DPhil Thesis

Using Deep Learning on Histology to Inform Colorectal Cancer Patient Treatment

Ruby Wood

Keble College

University of Oxford

Department of Engineering Science

supervised by Jens Rittscher and Tim Maughan

Summer 2024

Acknowledgements

Personal

Thank you to my family for always being my biggest cheerleaders, and thank you to Woody for all his support.

Institutional

Thank you to my supervisors Prof. Jens Rittscher and Prof. Tim Maughan for everything, to Prof. Viktor Koelzer for constantly guiding and helping me, and to all my other colleagues who have taken a supervisory role and advised me over the course of my DPhil.

This research was supported by the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1) and Cancer Research UK.

Abstract

Colorectal cancer is a serious health problem in the UK, with 11,500 patients diagnosed with rectal cancer each year. Neoadjuvant chemoradiotherapy treatment can be given to a patient prior to surgery to shrink the tumour, but roughly one third of patients will have a poor response to this treatment. In this work we develop deep learning approaches to predict how rectal cancer patients will respond to radiotherapy treatment, based off the routinely taken histology slides of pre-operative biopsies, in order to help clinicians make better personalised treatment decisions. We integrate the context of the imaging modality and the nature of the cancer into our approaches, first by including morphological and positional information into a Vision Transformer network. Secondly, we use tissue graph neural networks and multi-task learning with spatial and molecular endpoints to improve interpretability of our prediction model. Finally, we develop a domain adaptation method to help our proposed approach generalise to different cohorts of patients.

Contents

List of Abbreviations **viii**

1 Introduction **1**
1.1 Background . . . 2
1.1.1 Colorectal Cancer Treatment . . . 2
1.1.2 Pathology . . . 3
1.1.3 Response to Radiotherapy . . . 4
1.1.4 Data Overview . . . 5
1.2 Thesis Structure and Contributions . . . 5
1.2.1 Aims . . . 5
1.2.2 Chapters . . . 6
1.2.3 Publications . . . 7

2 Literature Review **10**
2.1 Colorectal Cancer Biomarkers . . . 11
2.1.1 Biology of Colorectal Cancer . . . 11
2.1.2 Predicting Treatment Response . . . 14
2.2 Predictions on CRC Histology . . . 15
2.3 Deep Learning Approaches on Histology . . . 19
2.3.1 Weakly Supervised Learning on WSIs . . . 19
2.3.2 Attention Models . . . 21
2.3.3 Graph Neural Networks . . . 26
2.3.4 Spatially Aware Approaches . . . 31
2.3.5 Foundation Models . . . 33
2.4 Interpretability . . . 34
2.5 Biomarkers in the Clinic . . . 36
2.5.1 Measuring Performance . . . 37
2.5.2 Stages of Biomarker Development . . . 39
2.5.3 Example of Biomarkers in the Clinic . . . 39
2.6 Summary . . . 41

3 Data **42**
3.1 Datasets . . . 43
3.1.1 Grampian Dataset . . . 44
3.1.2 Aristotle Dataset . . . 45
3.1.3 Salzburg Dataset . . . 46
3.2 Exploratory Data Analysis . . . 47
3.2.1 Image Analysis . . . 47
3.2.2 Outcome Analysis . . . 49
3.3 Slide Processing . . . 53
3.3.1 Slide Magnification . . . 53
3.3.2 Patching . . . 54

4 Predicting Treatment Response From Histology **56**
4.1 Introduction . . . 58
4.2 Methods . . . 60
4.2.1 Preprocessing . . . 60
4.2.2 Baseline Model . . . 61
4.2.3 Vision Transformers . . . 62
4.2.4 Patch Restoration Embedding ViT (PREViT) . . . 62
4.2.5 ClusterViT . . . 63
4.2.6 ClusterPREViT . . . 63
4.2.7 Clustering Approach . . . 63
4.2.8 Interpretability . . . 65
4.3 Experiment Results . . . 65
4.3.1 Model Hyperparameters . . . 66
4.3.2 Predicting RSS . . . 67
4.3.3 Predicting Response to Radiotherapy . . . 70
4.3.4 Visualising Clusters . . . 72
4.3.5 Attention Heatmaps . . . 74
4.3.6 Odds Ratios . . . 76
4.4 Discussion &Conclusion . . . 76

5 Interpretability with Molecular Traits and Spatial Organisation **81**
5.1 Introduction . . . 83
5.2 Methods . . . 84
5.2.1 Feature Extraction . . . 85
5.2.2 WSI Graph Design . . . 86
5.2.3 Data . . . 89
5.3 Experiments . . . 90
5.3.1 Superpixels . . . 90

5.3.2 Graph Connectivity . . . 93
5.3.3 Implementation . . . 96
5.4 Results . . . 99
5.4.1 Node-level Epithelium . . . 100
5.4.2 Visualisation . . . 101
5.4.3 Slide-level Epithelium Proportion . . . 102
5.4.4 Balancing Across Cohorts . . . 104
5.4.5 External Test Set . . . 107
5.4.6 Ablation Studies . . . 108
5.5 Analysis of Gradients . . . 109
5.6 Conclusion . . . 113

6 Domain Adaptation **114**
6.1 Introduction . . . 116
6.2 Related Work . . . 118
6.2.1 Unsupervised Domain Adaptation . . . 118
6.2.2 Histology Domain Adaptation . . . 120
6.3 Methods . . . 122
6.3.1 Source Model . . . 123
6.3.2 Clustering . . . 124
6.3.3 Cluster Triplet Loss . . . 125
6.4 Experiments . . . 128
6.4.1 Data . . . 128
6.4.2 Results . . . 129
6.4.3 Comparison with State-of-the-Art . . . 133
6.4.4 Ablation Studies . . . 134
6.5 Discussion . . . 137
6.5.1 Advantages . . . 137
6.5.2 Limitations . . . 137
6.5.3 Conclusion . . . 139

7 Conclusion **140**
7.1 Contributions . . . 141
7.2 Translation to Clinic . . . 141
7.2.1 Patient Perspectives . . . 141
7.2.2 Biomarker Validation . . . 143
7.3 Limitations . . . 144
7.3.1 Data . . . 144
7.3.2 Research . . . 144
7.4 Future Work . . . 145

7.4.1 Tumour Microenvironment 145
7.4.2 Temporal Modelling 145
7.4.3 Domain Adaptation 146
7.4.4 Patient Treatments 146

References **148**

List of Abbreviations

MSI	Microsatellite instability
CRC	Colorectal cancer
RT	Radiotherapy
CMS	Consensus Molecular Subtypes for colorectal cancer
H&E	Haematoxylin and eosin staining
RSS	33 gene expression RadioSensitive Signature
pCR	Pathological complete response
WSI	Whole slide image
TME	Tumour microenvironment
TGF-β	Transforming growth factor beta
TNM	Tumour, node, metastasis cancer staging system
AUC/AUROC	Area under the receiver operating characteristic curve
TILs	Tumour infiltrating lymphocytes
CNN	Convolutional neural network
ViT	Vision Transformer
GNN	Graph neural network
MIL	Multiple instance learning
RNN	Recurrent neural network
GIN	Graph isomorphism network
MLP	Multilayer perceptron
NICE	National Institute for Health and Care Excellence
CapRT	Radiotherapy with capecitabine
CR/NoCR	Complete response or no complete response to radiotherapy treatment

BCE . . . . . Binary cross entropy

MSE . . . . . Mean sqaured error

MAE . . . . . Mean absolute error

SLIC . . . . . Simple Linear Iterative Clustering superpixel algorithm

UDA . . . . . Unsupervised domain adaptation

1

Introduction

Contents

1.1 Background 2

1.1.1 Colorectal Cancer Treatment 2

1.1.2 Pathology 3

1.1.3 Response to Radiotherapy 4

1.1.4 Data Overview 5

1.2 Thesis Structure and Contributions 5

1.2.1 Aims 5

1.2.2 Chapters 6

1.2.3 Publications 7

Over the past decade the field of deep learning has developed rapidly, and advances in deep neural networks allow us to extract complex information from images.

These neural networks can be applied to images for segmentation, classification, or even prediction. The application of these deep learning methods in the field of cancer imaging has been very successful for determining clinically useful outcomes from medical images, including histology slides. Recent research has demonstrated the ability to predict a patient's response to therapy or indeed biologically relevant information from morphological features extracted from images taken from standard pathology slides. For example, determining the presence of microsatellite instability (MSI) in colorectal cancer (CRC) has been shown to help with treatment decisions, and research has shown it is possible to predict MSI presence using deep learning on histology images [1]. Sirinukunwattana *et al.* [2] have proven it is possible to predict the CRC molecular classification system, CMS, from histology using deep learning approaches. This work paves the way for my research, to predict response to treatment for CRC patients from histology slides.

1.1 Background

1.1.1 Colorectal Cancer Treatment

CRC is defined as cancer that starts in the colon or the rectum [3], and is a serious health problem in the UK and around the world. In the UK, it is the 4th most common cancer [4] and is the 8th most common cancer internationally [5]. Approximately 11,500 patients are diagnosed with rectal cancer in the UK each year [6].

Surgery is the most common form of treatment for CRC, but other common forms of treatment include chemotherapy and radiotherapy (RT), which can be used in combination with each other in chemoradiotherapy treatments, and can be used both before and after surgery (in neoadjuvant and adjuvant settings respectively). The most common chemotherapy drugs for CRC are fluorouracil, capecitabine, oxaliplatin and irinotecan [7].

Neoadjuvant treatment in the form of RT is used more in the treatment of rectal cancer than colon cancer [7], and is commonly given to patients with locally advanced

rectal cancer to shrink the size of the tumour, either before surgery to prevent further tumour growth or instead of surgery if the cancer is small enough [8]. Recent evidence suggests that 10-20% of patients will have a complete pathological response to neoadjuvant therapy and can therefore avoid surgery altogether [9, 10]. However, one third of rectal cancer patients have a poor response to conventional RT treatment [4, 11], and subsequently more than 10% of patients experience severe long-term complications requiring hospitalization following treatment [5]. In an assessment of the watch and wait strategy for rectal cancer patients who had a complete response to neoadjuvant RT (prior to surgery), Smith *et al.* summarised that though this strategy could work well for many patients, better patient stratification approaches are needed for optimal treatment decisions [12]. Determining patient response to RT with a is personalized approach is therefore critical to avoid overtreatment.

1.1.2 Pathology

Pathology plays a crucial role in confirming the cancer diagnosis and subsequently informing the patient's treatment decision. CRC patients typically undergo a tissue biopsy to remove the cancerous tissue from the colon. The aim of this surgery is to remove as much cancerous tumour tissue as possible. The excised tissue and samples of the tumour margin will be analysed by a pathologist. Prior to analysis any tissue is sliced and stained with haematoxylin and eosin (H&E), which is a commonly used stain for cancer biopsies in histology. The haematoxylin detects cell nuclei, staining them purple, and the eosin detects cytoplasm and muscle fibres among other things, staining them pink [13] [14]. The stained slices are put onto glass slides and scanned to create a digital image for each slice. Pathologists can then study these slides using computer software and can contribute their findings towards the patient's treatment decision. The patient's treatment is decided in a Multidisciplinary Team (MDT) meeting, where clinicians from multiple fields come together to contribute their expertise. The pathologist would therefore contribute their interpretation gained from the biopsy slides.

1.1.3 Response to Radiotherapy

The overall goal of this work is to provide a tool to clinicians that can accurately predict an individual CRC patient's response to treatment, and also explain the prediction in context of their existing specialist knowledge. Using deep learning to identify morphological features from the cancerous tissue in the histology slides, we aim to predict a patient's response to treatment, focusing on response to RT in particular. This prediction could be used to stratify patients in a trial or be integrated into a patient's portfolio in a medical MDT meeting to help determine the best treatment option for an individual patient.

The RadioSensitive Signature (RSS) gene expression signature is a genetic set developed by Domingo *et al.*, calculated from 33 gene expressions derived from gene sequencing performed on the CRC biopsies [15]. This gene score has shown promise in predicting a patient's pathological complete response (pCR) to RT, achieving 84% accuracy in a machine learning model [15]. They find that the tumour of cancer patients who had a pCR had certain characteristics in the stromal and immune cell compartments, which they propose could potentially be identified with their RSS biomarker.

In this research we will try to predict this RSS gene score from the histology images, as a way of quantifying a patient's response to RT. This predictor should mimic RSS but as it is based on standard H&E images it could be deployed much faster, since this prediction model would be able to determine the RSS gene score purely from images of the biopsies, whereas to calculate the RSS gene score from the biopsy tissue via gene sequencing would take a significant amount of time. Predicting a patient's response to RT from digital biopsy images is also much cheaper and more efficient than trying to predict it from gene assays, so we may be able to forgo the RSS score by predicting a patient's response to RT directly.

1.1.4 Data Overview

The data available for this project is H&E-stained histology slides of CRC biopsies from three separate trial cohorts, Grampian, Aristotle and Salzburg. For most of these slides we have additional data on the patient's response to RT, as well as the RSS gene score, derived from gene expressions from the biopsy tissue. When the biopsies are sliced it introduces the unwanted side effect of tissue stretching, where the tissue is deformed from its original shape, potentially changing the appearance of cells in the WSIs. Furthermore, when tissue slices are stained in different hospital labs, it can introduce a batch effect because each hospital may have a slightly different workflow.

The digital images of the biopsies or surgical resections, called Whole Slide Images (WSIs), are scanned at a very high resolution so that the resulting WSIs are massive files. Though the tissue itself may only be 15-20mm in diameter, the resulting WSI could be 15GB in size [16]. The medical imaging community has considered various approaches to dealing with these large images in the modelling process, which we discuss in the literature review in Chapter 2.

Additionally, these datasets present their own challenges since the outcomes are highly imbalanced, because most patients unfortunately do not respond fully to the treatment reviewed here, and the number of patients for which we have data is limited.

1.2 Thesis Structure and Contributions

1.2.1 Aims

The research goals for this thesis can be summarised into three aims, consistently focusing on how we use the context of the H&E slides to enhance predictions and interpretation of results. The first aim is to develop a deep learning model that can predict a patient's response to RT from their digital histology images, exploring whether our model can predict either the RSS gene score or directly the response to

RT. The second aim is to provide biologically useful interpretations for this model prediction to provide context and build trust with the clinicians who will be using this model in practice. The third aim is to consider the steps for how to translate this model into the clinical setting, to make use of the model predictions to help determine a patient's treatment. This includes validating the model on external datasets to evaluate its applicability on various patient cohorts, and developing methods to help the model adapt to new domains.

1.2.2 Chapters

- **Chapter 1** introduces the clinical problem and some background, and then provides some details on the contributions from this work.
- **Chapter 2** provides a comprehensive literature review of relevant research to this thesis. It covers biology of CRC and how biomarkers from histology slides can be used to make prognostic or diagnostic predictions. We then introduce popular deep learning approaches for histology, from weakly supervised multiple instance learning to attention models and graph neural networks. We present relevant research on how deep learning biomarkers can be translated to the clinic, stressing the importance of interpretability and understanding from clinicians, and review some biomarkers which are (or are close to being) used in the clinical setting.
- **Chapter 3** provides an overview of the data we use in this project. We provide detailed descriptions of the three cohorts of CRC patients and their corresponding treatments, as well as demonstrating features of the digital H&E-stained biopsy slides and some preliminary exploratory data analysis.
- **Chapter 4** presents our first attempt to predict a patient's response to RT and RSS gene score from the histology slides. We develop novel variations on the Vision Transformer model which both preserve spatial information from the layout of the tumour tissue in the WSI, and incorporate a morphological prior via a clustering to allow the model to improve its predictions.

- **Chapter 5** develops a more intuitive and interpretable method for predicting response to RT from histology images. We design graphs to represent the WSIs, using segmented tissue regions instead of square patches for a more natural representation of the tumour. We implement a graph neural network with a multi-task learning approach to incorporate both molecular traits and the spatial organisation of the tumour, providing visualisations which enhance interpretability of our primary prediction of therapy response. Additionally, we analyse the output graph predictions to observe changes in predicted features across tissue boundaries.

- **Chapter 6** builds on the work in the previous chapter, tackling the issue of model generalisation to unseen external datasets. We develop an unsupervised domain adaptation technique to adapt our pre-trained graph model to a new dataset, only using a lightweight representation of the source domain to do so.

- **Chapter 7** summarises our contributions from this work, and explores how our work meets criteria regarding translatability to the clinical setting, including limitations of the data and the research. We discuss various interesting avenues for future work, stemming from this research.

1.2.3 Publications

Chapter 4

Ruby Wood, Korsuk Sirinukunwattana, Enric Domingo, Alexander Sauer, Maxime W. Lafarge, Viktor H. Koelzer, Timothy S. Maughan & Jens Rittscher. Enhancing Local Context of Histology Features in Vision Transformers. *In: Artificial Intelligence over Infrared Images for Medical Applications and Medical Image Assisted Biomarker Discovery. Cham: Springer Nature Switzerland, 2022, pp. 154–163.* Selected for oral presentation.

Chapter 5

Ruby Wood, Enric Domingo, Korsuk Sirinukunwattana, Maxime W. Lafarge, Viktor H. Koelzer, Timothy S. Maughan & Jens Rittscher. Joint Prediction of Response to Therapy, Molecular Traits, and Spatial Organisation in Colorectal Cancer Biopsies. *In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Ed. by Hayit Greenspan* et al. *Cham: Springer Nature Switzerland, 2023, pp. 758–767.* Selected for oral presentation.

Chapter 6

Ruby Wood, Enric Domingo, Viktor Hendrik Koelzer, Timothy S. Maughan & Jens Rittscher. Cluster Triplet Loss for Unsupervised Domain Adaptation on Histology Images. *In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2024, pp. 5122–5131.* Selected for oral presentation.

Other publications

Yang Hu, Korsuk Sirinukunwattana, Kezia Gaitskell, **Ruby Wood**, Clare Verrill & Jens Rittscher. Predicting molecular traits from tissue morphology through self-interactive multi-instance learning. *In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 130-139. Cham: Springer Nature Switzerland, 2022.*

Maxime W. Lafarge, Enric Domingo, Korsuk Sirinukunwattana, **Ruby Wood**, Leslie Samuel, Graeme Murray, Susan D. Richman et al. Image-based consensus molecular subtyping in rectal cancer biopsies and response to neoadjuvant chemoradiotherapy. *NPJ precision oncology 8, no. 1 (2024): 89.*

Hosuk Ryou, Korsuk Sirinukunwattana, **Ruby Wood**, Alan Aberdeen, Jens Rittscher, Olga Weinberg, Robert Hasserjian et al. Quantitative Analysis of Bone

Marrow Features Highlights Heterogeneity in Myelofibrosis Patients Treated with Zinpentraxin Alfa in a Phase II Clinical Study. *Blood 142 (2023): 4558.*

2

Literature Review

Contents

2.1 Colorectal Cancer Biomarkers 11

2.1.1 Biology of Colorectal Cancer 11

2.1.2 Predicting Treatment Response 14

2.2 Predictions on CRC Histology 15

2.3 Deep Learning Approaches on Histology 19

2.3.1 Weakly Supervised Learning on WSIs 19

2.3.2 Attention Models 21

2.3.3 Graph Neural Networks 26

2.3.4 Spatially Aware Approaches 31

2.3.5 Foundation Models 33

2.4 Interpretability 34

2.5 Biomarkers in the Clinic 36

2.5.1 Measuring Performance 37
2.5.2 Stages of Biomarker Development 39
2.5.3 Example of Biomarkers in the Clinic 39
2.6 Summary 41

In this literature review we will first introduce the key biomarkers used in colorectal cancer (CRC) prognosis, showcasing those which have a positive association with response to treatment, as well as a negative response. We will then demonstrate research on some of these prognostic features which can be found in histology H&E stained biopsy slides, demonstrating the use of this stain when looking for features in the images. Recent research has shown that it is possible to assess or enhance some of these CRC pathological features using deep learning techniques, which we explore below, as well as work looking specifically at predictions of a patient's response to treatment. Finally, we also consider deep learning approaches on other forms of cancer, and explore the technical deep learning approaches that are used.

2.1 Colorectal Cancer Biomarkers

In this section we review some background on the underlying biology of CRC, and introduce some established molecular and genetic biomarkers in the field. We review approaches to predict CRC patient treatment responses from various data modalities. Finally, we consider other important works which have applied deep learning on histology slides of CRC.

2.1.1 Biology of Colorectal Cancer

TGF-β

In cancer, transforming growth factor beta (TGF-β) is a protein which can both suppress (in earlier stages) and induce (in later stages) tumour growth [17]. Normal cells, in particular epithelial cells, need to avoid this growth inhibitory signal

to become cancerous [18]. Generally speaking though, activation of TGF-β in the tumour microenvironment (TME) supports survival of cancer cells in CRC [19], where the TME consists of the stromal tissue and extracellular matrix and cancer cells themselves surrounding the tumour. Infiltrating immune cells, such as macrophages and lymphocytes, are also found in the TME, and demonstrate prognostic value [20]. Zaborowski *et al.* claim that the spatial organisation of these cells may also be associated with treatment responses [21], and it's been hypothesised that the interactions between the cancer cells and their TME could affect the progression of the tumour [20].

Microsatellite instability

Some colorectal cancers are induced by a defective DNA mismatch repair (MMR) which leads to high microsatellite instability (MSI) [21]. Microsatellites are small repeating sequences of DNA, and MSI refers to mutations in these sequences [22]. MSI tumours are not as likely to spread to other parts of the body, and high MSI is correlated with lymphocytic infiltration, which can lead to improved survival [21]. However, Tran *et al.* showed that in CRC that has metastasised, MSI actually associates with poorer survival, though this is due to its association with the BRAF mutation [23].

p53

The p53 protein, also called the TP53 gene in humans, has multiple functions and can repair cells, stop them multiplying or even kill them [18]. It can act as a tumour suppressor in cancer [24], and mutations of this gene are known to drive cancer progression [25]. Mutations of TP53 are often found in CRC [24], and have been associated with resistance to cancer treatment [25].

Consensus Molecular Subtypes (CMS)

The consensus molecular subtypes (CMS) classification system derived from gene expressions [26] has been developed to provide biological insight into metastatic

CRC. The classification system consists of four CMS classes of CRC, each with distinguishing features. For example, the CMS1 class is associated with high MSI and cytotoxic lymphocytes, and the CMS4 class is associated with high TGF-β activation [27]. The CMS classes also have distinct morphological features that can be interpreted from histology images of the cancer. Tran *et al.* [23] found that high MSI is associated with poor survival in CRC for later stages of the disease, though this is partly due to its association with the BRAF gene mutation, which is also linked to poor survival. Itatani *et al.* [19] found that TGF-β activation in the TME, as found in CMS4, can result in poorer prognosis.

Various studies have investigated the link between CMS and patient outcomes, suggesting that patients with tumour classified as CMS4, which features stromal invasion [26] and shows significantly higher stroma content [28], have worse survival rates compared to the other CMS classes [10].

Prognostic pathology features

In the past couple of decades there have been developments in the research on pathology features that can be identified as key indicators of CRC. The ABC of Colorectal Cancer book [29] defines prognostic features in CRC, separating poor and good prognostic features. In the TNM (tumour, node, metastasis) cancer staging system, a pathological tumour stage 4 (pT4) is a very poor prognostic feature in CRC [30]. Histological features of pT4 include tumour perforation, breaching of serosal surface and invasion to adjacent organs [29], [31]. Other poor prognostic features in CRC include poor tumour differentiation, recovery of a small number of lymph nodes, extramural vascular invasion and an infiltrative tumour edge and tumour budding [29]. Good prognostic features in CRC include MSI, dense intra-tumoural lymphocytes and a pushing tumour edge [29].

