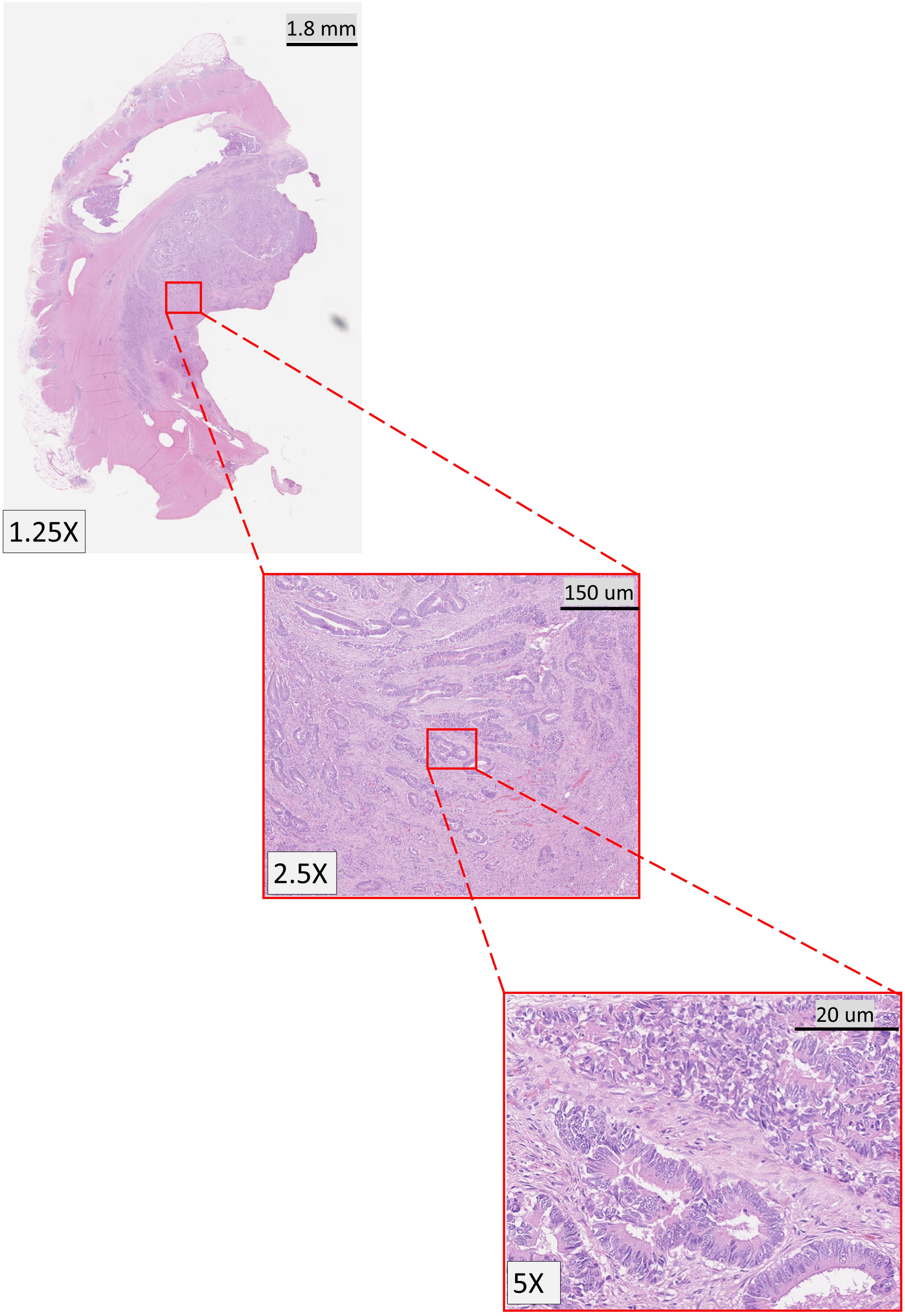
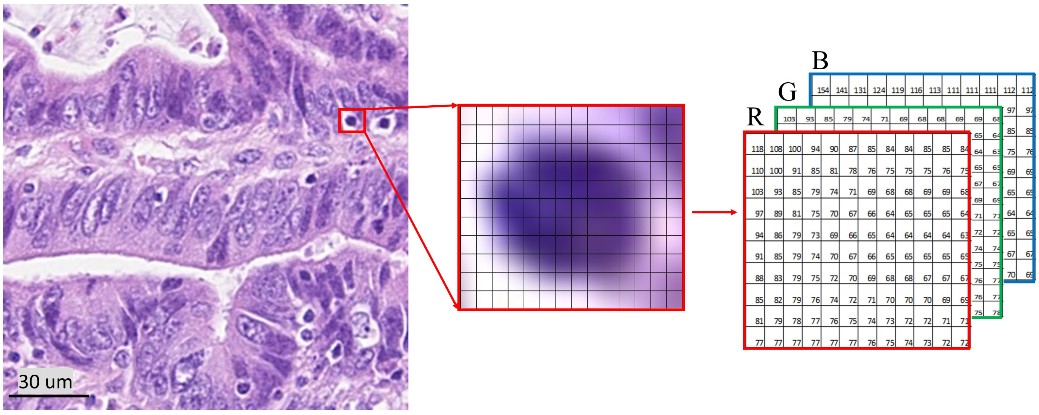


Figure 1.5: An illustration of a whole-slide image and the pyramidal structure containing different magnifications.



118	108	100	94	90	87	85	84	84	85	85	84
110	100	91	85	81	78	76	75	75	75	76	75
103	93	85	79	74	71	69	68	68	69	69	68
97	89	81	75	70	67	66	64	65	65	65	64
94	86	79	73	69	66	65	64	64	64	64	63
91	85	79	74	70	67	66	65	65	65	65	65
88	83	79	75	72	70	69	68	68	67	67	67
85	82	79	76	74	72	71	70	70	70	69	69
81	79	78	77	76	75	74	73	72	72	71	71
77	77	77	77	77	77	76	75	74	73	72	72

Figure 1.6: Example showing image as *matrix* of red, green, and blue channels.

1.2.1 Whole Slide Images

A whole-slide image (WSI) is a multi-resolution gigapixel image typically stored in a pyramid structure, formed by scanning a conventional glass slide at microscopic resolution. A WSI contains a detailed pixel level representation of tissue slide and typically contain 10^{10} pixels. The pyramid structure consists of multiple magnification levels, generally ranging between 40× (with scan resolution of approximately 0.25µm/pixel) to lower than 0.625×, see Figure 1.5. On average, an uncompressed version of a WSI at 40× requires approximately 56 GB [35] of memory. Loading an entire WSI into GPU or CPU memory is a non-trivial and computationally demanding task. Therefore, WSIs are stored in different file formats equipped with lossless compression including tiles produced by Aperio (.svs), Hamamatsu (.ndpi), Leica (.scn), MIRAX (.mrxs), Omnyx (.jp2), and Philips (.tiff) scanners. Reading and retrieving data from these WSI formats requires custom libraries, like OpenSlide [36] (written in C) or Kakadu [37] (C++).

Overall, this paradigm shift towards slide scanning has not only revolutionised the process of cancer diagnosis by replacing the optical microscope with WSIs, but it has also paved the way for computer-assisted diagnosis. Some other potential advantages are the development of telepathology workstations which may assist in remote diagnosis and telepathology training. With the adoption of slide scanners and the increasing number of cancer cases, it is imperative to develop computational pathology algorithms that can assist the histopathologists in their diagnosis of cancer.

1.2.2 Computational Pathology Algorithms

Computational pathology refers to a discipline that utilises varied forms of relevant data (including imaging and clinical data) to design computational models that analyse the data and presents clinically relevant knowledge to experts for improving histology based diagnosis and patient treatment [38]. Due to the nature of the data (images) used in digital pathology, the computational models are typically built on the concepts of digital image analysis and computer vision. More specifically, these models learn morphological and textural characteristics either using hand-crafted (shallow learning) or data-driven (deep learning) approaches. We have leveraged some of these approaches within this thesis for automatic WSI analysis, with the aim to improve the precision and reproducibility of cancer diagnosis.

While the adoption of slide scanners has greatly advanced the progress in computational pathology, it has also inherited some of the challenges faced by the domains of computer vision and image processing. The fundamental disparity is how human and machines perceive the provided information. Machines are designed to understand data in the form of numbers (binary, octal, hexadecimal, etc.) and similarly, each image is presented in the form of matrices, as shown in Figure 1.6. Whereas we as humans inherit knowledge to process visual information. Therefore, varying standards in histology images may result in producing low quality images which may adversely affect the performance of the underlying model (two such examples are shown in Figure 1.7).

There is a wealth of information available in each cancerous WSI containing thousands of nuclei with varying levels of heterogeneity. Processing, analysing, and retrieving meaningful outcome from these large scale images is a computationally intensive task. Therefore, automated algorithms are generally structured into a sequence of algorithms. The pipeline of automated workflow is crucial for efficient training, inference, and deployment of frameworks. The most common approach for automated analysis of WSIs involves a pre-processing step comprising of tissue segmentation and tiling tissue regions into small patches for training the underlying model. The pre-processing step is generally followed by training a patch-based supervised, unsupervised, or reinforcement model to learn relevant information and predict the outcome of each input patch. The work presented in this thesis utilises deep learning models in conjunction with algebraic topology and visual attention mechanisms. Persistent homology is an algebraic tool for studying topological features using the structure of a given space. Generally, persistent homology is computed using simplicial homology by forming n-dimensional simplices. For analysing images, we can avoid the use of simplicial homology by forming filtration space that essentially means a growing sequence of subspaces. For a given visual scenario, we (humans) can effectively localise potential regions of interest to understand the overall scenario. Similarly, the concept of visual attention is generally used in computer vision to localise potential regions of interest, for a better understanding of a given image. Both these concepts are further explained in relevant sections of this thesis. The last step, in general, involves aggregation of patch-based results to the WSI level which is typically done by aggregating the statistics from the output probability map of each WSI. Depending on the availability of clinical data, morphological features are further explored for potential link with clinical outcomes and survival to facilitate better prognostication and treatment of cancer patients.

1.3 Aims and Objectives

This thesis aims to develop automated algorithms for quantitative assessment of histology images. More specifically, we investigate the significance of algebraic topology and visual attention based models to improve the robustness and accuracy in histology image analysis. Applicability of both these concepts (attention and topology) has not been fully explored in the domain of computational pathology. Algorithms in the area of deep learning, computer vision, and image processing approaches will serve as the main tools for achieving the aims of this thesis. In line with the aforementioned aims, the following contributions have been made:

1.3.1 Main Contributions

- We derive an algorithm for computing persistent homology profiles (topological features) which effectively transform an input patch into two 1-D statistical distributions. These distributions capture the degree of nuclear connectivity in a given patch of a WSI. The proposed efficient computation of persistent homology provides an alternative to simplicial homology for 2D images.

- We propose two approaches for tumour segmentation based on the homology profiles. One that relies on the selection of exemplar patches using convolution neural network (CNN) and patch-level classification using a variant of k-nearest neighbours (k-NN). The second approach relies on the best of both worlds by ensembling persistent homology profiles with deep convolutional features.

- We report on the first-ever systematic study (contest) for scoring the overexpression of HER2 biomarker on BC histology WSIs that has instigated a

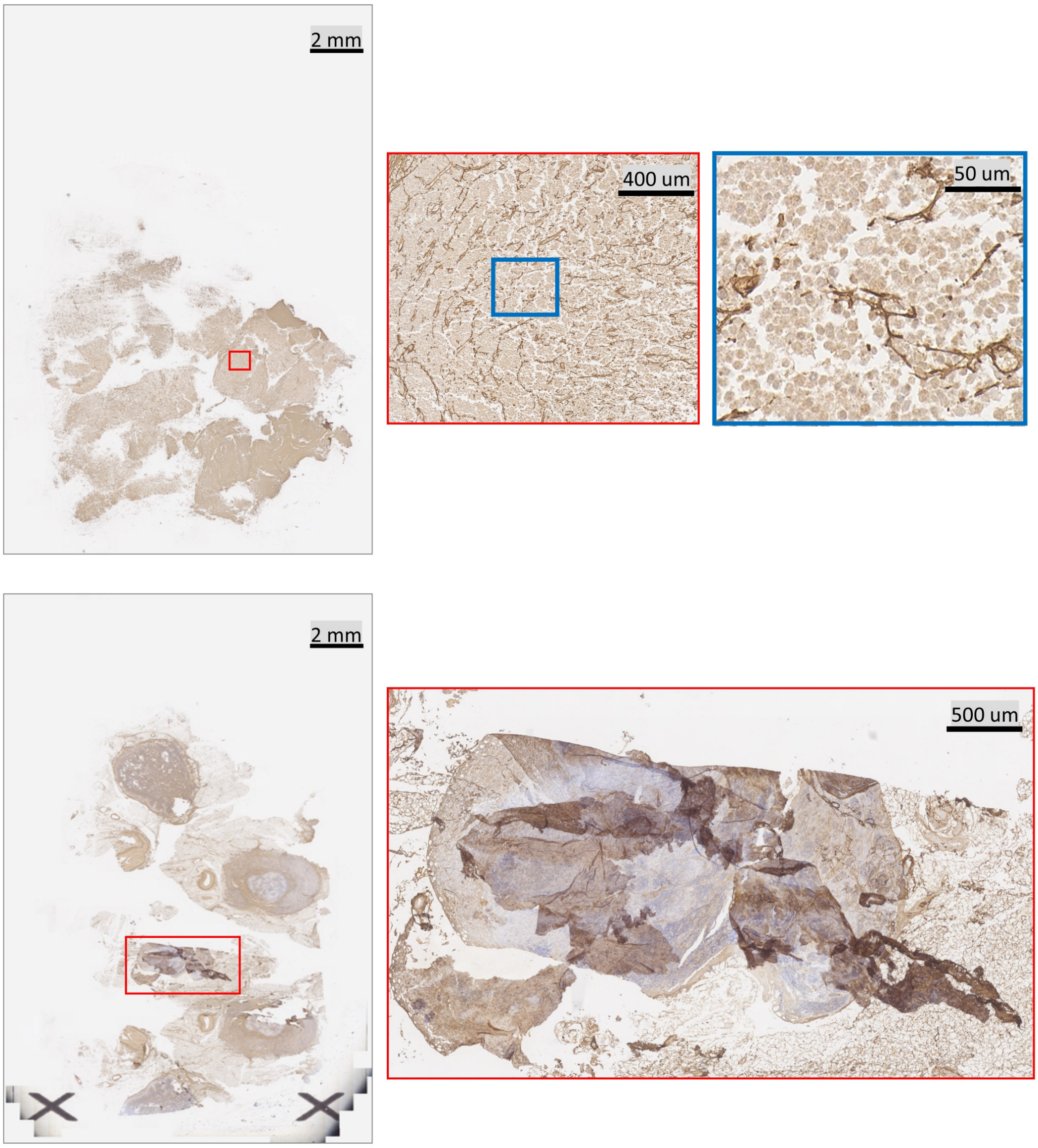


Figure 1.7: Examples showing different types of artefacts in histological images: (*top*) an over-stained histology samples and (*bottom*) tissue folding.

dedicated platform for researchers to contribute and assess the performance of their automated algorithms.

- We devise a reinforcement learning based attention mechanism that sequentially identifies and analyses the diagnostically relevant regions within a given image. To inhibit the model to reiterate previously selected regions, we propose an inhibition of return (IoR) strategy, giving higher priority to regions that have not previously considered for learning.

- We propose an image based biomarker, which we refer to as *Digital Proximity Signature* (DPS), to predict the overall survival in DLBCL patients. For tumour cell detection and classification, we propose a multi-task learning framework, which is named Hydra-Net.

1.4 Thesis Organisation

Chapter 2. Persistent Homology for Fast and Accurate Tumour Segmentation. This chapter presents a tumour segmentation framework based on the concept of *persistent homology profiles* (PHPs) and deep learning. Persistent homology is an algebraic tool from homology theory that computes topological features of a given filtration space. The PHPs are devised to distinguish tumour regions from their normal counterparts by modelling the atypical characteristics of tumour nuclei. The framework contains two variants of our method: one that targets speed without compromising accuracy and the other that targets higher accuracy. The fast version is based on a selection of exemplar image patches from a CNN and patch classification by quantifying the divergence between the PHPs of exemplars and the input image patch. Comparative evaluation shows that the proposed algorithm is significantly faster than competing algorithms while achieving comparable results. The accurate version combines the PHPs and high-level CNN features and employs a multi-stage ensemble strategy for image patch labelling. Experimental results demonstrate that the combination of PHPs and CNN features outperforms competing algorithms. This study is performed on two independently collected colorectal datasets containing adenoma, adenocarcinoma, signet and healthy cases. Collectively, the accurate tumour segmentation produces the highest average patch-level F1-score, as compared with competing algorithms, on malignant and healthy cases from both the datasets.

Chapter 3. A Systematic Study on IHC HER2 Scoring Algorithms. Estimating over-amplification of HER2 on invasive BC is regarded as a significant predictive and prognostic marker. This chapter reports on a systematic study based on an automated HER2 scoring contest, held in conjunction with the Pathological Society of Great Britain & Ireland (PathSoc) meeting held in Nottingham in June 2016, aimed at systematically comparing and advancing the state-of-the-art automated methods for HER2 scoring. The contest dataset comprised of WSIs of sections from 86 cases of invasive breast carcinoma stained with both H&E and IHC for HER2. The ground-truth (GT) for the contest was collected by a consensus score from at least two experts. It also contains a simple Man vs Machine contest for the scoring of HER2 and shows that the automated methods could outperform the pathology experts on this contest dataset.

Chapter 4. Learning where to see: Attention Model for Automated IHC Scoring. This chapter presents a novel deep reinforcement learning (DRL) based model that treats IHC scoring of HER2 as a sequential learning task. For a given image tile sampled from multi-resolution giga-pixel WSI, the model learns to sequentially identify some of the diagnostically relevant regions of interest (ROIs) by following a parameterised policy. The selected ROIs are processed by recurrent and residual convolution networks to learn the discriminative features for different HER2 scores and predict the next location, without requiring to process all the sub-image patches of a given tile for predicting the HER2 score, mimicking the histopathologist who would not usually analyse every part of the slide at the highest magnification. The proposed model incorporates a task-specific regularisation term and inhibition of return mechanism to prevent the model from revisiting the previously attended locations. We evaluated our model on a publicly available dataset from the HER2 scoring challenge contest. We demonstrate that the proposed model outperforms other methods based on state-of-the-art deep convolutional networks. The proposed algorithm could potentially lead to wider use of DRL in the domain of computational pathology reducing the computational burden of the analysis of large multi-gigapixel histology images.

Chapter 5. Tumour-Collagen Proximity Analytics. This chapter presents an automated method to investigate the spatial proximity of collagen (type VI) and tumour cells in DLBCL cases. For each WSI, we calculate the DPS, which represents summary-level statistics of tumour-collagen proximity. The core components of the proposed framework involve: a) cell detection and classification, (b) finding the reference points from the collagen fibres, and c) calculation of DPS. To the best of our knowledge, this is the first study that performs automated analysis of tumour and collagen on DLBCL to identify potential prognostic factors. Experimental results favour our cell classification algorithm over conventional approaches. In addition, our results show that strongly associated tumour-collagen regions are statistically significant in predicting overall survival (OS) in DLBCL patients.

Chapter 6. Conclusions and Future Work This chapter summarises the main contributions of this thesis and discusses future research direction for extending this work.

Chapter 2

Persistent Homology for Fast and Accurate Tumour Segmentation

Localisation of malignant tumour regions in H&E stained slides is an important first task for a pathologist while diagnosing CRC. Manual segmentation of tumour regions from glass slides is a challenging and time consuming task. Therefore, automated localisation of tumour-rich areas is a vital step towards a computer-assisted diagnosis and quantitative image analysis. Accurate segmentation of tumour-rich areas may also assist pathologists in understanding disease aggressiveness and selection of high power fields for tumour proliferation grading and scoring. In a recent study [39], it has been shown that precise localisation of tumour epithelial regions in CRC can overcome the association of non-malignant stroma regions in gene expression profiling, which also provides substantial prognostic information for individual cases. Hence, automated tumour segmentation of CRC tissue slides could potentially speed up the diagnostic process and overcome the inter-observer variability of conventional methods [40, 41].

Tumour regions can be distinguished from normal regions using the appearance of cell nuclei [42, 43]. In tumour regions, epithelial nuclei have atypical characteristics — relatively large nuclei, with heterogeneous chromatin texture and irregularities in their shape and size. Due to uncontrolled cell division, tumour nuclei sometimes form clusters filling inter-cellular regions, as shown in Figure 2.1. In some cases with moderately and poorly differentiated grades, the structure of individual nuclei is difficult to discern. In contrast, nuclei retain their structure and morphological appearance in normal regions including stroma, lymphocytes, normal mucosa, and adipose tissue regions.

We show that important morphological differences between normal and cancer nuclei can be measured using persistent homology, a mathematical tool explained in Section 2.2. We propose two persistent homology based methods for tumour segmentation of H&E stained WSIs including *a)* homology based fast and reliable tumour segmentation *b)* accurate tumour segmentation by combining the homological and deep convolutional features to enhance the classification accuracy of a deep CNN. We validate the proposed methods with relatively large datasets containing both malignant and healthy cases from two independent institutions. Generally, in a clinical setup, a WSI scanner processes 500 to 1000 glass slides each day and so analysing data from a single scanner may require the processing of several terabytes of new data each day. We, therefore, make a particular point in quoting run times, and show that our fast and reliable tumour segmentation algorithm is significantly faster than a conventional CNN and other competing approaches.

2.1 Related Work

Existing literature on tumour segmentation in histology images can be broadly classified into two categories: 1) hand-crafted feature learning methods and 2) data-driven deep feature based algorithms. In this section, we review previous work regarding the tumour segmentation on images of H&E stained slides. Literature review on other related methods is covered in relevant sections, where appropriate.

A wide range of studies have been published on the use of texture, morphological and colour features for tumour segmentation. Perception-based features [44], local binary patterns (LBP) along with contrast measure features [45], colour graphs [46], Gabor and histogram features [47, 48],and bags-of-superpixels pyramid [49], have been used to segment tumour rich areas. Weakly supervised multiple clustered instance approaches [50, 51] have also been proposed for segmentation of tumour in tissue micro arrays (TMAs) of colon cancer images, by which bags of selected patches are generated to learn the model in a multiple instance framework. The multiple clustered instance model learns from a set of general features like the L*a*b colour histogram, LBP, multiwavelet transforms, and scale invariant feature transforms. However, the selection of an optimal set of features for fully or weakly supervised learning is an onerous task and poses the risk of over-emphasising some particular features of a dataset. Moreover, a major shortcoming of the above algorithms is the fact that their scope is mainly limited to hand-picked visual fields or TMAs. In a clinical setup, a tumour segmentation solution should be capable of

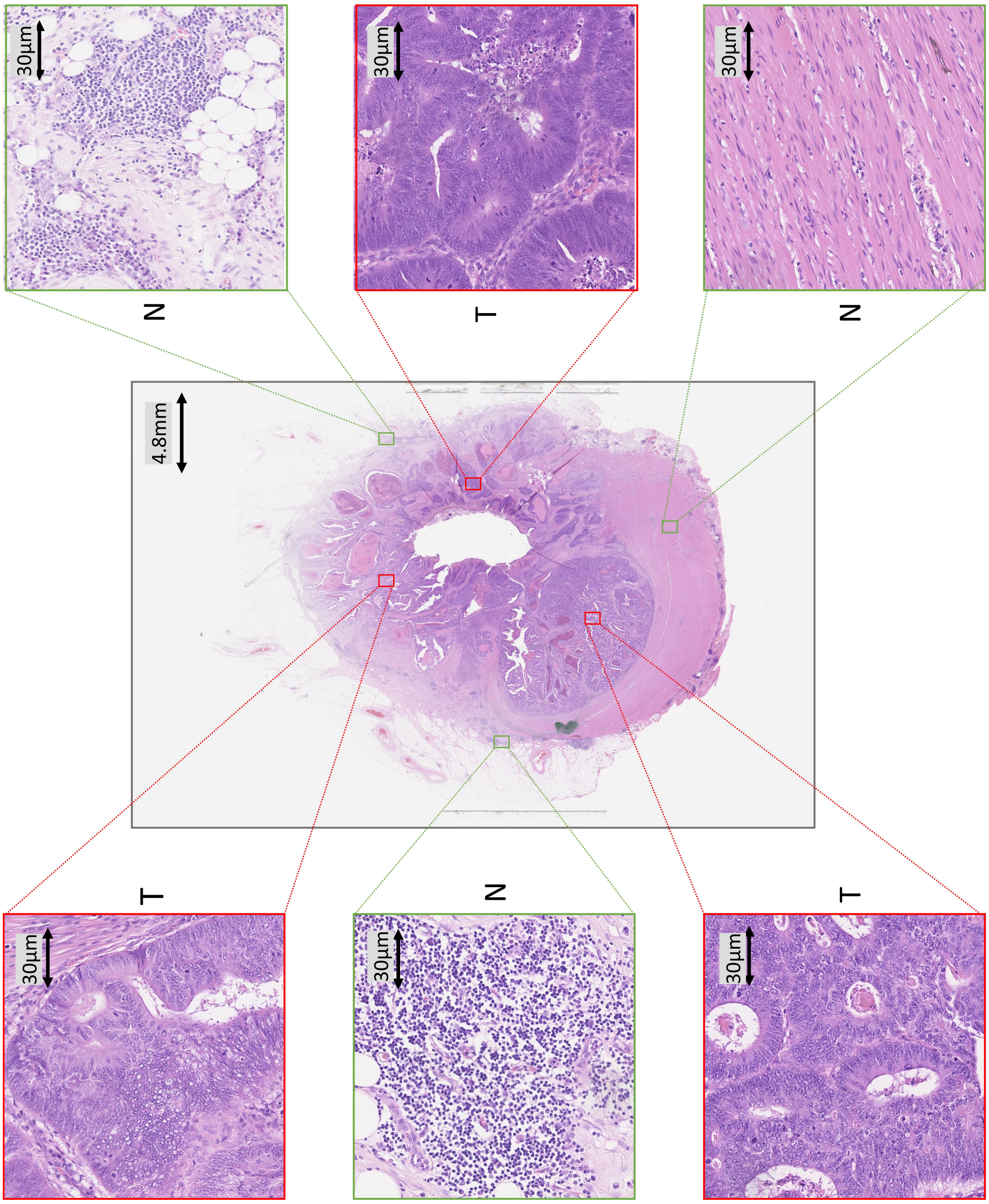


Figure 2.1: An example of a whole slide image with 6 regions-of-interest (ROIs) to illustrate the degree of connectivity between nuclei in tumour and normal regions. The zoomed-in regions are of size 140.8 × 140.8 μm^2, which is equivalent to 20× magnification. ROIs with green rectangles (with label N) shows non-tumour whereas sub-region with red rectangles (with label T) shows tumour areas.

scaling the results to the WSI level.

Deep learning has recently produced exceptional performance on tasks in computer vision [52] and in medical image processing [32, 53, 54]. One of the well-known methods for segmentation is to learn a set of hierarchical features by employing a combination of down and up sampling convolution layers, such as U-net [55]. These approaches perform reasonably well for pixel level segmentation but are computationally expensive and may encounter the vanishing gradients problem while training. Cruz-Roa, *et al.* [56] presented a CNN framework for tumour detection in breast histology images. Our CNN architecture for tumour segmentation has some basic similarities with the proposed framework in [56]. However, our CNN model is relatively deep, enabling the model to learn a set of features at various levels of abstraction. In a supervised learning environment, one can think of deep learning features as a set of data-driven features, learnt by using back propagation (BP). The BP algorithm penalises the kernel maps by forcing them to learn from their mistakes on the training dataset. Instead of relying on a set of handcrafted features, deep learning models learn the set of optimal features without human intervention.

Each of the several physical processes involved in creating a WSI, starting with the original biopsy or resection and ending with laying a 5µm thick section on a glass slide, is random with respect to orientation. Therefore the textural and geometric features in a WSI will have a random orientation, though the orientations of different features may well be correlated with each other. However, deep learning models and especially CNNs find difficulty in learning the rotationally invariant characteristics of an input image [57]. In contrast, our biologically interpretable PHPs not only capture the degree of connectedness among nuclei but are also invariant to rotational transformations.

2.2 Persistent Homology

Persistent homology is an algebraic tool, whereby, given a topological space, certain *algebraic invariants* are computed using the structure of that space. It is a fairly recent concept of homology theory, with a wide range of applications in different domains of data analytics including protein structure [58, 59], robotics [60, 61], neuroscience [62], shape modelling [63], analysing brain arteries [64], classification of endoscopy images [65], mutational profile and survival analysis [66], video surveillance [67], time series modelling [68] and natural language processing [69]. The concept of persistent homology is relatively new for medical image analysis in general and for histology image analysis in particular.

Persistent Homology theory is the study of the homology of a filtered space, by which we mean a sequence

$$\{\emptyset = X_0 \subseteq X_1 \subseteq X_2 \subseteq \cdots \subseteq X_k = \mathbf{X}\} \tag{2.1}$$

This is referred to as a *filtration* of the topological space X. Readers are referred to [70, 71, 72] for a description of the general theory.

In our case, we are analysing 2D greyscale images, and each X_i is the union of closed pixels in a single image. $X_k = X$ is the entire image. These subspaces are so special, and with such nice properties, that dramatic simplifications are possible in computing the persistent homology. The only homology groups that are non-zero are in dimensions 0 and 1. We have no need to consider what is often a significant aspect of persistent homology, namely the birth and death of homology classes. It is sufficient for our purposes to consider only the Betti numbers $\beta_0(X_i)$ and $\beta_1(X_i)$ for $1 \leq i < k$. Betti numbers are represented by non-negative integers for discriminating topological spaces based on their connectivity. Moreover, these Betti numbers can be computed using basic topological ideas, namely a count of connected components, which is a simple and rapid computational procedure. In other words, β_0 represents the number of connected components in the foreground and similarly, β_1 means counting background connected components without holes.

To explain how we generate the filtration of a given greyscale image, we suppose for definiteness that the intensity of each pixel is an integer in the range $[0, 255]$. We then select a sequence of integers $0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = 256$. These integers are various threshold levels, at which the image is binarised. We define X_i to be the union of closed pixels p, with intensity $I(p) < t_i$, so that $X_0 = \emptyset$; and $X_k = X$. Each greyscale image gives rise to a single filtration. A careful choice[1] of $t_1, t_2, \ldots, t_{k-1}$ balances density in $[0, 255]$ to give all (or nearly all) the information needed from the original greyscale image. We compute two relevant homology groups $H_0(X_i)$ and $H_1(X_i)$ for $1 \leq i < k$ as follows. Instead of using computationally expensive constructs of simplicial topology, as is normal in the literature using persistent homology [73, 74], we compute $H_0(X_i)$ by counting the connected components of X_i, giving the Betti number $\beta_0(X_i)$ and $H_1(X_i)$ by counting the components of $X \setminus X_i$, giving the Betti number $\beta_1(X_i)$. For our purposes, it is therefore sufficient to calculate, for each i with $1 \leq i < k$, these two numbers, giving a total of $2k - 2$ length of feature vector.

[1] We empirically decided on choosing every third possible value in the range 1..255, giving 84 levels instead of 255.

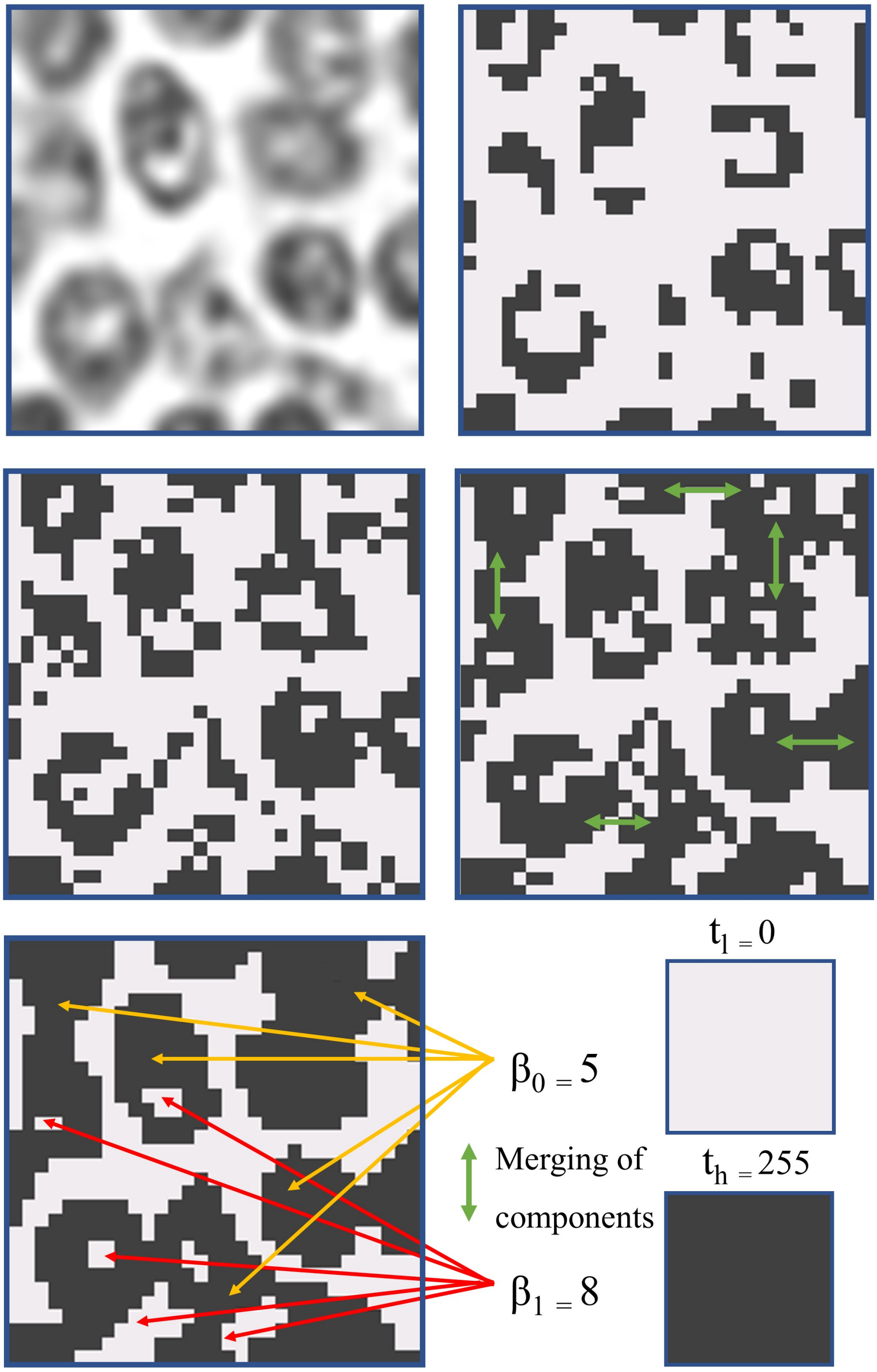


Figure 2.2: An illustration of the filtered space associated to an image. We first show an input image at 40× magnification, and then four images showing growing sequence of subspaces, obtained by thresholding at increasing values of threshold. β_0 represents the number of connected components and, similarly, β_1 shows the number of one-dimensional voids (red arrows only show few of them).

A given small patch as in Figure 2.2 represents a filtered space extracted from a WSI. For a range of t it gives a list of subspaces such that all pixels of previous subspace X_{i-1} are present in X_i. To construct PHPs we recorded the rank of homology groups $(H_0(X_i), H_1(X_i))$ for an entire range of t, such as $0 < t_1 < t_2 < \cdots < t_{k-1} < t_k = 256$. We trade a considerable improvement in speed of computation for a negligible loss in information. The algorithm for computing PHPs is the main foundation for fast tumour segmentation. In this case, the *algebraic invariants* turn out to be nothing more complicated than whole numbers (ranks of homology groups, which are 0^{th} and 1^{st} Betti numbers). Hence we do not need to build computationally expensive simplical complexes in order to compute persistent features. This provides an alternative approach to simplicial homology for 2D images.

We now present two proposed approaches based on PHPs. We first describe the workflow of the fast tumour segmentation and then explain the algorithm for accurate tumour segmentation. For a given WSI, we divide it into patches of 256×256 at 20× magnification. The problem then reduces to classification of each patch as either tumour or non-tumour.

2.3 Fast Tumour Segmentation

The algorithm for fast tumour segmentation is established on three pivotal steps: *1)* an efficient way of computing PHPs, *2)* selection of representative images from the activation maps of a convolutional network, and *3)* an algorithm for patch classification.

2.3.1 Persistent Homology Profiles

As mentioned earlier, tumour nuclei carry atypical characteristics and exhibit chromatin texture allowing us to distinguish them from non-tumour nuclei. Here, we characterise these phenomena with the help of persistent homology for tumour regions in CRC histology images. Our topological features provide a global description of image patch I by finding the relationships among data points (pixels), contrary to textural and geometrical features where precise distances, angles and spatial arrangement are important. For a given patch I, we derive two statistical distributions by computing the ranks of their 0^{th} and 1^{st} dimensional homology groups, denoted by β_0 and β_1. We refer to these statistical distributions as *persistent homology profiles.*

Highly variable colour of histology specimens is mainly due to non standardisation in staining protocols. To overcome this problem we first perform stain deconvolution [75] separating an RGB image patch into three channels, Haematoxylin, Eosin, and background. Generally, eosin channel contains information regarding cytoplasmic regions, including muscles and connective tissues whereas haematoxylin stain binds with nucleic acids and dyes dense nuclei as dark blue. Therefore, for follow up analysis, we only use the Haematoxylin channel to improve consistency in intensity appearances. To extract the topological features, we binarise the Haematoxylin channel to record the corresponding Betti numbers (β_0, β_1). In our case, the topological features were only recorded for one third of possible threshold values - see Footnote 1. These limited threshold values retain the discriminative characteristics of PHPs by using three times fewer parameters as compared to using all threshold values [76]. For each t, the Betti numbers are computed by counting the connected components and one-dimensional voids. Rather than relying on hand-picked threshold values, we observe the topological features at $0 = t_0 < t_1 < \cdots < t_{k-1} < 256$. Later, we convert each statistical curve into a discrete probability distribution, scaling the values so that the area under each curve is one.

Differences between tumour and non-tumour regions are reflected in their homology invariants. This can be seen in Figures 2.3 and 2.4, which show the curves representing median of our PHPs for selected exemplar patches from a CNN (explained in Section 2.3.2) for both tumour and normal classes, with first (Q_1) and third (Q_3) quartile. The green dotted line shows the PHPs (β_0, β_1) for image patches as shown in the first column. It is worth mentioning here the magnitude of derived PHPs is less relevant instead, the pivotal aspect is the noticeable trend in growth of homology classes. As we start increasing the threshold t from the lower limit (t_0) to upper limit (t_{k-1}), the filtering subspace propagates from an empty set to the entire topological space. Since tumour regions carry more irregularities in terms of their shape, size and tumour nuclei lie relatively close to each other filling the inter-cellular cytoplasmic space, their homology ranks (β_0, β_1) do not show rapid change while merging and forming into new classes as compared to those for non-tumour regions.