The following studies provide a first indication that biological features of the primary tumour can be captured in biopsy fragments and allow prediction of clinical disease behaviour. Anitei *et al.* propose that tumour immune infiltrate evaluated with their Immunoscore method is a useful prognostic marker for rectal

cancer patients prior to surgery [32]. Furthermore, Jones *et al.* [33] find that increased stromal content in early rectal cancer, as is prominent in CMS4, is a predictor for an increased risk of recurrence. These studies, however, focus on single features rather than visual assessment, and are based on limited cohorts with heterogeneous treatment conditions.

The spatial organisation of the cancerous tissue has been identified as a biomarker for aggressiveness or recurrence [34], and Qi *et al.* [28] found that the features they developed representing spatial organisation reflected characteristics of the four CMS classes. Interactions between the epithelial tissue (cellular tissue lining) and other prevalent tissue types in the TME are also indicators of prognosis [28], since progression of CRC is dependent on both the epithelial and stromal tissues [35].

2.1.2 Predicting Treatment Response

Alkan *et al.* [9] published a review of several papers containing findings of biomarkers to predict a patient's response to treatment, specifically for rectal cancer. They review multiple papers, some of which look at immunological markers, such as TGF-β1, and some of which consider molecular genetic markers, such as MSI. From this review it is clear that there is a lot of research being done to find biomarkers to predict response to treatment for rectal cancer tumours, though only two papers referenced in this review focus on imaging data modalities. Koelzer *et al.* found that CD8 T-cell infiltration is associated with positive outcomes, and as a result they recommend assessing CD8i infiltration in biopsies taken prior to surgery for prognostic value [36]. The CBLL1 gene could also be a useful biomarker in cancers classified as CMS2, since these patients show worse survival rates [37]. However, only a few genetic mutations are currently used for patient treatment [37].

In a review of prognostic biomarkers in CRC, Zaborowski *et al.* observed that the spatial organisation of the immune cells in the TME could be predictive of treatment response, such as the proportion of lymphocytes in the tumour tissue [21]. In fact, the spatial variability found in tumours can cause variability in the flow and therefore in the effect of drugs in this system [38].

Other imaging modalities such as CT scans can be used to measure the tumour volume, and Laleh *et al.* tried to mathematically model treatment responses via tumour growth and decay using these observations in partial differential equations. However, the models they used have a large number of hand-selected parameters which require at least as many observations to fit such a model. Though their mathematical models work reasonably well for predicting early treatment response, they observe that in the long term the early treatment response is not highly correlated with the final response, suggesting the need for more complex models to better predict treatment response [39].

MRI scans have been used to predict treatment response in rectal cancer, training a Siamese deep learning network to predict neoadjuvant chemoradiotherapy response on locally advanced rectal cancer [40, 41]. While achieving good results on the validation data, the model had trouble generalising to perform well on an external test cohort, with AUC (area under the receiver operating characteristic curve) of 0.54-0.6, highlighting the importance of the data quality and the difficulty of generalising to images from multiple centres [40].

Concerning pathology, there is currently limited research on predicting treatment response for CRC, though tumour budding, when observed in pre-treatment histology slides of rectal cancer, has been identified a negative predictor of response to radiotherapy [42]. Furthermore, other research has proven it is possible to directly predict a patient's response to treatment from the digital biopsy slide histology images. Although Zhang *et al.* have demonstrated the ability to predict the chemoradiotherapy response in locally advanced rectal cancer from images of H&E stained biopsies, they only use standard machine learning approaches on hand-crafted image features [43], and do not provide contextual interpretations.

2.2 Predictions on CRC Histology

Recent research has demonstrated the application of deep learning to predict a patient's response to therapy and to predict biologically relevant information from

morphological features extracted from standard histology slides.

Predicting microsatellite instability

MSI is one of the key markers informing the treatment decision in CRC [44, 45]. Research has shown it is possible to predict MSI status in CRC using deep learning from standard H&E stained slide images [1, 46, 47]. Echle *et al.* [47] successfully modify a deep learning network to detect MSI from H&E images of CRC, though they did find that they had greater success in single-centre cohorts of data, and had a harder time generalising the model to other patient cohorts. Bilal *et al.* [46] develop a deep learning pipeline to also predict the status of MSI from CRC histology, as well as some other key molecular pathways and mutations such as chromosomal instability, CpG island methylator phenotype and tumour-infiltrating lymphocytes (TILs). They use a ResNet-18 [48] convolutional neural network (CNN) as the baseline model, to first simply separate the tumour and non-tumour tiles. As well as MSI, they predict some other key molecular pathways and mutations such as chromosomal instability, CpG island methylator phenotype and TILs from CRC histology slides [46].

Guo *et al.* use two Shifted Window Hierarchical Vision Transformer (Swin-T) models, one to detect tumour tissue, and a second to classify six CRC biomarkers: hypermutation, MSI status, chromosomal instability, CpG island methylator phenotype, BRAF, and TP53 mutations [49]. They use the CRC histology slides in the publicly available dataset The Cancer Genome Atlas (TCGA) as well as a private CRC dataset from Australia, which requires some preprocessing. When using multiple cohorts of data, it can help to use colour normalization to reduce bias from the different staining methods across cohorts, so they applied Macenko's method here. Additionally, they used Canny edge detection to help remove background or blurry tissue, reducing noise in the dataset. From the prediction of MSI status, they discussed how this should be used for a diagnostic outcome, and found that a cutoff of 0.16 could provide 95% sensitivity for predicting MSI-H in patients, an important biomarker used in CRC treatment decisions.

imCMS

Sirinukunwattana *et al.* [2] developed an image-based deep learning tool for predicting the CMS classifications of CRC, a stratification system with clear biological interpretation [26], from H&E slides, removing the need to determine the CMS classification of the cancer using expensive gene expression profiling. Their research is performed on histology slides of rectal cancer biopsies, testing on the Grampian dataset which is a private dataset also available for our research (details given in Chapter 3). They train an Inception V3 CNN model [50] using a domain adversarial approach, which encourages the model to work well across multiple cohorts. They also study the prognostic associations of both the imCMS predictions and the CMS classifications. They find that both CMS4 and imCMS4 indicate worse prognosis, while CMS1 associates with adverse outcome, both with statistically significant trends. The imCMS1 classification has a statistically significant trend towards worse overall survival when compared with the CMS1 classification.

A more recent version of the imCMS model calculates odds ratios and confidence intervals to find associations between imCMS and pathological complete response (pCR) to RT directly. They found that the imCMS1 class is significantly positively associated with pCR, and that imCMS4 is positively associated with the inverse, the outcome where patients do not have a pCR [51].

Tsai *et al.* also tried predicting CMS from histology images of CRC, and found the AUC for the CMS prediction (between CMS2 and CMS4 only) improved to 0.75 when incorporating information on whether the cancer was found in the colon or rectum [52]. Their pipeline combined a ResNet-50 to extract features, K-Means to cluster these features into ten tissue types, and Vision Transformers (ViTs) [53] to find attention-based information feature vectors for each cluster. They found that presence of stroma and mucus were strong indicators of both CMS2 and CMS4, and that these regions are crucial for predicting survival outcomes, consistent with previously established prognostic observations of tumour invasiveness and tumour-stroma interactions.

Other molecular markers

As has been mentioned already, the spatial organisation of cells in the TME can be predictive of prognosis, and these molecular correlations can be explored using deep learning. Specifically, CNNs can be used to classify tissue regions as containing TILs, which can be refined using an iterative approach where at the end of each cycle the segmentations are reviewed by pathologists and the model is refined accordingly. Due to the visual similarities between the nuclei of TILs and necrosis, a second CNN can be used to segment necrosis regions, which can help to avoid false positives in the final TIL segmentation. For medical interpretation of the classified TIL regions, the authors used affinity propagation to find a spatially coherent clustering of the regions, which revealed a spatial structure in the TIL patterns and was linked to survival outcomes using Cox regression models [54]. The data used here was histology slides of 13 different cancer types, including rectal cancer.

Focusing only on CRC histology slides, Kiehl *et al.* used deep learning to directly predict the lymph node status from the images, incorporating clinical data to achieve 0.74 AUC on an internal test set [55]. In their work on classifying nine CRC tissue classes from histopathological WSIs with the popular VGG-19 model, Martínez-Fernandez *et al.* find that the learning rate is the most important hyperparameter to optimise during training for best results, when classifying epithelial vs stromal tissue for example [56]. Other research on immune cell populations in CRC has been motivated by the fact that while high T-cell density has been established as a prognostic factor, the impact of tumour-associated plasma cells, eosinophils and neutrophils is more unknown. Väyrynen *et al.* identify these and more areas of CRC histology slides, and explored spatial patterns of immune cell infiltration using GTumor, a function that evaluates the likelihoods of other immune cells within a certain radius of tumour cells [57]. High densities of stromal lymphocytes and eosinophils were associated with better survival outcomes via a multivariable Cox proportional hazards regression.

Skrede *et al.* [58] use deep learning to also predict survival from CRC histology images. They use an ensemble approach, training multiple deep learning models on

the images at two different magnifications (40x and 10x), then take an average of the model predictions and define patient survival outcome based on a learned threshold.

2.3 Deep Learning Approaches on Histology

Thus far, the approaches introduced for prediction problems in CRC have been fairly traditional, for example using hand-crafted image features, clustering, statistical models or fundamental deep learning techniques such as classical CNNs. More recently in the field, more complex vision deep learning models have been developed and their use has been demonstrated on histology slides. In this section we explore more complex model architectures such as attention-based models and graph neural networks (GNNs), and discuss how these general computer vision models can and should be applied onto the large histology slides. We move away from CRC for a more general look at deep learning approaches in cancer histopathology.

2.3.1 Weakly Supervised Learning on WSIs

As briefly mentioned in the Introduction in Chapter 1, histology slides are scanned as digital WSIs, which are massive images. Applying deep learning to WSIs of histology samples is not computationally straightforward owing to their large pixel numbers. In fact, they are generally too large to be processed all at once by a GPU, and so they are split into smaller images to be processed separately, known as patches or tiles. To label each patch for use in the model learning process, we use the label from the whole slide and apply it to each patch within that image. This is known as weakly supervised learning, since we have weak inherited labels for each patch. A deep learning model will learn to predict an outcome for each patch, so that each patch is assigned its own label from the parent slide and subsequently receives its own output. There are different methods to aggregate these patch predictions into a slide-level prediction for the WSI, known as multiple instance learning (MIL).

One of the simplest methods of aggregation is to take the mean over all the patch predictions in one slide, as done in [55], known as mean-pooling. Another

straightforward approach is max-pooling, where the maximum patch prediction in a slide is selected for the slide-level prediction [59]; however, this can be influenced heavily by just one wrong false positive prediction [60]. [61] predict a patient's complete response to neoadjuvant chemotherapy on breast cancer histology images. They use deep learning to first segment the tumour epithelium from the digital images, and then use contrastive learning to score the tiles with regards to treatment response. Only the top-k tiles are used for the final overall slide prediction using a mean-pooling approach.

Campanella *et al.* explore different patch aggregation methods on histology WSIs of various cancers at different magnifications [59]. They find that max-pooling, using the maximum function, is not a robust method of aggregating patch predictions, but they have success using a recurrent neural network (RNN) to combine the predictions. Similarly, Kanavati *et al.* also try both max-pooling and an RNN to aggregate their patch predictions from histology WSIs to classify the cancer subtype [62]. As in [59], they found that the max-pooling approach gives more false positives in some cases, whereas the RNN is a more successful method. Similar research has followed on from the work of Campanella *et al.*, using the max-pooling and RNN techniques to aggregate WSI patch predictions on colon tumours [60]. All of these examples use a CNN as their deep learning model to generate the patch predictions.

However, the downside of using RNNs for this patch aggregation is that an RNN assumes a flat input sequence, which does not capture the 2-dimensional nature of the images. Furthermore, RNN models typically have difficulty capturing longer range information, so would not be able to simultaneously capture information spatially distant on the image [63]. [64] get around the 1-dimensional nature of RNNs and 2-dimensional input problem by using a 2D long short-term memory model (LSTM), a type of RNN, to better capture the spatial context of the patch within the WSI. They use this LSTM to aggregate patch predictions from a VGG-16 CNN, in order to predict the probability of 5-year survival for CRC patients. Despite this, the RNN approach to patch prediction aggregation still has the issues of not capturing longer range information in the sequence.

Using patches at multiple scales requires further consideration of aggregation techniques in another dimension. Hong *et al.* use patches at 2.5x, 5x and 10x magnification to predict endometrial cancer subtypes. They started with separate CNN models, called branches, for each magnification input, ultimately concatenating the branches and applying global average pooling to aggregate, before passing that through a final fully connected layer. The combined model was trained end to end, backpropagating the loss gradients through the branches, which the authors claim preserves the spatial information in the image [65].

2.3.2 Attention Models

To date, most deep learning approaches utilise this tile-based approach, aggregating independent tile predictions to obtain a slide-level prediction. We contend that there is a benefit to integrating local context information as early as possible in the learning process. Such research has explored using attention models, such as the ViT [53], to get a slide-level prediction from WSIs. Attention models work by assigning importance weights to different patches in the image, and use those weights to get an overall slide prediction.

Lu *et al.* [66] develop the CLAM model (Clustering-Constrained-Attention Multiple-instance learning), combining clustering and attention models to aggregate their patch predictions. As other researchers have done, they train a CNN to extract features for the image patches. They then apply the attention-based pooling function developed by Ilse *et al.* [67] to aggregate patch predictions, using clustering on the patch features to constrain the feature space and refine predictions. Sharma *et al.* use the K-Means clustering technique to cluster and sample patches extracted from a CNN, on which they apply the attention-based pooling function from [67] for a slide-level prediction [68]. They cluster across each slide as opposed to across all slides in the dataset, because they find that the global clustering approach is more prone to cluster patches based on visual cues such as differences in staining, rather than morphological tissue features.

After criticising the RNN method of patch aggregation, [63] propose to use a ViT to aggregate its model predictions, applying the Transformer in histopathological image analysis for the first time. They use an EfficientNet B0 (pre-trained on the ImageNet database) to extract features from the patches, and then use the ViT model to predict the slide-level outcome from these features.

Kwon *et al.* use attention on top of a ResNet-101 classifier for CRC histology, and choose to train using a localization loss to improve the accuracy of the resulting attention maps [69]. Broad *et al.* develop their own attention-inspired approach to predict the tumour-stroma ratio on CRC pathology [70]. Their algorithm iteratively resamples informative patches in the WSI, initialising randomly and sparsely but gradually sampling from higher density regions where tumour patches are being classified against normal epithelium. Nahhas *et al.* use an attention-based MIL method to argue that using a regression-based deep learning approach can outperform classification-based methods when predicting continuous biomarkers and prognostic factors from CRC pathology slides [71].

Tackling the problem of existing patch-based MIL approaches not considering the heterogeneity of the tumour and the fact that important patterns could extend beyond the boundaries of a single patch, Fourkioti *et al.* develop a new MIL framework for contextual WSI classification. They propose Context-Aware Multiple Instance Learning (CAMIL), where they make use of a matrix mask which acts as a spatial prior indicating which pairs of patches are adjacent to one another, and use this to incorporate attention scores of neighbouring patches [72]. This concept of recognising the importance of the spatial structure and surrounding tissue beyond an individual patch is a common developing theme in histology, and one which we will explore further both in this literature review and in the research in this project.

Vision Transformers

The ViT model was developed to adapt the original Transformer model from the field of natural language processing, considering sequences of text as input to the model, to the field of computer vision [53, 73], considering images. A distinguishing feature

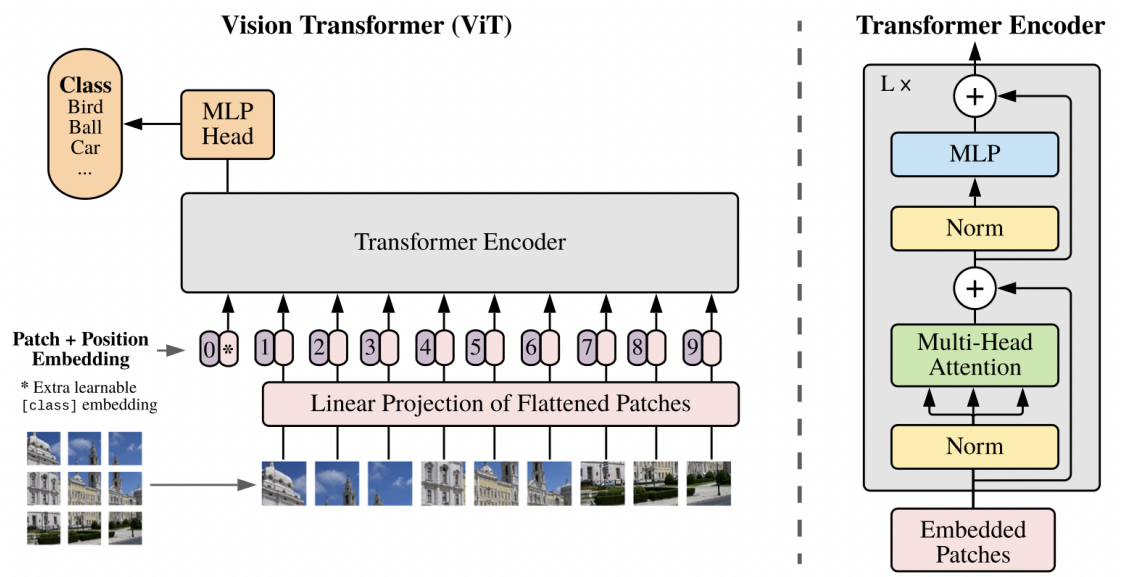


Figure 2.1: Overview of the Vision Transformer model taken from the original paper [53]. The Transformer Encoder used is the same one used in the original Transformer paper [73].

of the Transformer framework is its self-attention modules, which use the so-called query, key and value (Q, K and V) approach to calculate attention weights based on multiple linear transformations of the same input embeddings, hence the attention being on the 'self'. The scaled dot product attention can be formally defined as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.1}$$

where Q, K and V are different linear transformations on the input patch embeddings, and d_k is the dimension of K [73].

The ViT model uses many of these self-attention modules in parallel, but prior to the input entering this main part of the model, two tokens are added. One token is a position embedding, indicating the position of the patch in the input sequence, as we discuss further below, and the other token is a classification token, which is only used for the final prediction. An overview of the model architecture can be viewed in Figure 2.1.

Compared to CNNs, ViTs have less inductive bias. While CNNs assume equal kernel weights can be applied across an input image i.e. translation equivariance,

the self-attention layers in a ViT are global and no similar assumptions are required.

ViTs have been used to aggregate predictions across different scales, taking different size patches (256 x 256 and 4096 x 4096) at a single resolution [74]. The motivation here is to capture coarser-grained features such as tumour invasion and infiltrating lymphocytes, whilst also capturing the broader spatial patterns present in the TME. This research found that the multihead self-attention modules in ViTs could learn morphological features of the histopathology tissue, even in the pre-training self-supervised setting [74]. Two separate ViT models were trained for each patch size and then combined with another ViT model for final predictions. Other research explores adjusting the ViT network itself for more interpretable histology features. Tran *et al.* replace the linear layers in a ViT with a nonlinear B-cos transform for better alignment between the weights and input data, and find that their approach results in attention heat maps which can better highlight morphological structures in the histopathology such as nuclei and cytoplasm [75].

ViT Position Embedding

The original ViT model [53] was designed to split a standard size image into smaller patches of only 16 x 16 pixels. It is not designed to work with massive WSIs that are the standard medium in histology. Most of the ViT network can adapt well to scaling up to process larger patches, but the problem with WSIs is that the size of the WSI can vary, and hence the number of patches per slide can vary too (if the patch size is kept consistent).

This is a problem because the ViT network contains a position embedding token, which is added to the input at the start of the model network, and is used to embed the position of a patch within the input sequence. In the original Transformer paper the main position embedding they use is a simple learnable 1D Pytorch Parameter embedding, to capture where words lie in a sequence of text. The ViT paper also implements this 1D position embedding, which is randomly initialised and therefore adds no implicit bias regarding spatial positioning of the patch.

They also experiment with other position embeddings, including 2D embeddings using the sine and cosine functions.

The problem with absolute embeddings, however, is that to define the position embedding requires the user to define the number of patches expected per image. They are not designed for WSIs, but rather standard size images which can be split into the same number of 16 x 16 pixel size patches to process. This is apparent in the definition of the position embedding, which in the original ViT paper [53] is implemented as a learnable Parameter embedding of fixed length. The length of this parameter is initially defined as the maximum number patches across the whole dataset, and therefore is not flexible dependent on the varying size of the WSIs. During training and evaluation of the standard ViT model on histology WSIs, one solution to this problem is that for each image the length of this position embedding is truncated to the number of patches in that particular WSI. The issue here is that the latter few values of the embedding are learnt on very few slides, because most slides have a number of patches far fewer than the maximum, and in these cases the latter end of the position embedding parameter is simply not trained.

Huang *et al.* [76] perform survival analysis on WSIs using the Transformer model network for three different types of cancer. Their solution to the problem of the position embedding length is to sample 600 patches from each slide, in order to make the number of patches per slide constant. They also use a 2D position embedding, which embeds the original coordinates of the patch on the WSI [76]. [63] also use a 2D position embedding for their WSIs, choosing to use sine and cosine functions instead of a learnable entity that can be trained.

Another solution to get around the problem of an absolute position embedding for WSIs is to get rid of it completely, as Zheng *et al.* [77] do to classify cancer WSIs. Instead, they use a graph-transformer model, which relies on graph convolutions to capture the local and global context of the tissue rather than a position embedding.

Some research has been done on using CNNs for the position embedding, to capture both local and global context of the patches [78]. Chu *et al.* [79] propose a conditional position encoding for the ViT model, which focuses the learning of

the encoding on the local neighbourhood of each patch only, meaning that their position encoding can in fact generalise to any size WSI, even those larger than seen during training the model. Specifically, they propose positional encoding generators which restore the 1D input sequence into a 2D shape, mimicking the image, and then apply 2D convolutions with zero padding on the array where necessary. As shown in [80], the addition of zero padding can help to implicitly learn absolute position information in CNNs. However, the work by Chu *et al.* [79] is not applied to WSIs, and so they don't consider the fact that the patches may come from various regions of the image due to tissue masks and empty space, and hence to simply convert the 1D input sequence into a 2D array will not restore the tissue patches to their original position on the WSI.

Shao *et al.* [78] propose the TransMIL model (Transformer based Correlated Multiple Instance Learning) for WSI classification, which is a development on the ViT model using correlated MIL to aggregate ViT model predictions. They introduce a novel position embedding for their ViT model, purposefully built to adapt to patches in WSIs. They propose the Pyramid Position Encoding Generator (PPEG) module, designed to capture positional information of the tissue patches and encode both local and global spatial information. This work is similar to that seen in [79], since they first restore the input sequence into a 2D image space, and then apply CNNs to encode the position information. Building on this, Shao *et al.* [78] also apply three separate group convolutions to encode positional information which can be learnt by the model, using convolution kernels of different sizes to capture spatial information with different granularity. The outputs are then fused and flattened to the original input shape. However, as in [79], before applying their PPEG module they convert the input sequence to lie in a 2D image space, without considering the original position of the patches in the WSI.

2.3.3 Graph Neural Networks

The field of geometric deep learning aims to use the known physical structure of data in the deep learning process instead of discarding it, to help simplify the

high-dimensional optimisation landscape. CNNs already use geometric priors, since they apply and learn the weights of the same convolutional kernels across the whole of the image data, assuming geometric properties such as shift-equivariance and translational symmetry [81]. Use cases extend far beyond images, and geometric deep learning has been used to model manifolds of the brain, social networks and molecular structures [82–84].

In this work we focus on GNNs, which derive from the field of geometric deep learning, since the graph itself has inherent geometric properties. Firstly, a graph consists of nodes and edges between pairs of nodes [81]. The nodes can be associated with any number of features, and the edges between nodes can be directional or not. A key geometric property of the graph is that the nodes are not ordered, and therefore any functions applied to the graph should be permutation invariant, meaning the output will be the same regardless of the order of the nodes when input into the function [81]. This property is also known as isomorphism, and can be tested using the traditional Weisfeiler–Leman (WL) test [85]. To ensure permutation equivariance on the graph, meaning the output of a function changes in the same way as the input if also changed, we can specify a function ϕ that operates on each node over its local neighbourhood, usually defined by having a directly connected edge.

GNNs can be either spectral or non-spectral (i.e. spatial). Spectral approaches are based on a spectral representation of the graphs and depend on the graph structure, whereas spatial approaches work directly on the graph, updating based on neighbouring nodes. However, since spectral GNNs cannot be applied to graphs with different structures, and histology WSIs are highly variable from one to another, in this work we focus on spatial GNN approaches [86].

The local ϕ function over each node's neighbourhood is known as the update function. In learning a GNN, this function is applied to calculate each node's updated value. Prior to applying this function, we can also apply a transformation function to each node, and then a nonparametric aggregation function over the neighbourhood of the node to allow for neighbourhood contributions to the updated node value. This is repeated for each layer of the GNN, and with each layer a

Algorithm 1: Training a Graph Neural Network

Input : pre-defined graph G

repeat
  foreach *node u in G* **do**
    Optionally transform features in node neighbourhood N_u with learnable transform function ψ e.g. linear layer with activation;
    Aggregate features from N_u with nonparametric permutation-invariant aggregation function $\bigoplus$ e.g. mean, sum or maximum;
    Update features of node u with learnable update function ϕ e.g. linear layer with activation and optional skip-connection or multilayer perceptron
  end
until *enough layers in the model or reach across the nodes*;
Optional global pooling on final graph for a scalar output;

Output : node or graph prediction

single node's information spreads further across the graph. Finally, depending on whether the desired output is at the node- or graph-level, a global pooling operation can be applied to the ultimate graph for a single scalar output per graph. The generalisable algorithm for a GNN can be found in Algorithm 1 [81].

The Graph Isomorphism Network (GIN) [87] aims to make a powerful GNN by providing a simple injective function for updating the node values, n. It uses the updating function

$$n_u = \phi\left((1+\epsilon)\cdot n_u + \sum_{v\in N_u} n_v\right), \tag{2.2}$$

where ϕ is a neural network such as a multilayer perceptron (MLP), and in this work we set $\epsilon = 0$. Graph Attention Networks are another popular GNN we explore in this work, which employ masked self-attention layers, attending over a node's neighbours [86]. It's possible to also learn edge features in a graph representation, describing relationships between neighbouring nodes by applying convolutions on the edges themselves [88], but this extends beyond the scope of this work.

Classical Graph Techniques

Prior to the development of GNNs and their use in histopathology, graphs were still being used in a more classical sense, where hand-crafted features extracted from the graphs could be informative for image analysis. Sharma *et al.* published a review on these graph-based methods, highlighting commonly used global graph measures such as number of nodes, edges and trees to indicate size, number of neighbours for the degree of a node and clustering coefficients [89]. Graph-based clustering has been used for the purpose of quantifying immune infiltration in cancer tissues, where segmented cells were used as the nodes of the graph, to capture the spatial relationships between cell types such as tumour, stroma and CD8 cells [90].

To help model CRC, Sirinukunwattana *et al.* first construct cell networks to represent the TME, consisting of four segmented cell types: malignant epithelial cells, fibroblasts, inflammatory cells and necrosis. They then use unsupervised learning to cluster the local tissue regions into phenotypic signatures, based on the connections between cell types and the frequency ratios of these found in the tissue region. The resulting phenotypes can distinguish between muscle, inflammation, tumour, stroma and necrosis, which the authors claim could be used to predict the risk of the cancer metastasising [91].

Deep Learning Graph Techniques

Designing the graph representation of the WSI is an essential step which happens before any GNN is used. There are many different choices for defining nodes, edges and features of both, within the WSI.

Edges between the nodes are usually defined by a spatial distance metric, which helps model the spatial organisation of the tissue. To define which nodes should be connected with edges, one research paper used a threshold on the Euclidean distance between patch features from a VGG-16 model [92]. However, it's more common to use some sort of spatial measure which encapsulates the layout of the tissue within the image. For example, Ding *et al.* randomly sample tumour patches for their graph, but define the edges based entirely on geometric coordinates of

those patches within the WSI [25]. Lee *et al.* argue that sampling patches in this manner causes information to be lost on spatial interactions, but admit that it may be a necessary step to model the entire WSI in a scalable manner [93]. They use a distance threshold for edges, defining an edge between two nodes if there are less than five patches between them. This approach to defining edges seems to be too inclusive, potentially producing more connections than required to capture important spatial interactions on a smaller scale. However, they adjust their network to include information on both the distance and the angle between nodes, so that more explicit fine-grained positional information is incorporated. Aryal *et al.* use the popular k-nearest neighbours algorithm to define edges between nodes, and also explicitly incorporate positional information of the nodes into their network, using B-splines in a Spline CNN to learn the node positions [94]. Since the graphs we consider here are inherently unstructured and permutation-invariant, it can be useful to add explicit positional information in such a manner. Finally, using Delauney triangulation is another popular approach to defining spatially close node neighbours, similar to k-nearest neighbours [95].

The scale of the nodes themselves is another consideration in WSI graph design, since choices for nodes span multiple scales, including pixels, nuclei or cells, patches or even superpatches, which are collections of standard size patches. These superpatches can be combined by collecting patches with similar features from a CNN, for example [93]. Other approaches to setting the graph nodes include using subgraphs to represent regions [96]. To incorporate multiple scales, one paper chose the nodes of their graph as tissue patches at multiple resolutions, defining spatial edges between patches on the same level as well as scaling edges between patches in the same position at different resolutions [49].

Once the graph has been designed, a GIN is a popular version of a GNN for modelling a histology slide [25, 97, 98]. However, due to the complexities of the tumour tissue, a jumping knowledge structure can be added to the GIN framework to incorporate and carry forward node information from all layers of the network [25, 97]. The SlideGraph pipeline implements this jumping connectivity GIN

model, clustering nuclei for the graph nodes, and providing node-level predictions to make their model more interpretable [97]. An MLP is a common choice for the updating function applied to each node after aggregating information from its neighbours [25]. Another type of GNN that's been used to model a WSI is a Graph Attention Network such as GATConv [86], which aggregates information from node neighbours using different weights or attention scores [93, 99], helping to model spatial interactions between different tissue types.

In the literature, graphs have been used to model and make predictions from CRC WSIs specifically, the cancer type we consider in this research. For example, Raju *et al.* implement a Graph Attention mechanism on top of a baseline ResNet-50 model [48] to extract texture features from WSIs of CRC [100]. Once they have found the texture features, they then cluster these patch-based features and use them as nodes in a graph network to predict the stage of the tumour. Another approach uses a graph CNN to predict the grade of the CRC tumour from CRC histology images, where the nodes of the graph are the cell nuclei, derived from a segmentation network [101]. Their novel graph CNN utilises the local context of the cells, as they work with the assumption that neighbouring cells are more likely to interact with each other. Ding *et al.* use a GNN to predict genetic mutations using histopathological images of colon cancer, enhancing their network with features to incorporate local information as well as the global topological structure of the histopathology images [102]. In their later work, their patch-based GNN approach to predict genetic mutations in CRC from H&E slides generalises well to rectal cancer after the model was trained on colon cancer [25].

2.3.4 Spatially Aware Approaches

Whilst ViTs and GNNs can go a long way to help to inherently model spatial interactions and importance between regions of the WSI, other approaches have been developed to address this challenge more directly. Shaw *et al.* have actually combined ideas from the Transformer and graph approaches, extending the Transformer to

incorporate explicit positional information within its structure by considering possible edges between all inputs [103].

Many works aim to capture interactions between the cells, which is a prognostic factor in CRC [20], and use spatial proximity to model this [57, 104, 105]. However, these works are often based on the assumption that spatially close cells will interact, which is not necessarily true and should be carefully considered [106]. To verify the truth behind this, Fu *et al.* propose the need for more studies on spatial interactions both in vivo and in vitro.

Other methods claim a different type of data entirely is required, allowing spatial omics to be used for better understanding of the intricacies of the TME, making predictions and embeddings arguably more explainable [107, 108]. Spatial omics can be used to analyse neighbourhoods of cells, usually quantified by co-occurence or colocalisation of cell types within a user-defined radius [105, 109]. However, this only works under the assumption of linear neighbouring structures. Tanevski *et al.* propose a new nonlinear method called Kasumi which predicts neighbourhoods at different scales across all available tissue patches [107], and argue that their method is a more interpretable alternative to using GNNs.

Segmentation before prediction

The approach of considering segmented regions or cell types within the tissue before using further deep learning models for prediction is a common one. For example, segmenting the nuclei of cells could be useful when wanting to resolve detail on the cell level. Spatially constrained CNNs can be used for this task, which evaluate the probability of each pixel being the centre of a nucleus, while forcing pixels close to the centre to have higher probabilities with the spatial constraint [110]. Segmenting the epithelial and stromal regions can be useful, since changes in the stroma, though not considered particularly malignant, can promote spreading of the tumour, and the epithelium can provide information on the spatial arrangement of the tissue, which is also a prognostic factor in CRC [35, 111]. For reference, the stromal tissue consists mainly of fibrous, fatty connective tissue, and the epithelium is the cellular

lining, and can be segmented and classified using deep CNNs [35]. The tumour can be segmented further, and a CNN can be used to find contents such as necrosis and lymphocytes as well as stroma and tumour. This breakdown of the TME can then be correlated with clinical outcomes using traditional approaches such as Cox regression for survival [112]. The authors found that the tumour-stroma ratio is a significant prognostic factor in colon adenocarcinoma [112]. Finally, glands can be segmented from histology slides by modelling them as ellipsoids, and these can then be used for cancer grading by extracting radiomics features capturing spatial patterns and texture, and feeding these into a Support Vector Machine model for prediction [14].

2.3.5 Foundation Models

Since beginning this research project three years ago, the field of deep learning on histopathology has rapidly evolved with the introduction of histopathology foundation models. These are self-supervised models trained on vast amounts of data, and provide a strong feature embedding for any downstream task, which can be learnt in relatively little time using transfer learning or fine tuning with any MIL method.

ViTs are a very popular choice of model for these foundation models, since they can be trained effectively using self-supervised learning and are provided with enough data to sufficiently learn the large number of parameters within these models. The Bidirectional Encoder representation from Image Transformers (BEiT) method uses masked modelling to pre-train ViTs, trained to recover the original tokens from masked patches [113]. The DINO method (self-distillation with no labels) uses architecturally identical student and teacher models and trains them to return similar outputs using a cross entropy loss, where the two models are fed different cropped regions of an input image [114]. This approach builds on the SimCLR method, which again trains using a contrastive loss in the latent space on augmented image pairs, but uses a ResNet model as its baseline instead of a ViT [115].

These methods moved the field of general computer vision towards foundation models by proving that self-supervised methods can provide better feature embeddings than supervised or semi-supervised alternatives. It wasn't long before

researchers developed such methods for histopathology images, with CTransPath being one of the first methods using unsupervised contrastive learning with a ViT model, trained on a large number of histology slides from multiple cancer types, to develop a robust feature extractor [116]. In a 2021 study, Chen *et al.* found that DINO performed well on histopathology compared to the self-supervised SimCLR method, hypothesising that this could be due to capturing the hierarchical local-global context of the tumour tissue with the DINO augmentations in the ViT model [117]. More recently, members of the Mahmood lab in Harvard published their UNI model, trained on over 100,000 WSIs [118] using the DINOV2 approach [119]. They demonstrated the quality of their feature embeddings compared to other histopathology self-supervised feature extractors on downstream tasks such as tissue type classification on the NCT-CRC-HE-100K dataset of CRC WSIs, and have shown comparable results.

2.4 Interpretability

With larger models than ever, interpretability is becoming more and more important for understanding the reasoning and motivation behind these deep learning model predictions. Many people work on the issue of interpretability, but this term is used in a rather loosely defined manner, and its meaning can differ drastically from one area of research to another. Later on in this thesis we will address this ambiguity in interpretability methods and propose our own method, tailored to the problem at hand.

Most people agree on the reasoning that simpler, transparent models are more interpretable, including more traditional machine learning methods such as logistic regression, where each covariate is assigned a certain weight, and decision trees, which make interpretable rule-based decisions (provided they are not too deep) [120, 121]. These approaches could be classified as having model-intrinsic explainability, since the explanations are inherently built into the model architecture [121]. Simpler rule-based algorithms are particularly popular in clinical decision making, as they

can be easily interpreted by a clinician, such as the Gleason grading system for prostate cancer [122]. There is also the advantage that such methods have been around for a long time, meaning practising clinicians today will have been trained to interpret such algorithms throughout their career, and will therefore are more likely to be familiar and comfortable with these decision tools, addressing the issue of trust.

Using gene-based signatures is one popular approach for clinical rule-based algorithms. Kim *et al.* establish a nineteen gene-based risk score to predict which CRC patients would benefit from adjuvant chemotherapy [123]. They use statistical methods such as clustering and generalised linear models, the latter of which provide inherent interpretability. ColoType is another gene-based signature from forty genes, which was demonstrated to be able to predict the CRC CMS classifications [124]. To predict response to therapy from these CMS classes, the paper mentions some reports of specific drugs being predicted, however further subtyping was required.

Other approaches to interpretability are, however, usually required for deep learning models, due to the difference in scale of the number of parameters in these models. As taxonomised in the survey by Das *et al.*, some explainability methods can be categorised by their methodology, based on either backpropagation over the gradients or perturbations in the input data [121]. Both methods are popular for model-agnostic, post-hoc explanations, such as Local Interpretable Model-agnostic Explanations (LIME) [125] and Gradient Class Activation Mapping (GradCAM) [126]. LIME is a perturbation-based method which provides explanations for individual predictions by approximating a local model using permutations in the input data [125]. GradCAM is a backpropagation-based method which provides explanations for individual classes within an input image, and allows the user to visualise the higher valued activations in a heatmap [126]. While these methods have no guarantees of accurately determining the true underlying reasons for a certain prediction, they have been shown to be able to identify previously unknown imaging biomarkers [127].

Lipton *et al.* argue that true interpretability means finding causal associations between the input data and the predicted outcome, similar to the definition by Das

et al., as features which provide enough understanding as to how the algorithm works [120, 121]. However, Lipton *et al.* point out that, in most settings, deep learning models are trained using an unsuitable loss function whose goal is simply to minimise error between predictions and true labels, as opposed to finding any sort of causality or explainable latent features [120]. We need to be aware that the explainability of models is limited, motivating my work in this thesis on predicting an event together with some external correlates, solidifying the features which provide the underlying feature set for our predictions.

2.5 Biomarkers in the Clinic

While we have introduced many successful methods for predicting biomarkers from histology images, very few of these have ever made it into the clinical setting [128]. Extensive validation is (and should) be required before patient treatment decisions are made based on these algorithms, though Burke *et al.* argues that this lack of translatability to the clinic is instead due to a simple failure to understand these biomarkers [128]. As discussed in the previous section on Interpretability in Section 2.4, categorical biomarkers (as opposed to numerical) have historically been more prevalent in the clinic due to their comprehensible nature, such as objective response defined by the RECIST 1.1 criteria [129].

The TNM (tumour, node, metastases) system was one of the first rudimentary biomarkers for cancer staging and was also used for determining patient treatment decisions [128]. This popular system, still used today and clinically quantified on CT and PET images [130], categorises cancer based on the tumour size (T), state of the lymph nodes (N), and presence of distant cancer metastasis (M) [128]. While prognostic in some cancers such as prostate cancer [130], the TNM system cannot perfectly predict prognosis since due to its simplicity it cannot model nuances in individual treatment response [128].

Burke *et al.* stress that time is a crucial yet frequently ignored aspect of cancer progression and predictions of prognosis. While time may be correlated to the

biology of cancer progression, it is not a direct linear determinant in its causality, as is often assumed in biomarkers such as TNM staging [128]. Therefore, Burke *et al.* recommend, or even require, a time interval to be a unit of measure for any prognostic prediction, such as survival within five years or no recurrence after therapy within two years, depending on the clinical data and endpoints used to develop the biomarker.

Biomarkers are also not independent from each other, and cancer can be an incredibly complex disease. It should be acknowledged at all points in the development of a biomarker to use in clinical treatment decisions that the components which contribute to quantifying a biomarker do not act independently, but interact with all aspects of the system around them [128]. Performance of biomarkers can also be affected by the tools used to measure them, such as the image scanners, but how these changes in tools affect individual biomarker measurements is not of primary concern for the retailers and their customers [130].

2.5.1 Measuring Performance

It's very important to consider how the performance of biomarkers is measured. For prognostic and predictive studies that wish to demonstrate an improvement using the biomarker over current approaches, metrics such as hazard ratios should be reported. For screening and diagnostic studies, metrics such as sensitivity and specificity should be reported, since these metrics can provide insights into how important subgroups of patients will be dealt with, not just the majority. The area under the receiver operating characteristic curve (AUROC) is another good nonparametric and nonlinear metric, since it does not depend on classification thresholds and therefore provides a good representation of predictions over imbalanced datasets [128, 130].

There are many possible metrics to use which are formulated from the number of true positives, true negatives, false positives and false negatives from a prediction model. These values can be portrayed in a confusion matrix, as seen in Table 2.1. Common metrics derived from the confusion matrix can be used with different names across the medical and deep learning spaces, so we provide definitions here for clarity. Where feasible, we try to define related metrics in terms of specificity

Confusion Matrix	Predicted Positive	Predicted Negative
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

Table 2.1: A table showing the format of a confusion matrix, categorising predictions from a model based on the predicted value and the actual ground truth value, with each entry showing the total number of predictions in that category.

and sensitivity, since these are popular metrics in the clinical domain and necessary for clinical translation and understanding.

Firstly, we define sensitivity and specificity with regards to values from the confusion matrix,

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{2.3}$$

Sensitivity is also known as the recall or the true positive rate (TPR). Similarly, specificity is also known as the true negative rate (TNR). Recall is often presented alongside precision, which is also known as the positive predictive value (PPV), as defined by

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2.4}$$

Then the F1 score can be defined from recall and precision, such that

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{2.5}$$

Returning to sensitivity and specificity, we can defined the balanced accuracy metric as the mean of these measures, such that

$$\text{balanced accuracy} = \frac{\text{sensitivity} + \text{specificity}}{2}, \tag{2.6}$$

however accuracy itself is quantified directly from the confusion matrix, such that

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}, \tag{2.7}$$

where P and N are the counts of true positive and negative values in the dataset, giving the total number of data points when summed.

Finally, the receiver operating characteristic (ROC) curve can be defined by calculating sensitivity and specificity for model predictions using multiple thresholds, spanning the range of 0-1 for a binary classification. These values can then be plotted, with sensitivity on the y-axis and $1 -$ specificity on the x-axis, which is also known as the false positive rate (FPR). The area under the receiver operating curve (AUROC) is exactly that, given by the area underneath the plotted curve.

2.5.2 Stages of Biomarker Development

O'Connor *et al.* set out the stages of imaging biomarker development in their roadmap, defining three main stage domains: discovery of the biomarker, validation, and finally qualification and ongoing technical validation. The discovery of the biomarker establishes a signal in a research setting. The biomarker should be technically validated with regards to precision, biases and availability. The authors highlight the need for biomarkers to be generalisable and reproducible across geographies and centres. The biomarker should also be biologically and clinically validated in the context of a meaningful intervention in a clinical trial. Finally, it's necessary that the biomarker should improve health benefits over existing practices but still be cost effective, delivering an economic motivation for healthcare providers in terms of a better cost per quality of adjusted life year (QALY) [130].

2.5.3 Example of Biomarkers in the Clinic

Here we review both classical and deep learning based methods which have been implemented in the clinical setting.

Immunoscore is a biomarker based on digital pathology of resected CRC tumours, which has been shown to be prognostic in localised colon cancer when used alongside the TNM system, and can inform on chemotherapy treatment decisions [32, 131, 132]. It uses deep learning techniques to measure the densities of TILs found at specific tumour sites, quantifying immune cell infiltration into five Immunoscore values, 0 to 4. It has been extensively tested in multiple independent studies consisting of thousands of patients [131]. The intellectual property is CE marked and the score has been used by private hospitals in the UK, at the cost of £2,250 per test [131].

The company Owkin has developed their own CE-marked biomarker for CRC, called MSIntuit [133]. Their deep learning based biomarker screens CRC patients for MSI based on the H&E stained slides, as is recommended by the National Institute for Health and Care Excellence (NICE) regulatory body in the CRC patient pathway [134]. MSI tumours can identify patients with Lynch syndrome, who can therefore benefit from different treatment options and closer observation, since this patient population is at higher risk of developing cancers in other organs. The biomarker has been recently validated in an independent dataset of 600 patients, achieving sensitivity of over 0.96, minimising the risk of misclassifying a patient with MSI tumours [135].

ArteraAI is a pathology-based biomarker for determining which prostate cancer patients will benefit from short-term hormone therapy in addition to radiotherapy treatment [136]. It has been clinically validated in five Phase 3 randomised trials, and is recommended for use in the USA by the National Comprehensive Cancer Network [137], despite not yet being approved by the governing body in the USA, the FDA.

Paige, on the other hand, is a company with an FDA-approved diagnostic biomarker for prostate cancer, the first one approved for detecting cancer in prostate needle biopsies [138]. Their underlying technical approach builds on weakly supervised deep learning methods developed by Campanella *et al.* [59]. Their FDA-approved Paige Prostate Detect provides a score on the cancer grading, and another biomarker evaluates the Gleason score for each pathology slide [139]. In the UK, however, NICE found in their 2021 report that evidence was lacking, and

were not convinced by the generalisability of the method to different populations, especially since none of the studies were based in the UK [140].

2.6 Summary

There has been a lot of work in recent years in the field of medical imaging, in particular for cancer research. Similarly, there have been developments in the field of computer vision that create opportunities to apply the latest deep learning models onto medical images. Despite this, there has been little work done on using deep learning to predict a colorectal cancer patient's response to radiotherapy from histology biopsy slides, which is what we aim to do in this project. Furthermore, the computer vision models are not designed with histology slides in mind, and so we work on developing deep learning models and approaches to training which are tailored to this modality and the underlying cancer biology.