2.3.2 Selection of Exemplar Patches

To select exemplar patches for fast tumour segmentation, we first train a deep CNN model to predict whether a patch should be labelled as tumour or non-tumour. The CNN architecture is inspired by [52] with some modifications, as shown in Figure 2.5. The objective here is to infer a set of representative patches from the entire

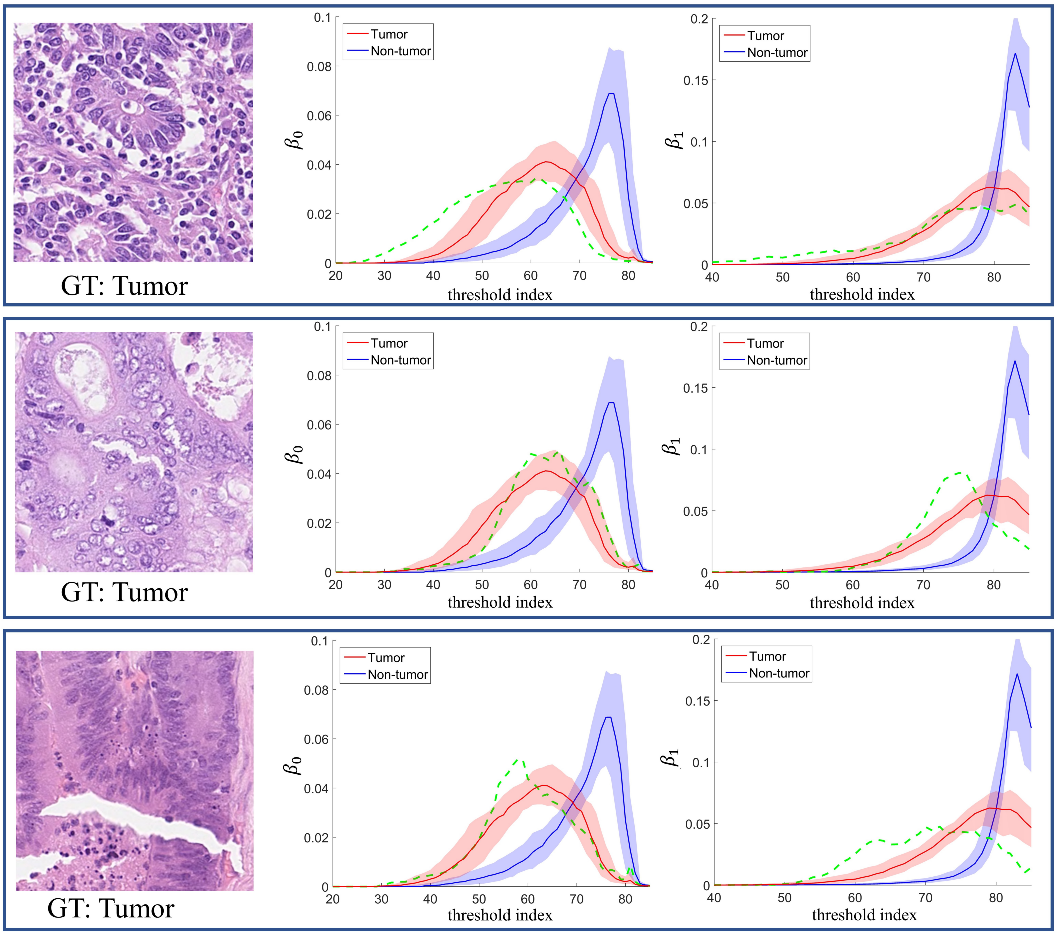


Figure 2.3: An example of persistent homology profiles (PHP) for the selected tumour patches (left): original images with ground truth (GT), (middle): PHP for β_0, (right): PHP for β_1. The shaded regions in β_0 and β_1 show the first and third quartile of the exemplar patches, whereas the green dotted line shows the PHP of selected patch.

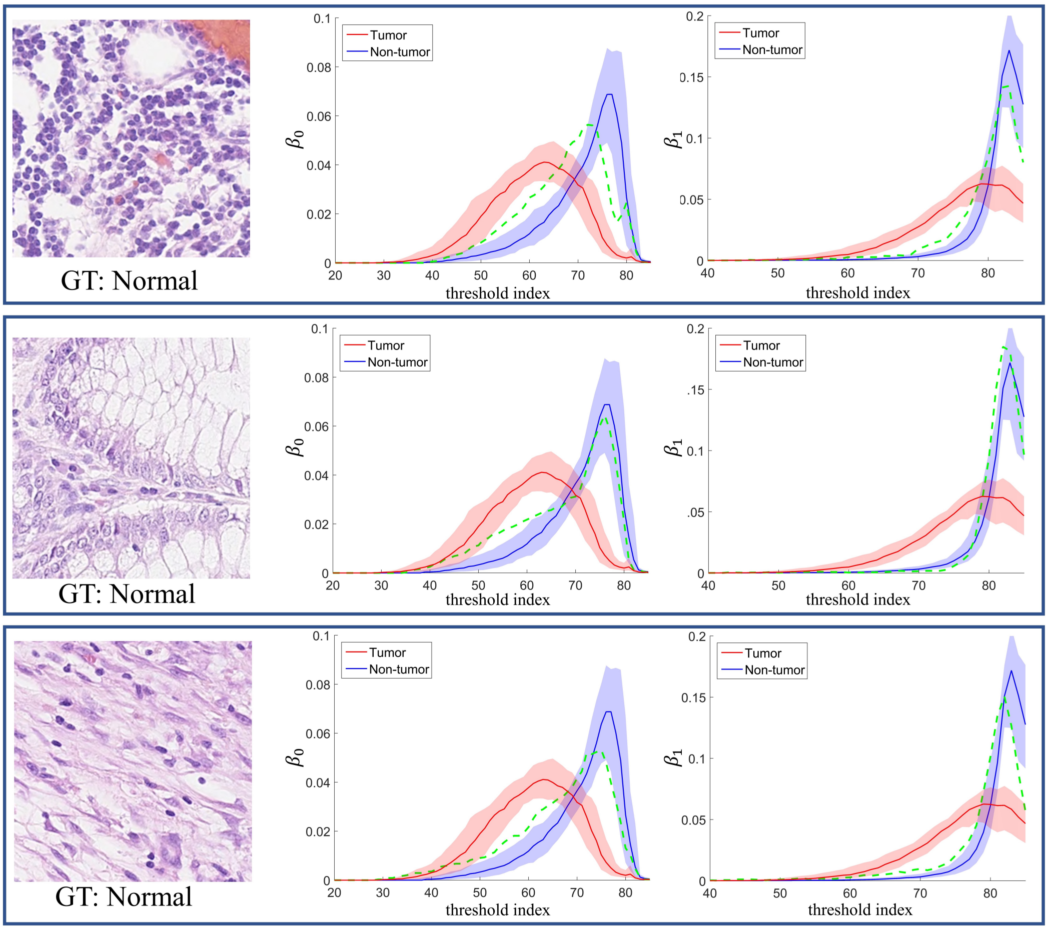


Figure 2.4: Another example of PHP for selected non-tumour patches (left): original images with ground truth (GT), (middle): PHP for β_0, (right) PHP for β_1 .

training dataset for both tumour and non-tumour classes by exploring the learned activation maps from the last convolution layer. We then compute the PHPs of selected patches in order to measure the value of divergence from an input patch as described in the next section.

Let us consider a convolution layer with its corresponding activation maps $\alpha \in \mathbb{R}^{W \times H \times Z}$, where Z represents the depth of activation maps of spatial dimension $H \times W$. The activation maps from convolution layers emphasise different tissue parts for different layers [77]. Generally, in the first layer, neurons activate for a combination of low-level features like edges on nuclei boundaries and chromatin material found within nucleus. The middle layers get more sense of object localisation by learning the most discriminative regions within the tissue patch. The top convolution layer neurons reflect higher level features of tissue components and have high activations around the object within an input image. The idea here is to collapse the 3D activation maps of the last convolution layer into a scalar value that can be later used as an indication of the significance of each patch with respect to the activation maps. We first define a mapping function closely related to [78] to flatten the 3D activation maps to 2D across the Z dimension as below,

$$F(w,h) = \sum_{z=1}^{Z} \left| \alpha_{(z)}^{w,h} \right|^2, \tag{2.2}$$

Here, the mapping function gives more weight to neurons with high activations. It is worth noting that normalising the flattened 2D activation map is important for follow-up analysis. We compute the median value of $F(w,h)$ to find the central tendency of the 2D activation map, that later assist in exclusion of unimportant patches in the training dataset. Similarly, for the entire training dataset, we get an $M \times 1$ vector by computing the $median(F(w,h))$ for each patch, where M represents the number of patches.

The last step is to find a set of exemplar patches for tumour and non-tumour classes, separately. One way of choosing the exemplar patches is to find the highly activated patches from the $M \times 1$ matrix. The only pitfall of following such a method is that we may end up selecting patches representing a certain type of tumour or normal tissue. In order to avoid this scenario, we compute the interquartile range (IQR) [79] of $M \times 1$ matrix separately for the tumour and non-tumour classes. Furthermore, we split the IQR of each class into Q same size bins and select the value that lies closest to the median of corresponding bin, where Q represents the number of exemplar patches for each class. Comparative results for the selection of

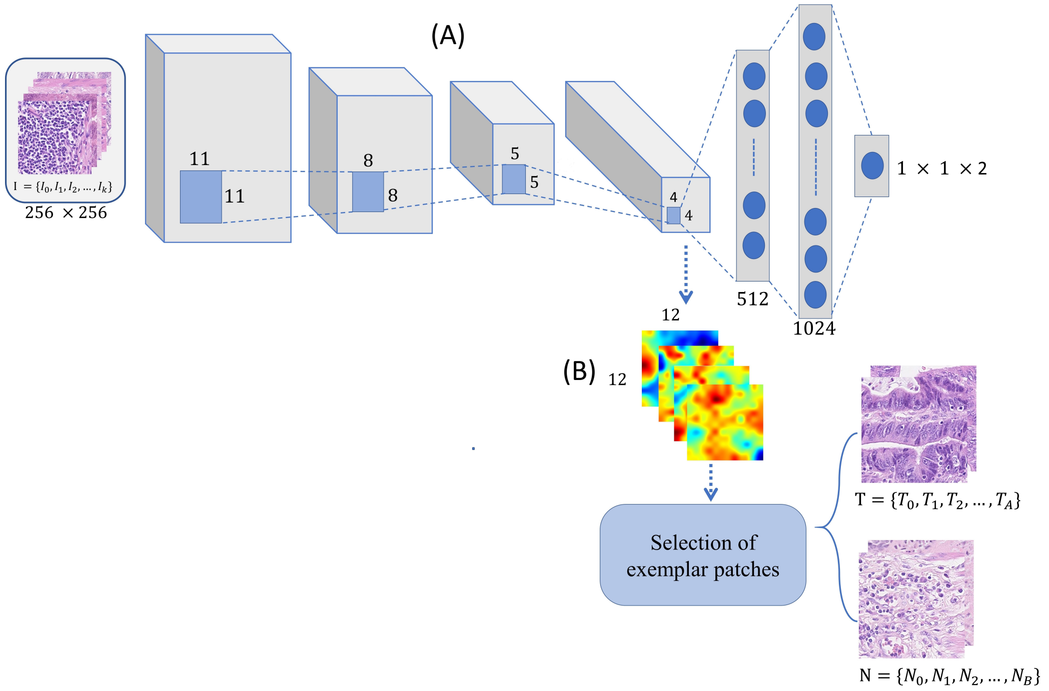

(A)
I = $\{I_0, I_1, I_2, \dots, I_k\}$
256 × 256
11
11
8
8
5
5
4
4
512
1024
1 × 1 × 2
(B)
12
12
Selection of exemplar patches
T = $\{T_0, T_1, T_2, \dots, T_A\}$
N = $\{N_0, N_1, N_2, \dots, N_B\}$

Figure 2.5: (A) A schematic illustration of our deep convolutional neural network (B) An overview of the proposed exemplar selection algorithm based on activation maps of the last convolution layer.

exemplar patches is discussed in Section 2.5.2.

2.3.3 Patch Classification

For patch classification, we first transform the PHPs into discrete probability distributions. We then compute the symmetric Kullback-Leibler divergence (KLD) to measure the distance between the PHPs of an input patch I and the PHPs of an exemplar patch E as defined below,

$$D_{KL}(I \parallel E) = \sum_{i} I(i) \log \frac{I(i)}{E(i)}, \tag{2.3}$$

where I represents the *true* representation of data and E represents an *approximation* of I. The symmetrised, non-negative KLD is defined as

$$d_{I,E} = D_{KL}(I \parallel E) + D_{KL}(E \parallel I). \tag{2.4}$$

We compute a vector of distance values to exemplar patches $D(I) = (d_{I,T_1}$, $d_{I,T_2}, ..., d_{I_A,T_A}$, $d_{I,N_1}, d_{I,N_2}, ..., d_{I,N_B})$, containing divergence values for input patch

PHP profiles of I and all exemplar tumour $T = \{T_1, ..., T_A\}$ and non-tumour exemplar patches $N = \{N_1, ..., N_B\}$. We derive a similarity measure from the KLD values computed in (2.4) and compare the total similarity scores to the k nearest tumour and non-tumour patches as follows:

$$\sum_{j=0}^{k_t} e^{-d_{I,T_j}} > \sum_{j=0}^{k_n} e^{-d_{I,N_j}}, \tag{2.5}$$

where k_t, k_n denotes the number of nearest tumour and non-tumour patches according to (2.5), and $k = k_t + k_n$. In order to classify a given image patch as tumour, the total similarity scores of its k_t nearest tumour patches should be greater than that of the k_n nearest normal patches.

2.4 Accurate Tumour Segmentation

In this variant of our algorithm, we combine deep convolutional and persistent homology features. For this we trained a CNN as shown in Figure 2.5 to extract features from the last fully connected layer. We then fed the extracted features to a Random Forest (RF) regression model separately for the topological and deep convolutional features. Finally, we propose a multi-stage ensemble strategy to combine the two RF regression models. The main objective of this method is to combine *the best of both worlds*. The key contribution of the topological features is to capture the underlying connectivity whereas CNN tends to learn the data driven features.

2.4.1 Deep Convolutional Features

The CNN architecture contains four convolutional layers followed by an activation function and a max-pooling operation. Additionally, it contains two fully connected layers and the softmax classification layer to predict the label of each patch as a tumour or non-tumour. We use rectified linear unit ReLU as activation function, which enables faster convergence and also reduces the vanishing gradient problem. A dropout layer at the end of the second fully connected layer is placed to overcome the overfitting problem. The CNN was trained to minimise the overall cross entropy loss L, as given below

$$L(g, y) = -(y(x)log(g(x)) + (1 - y(x))log(1 - g(x))) \tag{2.6}$$

where for input x, g represents the ground-truth label (0 for tumor and 1 for non-tumor) and y is the probability of the tumor predicted by the CNN. The fully connected layers contain non-linear combinations of learned features from the convolution layers. We extracted CNN features for the training dataset after the last fully connected layer just before the softmax classification layer. For each patch of the training dataset, we obtained a feature vector of size $(1 \times 1 \times 1024)$.

2.4.2 Ensembling Strategy

After obtaining topological and convolutional features, we concatenate both PHPs $(\beta 0, \beta 1)$ to form a combined topological feature vector. We then train the RF regression model separately for both types of feature. We optimise the RF model with an ensemble of 200 bagged trees, randomly selecting one third of the variables for each decision split and setting the minimum leaf size to 5.

We combine the probability of both regression models (O_1, O_2) as in (2.7), where O_1 represents a regression model of topological features and similarly O_2 represents a regression model of convolutional features. The multi-stage ensemble strategy follows two alternative routes: a) averaging the outcome probabilities of O_1 and O_2 to predict the output label where both regression models agree b) for the remaining few patches ($\approx 1\%$ from the test data) where the average probabilities lies in range $0.49 - 0.51$. We refer to these as *critical patches* and we assign the output label for those patches by rounding the probabilities from $O_1(x)$ as in our experiments based on two datasets we observed the comparatively high F1-score with the setting.

$$\hat{O}(x) = \begin{cases} 0, & \text{if } \frac{(O_1+O_2)}{2} < 0.49. \\ 1, & \text{else if } \frac{(O_1+O_2)}{2} > 0.51. \\ \lfloor O_1(x) \rceil & \text{otherwise (rounding)} \end{cases} \tag{2.7}$$

2.5 Experiments and Results

2.5.1 Dataset and Experimental Setup

The Warwick-UHCW Dataset.

This dataset consists of 75 WSIs of H&E stained colorectal adenocarcinoma tissues. At the highest resolution, each WSI normally contains more than 10^{10} pixels. The WSIs were digitally scanned at a pixel resolution of 0.275µm/pixel (40×) using an Omnyx VL120 scanner. The ground-truth for tumour regions were hand-marked by an expert pathologist. For each WSI, we randomly selected 1,500 patches including 750 from non-tumour and 750 patches from tumour regions. In total, we extracted, 75,000 patches for training from 50 WSIs and 37,500 patches for testing from 25 WSIs. The collected dataset for this study is roughly 20 times more than that in [80] and at least 2 times more than in [76]. For generating the tumour probability map of a WSI, we first split the given WSI into patches and then applied our methods to each patch.

The Warwick-Osaka Dataset.

This dataset contains 50 H&E stained histology WSIs of colorectal tissue. The WSIs were scanned at a pixel resolution of 0.23µm/pixel (40×) using a Hamamatsu NanoZoomer 2.0-HT scanner. The ground-truth for this dataset were hand-marked by two expert pathologists and the cases were identified as belonging to 6 categories, including 11 cases of adenoma, 14 moderately differentiated, 6 poorly differentiated, 10 well-differentiated, 8 healthy and 1 signet case. The inclusion of normal cases in this dataset helps in evaluating the robustness of the proposed methods as discussed in Section 2.5.2. Similarly to the Warwick-UHCW dataset, we randomly selected 1,500 patches (750 tumour, 750 non-tumour) from each WSI.

Experimental Setting.

For our proposed methodologies we split a WSI into manageable patches of 256×256 for training as well as testing. In order to counter overfitting, we performed data augmentation by rotating $(0°, 90°, 180°, 270°)$, flipping (horizontal or vertical axis), and perturbing the colour distribution (hue variation) of both the training datasets (Warwick-UHCW and Warwick-Osaka). The colour perturbation was achieved by randomly varying the hue in the interval $[0, 0.5]$ and saturation between 0.5 and 1.5. The weights for a CNN were initialised by using Xavier initialisation, as given below

$$l = \sqrt{\frac{3}{(N_{in} + N_{out})}} \tag{2.8}$$

where N_{in} and N_{out} represents the number of input and output neurons, respectively. The selected weight initialisation approach tends to restrict the magnitude of gradients from excessive shrinking or growing during the training process. The network learns the weights by using the mini-batch gradient descent algorithm by selecting a batch size of 100. During the training phase, the initial learning rate was set to 0.0001 and an Adam optimiser was employed instead of conventional gradient descent algorithm. In addition, a dropout layer (dropout rate 0.5) was placed between the two fully connected layers to overcome the interdependence among intermediate neurons and to increase the robustness of the trained network. For the fast tumour segmentation, we separately chose 128 exemplar patches for both tumour and non-tumour, and $k = 11$ for k-NN.

Evaluation.

We compute the F1-score to evaluate the performance of different approaches as a harmonic mean of precision and recall as defined below,

$$F_1 = 2 \times \frac{P_r \times R_e}{(P_r + R_e)}; \qquad P_r = \frac{T_p}{T_p + F_p}; \qquad R_e = \frac{T_p}{T_p + F_n} \tag{2.9}$$

where T_p, F_p, and F_n represents the number of true positive, false positive, and false negative patches. For a given test dataset, the correctly identified patches are classified as either true positives or true negative, misclassified predictions are categorised as false positives, and false negatives.

2.5.2 Comparative Analysis

Selection of Exemplar Patches

One of the most critical parts in the fast tumour segmentation is the selection of exemplar patches. Hence the objective of this experiment is to investigate a few options for selection of exemplar patches. The experiment is conducted on 75 colorectal adenocarcinoma WSIs from the Warwick-UHCW dataset. The following algorithms were selected for comparison and as an alternative to the proposed method of exemplar selection. We start with a random selection of exemplar patches from the training dataset for both tumour and non-tumour classes. We repeat this process 10 times before concluding the final results and the reported results (Table 2.1) show the mean precision, recall and F1-score. In k-means clustering the training dataset was partitioned into k clusters with respect to their RGB intensities. Then we selected those patches that lie closest to the cluster centroids as exemplar patches, one patch per centroid. The third algorithm for comparison is to select highly activated patches as exemplars from a CNN as proposed in [80]. For a fair comparison, an equal number of exemplar patches were selected for each of the above mentioned algorithms.

Table 2.1 reports the patch based tumour segmentation results for different

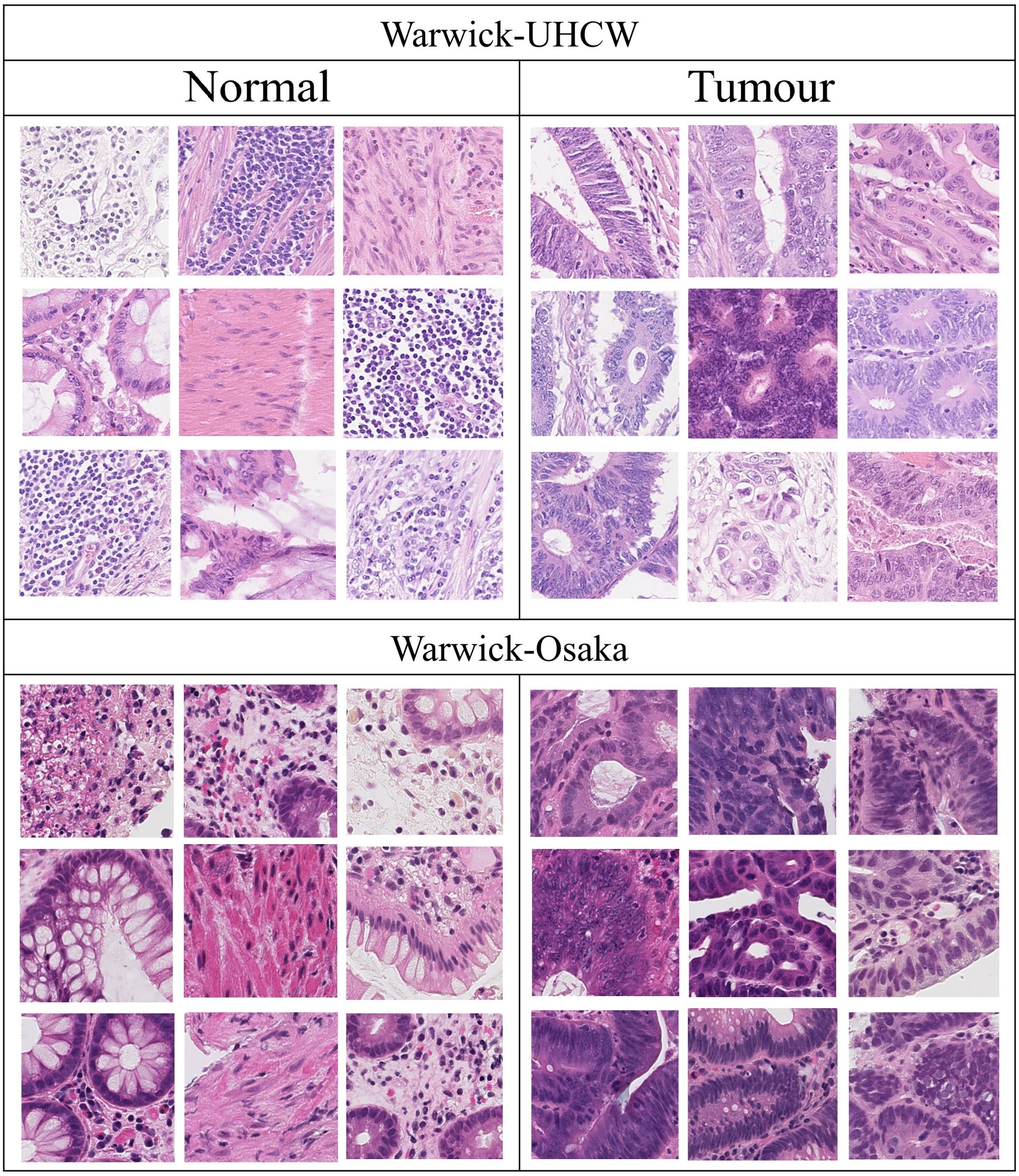


Figure 2.6: Representative tumour and non-tumour patches from the Warwick-UHCW and the Warwick-Osaka datasets.

Table 2.1: Comparison of various options for selecting the exemplar patches; results on the Warwick-UHCW dataset.

Method	Precision	Recall	F1-score	Accuracy
CNN (Section 2.3.2)	0.9272	**0.8513**	**0.8999**	**0.8924**
CNN (highly activated) [80]	0.9112	0.8309	0.8692	0.875
k-means	**0.968**	0.7559	0.8489	0.865
Random selection	0.8812	0.8309	0.8553	0.8594

approaches. Overall the results are in favour of the proposed method. Figure 2.6 shows a sample of 9 representative patches for tumour and non-tumour classes from both the datasets using the proposed method. The CNN (highly activated) [80] seems a straightforward approach that selects only patches where its neurons produce high activations. However, for a relatively large dataset, the exemplars may be overemphasised by a particular atypical WSI, where we normally have thousands of patches from a single case. Thus, this kind of approach is more suitable for a small dataset containing a limited number of image patches from which to select exemplars. Another downside is the inclusion of outliers as exemplars which can easily happen due to lack of precisely marked ground-truth. This argument remains valid for *k*-means and random selection approaches.

Tumour Segmentation on Adenocarcinoma Cases

In this experiment, we evaluate the performance on the Warwick-UHCW dataset of the proposed algorithms in comparison to some recently published algorithms for tumour segmentation. The experiment is conducted on all 75 adenocarcinoma WSIs where we used 75,000 randomly selected patches from 50 cases for training and the remaining 37,500 patches from 25 WSIs for testing. For a fair comparison, we retrained the selected algorithms on our dataset except for HyMap [47], which is an unsupervised method. For comparative analysis we selected only algorithms that are closely related to CRC image analysis, provided that their source codes were released with proper implementation details and comments. The features of the selected algorithms are as follows:

- Multi-class texture analysis [48] or MCTA: This method first computes a set of textural features on a given image patch including lower-order and higher-order histogram statistics, local binary patterns, grey-level co-occurrence matrix, Gabor filter, preception-like features and then fed these features into radial-basis function (RBF) support vector machine (SVM).

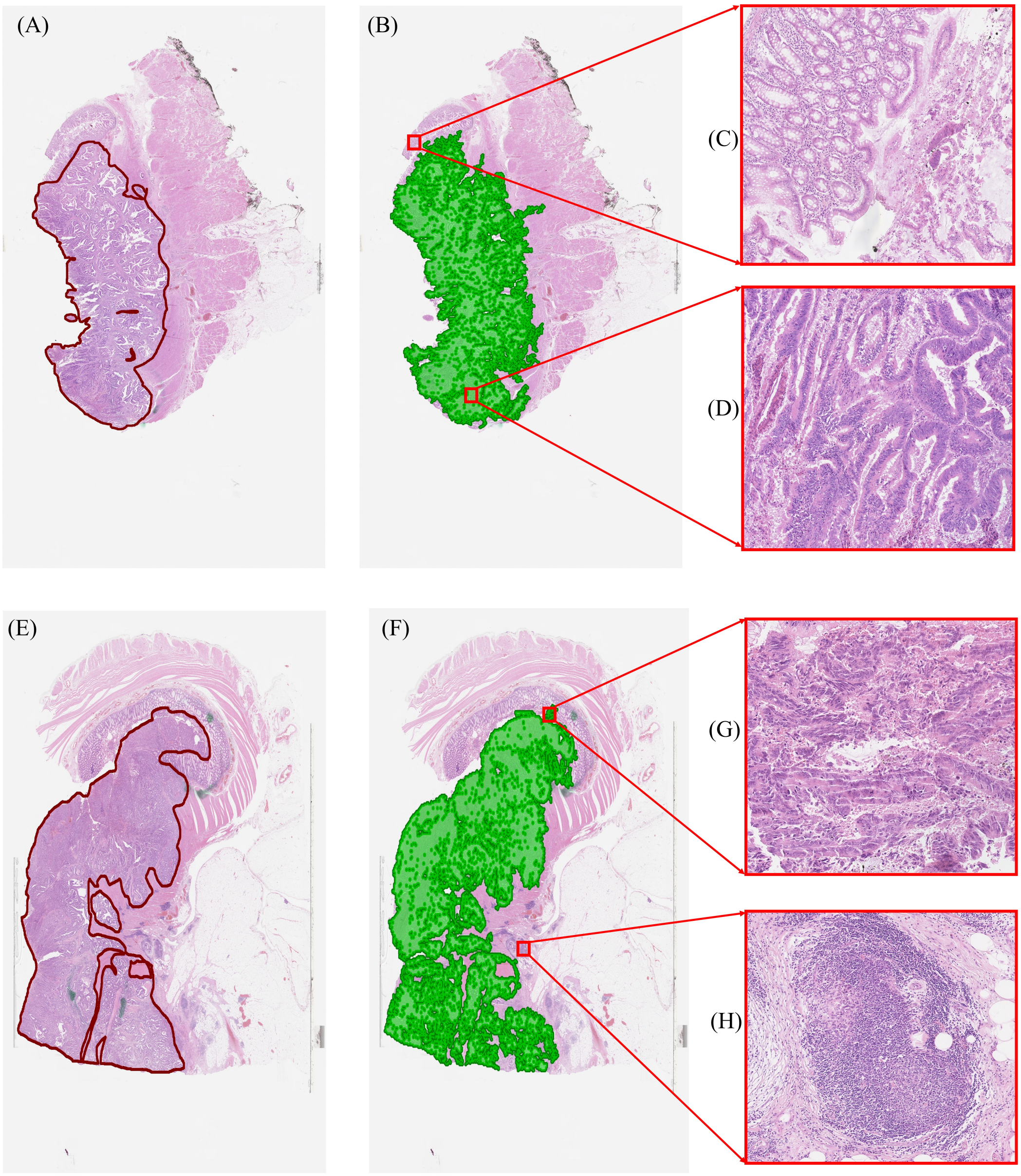


Figure 2.7: Results of accurate tumour segmentation on the whole-slide image (WSI) level. (A)&(E) input WSIs with annotated ground-truth, (B)&(F) tumour segmentation results, green region showing the predicted tumour areas, (C)&(H) zoom in regions containing true negatives (D)&(G) showing a sample of the true positive tumour segmented regions.

Table 2.2: Tumour segmentation results on the Warwick-UHCW dataset.

Method	Precision	Recall	F1-score	Accuracy
Fast Tumour Segmentation (PHP)	**0.9272**	0.8513	0.8999	0.8924
Accurate Tumour Segmentation (PHP+CNN)	0.9267	**0.922**	**0.9243**	**0.9245**
HyMap [47]	0.6851	0.8600	0.7626	0.7323
ConvNet CNN_3 [56]	0.856	0.867	0.8615	0.8606
MCTA [48]	0.8701	0.8834	0.8767	0.8843
TVIA [81]	0.8224	0.8641	0.8427	0.8387

- HyMap [47]: This is an unsupervised algorithm for tumour segmentation that classifies each pixel as hypo or hyper cellular by computing Gabor filter, texture energy and phase-gradient features followed by ensembling the projections for each feature. Since this is an unsupervised method, we did not retrain it for our experiments.

- ConvNet CNN_3 [56]: This method poses cancer detection as a two-class problem by assigning an 'invasive or non-invasive' label to each patch. The framework is a combination of convolutional layers followed by pooling operations and fully connected layers. The reported results suggest the CNN_3 outperformed the other counterparts on breast tissue so we used only CNN_3 for our experiment.

- Texture-based tumour viability image analysis [81] or TVIA: It classifies a given patch as viable or non-viable or as 'other tissue parts'. They computed local binary patterns and local contrast measures at the patch level and fed the extracted features into a SVM model. This algorithm is not directly related to tumour segmentation. We, therefore, evaluate it only as a two-class problem, that is, as a decision whether to classify a patch as tumour or non-tumour.

Table 2.2 reports comparative results from the experiment and Figure 2.7 shows qualitative results for the accurate tumour segmentation on the WSI level. Our PHP+CNN based accurate tumour segmentation produces the best results in terms of recall and F1-score, whereas the PHP based fast tumour segmentation performs best for the precision metric. The PHP+CNN outperformed the other competing methods by a reasonable margin. It is encouraging that incorporating deep features along with topological features boosts the overall performance. Generally, CNN models struggle to capture the rotation or viewpoint invariance of the learned object. To overcome this deficiency, we need to perform flipping or sev-

Table 2.3: Tumour segmentation results on the Warwick-Osaka dataset.

Method	Precision	Recall	F1-score	Accuracy
Fast Tumour Segmentation (PHP)	0.8259	0.8019	0.8137	0.8164
Accurate Tumour Segmentation (PHP+CNN)	**0.8311**	**0.8235**	**0.8273**	**0.8280**
HyMap [47]	0.6469	0.7228	0.6827	0.6641
ConvNet CNN_3 [56]	0.6927	0.8446	0.7612	0.735
MCTA [48]	0.7050	0.7419	0.7229	0.7157
TVIA [81]	0.6993	0.7240	0.7114	0.7063

eral arbitrary rotations on our images in the form of augmentation to make our model more generaliseable. Data augmentation overcomes the rotational invariance to some extent but it is still a non-trivial task to produce all possible rotations while training a CNN classifier. In contrast, PHP captures the rotational invariance by emphasising the merging and forming of homology classes so no matter how much a patch is rotated the PHP will remain persistent. The PHP also captures the biological phenomenon that the connectivity among tumour and non-tumour nuclei is significantly different. As compared to CNNs, the PHP based fast tumour segmentation algorithm only consults a number of exemplar patches from both classes in predicting the outcome of a patch. In our previous work [80], we observed that a similar approach performs marginally better than CNN. This offers a trade-off between accuracy and efficiency with reliable outcomes.

Tumour Segmentation on Adenoma, Carcinoma and Healthy Cases

The goal of this experiment is to test the generalisation of the proposed algorithms on another dataset that consists of different types of epithelial tumours and healthy cases. The experiment is conducted on 50 WSIs of the Warwick-Osaka dataset. We perform 2-fold cross-validation by selecting half of the dataset for training and the remaining half for testing. We then perform the same experiment by switching the training and test datasets. Similarly to the experiment with the Warwick-UHCW dataset we perform comparative analysis on the aforementioned selected algorithms and by retraining them on the Warwick-Osaka dataset.

Table 2.3 shows the results from 2-fold cross-validation. It can be observed that the proposed methods for all 3 metrics produce comparable results on an independent dataset. One of the arguments for ascribing relatively low performance by MCTA, HyMap, and TVIA is the selection of textural features, especially local binary patterns, local contrast measures, histogram statistics, and the grey-level co-occurrence matrix, that are sensitive to stain variations and image blurring. An additional cause for the poor performance of these programs is that, due to varying staining protocols at different centres, the morphological appearance of lymphocytes and benign epithelial nuclei from one centre can resemble that of malignant epithelial nuclei from another.

Results for Healthy Cases

This experiment compares the performance of different tumour segmentation algorithms for healthy cases in the Warwick-Osaka dataset. This is important for routine clinical practice as well as in cancer screening studies, which involve a large number of normal cases. The most challenging sections are healthy epithelial and lymphocyte regions where nuclei are densely packed and pose difficulties in identifying those regions as non-tumour. We evaluate the performance of this experiment by the specificity metric (also known as the true negative rate), that measures the involvement of negative samples misclassified as positive. Figure 2.8 shows the summarised results for all the healthy cases involved in this study. The specificity analysis demonstrates the effectiveness of the proposed ensemble strategy for accurate tumour segmentation, showing that a model relying on an agreement between topological and deep features outperforms all competing approaches.

Robustness Analysis

The aim of performing this experiment is to evaluate the robustness of the proposed tumour segmentation algorithms by training on the Warwick-UHCW dataset and testing on the Warwick-Osaka dataset which contains cancerous and healthy cases. One of the most challenging aspects of processing H&E WSIs is to overcome the non-standardised parameters involved in slide preparations like tissue sectioning, staining duration, dyes, and formalin concentration [82]. In order to become a part of the routine diagnosis, an automated tumour segmentation algorithm should be resilient to these data variations. With that in mind, we perform this experiment by keeping the parameters of the algorithms fixed during the training and testing.

Table 2.4 reports the results for robustness analysis and Figure 2.9 shows some qualitative results. The proposed methods attain the highest accuracy among different methods, demonstrating their robustness on strongly cross-validated data. The stain variability in both the datasets can be observed in Figure 2.6, which shows the selected exemplars from both the datasets. Regardless of stain variations in both the datasets, the degree of connectivity among tumour and non-tumour class is no-

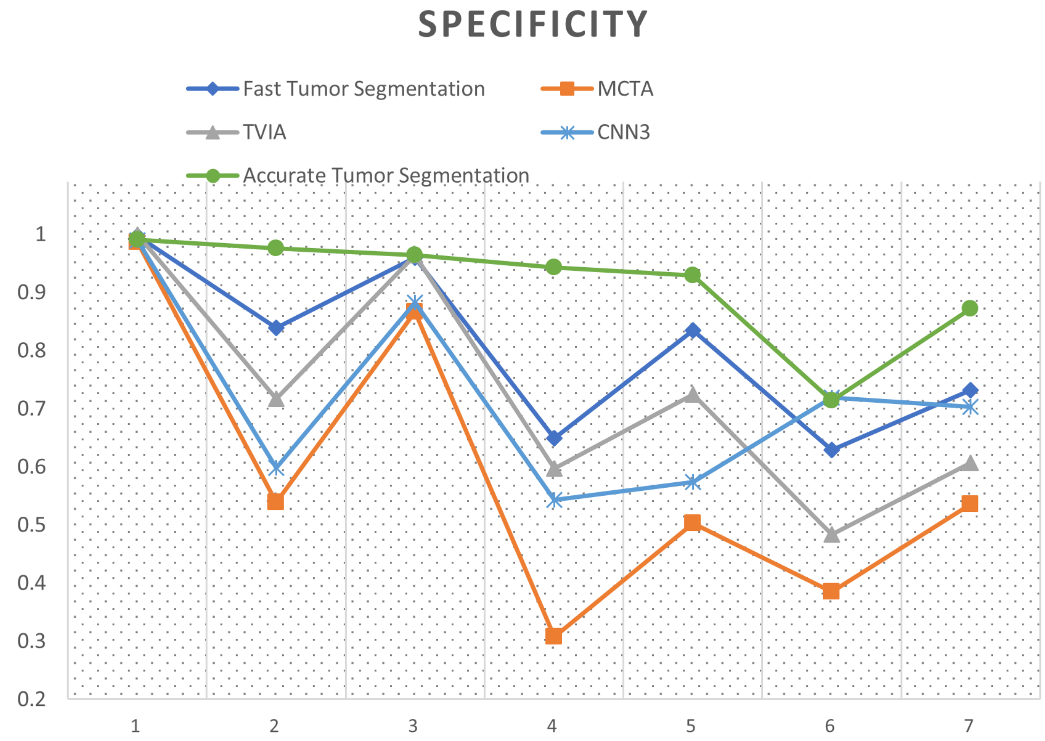


Figure 2.8: Specificity curve for tumour segmentation algorithms on healthy cases from the Warwick-Osaka dataset. The horizontal axis shows the index of healthy cases selected, and the vertical axis shows the specificity.

tably distinguishable and that leads to better performance of the proposed methods. It is interesting to note that the PHP based fast tumour segmentation marginally outperformed the PHP+CNN based accurate tumour segmentation (0.3%). This bodes well for the potential generalisability of fast tumour segmentation. Relatively smaller values of F1 measure for other competing algorithms can also be noticed in Table 2.4. Excluding HyMap, an obvious indication for a considerable drop in F1-score is due to the precision measure. On the one hand, the competing algorithms perform well in predicting tumour patches but at the cost of a large number of false positives. In contrast, regardless of stain variations in the two datasets, the fast tumour segmentation algorithm maintains a balance between precision and recall. On the whole, there is a decrease in performance accuracy for all algorithms as compared to their performance when trained and tested on data from the same centre, although the proposed algorithm still gives almost 8-10% higher F1-score than other competing algorithms. One potential strategy to increase the robustness of an underlying model is to train it on datasets from a number of different centres. Eventually, this will reduce the chances of overfitting and enables the learnable weights

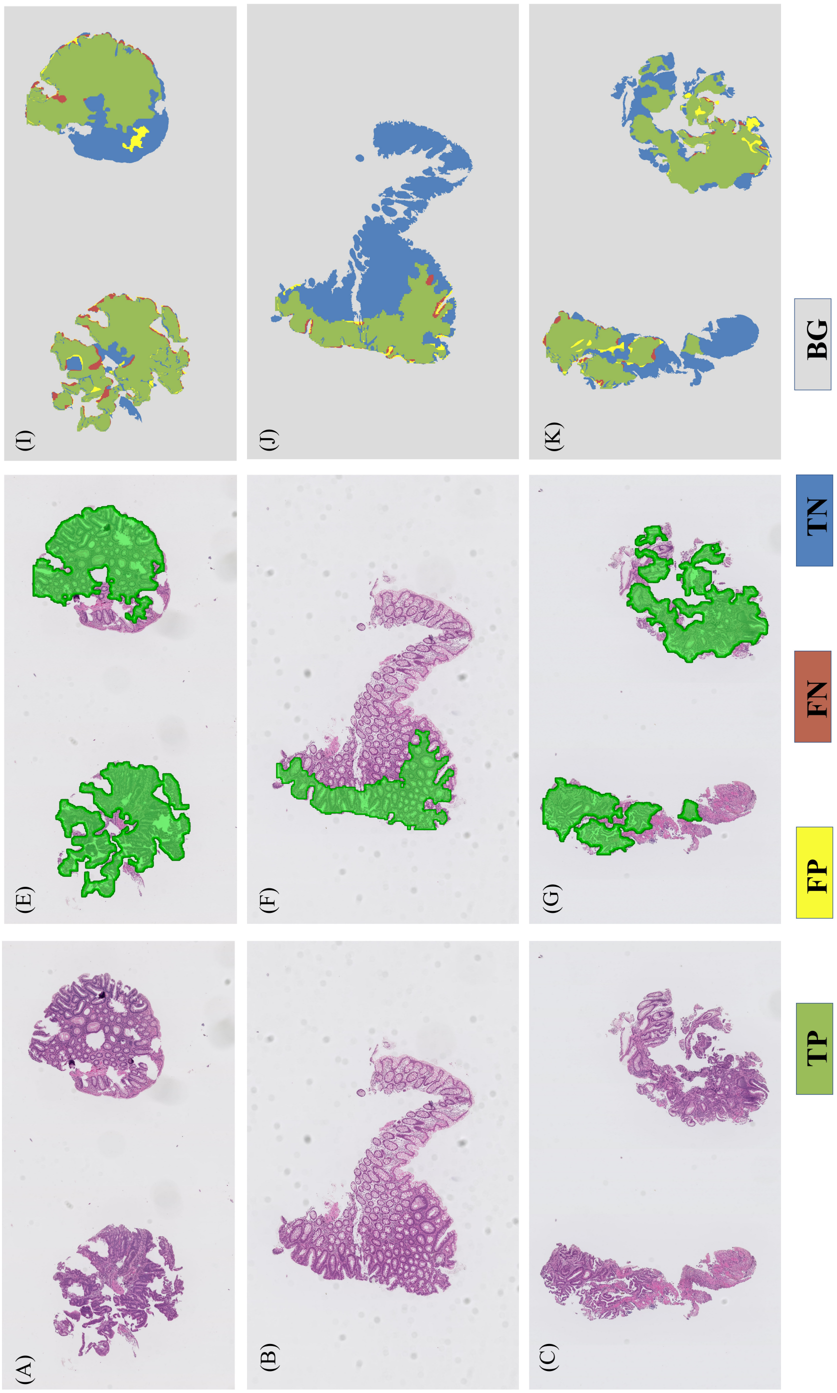


Figure 2.9: Robustness Analysis: results for the fast tumour segmentation on selected whole-slide images (WSI) from the Warwick-Osaka dataset. (A)-(C) input WSIs, (E)-(G) predicted tumour regions, (I)-(K) results for true positives (TP), false positive (FP), false negative (FN), and true negative (TN) regions.

to optimise on a variety of datasets with varying contents and staining protocols.

Table 2.4: Results for robustness analysis of different tumour segmentation algorithms.

Method	Precision	Recall	F1-score	Accuracy
Fast Tumour Segmentation (PHP)	**0.772**	0.7890	**0.7804**	**0.778**
Accurate Tumour Segmentation (PHP+CNN)	0.7413	0.8172	0.7774	0.7661
HyMap [47]	0.6469	0.7228	0.6827	0.6641
ConvNet CNN_3 [56]	0.6234	0.8167	0.7071	0.6616
MCTA [48]	0.5429	0.9586	0.6932	0.5758
TVIA [81]	0.5334	**0.9747**	0.6895	0.5610

Runtime Analysis

This section contains the runtime analysis of different tumour segmentation algorithms. Digitised WSIs are giga-pixel images, so fast tumour segmentation algorithms could play a crucial role in delivering 'real world' diagnostics. For all the algorithms, runtime analysis at the test stage is performed on a static machine having an 8-core processor with 3.1 GHz clock speed, 128 MB of memory and a GTX 1080 Ti graphical processing unit.

Table 2.5 reports the processing time for different algorithms on a patch of size 256×256×3 at 20×. The fast tumour segmentation is less computationally complex and an order of magnitude faster than competing algorithms, specifically ≈ 4.2 times faster than the CNN and ≈ 5.2 times faster than the accurate tumour segmentation. The algorithm for computing the PHPs is the foundation of fast tumour segmentation. The topological features (β_0, β_1) are computed by enumerating the connected components for a given filtered space. Consequently, the fast tumour segmentation algorithm is far less computationally complex than a multi-layer convolutional network.

2.6 Discussion and Conclusions

This study aimed at improving the diagnostic workflow by introducing a novel automated tumour segmentation framework for colorectal cancer histology WSIs. Experimental results conducted on fairly challenging datasets collected from two independent pathology centres demonstrate the efficacy and generalisability of persistent homology in histopathological image analysis. In this work, we present novel

Table 2.5: Runtime analysis in milliseconds for different tumour segmentation algorithms

Method	Time
Fast Tumour Segmentation (PHP)	**28.8** ms
Accurate Tumour Segmentation (PHP+CNN)	151 ms
HyMap [47]	931 ms
ConvNet CNN_3 [56]	97.1 ms
MCTA [48]	228.12 ms
TVIA [81]	109 ms

topological signatures (PHPs) that, in some respects, resemble clinicians' approach for identification of tumour enrich areas. It is evident from the performance that incorporating topological features and deep convolution features can enhance the overall performance of a CNN for the task of accurate tumour segmentation. The proposed framework was shown to work well on colorectal epithelial tumours of different histology grades.

Histological assessments are generally estimated visually and producing subjective measures for quantification of morphological features [83]. Computer-assisted image analysis requires precise annotations at high resolution to effectively train an underlying model. The inevitable fact is that the domain experts are generally not available for this laborious task of providing precise ground-truth at high resolutions. In such circumstances, the performance of a trained model may be affected by expert annotation that is carried out too rapidly or without close attention to detail, or by inexpert annotation. After careful consideration, we decided that this could be a relevant factor in both datasets, resulting in a difference in performance of the tumour segmentation algorithms (Table 2.2 & 2.3), regardless of stain and morphological variability. Nearly all the tumour segmentation algorithms experience difficulties in some of the benign epithelial and dense lymphocytic regions, especially where the intra-cellular region displays morphology that is similar to the cancer-distorted nucleoplasm. In this work, we captured important morphological differences between normal and cancer nuclei using persistent homology. However, these morphological differences may also be explored using level-set method [84, 85].

WSI scanners are becoming more viable for routine analysis [86], capable of producing hundreds of terabytes of data daily [35, 87]. The proposed framework offers a decent trade-off between speed and accuracy. With this change of paradigm, the fast tumour segmentation has enormous potential to overcome this ongoing challenge, reducing subjectivity and the pathologists' workload. It can be observed from the experimental results that careful selection of the exemplar patches can nearly obviate the need for retraining of our algorithm on a new dataset as shown in Table 2.4. One limitation of this work is that parameters like the number of exemplar patches and k (as in k-means) are empirically selected for this study and may need proper fine tuning, depending upon the data. The selection of a representative subset from a dataset is an active area of research and has several applications in computer vision and natural language processing. The proposed method for selection of exemplar patches is presented as an application to deep convolution networks. In the literature, there exist dissimilarity based subset selection methods [88, 89], but the selection of a dissimilarity measure is a non-trivial task and computing a dissimilarity matrix for a large dataset is computationally expensive. Other approaches like loopy belief propagation [90] can handle a large dataset but do not offer any guarantee of convergence. In contrast, the proposed fast tumour segmentation method exploits the learned activation maps to deduce the representative patches and is less computationally expensive and more robust to outliers. It is also evident from the experimental results that the accurate tumour segmentation algorithm presents a simple yet meaningful way of combining our novel topological signatures with deep convolution features.

In conclusion, we presented an automated tumour segmentation framework for colorectal cancer histology WSIs based on persistent homology and deep learning. The proposed framework is validated on two independent datasets, consisting of both malignant cases and healthy cases. Extensive comparative analysis demonstrated better than the state-of-the-art performance of the proposed algorithms. This study provides an insight into the topological persistence of an image and may constitute the first step towards interpretable incorporation of homology features in the domain of histopathology image analysis. The proposed homology profiles and the intrinsic phenomena of cell connectivity may be applicable to other similar problems in computational pathology.

Often after identification of tumour in a patient, pathologist needs further information regarding the overexpression of specific proteins. The quantification of antibodies carries huge clinical significance as it may help in planning appropriate treatment, or serve as a predictive or prognostic indicator. In the next chapter, we present a systematic study (HER2 Scoring Contest) on automated scoring of HER2 biomarker of invasive breast carcinoma cases.

Chapter 3

A Systematic Study on IHC HER2 Scoring Algorithms

The increasing importance of tissue-based biomarkers in stratified medicine [91] has recently received significant attention in the domain of computational pathology. Evaluating expression of the HER2 gene on IHC stained cases of invasive BC is a key part of the diagnostic assessment mainly due to its recognised importance as a predictive and prognostic marker in clinical practice [92]. This chapter presents a systematic study of the automated HER2 scoring contest, held in conjunction with the annual PathSoc meeting in Nottingham, June 2016, aimed at systematically comparing and advancing the state-of-the-art automated methods for HER2 scoring. In recent years, a number of systematic studies (competitions) have been designed in the domain of computational pathology to speed-up the development of artificial intelligence based diagnosis [32, 54, 82, 93, 94, 95]. This was the first-ever contest organised on IHC stained WSI which makes it distinctive from previously conducted studies, which were mainly based on H&E stained images. The algorithmic description of the automated methods presented in this chapter and in Appendix A is provided by the teams participated in this systematic study. Our contributions were more centred towards analysing/evaluating the presented results and designing the overall workflow of this study.

HER2 is a protein that influences the growth of malignant epithelial cells. The over-amplification of the HER2 gene is observed in BC cases having invasive tumour regions. Cancer cells in HER2+ (positive) cases of BC encounter neoplastic transformations which lead to uncontrolled growth of tumour cells as compared to HER2- (negative) cases. Approximately all the invasive breast carcinomas cases are recommended for HER2 testing [96] and nearly 20-30% of cases have over-expression

Table 3.1: Recommended HER2 scoring criteria for IHC stained breast cancer tissue slides [34].

Score	Cell membrane staining pattern	Staining assessment
0/1+	No membrane staining or incomplete membrane staining in $< 10\%$ of invasive tumour cells (0+) OR faint/barely perceptible or weak incomplete membrane staining in $> 10\%$ of tumour cells (1+)	Negative
2+	A weak to moderate complete membrane staining is observed in $> 10\%$ of tumour cells OR strong complete membrane staining in 10% of tumour cells	Borderline (Equivocal)
3+	A strong (intense and uniform) complete membrane staining is observed in $> 10\%$ of invasive tumour cells	Positive

of HER2 protein [15] which is associated with poor prognosis, lower survival, and high recurrence [97]. Recent studies have reported the HER2 status as a predictive factor for anti-HER2 and hormonal therapies and also a prognostic factor to associate invasive tumours with mortality and duration of recurrence free survival [92]. Therefore, precise quantification of HER2 over-expression is crucial for ensuring that HER2+ patients receive appropriate anti-HER2 treatment.

Gene amplification can also be identified through Fluorescence In Situ Hybridisation (FISH). Given the technical ease of performing IHC, it has become the preferred test and FISH is usually only performed when the IHC is equivocal. In a routine clinical practice, an expert histopathologist reports a score between 0 and 3+ and cases scoring 0 or 1+ are classified negative whilst cases with a score of 3+ are classed as positive. Cases with score 2+ are classified as equivocal and are further assessed by FISH to test for gene amplification. Examples of the four different HER2 scores (0 to 3+) are shown in Figure 3.1. A summary of recommended guidelines for HER2 IHC scoring criteria [34] is shown in Table 3.1. Historically, it was reported that up to 20% of the HER2 IHC results may contain inaccuracies [98] due to variations in the technical quality and the subjective nature of scoring. Although adoption of the HER2 guidelines and recommendations [34], have served to improve standards in HER2 testing, there remain challenging cases especially

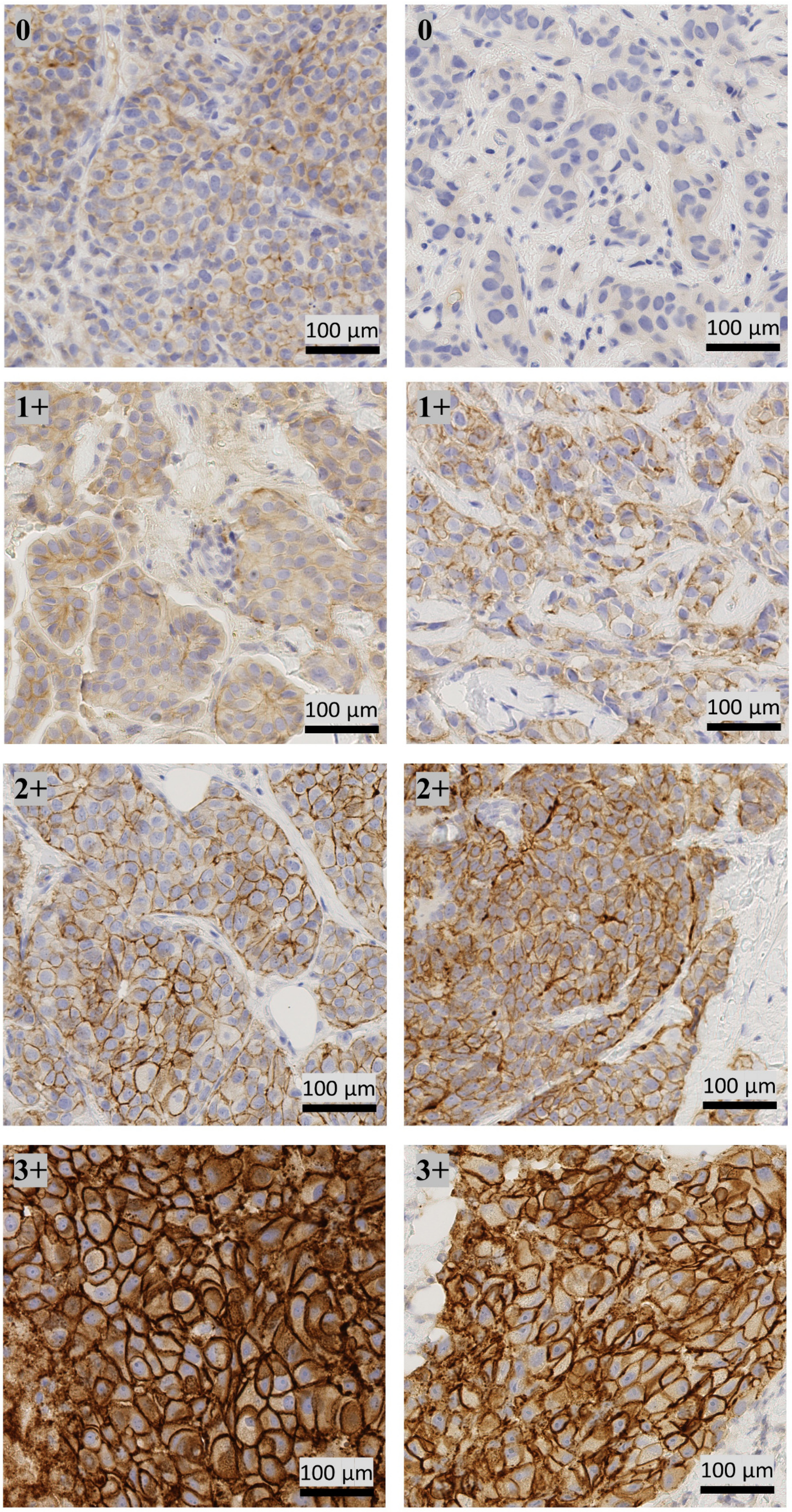


Figure 3.1: Top to bottom: example of regions of interest from WSIs of the training dataset.

with HER2 scores deemed borderlines between categories. Automated IHC scoring of HER2 carries a promise to overcome the existing problems in conventional methods. Automated scoring methods are relatively less prone to subjective bias and can provide precise quantitative analysis which may assist the expert pathologist to reach a reproducible score.

3.1 Materials and methods

3.1.1 Ethics

The ethics approval was provided by Nottingham Research Ethics Committee 2 [Approval No: REC 2020313]; R&D reference (N) 03HI01.

3.1.2 Image Data Acquisition and Ground Truth

The histology slides for this contest were scanned on a Hamamatsu NanoZoomer C9600 enabling the scanned images to be viewed from a 4× to a 40× magnification, making the process comparable to a clinicians standard microscope. The contest dataset entailed 172 WSIs extracted from 86 cases of invasive breast carcinomas and included both the H&E and HER2 IHC stained slides. The actual HER2 scoring is normally done on the IHC stained slides whilst the H&E slides assist the expert pathologist to identify the areas of invasive tumour and discriminate these from areas of *in situ* disease. Figure 3.2 shows an example of the two types of WSIs (with a corresponding zoomed-in region of interest) from the contest dataset.

The ground-truth was taken from the clinical reports issued on the cases at a tertiary referral centre for breast pathology (Nottingham University Hospitals, NHS Trust). At this centre, each case had been reported or reviewed by at least 2 specialist consultant histopathologists as part of their routine practice (preliminary reporting and MDT review). The centre provides regular internal quality control for HER2 assessment for immunohistochemistry runs and regularly contributes and participates in the UK NEQAS (National External Quality Assessment Scheme) for immunocytochemistry and *in situ* hybridisation (ICC & ISH).

3.1.3 Contestants

A total of 105 teams from more than 28 countries registered to access the training dataset before the end of the registration deadline. By the end of submission deadline (for the off-site contest), a total of 18 submissions from 14 teams were received for evaluation. We provided an opportunity to each of the 14 teams for presenting their

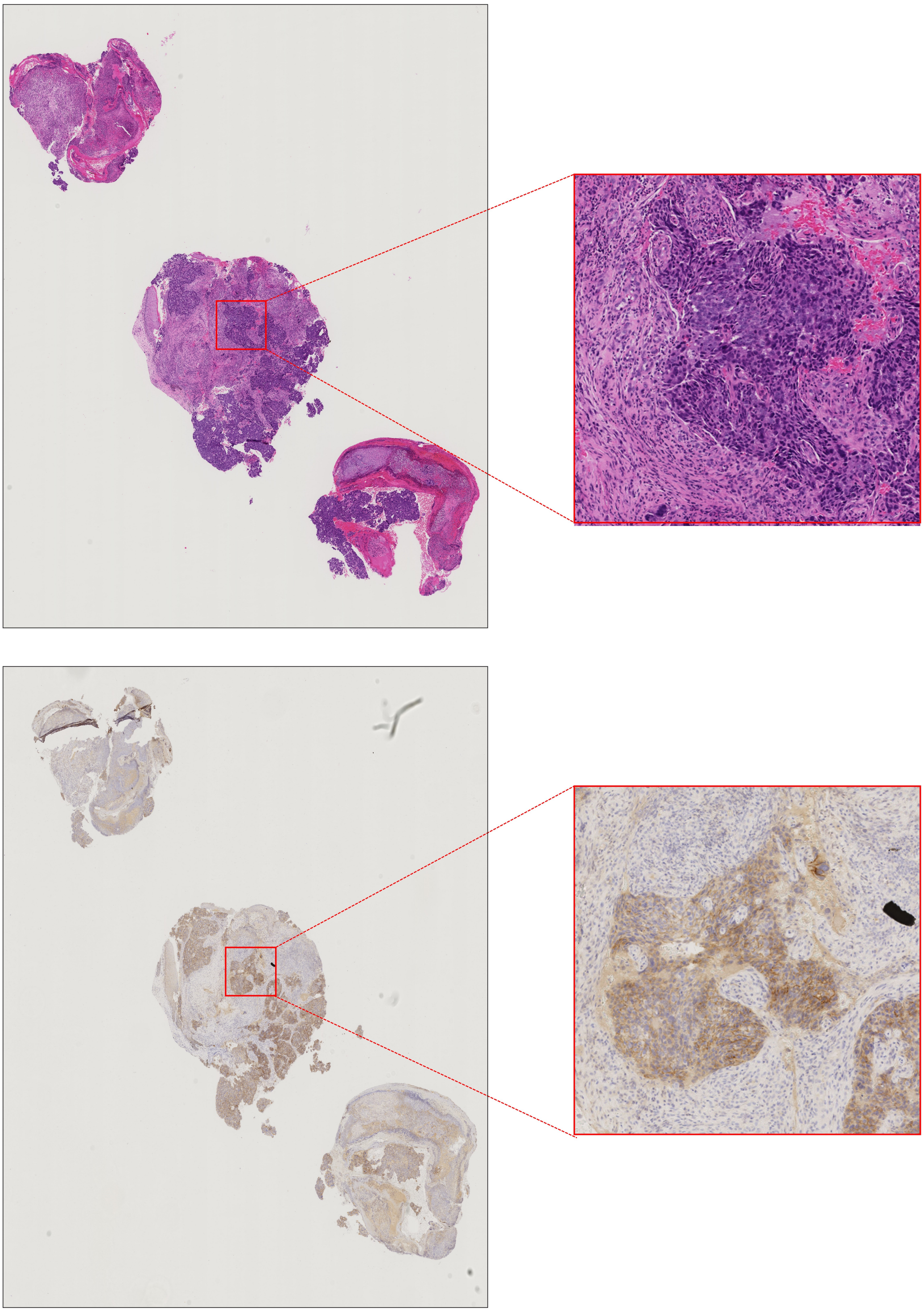


Figure 3.2: An example of WSIs along with a zoomed-in cross-sectional area showing the tumour region (*top*) H&E stained slide, and (*bottom*) IHC stained slide.

Table 3.2: Agreement points for predicted calls of ground-truth (GT).

	Erroneous predicted scores				
Ground Truth	Score	0	1+	2+	3+
	0	15	15	10	0
	1+	15	15	10	0
	2+	2.5	2.5	15	5
	3+	0	0	10	15

Table 3.3: Bonus point criteria, when PCMS lies in a certain range of the GT value of the PCMS.

Ground truth	Percentage with Complete Membrane Staining (PCMS)	
0	0	0
1+	1 (PCMS <3%)	3 (PCMS ± 2)
2+	5 (PCMS ± 5)	2.5 (PCMS ± 10)
3+	5 (PCMS ± 5)	2.5 (PCMS ± 10)

approach in the contest workshop and 6 teams chose to present. For the *Man vs Machine* contest, we received the markings from 4 pathologists. The contest website was reopened for new submissions after concluding the workshop.

3.1.4 Evaluation Measures

The performance of each submitted algorithm was evaluated based on three criteria: 1) agreement points, 2) weighted confidence, and 3) combined points. Each assessment criterion has a separate leaderboard. The evaluation criteria were rationalised according to the clinical significance and implications of HER2 IHC scoring as follows: in everyday clinical practice, for a score of 0 and 1+: No Herceptin is offered to the patient; for 3+ score, Herceptin is offered. For an IHC 2+ score, a FISH test is performed; if positive (i.e.) there is evidence of gene amplification and Herceptin is offered while for a negative result, it is not offered. The evaluation considers the impact of erroneous classification. For example, a score of 0/1+ being interpreted as 3+ or vice versa is a serious error while a 2+ scored as 0/1+ denies a few patients of valid treatment; a score of 3+ for a 2+ case bypasses the FISH test and may erroneously treat few cases (which would have been FISH negative) with toxic drugs while a score of actual 3+ downgraded to a 2+ calls for additional expense of FISH testing but the end result will probably be the same and hence this should not be regarded as that serious an error. These have been summarised in Table 3.2.

For agreement points, a penalty method was employed whereby each erroneous prediction is penalised with respect to its deviation from the GT as shown in Table 3.2. In routine clinical practice, pathologist also estimates the percentage of cells with complete cell membrane staining (PCMS) along with HER2 score. To increase the objectivity in evaluating the automated algorithms, we devise a bonus criterion as shown in Table 3.3, where the decision was made on the PCMS regardless of the intensity. The bonus points were primarily introduced for score 2+ and 3+ as they attain more clinical significance. For the IHC score 1+, 1 bonus point was awarded if there was an accurate prediction of the IHC score and PCMS $< 3\%$, while 3 bonus points were awarded if there was an accurate prediction of the IHC score and PCMS $> 3\%$ but the predicted PCMS value only deviated $\pm 2\%$ from the GT. For the IHC scores 2+ and 3+, 5 bonus points were awarded if there was an accurate prediction of the IHC score and PCMS only deviated $\pm 5\%$ from the GT. Similarly, 2.5 bonus points were awarded for score 2+ and 3+, if there was an accurate prediction of IHC score and PCMS only deviated $\pm 10\%$ from the GT.

For the development of an interactive diagnostic tool, it is important that automated algorithms identify such cases which require further examination by pathologists before concluding the final outcome. In this regard, we devise a weighted confidence criterion that may indicate those cases where further examination is required. The criterion to measure the weighted confidence w_c was distinct for both truly and wrongly classified cases. In cases where the predicted HER2 score p_s matched with the GT with higher confidence c, the weighted confidence amplified the confidence value for true prediction whereas wrong predictions with high confidence were penalised accordingly, as given in equation (3.1).

$$w_c = \begin{cases} \frac{(2c - c^2 + 1)}{2}, & \text{if } p_s == \text{GT} \\ \frac{(-c^2 + 1)}{2} & \text{otherwise} \end{cases} \tag{3.1}$$

The third assessment criterion is a combination of both agreement points and weighted confidence based evaluations. The combined points were calculated by taking the product of two assessment criteria for each case.

3.2 Organisation

3.2.1 Stage 1: Release of the Training Data

In the first stage, the training dataset comprising 52 cases were released to the registrants, through a secure website portal. The dataset consisted of IHC and H&E stained images and the GT. The GT score and PCMS were released for the training dataset. At this stage, most of the details regarding contest (like tasks, contest rules, contest forum details, etc) were already posted to the contest website and the registered teams started developing their automated algorithms for HER2 scoring. The registration process remained open for five weeks. We also created a social-forum (a Google group) for the participants to share their queries and to communicate with the organisers.

3.2.2 Stage 2: Release of Off-Site Test Data

A dataset comprising 28 cases was selected for the off-site testing. The test dataset consisted of IHC and H&E stained WSIs, the GT for the test datatset was not released to the participants to ensure a fair evaluation. Source code for performance assessment in both MATLAB and Python languages were also released to the registrants. The registrants were given more than a month after releasing the test data to finalise and submit their scoring methods for the announced tasks.

3.2.3 Stage 3: Submission of Results (Off-Site)

The deadline for submission of results for the test dataset was set to be a week before to the contest workshop. Each team had to submit results in a comma-separated values (CSV) file along with a maximum 2-page summary of their algorithms, a description of the experimental setup, and some preliminary results. The participants were advised that the CSV file should contain the predicted HER2 score, the confidence value for predicted score, and the PCMS for each WSI in the test dataset. Each registrant was allowed to submit up to three sets of results. The submitted results were evaluated but outcomes were not announced until the contest workshop was held.

3.2.4 Stage 4: Contest Workshop

The contest workshop covered three main events: a) a brief talk from the organisers and the participants where 6 teams were invited for a brief presentation to give an overview of their approaches and experiments, b) announcement of the comparative results for both off-site, and c) for the *Man vs Machine* comparison as a part of the on-site contests. The remaining 6 cases (of the 86) were used for an on-site competition (although they were released one day before the contest workshop due to the computational requirements of some of the automated algorithms and their results are not discussed). The complete tables of results are available on the contest website.

3.3 Results

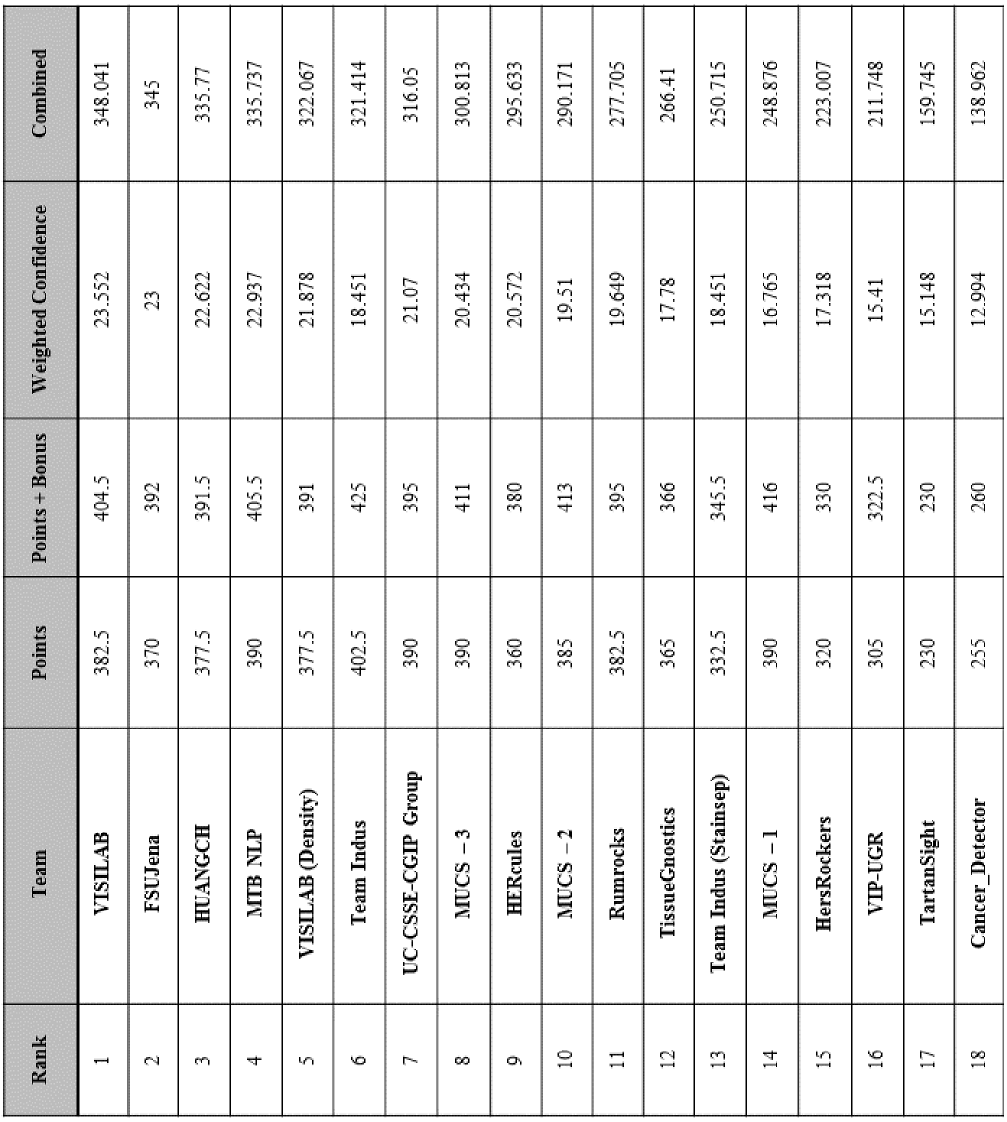
3.3.1 Leaderboards

Comprehensive results comprising all the submissions for automated methods are shown in Figure 3.3. The teams were sorted in descending order with respect to the combined and bonus points based assessment. For the off-site contest, the total possible points were 420 (28 cases with a maximum of 15 points each) whereas, for weighted confidence, the maximum points were 28, 1 for each case. The top three teams with respect to point based assessments were Team Indus, MUCS-1, and MUCS-2 whereas, according to weighted confidence assessment the top-ranked teams were VISILAB, FSUJena, and MTB NLP. The combined results rank the top three teams in the following order: VISILAB, FSUJena, and Huangch. The performance of top-ranked teams including bonus points and the trend for total points (without the bonus points) can be seen in Figure 3.4. MUCS-1, MUCS-3, CS-UCCGIP, and MTB NLP achieved equal points but MUCS-1 secured more bonus points as their PCMS was more accurate as compared to the remaining counterparts. Similarly, VISILAB and Rumrocks ended up in a tie where both teams attained equal points but the VISILAB method was more precise in predicting PCMS. Comprehensive tables for all three leaderboards are available for download from the contest website[1].

3.3.2 Summary of Automated Methods

Most of the automated methods (described in Appendix A.2) applied a supervised patch based classification approach for solving this problem. The most common pipeline was based on three main components: 1) pre-processing including some sort of automated or interactive method to identify the potential regions of interest for training the underlying model, 2) classification based on handcrafted or neural

[1] https://warwick.ac.uk/tialab/her2contest/outcome



Rank	Team	Points	Points + Bonus	Weighted Confidence	Combined
1	**VISILAB**	382.5	404.5	23.552	348.041
2	**FSUJena**	370	392	23	345
3	**HUANGCH**	377.5	391.5	22.622	335.77
4	**MTB NLP**	390	405.5	22.937	335.737
5	**VISILAB (Density)**	377.5	391	21.878	322.067
6	**Team Indus**	402.5	425	18.451	321.414
7	**UC-CSSE-CGIP Group**	390	395	21.07	316.05
8	**MUCS – 3**	390	411	20.434	300.813
9	**HERcules**	360	380	20.572	295.633
10	**MUCS – 2**	385	413	19.51	290.171
11	**Rumrocks**	382.5	395	19.649	277.705
12	**TissueGnostics**	365	366	17.78	266.41
13	**Team Indus (Stainsep)**	332.5	345.5	18.451	250.715
14	**MUCS – 1**	390	416	16.765	248.876
15	**HersRockers**	320	330	17.318	223.007
16	**VIP-UGR**	305	322.5	15.41	211.748
17	**TartanSight**	230	230	15.148	159.745
18	**Cancer_Detector**	255	260	12.994	138.962

Figure 3.3: A summary of results of all three assessment criteria for the automated HER2 scoring contest, ordered by the combined points criterion.

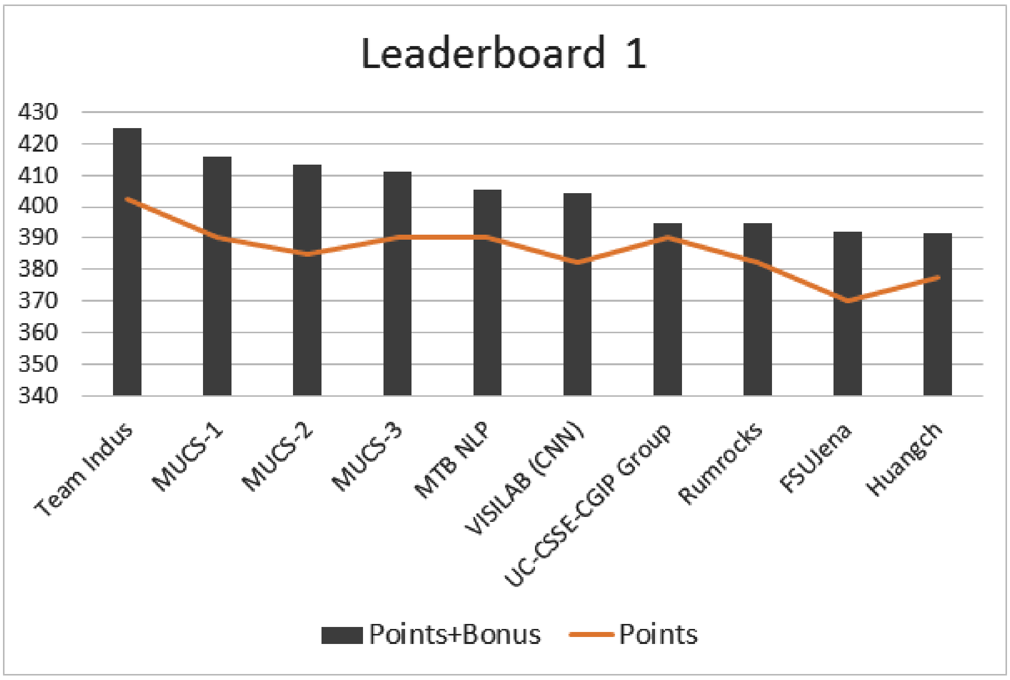


Figure 3.4: Combined results for top ranked teams with respect to agreement and bonus points.The trend shows the significance of correctly predicting the percentage of cell membrane.

network learned features, and 3) post-processing techniques to aggregate the HER2 score at WSI level and to estimate the PCMS. Deep learning, especially CNN based approaches dominated as most of the methods were based on CNN. The majority of the CNN architectures (Team Indus, MUCS-(1-3), MTB NLP, VISILAB, RumRocks, FSUJena) were inspired by state-of-the-art deep neural networks [52, 99].

In the pre-processing and patch extraction stage, most of the teams followed the conventional thresholding techniques with a combination of morphological operators. These techniques are computationally less expensive and generally work well as background regions lack any texture contents in contrast with other tissue components. The MUCS-(1-3), MTB NLP, VISILAB, and FSUJena manually probe the regions of interest through some calibration or customised methodologies. These methods aimed to pick the best possible regions for training their algorithm, generally without affecting the testing phase. To segment tissue regions, RumRocks team implemented a deconvolutional neural network (DCNN) and a 2D CNN for selection of patches based on their texture. The Huangch team performed mean filtering and stain normalisation by using the control tissue intensity values to calibrate the stain colour intensity as a pre-processing step.