3

Data

Contents

3.1 Datasets 43

3.1.1 Grampian Dataset 44

3.1.2 Aristotle Dataset 45

3.1.3 Salzburg Dataset 46

3.2 Exploratory Data Analysis 47

3.2.1 Image Analysis 47

3.2.2 Outcome Analysis 49

3.3 Slide Processing 53

3.3.1 Slide Magnification 53

3.3.2 Patching 54

3.1 Datasets

There are three retrospective colorectal cancer (CRC) histology datasets that are available for this research: Grampian, Aristotle and Salzburg. All three datasets contain haematoxylin and eosin (H&E) stained slides. Both the Grampian and Aristotle datasets come from the Stratification in COloRecTal cancer (S:CORT) programme, a UK-wide consortium funded by the Medical Research Council (MRC) and Cancer Research UK (CRUK) with the aim of gathering a large high quality multi-omic dataset on CRC. The Salzburg dataset is completely independent to the other two datasets, and comes from the University Hospital Salzburg in Austria.

Treatment

All patients here received the same treatment, standard chemoradiotherapy of pelvic irradiation (45-50.4Gy in 25 fractions over 5 weeks) with capecitabine 900mg/m^2, regardless of cohort.

Chronologically, the patients underwent a pre-treatment biopsy to detect and analyse the tumour, followed by chemoradiotherapy treatment to shrink the tumour, and some time later a post-treatment resection to evaluate the response to therapy. The images we consider in this work are from the pre-treatment biopsy, taken prior to radiotherapy treatment.

Crucially, all three datasets have corresponding outcome labels which we are interested in. Specifically, for the same treatments, we have the patient's recorded response to radiotherapy (RT), which we consider as a binary outcome: whether the patient had a pathological complete response (pCR) to RT or not. Unfortunately, most patients do not have a pCR, meaning that our datasets are very unbalanced with regards to the outcome label.

This data label imbalance presents a challenge for us throughout this research, along with the demographic limitations of the data, since we have only three patient

cohorts to work with. However, these are real-world problems that need to be addressed, since histology slides labelled with this outcome are hard to come by, and the outcomes will always be imbalanced due to the nature of the treatment and the cancer itself. We attempt to be realistic with our approaches and inference, considering these inherent properties of the data throughout our research.

CMS

The CMS labels for this data are derived from three different transcriptomic versions (single cohort, combined cohort correcting batch effects and combined cohort including 2036 cases run with the same platform) in order to generate robust classifications. In all cases the CMS call was calculated using the CMSclassifier random forest and single sample predictor [26]. Final CMS calls are based on matching calls between the three transcriptomic versions, and can be defined as 'Unmatched' if the calls are not matching. Some calls can also be defined as 'Unclassified' due to problems with the biological transcriptomics process. Despite efforts to minimise the noise from RNA sequencing, we still expect a certain level of noise in our ground truth data, which we discuss in the corresponding research sections.

3.1.1 Grampian Dataset

The Grampian dataset contains 527 H&E slides of pre-treatment endoscopy biopsies of rectal cancer. For most patients there are two slides from the same biopsy, sliced a few microns apart. In the quality control stages, led first by Susan Richman and Phil Quirke of the S:CORT consortium, and subsequently checked by the pathologist Viktor Koelzer, samples were excluded due to unverified clinical data and low quality. This quality check excluded slides with poor quality of staining, out-of-focus images and those without sufficient tumour content. In a subsequent quality control stage led by myself, I excluded corrupted image files and slides with no corresponding tumour mask. These annotated masks are generated from black

pen circles drawn on the physical slides, to highlight the relevant tumour area on the slide to remove for gene sequencing. The masks are used here to filter down which regions of the image are patched for use in the analysis. After quality control, there are 499 slides remaining with matching tumour masks.

Slides without any recorded treatment response outcome data were excluded, leaving 452 slides from 231 patients. In the dataset used for inference on the RSS gene score, some slides were excluded further due to lack or low quality of genetic data where the transcriptome or next-generation sequencing failed, determined by Enric Domingo of the S:CORT consortium. After this step there are 437 slides from 223 patients remaining for the research on RSS.

In the Grampian clinical trial, patients received different forms of treatment. For our inference on a patient's response to RT, we consider a single type of chemoradiotherapy treatment, RT and Capecitabine (CapRT). Therefore we must only use data from patients who all had this same type of RT treatment, and therefore once we remove those patients who did not receive CapRT, our dataset reduces down to 133 patients for whom we have 258 slides.

Later in this thesis when we consider the CMS classification as an additional response, we remove slides which do not have this label and the number of available slides goes down slightly to 247.

3.1.2 Aristotle Dataset

The Aristotle dataset contains 610 H&E slides of pre-treatment endoscopy biopsies of rectal cancer. As in Grampian, slides were excluded in an initial quality control stage by members of the S:CORT consortium due to unverified clinical data, reducing the size of our dataset to 606 H&E slides from 303 patients. For each patient there exist two slides from the same biopsy, however we only have tumour masks for one slide per patient, and hence only use one slide per patient in our analysis. This leaves 303 slides from 303 patients for our analysis.

In the dataset used for analysis on the RSS gene score, some slides were excluded further due to lack or low quality of the genetic data where the transcriptome or

next-generation sequencing failed, determined by Enric Domingo of the S:CORT consortium. After this step there are 298 slides from 298 patients remaining.

In the Aristotle clinical trial, 28 patients did not have surgery which means we cannot in the same way measure their response to RT. The response to RT data was simply not provided in our dataset for a further 151 patients, meaning for our analysis in predicting response to RT we have 124 slides from 124 patients available. All of these patients received neoadjuvant CapRT therapy.

For later analysis we further remove slides without the CMS classification, reducing the number of available slides slightly to 121.

For the Aristotle data the tumour masks are generated by a clinician reviewing the slides, considering the black pen marks drawn around the parts of tissue to be used for the gene sequencing. In Aristotle, not all of the tumour was originally circled on the slide to be taken for gene sequencing, so the clinician reviewing these slides edited the tumour masks to include these additional regions of tumour tissue.

3.1.3 Salzburg Dataset

The Salzburg dataset is smaller than the other two, containing 61 H&E slides from rectal preoperative biopsies. Each slide contains multiple serial sections of the same biopsy specimen, resulting in generally more sections per slide than is seen in the S:CORT datasets.

Some slides were excluded from our research in a quality control phase, due to poor staining, high grade dysplasia and lack of invasive cancer, and insufficient tumour content, leaving us with 55 slides for use in our research. Tumour masks were also provided for all slides, manually generated by an expert pathologist.

All patients received neoadjuvant CapRT chemoradiotherapy treatment, as in the other cohorts. The patients who received this treatment, and therefore are in this cohort, were selected by being considered at high risk for rectal cancer. Between 6-12 weeks after treatment, a further resection was taken to measure pathological response, recorded using the Dworak tumour regression grading system

[51]. In this research, we accept a Dworak tumour regression grade of 4 to define our positive outcome label, CR to RT.

It should be noted that we did not have access to this Salzburg dataset from the beginning of this research, and therefore the use of this dataset appears only in the later chapters. Hence we also do not consider the RSS gene scores of this cohort.

3.2 Exploratory Data Analysis

To provide some insight to the images we are considering in this research, we provide thumbnail images of three randomly sampled slides for each cohort. These can be viewed in Figure 3.1, highlighting the potential differences and similarities in the images across the three cohorts.

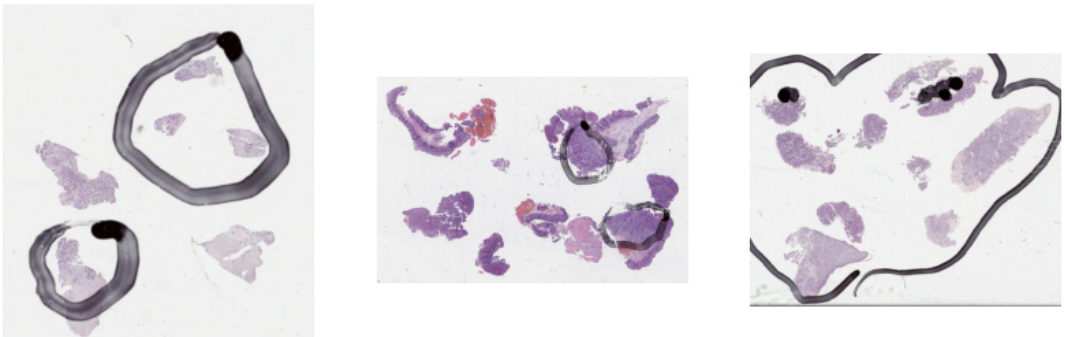
3.2.1 Image Analysis

Staining across cohorts

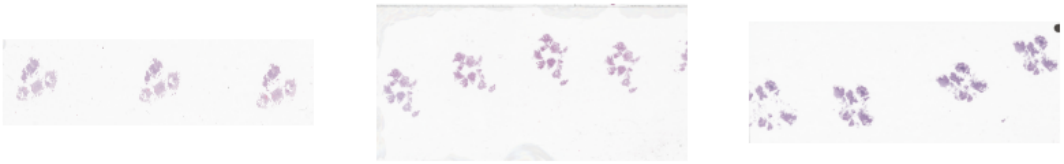
To explore the different staining effects across cohorts, we separated the RGB (red, green and blue) channels of the WSIs into their contributing haematoxylin (H) and eosin (E) channels, and plotted a histogram of the mean value of each H&E channel per WSI, stratified by patient data cohort. The histograms can be seen in Figure 3.2. Haematoxylin shows up as purple in the WSI and highlights nuclei, whereas eosin shows up pink in the RGB image and highlights cytoplasm and the extracellular matrix.

In both the H and E channels, generally speaking Grampian has the higher presence of each stain. Salzburg has the lowest values for the H channel, and Aristotle has the lowest values for the E channel. This figure highlights the differences between the images due to the staining procedures, which could also be due to inherent differences in the tissue itself, either due to patient demographics or sampling procedures.

(a) Grampian

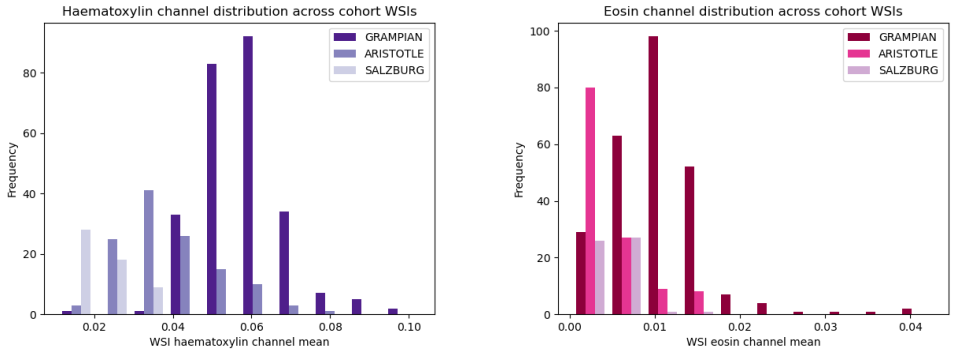


(b) Aristotle



(c) Salzburg

Figure 3.1: Randomly sampled WSIs from each of our three dataset cohorts. The black pen marks were drawn by a pathologist circling the tumour. These samples represent the diversity of histology slides that can be found both between and within cohorts, with variations in colour and size of tissue. The Salzburg slides generally contain more biopsy tissue samples than the Grampian and Aristotle slides, hence those samples look smaller at the zoomed out level, but they are also scanned at a higher magnification and so can still be reasonably compared.



(a) Haematoxylin **(b)** Eosin

Figure 3.2: Histograms of the mean values of each staining colour channel, (a) haematoxylin and (b) eosin, in each slide, separated by cohort.

3.2.2 Outcome Analysis

RSS distribution

The distributions of the RSS gene score outcome appear different across the two datasets, as seen in Figure 3.3. In the Grampian dataset the distribution of the RSS score is bimodal, with both peaks in the upper half of the distribution range. The RSS score in the Aristotle dataset has just one defined distribution peak, and a much larger tail towards the lower end of the distribution range, where the RSS score from Grampian is not as well represented.

Exploring this further, we review the known dissimilarities in the cohorts of Grampian and Aristotle, to potentially explain the difference in distributions of the RSS score seen in Figure 3.3. Geographically, the patients in the Grampian cohort live in a specific area of the UK, North East Scotland, whereas the samples in Aristotle come from multiple different sites around the UK. While samples from both cohorts were processed identically at a central lab, they were processed roughly two years apart [132]. Finally, in their analyses across the two cohorts, Domingo *et al.* observed significantly different measurements in some genetic variables, including APC mutation, TP53 mutation and Chromosomal Instability [132].

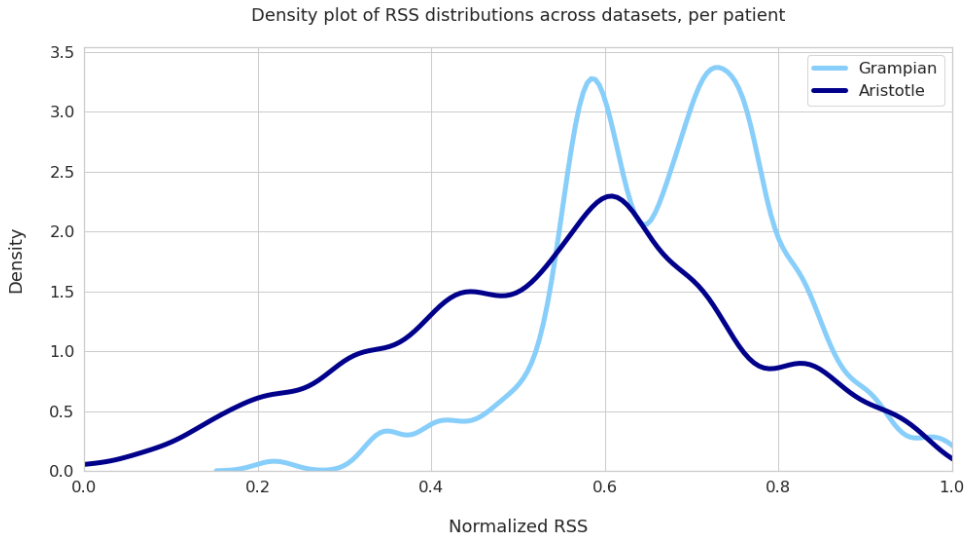


Figure 3.3: Density plot of the RSS gene score distribution, stratified by dataset (Grampian or Aristotle). This plot shows the smoothed distribution of the normalized RSS scores for each patient in the datasets. The RSS gene score shows a unimodal distribution in the Aristotle dataset, with a larger tail towards the lower end of the RSS range. The RSS gene score in the Grampian dataset follows a bimodal distribution, and has very few values at the lower end of the RSS score range.

Response to radiotherapy distribution

The counts of the classifications of the response to RT outcome for Grampian and Aristotle can be found in Table 3.1. The classification system differ across the three datasets, but in order to use them together in our analysis we need to match the classes across the datasets. The Aristotle dataset has two classes for complete and no complete response already, whereas the Grampian dataset contains four classes measuring complete, good partial, partial and minimal responses. In order to align the two datasets, we define a no complete response (NoCR) class in Grampian to include good partial, partial and minimal responses. In other words, everything other than a complete response (CR) is defined as NoCR.

The Salzburg dataset uses the Dworak tumour grades primarily, from 0-4, and we define a CR as a Dworak grade of 4 and a NoCR as all other grades, since Dworak grade 4 means lack of tumour cells, same as pCR, but a Dworak grade 3 could mean a few tumour cells remaining, similar to the good partial response class in Grampian. The distribution of the Dworak grades across the Salzburg cohort can be found in

Response	complete	good partial	partial	minimal	no complete	no surgery	not provided
Grampian	67	98	80	13	-	-	-
Aristotle	24	-	-	-	100	28	151

Table 3.1: Counts of response to radiotherapy classifications found in the Grampian and Aristotle datasets, for a total 258 slides in Grampian and 303 in Aristotle.

Dworak grade	0	1	2	3	4
# in Salzburg	1	11	18	19	6

Table 3.2: Counts of post-treatment Dworak grade classifications found in the Salzburg dataset, where Dworak grade 4 indicates a complete response to radiotherapy, for a total of 55 slides.

Table 3.2, where it can be observed that a high proportion of patients were classified with Dworak grade 3, meaning that they almost had a CR to RT, but not quite.

The counts of the binary classifications of response to RT for all cohorts can be found in Table 3.3. In Grampian the CR class represents 26% of the binary response data, in Aristotle the CR class represents 19% of the binary response data, and in Salzburg CR represents 11% of the data.

CMS distribution

Of the patients who received neoadjuvant CapRT, filtering slides after the preliminary quality control but before selecting for recorded response to RT (n=258 in

Binary Response	complete	no complete
Grampian	67	191
Aristotle	24	100
Salzburg	6	49

Table 3.3: Counts of binary response to radiotherapy classifications found in the Grampian and Aristotle datasets, for a total 258 slides in Grampian, 124 in Aristotle and 55 in Salzburg. The 'no complete' response class includes the original response classes 'good partial', 'partial', 'minimal' and 'no complete' and Dworak grades from 0 to 3.

Cohort	CMS1	CMS2	CMS3	CMS4	Unclassified	Unmatched	NaN
Grampian	21	74	52	29	48	26	8
Aristotle	16	39	18	19	43	17	3
Salzburg	4	14	11	14	12	-	-
Total	41	127	81	62	103	43	11

Table 3.4: Counts of CMS calls found in the Grampian, Aristotle and Salzburg datasets, for a total 258 slides in Grampian, 155 in Aristotle (including patients without a recorded response to radiotherapy) and 55 in Salzburg.

Grampian, n=155 in Aristotle, n=55 in Salzburg), we present the distributions of the CMS calls in each cohort in Table 3.4. CMS2 is consistently the most frequently classified subtype across cohorts, followed by either CMS3 or CMS4, with CMS1 being the minority subtype. Unfortunately a number of slides have not been classified with a valid CMS call, due to the difficulties in extracting this molecular subtype.

Size of tumour mask

We wanted to explore whether the response to RT outcome was correlated with the quantity of the tumour tissue in the biopsy, as outlined by a pathologist, since this could be an indicator of a larger tumour mass in vivo, or alternatively less remaining tumour post-biopsy. To evaluate this, we used the tumour masks defined by the pathologist and calculated the area of these, keeping the same resolution across cohorts. However, as previously mentioned, there are many more biopsy samples in each of the Salzburg slides than the other two cohorts, and so the raw amount of tumour across responses should only be compared within cohort, not across cohort. The box plots showing the amount of tumour in each WSI, stratified by cohort and response, can be seen in Figure 3.4. We've excluded the outliers of the box plot intervals since there are sparse but large outliers across all cohorts which corrupt the visualisation aspect ratio.

Within Grampian and Aristotle, visibly there's not a big difference in tumour size between responder groups, but within Salzburg there is a more of a difference between responders, with complete responders seeming to potentially show larger

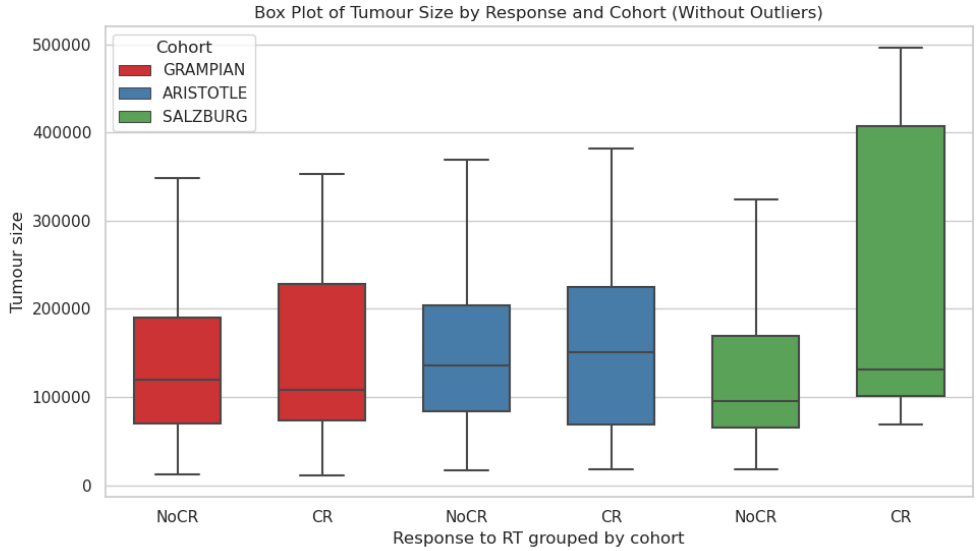


Figure 3.4: Box plots of the size of the tumour in each WSI, stratified by cohort and response (CR or NoCR). Since there can be different numbers of biopsy samples in each cohort, the box plots should only be compared within each cohort and not between them. The tumour size was calculated from the area within the pathologist-annotated tumour masks.

tumour sizes than non-responders. However, there are only six patients with CR in the Salzburg dataset, so there is not enough power here to make claims of significance.

3.3 Slide Processing

3.3.1 Slide Magnification

The pre-treatment biopsy slides in the S:CORT datasets, Grampian and Aristotle, were all sectioned and stained in the same laboratory and scanned at 20x magnification (0.5 μm^2/pixel) on an Aperio scanner. The Salzburg slides were sectioned and stained within the University Hospital in Salzburg, and were scanned at a higher magnification of 40x (0.25 μm^2/pixel). This allows us to study all images at 5x, 10x and 20x magnification levels, giving us different options of how zoomed in we want the images to be for our analysis. We tend to use either the 10x or 20x magnification levels in this work, to provide more detail in the images.

The images are saved in the standard Aperio *.svs slide format, and hence we can use the popular Openslide library [141] to open and process our WSIs, defining what level of magnification we would like to view the images at.

3.3.2 Patching

Before training on these WSIs, we divide the images into patches (or tiles) so that they can be of a reasonable size to be processed by the GPU. The size of these patches is defined as 256 x 256 pixels, and the stride of the patches is defined as either 128 pixels, so that each patch overlaps with the adjacent patch by 128 pixels, or 256 pixels, meaning there is no overlap between patches. Only parts of the image that are included in the tumour mask are patched, and parts of the image which are deliberately excluded from the mask are ignored, using an inclusion threshold of 50%.

Depending on the tumour mask and the amount of tissue in the WSI, the number of patches per slide varies massively. This is visualised in the histogram in Figure 3.5, which shows the number of patches per slide across each cohort of patients, at 20x magnification with a stride of 50%. Most slides have up to around 5000 patches, but there are some outliers which have over 10,000 patches per slide. This could be due to the amount of tissue taken in the biopsy and seen on the digital slide, as well as how much of this tissue is selected in the tumour mask. The distributions of the number of patches across cohorts are roughly similar, with lower frequencies for Salzburg since it is the smallest dataset. In our datasets the maximum number of patches taken from a slide is 50,770, and the minimum number of patches from one slide is 73. This demonstrates the huge range in the number of patches per slide in our datasets.