In the second step, most of the teams employed deep learning approaches whereas other teams such as CS-UCCGIP and Huangch used shallow machine learning algorithms on handcrafted characteristic curves derived from pixel intensities. Team Indus used a combination of data-driven and handcrafted features. They combined the average control tissue intensity with activation maps of the last convolutional layer, before feeding them into fully connected layers. Some of the top-ranked teams deployed off-the-shelf models such as Alexnet [52] and GoogLeNet [99] for predicting the HER2 score. The FSUJena team computed a set of bilinear features from the convolution activation maps of the AlexNet. The derived activations contain the learned feature maps representing a d-dimensional $w \times h$ spatial grid. This approach enables them to perform their analysis on top of the convolutional features. In combination with standard approaches for data regularisation, MTB NLP and RumRocks trained multiple models. The final HER2 score and PCMS were estimated by averaging over all the models. Additionally, a wide range of data augmentation and regularisation techniques were employed to overcome the overfitting issues. As in practice, the standard data augmentation techniques such as affine transformations (e.g. rotation, flip, translation), random cropping, blurring, and elastic deformations were applied to train the network. MUCS-2, MTB NLP, and RumRocks extensively used the data augmentation techniques for improved generalisation on unseen data.

In the final stage of post-processing and predicting the PCMS, most of the teams employed standard image processing and shallow machine learning approaches to the results attained from the last step. A Random Forest classifier was trained by MTB NLP to produce the final class probabilities and to estimate the PCMS. FSUJena simply used the mean tumour cell percentage as seen in the training dataset for a particular class as an estimate. Team Indus was the only participants to used both IHC and H&E stained slides to estimate the PCMS using standard image processing approaches like contour detection, thresholding, and morphological features. Nearly, all the proposed methods extract image patches from high magnifications (10× or above) excluding MUCS and Rumrocks, as their proposed methods identify ROIs from a lower magnification.

3.3.3 *Man vs Machine*

Organisation

One way of evaluating the automated algorithms for IHC (HER2) scoring is to perform a comparative analysis between the assessment of expert pathologists and

Table 3.4: Summary Results for the *Man vs Machine* event. The evaluation was carried out according to the contest criteria as described in the Evaluation Section.

Rank	Automated methods & pathologists	Score	Bonus	Score+bonus
1	Team Indus	220	12.5	232.5
2	Expert 2	210	20.5	230.5
3	VISILAB	212.5	15	227.5
4	MUCS_1	205	20.5	225.5
5	Expert 1	185	10	195
6	Expert 3	180	13	193

predictions of automated methods for a handful of cases as compared to the scores for those cases as agreed by at least two consultant breast-pathologists (GT). On the day of the contest workshop, we organised an event called *Man vs Machine*. The main aim of this event was to analyse the performance of the proposed methods and to explore the disagreements among conventional and automated methods. This kind of analysis can lead us to a more sophisticated protocol for automatic HER2 scoring and to overcome the inter- and intra-observer agreements that can be found in normal practice.

This analysis was performed on a subset of 15 cases from the off-site test dataset. We also set up an online web-page for the pathologists. The web-page enabled the experts to load and navigate (including pan and zoom) through the WSI of those cases. Both IHC (HER2) and H&E stained digital images were made available to mimic the conventional scoring environment. We requested the expert pathologists on the contest day at PathSoc to score each case by providing the HER2 score, PCMS and a confidence value.

Comparative Results

Table 3.4 summarises the overall evaluation scores achieved by top-6 participant for this event. Each table entry gives the cumulative score for all 15 cases, which indicates the overall performance. The agreement-points based assessment was used to evaluate the performance of this event. In total, we received 4 responses from expert pathologists and we ranked the top-6 submissions including the top-3 automated methods. From submitted responses, three participant pathologists reported themselves as Consultant Pathologist and one as Trainee Pathologist and all three of them marked breast pathology as a sub-speciality.

As can be seen in Table 3.4, one of the automated methods slightly outperformed the top-performing participant pathologist[2]. These results point to the potential significance of automated scoring methods and the recent advancements in digital pathology. Its worth mentioning that automated HER2 scoring algorithms submitted in this contest are not ready to deploy in their current form, as they will require extensive validation on a significantly large-scale data and also plenty of input from experts to prepare the GT on the larger data.

Table 3.4 shows pooled data for HER2 scoring among top-3 automated methods, the scores from top-3 participant pathologists and comparison with the GT. Table 3.5 was determined for the 15 cases selected from the off-site contest dataset. On the basis of HER2 scores, a 100% agreement with the GT was observed for score 3+ among the participant pathologists and the automated methods. For the scores of 1+ and 2+, there were disparities between the GT and the new scores. In all cases except one, for both man and machine, the error resulted from overcalling the score. Thus, for the score 1+, on 6/9 (67%) were overcalled as 2+ by humans whilst 4/9 (44%) were overcalled by the machine algorithms. For the score of 2+, 7/15 (46%) were overcalled as 3+ by humans whilst machines overcalled 1/15 (6%) as 3+ and 1/15 (6%) was undercalled as 1+. Clinically, the score of 2+ is critical, as in routine practice, cases of score 2+ are recommended to go through FISH testing. Its equally important to avoid predicting the score 2+ as 1+ or 0, cases such erroneous prediction will deny the further assessment of HER2. As it can be seen in Table 3.5, none of the cases with score 2+ was misclassified by the participant pathologists as either 1+ or 0 whereas for one of the case an automated method wrongly predicted a score of 2+ as 1+.

Most of the incorrect predictions by the participant pathologists were found to be in cases where there was considerable heterogeneity. Two such examples are shown in Figure 3.5 and 3.6. In tumour cells of HER2 score 2+, a pattern of weak to moderate complete membrane staining is observed whereas, for score 3+, an intense (uniform) complete membrane staining is observed. Estimating the complete membrane staining is a difficult and highly subjective process especially for score 2+ and 3+, as it is extremely hard to pick up subtle differences in the morphological appearance for those cases.

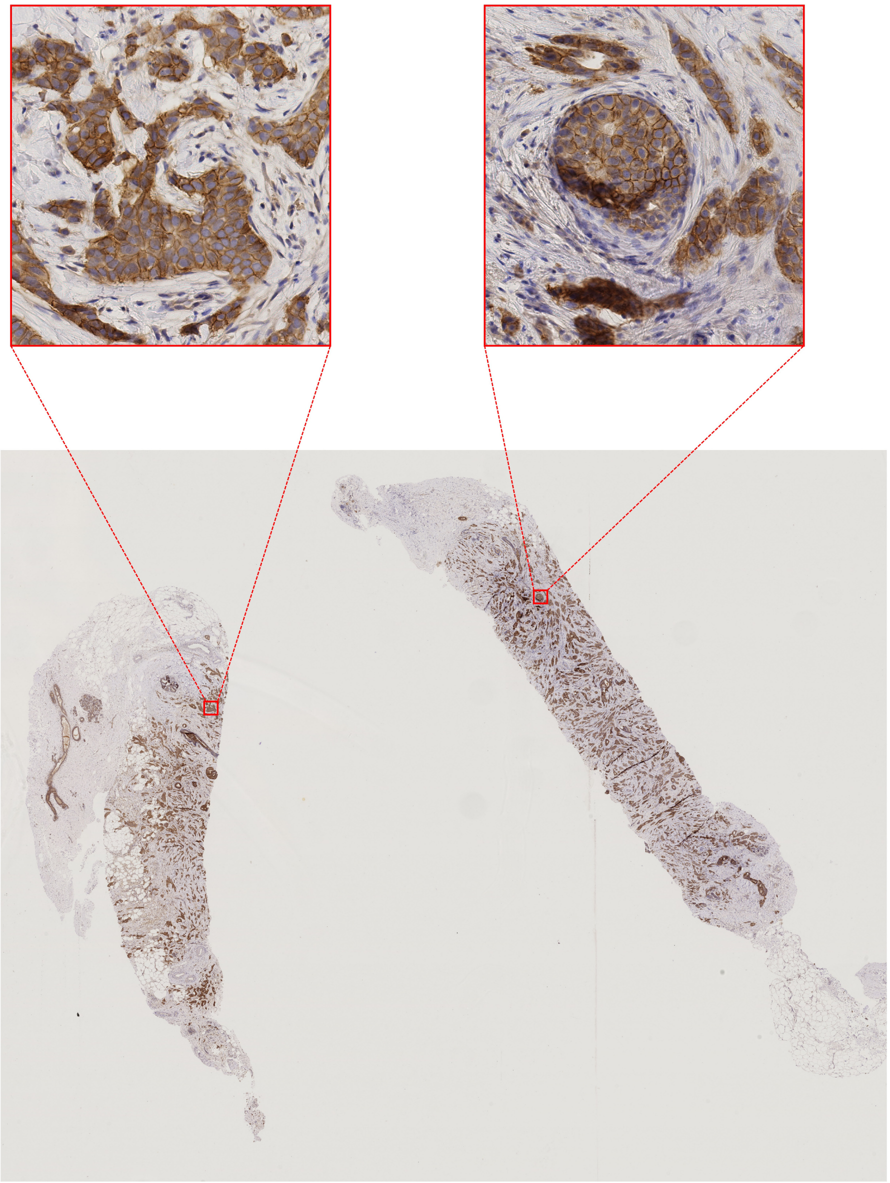
[2] The reported results are based on 15 HER2 cases so any change in the dataset or in evaluation criteria may influence these results.

3.4 Discussion

A major aim of organising this study was to provide a platform for computer scientists and researchers to contribute and to evaluate the performance of their computer algorithms for automated IHC scoring of HER2 images from BC tissue slides. Automated scoring may significantly overcome the subjectivity found due to varying standards adopted by different diagnostics labs. There is a current wealth of literature [100, 101] using individual platforms (both freely and commercially available) for digital analysis of HER2 in BC. This, however, was the first comparison of platforms and algorithms and provides a pilot for an independent comparison of computing algorithms for HER2 assessment on a benchmark dataset. Overall, the contest highlights the wealth of potential carried by AI techniques for assessment of IHC slides.

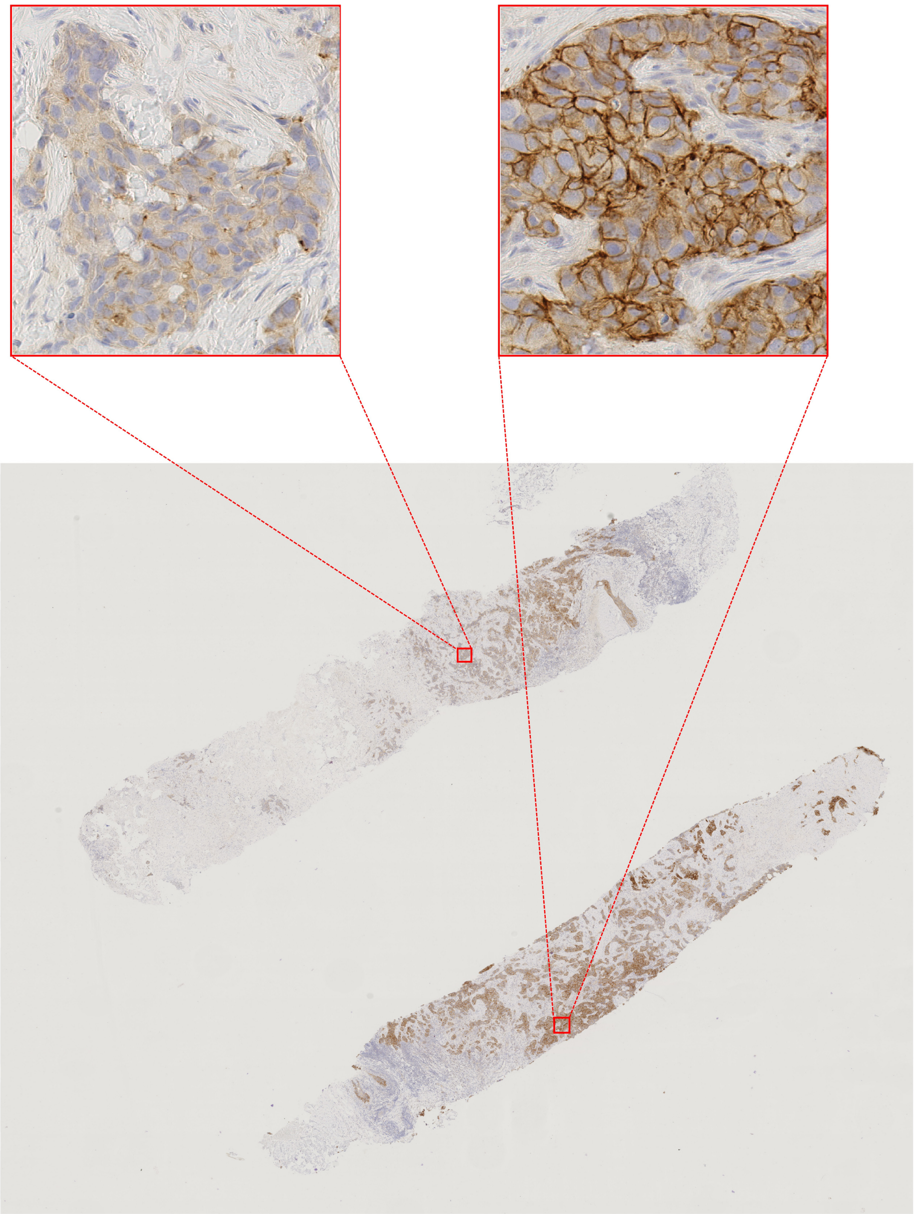
The contest training dataset was deliberately selected in a way that it contained a reasonable number of cases from all HER2 scores, bearing in mind the need for the training algorithms to learn features for each score. For the test dataset (both off-site and on-site), the GT was withheld at the time of results evaluation. Results showed that the automated analysis performed comparably to histopathologists. Many of the algorithms achieved high accuracy often close to the maximum. Our main objective was to analyse the performance of algorithms based on clinical relevance and hence the three particular evaluation criteria described above were chosen. The evaluation criteria were decided according to the clinical significance and implications of HER2 with the help of two pathologists, who were also involved in organising this study. Although, agreement (Table 3.2) and bonus points (Table 3.3) criteria involved clinicians input and may have clinical relevance but both criteria were not inspired by any quantitative studies. One of our objectives in organising this study was to provide a publically available dataset for benchmarking and accelerating the development of automated scoring algorithms. It is quite possible that other assessment criteria may influence the ranking (as given in Figure 3.3) of comparative results. Therefore, it would an added advantage if some sort of statistical measure (like concordance metric) would also be incorporated along with clinically relevant measures to evaluate the correlation between ground-truth and predicted HER2 score.

The data from the *Man vs Machine* comparison showed that, reassuringly, all participants (whether human or computer) correctly identified cases with the GT score of 3+. This means that no-one in the category would have been denied treatment. Similarly, for the cases with a score of 0 or 1+, although there was some



Ground Truth	Path_1	Path_2	Path_3	Team Indus	MUCS	VISILAB
2	3	2	3	2	3	2

Figure 3.5: An example showing IHC stained WSI and zoomed-in cross sectional areas with corresponding HER2 GT scores marked by expert pathologists and predictions from top 3 automated methods.



Ground Truth	Path_1	Path_2	Path_3	Team Indus	MUCS	VISILAB
2	2	2	3	2	2	2

Figure 3.6: An other example of WSI and zoomed-in cross sectional areas with corresponding HER2 GT scores marked by expert pathologists and predictions from top 3 automated methods.

overcalling, this never exceeded 2+ and thus none would have received treatment without further testing. The most problematic category was, not unexpectedly, cases with a score of 2+; in both human and machine evaluations. If overcalled as 3+, the FISH negative subset would be over-treated. The GT information for the FISH results was not released to the participants as the contest was aimed just at comparing the interpretation of HER2 IHC results. Hence, most of the automated algorithms aimed at predicting the equivocal cases as 2+. Table 3.5 incorporates the FISH results for all the cases that were marked as 2+ in the test data GT (including *Man vs Machine* dataset). From *Man vs Machine* cases (15 in total), a score of 2+ (subsequently FISH negative) was overcalled by the machines as 3+ in just one instance (VISILAB). In contrast, on three occasions (subsequently FISH negative) the participant pathologists overcalled the score 2+ as 3+. Moreover, for the remaining test dataset (13 cases), on three instances the score of 2+ (subsequently FISH positive) were erroneously predicted as either 1+ and 0 by the automated algorithms. Overall, the results indicate that further fine-tuning will be required for 2+ cases with AI. While it is encouraging that automated HER2 scoring algorithms may have sufficient potential, as direct comparison to human diagnosis, it is probably worthwhile to reflect that the number of pathologists actually joining the contest was small (only four) and it would have been better to compare the pathologists assessment of the slides on a reporting microscope rather than a computer for a fairer comparison to real-life practice.

Conventionally, expert pathologists often switch back and forth between the IHC and H&E slides to map the invasive tumour regions for estimating the percentage of membrane staining. Beside one of the participants (Team Indus), most of the algorithms reported in this chapter have avoided the use of H&E slides, though one cannot rule out the significance of H&E slide for automatic detection of DCIS regions. In addition, the task of predicting the PCMS is extremely subjective, as the expert has to make estimation on the basis of physical appearance of the stained invasive tumour region. The semi-automated methods could provide a comprehensive quantitative analysis on selected regions of interest to assist the experts in estimating the PCMS and HER2 score, especially on borderline cases. As HER2 immunoscoring relies not only on intensity but the completeness of membrane positivity, automated scoring may be helpful as demonstrated by Brgmann *et al.* [102] who proposed scoring of HER2 based on an algorithm evaluating the cell membrane connectivity.

This study shows that automated IHC scoring algorithms can provide a quantitative assessment of morphological features that may assist in objective computer-assisted diagnosis and predictive modelling of the outcome and survival. In the context of breast histopathology, whereby almost all the invasive tumour cases are considered for HER2 testing, an automated or semi-automated scoring method has the potential for deployment in routine practice. Despite all these advancements, several challenges remain for the AI algorithms to be optimised and to become part of the routine diagnosis. It is worth noting that serious optimisation will be needed for automated methods while processing a whole-slide image. Some methods required more than three hours per case, which, in the *real-world* of diagnostic service delivery is not feasible. Another limitation of this contest was that the image data were collected from a single site using a single scanner.

A potential extension would be to collect data from multiple pathology laboratories with HER2 scores marked by different experts and images scanned using a variety of different machines. This would also test the differences inherent in staining quality and appearance that may affect such procedures. Such enhancements could significantly overcome the overfitting to one particular dataset that may occur in the automated scoring methods. In moving across systems, other laboratories for example, have acknowledged the challenges in reaching the optimum Aperio algorithm parameters to provide results that were equivalent to those of the Automated Cellular Imaging System (ACIS) or Cell Analysis System (CAS 200) quantitation systems [103], which are fully automated environments for detecting cells based on intensity characteristics and handcrafted features found in IHC stained images. Therefore, there is a need to learn across comparative systems for which the current study provided a valid starting point. Also, the study highlights the need of dialogue between histopathologists and informaticians to understand correct identification of tissue compartments relevant for assessment, correct morphology (normal vs *in situ* vs invasive) and stromal stain vs tumour stain. Algorithms will also need to be trained to the natural acceptable variation in staining hues and intensities (intra and inter-laboratories) to work effectively during routine practice.

This contest provided a baseline for computer scientists and computational pathology researchers to develop automated/semi-automated algorithms for scoring HER2 WSIs. The contest has ended now but the registration and the web-portal is open for future participants to make their novel contribution in automated HER2 scoring. It is worth emphasising that in general, for IHC scoring, the GT information is only provided on the WSI level. However, computational pathology algorithms usually require detailed annotated datasets to train the underlying model. Most of the previously published automated IHC scoring methods and all the proposed algorithms in this study were using some sort of pre-processing method to select potential ROIs for training. The inevitable fact is that such hand-crafted approaches introduce a bias to the model predictions. With that in mind, in the next chapter, we present an attention mechanism that overcomes this bias and automatically ignores the irrelevant information by *learning where to see.*

Table 3.5: Combined matrix for agreement among top-3 experts and top-3 automated methods based on the agreement points and the GT for 15 cases in the *Man vs Machine* event.

Case	Ground Truth	FISH Results	Expert 1	Expert 2	Expert 3	Team Indus	VISILAB	MUCS-1
1	2+	Negative	3+	2+	2+	2+	2+	2+
2	0	-	0	1+	1+	1+	1+	0
3	3+	-	3+	3+	3+	3+	3+	3+
4	0	-	1+	1+	1+	0	1+	1+
5	1+	-	2+	1+	2+	1+	2+	1+
6	3+	-	3+	3+	3+	3+	3+	3+
7	2+	Borderline amplified	3+	3+	3+	2+	2+	2+
8	2+	Negative	3+	2+	3+	2+	3+	2+
9	3+	-	3+	3+	3+	3+	3+	3+
10	3+	-	3+	3+	3+	3+	3+	3+
11	1+	-	1+	1+	2+	0	1+	1+
12	2+	Positive	2+	2+	3+	2+	2+	2+
13	1+	-	2+	2+	2+	2+	2+	1+
14	2+	Negative	2+	2+	2+	2+	2+	1+
15	0	-	0	1+	0	0	1+	0
16	2+	Borderline amplified	-	-	-	0	1+	2+
17	2+	Negative	-	-	-	2+	2+	2+
18	2+	Positive	-	-	-	2+	1+	2+
19	2+	Borderline amplified	-	-	-	2+	2+	2+
20	1+	-	-	-	-	1+	1+	1+
21	1+	-	-	-	-	1+	1+	2+
22	0	-	-	-	-	1+	0	1+
23	1+	-	-	-	-	0	1+	1+
24	1+	-	-	-	-	0	1+	2+
25	3+	-	-	-	-	3+	3+	3+
26	0	-	-	-	-	1+	0	1+
27	0	-	-	-	-	0	0	1+
28	0	-	-	-	-	0	0	0

Chapter 4

Learning where to see: Attention Model for Automated IHC Scoring

Despite the recent progress that highlights the significance of image analysis in the domain of computational pathology [32], there are several challenges that hinder the adoption of algorithms in routine clinical practice. Computational pathology algorithms usually require detailed annotated datasets to predict the slide label. For the task at hand, the GT label for IHC score (HER2 score in our case) is generally provided at the WSI level and there are no detailed annotations provided about which ROIs from the tissue slides are consulted for the final HER2 score. Amongst existing automated approaches, the simplest approach is to manually or randomly extract patches from desired ROIs of a WSI and train a supervised model to predict the required HER2 score. Such approaches introduce an inevitable bias to the model predictions and disjoint selection of small patches may also suffer from the loss of visual context. Another potential shortcoming is computational inefficiency, as these models need to process all the regions of a given image, where some of the tissue regions may not be diagnostically relevant for the prediction of the correct IHC score.

With regards to the aforementioned challenges, we ask the question: can we train a model that ignores the irrelevant information and *learns where to see?* To answer this question, we propose a novel deep learning approach for automated scoring of IHC stained HER2 slides of invasive breast carcinoma, based on the concept of policy gradients. Given a large tile from a WSI, the proposed model identifies some of the diagnostically relevant locations from a low resolution (2.5×)

coarse representation of a given image by learning a parameterised policy over the interaction sequences of ROIs locations. The model sequentially samples the multi-resolution ROIs 40× and 20×, from the relevant locations to learn the discriminative features for different HER2 scores (0 to 3+).

The core components of the proposed model are a residual convolutional neural network (ResNet) and a recurrent neural network (RNN). The role of ResNet in this model is to learn discriminative features whereas the RNN sequentially analyses the provided features to predict the outcome and the next location. Since the GT information was only provided for WSI level with no prior knowledge of ROI locations, we train the model with policy gradients. Our model is designed to explore spatially distinct locations and learn features from visually discriminative regions. In cognitive psychology, this phenomenon is known as inhibition of return (IoR) [104] that prevents the previously attended regions to be attended again. Our model incorporates the concept of IoR in order to encourage the model to attend non-overlapping diagnostically relevant regions. Another important issue is that an erroneous scoring of 3+ as 0/1+ or vice versa may have far-reaching effects for a patient. In order to avoid such large errors, we propose a task-specific regularisation term that penalises such predictions. This study was conducted on a publicly available dataset from the HER2 scoring contest (as explained in the previous chapter) [41] containing 172 WSIs from 86 cases. Extensive experiments on the contest dataset show the efficacy of our proposed model, for guiding deep learning models to ignore irrelevant regions and scaling up to large images. The proposed method outperforms all the 18 algorithms that participated in the HER2 contest, most methods using state-of-the-art CNNs.

4.1 Related Work

Automated IHC scoring has been approached with a variety of handcrafted features and deep learning based methods. The most common approach for automated IHC scoring involves a pre-processing step to identify the potential tissue regions for training the underlying model. Then, a handful of small patches are sampled from selected tissue regions, either randomly or by using sliding window approach. The identification of potential tissue regions is generally accomplished by manual selection [105, 106], semi-automated [102] or thresholding based automated methods [107]. The pre-processing step is generally followed by training a patch-based supervised model, to learn the discriminative features and predict the outcome of each input patch. A range of hand-crafted [107, 108], approaches have been proposed to improve the IHC scoring of hormone receptors in breast cancer. For HER2 scoring, Rodner *et al.* [109] recently proposed an algorithm that computes a set of bilinear filters using convolutional layers. For classification of HER2 scores on patch level, they use bilinear features to train a multi-class logistic regression model. A deep neural network has been presented by Saha *et al.* [110] for HER2 quantification by segmenting nuclei and cell membranes. Mukundan *et al.* [111] introduced a set of characteristic curves by varying the intensity of saturation channel with a hand-picked threshold for classification of HER2 score. The final step of HER2 scoring in general involves aggregation of patch level scores to the WSI level score which is typically done by finding the most dominant class within a WSI or by training a shallow classifier on features selected from the output probability map of a WSI. Supervised patch-based approaches have established well for problems where tissue level GT is readily available. However, in IHC scoring where tissue level GT is generally not available, it is imperative to explore how deep learning models can be trained to ignore unnecessary information from the given image and focus only on regions that eventually helps in predicting the correct outcome.

Recent studies have shown that DRL has been employed in widespread applications. For object detection, Caicedo *et al.* [112] proposed a deep Q-Network (DQN) for multi-class object localisation. The model localises target objects by following a search strategy, which starts with analysing the input image and then the agent guides the model to narrow down the field of view for precise object localisation. The reward function was calculated by computing the intersection-over-union between the GT and the predicted bounding box for the object. This work was further extended for medical images including automated anatomical landmark [113] and breast lesion detection [114] for DCE-MRI images, whereby an agent localises the potential ROI containing the lesion by iteratively adjusting the bounding box. The reward function was computed by using the Dice coefficient between the GT and the predicted box. There also exist some works that incorporate DRL for genomics data to enhance the annotation of biological sequences in genome sequencing [115] and construction of protein interaction network for prostate cancer [116]. Another interesting extension is the incorporation of attention models with policy gradients that enable the model to learn a parametrised policy based on spatial dependencies. This combination has been explored for a number of applications including object detection [117], action recognition [118] and image captioning [119]. However, having precisely annotated GT for computing the reward function limits the use of deep Q-learning for IHC scoring.

In this work, we treat IHC scoring as a sequential learning task to learn discriminative features and select informative regions within a large image tile of a WSI. To the best of our knowledge, this is the first study that uses DRL for IHC scoring of histology WSIs. In terms of policy learning methodology, our work has some similarities to the methods proposed by Mnih *et al.* [120] and Ranzato *et al.* [121]. Our proposed model contains a context module that incorporates the coarse representation of input image, before predicting attentive locations. The end-to-end inhibition of return (IoR) mechanism encourages the model to explore spatially distinct attentive locations. Moreover, the scope of existing attention methods is limited to relatively small natural images whereas tumours in IHC stained WSIs are heterogeneous in terms of their morphological appearance, colour variability, shape, and temporal locations.

4.2 Learning Where to See

Given an image I, the task is to predict the HER2 score ranging from 0 to 3+ by selecting a set of diagnostically relevant regions as well as learning discriminative features from those regions. The schematic diagram of the proposed model is shown in Figure 4.1. At each time step t, the model receives two ROIs $i_t = (i_t^0, i_t^1)$, where $i_t^0, i_t^1 \in I$ are regions of width 128 and height 128 sampled at the region centre l_t at different magnification levels 40× and 20×, respectively. The convolutional network f_{c1} with learnable parameters θ_{c1} analyses i_t and transforms it into a fixed length feature vector $v_t \in \mathbb{R}^m$. The recurrent model f_h with learnable parameters θ_h sequentially processes the aggregated ROI features to update its internal state. Besides, the context model processes the down-sampled version $I^{\downarrow 16}$ (down-sampled by a factor of 16 in both directions) of the input image and perform the IoR operation, as described in Section 4.2.4. The next location $l_{(t+1)}$ is predicted by analysing the hidden state $(f_h(\theta)_h))$ from the RNN that reflects where we currently are, and the output $v^{\downarrow 16}$ of CNN f_{c2}, with learnable parameters θ_{c2}, that represents the context. The whole process is repeated for T iterations and at the end of the sequence $(i_1, i_2, ., i_T)$, the model predicts the final output score Y_T.

This iterative process wrapped around an RNN model forms a classical environment-agent interface that can be formalised by the partially observable Markov decision process (POMDP). In the current setup, the RNN and CNNs collectively act as a decision maker, which is formally known as an *agent* in the reinforcement learning (RL) literature. The agent sequentially interacts with the *environment*, which in our case is the image I. For each time step t, the agent receives a state from the environment. It then processes the given state and responds with appro-

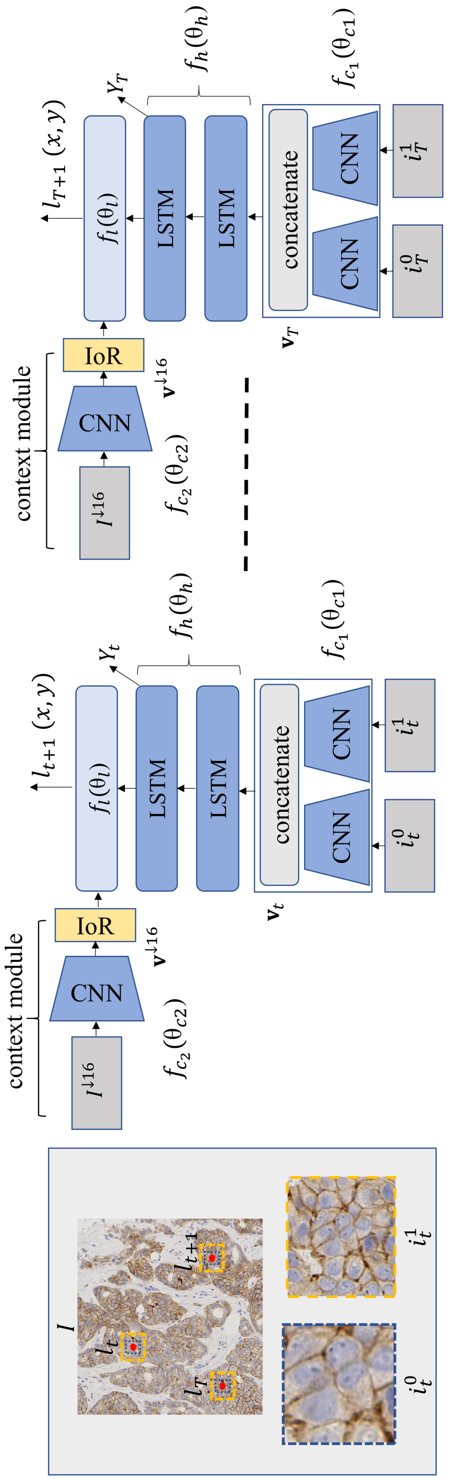


Figure 4.1: The schematic diagram of our deep recurrent model. The regions of interest sampled around l_t at 40× and 20× are shown with blue and yellow dashed bounding boxes, respectively. The dashed line in the recurrent model shows the **T** sequential iterations.

priate actions, which in our cases is the next location ($l_{(t+1)}$), directing the model where to see and eventually deciding on the HER2 score. Overall this process of predicting the next location is partly stochastic and non-differentiable, requiring the use of policy gradients. The ultimate task for an agent is to map the given states into actions by learning a parametrised policy (π) through trial and error. At the end of this sequential process, the agent receives a scalar reward R based on its actions, which in our case is linked to correct prediction of the HER2 score as described in Section 4.2.3.

At a high level, our model mimics the histopathologist practice treating a given image as an environment and the histopathologist as an agent who acts as a decision maker. Given the environment, the agent glances through different tissue components at a high level (low magnification) and then selects certain visual fields (ROIs) at low level (high magnification) to observe and store the relevant morphological features into the memory. The agent repeats this process for a certain number of iterations on different states to build an internal representation of the overall environment before coming up with the final decision, i.e. assigning the final score. For the sake of simplicity, in the remainder of this chapter, we refer to the combination of RNN and CNNs as an agent and the selected ROIs as the state.

4.2.1 Sequential Modelling

The recurrent component (RNN) acts as the backbone of the proposed model. At each time step t, the recurrent model updates the parameters of hidden states and predicts the HER2 score for a CNN based feature representation of $i_t = (i_t^0, i_t^1)$. The input to the RNN is v_t, the CNN based feature vector for multi-resolution ROIs (i_t) centred at location l_t. The RNN model also updates internal states (memory) that capture information from previous time steps. The output $f_h(\theta_h)$ of the RNN model is defined as below,

$$f_{h^*}(\theta_{h^*}) = f(h_{t-1}, f_{c1}(\theta_{c1}); \theta_{h^*}) \tag{4.1}$$

$$f_h(\theta_h) = f(h_{t-1}, f_{h^*}(\theta_{h^*}); \theta_h) \tag{4.2}$$

We choose long short-term memory (LSTM) [122] as a preferred choice for RNN to learn spatial dependencies between the ROIs. LSTM has proven to be more robust as compared to vanilla RNN and less likely to suffer from the problem of vanishing gradients. An important task of the model is to predict the next location $l_{(t+1)}$ by using $v^{\downarrow 16}$ provided by CNN and hidden representation $f_h(\theta_h)$ of

ROI images processed by CNN and LSTM. We computed the Hadamard product of $f_h(\theta_h)$ and $v^{\downarrow 16}$ to obtain a combined feature vector. Finally, the location module $f_l(\theta_l)$, linearly transforms the combined representation to predict the normalised coordinates of the next location $l_{(t+1)}(x, y)$. During the training process, the model eventually learns to encode the information from the past sequences and decide *where to see.*

4.2.2 Residual Convolutional Network

In the proposed model, the convolutional network serves as a non-linear function that maps a given RGB image into a fixed length vector representation. More specifically, we are using a variant of residual CNN [122] that contains residual connections to reroute the input information into deeper convolutional layers. Residual blocks ensure the end-to-end training of deep models by preventing the gradient from vanishing within lower layers of a CNN. It also enables the underlying model to reuse the low level features along with deeper convolutional (high level) features. For a given input p_k, the residual block function is defined as below

$$p_{k+1} = \sigma(F_k(p_k, w_k) + q(p_k)) \tag{4.3}$$

where $F_k(p_k, w_k)$ is representing a sequence of convolutional operations, w_k denote the trainable weights (biases are omitted), k represents the k^{th} residual block of the CNN, $p_{(k+1)}$ is the output of residual block and $q(p_k) = p_k$ is an identity function. The function $\sigma(.)$ denotes a non-linear activation function, which in our case is a ReLU. The $F_k(p_k, w_k) + p_k$ operation is representing element-wise addition of two activation maps. A schematic illustration of the residual CNN is shown in Figure 4.2.

We use two residual CNNs, f_{c1} and f_{c2} where f_{c1} learns discriminative features by producing a fixed length non-linear vector representation v_t for further processing by the recurrent network and f_{c2} incorporates the context information. The input to $f_{c1}(\theta_{c1})$ is the corresponding ROIs (at 40× and 20×) sampled at $l_t(x, y)$ from the input image I. The network separately processes the selected ROIs and concatenates the feature representative of i_t^0, i_t^1. The concatenation operation is previously used in [55] for a hierarchical CNN to combine the feature representation of corresponding up/down convolution layers. The CNN features of extracted ROIs mainly assist our model to learn the discriminative patterns of different HER2 scores. The task for $f_{c2}(\theta_{c2})$ is to embed the contextual awareness within the proposed model.

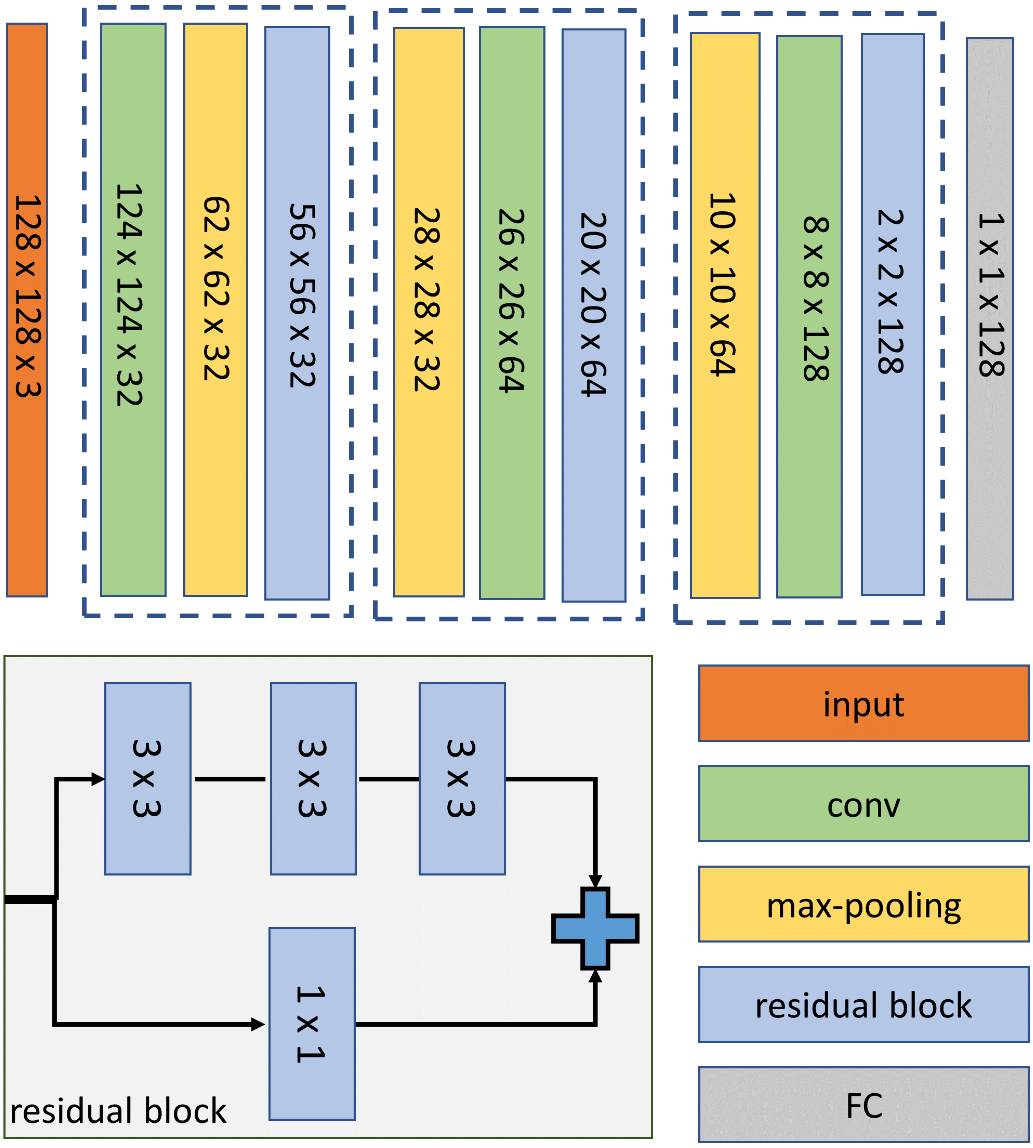


Figure 4.2: A schematic illustration of residual convolution neural network.

4.2.3 Model Training

The proposed model (Figure 4.1) is trained end-to-end by maximising the performance over the parameters $\theta = (\theta_c, \theta_h, \theta_l)$ of the residual CNNs, recurrent network and the location module. By interacting with the environment, the model forms a special case of the POMDP framework with an episodic sequence of states, actions, and rewards. The task for the model is to learn a parameterised DRL policy (π) that maps a given state into action(s) by maximising the sum of expected reward while following the parameterised policy π. The parameterised policy π for calculating the probability of a certain action a from the action space A at iteration t for a given state s with parameters θ can be defined as follows,

$$\pi(a|s) = P(A_t|S_t; \theta) \tag{4.4}$$

where S_t denotes the set of possible states at time t. Similarly, for the problem at hand, the model needs to learn a parameterised policy $\pi((l_{(t+1)}, Y_t)|(i_t, I^{\downarrow 16}); \theta)$ to predict the HER2 score (Y_t) and coordinates of the next location $l_{(t+1)}(x, y)$, given the selected ROIs (i_t) from $l_t(x, y)$ and down-sampled input image $I^{\downarrow 16}$. The model receives a scalar reward r_t after interacting with each selected ROIs ($i_{1,2,..T}$) of the given image,

$$r_t = \begin{cases} 1 & g = \underset{s}{\operatorname{argmax}} \, Y_T^{'} \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

as provided in 4.5, where g denotes the GT score, $Y_T^{'}$ represents the output of the softmax layer and s denotes the output label. The total sum of reward is computed after analysing all the selected ROIs, as defined below,

$$R = \sum_{t=1}^{T} \gamma^{t-1} r_t \tag{4.6}$$

where $\gamma^{(t-1)}$ is the weight factor for reward r_t at time t. For a finite horizon problem such as classification, we set $\gamma = 1$. Here the learning task is to optimise the parameters θ that maximise the overall performance $L(\theta)$, associated with reward r_t. A straightforward approach for handling this maximisation task is by using the gradient ascent. The update rule follows the standard back-propagation and is defined as,

$$\theta_{n+1} = \theta_n + \alpha \nabla_\theta L(\theta_n) \tag{4.7}$$

where n is the iteration index and α is the learning rate. In order to maximise $L(\theta)$, we employ the REINFORCE rule [123] from the class of policy gradients to adjust the model parameters. In this episodic scenario, on average, the model computes the gradients for actions that lead to higher rewards and consequently, the log probability of actions with low rewards will be decreased. The policy gradient, as described above, can be mathematically expressed as follows,

$$\nabla L_\theta = \sum_{t=1}^{T} \nabla_\theta \log \pi((l_{t+1}, Y_t)|\,(i_t, I^{\downarrow 16}); \theta) R_t \tag{4.8}$$

or

$$\nabla L_\theta = \sum_{n=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi((l^n_{t+1}, Y^n_t)|\,(i^n_t, I^{\downarrow 16}); \theta) R^n_t \tag{4.9}$$

One limitation of the above formulation is that the model convergence can be challenging if intra-class variance in the training dataset is relatively high. To generalise the policy gradient algorithm, we include a baseline function $b_t = E_\pi[R_t]$ for comparing the action values to the cumulative reward [124],

$$\nabla L_\theta = \sum_{n=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi((l^n_{t+1}, Y^n_t)|\,(i^n_t, I^{\downarrow 16}); \theta)(R^n_t - b_t) \tag{4.10}$$

4.2.4 Inhibition of Return

An important factor in adequate selection of diagnostically relevant regions is to inhibit the model from visiting the previously attended regions. We observe that for some of the selected locations during the sequential process, the sampled ROIs are not spatially distinct. Figure 4.3 (1st column) shows three such examples where the selected ROIs lie relatively close to each other, resulting in overlap with previously attended regions without any significant performance gain. This argument also applies to images where the diaminobenzidine (DAB) stain expression is relatively sparse as shown in Figure 4.3 (2nd row). A straightforward strategy to address this issue would be to suppress the texture information [104] of previously attended locations that would encourage the model to rapidly explore spatially distinct locations. This simple IoR strategy leads to the model giving higher priority to regions that it has not previously considered for learning the discriminative features. The IoR strategy is computationally efficient, widely studied in cognitive psychology [125] and in sequential learning [104, 112].

Further, we introduced an additional constraint L_{IoR} in the loss function. At

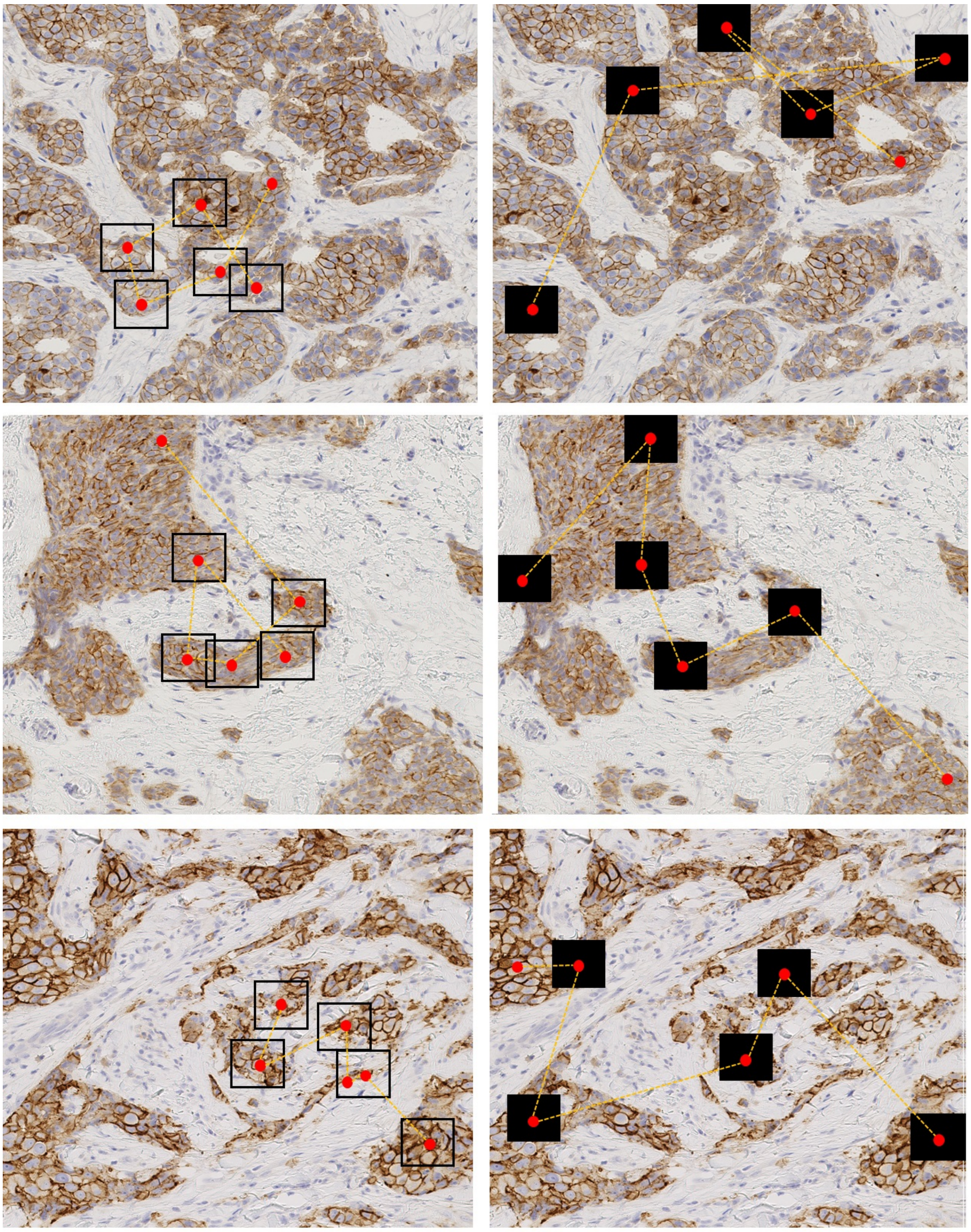


Figure 4.3: Three sample images representing the effect of inhibition of return (IoR), showing the selected ROIs without the IoR penalisation (Left) and with the IoR penalisation (Right). As can be seen, the selected ROIs after the IoR penalisation are relatively distinct from each other. Filled rectangular regions (black) show the suppressed texture.

the end of the iteration sequence $t = (1, , T)$, L_{IoR} computes the overlap between the coordinates of selected ROIs, penalising selection of image patches relatively close to each other. The scope of IoR penalisation vanishes (its value becomes 0) if selected locations are spatially distinct from each other. The L_{IoR} term is defined as below,

$$L_{IoR} = \frac{1}{C_2^T} \sum_{t=1}^{T} \sum_{j=t+1}^{T} rect[l_t(x, y)] \cap rect[l_j(x, y)] \tag{4.11}$$

where $rect[l_t]$ and $rect[l_j]$ represent rectangular coordinates sampled from the input image I and C_2^T is the number of combinations of different ROIs, in turn helping in limiting the intersection values between 0 and 1. The above loss function updates the policy parameters to correctly predict the HER2 score by penalising the model for locating spatially overlapping regions.

4.2.5 Task-Specific Regularisation

The clinical impact of large erroneous predictions of HER2 score is highly significant and should be avoided as much as possible. Inaccurate prediction for patients of score 0/1+ as 3+ will lead to giving treatment with toxic anti-HER2 drugs to patients who do not need it, while predicting cases with score 3+ as 0/1+ will lead to the patient not given the appropriate treatment needed. To avoid such scenarios, we added a task-specific regularisation term.

$$L_{sc} = |\operatorname*{argmax}_{s} (Y_T^{'}) - g| \tag{4.12}$$

The final loss function combines the parameterised loss with task-specific regularisation and IoR as given below,

$$L = L_\theta + \lambda_1 L_{sc} + \lambda_2 L_{IoR} \tag{4.13}$$

where $\lambda = \lambda_1 + \lambda_2$ controls the sensitivity (scope of penalty) for both L_{sc} and L_{IoR}. For our experiments, we opted to keep $\lambda_1 = \lambda_2$.

4.3 Experiments and Results

4.3.1 Dataset

This study is conducted on a publicly available dataset from the HER2 scoring contest, as explained in the previous chapter. The contest dataset consists of WSIs from 172 histology slides of 86 invasive breast carcinomas cases scanned using Hamamatsu NanoZoomer C9600 at the highest resolution (40×), two slides per case (one IHC stained with HER2 and another with the standard H&E). On average, each scanned WSI contains more than 10^{10} pixels. The GT for the contest was marked by a minimum of two expert histopathologists. For each case in the training dataset, the GT consists of a HER2 score and a PCMS, both at the WSI level. The training dataset is made of 52 cases, 13 cases from each HER2 score (0-3+) and the test set consists of 28 cases. The remaining 6 cases were not included in the test/training dataset and only reserved for the on-site part of the competition.

4.3.2 Experimental Setup

The ROIs were cropped at 40× and 20× resolutions, each ROI of size 128×128×3 pixels. The size of the input image I was 2048×2048×3 (471.1×471.1 μm^2) sampled at 40×. In total, we extracted 58,500 image tiles (with an overlap of 50%) from the 52 training WSIs after tissue segmentation and a simple DAB intensity based thresholding. The number of neurons in hidden layers of RNN was set to 256 and 128, respectively. The CNN transforms the given image into the feature representation of size 1×128. ReLU activation function was used after each residual block. To overcome the overfitting problem, we performed data augmentation by random rotations (0°,90°,180°,270°), horizontal and vertical flipping, and the transpose of all the images in the training dataset. The regularisation parameter λ controls the sensitivity of the task-specific regularisation and IoR penalisation. Through empirical observations based on the validation data, we found that the best performance was achieved with a value of 0.04 for λ , which we used for all the experiments. Some other values of λ we used to evaluate the performance of the model are $0.01, 0.02$, and 0.06. The initial learning rate was set to 0.001 with exponential reduction of 0.97 and the momentum was adjusted at 0.9. The batch size was selected as 10. The location of the first ROI was randomly selected. The number of ROIs per image was selected as 6. The learning parameters were initialised as Gaussian random numbers with 0 mean and 10^{-2} standard deviation and biases were set to 0.

4.3.3 Comparative Analysis

In this section, we discuss a variety of experiments to demonstrate the efficacy and evaluate the performance of the proposed method. For the following experiments, we performed 4-fold cross validation across 52 cases. We split the 52 cases into 4 subsets, with nearly equal representation of all four HER2 scores, and used 3 subsets

Table 4.1: Acc$_{comb}$ denotes combined accuracy. In DAB ROIs, the locations were randomly selected from the diaminobenzidine (DAB) regions.

Method	**0**	**1+**	**2+**	**3+**	**Acc$_{comb}$**
RMVA	0.702	0.446	0.275	0.275	0.355
random ROIs, L_θ, L_{sc}	0.868	0.671	0.632	0.803	0.743
random DAB ROIs, L_θ, L_{sc}	0.822	0.615	0.677	**0.874**	0.764
Proposed - without context module	**0.982**	0.452	0.568	0.721	0.652
Proposed - L_θ	0.982	0.538	0.703	0.782	0.733
Proposed - L_θ, L_{sc}	0.963	0.532	**0.721**	0.825	0.753
Proposed - L_θ, L_{sc}, L_{IoR}	0.919	**0.772**	0.661	0.837	**0.794**

for training and the remaining one subset for validation. The GT for the test dataset is not publicly available and, in this section, we have reported the performance of different variants of the proposed model on the validation dataset. We report the results for the test dataset in Section 4.3.4. Generally, a large part of WSI contains background (glass) regions with no tissue components. For tissue segmentation, we perform local entropy filtering on a lower resolution (2.5×) version of the WSI.

Comparative Results

In this experiment, we evaluate the significance of different sub-components of the proposed model, including L_θ, L_{sc}, and L_{IoR}. Another important aspect is to evaluate the significance of the parameterised policy for selecting relevant ROIs and how it affects the performance if we select ROIs randomly instead of following a certain policy.

For random selection of ROIs, we perform two main experiments: *a)* select ROIs randomly from the entire I and *b)* select ROIs randomly from only DAB regions of I. In the first approach of random selection, we predict the HER2 score of given ROIs (i_t) and select the next location randomly without consulting the context and current state of the model. Random selection correctly predicts HER2 score for images where most of the area is covered by discriminative tissue regions. However, it is susceptible to selecting regions that contain mostly background and sparse DAB regions. For the second approach, we perform stain deconvolution [126] on I by estimating the stain matrix using [127]. DAB regions contain low luminance and therefore for binarising the DAB channel using $\tau(I_{DAB})$, we empirically chose a relatively high threshold value of 0.8. The $\tau(I_{DAB})$ is then followed by morphological operations to exclude the noisy (small) components of the DAB channel. Table 4.1 shows the results for both experiments. Evidently, the second method is a relatively

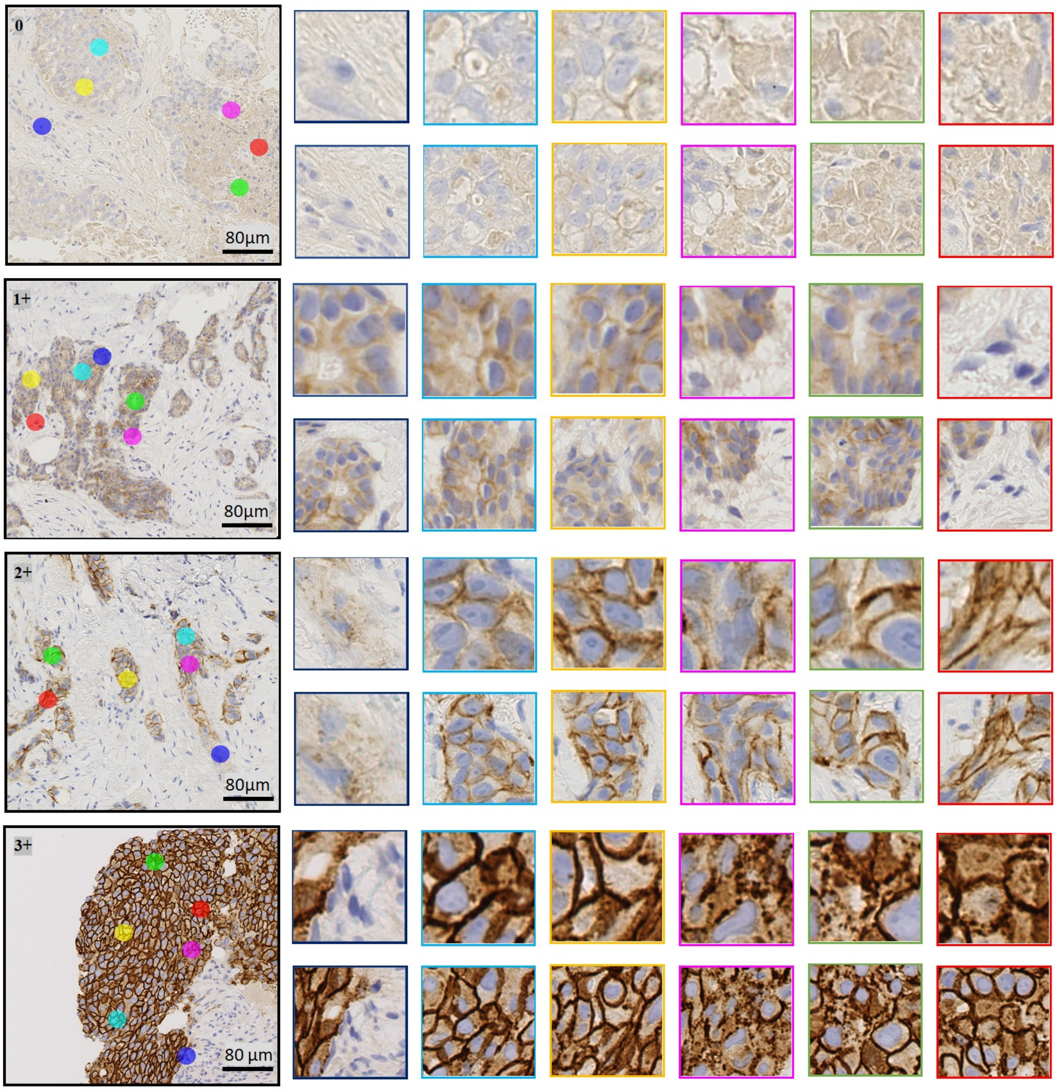


Figure 4.4: Example of four images with selected regions-of-interest (ROIs) predicted by our method, for each HER2 score (0-3+), respectively. The first column shows the input images and coloured disks shows the predicted locations. The remaining columns show the selected regions at 40× and 20× around the selected locations $l_t, t = 1, 2, , 6$. The first selected region is shown with blue bounding box and the last selected region is shown with red bounding boxes.

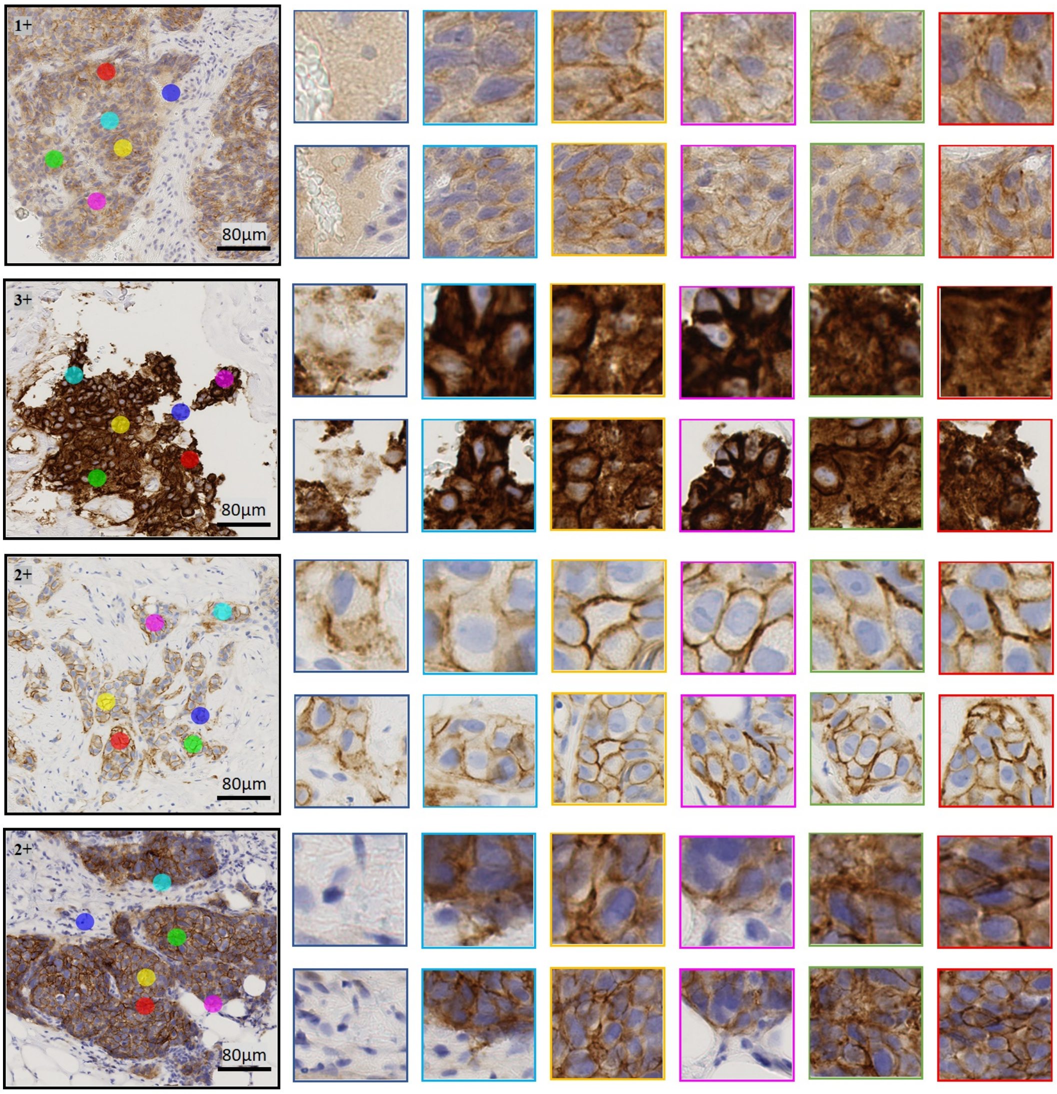


Figure 4.5: Another example of four images with selected regions-of-interest (ROIs) predicted by the algorithm. Correctly predicted HER2 scores are 1+, 3+, 2+, and 2+. The first column shows input images and coloured circles shows the predicted locations. The first location is shown with blue bounding boxes and similarly the last location is shown with red bounding boxes.

direct way of selecting i_t and therefore yields higher performance as compared to the first approach for random selection. However, it offers some limitations as compared to the proposed model: the first major limitation is the absence of L_{IoR}, that enables the model to explore spatially distinct locations. The L_{IoR} restrains the proposed model from overemphasising particular locations by penalising the learnable parameters and encourages the model to learn discriminative features from different tissue regions.

One way of handling the absence of L_{IoR} is to introduce a set of hard-constraints for selecting spatially distinct $l_t(x, y)$. However, defining a set of generalised hard-constrains is a non-trivial task and it may influence the model performance on images where we have sparse DAB representation (3rd row of Figure 4.4). The comparative analysis for combining randomly sampled ROIs with IoR could be achieved by rejecting ROI samples where there is an overlap with the previous ROIs. For the sake of simplicity, we have not performed such comparison with the proposed method and it may be considered as a limitation of this comparative analysis. Secondly, it has no longer access to the overall $I^{\downarrow 16}$ and the context module. Consequently, this variant of our proposed model could be considered a departure from the routine clinical practice, where a pathologist glances through a coarse representation of input image at a higher level and then selects ROIs at lower levels before concluding the outcome. Therefore, it is imperative to follow a parameterised policy that incorporates the context and offers a temporal connection (via LSTM) between the selected locations.

Further, we investigate the implications of the context module, task-specific regularisation (L_{sc}) and IoR (L_{IoR}). The context module allows the model to analyse a coarse representation of the overall environment and use that to predict $l_{(t+1)}(x, y)$. Existing attention based models [120, 128] including RMVA (recurrent model for visual attention) have no mechanism to prevent models from revisiting the previously attended regions. Besides, RMVA also incorporates a strong location prior to the model, which is irrelevant in histology image analysis mainly due to the random orientation and morphological appearance of underlying tissue. Table 4.1 gives patch-based classification results of the proposed model with different settings. Overall the results are in favour of the proposed method (L_θ, L_{sc}, L_{IoR}). Figure 4.4 shows a sample of representative patches with selected ROIs. The most challenging images were from HER2 class 1+ and 2+, where resemblance in morphological appearance was present due to tumour heterogeneity, as shown in 2nd and 3rd rows of Figure 4.4. In Figure 4.4, the image patch for score 2+ contains relatively smaller extent of DAB expression and most of the tissue region belongs to 0/1+. In that case, the model starts with a relatively less informative region and sequentially learn to focus on the informative regions of the image to predict the correct outcome.

Number of ROIs

The aim of this experiment is to investigate the effect of number of ROIs for a given image I. We evaluate the performance of the proposed model by using 4,5,6 and 8 number of ROIs. For this contest dataset, we observed the best performance with 6 ROIs, as shown in Figure 4.6. One of the main reasons for relatively low performance with 8 ROIs is the images containing tissue boundary regions, where most of the image region is covered by background glass. Therefore, in those cases, selecting more locations may confound the model in predicting the correct outcome. Another interesting aspect is the reduction of inference time and gain in the overall performance (as discussed in Section 4.3.4), in a conventional patch-based setting, for a given image I of size 2048×2048×3 at 40×. The model needs to process all 64 ROIs (each of size 256×256×3). In contrast, the proposed model can predict the HER2 score after consulting a handful of ROI patches accompanied with down-sampled version of I. However, although computing time may not be the most decisive aspect in clinical practice, it may be an important factor in high-volume diagnostic settings and for high-throughput IHC screening.

Selection of multiple resolution regions

This experiment compares the performance of the proposed model for selecting a suitable combination of magnification levels. We perform this experiment with three sets of magnification levels including 40×, 20×, and 10×. We observed that for IHC HER2 scoring, ROIs selected from 40× and 20× yield the best performance. The results for mean scoring accuracy for all 4 classes are shown in Table 4.2. ROIs selected at higher resolutions offer more detailed information regarding HER2 expression at cell levels. On the other hand, ROIs selected at 10× or lower magnification offer more context information, but they are also more likely to contain background or irrelevant tissue regions, including non-invasive haeamatoxylin stained regions.

Size of ROIs

The main objective of this experiment was to evaluate the performance of the proposed model on different sizes of ROIs. After selecting the location $l_t(x, y)$, it is worthwhile to quantify the extent of context required for predicting the correct outcome. Recent studies in computational pathology have also emphasised the sig-

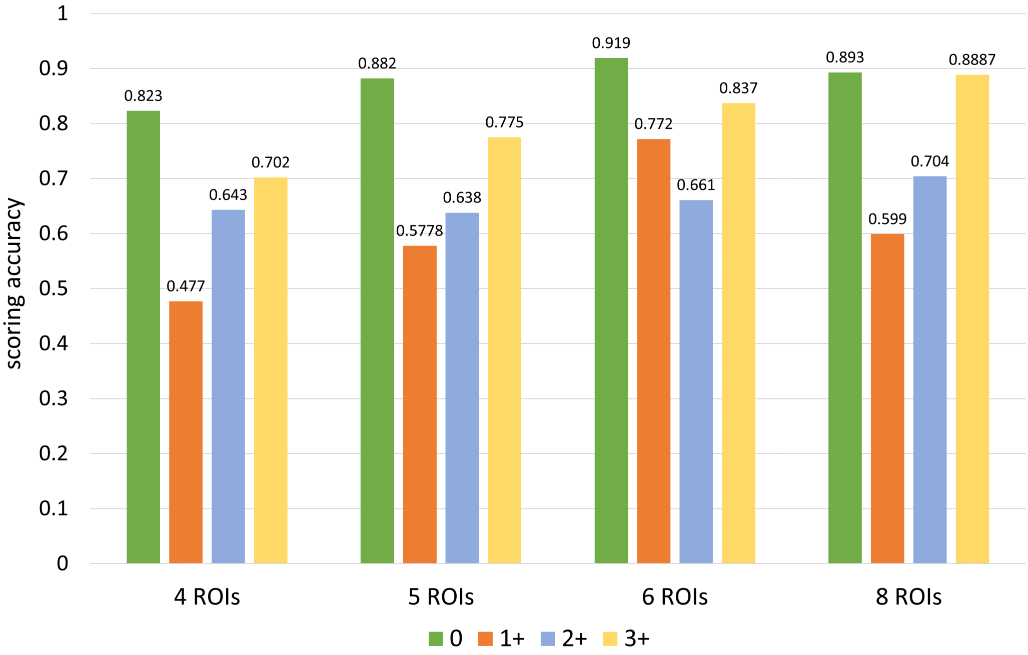


Figure 4.6: Comparative results for different numbers of ROIs.

Table 4.2: Significance of context in the proposed method. Acc$_{comb}$ denotes combined accuracy.

Method	**0**	**1+**	**2+**	**3+**	**Acc$_{comb}$**
Proposed 40×, 20×	**0.919**	**0.772**	0.661	**0.837**	**0.794**
Proposed 40×, 10×	0.881	0.613	**0.676**	0.807	0.742
Proposed 20×, 10×	0.802	0.592	0.608	0.711	0.65

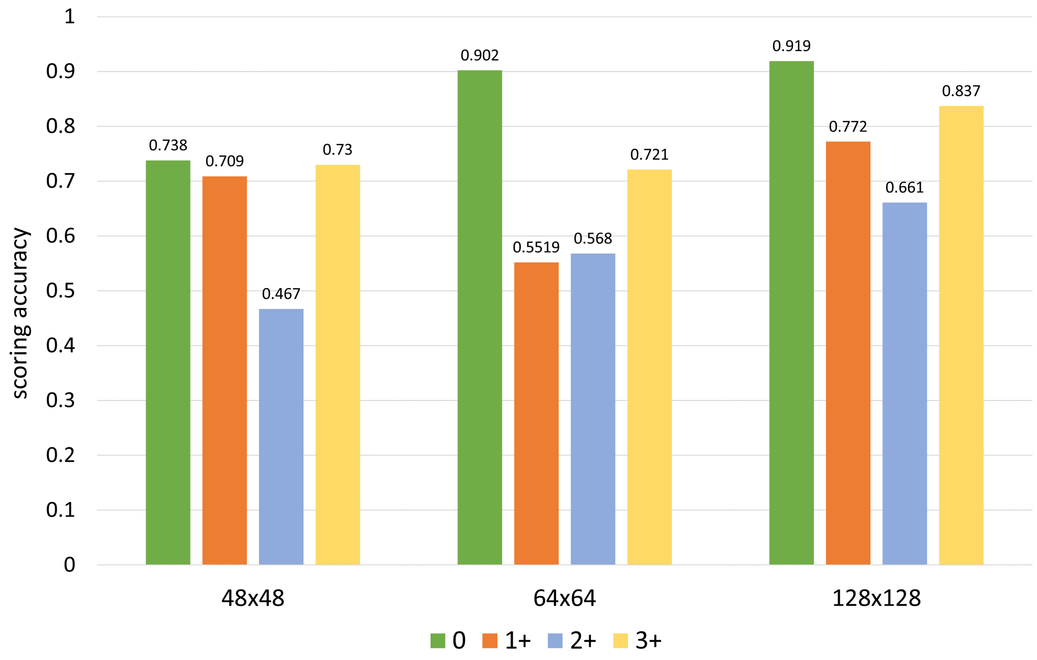


Figure 4.7: Comparative results for different patch sizes of ROIs.

nificance of visual context [129]. We perform this experiment on three different patch sizes, including 48×48, 64×64, and 128×128, with results in Figure 4.7. We noticed the best performance with ROIs of size 128×128. As expected, the performance of the proposed model increases with increase in the context. However, in HER2 scoring, it is important to limit the ROI size to prevent the inclusion of irrelevant tissue regions.

4.3.4 Contest Leaderboards

This subsection covers the description for scaling the patch level results to WSIs and performance of the proposed algorithm on the contest tasks.

Contest Tasks

A detailed description regarding the evaluation criteria is explained in the previous chapter. The performance of the proposed algorithm on the WSI level is evaluated by using 3 different criteria, as suggested in the contest guidelines: *a)* agreement points, *b)* weighted confidence and *c)* combined points. In agreement points, a penalty method was introduced that assigned points between 0 and 15 to each case based on the clinical significance of the difference in predicted and actual scores.

To resolve tie situation in the first criterion, bonus points were also awarded based on a correct prediction of PCMS. A weighted confidence was devised to estimate the credibility of WSI results predicted by the algorithm. This measure may also help in stratifying cases that need further input from pathologists. And finally, for each case, a combined point was calculated by taking the product of the other two assessment criteria.

PCMS Estimation

In routine clinical practice, a pathologist visually estimates the PCMS on the WSI level, indicating the strength of invasive carcinoma cells stained to HER2 protein. For our experiments, we split the WSI into manageable image tiles I, depending on the computational resources. The model then predicts a HER2 score for each image tile and aggregates the results on the WSI level by simply choosing the most dominant class as the HER2 score $s = \text{argmax}(s_0; s_{1+}; s_{2+}; s_{3+})$ where s_0; s_{1+}; s_{2+}; s_{3+} represent the number of image tiles predicted as 0, 1+, 2+, and 3+, respectively. The PCMS for invasive breast cases was estimated by the WSI output maps. For each score map, we simply compute the ratio between the area covered by each predicted score over the total area of tissue region within the WSI.

Combined Leaderboards

One of the most challenging aspects in analysing histology images is to limit the automated analysis to diagnostically relevant tissue parts. A fully supervised method processes the given image regardless of noisy or irrelevant contents. In contrast, recurrent models can appropriately tackle such scenarios by only processing the relevant ROIs. Figure 4.5 shows some of the visual results on the image patches from the validation dataset. Table 4.3 reports the WSI level performance of the proposed method on all 3 evaluation criteria. It also contains the results of top-10 performing algorithms in the contest. The proposed method ranked as 1st amongst all 18 submissions in the contest, including the combined points criterion and point based scoring. The contesting algorithms were based on a wide range of the state-of-the-art deep convolutional networks including GoogleNet [99], AlexNet [52] and LeNet [130]. On 28 WSIs from the test dataset, the proposed algorithm correctly classified 26 WSIs. The most difficult cases were from borderline class 2+ and 3+. Both the false predictions belonged to those classes. It is worth noting that the scores of 2+ and 3+ are the most difficult to call for expert histopathologists as well. On the other hand, the proposed model correctly predicted the scores of 12 out of 14 WSIs

Table 4.3: Comparative results with other participant teams on the test dataset of HER2 scoring contest.

Teams	**Points**	**Points+ Bonus**	**Weighted confidence**	**Weighted Points**
The proposed method	**405**	**419**	**24.1**	**359.1**
VISILAB-I (GoogLeNet [99])	382.5	404.5	**23.55**	**348**
FSUJena [109]	370	392	23	345
HUANGCH (AdaBoost)	377.5	391.5	22.62	345.7
MTB NLP (AlexNet [52])	390	405.5	22.94	335.7
VISILAB-II (contour analysis)	377.5	391	21.88	322
Team Indus (LeNet [130])	**402.5**	**425**	18.45	321.4
UC-CCSE [111]	390	395	21.07	316
MUCS-III [106]	390	411	20.43	300.8
HERcules (SVM)	360	380	20.57	295.6
MUCS-II (GoogLeNet [99])	385	413	19.51	290.1

with scores 2+ or 3+.

4.3.5 Glyoxalase-1 protein (Glo1) Scoring

In this experiment, we evaluate the performance of the proposed method on IHC stained gastroenteropancreatic neuroendocrine tumours (GEPs). The over-amplification in glyoxalase-1 protein (Glo1) is associated with resistance in multidrug tumour chemotherapy [131]. In routine practice, an expert pathologist visually examines IHC stained GEP slides and reports a score between 1+ and 3+, which represents weak, moderate and intense Glo1 immunostaining, respectively. This experiment is conducted on 82 WSIs from 39 patients with 25 midguts and 14 pancreatic cases, with a total number of 22, 33 and 27 WSIs scored by an expert pathologist as 1+, 2+ and 3+ respectively. Further details regarding the dataset and GEP tumours can be found in [132].

The main objective of this experiment is to test the efficacy of the proposed IHC scoring method on another IHC stain by quantifying the concordance between the pathologist GT score and Glo1 score predicted by the proposed method. We first split the dataset into two folds and perform cross-validation by selecting half of the dataset for training and the remaining half for testing. We then repeat the experiment by swapping the training and test datasets. For each fold, the training dataset was further spilt into training and validation subsets by selecting 35 WSIs for training and remaining 6 (15% of the training data) for validation. Tissue regions from WSIs were segmented by performing local entropy filtering on a lower resolution

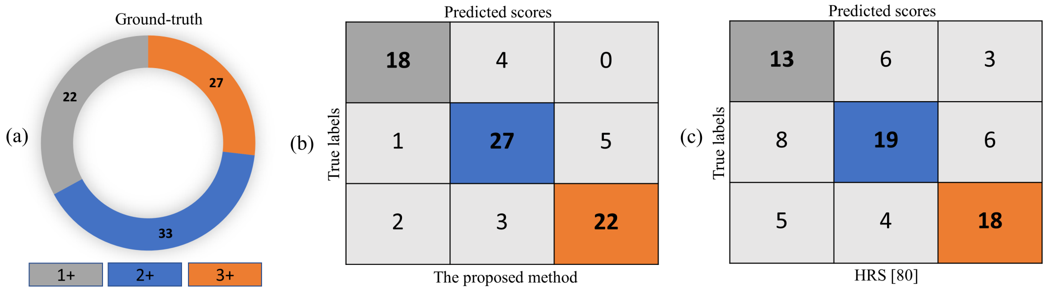


Figure 4.8: (a) Description of the number of Glo1 scoring WSIs (b) confusion matrix for the proposed method and (c) confusion matrix for the hormone receptors scoring (HRS) method [108].

(2.5×) version of the WSI. To avoid the class imbalance problem, we extracted all the image tiles (containing tissue) from 1+ WSIs and randomly sampled equal number of image tiles from classes 2+ and 3+. In total, we extracted 86,700 tiles each of size 2048×2048×3 at 40×. We then retrained the proposed model by retaining the same data augmentation methods and hyperparameters (including learning rate, lambda, number of ROIs, etc.), as explained in Section 4.3.2. Optimal weights were selected based on the validation set. The final Glo1 score on the WSI level was estimated by aggregating the scores of image tiles, as explained in Section 4.3.4.