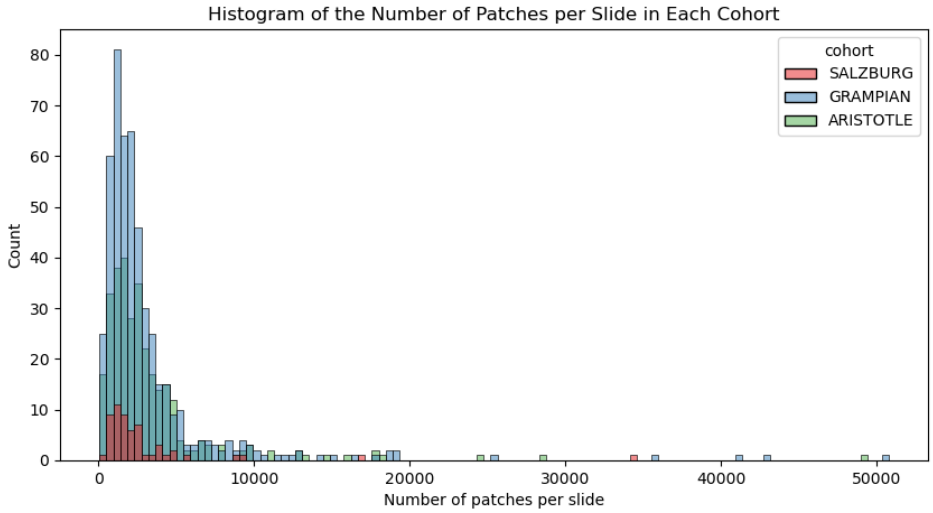


Figure 3.5: Histogram of the number of patches taken per slide across the Grampian, Aristotle and Salzburg datasets. The patches were taken with a stride of 50% at 20x magnification, excluding background and unwanted tissue outside the tumour mask.

4

Predicting Treatment Response From Histology

Contents

4.1 Introduction 58

4.2 Methods 60

4.2.1 Preprocessing 60

4.2.2 Baseline Model 61

4.2.3 Vision Transformers 62

4.2.4 Patch Restoration Embedding ViT (PREViT) 62

4.2.5 ClusterViT 63

4.2.6 ClusterPREViT 63

4.2.7 Clustering Approach 63

4.2.8 Interpretability 65
4.3 Experiment Results 65
4.3.1 Model Hyperparameters 66
4.3.2 Predicting RSS 67
4.3.3 Predicting Response to Radiotherapy 70
4.3.4 Visualising Clusters 72
4.3.5 Attention Heatmaps 74
4.3.6 Odds Ratios 76
4.4 Discussion &Conclusion 76

Contributions

In this work, we will demonstrate the ability of the state-of-the-art Vision Transformer (ViT) models to predict response to radiotherapy in locally advanced rectal cancer patients, which to our knowledge have never been applied for this purpose before, and will demonstrate their ability to provide visual interpretation to image features associated with the treatment outcomes. We also experiment with predicting the RSS gene score from the histology slides, which provides less fruitful outcomes.

We propose two novel variations of ViT to enhance the local context of the tissue patch features by introducing prior information to the model. The first network is the PREViT which uses a novel trainable position restoration embedding (PRE) to restore patches to their original position in the WSI, preserving the spatial relationship between tissue patches. The second is the ClusterViT which introduces a trainable cluster label encoding derived from the baseline model feature embeddings as prior information for the ViT model to capture the different tissue motifs. Two independent clinical trial cohorts, Grampian and Aristotle, are used in the development and validation of the models. Experiments demonstrate that the proposed models improve predictions over the baseline model and the interpretation of predictions is aided by including the PRE and cluster token.

Sections of this work have been published in the proceedings for the Workshop on Medical Image Assisted Biomarker Discovery (MIABID) at the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2022 conference [142].

4.1 Introduction

This work attempts to fill a gap in the literature regarding the possibility of using deep learning to predict a colorectal cancer (CRC) patient's response to radiotherapy (RT) from their histology. Various work has been done on different cancer types and on different image mediums to predict different outcomes, and this is all considered as a foundation for this research to develop on.

Based on the literature we have explored the applications of various deep learning imaging techniques, including traditional CNNs and more recently developed attention networks. In this work we explore predicting both the RSS gene score and the patient's response to RT. A lot of work has been done in the literature regarding predicting biomarkers from cancer histology images, but there is a gap in the knowledge regarding both predicting the RSS gene score (which was unpublished at the time of this research but is now published [15]) and predicting response to RT for CRC using deep learning. We propose to apply and expand upon the latest deep learning techniques to explore if it is possible to predict this information from the histology images.

To refine these predictions we use Vision Transformer (ViT) attention networks [53], which have not been applied for this purpose before. We propose two novel deep learning networks, both variations on the ViT network architecture, and show how applying these models in this scenario improves our prediction results. The first novel network introduced in this report is the PREViT which uses a special position embedding to restore patches to their original position in the WSI. The second novel network introduced is the ClusterViT which includes a cluster label from the baseline model feature embeddings as prior information for the model input. Our results shows that both the PREViT model predictions and the ClusterViT model predictions are an improvement on the baseline models.

Not only does applying the ViT network and its variations improve our model prediction results, but it also opens up the opportunity for more interpretability of the predictions. In this chapter we demonstrate the use of heatmaps from

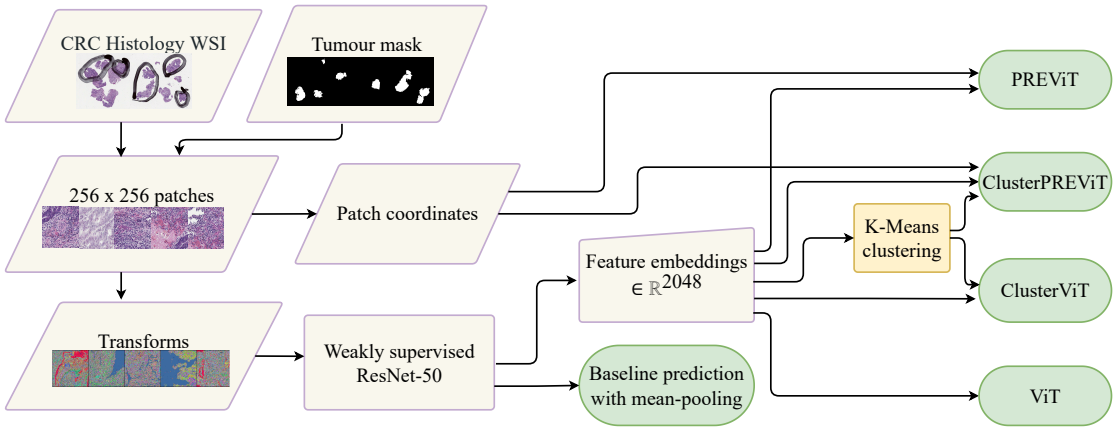


Figure 4.1: This figure shows the data and model pipeline for this work. We start with the CRC histology WSI, and then apply preprocessing steps to the images including splitting the WSIs into patches of size 256 x 256 pixels. These patches are used to train a ResNet-50 model, using the slide-level annotation as a label for each patch. From here we can generate the baseline predictions taking the mean patch predictions over each slide. We also extract features from this baseline ResNet model and use these to train the attention models: the ViT, the ClusterViT, the PREViT and the ClusterPREViT. We fit a K-Means clustering algorithm to the features to use as a cluster token embedding in the ClusterViT models. The PREViT models require additional information in the form of the patch coordinates.

attention networks to explain model predictions, and demonstrate visualizations of the clustered features used in one ViT model variation.

This chapter explains the methods explored and offers results and findings from these different approaches. The pipeline for this work is visualised in Figure 4.1. In Section 4.2 we introduce the methods used for the prediction models, including preprocessing steps, the baseline model for feature extraction, and the ViT model. We introduce the novel ViT model variations, PREViT and ClusterViT, explaining their model architecture and motivation. We discuss the outcome variables used in these prediction models, and give the details of all hyperparameters used in model training. In Section 4.3, we show the results from applying these methods. Finally, in Section 4.4, we discuss the results and conclude what work can be done following on from this. The work on predicting response to RT was published in the Workshop on Medical Image Assisted Biomarker Discovery (MIABID) at the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2022 conference proceedings [142].

4.2 Methods

First, we discuss the preprocessing steps required on the WSIs and provide an outline of the baseline model. Subsequently, we introduce the ViT model and how it is extended with our proposed PRE. We also introduce the ClusterVIT, where we specify how relevant morphological motifs are found through standard clustering methods. Naturally, we combine the two proposed methods into the ClusterPREViT model. We then detail the hyperparameters used and explore the clustering methods used, as well as provide a brief discussion on interpretability. An overview of the resulting imaging pipeline is shown in Figure 4.1.

4.2.1 Preprocessing

As seen in the Literature Review (Chapter 2), it is common practice in the medical imaging field to split the WSIs into patches, so that the size of the model input is not too large to fit into computer memory. We split all the available WSIs in the two patient cohort datasets, Grampian and Aristotle, into patches of size 256 x 256 pixels, with a stride (i.e. overlap) of 128 pixels. 256 is the size of a standard image and a commonly used patch size in the field of histology to capture a reasonable region of interest, and a stride is often used to avoid inducing fake boundaries into the inference. We use tumour masks provided by a pathologist to filter out unwanted tissue and background, excluding a patch if more than 50% of the patch does not contain relevant tissue as defined by the pathologist's mask. We then apply data augmentations to these images, to make the network more generalisable to data from other cohorts, in the following order: resize the images to 224 x 224 pixels; apply a random vertical flip with a probability of 0.5 and a random horizontal flip with a probability 0.5; apply a colour jitter with parameters brightness 0.1, contrast 0.05 and hue 0.1; randomly rotate the image by multiples of 90 degrees with equal chances of each; transform to a Pytorch Tensor and normalise with respect to the ImageNet mean and standard deviation ([0.485, 0.456, 0.406] and [0.229, 0.224, 0.225] respectively). This final step is done since our neural network is pre-trained

on the ImageNet dataset, as it is less computationally expensive to use transfer learning from a pre-trained model that has already seen some images and is familiar with extracting generic features such as edges. This set of transforms is applied to the training dataset. To the validation dataset we apply a reduced set of transforms which consists only of: resize the images to 224 x 224 pixels; transform to a Pytorch Tensor; normalise with respect to the ImageNet mean and standard deviation.

4.2.2 Baseline Model

Once the WSIs are patched, we can use weakly supervised learning to train a neural network using the slide-level outcome as a label for each patch in that slide. For all our experiments, we train a ResNet-50 [48] as a baseline model on the WSI patches, pre-trained on ImageNet. The ResNet is also known as the Residual Network model, and the version we are using, the ResNet-50, has 50 layers. The reason the ResNet network can successfully train with so many layers and such a deep network is that they use skip connections, which connect layers directly to others much deeper into the network [48]. These skip connections help avoid the vanishing gradient problem found in deeper neural networks [48]. The layers in the ResNet-50 model include convolution layers, identity blocks, batch norm layers, activation layers, pooling and a final fully connected layer to put the output into a shape the user requires.

The separate patch predictions from the baseline model can then be aggregated to a slide-level prediction using a method of MIL. As seen in the literature review, this can be done by simply taking the mean, or in some cases applying another neural network such as a recurrent neural network [59]. Alternatively, this can be done with an attention network, which aggregates patch predictions into a slide-level prediction by giving different weights or attention to what the network deems are the more important parts of the image for the prediction. For all models in this research we evaluate them initially using the mean-pooling MIL method of taking the mean across patch predictions in each slide for a slide-level prediction, and then use this as a baseline for comparing the attention models to. In addition,

we extract feature embeddings $\in \mathbb{R}^{2048}$ from the penultimate layer of the ResNet model for use in further models.

4.2.3 Vision Transformers

We focus on a particular type of attention network for this research, the ViT [53], which we introduce in in the Literature Review in Section 2.3.2. The ViT is well suited for this problem due to its ability to capture the local context surrounding each tissue region, in addition to longer range interactions, which is important for this prediction problem. The self-attention modules can 'attend' to each and every patch in the WSI through the matrix multiplications in the attention calculations, as seen in Section 2.3.2. However, the ViT is not built well for WSIs with regards to the position embedding, which according to the original implementation would be a fixed length learnable embedding, with a single position entry for each patch, that would have to be truncated for each slide depending on how many patches are present in that WSI.

4.2.4 Patch Restoration Embedding ViT (PREViT)

In the Literature Review in Section 2.3.2 we introduced the TransMIL model, which attempts to get around this position embedding problem for WSIs by proposing the PPEG module which uses convolution kernels to capture positional information.

However, as briefly mentioned in the Literature Review, the original PPEG module used in TransMIL for the position embedding of the patches aims to restore the 2D layout of the image, but this does not apply well when the patches originate from different parts of the WSI such as different parts of the biopsy tissue, and background patches have been excluded meaning the full 2D set of patches is not required. In this case, therefore, naively converting the input patches into 2D space is not contextually meaningful and arguably less so that keeping them in a 1D shape, since the approach does not consider the original position of the patches in the WSI.

We propose a novel patch Position Restoration Embedding (PRE), which restores the original position of the patch relative to the WSI, using zero padding for areas of the image which do not qualify for training [80], and then applies convolutions in the same manner as the PPEG module. The motivation for the patch restoration embedding is to incorporate local information around the patch, acting as a prior for the model representing the surrounding tissue environment.

4.2.5 ClusterViT

We propose a novel method for incorporating a cluster token into the ViT network, ClusterViT, which incorporates a learnable cluster token which is added to the input at the stage when the position embedding and classification token are added but before the combined input is passed to the main Transformer attention network, in a similar vein to the position embedding. This idea is inspired by Gao *et al.* [143], who add a nuclei grade embedding to their ViT model to capture prior information on the nuclei for subtyping of papillary renal cell carcinoma. We encode the cluster labels using an encoding from Gao *et al.* [143]. The aim of the cluster token is to consider the tissue morphology for each patch in the WSI, and use this as prior information for the model prediction. Each patch is assigned a cluster label, which is then embedded and concatenated to the input as a learnable cluster token.

4.2.6 ClusterPREViT

The ClusterPREViT model is a ViT model combining both the patch PRE and the cluster token from the ClusterViT model. The ViT classification token is appended to the input first, followed by the cluster token and finally the PRE.

4.2.7 Clustering Approach

To capture the different tissue motifs across the WSIs we perform a clustering analysis, using a K-Means clustering model on the feature embeddings from the baseline model. The purpose of the clustering is to detect the morphological patterns

that capture interactions between tissue types including stroma, epithelium and immune cells. K-Means is a classical clustering method, often chosen as a clustering method in the medical imaging field over other more complex clustering methods [68, 144]. The training set (used for training the baseline ResNet model) is used to estimate the parameters of the K-Means model which is then used to predict the cluster labels for the validation set of slides. The K-Means model is learnt on the L_2-normalised feature embeddings from the trained ResNet model, and the clustering model is trained across the whole training set (as opposed to per slide). The K-Means model is trained on the same training set as the ResNet that generates the features, and then the model is used to predict the cluster labels for the corresponding validation set. Each patch receives its own cluster label from the clustering model predictions, with the idea that the clusters detect the different tissue types within the WSI e.g. tumour stroma, epithelial tissue, muscle and surface tissue.

We visualise these clusters in different ways. Firstly, we perform a nearest neighbour analysis on the clusters, finding the patches which are most similar to the K-Means cluster centres in the Euclidean space. These patches can be visualised, giving an idea of which tissue types are dominant in the respective clusters. A clear distinction is seen between some clusters, but it's unclear whether these differences are biologically meaningful or rather purely visual differences in the images. An example can be seen in Figure 4.4, and we discuss the clusters more in the Results in Section 4.3.4.

To choose the number of clusters, k, to group the tissue patches into we consult the imCMS paper by Sirinukunwattana *et al.* [2], since they use these very same images to detect the CRC CMS classification and perform clustering analyses to explore the model results. Since there are four CMS classifications, each with marked morphological correlation, we fit one K-Means algorithm with $k = 4$ clusters. The clustering analysis performed by Sirinukunwattana *et al.* [2] found an optimal number of 13 clusters could be applied on their feature set to differentiate across the CMS classes. Hence we also explore fitting a K-Means algorithm with $k = 13$ clusters.

The clusters can also be visualised for each slide, superimposed on the WSI showing which parts of the tissue belong to different clusters. Some cluster labels may be more prominent in some WSIs more than others, which could indicate a stronger presence of certain types of tissue in these images, and in turn this could indicate a patient's response to RT. Hence, we use the cluster labels as a cluster token in the ClusterViT model variation, to try to incorporate this contextual tissue information.

4.2.8 Interpretability

The benefit of using attention models is that they can help us to interpret the slide-level prediction by visualising the weights given to different tissue patches in the image. The weights are between 0 and 1, and so using continuous colour maps we can colour higher and lower weights differently, which provides a nice visualisation on the individual slide level. These visualisations would be available for all predictions of the model, which could be informative for clinicians, to judge whether the parts of the tissue highlighted as important for the attention model prediction are indeed biologically meaningful. An example can be seen in Figure 4.2.

4.3 Experiment Results

Thus far, we have proposed various models to predict both the RSS gene score from the images, and the patient's response to RT. The same underlying neural networks can be used to predict these two outcomes, but considering the nature of the outcomes, in that one is a binary classification prediction and the other a continuous regression problem, a few modifications are required in the training process.

The proposed methods are demonstrated on two large independent rectal cancer datasets, Grampian and Aristotle, consisting of histology slides from patients selectively treated with RT and capecitabine in two UK clinical trials. Further details can be found in Sections 3.1.1 and 3.1.2.

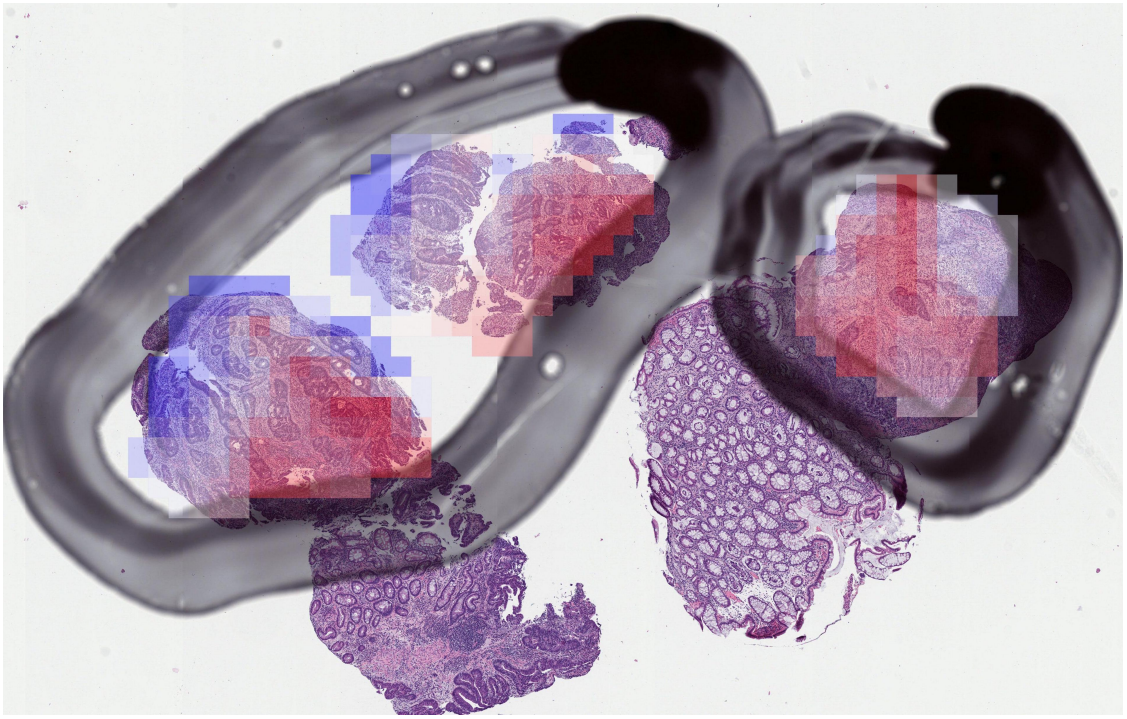


Figure 4.2: A single histology slide from the Aristotle validation dataset with a heatmap overlaid onto the image, showing the attention scores from the binary PREViT model predictions for this slide, where red means higher attention values and blue means lower. The black pen marks are drawn on the original biopsy slide to select the tumour tissue which should be sampled for gene sequencing. Each patch in the slide (within the tumour mask) has its own attention value, hence the heatmap appears slightly pixelated. The patient from which this biopsy was taken had a complete response to radiotherapy, and the model predicted they would have no complete response to radiotherapy.

4.3.1 Model Hyperparameters

In the baseline ResNet training process, the final layer of the ResNet model is set to return a single value, using a linear layer with 2048 features as input and one feature as output. For all models shown in the Results in Section 4.3 we use the patched images at 10x magnification (1 μm^2/pixel), but we have also explored using 5x and 20x magnification in other models. We use a training and validation split of 70% training vs 30% validation, splitting the data such that all slides from a single patient are in the same dataset, whether that be training or validation. We use stochastic gradient descent as the optimiser function, and binary cross entropy (BCE) as the loss function, with a learning rate of $1e-3$, momentum of 0.9 and weight decay of $1e-4$. The models train on a batch size of 256 patches. All ResNet

models are trained for 20 epochs, allowing sufficient time for training and validation losses to converge without using unnecessary resources and power.

In the ViT training process, the training and validation datasets from the ResNet training are conserved and used in the same manner for the ViT training. Again, all inference shown in this report is from models trained on patches at 10x magnification, unless specified otherwise. Again we use stochastic gradient descent as the optimiser function for the ViT models, with a learning rate of $1e-3$, momentum of 0.9 and weight decay of $1e-4$. The model trains on a single WSI at once, using the feature embeddings from all patches in a slide as input. In the case when the number of patches per slide is unusually large, for computational reasons we randomly select 10,000 patches from the slide to train or validate on. All ViT models are trained for 50 epochs, since they are quicker to train in comparison to the ResNet models. For all the ViT models (including the PREViT model, with the patch restoration position embedding instead of the learnable parameter position embedding, and the ClusterViT model) we set the number of classes to 1, the internal dimension to 512, the patch dimension to 2048, the depth to 4, the number of heads to 4, the dimension of the MLP to 512, the dimension of each attention head to 64, the dropout to 0.3 and the embedding dropout to 0.3.

To predict the response to RT we use the BCE loss, and to predict the RSS gene score we apply either the mean squared error (MSE) loss or shrinkage loss in training.

4.3.2 Predicting RSS

The models described in the Methods in Section 4.2 are used to try to predict the 33 gene RSS score from the digital histology WSIs. The gene score is scaled to lie between 0 and 1, preserving the original distribution. As a baseline model, we train a ResNet-50 CNN to predict the normalised RSS gene score. This is therefore a regression problem, so the final layer of the ResNet model is edited to return a single class prediction, and then a sigmoid function is applied to the model predictions to scale the output between 0 and 1. The separate patch predictions

ResNet-50 for RSS	Mean Absolute Error	Pearson Correlation
Grampian	0.100	0.239
Aristotle	0.159	0.327

Table 4.1: Validation results from the best epoch of two ResNet-50 models trained to predict the normalised RSS gene score. One model is trained and validated on the Grampian dataset and the other model is trained and validated on the Aristotle dataset. The mean absolute error (MAE) and the Pearson correlation are evaluated between the true RSS values and the model predictions of the RSS values. A good model would have a MAE closer to 0, and a Pearson correlation closer to 1. From these results we can determine that the ResNet-50 model cannot predict the normalised RSS gene score very well.

are aggregated to a slide-level prediction by taking the mean RSS score across the patch predictions for each slide.

One baseline model is trained on the Grampian dataset only, by splitting the Grampian dataset into training and validation sets. Another baseline model is trained only on the Aristotle dataset, by similarly splitting the Aristotle dataset into training and validation sets. The validation results from the epoch with the best validation metrics for both of these models are found in Table 4.1. The mean absolute error (MAE, also known as the L1-loss) is evaluated between the true RSS values and the model predictions, and the Pearson correlation is also evaluated between the true RSS values and the model predictions of the RSS values. A good model would have a MAE closer to 0, and a Pearson correlation closer to 1. These results do not convince us that the RSS gene score can be predicted from these images with our ResNet baseline model approach.