Figure 4.8 shows two confusion matrices (CF) containing WSI results of Glo1 scoring. The CF in Figure 4.8(b) shows the results for the proposed method and the results for automated hormone receptors scoring (HRS) method [108], which was initially used in [132], are shown in Figure 4.8(c). The overall agreement between the GT and the proposed method is 81.7% (Glo1 scores for 67 WSIs were correctly predicted), whereas the agreement is 60.9% in case of HRS. For IHC scoring, HRS relies heavily on pixel intensities for quantifying the chromatin and protein content. Therefore, heterogeneous morphological characteristics of neuroendocrine cells within tumour regions pose the risk of confusing the algorithm in extracting the desired features for IHC scoring. Figure 4.9 shows some of the qualitative results on image tiles with different Glo1 scores. Overall, it is encouraging that the proposed method outperforms the HRS with a noticeable margin. It is worth mentioning that the performance of the proposed method may improve by appropriately tuning the hyperparameters. Our intention here is to demonstrate to some extent the generalisability of the hyperparameters selected from HER2 scoring. The IHC scoring of HER2 and Glo1 cases have some fundamental similarities: *a)* in both cases the GT was provided on the WSI level and therefore, it is imperative that the underlying model learns a stochastic policy and identifies some of the diagnostically relevant regions in predicting the final outcome, and *b)* similar to HER2 scoring, an erroneous scoring of 1+ as 3+ or vice versa may have far reaching effects for a patient. It is somewhat necessary to have a mechanism that penalises the learnable parameters for overcalling those classes. We also observed that in some tissue regions, cells are densely packed and pose difficulties for the model in selecting ROIs from the invasive tumour regions.

4.4 Discussion

With the adoption of digital slide scanners in routine pathology labs, large-scale WSIs or virtual slides (10^{10} pixels) have emerged as a reliable alternative to conventional glass slides [86]. Typically, for effective training of deep learning models, it is incumbent to train computational models with large-scale precisely annotated regions. The ineludible fact is that sourcing precise high-resolution annotations of scanned slides is a laborious task and not considered as a part of the routine clinical practice. Hence, attaining tissue-level annotations for a significantly large dataset is one of the factors that may delay the acceptability of automated methods in clinical practice [29].

This study proposes a novel recurrent model for IHC scoring of HER2 slides of breast cancer. In some respects, the model mimics the pathological behaviour by learning a parameterised policy to select diagnostically relevant regions. Experimental results conducted on a challenging contest dataset demonstrate the efficacy of the proposed model by outperforming the state-of-the-art methods, as shown in Table 4.3. It is worth noting that the proposed model only identifies a small number of regions required for predicting the correct outcome. Depending on the requirements, the model can be extended to assign different scalar rewards to each selected ROI. Different rewards may also be interpreted as the significance of each selected ROI. Nevertheless, a challenge remains in that sometimes the model selects regions that may not appear to be diagnostically relevant (e.g., non-invasive tumour regions) due to tumour heterogeneity. An example is shown in Figure 4.4, the last location of score 1+.

In computational pathology, diagnostic efficiency may be based on two main factors: *1)* selecting potential ROIs from a given WSI, and *2)* extracting discriminative features from within the selected ROIs. Some studies on eye movement have shown that with the passage of time trainee pathologists eventually gain experience in the selection of diagnostically relevant ROIs and learn discriminative features [133]. Another recent study [134] shows that the overlap ratio between

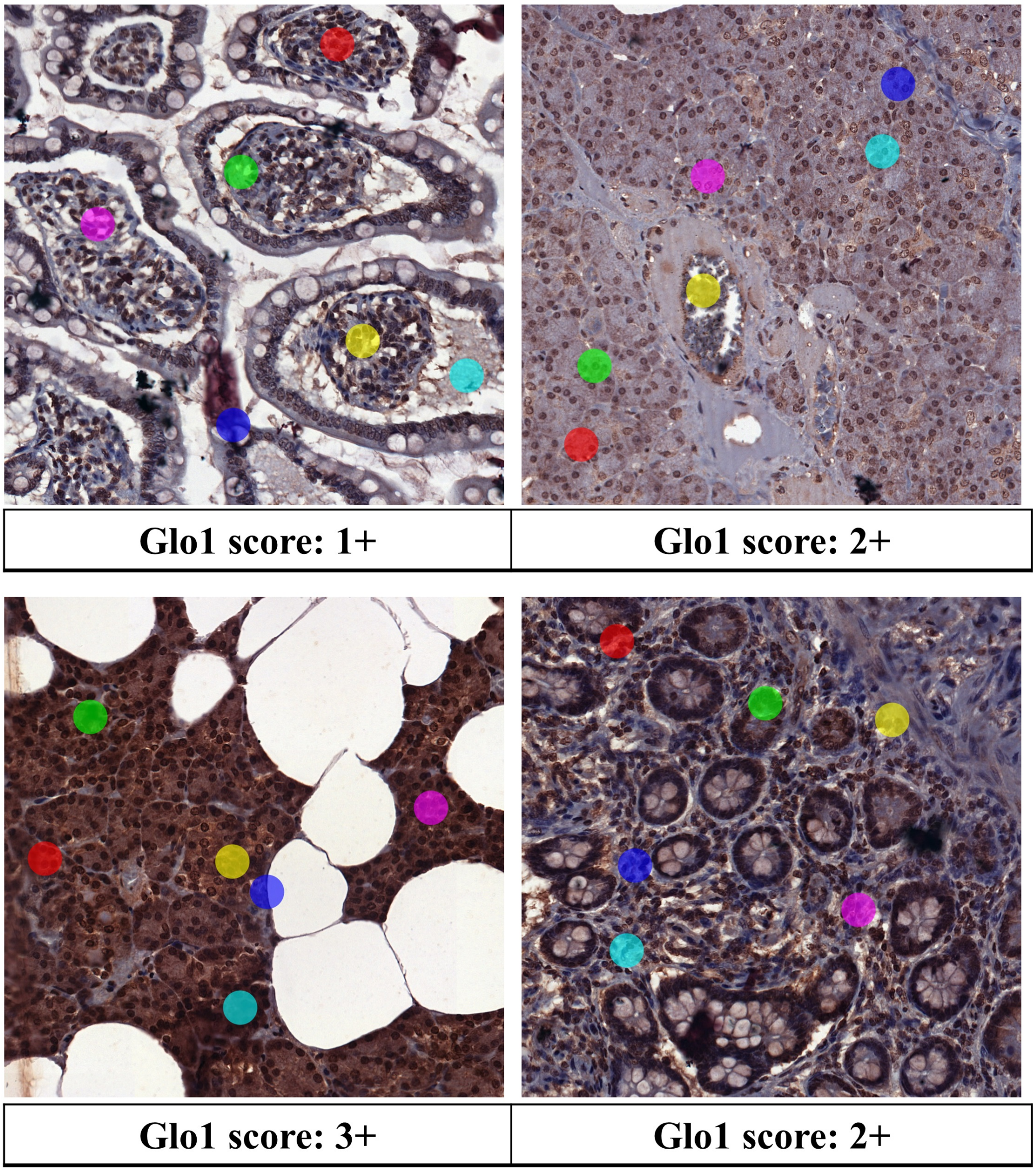


Figure 4.9: Sample of four IHC stained Glo1 images with selected regions-of-interest (ROIs) predicted by our method. Coloured disks show the predicted locations and the sequence is as follows: blue, cyan, yellow, magenta, green, and red.

ROIs selected by different pathologists is associated with higher diagnostic accuracy. An interesting extension of our work would be to find correlation between the diagnostically relevant ROIs selected by machines and experts. This may help in evaluating the diagnostic reliability of computer-assisted diagnostic systems. The proposed model also offers several advantages over the conventional fully supervised approaches. First, the model is capable of handling unwanted background regions that are often quite common in histology images due to several non-standardisation factors in slide preparation. Second, the proposed model is capable of scaling up to WSIs or significantly large tiles of a WSI as the number of trainable parameters are directly linked with the size of ROIs instead of image tiles. However, in order to perform end-to-end training of the proposed model, we need the uncompressed representation of all required tiles of a WSI to be hosted into the memory and it is worth mentioning that on average, a 24-bit representation of an uncompressed WSI at 40× requires approximately 56 GB [35] of memory, making it intractable to load an entire WSI into the GPU memory.

The scope of the proposed model is not only limited to HER2 antibody quantification. The scoring of other prognostic and predictive markers like ER, PR, and proliferation of Ki-67 on the WSI level may also be possible with a similar approach. Last year i~~I~~n the United States ~~alone previous year~~ (2018), it was expected that there will be 266,120 new invasive BC cases[1]. Whereas visual examination of histological specimens is generally influenced by subjectivity measures for learning discriminative features. The proposed model may assist experts in reducing subjectivity and serve as a semi-automated tool to find diagnostically relevant regions that may require the pathologist attention. Until now, we have described our proposed automated tools for cancerous regions of H&E and IHC stained histological images. During tumour progression, tumour stromal regions also experience structural variations and in the next chapter, we will describe an automated method that performs proximity analysis of tumour and stromal (collagen) regions and their significance for predicting overall survival (OS) in diffuse large B cell lymphoma (DLBCL) patients.

[1] https://www.breastcancer.org/symptoms/understand_bc/statistics

Chapter 5

Tumour-Collagen Proximity Analytics

Tumour-associated macrophages, tumour infiltrating lymphocytes and immune checkpoint expression have all been associated with tumour behaviour [25, 26, 135]. In contrast, there are a limited number of studies that have investigated the role of the acellular stromal microenvironment on outcome in diffuse large B cell lymphoma (DLBCL) [136]. Stromal gene expression signatures and extracellular matrix components such as fibronectin have previously been associated with outcome in DLBCL. Collagen is an important component of the acellular stromal microenvironment, and the presence of particular collagen subtypes has been linked to outcome in other cancers, including both urothelial and colorectal carcinoma [27, 137]. These findings warrant further investigation of the role of collagen in DLBCL.

In this work, we investigate the spatial proximity of tumour cells and collagen VI in DLBCL and describe a novel digital proximity signature (DPS), which serves as a marker of regions of the tumour likely to be enriched for active collagen signalling. The core components of the proposed framework involve: *a)* cell detection and classification, *b)* selection of collagen points and *c)* estimation of DPS. Figure 5.1 shows an example of a ROI selected from a DLBCL WSI and Figure 5.2 represents some exemplar tumour and normal cells. In this regard, we propose a novel deep learning framework, Hydra-Net, for simultaneous cell detection and classification, enabling the end-to-end learning on a multi-task problem. During testing, we introduce a multi-stage ensembling predictor that combines cell detection and classification predictions by leveraging information about the local neighbourhood of cells. The comparative analysis of cell classification demonstrates the efficacy of Hydra-Net over single-task learning models. The tumour-collagen proximity analy-

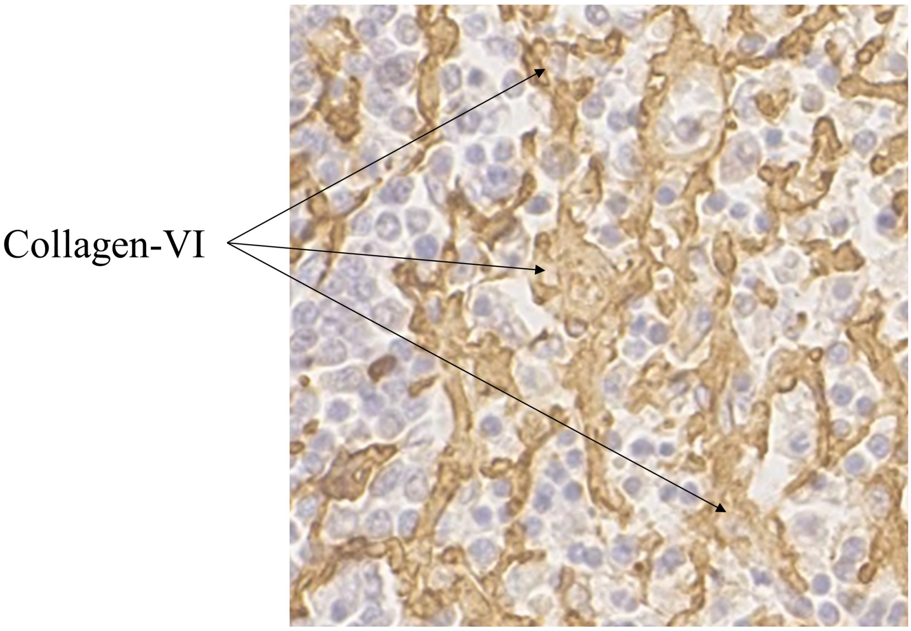
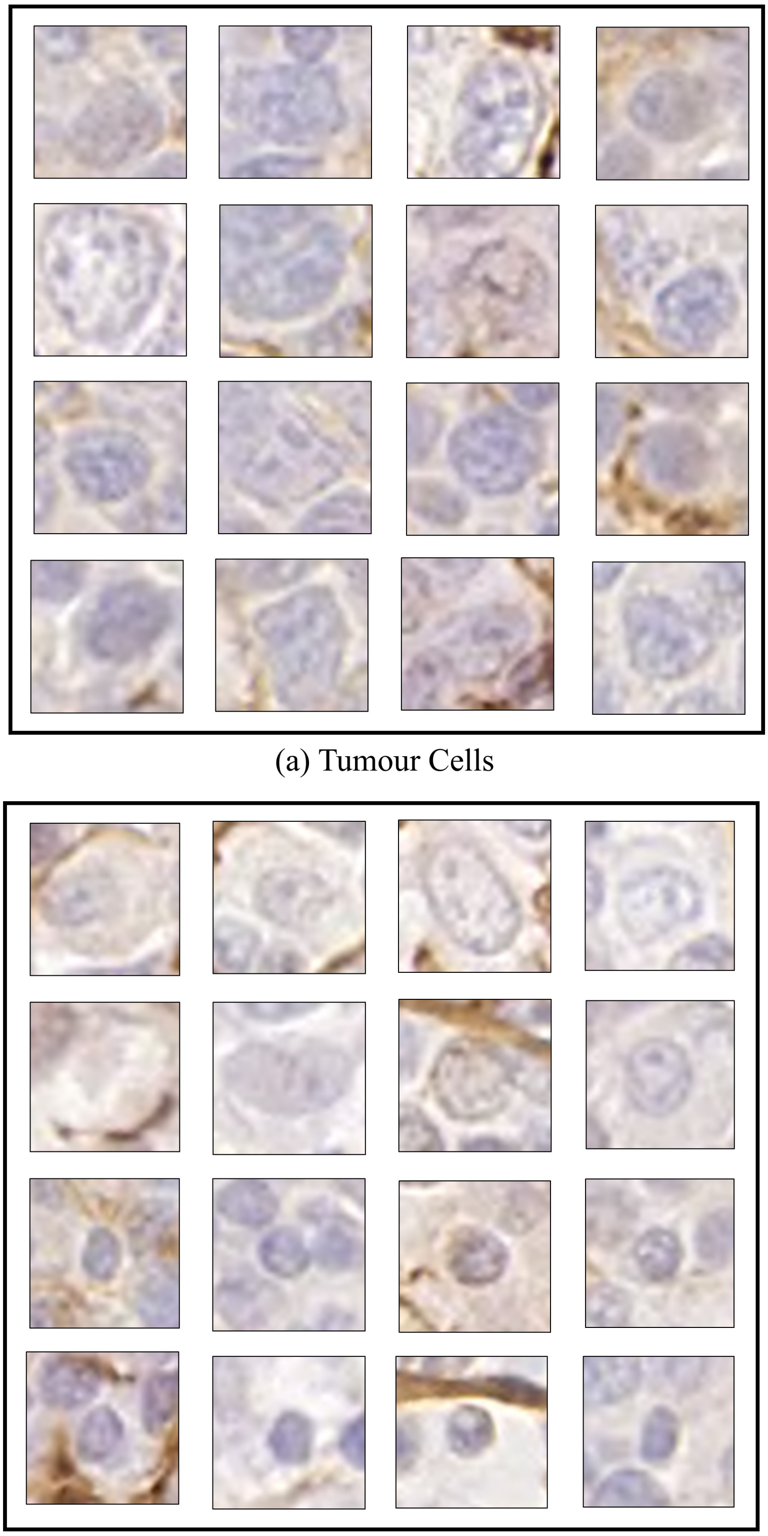


Figure 5.1: Example of a region of interest selected from diffuse large B-cell lymphoma whole-slide image.

sis is then performed by aggregating the tumour cell statistics within the vicinity of collagen VI. We further show that strongly associated tumour-collagen regions are linked with OS in DLBCL patients. To the best of our knowledge, this is the first study that employs automated analysis to identify prognostic factors in DLBCL.

5.1 Related Work

Computational pathology has paved the way for histology based patient survival analysis. Recently, Zhu *et al.* [138] studied geometry and texture features to detect and segment cells in non-small cell lung cancer (NSCLC). Selected handcrafted features were fed into a supervised principal component regression model to improve predictive performance. Weng *et al.* [139] demonstrated that histology-driven imaging data can better describe the tumour morphology and outperformed conventional biomarkers for predicting survival in NSCLC patients. They employed a deep learning model for cell sub-type classification and imaging biomarkers were then identified using cellular features. More recently, a survival analysis based on whole slide histopathological images survival analysis (WSISA) framework of glioma and NSCLC cancer patients was performed [140]. They trained a deep learning survival predictor to aggregate the patient-level predictions from clustered data. One



(a) Tumour Cells

(b) Normal Cells

Figure 5.2: Example of tumour and normal cells from diffuse large B-cell lymphoma.

of the potential shortcomings of prior approaches has been the random selection or uniformly sampled visual fields for survival analysis. The inevitable fact is that randomly sampled regions may not be strongly associated with disease outcome or may belong to a certain region of a WSI. In contrast, we computed summary-level statistics from the entire WSI and define a novel tumour-collagen proximity signature.

Regarding cell detection and classification, a variety of recently proposed methods handle cell detection and classification separately. Ciresan *et al.* [141] presented a deep learning based vanilla classifier to discriminate between mitotic and non-mitotic cells in breast histopathology images. Sirinukunwattana *et al.* [142] proposed a locality sensitive deep learning framework to separately tackle the cell detection and classification on colorectal adenocarcinoma histology images. More recently, Koohababni *et al.* [143] applied a mixture density network that maps the image tile into a probability density function to sample the observed locations of cells. The distribution of cells within the image is modelled using a mixture of Gaussians, where parameters are learned through back-propagation. Unlike previously published approaches, the proposed model unifies these two tasks into one model.

5.2 Methods

5.2.1 Hydra-Net: Cell Detection and Classification

Model Description

Given an image patch $i_n \in \mathbb{R}^{H \times W \times D}$ with height (H), width (W), and depth (D), where image tile, $I = \{(i_n)\}_{n=1}^{N}$, our model utilises the deep convolutional features to simultaneously predict the class probabilities c_n^k, where k is the number of classes, and the centre location $l_n(x, y)$ of a cell.

Figure 5.3 illustrates the proposed Hydra-Net architecture. The i_n is processed by a stack of convolutional layers (CL) followed by a ReLU activation function. For classification, we use convolutional kernels of relatively small receptive fields including 2×2 and 3×3 kernels, whereas stride for all CLs is adjusted as 1 pixel, preserving the spatial resolution after the convolution operation. The last pooling layer is followed by a spatial dropout layer and the softmax classification layer that predicts the label of each i_n into c_n^k where $k \in \{1, 2, 3\}$, including background (with collagen regions), normal and tumour classes. The other head of the Hydra-Net is responsible for predicting the centre of each cell $l_n(x, y)$. The intermediate convolutional features are fed into the attention module that enables the model

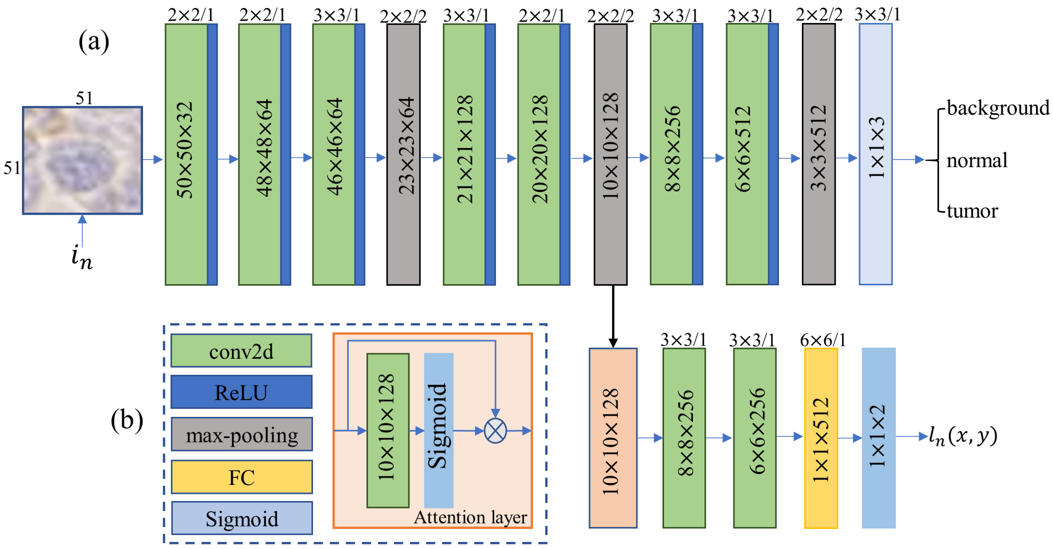


Figure 5.3: (a) A schematic illustration of our proposed Hydra-Net and (b) description of different sub-components.

to learn the structural dependencies lie within the provided activation maps. We use a simple yet effective sigmoid (or soft) attention layer, which can be represented as $C_p = f_p(C_{p-1}, w_p)$ and $A = C_p \odot \sigma(C_p)$, where $\odot$ denotes the Hadamard product, A denotes the output of the attention layer, C_p represents the convolutional features from the p^{th} layer with w_p trainable weights (biases are omitted). The output of the attention module is further processed by a group of convolutional and a fully connected layer of 512 channels.

Model Training

The learning mechanism of the Hydra-Net jointly optimises the combined loss function for multi-class cell classification and regression to predict the location tuple. It optimises 3 different types of weights $\boldsymbol{w} = (\boldsymbol{w_c}, \boldsymbol{w_r}, \boldsymbol{w_s})$ including cell classification, regression and shared multi-task learning (MTL) weights, respectively. The parameters $\boldsymbol{w_s}$ are jointly optimised for both the tasks, and the parameters $\boldsymbol{w_c}$ and $\boldsymbol{w_r}$ are optimised by using the combined loss L as defined below,

$$L = L_c(c, c') + \lambda L_r(l(x, y), l'(x, y)) \tag{5.1}$$

where $L_c(c, c')$ represents *log* loss for true class c'_i. The second part of the loss function is formulated over the true centre location of a cell $l'(x, y)$ and the predicted location $l(x, y)$. The L_r is the l_1 norm between the true and predicted locations, which is more robust to outliers as compared to the l_2 norm. Preferably, we want our model to predict location only for image patches that contain cell, and generally, for background patches, there is no information provided regarding the $l'(x, y)$. Therefore, during training, we introduced a flag λ that is set to 1 for tumour and normal classes and 0 for the background class. During inference, each i_n is processed by a multi-stage ensembling predictor (MEP).

Multi-stage Ensembling Predictor

During test, we use the sliding window strategy with a fixed stride to extract $\{(i_n)\}_{n=1}^{N}$ for a given I. In general, ensemble strategies assist in better understanding of data distribution and overcome the generalisation error [144, 145]. We split our dataset into 3 sub-groups and for each sub-group we separately train the Hydra-Net $(f(i_n))$. Each i_n is processed by a multi-stage ensembling method, as shown in Figure 5.4. Similarly, each sub-group model $(f_1(i_n)$,$f_2(i_n)$,$f_3(i_n))$ separately predicts cell location and class probabilities. For background patches, the $l_n(x, y)$ information is irrelevant and therefore, for follow-up analysis we only consider i_n where $\text{argmax}(c_n^k) \neq 1$ (background class). Further, we defined a probability map $\hat{i}_n \in \mathbb{R}^{H\times W\times D}$ for the output $(l_n(x, y)$, $c_n^k)$ of each sub-group as below

$$\hat{i}_n = g(c_n^k; l_n(x, y)) = \begin{cases} \max\limits_k(c_n^k)\dfrac{1}{1 + \left\|\hat{l}(\hat{x}, \hat{y}) - l(x, y)\right\|_2^2}, & \text{if } \left\|\hat{l}(\hat{x}, \hat{y}) - l(x, y)\right\|_2 \leq r \\ 0, & \text{otherwise} \end{cases} \tag{5.2}$$

$\max_k(c_n^k)$ is a weight that represents the height of $\hat{i}_n$ and $\hat{l}(\hat{x}, \hat{y})$ denotes the spatial coordinates of $\hat{i}_n$.

The value of $\hat{i}_n$ is 0 if the spatial coordinates lie outside a predefined radius r, for our experiments, we set the radius as 9 pixels. Certainly, $\hat{i}_n$ has the highest value at the predicted $l_n(x, y)$, which denotes the centre of a cell. To aggregate $\hat{i}_n$ from all the sub-groups, we compute the pixel-wise average over the computed probability maps. The final results on $\hat{I} = \{(\hat{i}_n)\}_{n=1}^{N}$ is obtained by an empirically chosen threshold, which in our case is 0.45.

5.2.2 Digital Proximity Signature

Computation of DPS is based on two core components: *a)* select a set of reference points $G = \{G_m\}_{m=1}^{M}$ within the collagen regions, where M is the number of sampled points, and *b)* perform proximity analysis between tumour cells and G_m.

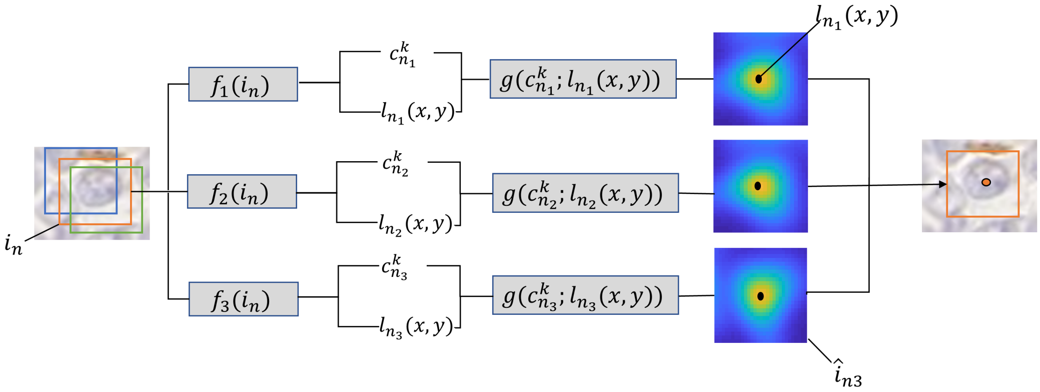


Figure 5.4: An illustration of the multistage ensembling prediction strategy. The coloured rectangular boxes in input image show different i_n using a sliding window strategy.

Collagen Localisation

The main objective of localising G_m is to select a set of representative points from the collagen regions. One straightforward approach is to randomly sample M points from the collagen regions of I. The major pitfall of following such a method is that we may end up selecting G_m from a particular region of I or they may lie close to each other. However, for proximity analysis, it is crucial to sample G_m from spatially distinct regions of I. Therefore, we first segmented the collagen regions by performing stain deconvolution separating I into 3 channels, Haematoxylin, DAB, and background. Collagen VI antibody generally binds with DAB channel having low-intensity values and therefore we empirically choose a relatively high threshold $\tau = 0.82$ to binarise the DAB channel I_{DAB}. The highlighted region in Figure 5.5(C) (teal colour) shows the segmented collagen region. We then compute the medial axis of $I_{DAB}(\tau)$ to retain the connectivity structure of collagen fibre, as defined below

$$Z = (I_{DAB}(\tau) \ominus tS) - (I_{DAB}(\tau) \ominus tS) \circ S \tag{5.3}$$

where $\ominus$ and $\circ$ denote the morphological erosion and opening respectively, and S is the structuring element with size t. Further, we split Z into small grids of size $p \times p$, where $Z = \{(z_m)\}_{m=1}^{M}$ and $p = 256$. The next task is to select a G_m from each z_m, in this regard, we compute the Euclidean distance between the centre point $z_m(p/2, p/2)$ and the spatial coordinates of medial axis $z_m(x_\alpha, y_\alpha)$, where α

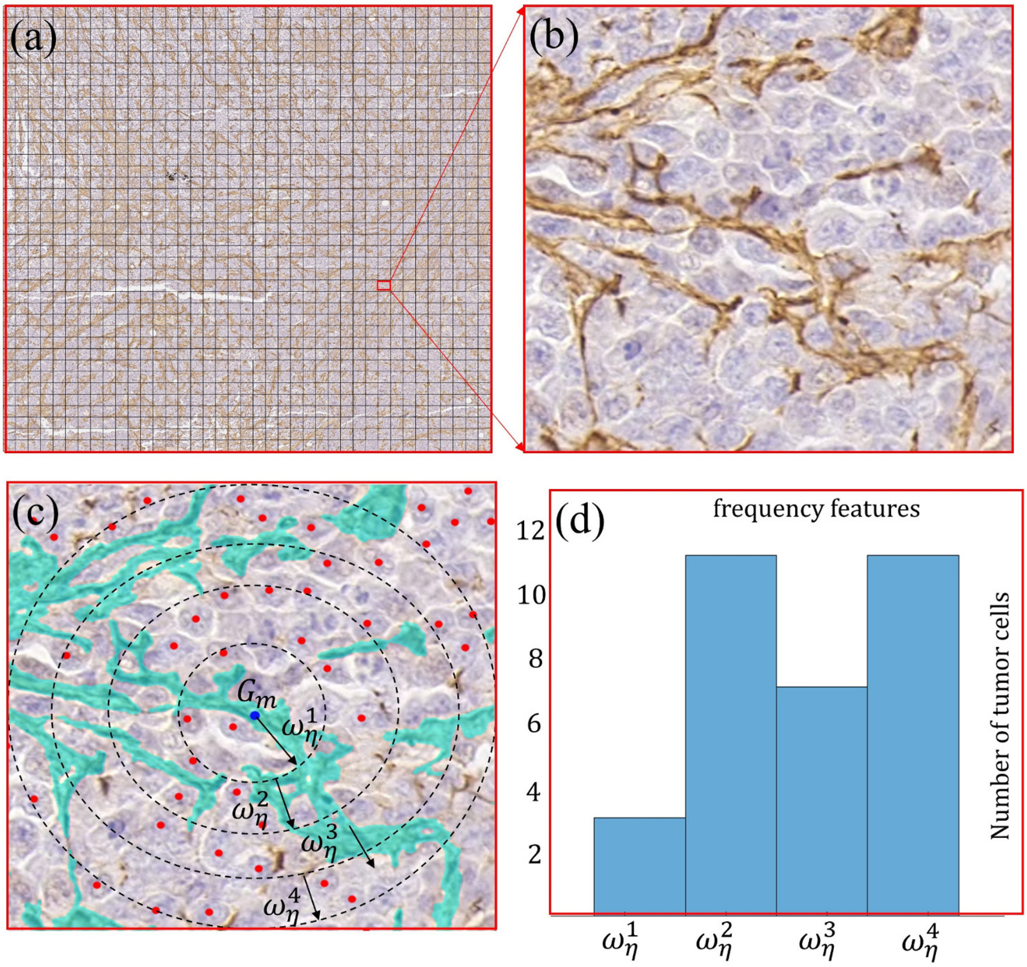


Figure 5.5: (a) Given image I with $p \times p$ small grids, (b) a zoomed-in region from I, (c) highlighted region shows collagen segmentation, black-dashed circles represent different proximity regions for DPS, red disks shows predicted tumour cells and (d) extracted frequency features.

represents the total points in medial axis, as given below

$$G_m(x,y) = \min_{x,y} \big(E(z_m(x_\alpha, y_\alpha), z_m(\frac{p}{2}, \frac{p}{2}))\big) \tag{5.4}$$

where $E(.)$ denotes the Euclidean distance. Finally, the collagen region lying close to the centre of z_m is selected as the reference point, as shown in Figure 5.5(c).

Collagen-tumour Proximity Analysis

For each G_m, we compute a set of frequency features by counting the number of tumour cells within its vicinity, as shown in Figure 5.5(d). The implicit advantage offered by frequency features are rotation, translation, and scale invariance. Similarly, we compute the features for the entire dataset $\omega = \{(\omega_\eta)\}_{\eta=1}^{\nu}$, where ν represents the number of collagen reference points from all the WSIs of the dataset. We then perform clustering on the computed features in order to assign labels to the segmented collagen regions.

We use a simple yet effective method, k-means clustering ($k = 4$) to arrange the frequency features into clusters. The algorithm randomly selects the first centroid and the remaining centroids are chosen based on the largest minimum distance to the preceding centroids, as defined below

$$\hat{\eta} = \text{argmax}_\eta \big(\min_\kappa \|\omega_\eta - S_\kappa\|_2^2\big) \tag{5.5}$$

where S_κ denotes the centroid of the κ^{th} cluster. The algorithm iteratively updates the cluster centroids by computing the distance between centroids and the data points using the Jensen-Shannon divergence, as defined below

$$JSD(S_\kappa \parallel \omega_\eta) = \frac{1}{2} KLD(S_\kappa \parallel \Omega) + \frac{1}{2} KLD(\omega_\eta \parallel \Omega) \tag{5.6}$$

where KLD represents Kullback-Leibler divergence and $\Omega = \frac{1}{2}(S_\kappa + \omega_\eta)$. On the basis of clustering results, we assign the clustering labels to the collagen region in the vicinity of G_m. We further normalise the clustering labels to limit their values between 0 and 1. Finally, in order to compute the DPS, we split the normalised clustering labels into 4 categorises and individually aggregate the statistics. The categories include very weak, weak, moderate, and strong association between tumour and collagen regions. The qualitative results of proximity analysis and its corresponding DPS can be seen in Figure 5.6. Our main intuition of identifying DPS into 4 categories is to provide a concise summary of statistics regarding the proximity between tumour and collagen across the WSI.

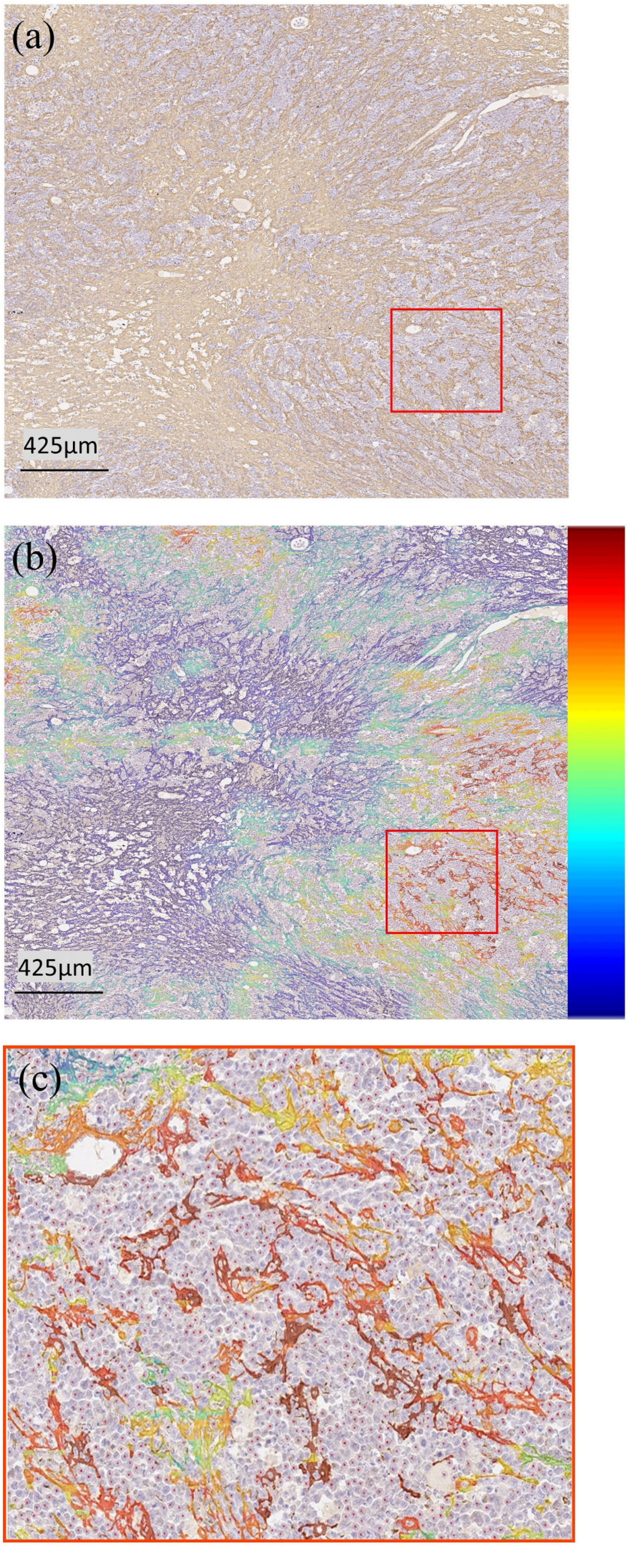


Figure 5.6: (a) A sample visual field extracted from a whole slide image, (b) highlighted collagen regions demonstrate the association between collagen and tumour (c) shows the area marked by rectangles in (a) and (b).

5.3 Experimental Results

This study was performed on two cohorts of DLBCL, each containing 32 and 27 patients. The selected cases contain histopathologist-verified DLBCL WSIs stained using immunohistochemistry for collagen VI with a Haematoxylin counterstain to simultaneously detect collagen VI and nuclear morphology. Paraffin embedded samples of DLBCL (obtained from the UHB Trust, Birmingham, UK and approved by National Research Ethics Service, UK: REC reference 14/WM/0001) were analysed by immunohistochemistry, using citrate buffer antigen retrieval method. The slides were scanned at a pixel resolution of 0.275µm/pixel (40×) using an Omnyx VL120 scanner. We separately report the results of the Hydra-Net and OS analysis with tumour-collagen DPS.

5.3.1 Cell Detection and Classification

The GT for cell detection and classification was collected in 9 cases and marked by an expert. In total, we get 2,617 annotated cells including 2,039 tumours, 462 normal and 116 macrophages. To mitigate the effect of sparse (and limited) GT, we extensively perform the data augmentation by random rotations, cropping, flipping (horizontal or vertical axis), and perturbing the colour distribution, attaining a total of 30,416 patches including 12,100 tumour cells, 9,957 normal cells (including macrophages), and 8,359 randomly selected background patches. Generally, tumour cells and macrophages exhibit a high degree of morphological resemblance, having weakly stained boundaries and hollow structure. For multi-class cell classification task, we used micro averaged F1-score (the other variant is macro averaging) to compute the classification performance. Generally, for imbalanced data micro averages are preferred over macro averages for multi-class problems. In micro averaging, we combine the individual true positives, false negatives, and false positives of each class and compute the micro averaged precision and recall.

We performed 3-fold cross-validation by selecting 2 folds of the dataset for training and the remaining fold for testing. Table 5.1 reports the quantitative results for cell detection and classification, whereas Figure 5.9 shows the prediction accuracy for each class with image patch I of size 51×51 and 61×61. The model has not only identified the tumour and normal cells but it has also distinguished background(collagen) patches (as shown in Figure 5.9). Our results suggest that for the cell classification task, the patch size 51×51 yields the best performance. The patch size 61×61 offers more context information, but they are more likely to contain irrelevant collagen tissue regions which may confound the model in predict-

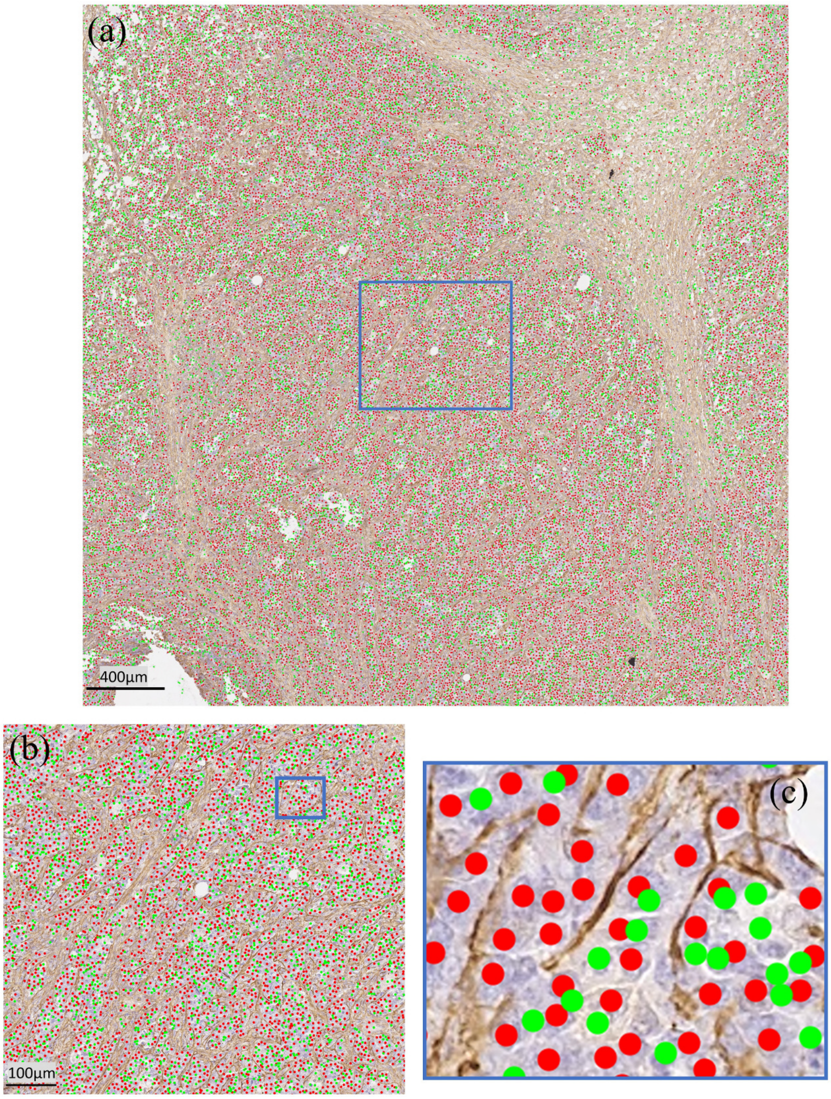


Figure 5.7: Cell detection and classification on an unseen image tile of a DLBCL WSI. Detected tumour cells are phenotyped as red disks and normal lymphocytes including macrophages are represented as green. (a), (b) and (c) shows the overlaid results on the image tiles at different magnification levels. The blue rectangle in (a) representing the region shown in (b), and (c) contains the zoomed-in region from (b), as shown with the blue rectangle. The cell detection and classification was performed on 40× magnification using the proposed Hydra-Net.

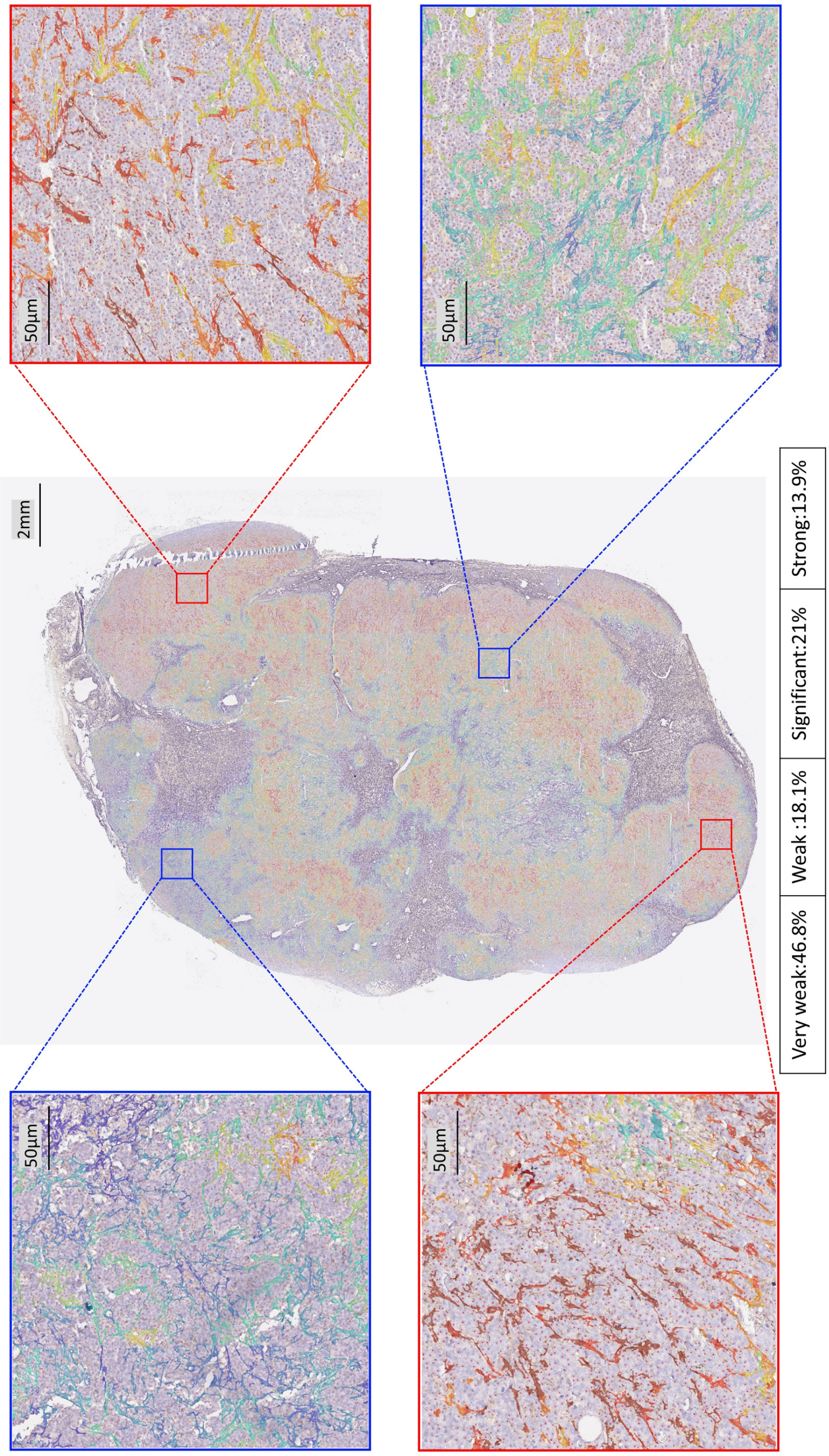


Figure 5.8: Results of collagen-tumour proximity analysis on the whole-slide image (WSI) and the estimated digital proximity signature.

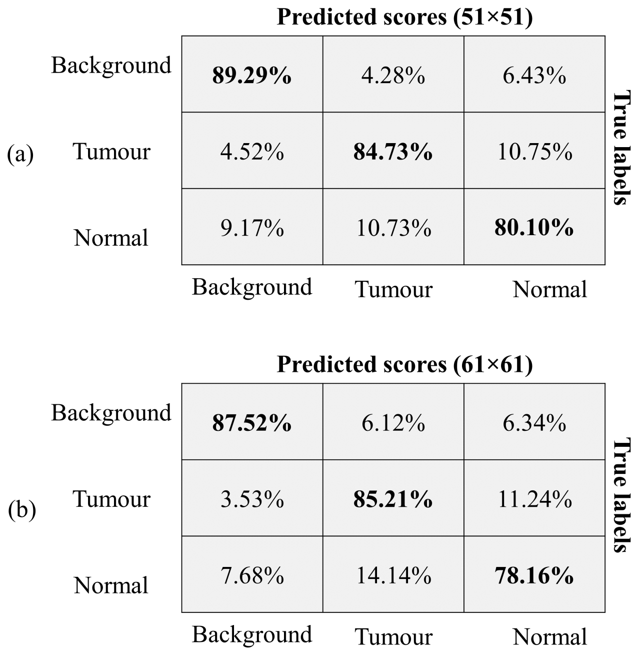
Table 5.1: Comparative analysis for cell detection and classification

Cell Classification			
Methods	Precision	Recall	F1-score
CRImage[146]	0.674	0.689	0.681
Super-pixels[147]	0.729	0.708	0.713
SC-CNN[142]	0.834	0.802	0.817
Hydra-Net without attention	0.828	0.825	0.826
Hydra-Net	**0.839**	**0.842**	**0.84**

ing the correct class of the given image. It is encouraging that the proposed MTL framework outperformed the other competing methods by a reasonable margin. To ensure a fair comparison, we retrained the selected algorithms on our dataset with the same number of image patches after augmentation. Figure 5.7 shows the detection and classification results on a large tile of a WSI where overlaid colour disks representing normal and tumour cells. Overall, the proposed Hydra-Net produces fewer false negatives as compared to other methods. Table 5.1 also highlights the significance of soft attention layer, which is computationally efficient and enables the model to focus only on a particular element (which in our case is $l_n(x, y)$) within the given input. Besides, the proposed MTL framework offers several advantages over single task learning models [142, 146, 147]: first, it overcomes the risk of overfitting by increasing the discriminative ability of shared weights. Second, the inference time for the MTL framework is relatively less as it involves single (forward) propagation for handling multiple tasks.

5.3.2 Survival Analysis

For survival analysis, we selected one of the DLBCL cohorts as discovery and the other one as the validation cohort. The discovery cohort contains 32 DLBCL samples included 10 from female patients and 22 from male patients. 11 had been categorised as germinal centre B-cell (GCB) type, 20 as activated B-cell (ABC) type and one was uncategorised. The age range of the patients was 24 to 90 years, with an average age of 63.6 years. There were 7 stage 1, 10 stage 2, 8 stage 3 and 7 stage 4 tumours. The average follow-up period was 1335 days, and 20 of the 32 patients died. The GT for cell detection and classification was also collected from this cohort. The validation cohort contains 27 DLBCL samples included 9 from female patients and 18 from male patients. Tow cases were later excluded from the analysis due to overstaining and overall quality of the scanned images. Out of those 27, 11 had been categorised as GC-type, and 16 as ABC-type. The age range of the patients was 30 to 86 years,



(a)

Predicted scores (51×51)

	Background	Tumour	Normal
Background	**89.29%**	4.28%	6.43%
Tumour	4.52%	**84.73%**	10.75%
Normal	9.17%	10.73%	**80.10%**

True labels

(b)

Predicted scores (61×61)

	Background	Tumour	Normal
Background	**87.52%**	6.12%	6.34%
Tumour	3.53%	**85.21%**	11.24%
Normal	7.68%	14.14%	**78.16%**

True labels

Figure 5.9: The Hydra-Net confusion matrices for two different patch sizes (a) 51×51 and (b) 61 × 61 for all the three classes of cell classification.

with an average age of 67.2 years. There were 8 stage 1, 2 stage 2, 6 stage 3 and 11 stage 4 tumours. The average follow-up period was 1130 days, and 14 of the 27 patients died.

To obtain the DPS for each WSI, we aggregate the DPSs obtained from image tiles. The qualitative results along with DPS for WSI results are shown in Figure 5.8. We observed that a large part of the tissue region contains very weak association (mainly due to normal regions) between tumour cells and collagen. Therefore, to avoid the over-localisation of very weak associations, we only report survival analysis on the remaining 3 categories excluding the very weak associations (which we named as DPS-3). It is worth mentioning that we separately perform *sum to unity* scaling on the statistics from DPS-3. To evaluate the prognostic significance, for each DPS category, a Kaplan-Meier (KM) and Cox proportional-hazards analysis were performed. The patients were stratified into two groups based on the optimal cut-off of the DPS proportion selected from the discovery cohort. The optimal cut-off on each DPS category was selected where the statistical significance between the two groups is the largest. We report the hazard ratio along with lower and upper 95% confidence interval. Statistical significance of each KM analysis was

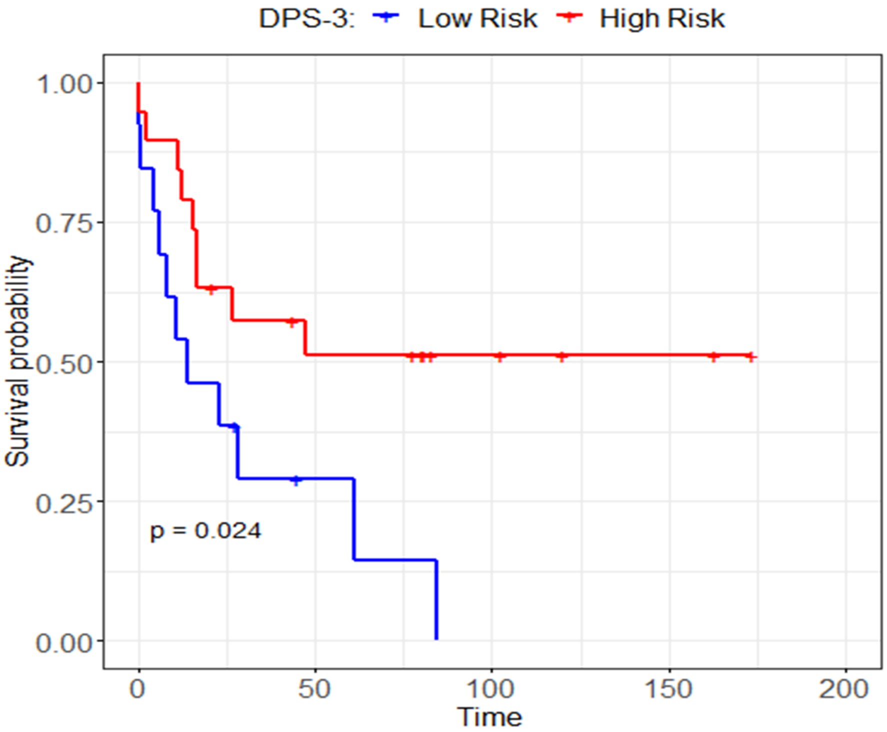
Table 5.2: Contains DPS-3 p-values from overall survival (OS) analysis on validation and discovery cohorts.

Validation cohort					
	p-value	hazard ratio	CI Lower 95%	CI Upper 95%	OS worse in
weak	0.28	2.3	0.49	11	High
moderate	0.0798	0.32	0.084	1.2	Low
strong	**0.0356**	**0.3**	**0.095**	**0.98**	**Low**
Discovery cohort					
	p-value	hazard ratio	CI Lower 95%	CI Upper 95%	OS worse in
weak	0.0031	15	1.4	170	High
moderate	0.0032	0.067	0.006	0.74	Low
strong	**0.0241**	**0.37**	**0.15**	**0.91**	**Low**

quantified using the log-rank test. For DPS-4, the univariate analysis revealed a non-significant trend for different associations of tumour-collagen regions with the inferior OS. The results obtained for DPS-3 are shown in Figure 5.10 and p-values of $0.0031, 0.0032, 0.0241$ were obtained on the discovery cohort and $0.28, 0.0798, 0.0356$ on the validation cohort for weak, moderate, and strongly associated regions, respectively. Moderate and strongly associated regions showed the opposite trend as compared to the weak association. This might be explained by the fact that weak associations are negatively correlated with the other DPS-3 proportions. Table 5.2 gives all p-values for DPS associations with survival. Our DPS-3 image biomarker suggests that patients whose tumours exhibit strong tumour-collagen associations appear to experience superior OS.

5.4 Discussion

Collagen is regarded as the most abundant protein and one of the core macromolecules in the extracellular matrix [148]. The extracellular matrix is a primary component of tumour-stroma and responsible for providing structural support to keep different tissues and cells connected. During tumour progression, along with malignant cells, stromal regions experience structural variations which include collagen degradation, and remodelling [149]. Conventionally, the role of collagen is to resist tumour proliferation whereas there are some studies that have reported the abundance of collagen expression may initiate and promote tumour progression [150]. It is also evident from a number of studies that collagen may acts as a double-edged sword in tumour progression [151]. However, the role of collagen expression



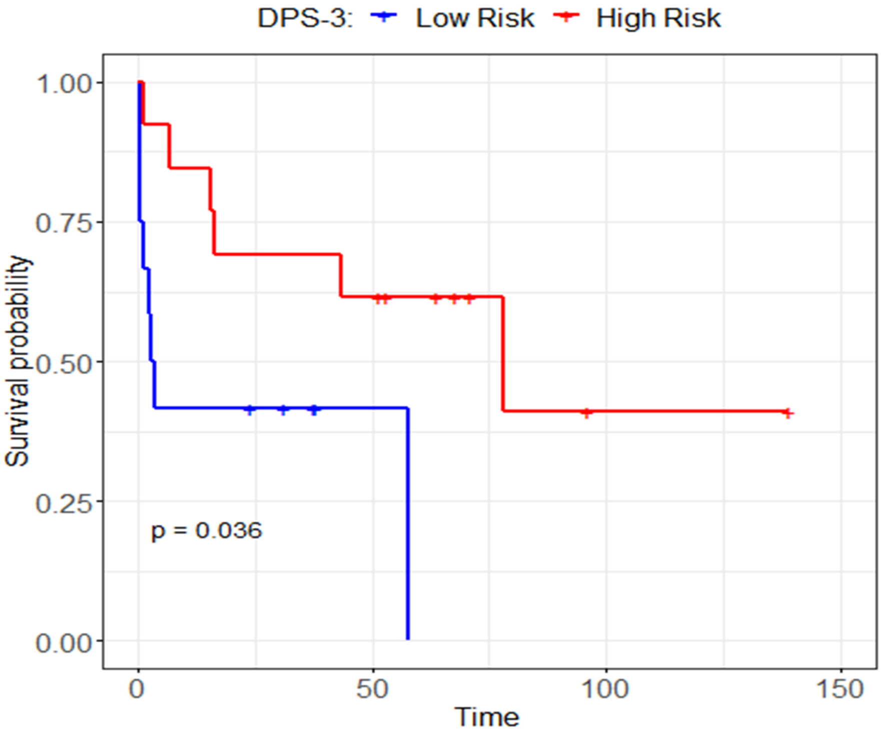


Figure 5.10: Kaplan-Meier survival curves from DPS-3 for strongly associated tumour-collagen regions of discovery (top) and validation cohorts (bottom). The horizontal axis shows the time in months and the vertical axis represents the probability of overall survival.

varies with different types of cancer and their stages. For instance, Mattix *et al.* [152] analysed the quantity of collagen for brain and prostate tumour cases. They observed the quantity of collagen increases with an increase in tumour proliferation for brain tissue. For prostate cancer patients, the opposite trend was observed, for earlier stages, there was an abundance of collagen content as compared to the advanced stages of the tumour where a decrease in collagen content was recorded. More recently, another study performed on DLBCL cases of mouse model shows an increased amount of collagen was observed for earlier stages of lymphoma as compared to later stages [153].

The adoption of computational pathology in routine clinical practice has increased the importance of imaging based biomarkers for predicting patients outcome. The proposed framework may serve as an independent prognostic marker to predict survival in DLBCL patients and may overcome subjectivity and pathologists workload. It can be observed from the experimental results (Table 5.2) that p-value for the strong association is largely comparable for validation and discovery cohorts. Data clustering is an active area of research and has several applications in a variety of domains like data science and computer vision. Our intent here is to select a simple yet effective clustering approach like k-means. It is worth mentioning that in most of the implementations (among different variants of k-means), the time complexity (TC) of k-means grows linearly $O(n)$ as opposed to most of the hierarchical clustering methods where their TC grows $O(n^2)$. One limitation of this work is the selection of parameters like k (in k-means), which are empirically selected for this study and may need appropriate readjustments, depending upon the requirements and datasets. Our main intent of selecting $k = 4$ is to relate the estimated DPS with clinician's practice as we have seen in the previous chapter where pathologist normally score HER2 cases into four categories $(0-3+)$. An interesting extension of this work would be to estimate proximity between tumour and collagen regions using graph-theories or graph CNN. However, such models would require a large scale of annotated regions representing different types of tumour-collagen associations to effectively train the selected graph CNN.

There are 28 known distinct types of collagen and overall they cover one-third of whole body protein [154] and around 90% of bone protein. In this study, we have investigated collagen VI in diffuse large B-cell lymphoma which normally contributes to the formation of muscle, bone, cornea, dermis, and cartilage. Our results show that strong proximity of collagen and tumour cells is linked to better OS in DLBCL patients (discovery cohort: p-value=0.0241, validation cohort: p-value=0.0356). Undoubtedly, in survival analysis, the plausibility of statistical findings increases with an increased number of patients involved in that study. However, to some extent, one major factor that influences the cohort size is the incidence rate of a particular disease. DLBCL is a type of non-Hodgkin lymphomas and their occurrence in combined UK and USA are reported to be 7 to 8 patients from 100,000 people [155]. The significance of the proposed DPS is not only limited to survival analysis. DLBCL is regarded as a heterogeneous malignancy and behaves differently from patient to patient. The DPS also highlights the inter- and intra-tumour heterogeneity within different DLBCL tumours, which might reflect its biological heterogeneity and merit further clinicopathological investigation. Lastly, the proposed model is the first study of its kind that performs automated analysis in exploring prognostic biomarkers for predicting OS in DLBCL patients and it offers an interesting prospect for upcoming studies with a large cohort from multiple centres.

Chapter 6

Conclusions and Future Work

This chapter provides a concise summary of the presented work. We also discuss some of the future directions for further exploration of the concepts and algorithms presented in this thesis.

In this thesis, we presented a range of automated methods for quantitative assessment of cancerous WSIs, to address some of the ongoing challenges in computational pathology. The developed algorithms cover both H&E and IHC stained histology WSIs and include (a) the employment of persistent homology and deep CNNs for tumour segmentation of CRC, (b) a systematic study to evaluate the performance of automated algorithms for HER2 scoring in BC patients, (c) a DRL based attention mechanism that enables the model to *learns where to see*, (d) a blend of deep learning and digital image processing based approaches to compute summary level statistics from a WSI which may serve as a potential biomarker for predicting OS in DLBCL patients.

The ultimate destination of a computational pathology algorithm is to become a part of routine clinical practice. This work has the potential to assist in building computer-assisted diagnostic tools. There are 3 main aspects which need to be considered before the deployment of automated tools. First, the integration of effective pre-processing approaches which adequately handle the artefacts generally occurring during slide preparation and digitisation. These artefacts may misspend precious computational time and influence model performance without any significant performance gain. Second, an extensive large-scale validation would be required on multi-centric datasets, prepared with a variety of slide preparation standards and scanners from different vendors. Third, the incorporation of interactive machine learning (IML) mechanism [156] which enables effective retraining and reduction in classification errors with the help of expert pathologists is needed.

6.1 Tumour Segmentation

In Chapter 2, we have shown an application of persistent homology (based on algebraic topology) in distinguishing cancerous regions from their normal counterparts. It also emphasises developing fast and accurate tumour segmentation algorithms for analysing large-scale histology images. In this work, we have restricted ourselves to deep convolutional networks which are better understood for images. An interesting direction could be to treat the homology profiles as a one dimensional (1D) signal.

There could be several options to explore such temporal data and one of the most well-known approaches is to employ the recurrent models. Unlike conventional models, RNNs build the temporal connections between different inputs $t_N = (t_0, t_1, ..., t_n)$ and pass the relevant contextual information to the proceeding inputs. One straightforward approach can be seen in Figure 6.1. It is worth mentioning that RNNs (especially vanilla RNNs) generally struggle with long-term data dependencies. To overcome the problem of vanishing gradients, LSTM was introduced with input, output and a forget gate. Depending on the requirements, such a setting would require a careful selection of threshold range to avoid the shrinking of the gradient in earlier layers of the LSTM.

6.2 HER2 Scoring Contest

Accessing high-quality medical imaging datasets has always been a challenge for computer science researchers. On the other hand, validation and evaluation of novel automated methods on state-of-the-art datasets is crucial for attaining generalisability. We believe that HER2 Scoring Contest has provided a baseline for researchers in computational pathology community for comparing their automated/semi-automated scoring methods and contribute towards advancing the automated analysis of IHC stained WSIs.

A potential extension to appropriately tackle the borderline (2+) cases of HER2 scoring is by incorporating FISH dataset. All cases with score 2+ are routinely recommended for further FISH testing to validate HER2 over-expression at the gene level. It would be an added advantage if the automated methods could be trained with FISH GT to predict the final outcome and the potential for automated algorithms in calling the actual final HER2 status with reproducible accuracy could be demonstrated. For this, a larger series with 2+ cases along with FISH data would need to be tested. Indeed, there have been promising studies that indicate that automated image analysis for HER2 instead of manual assessment may reduce

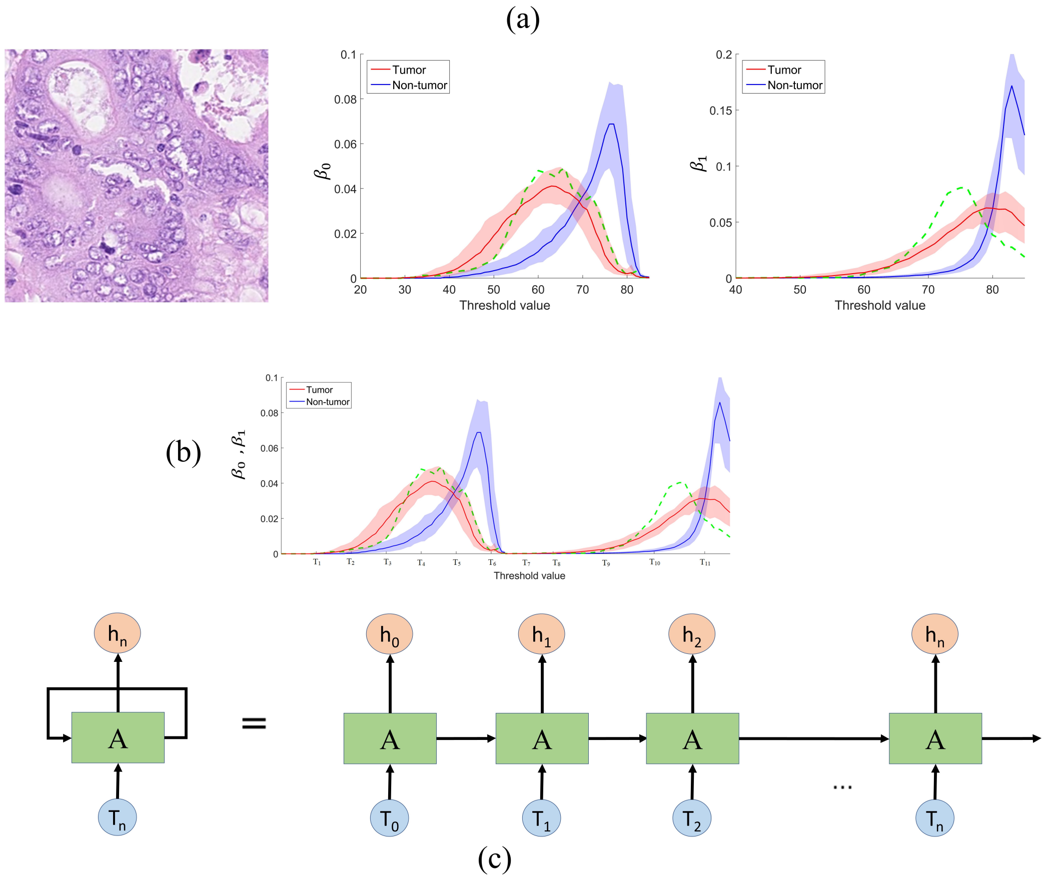


Figure 6.1: An illustration of using recurrent neural networks for handling homology profiles. (a) Independently compute the β_0, and β_1 (b) concatenate the feature representation of β_0, β_1 (c) feed the combined feature representation to a recurrent model.

the need for supplementary FISH testing by up to 68% [157]. In a diagnostic setting, this would significantly reduce costs and turn-around time. During the last decade, IHC staining has become ubiquitous in pathology labs around the world and the role of IHC evaluation in a high-throughput setting becomes key for IHC based companion diagnostics. Other possible extensions of digital pathology could be to automate the over-expression of the programmed death 1 (PD-1) receptor and its ligand (PD-L1), to evaluate anaplastic lymphoma kinase (ALK) protein and proto-oncogene tyrosine-protein kinase ROS1 in lung cancers [158]. The AI based algorithms would be more effective if IHC staining and scoring methods were treated as a composite assay [158, 159, 160]. The varying staining protocols and scoring parameters may restrain the effectiveness of AI based automated scoring algorithms including the HER2 scoring but with sufficiently variable data from different centres, AI algorithms could be trained to overcome that problem.

6.3 Learning Where to See

Reinforcement learning has greatly contributed to the development of self-driving autonomous cars and industrial automation. In fact, nowadays, DRL based algorithms are serving as the backbone for academic and industrial research on autonomous cars and robotics [161, 162]. This kind of learning enables the model to sequentially interact with its environment and respond with suitable line of actions to maximise the reward function. The usability and scope of DRL have been not fully explored in the domain of computational pathology. We proposed a DRL based attention mechanism that demonstrates the usability of DRL for automated IHC scoring.

One possible extension of our work to achieve end-to-end learning would be to devise a multi-stage (or multi-scale) attention mechanism. This can be achieved by formulating multi-stage cascaded models. In the first stage, the model would analyse the given WSI at a lower magnification (1.25× or lower) to identify the potential ROIs and in the second stage the attention model would analyse selected ROIs at higher resolutions (40×, 20×) to learn the discriminative features. However, such a model would require a prudent selection of reward function(s) and tuning of hyper-parameters to ensure the inclusion of diagnostically relevant regions in the first stage. In addition, the model would require a large number of WSIs to effectively optimise the weights for the first stage of the attention model.

6.4 Tumour-Collagen Proximity Analytics

In chapter 5, we have described a tumour-collagen proximity signature that can predict overall survival in DLBCL. We proposed the computing of the DPS by first detecting and classifying tumour cells using a multi-headed Hydra-Net. Second, we performed tumour-collagen proximity analysis by aggregating tumour cell statistics within the vicinity of collagen regions. DLBCL is a heterogeneous malignancy and tumour progression may vary among different patients. We have provided a baseline to highlight the inter- and intra-tumour heterogeneity, which may be helpful in better understanding of disease progression. One potential extension of this work could be to explore the collagen patterns and investigate why these patterns vary from case to case. Does it have any underlying molecular basis? Is it a cause or consequence of tumour behaviour? These findings are compelling observations which merit further clinicopathological investigations on larger patient cohorts.

6.5 Concluding Remarks

In this thesis, we have presented automated algorithms for quantitative assessment of cancerous whole slide images. The presented work provides a platform for further investigation on the role of algebraic topology and attention based reinforcement learning models. The presented studies demonstrate that machine learning based automated algorithms can be effectively used in the area of computational pathology. The journey towards developing computer-assisted diagnostic systems requires substantial resources for conducting large-scale clinical validation of automated methods for assisting experts in fighting one of the deadliest diseases of the current era.

Appendix A

The Proposed Algorithms in the HER2 Scoring Contest

The work presented in this appendix contains the description of the presented algorithms in the HER2 scoring contest.

A.1 Related Work

Automated image analysis is regarded as a solution [29, 163] to overcome the inter- and intra- observer variations found in the conventional assessment of tissue slides. In literature, a wide range of handcrafted features was proposed for IHC scoring algorithms [108, 164].

For instance, Choudhury *et al.* [165] proposed an averaged threshold measure (ATM) for scoring of digitised images of IHC stained tissue microarrays. A set of arbitrarily chosen thresholds were selected, whereby an optimal threshold using the ATM is used for calculating the percentage of the stained area. The proposed ATM statistic presented as a generalisation of the HSCORE [166] statistic for scoring IHC slides. Reyes-Aldasoro *et al.* [167] presented an alternative approach for automated segmentation of microvessels in IHC tumour slides. For segmentation, distinguishing hues of stained vascular endothelial nuclei and tissue regions were explored to extract the seeds for a region-growing model. Their post-processing of segmented microvessels from CD31 immunostaining contained three steps, closing morphological objects from tumour margins, combining isolated objects, and splitting objects into individual vessels with having multiple lumen. Although the thresholding approaches perform well on a specific dataset, they are likely to fare not as well on an unseen dataset as distinctive hues can be significantly varying. A potential reason for such variation lies in the staining process, as the histology slides normally stained at different occasions with inconsistent concentrations often exhibit large variations in colour and appearance. Such differences in slide preparation make the colour and morphological appearance of tissue components more unpredictable.

Kuse *et al.* [168] used local isotropic phase symmetry measure as a significant feature for beta cell detection and lymphocytes. By calculating the peak of median phase energy after stain normalisation but due to heterogeneous appearance and often-clumped structure makes nuclei segmentation a non-trivial task. Khan *et al.* [108] used stain quantisation for the scoring of the Estrogen Receptor (ER) and Progesterone Receptor (PR) by determining the amount of chromatin material and protein content from IHC stained WSIs. Ali *et al.* [169] used astronomical algorithms for the scoring of ER on IHC stained images of breast cancer. However, in this contest, the classical machine learning approaches have been outperformed by deep learning approaches. Most of the published algorithms are based on different approaches with different dataset whereas this contest provides a platform where participants can develop and validate the performance of their algorithms on the same dataset.

A.2 Description of the Proposed Automated Methods

This section provides a concise description of automated methods employed by some of the top-ranked teams.

A.2.1 VISILAB

In this method, a state-of-the-art GoogLeNet [99] was implanted to predict the HER2 score and the percentage of the complete cell membrane.

Data Preparation: A set of representative patches of the four HER2 scoring classes were extracted from the ground truth WSIs. Additionally, an extra class was employed to collect background samples. These extracted patches from training WSIs were 68 × 68 pixels size each. A total of 5750 patches were selected with an average of 1150 patches per class. The dataset was further split into training (75%) and validation (25%) dataset.

Training: Among several state-of-art CNNs, GoogLeNet was finally selected for submission according to the results on the validation dataset. The prepared dataset was used for training, by selecting 0.01 as the base learning rate, with a decreasing policy over 50 epochs, using the Stochastic Gradient Descent.

Classification: The algorithm takes a WSI and applies a grid technique to obtain the corresponding patches, with a similar size than the ones from the training dataset. These are later classified with the trained model, whose output is a class prediction and a percentage of confidence over that decision.

HER2 Scoring: Once every single patch is classified, a single class score is provided for the WSI. The decision rule takes into account the percentage of patches that belong to each class (omitting the background, which was treated as a separate class) using the following criteria: starting from class 3+ to class 1+, the first one to achieve at least 10% of patches is chosen as the final decision. Regarding the percentage of cells with full membrane staining, a customised expert rule was developed by calculating the staining density for different nuclei. As a result, a relationship between the classes percentage distribution and the percentage of membrane cell staining was discovered.

A.2.2 FSUJena

The algorithm for the automated HER2 scoring was based on Alexnet [52] CNN. In this method, an activation matrix was extracted after convolution layers to compute the bilinear filters.

At the first, ROIs were manually probed and patches of size 227 × 227 were randomly extracted at 20×. The pre-trained version of Alexnet was used from ImageNet dataset for further training on contest dataset. For each patch in the training dataset, an activation matrix was extracted after convolutional layers. The activations can be represented as a tensor $x \in \mathbb{R}^{w \times h \times d}$ comprised of d-dimensional vectors in a $w \times h$ spatial grid. The bilinear features [170, 171] were further computed as the Gramian G matrix by summing up dyadic products along the spatial dimensions: $G = \sum_{i,j} x_{i,j} x_{i,j}^T$. The matrix G contains the second-order statistics of the CNN features and have been found to be extremely useful for fine-grained recognition tasks. Then the square root and L_2 normalisation of G were employed to increase the numerical stability of further processing steps [171].

To differentiate among four scoring classes a multi-class logistic regression was used. It was also observed that using a pre-trained network on ImageNet dataset is also beneficial to avoid the overfitting issues. In preliminary results, the bilinear features approach outperformed the conventional CNN activations. During the test, an average was calculated for all the randomly cropped patches. To predict the

PCMS the mean tumour cell percentage seen in the training set of for a particular class as an estimate.

A.2.3 Huangch

In this approach, a range of handcrafted features extracted from the IHC stained slides after performing the stain deconvolution. The handcrafted features were then fed into a model of multi-class AdaBoost decision trees.

Pre-processing: Control tissues were extracted to developed a pseudo colour space for stain deconvolution [127] to obtain the two staining vectors. Further, mean filtering was performed to record the local maximal points. The patches were selected from each WSI on the basis of local maximal points as they were representing the strongest HER2 stained overexpression signals

Feature Extraction and Classification: A combined but numerically independent features vector space constructed by including Gabor Filtering, Features of Fractal Dimension by Differential Box-Counting, multi-wavelet methods, histogram statics methods, and grey-level (over all colour channels) co-occurrence based method [172, 173] etc.

For predicting the HER2 score and the PCMS, a model of multi-class AdaBoost decision-trees was employed to map the features vector of each patch to a predicted value. This model is known as Stagewise Additive Modelling using a Multi-class Exponential [174] loss function (SAMME). The model composed by a series of decision-trees by assigning a weight to each decision-tree. Whereas while training, a pool of decision-trees were generated and after each iteration the best decision-tree was selected with its corresponding weight. After certain iterations, a group of decision-trees were selected for the testing phase.

A.2.4 Team Indus

In this approach, a deep CNN was employed for predicting the HER2 score whereas for estimating the PCMS, a set of handcrafted morphological features were extracted from H&E and IHC stained slides.

Pre-processing: Patches with an average strength (intensity) lies higher then a certain threshold were selected for training the CNN.

HER2 Score Prediction: The presented CNN architecture contains five convolutional layers, one concatenation layer with following two fully connected and one classification layer. After each convolution and fully connected layer, a ReLu activation was performed whereas for classification layer a softmax activation was placed. After convolution layers, a concatenation layer was also positioned. The concatenation layer combines the activation maps from the convolution layers and the average control tissue intensity for the corresponding WSI from which the patches were originated. The weights for training CNN were updated using mini batch gradient descent (learning rate = 0.00015, weight decay = 10^6, Nesterov momentum = 0.95, batch size = 32). The CNN was trained over 41K patches generated each of size 224x224 from 52 training WSIs for 65 epochs. During testing, the trained network assigned a score to each patch of a WSI and to aggregate the patch scores into a single HER2 score following criteria was proposed. Let n_0, n_1, n_2 and n_3 be the number of patches scored as 0, 1+, 2+ and 3+ respectively and N be the total number of patches generated from a WSI.