In addition to this baseline model, we experiment with fitting some attention models to the feature embeddings extracted from the penultimate layer of the ResNet model. Each image patch is expressed as a vector of feature values $\in \mathbb{R}^{2048}$. We experimented with running this pipeline with WSIs at different magnifications, and with our different ViT model versions. We find that the ClusterPREViT model on WSIs at 10x magnification gives the best results here, as seen in the results table in Table 4.2. We also provide the results for the best model from the 5x and 20x magnifications, which was the standard ViT model. Now instead of separating

Magnification	Model	Mean Absolute Error	Pearson Correlation	Spearman Correlation
5x	ViT	0.154	0.239	0.223
10x	ClusterPREViT	**0.151**	**0.411**	**0.385**
20x	ViT	0.159	0.271	0.243

Table 4.2: Validation results from the best epoch of the VIT models using a baseline ResNet-50 model trained to predict the normalised RSS gene score from WSIs at different magnifications. The mean absolute error (MAE), the Pearson correlation and Spearman's rank correlation are evaluated between the true RSS values and the model predictions of the RSS values. A good model would have a MAE closer to 0, and correlations closer to 1. From these results we can determine that our we cannot predict the normalised RSS gene score very well from the histology slides with the approaches explored so far.

the cohorts, we use both Grampian and Aristotle in training, using roughly 70% of each in the training set and the other 30% of each in the validation set.

For all of our best models for RSS prediction, we train them using shrinkage loss, a loss function developed specifically for training deep regression models on imbalanced datasets [145]. It works by penalising the weighting of predictions from data that's easy to train on, giving heavier weighting to the harder predictions.

Applying the ViT model and its novel variants to this data does not result in a drastically improved prediction for the RSS gene score. From these results we can determine that our deep learning models cannot predict the normalised RSS gene score very well with the approaches explored so far.

It should be acknowledged, however, that directly predicting the gene signature from the images is not guaranteed to work, since there is some level of abstraction between the images and the RSS gene score. All of the 33 genes used in the calculation of the RSS gene score are not necessarily reflected in the tissue morphology in the image, and therefore could not reasonably be predicted from the image. In fact, to predict the gene signature successfully, a certain degree of tissue segmentation would be required, since not all pathways are active in each of the various tissue components. Fisher *et al.* explore this concept in their work on gene signature scores, where they find that small changes in the amount of stromal

tissue in a biopsy sample, as could be observed in a histology slide, can directly influence the value of gene expression signatures [146].

This point is exemplified in our datasets. The RSS gene expression score is calculated from 33 gene expressions taken from the biopsies for which we have digital images. The gene expressions are extracted from the tumour tissue regions, which have been circled in black marker pen on the biopsies, which are then scanned to get the digital images. Hence, we can see the tissue from which the RSS score is calculated. In the Grampian dataset, this is the same tissue that we use for training our models on. However, in the Aristotle dataset, there is tumour tissue that we use from the images for our models that was not used to calculate the gene scores of the tissue. Hence, for the prediction of the RSS gene score in Aristotle, this could add a lot of noise to our dataset, because the ground truth gene score for some tissue we are learning on could be different from what the gene score is for the rest of the tissue in the biopsy slide.

4.3.3 Predicting Response to Radiotherapy

The deep learning models described in the Methods in Section Section 4.2 are trained on our data to directly predict a patient's response to RT from the digital histology WSIs. Both datasets, Grampian and Aristotle, are split across the training and validation sets, and so both datasets are used in training and both datasets are used for validation, but with no overlap between the training and validation sets. The response to RT can vary across patients, and the classifications used to define this response varies across both of our datasets. However, across both datasets we are able to classify each patient as either having a complete response (CR) to RT, or no complete response (NoCR), where such data exists.

Due to the imbalanced distribution of the response data, we experiment with up-sampling the less prevalent class (CR to RT) in the training set so that there is a roughly equal distribution of complete responders and non-complete responders to RT. We also experiment with multiplying the loss by class weights to give more

weight to the minority class during training. These methods prove to be effective in improving the classification of the underrepresented class.

As mentioned previously, as a baseline model, we train a ResNet-50 CNN to predict the binary response to RT, encoded as either 0 (NoCR) or 1 (CR). The final layer of the ResNet model is edited to return a single class prediction, and the sigmoid function is then applied to the model predictions to scale the output between 0 and 1. This output is then rounded up or down to get an exact binary classification, per patch. The separate patch predictions are aggregated to a slide-level prediction by taking the mean across the patch predictions for each slide. On top of this, we train attention models on the features extracted from the predictions from the penultimate layer of this ResNet model.

We run the full model pipeline for five rounds, using different random seeds and different data splits across the datasets, determined by the random seed. All models are trained and validated on different training splits containing both the Grampian and Aristotle datasets together. The metrics used to evaluate this binary classification are the area under the curve (AUC), balanced accuracy, F1 score, precision and recall, all weighted by class-balanced sample weights due to the dataset imbalance. The results presented in Table 4.3 are the metrics on the validation set from the five rounds, including standard deviation scores in brackets. For each round, we choose the epoch with the best validation metrics. The default threshold of 0.5 is used for the metrics which require a binarised prediction. This provides a valid technical comparison, but specific metrics and thresholds for clinical translation could be explored further. The AUC is used as the primary metric here since it is more discriminating and consistent than the accuracy score [147]. These results demonstrate that it is possible to predict a patient's response to RT from these histology images, which opens up possibilities for our later work in the coming chapters.

In terms of our primary metric the best performing model is the ClusterPRE-ViT (with k=4 clusters), with a weighted AUC of 0.861 over five rounds and the lowest standard deviation in AUC. The PREViT and ClusterViT models

Model	Mean AUC (std)	Accuracy	F1 Score	Precision	Recall
ResNet	0.696 (0.116)	0.696	0.787	0.802	0.788
ViT	0.857 (0.017)	0.725	0.823	0.829	0.829
PREViT	0.859 (0.021)	0.754	**0.834**	0.844	**0.836**
ClusterViT	0.859 (0.028)	**0.767**	0.781	**0.845**	0.767
ClusterPREViT	**0.861** (0.013)	0.763	0.782	0.843	0.768

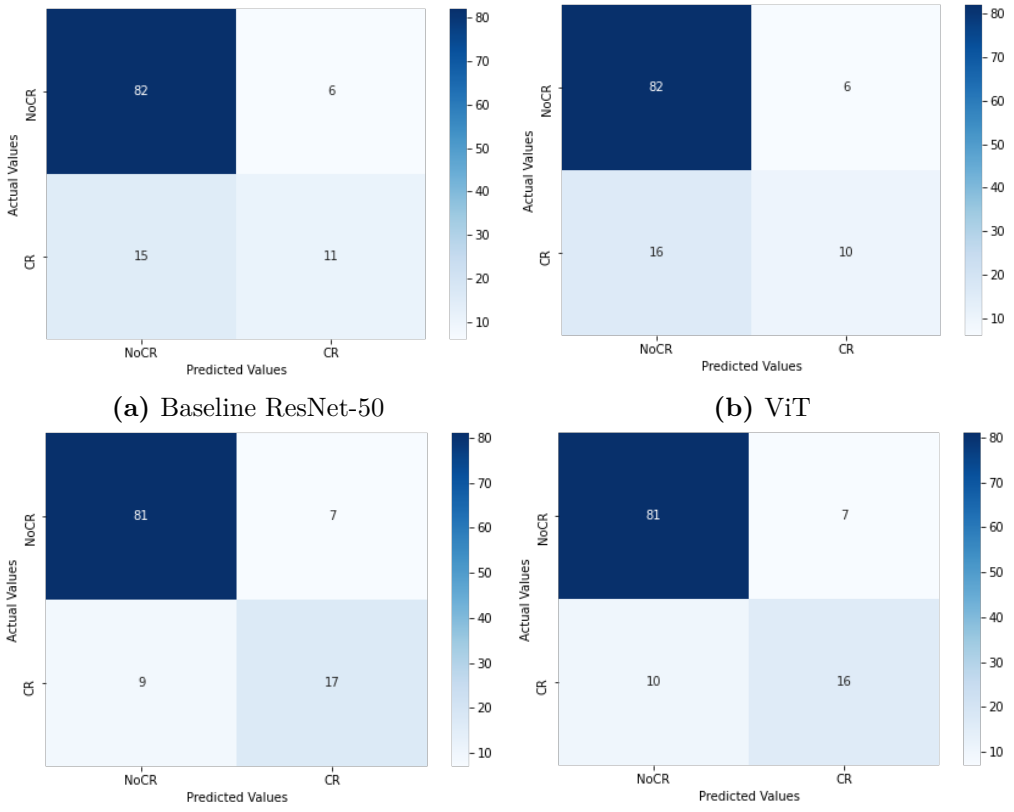
Table 4.3: Validation results from the best epoch of the models trained to predict the binary response to radiotherapy. Each model was run for five rounds using different random seeds and different data splits across the two combined datasets determined by the random seed. The mean weighted area under the curve of the receiver operating characteristic curve (AUC) and standard deviation (std) over the five rounds are presented as the primary metric, as well as accuracy, F1 score, precision and recall, all weighted by class-balanced sample weights. The best values for each metric are highlighted in bold font. Note results are given for the ClusterViT models with $k = 4$ clusters.

perform marginally better than the ViT, but all show a substantial improvement in performance over the baseline ResNet model, across all measured metrics.

The results for the models predicting response to RT are visualised in Figure 4.3 in confusion matrices from the validation set results from the best validation epoch for each model. From these results it is fair to interpret that the PREViT and ClusterViT give better results than the baseline ResNet-50 and ViT models.

4.3.4 Visualising Clusters

To visualise the clusters generated from the baseline ResNet-50 model features, we fit a nearest neighbours algorithm to the features from the training set of the model to predict response to RT. The 100 nearest neighbours (most similar patch features) to the cluster centres from the K-Means clustering model can then be found. The 100 nearest neighbours for the four clusters fit on the features from the baseline model to predict response to RT are seen in Figure 4.4. The four sets of images correspond to the nearest patch neighbours of the four cluster centres. The first cluster picks up on paler tissue with more empty space in the patch. The second cluster contains less white space, and the third cluster detects darker tissue staining. It is unclear from these clusters whether anything biologically meaningful is detected,



(a) Baseline ResNet-50

(b) ViT

(c) PREViT

(d) ClusterViT

Figure 4.3: Confusion matrices of the model predictions of (a) the baseline ResNet-50 model, (b) the ViT model, (c) the PREViT model with the patch restoration position embedding and (d) the ClusterViT with the additional cluster token. The counts in the confusion matrices are given on the validation sets from the model training process. The label 'NoCR' corresponds to no complete response, and the 'CR' label corresponds to complete response to radiotherapy. The model has performed well if most counts are in the top left hand box and the lower right hand box within the confusion matrices, as this means these slides have been predicted correctly. We can see that the PREViT model and ClusterViT model perform the classification better than the ResNet-50 and ViT model.

such as different tissue regions, or whether the clusters are based more on the other image features such as the colour of the tissue or the amount of empty space.

We also visualise the cluster labels on a single slide from the Aristotle dataset, seen in Figure 4.5. This figure shows four cluster labels represented by colours overlaid on the WSI. In this image it appears that two of the clusters contain mostly surface tissue, and the other two clusters could contain the regions with a higher density of tumour tissue.

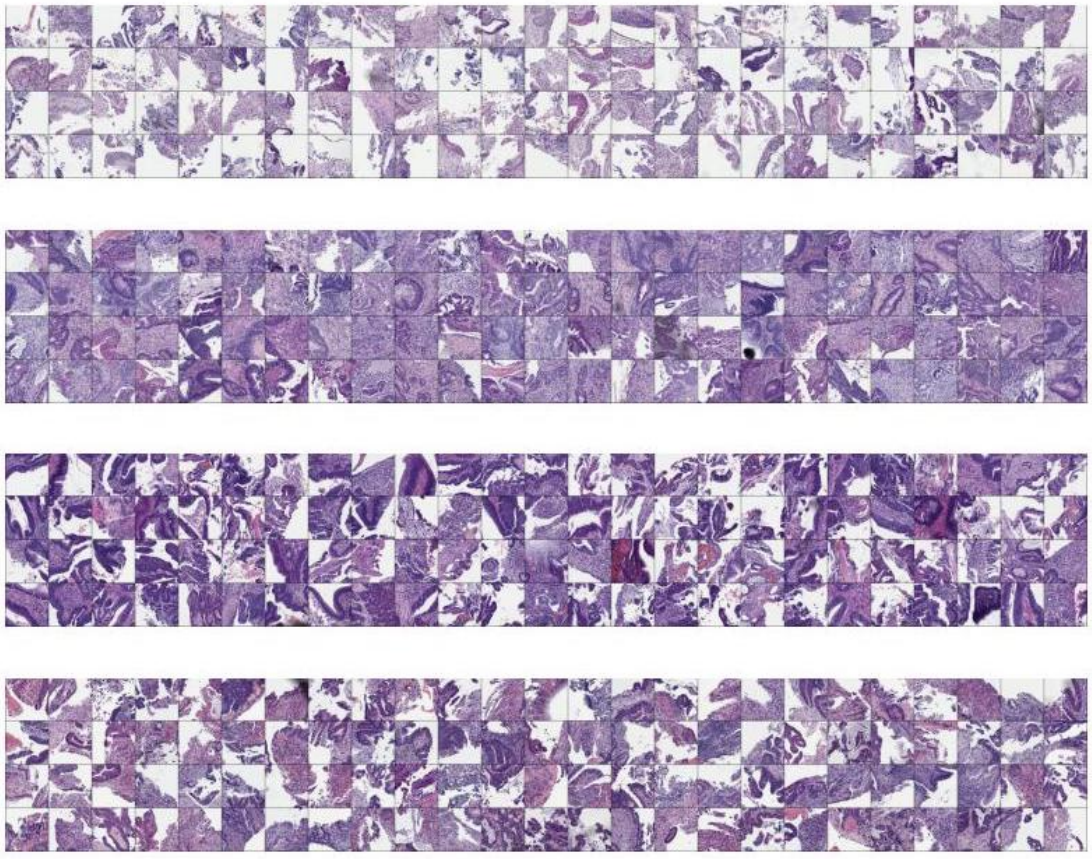


Figure 4.4: Each group of patches seen in this figure represents the 100 nearest neighbours to a different cluster centre. For each WSI patch, features are extracted from the trained baseline ResNet-50 model for the training set, and a K-Means clustering model is fitted to the features to categorise the patch features into four clusters. A nearest neighbours algorithm is then fit on the training feature set, and the 100 nearest neighbours are found to the four cluster centres from the K-Means model. It is unclear from this nearest neighbours visualisation whether anything biologically meaningful is detected in the clusters.

4.3.5 Attention Heatmaps

The benefit of both the clustering and position embedding approaches are that they provide more insight into the model and the model predictions. The clustering assigns a cluster label to each patch which can be visualised on the WSI as prior information for the attention model, and the PREViT model provides particularly informative attention heatmaps. Both approaches provide clues to morphological continuity in a slide which suppresses the ViT models from attending to small spurious regions and encourages the attention towards regions with morphological

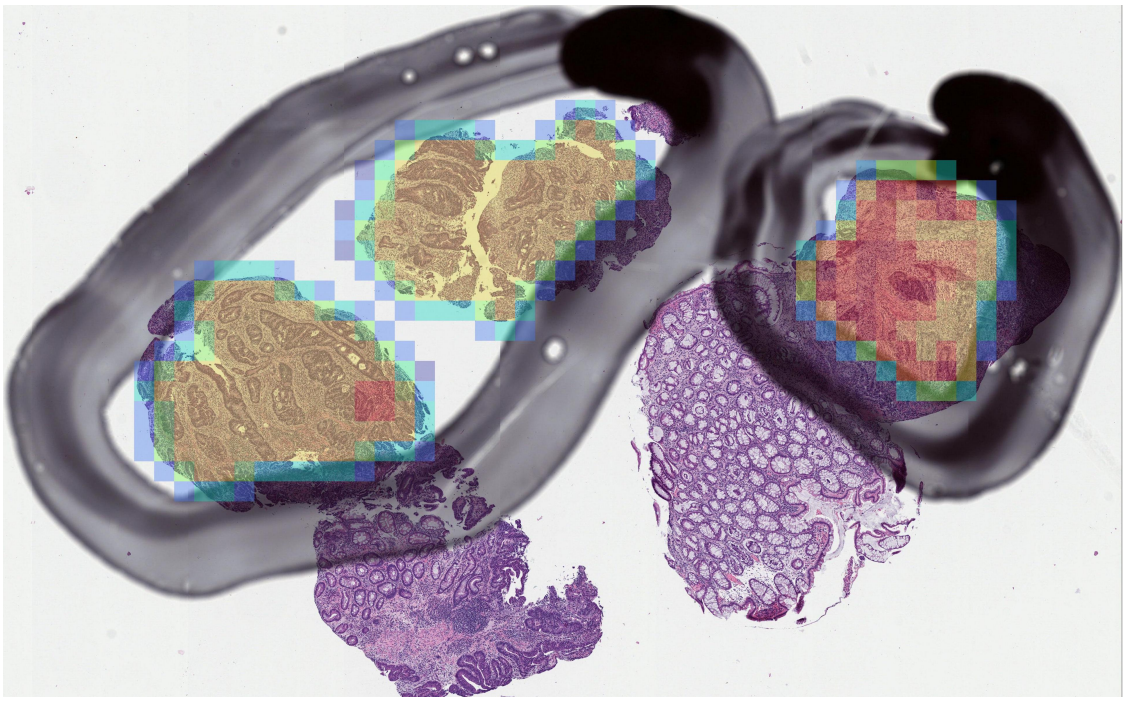


Figure 4.5: This figure shows a histology WSI from the Aristotle dataset, used for validation only in the model training process. The black pen marks are drawn on the original biopsy slide to select the tumour tissue which should be sampled for gene sequencing. On this digital WSI we have overlaid a heatmap, showing the cluster labels from the K-Means model fitted on the features extracted from the binary baseline ResNet model. Each patch in the WSI (within the tumour mask) has its own cluster label, hence the heatmap appears slightly pixelated. There are four clusters seen in this image, with each represented by a different colour. The order or specific label of the clusters does not matter since they are interchangeable. The outer edges of the tissue seem to mostly belong to two of the clusters, and the more internal tissue patches fit into the other two clusters. The patient from which this biopsy was taken had a complete response to radiotherapy.

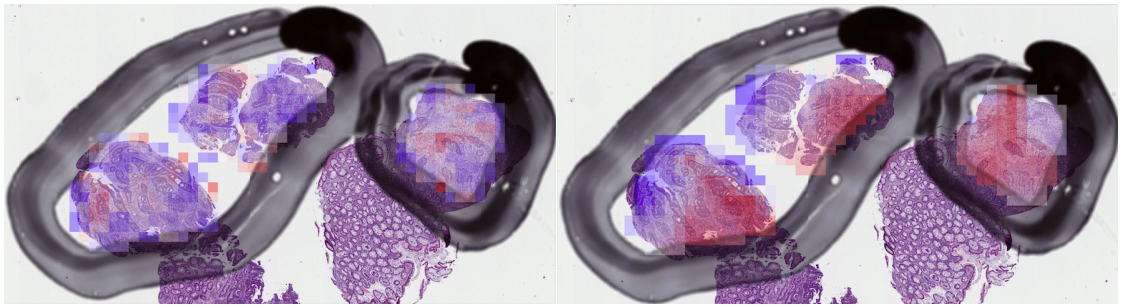


Figure 4.6: Heatmaps showing the attention weights from the model predictions for this slide. On the left is the heatmap from the ViT and on the right is the PREViT heatmap. A pathologist reviewing these heatmaps observed that the PREViT model demonstrates more attention to areas of invasive cancer, and less attention to non-informative artefacts.

meaning. As such, both approaches increase the interpretability of the heatmaps of the attention weights from the ViT model predictions. Figure 4.6 shows the comparison of a heatmap of the PREViT model attention weights for one slide (right) against the ViT model attention weights (left). Patches with higher attention from the model are coloured in red, and patches with lower attention are coloured in blue.

Another example of an attention heatmap on a histology slide is shown in Figure 4.7. This slide is from the Aristotle dataset, used in the validation part of the model training and validation process. Both heatmaps show that the PREViT model seems to give more attention to the biologically relevant parts of the tumour tissue, and less attention to the less informative parts of the tissue such as the surface tissue. These heatmaps can provide feedback to pathologists for a visual interpretation of the model results, though further exploration and model optimisation is required to determine whether the attention heatmaps could be informative in a clinical setting.

4.3.6 Odds Ratios

In order to demonstrate the clinical relevance of our prediction, we carry out a multivariate logistic regression for the validation set of each round in order to determine the odds ratio that a patient responds to RT. In this model, we find that the prediction made by the ClusterPREViT is the most predictive covariate, since the confidence intervals for the estimated effect are the furthest away from zero. This holds in all five rounds when compared to T stage, N stage, age and gender as shown in Figure 4.8.