PCMS Estimation: To estimate the PCMS, first tumour regions were identified by extracting the morphological features from tumour and normal regions of H&E images. After that stain normalisation [127] was performed by selecting the haematoxylin channel to segment the nuclei using Otsu thresholding. Further, nuclei contours were fit around each individual structure and filtered on the basis of area and eccentricity. This resulted in tumour identification regions by detecting tumour nuclei based on their roundness and size. In order to estimate the extent of membrane staining, the morphological features were extracted from an IHC image. In addition, a contagious chicken-wire pattern was observed for complete membrane stained regions whereas other tissue components result in a fragmented/broken-up skeleton. Further, by filling holes in the chicken-wire skeleton and by measuring similarity with the original binary image the extent of membrane staining was estimated. The PCMS is estimated by calculating the ratio between the extent of membrane staining and tumour identification regions.

A.2.5 MUCS

In this submission, the well-known neural networks Alexnet [52] and GoogLeNet [99] were adapted by adjusting the layer specific parameters, such as kernel size, stride, and padding. There were three submissions from the MUCS team with two submissions using Alexnet (MUCS-1 and MUCS-2) and one using GoogLeNet (MUCS-3).

Model Training: The training dataset was obtained by hand-picking the regions of interest from 52 training IHC images that were considered to contain the most representative samples from each class. The regions were selected from the low resolution (0.625×) and mapped to the highest resolution (40×) whereupon each region was divided into 128 × 128 pixel patches. The MUCS-1 trained network had four output classes with corresponding HER2 scores from 0 to 3+. MUCS-2 and MUCS-3 had an additional output class for the background. The background class contained the regions with a texture having only a weak appearance of nuclei (without blueish or brownish colour). The training dataset for MUCS-2 was extended by data augmentation (rotation and mirroring) and by adding the hand-picked regions from test images (without knowing the classification of the slide it originated from). The total patches for MUCS-1, MUCS-2 and MUCS-3 were 29,000, 31,9000 and 33,500, respectively. The training images were divided between actual training data (75%) and validation data (25%). For all three submissions, the base learning rate was set to 0.001, and the learning rate was dropped every one-third of the maximum iterations by a factor of 10 ($\gamma = 0.1$). The mean pixel value was subtracted from the training dataset.

Model Inference: For testing, the common regions from H&E and IHC were selected at a low resolution and those regions were mapped to maximum resolution to generate the patches for testing. Further, adaptive thresholding was applied to each patch, with an offset of 10, to produce a binary image. If the proportion of ones in the binary image was smaller than a factor of 0.9, then patch was classified with the trained model, otherwise, the patch was marked as background and therefore did not require classification. The HER2 score for a WSI was determined using the classified patches as follows:

- Score 3+, if patches with class 3+ was greater than or equal to 10% of total patches
- Score 2+, if patches with class 2+ was greater than or equal to 10%, or patches with class 3+ was between 1% and 10%, of total patches
- Score 1+, if patches with class 1+ was greater than or equal to 10% of total patches
- Score 0, otherwise

The confidence value for each WSI was calculated by averaging the confidence values of each patch. PCMS was calculated by summing the number of score 3+

and 2+ patches and dividing the sum by total number of patches (excluding the background) as given below

$$PCMS = 100(n_2 + n_3)(\sum_{s=0}^{3} n_s)^{-1} \tag{A.1}$$

where n is the number of patches given score $s, s \in \{0, 1, 2, 3\}$.

Bibliography

[1] Ross Cagan and Pablo Meyer. Rethinking cancer: current challenges and opportunities in cancer research, 2017.

[2] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[3] BWKP Stewart, Christopher P Wild, et al. World cancer report 2014. 2014.

[4] Robert Lanza, John Gearhart, Brigid Hogan, Douglas Melton, Roger Pedersen, E Donnall Thomas, James A Thomson, and Michael West. *Essentials of stem cell biology*. Elsevier, 2005.

[5] National cancer institute, ”tumor grade”., 2019. URL `https://www.cancer.gov/about-cancer/diagnosis-staging/prognosis/tumor-grade-fact-sheet?redirect=true`.

[6] Marisa T Carriaga and Donald Earl Henson. The histologic grading of cancer. *Cancer*, 75(S1):406–421, 1995.

[7] J Ferlay, DM Parkin, and E Steliarova-Foucher. Estimates of cancer incidence and mortality in europe in 2008. *European journal of cancer*, 46(4):765–781, 2010.

[8] Lindsey A Torre, Freddie Bray, Rebecca L Siegel, Jacques Ferlay, Joannie Lortet-Tieulent, and Ahmedin Jemal. Global cancer statistics, 2012. *CA: a cancer journal for clinicians*, 65(2):87–108, 2015.

[9] Carolyn C Compton. Updated protocol for the examination of specimens from patients with carcinomas of the colon and rectum, excluding carcinoid tumors, lymphomas, sarcomas, and tumors of the vermiform appendix: a basis

for checklists. *Archives of pathology & laboratory medicine*, 124(7):1016–1025, 2000.

[10] Stanley R Hamilton, Lauri A Aaltonen, et al. *Pathology and genetics of tumours of the digestive system*, volume 48. IARC press Lyon:, 2000.

[11] Jiemin Ma and Ahmedin Jemal. Breast cancer statistics. In *Breast Cancer Metastasis and Drug Resistance*, pages 1–18. Springer, 2013.

[12] Breast cancer statistics, cancer research uk., September 2017. URL `http://www.cancerresearchuk.org/cancer-info/cancerstats/types/breast/`.

[13] Jay R Harris, Marc E Lippman, C Kent Osborne, and Monica Morrow. *Diseases of the Breast*. Lippincott Williams & Wilkins, 2012.

[14] Emad A Rakha, Maysa E El-Sayed, Sindhu Menon, Andrew R Green, Andrew HS Lee, and Ian O Ellis. Histologic grading is an independent prognostic factor in invasive lobular carcinoma of the breast. *Breast cancer research and treatment*, 111(1):121–127, 2008.

[15] Mamatha Chivukula, Rohit Bhargava, Adam Brufsky, Urvashi Surti, and David J Dabbs. Clinical importance of her2 immunohistologic heterogeneous expression in core-needle biopsies vs resection specimens for equivocal (immunohistochemical score 2+) cases. *Modern Pathology*, 21(4):363, 2008.

[16] Sylvie Ménard, Stefania Fortis, Fabio Castiglioni, Roberto Agresti, and Andrea Balsari. Her2 as a prognostic factor in breast cancer. *Oncology*, 61(Suppl. 2):67–72, 2001.

[17] Walter P Carney. Her2 status is an important biomarker in guiding personalized her2 therapy. 2005.

[18] Shaoying Li, Ken H Young, and L Jeffrey Medeiros. Diffuse large b-cell lymphoma. *Pathology*, 50(1):74–87, 2018.

[19] A Smith, D Howell, R Patmore, A Jack, and E Roman. Incidence of haematological malignancy by sub-type: a report from the haematological malignancy research network. *British journal of cancer*, 105(11):1684, 2011.

[20] MARGARET P Sullivan and JAMES J Butler. Non-hodgkin's lymphoma of childhood. *Clinical pediatric oncology*, pages 313–336, 1973.

[21] Bertrand Coiffier, Eric Lepage, Josette Brière, Raoul Herbrecht, Hervé Tilly, Reda Bouabdallah, Pierre Morel, Eric Van Den Neste, Gilles Salles, Philippe Gaulard, et al. Chop chemotherapy plus rituximab compared with chop alone in elderly patients with diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346(4):235–242, 2002.

[22] AV de Jonge, TJA Roosma, I Houtenbos, WLE Vasmel, K van de Hem, JP de Boer, T van Maanen, G Lindauer-van der Werf, A Beeker, GJ Timmers, et al. Diffuse large b-cell lymphoma with myc gene rearrangements: Current perspective on treatment of diffuse large b-cell lymphoma with myc gene rearrangements; case series and review of the literature. *European Journal of Cancer*, 55:140–146, 2016.

[23] Andreas Rosenwald, George Wright, Wing C Chan, Joseph M Connors, Elias Campo, Richard I Fisher, Randy D Gascoyne, H Konrad Muller-Hermelink, Erlend B Smeland, Jena M Giltnane, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346(25):1937–1947, 2002.

[24] Roland Schmitz, George W Wright, Da Wei Huang, Calvin A Johnson, James D Phelan, James Q Wang, Sandrine Roulland, Monica Kasbekar, Ryan M Young, Arthur L Shaffer, et al. Genetics and pathogenesis of diffuse large b-cell lymphoma. *New England Journal of Medicine*, 378(15):1396–1407, 2018.

[25] Jingxuan Wang, Kun Gao, Wanting Lei, Lina Dong, Qijia Xuan, Meiyan Feng, Jinlu Wang, Xiangnan Ye, Tuan Jin, Zhongbai Zhang, et al. Lymphocyte-to-monocyte ratio is associated with prognosis of diffuse large b-cell lymphoma: correlation with cd163 positive m2 type tumor-associated macrophages, not pd-1 positive tumor-infiltrating lymphocytes. *Oncotarget*, 8(3):5414, 2017.

[26] Zihang Chen, Xueqin Deng, Yunxia Ye, Limin Gao, Wenyan Zhang, Weiping Liu, and Sha Zhao. Novel risk stratification of de novo diffuse large b cell lymphoma based on tumour-infiltrating t lymphocytes evaluated by flow cytometry. *Annals of hematology*, 98(2):391–399, 2019.

[27] Makito Miyake, Shunta Hori, Yosuke Morizawa, Yoshihiro Tatsumi, Michihiro Toritsuka, Sayuri Ohnishi, Keiji Shimada, Hideki Furuya, Vedbar S Khadka, Youping Deng, et al. Collagen type iv alpha 1 (col4a1) and collagen type xiii alpha 1 (col13a1) produced in cancer cells promote tumor budding at the

invasion front in human urothelial carcinoma of the bladder. *Oncotarget*, 8 (22):36099, 2017.

[28] Paolo P Provenzano, David R Inman, Kevin W Eliceiri, Justin G Knittel, Long Yan, Curtis T Rueden, John G White, and Patricia J Keely. Collagen density promotes mammary tumor initiation and progression. *BMC medicine*, 6(1):11, 2008.

[29] Metin N Gurcan, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2:147–171, 2009.

[30] Mike May. A better lens on disease. *Scientific American*, 302(5):74–77, 2010.

[31] Simon S Cross, Samar Betmouni, Julian L Burton, Asha K Dubé, Kenneth M Feeley, Miles R Holbrook, Robert J Landers, Phillip B Lumb, and Timothy J Stephenson. What levels of agreement can be expected between histopathologists assigning cases to discrete nominal categories? a study of the diagnosis of hyperplastic and adenomatous colorectal polyps. *Modern Pathology*, 13(9): 941, 2000.

[32] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM van der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, 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.

[33] Siyamalan Manivannan, Wenqi Li, Jianguo Zhang, Emanuele Trucco, and Stephen J McKenna. Structure prediction for gland segmentation with hand-crafted and deep convolutional features. *IEEE transactions on medical imaging*, 37(1):210–221, 2018.

[34] Emad A Rakha, Sarah E Pinder, John MS Bartlett, Merdol Ibrahim, Jane Starczynski, Pauline J Carder, Elena Provenzano, Andrew Hanby, Sally Hales, Andrew HS Lee, et al. Updated uk recommendations for her2 assessment in breast cancer. *Journal of clinical pathology*, 68(2):93–99, 2015.

[35] Lee AD Cooper, Alexis B Carter, Alton B Farris, Fusheng Wang, Jun Kong, David A Gutman, Patrick Widener, Tony C Pan, Sharath R Cholleti, Ashish Sharma, et al. Digital pathology: Data-intensive frontier in medical imaging. *Proceedings of the IEEE*, 100(4):991–1003, 2012.

[36] Adam Goode, Benjamin Gilbert, Jan Harkes, Drazen Jukic, and Mahadev Satyanarayanan. Openslide: A vendor-neutral software foundation for digital pathology. *Journal of pathology informatics*, 4, 2013.

[37] David Taubman and Michael Marcellin. *JPEG2000 image compression fundamentals, standards and practice: image compression fundamentals, standards and practice*, volume 642. Springer Science & Business Media, 2012.

[38] David N Louis, Georg K Gerber, Jason M Baron, Lyn Bry, Anand S Dighe, Gad Getz, John M Higgins, Frank C Kuo, William J Lane, James S Michaelson, et al. Computational pathology: an emerging definition. *Archives of pathology & laboratory medicine*, 138(9):1133–1138, 2014.

[39] Philip D Dunne, Darragh G McArt, Conor A Bradley, Paul G O'Reilly, Helen L Barrett, Robert Cummins, Tony O'Grady, Ken Arthur, Maurice B Loughrey, Wendy L Allen, et al. Challenging the cancer molecular stratification dogma: intratumoral heterogeneity undermines consensus molecular subtypes and potential diagnostic value in colorectal cancer. *Clinical Cancer Research*, 22(16):4095–4104, 2016.

[40] Geert Litjens, Clara I Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen-Van De Kaa, Peter Bult, Bram Van Ginneken, and Jeroen Van Der Laak. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports*, 6:26286, 2016.

[41] Talha Qaiser, Abhik Mukherjee, Chaitanya Reddy PB, Sai D Munugoti, Vamsi Tallam, Tomi Pitkaho, Taina Lehtimki, Thomas Naughton, Matt Berseth, Anbal Pedraza, Ramakrishnan Mukundan, Matthew Smith, Abhir Bhalerao, Erik Rodner, Marcel Simon, Joachim Denzler, Chao-Hui Huang, Gloria Bueno, David Snead, Ian O Ellis, Mohammad Ilyas, and Nasir 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, 2018. ISSN 1365-2559. doi: 10.1111/his.13333. URL `http://dx.doi.org/10.1111/his.13333`.

[42] Alecsandru Ioan Baba and Cornel Câtoi. *Comparative oncology*. Publishing House of the Romanian Academy Bucharest, 2007.

[43] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and ac-

curate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[44] Francesco Bianconi, Alberto Álvarez-Larrán, and Antonio Fernández. Discrimination between tumour epithelium and stroma via perception-based features. *Neurocomputing*, 154:119–126, 2015.

[45] Nina Linder, Juho Konsti, Riku Turkki, Esa Rahtu, Mikael Lundin, Stig Nordling, Caj Haglund, Timo Ahonen, Matti Pietikäinen, and Johan Lundin. Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. *Diagnostic pathology*, 7(1):22, 2012.

[46] Dogan Altunbay, Celal Cigir, Cenk Sokmensuer, and Cigdem Gunduz-Demir. Color graphs for automated cancer diagnosis and grading. *IEEE Transactions on Biomedical Engineering*, 57(3):665–674, 2010.

[47] Adnan M Khan, Hesham El-Daly, Emma Simmons, and Nasir M Rajpoot. Hymap: A hybrid magnitude-phase approach to unsupervised segmentation of tumor areas in breast cancer histology images. *Journal of pathology informatics*, 4(Suppl), 2013.

[48] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6:27988, 2016.

[49] Shazia Akbar, Lee Jordan, Alastair M Thompson, and Stephen J McKenna. Tumor localization in tissue microarrays using rotation invariant superpixel pyramids. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 1292–1295. IEEE, 2015.

[50] Yan Xu, Jun-Yan Zhu, I Eric, Chao Chang, Maode Lai, and Zhuowen Tu. Weakly supervised histopathology cancer image segmentation and classification. *Medical image analysis*, 18(3):591–604, 2014.

[51] Yan Xu, Jun-Yan Zhu, Eric Chang, and Zhuowen Tu. Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 964–971. IEEE, 2012.

[52] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[53] Hao Chen, Xiaojuan Qi, Lequan Yu, Qi Dou, Jing Qin, and Pheng-Ann Heng. Dcan: Deep contour-aware networks for object instance segmentation from histology images. *Medical image analysis*, 36:135–146, 2017.

[54] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

[55] Olaf Ronneberger, Philipp Fischer, and Thomas 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.

[56] Angel Cruz-Roa, Hannah Gilmore, Ajay Basavanhally, Michael Feldman, Shridar Ganesan, Natalie NC Shih, John Tomaszewski, Fabio A González, and Anant Madabhushi. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. *Scientific Reports*, 7:46450, 2017.

[57] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3859–3869, 2017.

[58] Kelin Xia and Guo-Wei Wei. Persistent homology analysis of protein structure, flexibility, and folding. *International journal for numerical methods in biomedical engineering*, 30(8):814–844, 2014.

[59] Zixuan Cang and Guo Wei Wei. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International Journal for Numerical Methods in Biomedical Engineering*, 2017.

[60] Subhrajit Bhattacharya, Robert Ghrist, and Vijay Kumar. Persistent homology for path planning in uncertain environments. *IEEE Transactions on Robotics*, 31(3):578–590, 2015.

[61] Florian T Pokorny, Majd Hawasly, and Subramanian Ramamoorthy. Topological trajectory classification with filtrations of simplicial complexes and

persistent homology. *The International Journal of Robotics Research*, 35(1-3): 204–223, 2016.

[62] Carina Curto. What can topology tell us about the neural code? *Bulletin of the American Mathematical Society*, 54(1):63–78, 2017.

[63] Katharine Turner, Sayan Mukherjee, and Doug M Boyer. Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA*, 3(4):310–344, 2014.

[64] Paul Bendich, James S Marron, Ezra Miller, Alex Pieloch, and Sean Skwerer. Persistent homology analysis of brain artery trees. *The annals of applied statistics*, 10(1):198, 2016.

[65] Olga Dunaeva, Herbert Edelsbrunner, Anton Lukyanov, Michael Machin, Daria Malkova, Roman Kuvaev, and Sergey Kashin. The classification of endoscopy images with persistent homology. *Pattern Recognition Letters*, 83: 13–22, 2016.

[66] Berhanu A Wubie. *Clustering Survival Data Using Random Forest and Persistent Homology.* PhD thesis, University of Alberta, 2016.

[67] Javier Lamar-Leon, Raul Alonso-Baryolo, Edel Garcia-Reyes, and Rocio Gonzalez-Diaz. Persistent homology-based gait recognition robust to upper body variations. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 1083–1088. IEEE, 2016.

[68] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1):17, 2017.

[69] Xiaojin Zhu. Persistent homology: An introduction and a new text representation for natural language processing. In *IJCAI*, pages 1953–1959, 2013.

[70] Herbert Edelsbrunner and John Harer. Persistent homology-a survey. *Contemporary mathematics*, 453:257–282, 2008.

[71] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[72] Robert Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[73] Kazuaki Nakane, Akihiro Takiyama, Seiji Mori, and Nariaki Matsuura. Homology-based method for detecting regions of interest in colonic digital images. *Diagnostic pathology*, 10(1):36, 2015.

[74] Kazuaki Nakane, Yasunari Tsuchihashi, and Nariaki Matsuura. A simple mathematical model utilizing topological invariants for automatic detection of tumor areas in digital tissue images. In *Diagnostic pathology*, volume 8, page S27. BioMed Central, 2013.

[75] Arnout C Ruifrok, Dennis A Johnston, et al. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, 23(4):291–299, 2001.

[76] Talha Qaiser, Yee-Wah Tsang, David Epstein, and Nasir Rajpoot. Tumor segmentation in whole slide images using persistent homology and deep convolutional features. In *Annual Conference on Medical Image Understanding and Analysis*, pages 320–329. Springer, 2017.

[77] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[78] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[79] Graham Upton and Ian Cook. *Understanding statistics.* Oxford University Press, 1996.

[80] Talha Qaiser, Korsuk Sirinukunwattana, Kazuaki Nakane, Yee-Wah Tsang, David Epstein, and Nasir Rajpoot. Persistent homology for fast tumor segmentation in whole slide histology images. *Procedia Computer Science*, 90: 119–124, 2016.

[81] Riku Turkki, Nina Linder, Tanja Holopainen, Yinhai Wang, Anne Grote, Mikael Lundin, Kari Alitalo, and Johan Lundin. Assessment of tumour viability in human lung cancer xenografts with texture-based image analysis. *Journal of clinical pathology*, pages jclinpath–2015, 2015.

[82] Mitko Veta, Paul J Van Diest, Stefan M Willems, Haibo Wang, Anant Madabhushi, Angel Cruz-Roa, Fabio Gonzalez, Anders BL Larsen, Jacob S Vestergaard, Anders B Dahl, et al. Assessment of algorithms for mitosis detection in

breast cancer histopathology images. *Medical image analysis*, 20(1):237–248, 2015.

[83] Joshua D Webster, Eleanor R Simpson, Aleksandra M Michalowski, Shelley B Hoover, and R Mark Simpson. Quantifying histological features of cancer biospecimens for biobanking quality assurance using automated morphometric pattern recognition image analysis algorithms. *Journal of biomolecular techniques: JBT*, 22(3):108, 2011.

[84] Stanley Osher and James A Sethian. Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations. *Journal of computational physics*, 79(1):12–49, 1988.

[85] James A Sethian. A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences*, 93(4):1591–1595, 1996.

[86] David RJ Snead, Yee-Wah Tsang, Aisha Meskiri, Peter K Kimani, Richard Crossman, Nasir M Rajpoot, Elaine Blessing, Klaus Chen, Kishore Gopalakrishnan, Paul Matthews, et al. Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology*, 68(7):1063–1072, 2016.

[87] Hassan MH Kamel. Trends and challenges in pathology practice: choices and necessities. *Sultan Qaboos University medical journal*, 11(1):38, 2011.

[88] Ehsan Elhamifar, Guillermo Sapiro, and S Shankar Sastry. Dissimilarity-based sparse subset selection. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2182–2197, 2016.

[89] Ehsan Elhamifar, Guillermo Sapiro, and Rene Vidal. Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery. In *Advances in Neural Information Processing Systems*, pages 19–27, 2012.

[90] Kevin P Murphy, Yair Weiss, and Michael I Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 467–475. Morgan Kaufmann Publishers Inc., 1999.

[91] Peter W Hamilton, Peter Bankhead, Yinhai Wang, Ryan Hutchinson, Declan Kieran, Darragh G McArt, Jacqueline James, and Manuel Salto-Tellez. Digital pathology and image analysis in tissue biomarker research. *Methods*, 70(1): 59–73, 2014.

[92] Jia-Mei Chen, Ai-Ping Qu, Lin-Wei Wang, Jing-Ping Yuan, Fang Yang, Qing-Ming Xiang, Ninu Maskey, Gui-Fang Yang, Juan Liu, and Yan Li. New breast cancer prognostic factors identified by computer-aided image analysis of h&e stained histopathology images. *Scientific reports*, 5:10690, 2015.

[93] Mitko Veta, Yujing J Heng, Nikolas Stathonikos, Babak Ehteshami Bejnordi, Francisco Beca, Thomas Wollmann, Karl Rohr, Manan A Shah, Dayong Wang, Mikael Rousson, et al. Predicting breast tumor proliferation from whole-slide images: the tupac16 challenge. *Medical Image Analysis*, 2019.

[94] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. Bach: Grand challenge on breast cancer histology images. *arXiv preprint arXiv:1808.04277*, 2018.

[95] Anne Martel. Spie-aapm-nci breastpathq: Cancer cellularity challenge 2019, 2019. URL `http://spiechallenges.cloudapp.net/competitions/14`.

[96] Gail H Vance, Todd S Barry, Kenneth J Bloom, Patrick L Fitzgibbons, David G Hicks, Robert B Jenkins, Diane L Persons, Raymond R Tubbs, and M Elizabeth H Hammond. Genetic heterogeneity in her2 testing in breast cancer: panel summary and guidelines. *Archives of pathology & laboratory medicine*, 133(4):611–612, 2009.

[97] Clifford A Hudis. Trastuzumabmechanism of action and use in clinical practice. *New England journal of medicine*, 357(1):39–51, 2007.

[98] Antonio C Wolff, M Elizabeth H Hammond, Jared N Schwartz, and Daniel F Hayes. Human epidermal growth factor receptor 2 testing recommendation-in reply. *ARCHIVES OF PATHOLOGY & LABORATORY MEDICINE*, 131 (9):1331–1333, 2007.

[99] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[100] Vilppu J Tuominen, Teemu T Tolonen, and Jorma Isola. Immunomembrane: a publicly available web application for digital image analysis of her2 immunohistochemistry. *Histopathology*, 60(5):758–767, 2012.

[101] Marios A Gavrielides, Catherine Conway, Neil OFlaherty, Brandon D Gallas, and Stephen M Hewitt. Observer performance in the use of digital and optical microscopy for the interpretation of tissue-based biomarkers. *Analytical Cellular Pathology*, 2014, 2014.

[102] Anja Brügmann, Mikkel Eld, Giedrius Lelkaitis, Søren Nielsen, Michael Grunkin, Johan D Hansen, Niels T Foged, and Mogens Vyberg. Digital image analysis of membrane connectivity is a robust measure of her2 immunostains. *Breast cancer research and treatment*, 132(1):41–49, 2012.

[103] Alton Brad Farris, Cynthia Cohen, Thomas E Rogers, and Geoffrey H Smith. Whole slide imaging for analytical anatomic pathology and telepathology: practical applications today, promises, and perils. *Archives of pathology & laboratory medicine*, 141(4):542–550, 2017.

[104] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194, 2001.

[105] Chuan-Yu Chang, Ya-Chi Huang, and Chien-Chuan Ko. Automatic analysis of her-2/neu immunohistochemistry in breast cancer. In *2012 Third International Conference on Innovations in Bio-Inspired Computing and Applications*, pages 297–300. IEEE, 2012.

[106] Tomi Pitkäaho, Taina M Lehtimäki, John McDonald, and Thomas J Naughton. Classifying her2 breast cancer cell samples using deep learning. In *Proc. Irish Mach. Vis. Image Process. Conf.*, pages 1–104, 2016.

[107] Nicholas Trahearn, Yee Wah Tsang, Ian A Cree, David Snead, David Epstein, and Nasir Rajpoot. Simultaneous automatic scoring and co-registration of hormone receptors in tumor areas in whole slide images of breast cancer tissue slides. *Cytometry Part A*, 91(6):585–594, 2017.

[108] Adnan M Khan, Aisha F Mohammed, Shama A Al-Hajri, Hajer M Al Shamari, Uvais Qidwai, Imaad Mujeeb, and Nasir M Rajpoot. A novel system for scoring of hormone receptors in breast cancer histopathology slides. In *2nd Middle East Conference on Biomedical Engineering*, pages 155–158. IEEE, 2014.

[109] Erik Rodner, Marcel Simon, and Joachim Denzler. Deep bilinear features for her2 scoring in digital pathology. *Current Directions in Biomedical Engineering*, 3(2):811–814.

[110] Monjoy Saha and Chandan Chakraborty. Her2net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. *IEEE Transactions on Image Processing*, 27(5):2189–2200, 2018.

[111] Ramakrishnan Mukundan. Image features based on characteristic curves and local binary patterns for automated her2 scoring. *Journal of Imaging*, 4(2): 35, 2018.

[112] Juan C Caicedo and Svetlana Lazebnik. Active object localization with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2488–2496, 2015.

[113] Amir Alansary, Ozan Oktay, Yuanwei Li, Loic Le Folgoc, Benjamin Hou, Ghislain Vaillant, Konstantinos Kamnitsas, Athanasios Vlontzos, Ben Glocker, Bernhard Kainz, et al. Evaluating reinforcement learning agents for anatomical landmark detection. *Medical image analysis*, 2019.

[114] Gabriel Maicas, Gustavo Carneiro, Andrew P Bradley, Jacinto C Nascimento, and Ian Reid. Deep reinforcement learning for active breast lesion detection from dce-mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 665–673. Springer, 2017.

[115] Celia Ghedini Ralha, Hugo Wruck Schneider, Maria Emilia Machado Telles Walter, and Ana Lucia Cetertich Bazzan. Reinforcement learning method for bioagents. In *2010 Eleventh Brazilian Symposium on Neural Networks*, pages 109–114. IEEE, 2010.

[116] Fei Zhu, Quan Liu, Xiaofang Zhang, and Bairong Shen. Protein-protein interaction network constructing based on text mining and reinforcement learning with application to prostate cancer. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 46–51. IEEE, 2014.

[117] Xiao Liu, Tian Xia, Jiang Wang, Yi Yang, Feng Zhou, and Yuanqing Lin. Fully convolutional attention networks for fine-grained recognition. *arXiv preprint arXiv:1603.06765*, 2016.

[118] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015.

[119] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural

image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[120] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

[121] Marc'Aurelio Ranzato. On learning where to look. *arXiv preprint arXiv:1405.5488*, 2014.

[122] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[123] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[124] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

[125] Juan Lupiáñez. Inhibition of return. *Attention and time*, pages 17–34, 2010.

[126] Adnan Mujahid Khan, Nasir Rajpoot, Darren Treanor, and Derek Magee. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering*, 61(6):1729–1738, 2014.

[127] Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110. IEEE, 2009.

[128] Aïcha BenTaieb and Ghassan Hamarneh. Predicting cancer with a recurrent visual attention model for histopathology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 129–137. Springer, 2018.

[129] Abhinav Agarwalla, Muhammad Shaban, and Nasir M Rajpoot. Representation-aggregation networks for segmentation of multi-gigapixel histology images. *arXiv preprint arXiv:1707.08814*, 2017.

[130] Yann LeCun et al. Lenet-5, convolutional neural networks. *URL: http://yann. lecun. com/exdb/lenet*, 20, 2015.

[131] Hiroshi Sakamoto, Tetsuo Mashima, Atsuo Kizaki, Shingo Dan, Yuichi Hashimoto, Mikihiko Naito, and Takashi Tsuruo. Glyoxalase i is involved in resistance of human leukemia cells to antitumor agent-induced apoptosis. *Blood*, 95(10):3214–3218, 2000.

[132] Mingzhan Xue, Alaa Shafie, Talha Qaiser, Nasir M Rajpoot, Gregory Kaltsas, Sean James, Kishore Gopalakrishnan, Adrian Fisk, Georgios K Dimitriadis, Dimitris K Grammatopoulos, et al. Glyoxalase 1 copy number variation in patients with well differentiated gastro-entero-pancreatic neuroendocrine tumours (gep-net). *Oncotarget*, 8(44):76961, 2017.

[133] Tad T Brunyé, Ezgi Mercan, Donald L Weaver, and Joann G Elmore. Accuracy is in the eyes of the pathologist: The visual interpretive process and diagnostic accuracy with digital whole slide images. *Journal of biomedical informatics*, 66:171–179, 2017.

[134] Dilip B Nagarkar, Ezgi Mercan, Donald L Weaver, Tad T Brunyé, Patricia A Carney, Mara H Rendi, Andrew H Beck, Paul D Frederick, Linda G Shapiro, and Joann G Elmore. Region of interest identification and diagnostic agreement in breast pathology. *Modern Pathology*, 29(9):1004, 2016.

[135] Xia Fang, Bing Xiu, Zhizhang Yang, Weizhe Qiu, Long Zhang, Suxia Zhang, Yunjin Wu, Xuyou Zhu, Xue Chen, Suhong Xie, et al. The expression and clinical relevance of pd-1, pd-l1, and tp63 in patients with diffuse large b-cell lymphoma. *Medicine*, 96(15), 2017.

[136] G Lenz, G Wright, SS Dave, W Xiao, J Powell, H Zhao, W Xu, B Tan, N Goldschmidt, Javeed Iqbal, et al. Stromal gene signatures in large-b-cell lymphomas. *New England Journal of Medicine*, 359(22):2313–2323, 2008.

[137] Jyri M Moilanen, Nina Kokkonen, Stefanie Löffek, Juha P Väyrynen, Erkki Syväniemi, Tiina Hurskainen, Markus Mäkinen, Kai Klintrup, Jyrki Mäkelä, Raija Sormunen, et al. Collagen xvii expression correlates with the invasion and metastasis of colorectal cancer. *Human pathology*, 46(3):434–442, 2015.

[138] Xinliang Zhu, Jiawen Yao, Xin Luo, Guanghua Xiao, Yang Xie, Adi Gazdar, and Junzhou Huang. Lung cancer survival prediction from pathological images and genetic dataan integration study. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 1173–1176. IEEE, 2016.

[139] Sheng Wang, Jiawen Yao, Zheng Xu, and Junzhou Huang. Subtype cell detection with an accelerated deep convolution neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 640–648. Springer, 2016.

[140] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7234–7242, 2017.

[141] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 411–418. Springer, 2013.

[142] Korsuk Sirinukunwattana, Shan e Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans. Med. Imaging*, 35(5):1196–1206, 2016.

[143] Navid Alemi Koohababni, Mostafa Jahanifar, Ali Gooya, and Nasir Rajpoot. Nuclei detection using mixture density networks. In *International Workshop on Machine Learning in Medical Imaging*, pages 241–248. Springer, 2018.

[144] Arjun Chandra and Xin Yao. Evolving hybrid ensembles of learning machines for better generalisation. *Neurocomputing*, 69(7-9):686–700, 2006.

[145] Naonori Ueda and Ryohei Nakano. Generalization error of ensemble estimators. In *Proceedings of International Conference on Neural Networks (ICNN'96)*, volume 1, pages 90–95. IEEE, 1996.

[146] Yinyin Yuan, Henrik Failmezger, Oscar M Rueda, H Raza Ali, Stefan Gräf, Suet-Feung Chin, Roland F Schwarz, Christina Curtis, Mark J Dunning, Helen Bardwell, et al. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Science translational medicine*, 4(157): 157ra143–157ra143, 2012.

[147] Korsuk Sirinukunwattana, David RJ Snead, and Nasir M Rajpoot. A novel texture descriptor for detection of glandular structures in colon histology images. In *Medical Imaging 2015: Digital Pathology*, volume 9420, page 94200S. International Society for Optics and Photonics, 2015.

[148] Christian Frantz, Kathleen M Stewart, and Valerie M Weaver. The extracellular matrix at a glance. *J Cell Sci*, 123(24):4195–4200, 2010.

[149] Pengfei Lu, Ken Takai, Valerie M Weaver, and Zena Werb. Extracellular matrix degradation and remodeling in development and disease. *Cold Spring Harbor perspectives in biology*, 3(12):a005058, 2011.

[150] Peiwen Chen, Matilde Cescon, and Paolo Bonaldo. Collagen vi in cancer and its biological mechanisms. *Trends in molecular medicine*, 19(7):410–417, 2013.

[151] Min Fang, Jingping Yuan, Chunwei Peng, and Yan Li. Collagen as a double-edged sword in tumor progression. *Tumor Biology*, 35(4):2871–2882, 2014.

[152] Brandon Mattix, Thomas Moore, Olga Uvarov, Samuel Pollard, Lauren O'Donnell, Katelyn Park, Devante Horne, Jhilmil Dhulekar, Laura Reese, Duong Nguyen, et al. Effects of polymeric nanoparticle surface properties on interaction with brain tumor environment. *Nano Life*, 3(04):1343003, 2013.

[153] Long Shen, Honghao Li, Yuzhi Shi, Dekun Wang, Junbo Gong, Jing Xun, Sifan Zhou, Rong Xiang, and Xiaoyue Tan. M2 tumour-associated macrophages contribute to tumour progression via legumain remodelling the extracellular matrix in diffuse large b cell lymphoma. *Scientific reports*, 6:30347, 2016.

[154] Sylvie Ricard-Blum. The collagen family. *Cold Spring Harbor perspectives in biology*, 3(1):a004978, 2011.

[155] Lindsay M Morton, Sophia S Wang, Susan S Devesa, Patricia Hartge, Dennis D Weisenburger, and Martha S Linet. Lymphoma incidence patterns by who subtype in the united states, 1992-2001. *Blood*, 107(1):265–276, 2006.

[156] Jerry Alan Fails and Dan R Olsen Jr. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45. ACM, 2003.

[157] Henrik Holten-Rossing, Maj-Lis Møller Talman, Martin Kristensson, and Ben Vainer. Optimizing her2 assessment in breast cancer: application of automated image analysis. *Breast cancer research and treatment*, 152(2):367–375, 2015.

[158] Emma Shtivelman, Thomas Hensing, George R Simon, Phillip A Dennis, Gregory A Otterson, Raphael Bueno, and Ravi Salgia. Molecular pathways and therapeutic targets in lung cancer. *Oncotarget*, 5(6):1392, 2014.

[159] Clive R Taylor. Predictive biomarkers and companion diagnostics. the future of immunohistochemistry–in situ proteomics, or just a stain? *Applied immunohistochemistry & molecular morphology: AIMM/official publication of the Society for Applied Immunohistochemistry*, 22(8):555, 2014.

[160] Marius Ilie, Véronique Hofman, Manfred Dietel, Jean-Charles Soria, and Paul Hofman. Assessment of the pd-l1 status by immunohistochemistry: challenges and perspectives for therapeutic strategies in lung cancer patients. *Virchows Archiv*, 468(5):511–525, 2016.

[161] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11): 1238–1274, 2013.

[162] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

[163] JD Webster and RW Dunstan. Whole-slide imaging and automated image analysis: considerations and opportunities in the practice of pathology. *Veterinary pathology*, 51(1):211–223, 2014.

[164] Shazia Akbar, Lee B Jordan, Colin A Purdie, Alastair M Thompson, and Stephen J McKenna. Comparing computer-generated and pathologist-generated tumour segmentations for immunohistochemical scoring of breast tissue microarrays. *British journal of cancer*, 113(7):1075, 2015.

[165] Kingshuk Roy Choudhury, Kevin J Yagle, Paul E Swanson, Kenneth A Krohn, and Joseph G Rajendran. A robust automated measure of average antibody staining in immunohistochemistry images. *Journal of Histochemistry & Cytochemistry*, 58(2):95–107, 2010.

[166] Yutaka Hatanaka, Kaoru Hashizume, Kazuo Nitta, Tomoyuki Kato, Ichiro Itoh, and Yoichi Tani. Cytometrical image analysis for immunohistochemical hormone receptor status in breast carcinomas. *Pathology international*, 53 (10):693–699, 2003.

[167] Constantino Carlos Reyes-Aldasoro, Leigh J Williams, Simon Akerman, Chryso Kanthou, and Gillian M Tozer. An automatic algorithm for the segmentation and morphological analysis of microvessels in immunostained histological tumour sections. *Journal of Microscopy*, 242(3):262–278, 2011.

[168] Manohar Kuse, Yi-Fang Wang, Vinay Kalasannavar, Michael Khan, and Nasir Rajpoot. Local isotropic phase symmetry measure for detection of beta cells and lymphocytes. *Journal of pathology informatics*, 2, 2011.

[169] Hamid Raza Ali, M Irwin, L Morris, SJ Dawson, FM Blows, E Provenzano, B Mahler-Araujo, PD Pharoah, NA Walton, JD Brenton, et al. Astronomical algorithms for automated analysis of tissue protein expression in breast cancer. *British journal of cancer*, 108(3):602, 2013.

[170] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016.

[171] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015.

[172] Po-Whei Huang and Cheng-Hsiung Lee. Automatic classification for pathological prostate images based on fractal analysis. *IEEE transactions on medical imaging*, 28(7):1037–1050, 2009.

[173] Matthew D DiFranco, Gillian OHurley, Elaine W Kay, R William G Watson, and Padraig Cunningham. Ensemble based system for whole-slide prostate cancer probability mapping using color texture features. *Computerized medical imaging and graphics*, 35(7-8):629–645, 2011.

[174] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360, 2009.