4.4 Discussion & Conclusion

Predicting CR to RT in rectal cancer patients using deep learning approaches from morphological features extracted from histology biopsies provides a quick, low-cost and effective way to assist clinical decision making. The proposed extensions to the ViT framework to improve the utilisation of contextual information present in WSIs demonstrate the potential of predicting response to RT from features

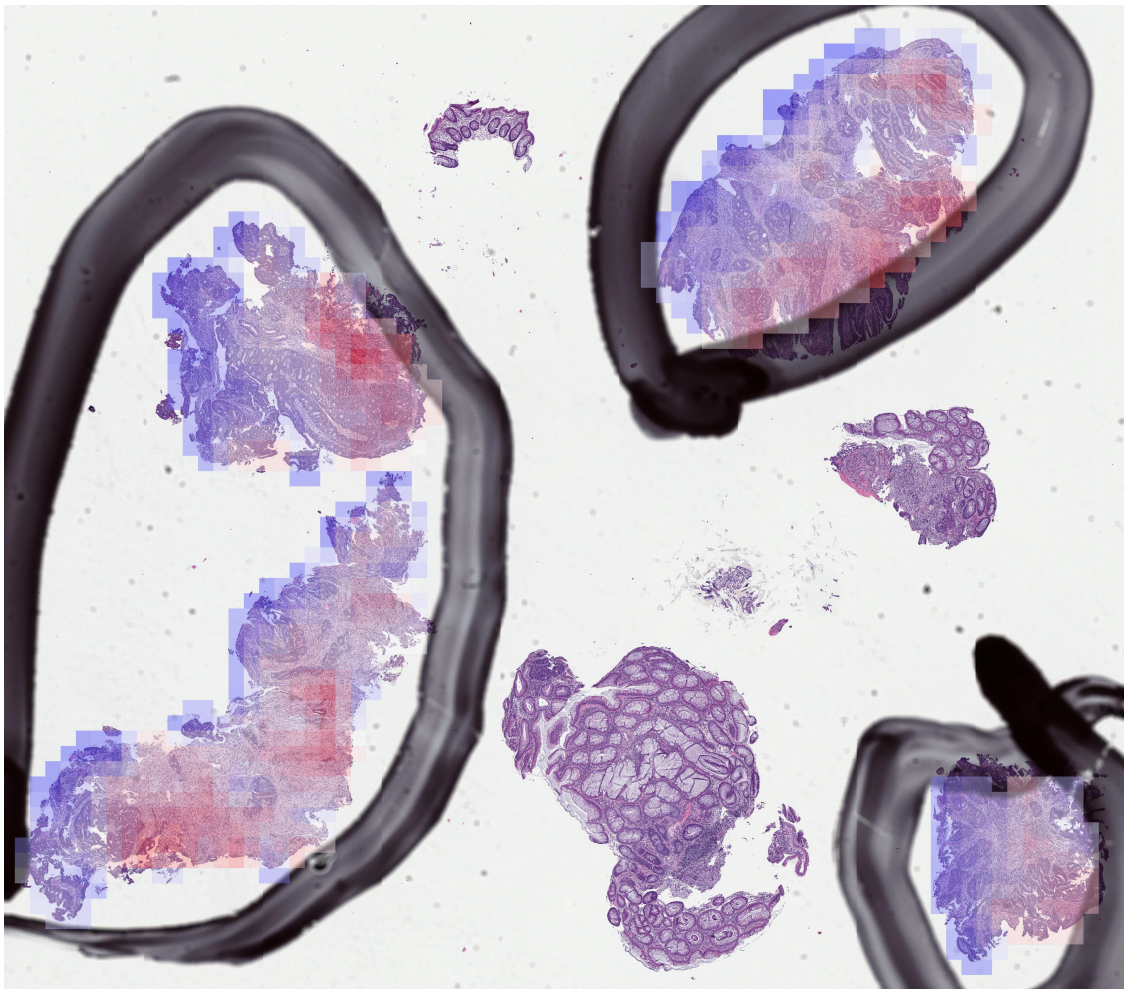


Figure 4.7: This figure shows a histology WSI from the Aristotle dataset, used for validation only in the model training process. The black pen marks are drawn on the original biopsy slide to select the tumour tissue which should be sampled for gene sequencing, and these pen marks also inform the digital tumour mask that is applied to all WSIs when patching. On this digital WSI we have overlaid a heatmap, showing the attention scores from the binary PREViT model predictions for this slide. Each patch in the WSI (within the tumour mask) has its own attention value, hence the heatmap appears slightly pixelated. Red patches indicate a higher attention score, and blue patches indicate a lower attention score. The heatmap shows that the PREViT model seems to give more attention to the biologically relevant parts of the tumour tissue, and less attention to the less informative parts of the tissue such as the surface tissue. The patient from which this biopsy was taken did not have a complete response to radiotherapy treatment, and the model correctly predicted they would have no complete response to radiotherapy.

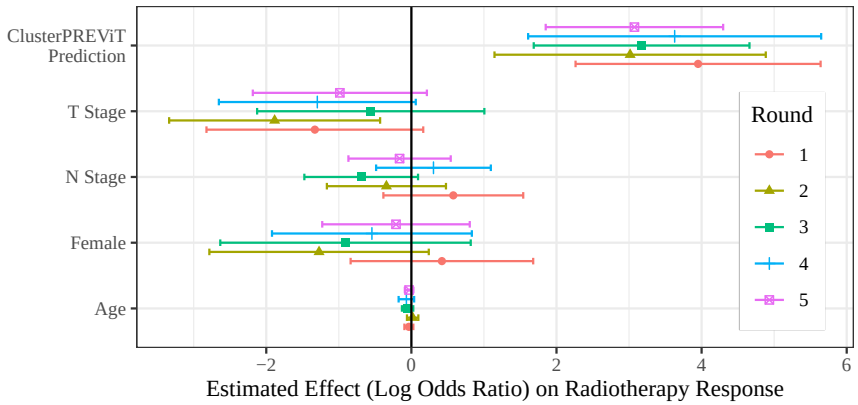


Figure 4.8: Log odds ratios for covariates in a logistic regression model predicting complete response to radiotherapy. The most predictive covariate is the ClusterPREViT model prediction when compared to T stage, N stage, age and gender. The validation set for each round was used in a separate model, hence five results for each covariate.

extracted from standard H&E slides. As such we are presenting a new and exciting application of computational pathology.

Expanding the ViT model in the proposed manner effectively enhances performance of the prediction over baseline models. Not only this, but using our position restoration embedding (PRE) we can preserve the spatial relationship between tissue patches, using their original positions on a WSI. Furthermore, our clustering analysis of extracted tissue features explores morphological motifs which capture fundamental biological processes found in the tumour microenvironment, introduced into the ViT network in the form of a cluster label token, helping the model to differentiate between tissue types. This clustering enhances the ability to provide visual feedback which ultimately makes our approach more usable in clinical translation.

Thresholds for binary predictions

For the binary model, predicting CR or NoCR, the model outputs a prediction between 0 and 1, and it is then rounded to get the binary outcome of 0 or 1 (i.e. NoCR or CR). Thus far we have been using a threshold of 0.5, but considering the distribution of model predictions it seems it would be worthwhile to explore another way of defining a different threshold which could better differentiate the two class predictions, particularly considering the imbalance in the dataset.

ViTs overfitting

Though the ViT models have worked reasonably well here, when training these models and observing the losses over multiple epochs, it seems that they are overfitting to the training data and hence performing less well on the validation data as the epochs increase. This is a known problem with ViTs since the number of parameters in these models is very large, as there are fewer inherent biases as compared to CNNs and therefore more flexibility which translates to more parameters. We could address this issue using traditional approaches to combat overfitting, such as by adjusting the built in dropout module of the ViT network, which we can change to be more severe to add more noise into the training process. Alternatively, we could consider other models asides from ViTs which have fewer parameters and therefore may be more suitable to this dataset.

Morphology is not square

Furthermore, the approach here is limited by the definition of the patches. The visualisations of the square patches can be hard for clinicians and pathologists to interpret, since there could be highly heterogeneous tissue within a single patch, and we would not be able to say which portions of this patch determined the overall patch attention score. Additionally, utilising the morphology of different tissue components could improve model predictions as well as interpretations.

Domain shift

Considering these results so far, we suspect a shift in the domain between the Grampian and Aristotle datasets. The RSS score is quite different between the two datasets, with Aristotle having a much larger tail in the lower end of the distribution. We confirm the statistical difference between the two cohort distributions of RSS using a Mann Whitney U test, which achieved a p-value of 3.45e-14. Despite this, clinicians with expert knowledge of these trials claim there should be no underlying difference between the two cohorts. However, simply the fact that the data comes

from two different hospitals in different parts of the country suggests there could be some unintentional domain shift between the two datasets.

Response instead of RSS prediction

In addition to the current downsides of the RSS gene score, that it can vary depending on which tissue regions within a biopsy sample are selected for RNA sequencing as discussed in Section 4.3.2, another limitation is from the range of the RSS score we have in our datasets. For the RSS prediction, we scale the RSS to lie between 0 and 1 before training on it, as is common in deep learning regression models. However, by doing this we are restricting the possible range of the RSS score that we can predict in future, defined by the minimum and maximum RSS values of the data we are currently dealing with. Based on the way RSS is calculated, essentially a sum of gene expression values, there is no guarantee that the range of RSS we have for the current data is exhaustive, and it is possible that RSS values could exist outside this range for other tissue biopsies. On the other hand, this also means that a prediction model wouldn't be extrapolating its predictions outside of the range of RSS that it has learnt on. Such predictions would be based on unverified observations, and so there would be less confidence in these predictions.

The results from this research so far indicate that it is possible to gain useful insight about a patient's predicted response to treatment from the biopsy images. Based on the work explored so far, it seems that the binary response to RT outcome is more easily detected in the images by our deep learning models than the RSS gene score. Hence, going forwards we focus more on predicting this binary response, which is a representation of the patient's true response to treatment, rather than the gene score derived from a partial sample of tissue.

5

Interpretability with Molecular Traits and Spatial Organisation

Contents

5.1 Introduction 83

5.2 Methods 84

5.2.1 Feature Extraction 85

5.2.2 WSI Graph Design 86

5.2.3 Data 89

5.3 Experiments 90

5.3.1 Superpixels 90

5.3.2 Graph Connectivity 93

5.3.3 Implementation 96

5.4 Results 99
5.4.1 Node-level Epithelium 100
5.4.2 Visualisation 101
5.4.3 Slide-level Epithelium Proportion 102
5.4.4 Balancing Across Cohorts 104
5.4.5 External Test Set 107
5.4.6 Ablation Studies 108
5.5 Analysis of Gradients 109
5.6 Conclusion 113

Contributions

Existing methods for interpretability of model predictions are largely based on technical insights and are not linked to clinical context. We use the question of predicting response to radiotherapy in colorectal cancer patients as an exemplar for developing prediction models that do provide such contextual information and therefore can effectively support clinical decision making. There is a growing body of evidence that about 30% of colorectal cancer patients do not respond to radiotherapy and will need alternative treatment. The consensus molecular subtypes (CMS) for colorectal cancer provide one such approach to categorising patients based on their disease biology. Here we select the CMS4 subtype as a proxy for stromal infiltration. By jointly predicting a patient's response to radiotherapy, the presence of CMS4, and the epithelial tissue map from morphological features extracted from standard H&E slides we provide a comprehensive clinically relevant assessment of a biopsy. A graph neural network is trained to achieve this joint prediction task, which subsequently provides novel interpretability maps to aid clinicians in their cancer treatment decision making process.

Sections of this work have been published in the proceedings for the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2023 conference [148].

5.1 Introduction

Histology-based digital biomarkers enable the possibility to predict a patient's response to therapy. However, as opposed to predicting response to radiotherapy (RT) alone, as we did in the previous chapter, we now aim to analyse this prediction in the context of the overall tissue architecture and the tumour biology as captured by CMS. In our Literature Review in Chapter 2 we presented research demonstrating the prognostic effects of stromal infiltration and spatial organisation of the epithelial tissue, and we develop our method here using these prognostic factors for interpretability.

Specifically, we use a multi-task learning approach as a method for interpretability. Instead of predicting response to RT alone, we use three separate classification layers in the final stage of the model to output three predictions, instead of one. In addition to response to RT, we also predict the presence of the CMS4 CRC molecular subtype, which we use as a proxy for stromal infiltration, and the presence of epithelial tissue to represent the spatial organisation of the tumour. We argue that this not only guides the therapy response prediction, but also provides contextual visualisations for better interpretability of the prediction.

Our previous approach using square patches was motivated by the need to overcome the memory limitations of existing GPUs, since the input to our model is a large H&E WSI. To achieve our goal of predicting response to RT in context of molecular traits and spatial organisation, we need to capture the heterogeneity at the slide level, by only considering meaningful neighbourhoods surrounding each tissue section, which is why the ViT is too flexible a model for this problem. Furthermore, applying full or semi-supervised approaches on individual tiles followed by a MIL aggregation method is not suitable since the interactions between neighbourhoods are not considered.

Instead, we build on recent graph neural network (GNN) approaches that allow us to model the entire WSI as a graph of connected neighbourhoods. Using square patches limited the interpretability value of each fine-grained prediction, and so here we use a different approach, using segmented tissue regions as the basis for our

model input. These local cell communities can form the nodes of our WSI graph, meaning we can effectively model the micro-anatomy of the tissue. At the same time it is possible to make predictions at the node-, graph-, and slide-level.

Our methodology proposes a novel and disease relevant approach to a more interpretable model that effectively supports a diagnostic task. Pathologists and oncologists can use this information to inspect the validity of the prediction result and interrogate key aspects of the spatial biology that is critical for patient management. Ultimately, this type of information that is not available today will help to characterise interactions between the tumour and the host tissue and therefore help to support choice of therapy. The initial framework combines self-supervised training of a Vision Transformer (ViT) to extract morphological features, a superpixel algorithm for determining nodes of a graph, and a GNN for predictions. We demonstrate how different methodologies affect performance at various stages of model development. After further optimising the model, we achieve 0.86 AUC predicting complete response (CR) to RT using deep learning on WSIs for CRC patients, whilst providing novel interpretability of the results.

The following work was accepted for publication as part of the MICCAI 2023 conference proceedings, and selected for an oral presentation at the conference. We integrate the paper's original supplementary materials into this chapter, as well as further research done since publication. We tend to refer to the published version of this work as the initial version, and the following version as the optimal or optimised version.

5.2 Methods

In this section we present the patch-level feature extraction, provide the detail of the superpixel segmentation of the WSI, and illustrate the resulting graph representation. A GNN with three branches for our output predictions is used to simultaneously make the three different predictions as shown in Figure 5.1.

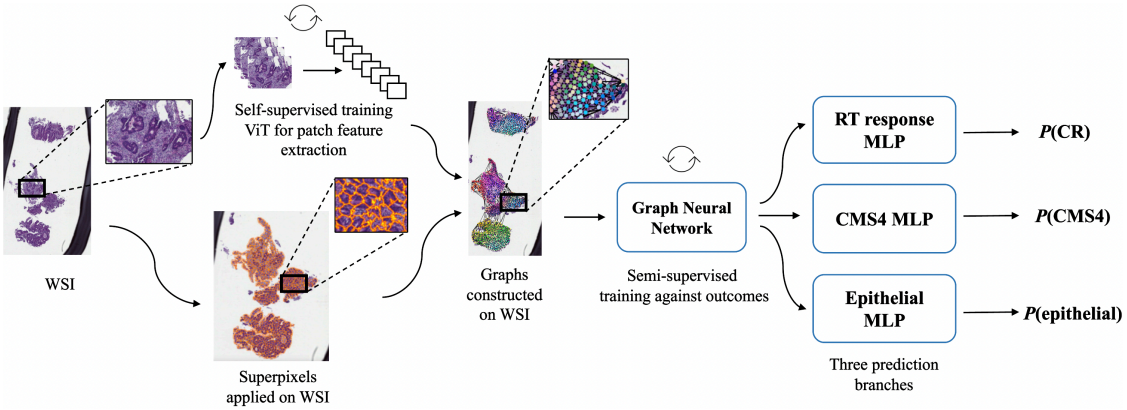


Figure 5.1: Approach We extract patch-level features from each WSI using self-supervised DINO training with a ViT model [114]. The SLIC superpixel algorithm segments the entire slide into smaller regions [149]. We calculate the mean patch features for these superpixel regions, and use the superpixel features and centers as our graph nodes, applying Delaunay triangulation to generate the edges of the graph. A GNN consisting of GINConv layers is trained on these fixed graphs, and the final layer splits into three separate MLP branches to provide predictions of three different outcomes, complete response (CR) to radiotherapy (RT), CMS4 classification, and epithelial tissue. An example output is visualized in Figure 5.8.

5.2.1 Feature Extraction

For computational reasons, all images are split into patches of size 256 x 256 pixels. In order to have a common feature set all the way up to the last layer of the GNN, individual patches should be represented by morphological features that are label-agnostic. This last layer of the GNN then splits into three branches to predict response to RT, the CMS4 subtype classification for CRC, and epithelial tissue regions. This way we can guarantee the common latent features and derivation across branches, maintaining the contextual importance of each branch.

We use patches which overlap with a stride of 50%, or 128 pixels. We also experiment with using no overlap and setting the stride window to 256, the same size as the patch itself. However, we find better performance in general when using a stride of 128 pixels, since this increases the quantity of data and balances the influence of boundary artifacts.

The DINO framework [114] is a method of self-supervised training for feature encodings, which implements a self-distillation training approach, using random data

augmentation and transforms to locally crop the input patches and train with a local-global student-teacher approach. The method uses cross entropy loss to compare the predictions of the student against the teacher's, and the weights of the teacher model are constantly updated with the exponential moving average of the student's weights. We use the DINO framework to train a ViT in a self-supervised manner on our H&E slides [150], representing each patch with 384 features. We use only the training dataset to train this model, and use the image patches at 20x magnification.

We also explore using another feature extractor to represent patch features in our histology slides. ViT models have many parameters and are known to require a lot of data in training. Since we have limited data to train our feature extractor model on, we also experiment with using a publicly available pre-trained model, CTransPath [116]. CTransPath was the first self-supervised feature encoder model trained on histology slides, but is still shown to be comparable to the latest histology foundation models [151]. The model architecture combines a standard CNN with a multi-scale Swin Transformer, which applies sliding windows on multiple scales in the self-attention layers [152]. We choose this model in particular because they use a considerable amount of CRC histology slides in training compared to other approaches, which should apply better to our CRC data. In particular, they train on the publicly available TCGA and PAIP data, the second of which contains 900 WSIs of CRC. They use further cohorts for validating their model by training downstream prediction tasks. Overall they use around 15 million image patches in training of the CTransPath feature extractor.

5.2.2 WSI Graph Design

Choosing the design of the graph to model the WSI is an important step, with lots of decisions to be considered, before training GNNs on the designed graph for outcome predictions. We discuss and provide justifications for our choices in this section below.

Superpixels

To find the nodes of the WSI graphs, we apply the Simple Linear Iterative Clustering (SLIC) superpixel algorithm [149] on the WSIs at 5x magnification to segment the tissue to capture cellular neighbourhoods that are roughly between 80-100 μm^2/pixels in size. The SLIC algorithm essentially clusters pixels based on the colour values of the pixel and the distance in the image to other neighbouring pixels, similar to the iterative K-Means clustering method. Specifically, the algorithm aims to minimise the distance D_s defined as

$$D_s = d_{lab} + \frac{m}{S} d_{xy}, \tag{5.1}$$

where d_{lab} is the Euclidean distance in the LAB colourspace, d_{xy} is the Euclidean distance in the spatial xy-plane, m is the compactness parameter and $S = \sqrt{N/K}$ where N is the number of pixels in the image and K is the approximate number of superpixels to aim for in the algorithm output [149].

It can be seen that the superpixel boundaries consistently align with the boundaries of tissue compartments, which was an observation also found by Achanta *et al.* when comparing the SLIC superpixel method to more recent state-of-the-art superpixel methods [153]. We choose to use superpixels to segment the WSI to give a more natural segmentation of the tissue instead of the arbitrary square tile which has no biological meaning. This is useful when visualising the predictions for each node, since the nodes correspond to more meaningful tissue sections than tiles would provide. We choose to use a GNN since a CNN would not work as well with superpixels due to their natural irregular shape.

Nodes and Edges

In designing the WSI graph representation, we then use these superpixels as the nodes on our graph. Specifically the centres of the superpixels are defined as the nodes, since a single point coordinate needs to be defined. We assign features to these nodes, based on the extracted morphological deep learning features found

on the underlying tissue in each superpixel. However, due to the expected input shape to the models, required to be consistent within each dimension, we cannot pass the superpixel directly to our feature extractor model, since its shape is not of a rectangular nature. One solution to this problem could be to use zero padding around the superpixel, creating a larger proxy square patch, though these would differ in size for each superpixel, which would render most deep learning methods unusable on such data. Our approach instead is to extract features for standard square patches, as discussed in Section 5.2.1, and then use the proportions of these square patches in each superpixel to calculate a weighted mean of the features which represent the tissue within the boundaries of that superpixel. More succinctly, the node features are the weighted mean of the corresponding patch features which overlap with the superpixel region. The edges of the graph between the nodes are determined by nearest neighbours from Delaunay triangulation, as in SlideGraph [97]. It's possible for edges of graphs to have fixed or learnable weights or features as well as the nodes, but in this case we didn't want to make assumptions about the interactions between the tissue types, and simply used fixed value edge connections. Experimenting with the values of the edges could be studied in future work.

Graph Neural Network

Building on the ideas introduced by SlideGraph [97], we use either GINConv [87] or GATConv layers [86], and explore using different depth (number of layers) and width (features in each layer) in our GNN. We add tempering to avoid overfitting, and replace their logistic regression scaler with a simple sigmoid function. We add three branches to the final layer of the GNN, in the form of three separate MLPs. Two of these MLPs return a graph-level prediction, for the response to RT and CMS4 predictions, and the final branch returns node-level predictions, predicting whether each node is epithelial tissue or not. Our loss function is defined as

$$\mathcal{L} = w_1 \mathrm{BCE}(\hat{y}_{RT}, y_{RT}) + w_2 \mathrm{BCE}(\hat{y}_{CMS4}, y_{CMS4}) + w_3 \mathrm{BCE}(\hat{\mathbf{y}}_{epi}, \mathbf{y}_{epi}) \ , \qquad (5.2)$$

where BCE is the binary cross entropy loss, $\hat{y}_{RT} \in \mathbb{R}$ is the slide-level prediction of response to RT, $\hat{y}_{CMS4} \in \mathbb{R}$ is the slide-level prediction of CMS4, $\hat{\mathbf{y}}_{epi} \in \mathbb{R}^{n_i}$ are the node-level predictions of epithelial tissue, n_i is the number of nodes in the i^{th} WSI graph, and w_1, w_2, w_3 are the loss weights for the respective outcomes. By default, we set all the weights equal such that $w_1 = 1, w_2 = 1, w_3 = 1$, though we also explore using these loss weights to guide the training process. In particular, since the final epithelial prediction is done at the node-level instead of the graph-level, there are many more samples for this prediction. Therefore we experiment with down-weighting this term in the loss function, to balance out the training across branches, using values such as $w_1 = 1, w_2 = 1, w_3 = 0.1$.

For each prediction branch, we can visualize the individual node predictions from the WSI graph, overlaid on the WSI itself, to get an idea of how the node predictions vary across the different tissue regions. Each graph-level prediction is derived from the corresponding branch node predictions, by applying pooling and dropout.

5.2.3 Data

We train and validate our methods on two retrospective rectal cancer datasets, Grampian and Aristotle. All patients in these cohorts received the same chemoradiotherapy treatment, consisting of RT with capecitabine. Pathological CR, which we use as a target outcome here, was derived from histopathological assessment from post-treatment resections.

The CMS classifications for each patient were derived from the pre-treatment biopsy samples. Here we concern ourselves only with the CMS4 call, used as a proxy for stromal infiltration, and define all other calls (CMS1-3, Unclassified and Unmatched) as not CMS4.

The epithelial labels for each graph node are calculated from epithelial masks for each WSI. These epithelial segmentation masks were generated at 10x magnification (1 μm^2/pixel) with a U-Net [154] which was previously trained and validated (outside of this project) on 666 full tissue sections belonging to 362 patients from

the FOCUS cohort [155]. The ground truth annotations for the training of this model were generated by an expert pathologist.

For consistency the tumour regions were marked up by an expert pathologist. We use these masks in our analysis to filter out background and irrelevant tissue from the images. Grampian and Aristotle are used in both training and validation, with a 70/30% training-validation split, keeping any WSIs from a single patient in the same dataset. We predict CR to RT against all other responses, such as partial response and no response. The datasets are unbalanced, since in Grampian only 61/247 slides have CR, and in Aristotle only 24/121 slides have CR. They are even more unbalanced for CMS4, since only 29/247 slides in Grampian and 17/121 slides in Aristotle are labelled with CMS4. We address this imbalance in the implementation details in Section 5.3.3. There are 368 slides total in our dataset, from 252 patients. See Chapter 3 for more details on the datasets.

5.3 Experiments

5.3.1 Superpixels

We experiment with various parameters of the SLIC superpixel algorithm to find an optimal application to our dataset, and we demonstrate the results here.

Number of superpixels

In our implementation of the SLIC superpixel algorithm we use the slic function from the skimage.segmentation Python package, version 0.19.3 [156]. This function has an option to define the approximate number of segments or superpixels you expect to see in the algorithm output, with the default value of 100. Due to the nature of the histology biopsy slides, the amount of tissue on each slide varies massively across the dataset (see Section 3.2), and therefore having a constant expected number of superpixels in each image across the whole dataset does not make sense. Hence we experiment with different approaches to choose the optimal number of

segments for each image to guide the algorithm, visualising and comparing the results on a subset of images.

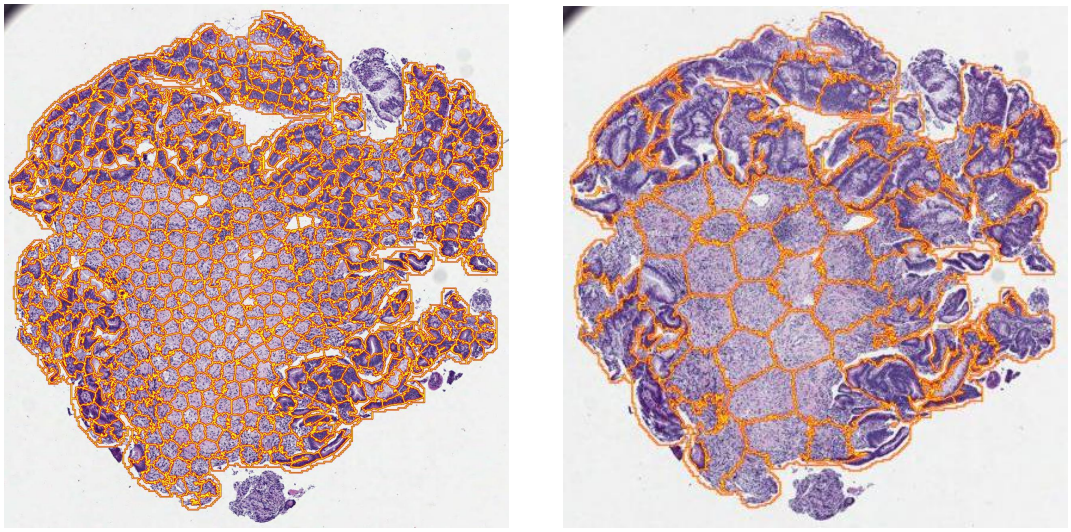
Firstly we try scaling the number of segments for each WSI by the size of the original image, taking the mean of the width and height of the image and dividing that by a scaling parameter, $scale\_slic = 2$. However, since the amount of tumour tissue (excluding background or other tissue) within the WSIs can vary and does not directly depend on the size of the image, we also implement the ability to use the number of patches in the WSI, again divided by a scaling parameter. The background and irrelevant tissue are already filtered out in the patching process, so this much better represents the quantity of tissue in the WSI.

We demonstrate these methods on a random sample of three WSIs from our datasets. These WSIs have sizes (6986, 3485, 3), (5080, 6474, 3) and (4002, 3983, 3) at 5x magnification, where the third dimension represents the three RGB colour channels. The resulting superpixel segmentations can be seen in Figure 5.2 for the number of segments correlated to the size of the image and Figure 5.3 for the number of segments correlated to the number of patches, both with varying values for the *scale_slic* scaling parameter. In all of these applications the compactness parameter is set to 20, otherwise default parameter values are used.

Initially, we set the expected number of segmentations to be the average size of the whole image scaled by $scale\_slic = 2$ (i.e. half the mean size of the WSI), as seen in Figure 5.2a, which results in smaller superpixels. However, we argue that scaling by the number of tissue patches provides more consistent relative superpixel sizes across images where the tissue proportions are different, and so we use this scaling in a later optimised version of the model.

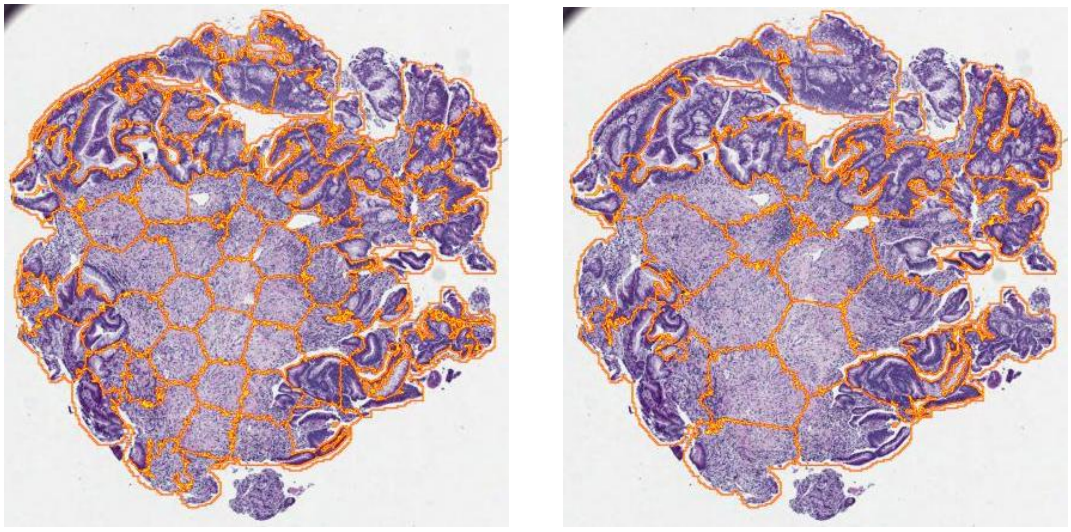
Compactness

In the SLIC algorithm, the compactness parameter is used to give more or less weight to the colour or space similarities. A higher compactness value gives more weight to the space proximity [156], meaning that superpixel shapes look more



(a) $scale\_slic = 2$ **(b)** $scale\_slic = 20$

Figure 5.2: The superpixel SLIC algorithm applied to an example WSI where the parameter defining the suggested number of segments is calculated by taking the average size of the image divided by a scaling parameter, *scale_slic*.



(a) $scale\_slic = 2$ **(b)** $scale\_slic = 5$

Figure 5.3: The superpixel SLIC algorithm applied to an example WSI where the parameter defining the suggested number of segments is calculated by taking the number of tissue patches in the image divided by a scaling parameter, *scale_slic*.

square, whereas a lower compactness values prioritises the colour values more in the segmentation, relaxing the spatial constraints of the algorithm.

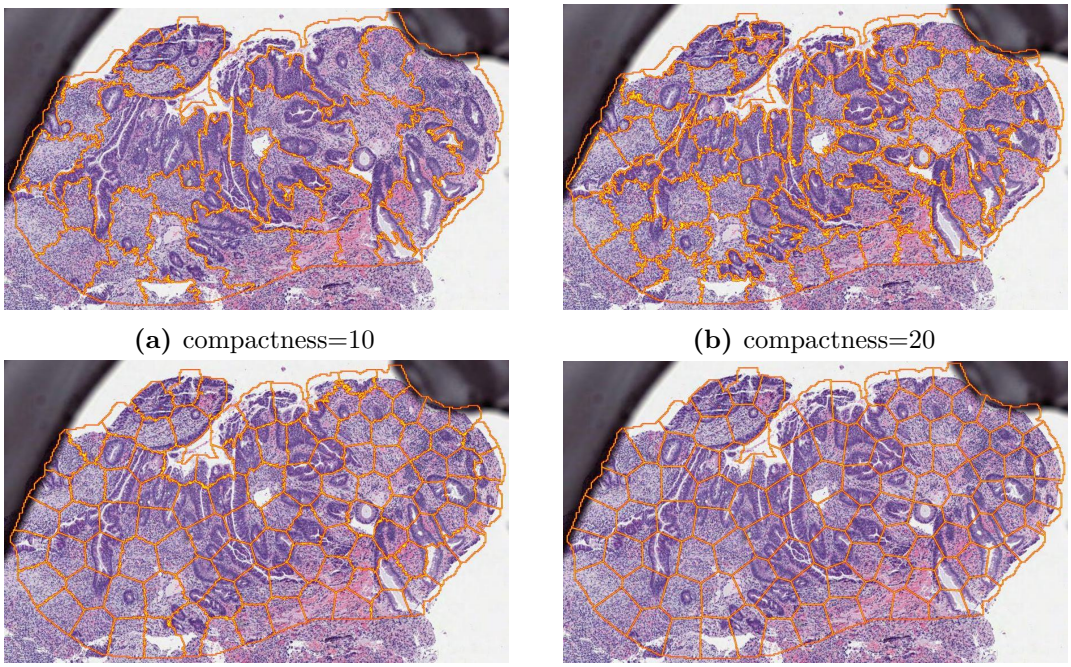
The visualisations from applying SLIC to a single slide with different compactness values (10, 20, 50, 100) can be found in Figure 5.4. The suggested number of segments for the algorithm is defined by the number of tissue patches scaled by $scale\_slic = 2$. In our experiments we found that when the compactness parameter was larger the superpixels were much more regular, whereas when it had a smaller value the superpixels could vary in size a lot more from each other, and look more irregular in shape. The lower compactness values work better for this application, since the tissue itself is highly irregular and the colours from the staining provide more information about the tissue type than the spatial proximity does. In applications where the compactness is higher (Figures 5.4c and 5.4d at compactness values 50 and 100 respectively), the segmentation boundaries are far less aligned with the underlying tissue boundaries, and in certain places cut across the tissue in arbitrary straight lines. The lower compactness values of 10 and 20 in Figures 5.4a and 5.4b respectively show better alignment with the tissue boundaries.

SLIC-zero

The SLIC-zero algorithm adapts the compactness parameter for each superpixel individually, which results in more regular and rounded superpixels across all tissue regions [153]. However, this results in worse segmentations, as seen in Figure 5.5, since the regions within the tissue are not regular and rounded but can, for example, be long and thin in places, which is not considered in this implementation.

5.3.2 Graph Connectivity

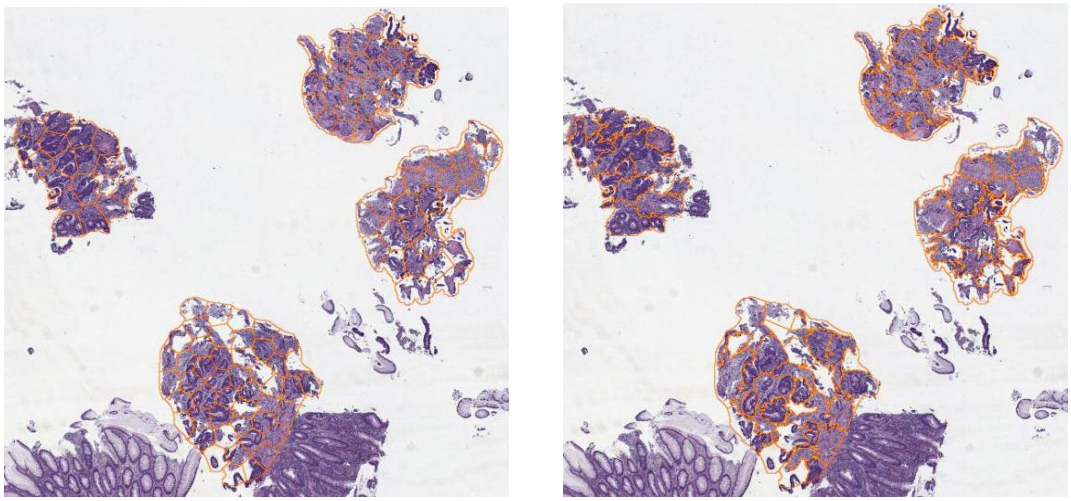
Once we have chosen the nodes of our graphs based on the centres of the selected superpixel regions, we experiment with how to connect these nodes within the WSI graph. One thing to consider when connecting these nodes is that the WSI can contain multiple tissue biopsies in one image, and therefore we need to consider



(a) compactness=10 **(b)** compactness=20

(c) compactness=50 **(d)** compactness=100

Figure 5.4: The superpixel SLIC algorithm applied to an example WSI with different values for the compactness parameter.



(a) SLIC-zero with initial compactness=10 **(b)** compactness=10

Figure 5.5: The SLIC algorithm applied on a single slide, comparing the effect when we set $slic\_zero = True$ in the algorithm, which uses the given compactness as an initial value and then runs the SLIC-zero algorithm to find a suitable compactness level.

whether to connect the tissue across these biopsies by defining edges between the corresponding nodes. However, due to the fact that any connections between these nodes would be more dependent on the layout of the biopsies in the glass slides when being scanned, as opposed to the adjacency expected to be seen in situ, we choose to find an algorithm which would not, in most cases, connect separate biopsies. Since these biopsies are adjacent in the unobserved third dimensional plane, it could be interesting to explore 3D graphs to model such connectivity between biopsy slices in future work.

We define the edges between nodes based on Delaunay triangulation, excluding edges which extend beyond a certain threshold. We explore different methods of defining this threshold which would generalise across different WSIs, connecting nodes well within samples but avoiding connections across biopsies. Initially, we use a similar approach to when defining the number of superpixels per slide, and set the distance threshold as the average range (i.e. maximum - minimum positions across x-y planes) found in the node positions divided by a scaling parameter, *connectivity_scale*, which we set to 8 initially.

Similar to the superpixel approach, we also experiment with using the size of the WSI, number of patches per WSI, and additionally the ratio of number of patches to the log area of the WSI, but we find that using the average range of node coordinates produces the most consistent results of these approaches. We also experiment with different values of the *connectivity_scale* parameter, including 8, 16, 20. Higher values of *connectivity_scale* (e.g. 16 vs 8) decrease the distance threshold for the Delaunay adjacency matrix and removes far away connections. We also experiment with using an absolute value for the distance threshold across the dataset, and found that this also gives good and reasonably consistent results.

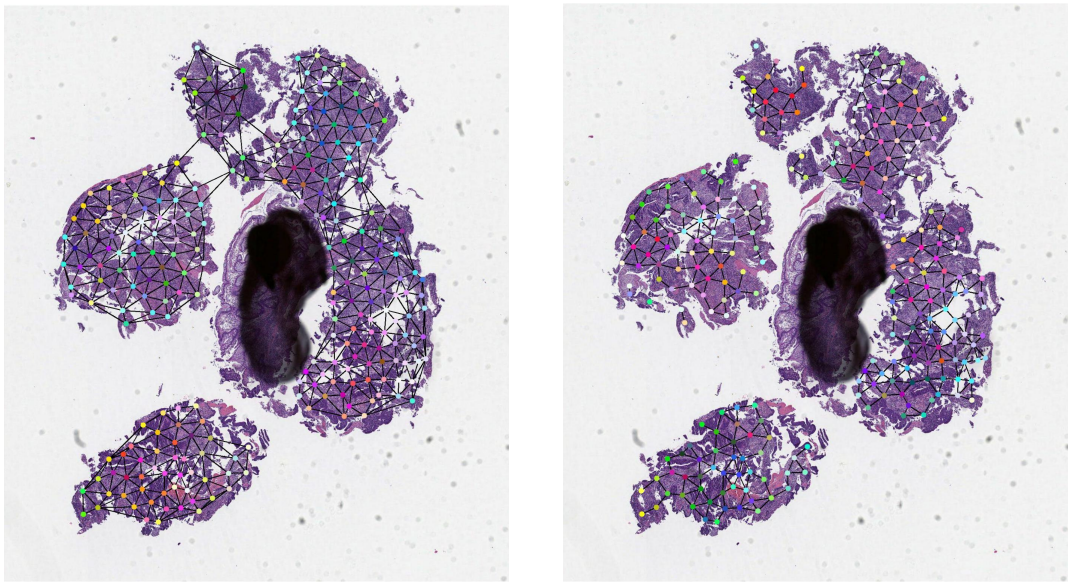
Examples demonstrating some of these approaches can be found in Figures 5.6 and 5.7, where for each sample slide we provide two graphs with the connectivity distance threshold set as the average range of the node coordinates scaled

by *connectivity_scale* = 8 or 20, and two graphs with the connectivity distance threshold set at absolute values of 800 and 1000. Despite initially using *connectivity_scale* = 8, we later prefer the choice of using the absolute *connectivity_distance* = 800 for better graph connectivity across samples. It can be seen that other choices result in too sparse edges across examples, such as *connectivity_scale* = 20, leaving some nodes not connected to any other nodes. On the other hand, using *connectivity_scale* = 8 gives more edges between nodes, but these can often jump across multiple biopsies which makes less sense in terms of modelling neighbouring tissue in situ. Despite using an absolute connectivity distance to be less intuitive since the size of the WSIs can vary, the resolution of the tissue remains comparable across samples and so we find this approach to be more robust in providing consistently meaningful graph connections.

5.3.3 Implementation

To train the self-supervised feature encoder we use the default DINO parameters, but train for 20 epochs with 5 warmup epochs. As discussed above, we apply the SLIC algorithm [149] with compactness of 20. In the original version we initially set the number of segments for each WSI as half the mean size of the WSI, and in the optimised implementation we later use half the number of patches. The parameters for the SLIC algorithm were initially chosen to provide meaningful segmentations confirmed by clinical experts.

Prior to fitting the graph model we normalize the node features relative to the training dataset. We train our graph model for 30 epochs using the Adam optimizer with learning rate $1e-3$ and weight decay $1e-4$. Our graph model has three GINConv layers [87] with dimensions 64, 32 and 16 respectively. We apply dropout of 0.5 in between graph layers, use minimum aggregation for message passing between nodes and initially use maximum pooling for concatenating the node activations. We apply tempering to the outcome of the graph model, dividing the output by 1.5. Graph hyperparameters were chosen from visualisation and model hyperparameters were determined by ablation studies on the validation set.



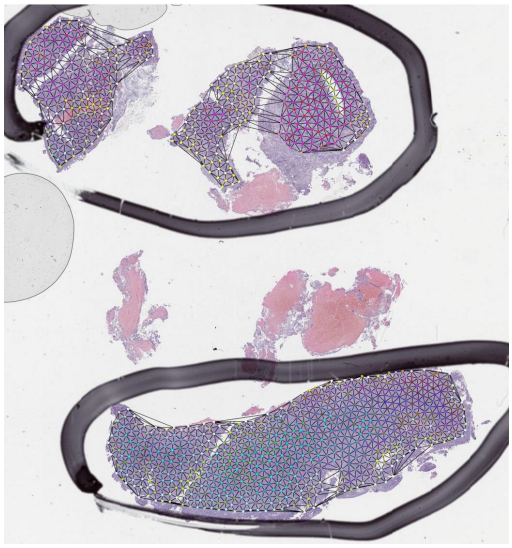
(a) connectivity scale = 8

(b) connectivity scale = 20

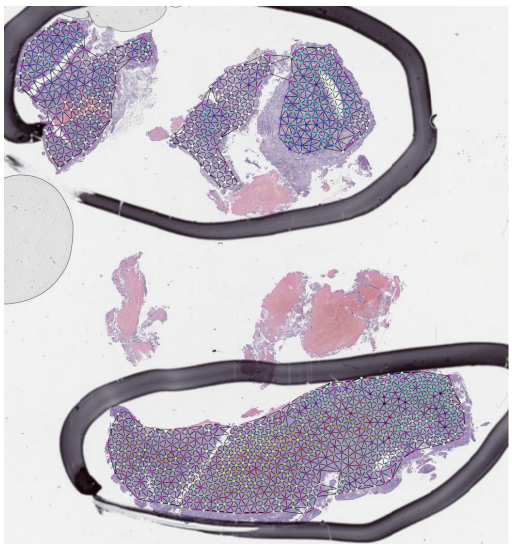
(c) connectivity distance = 800

(d) connectivity distance = 1000

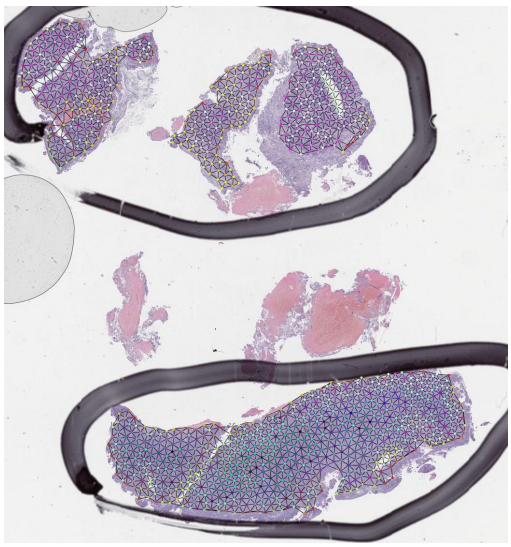
Figure 5.6: Graphs generated on a single WSI using different distance connectivity thresholds for the Delaunay triangulation algorithm to determine the graph edges. The colours of the nodes hold no value, their purpose is simply to better distinguish the nodes from one another. Where the connectivity scale is used, this value scales the average range of the node positions in the WSI, otherwise an absolute connectivity distance threshold is set. The nodes are consistent across graphs, determined by the centres of the superpixels.



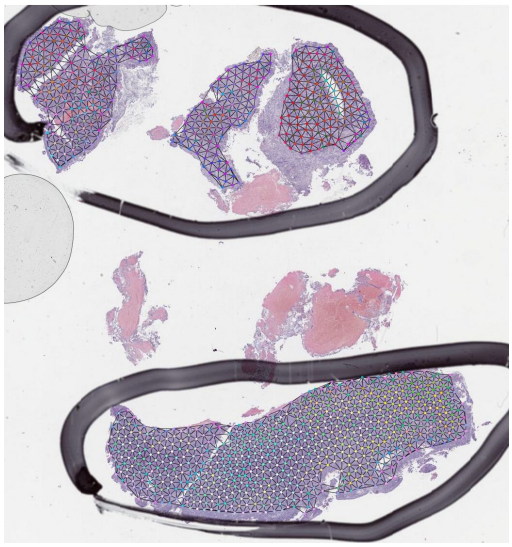
(a) connectivity scale = 8



(b) connectivity scale = 20



(c) connectivity distance = 800



(d) connectivity distance = 1000

Figure 5.7: Another example showing graphs generated on a single WSI using different distance connectivity thresholds for the Delaunay triangulation algorithm to determine the graph edges. The colours of the nodes hold no value, their purpose is simply to better distinguish the nodes from one another. Where the connectivity scale is used, this value scales the average range of the node positions in the WSI, otherwise an absolute connectivity distance threshold is set. The nodes are consistent across graphs, determined by the centres of the superpixels.

Response branch	Response to RT	CMS4	Epithelial
Mean AUC (std)	0.819 (0.04)	0.819 (0.04)	0.760 (0.01)
Mean accuracy (std)	0.795 (0.05)	0.750 (0.04)	0.691 (0.01)
Mean balanced accuracy (std)	0.774 (0.05)	0.719 (0.05)	0.691 (0.01)
Mean weighted F1 (std)	0.810 (0.04)	0.791 (0.02)	0.700 (0.01)
Mean weighted precision (std)	0.843 (0.02)	0.870 (0.02)	0.725 (0.00)
Mean weighted recall (std)	0.795 (0.05)	0.750 (0.04)	0.691 (0.01)

Table 5.1: Initial Results For each fold, we take the mean metrics for the three branch predictions from the best model on our validation data, with the best epoch chosen based on mean AUC for the three predictions. The standard deviation of the metrics across the four folds is provided in brackets. Each prediction uses an optimised threshold value determined from the validation set in order to round the output probabilities to a binary prediction. We use weighted metrics due to the class imbalance in our dataset.

We evaluate the best validation epoch by finding the best mean AUC across the three prediction branches. We use weighted metrics due to the class imbalance in our dataset, in order to better represent the performance on the less prevalent group, the patients with CR to RT. Specifically, we use the Python library scikit-learn 1.3.0. For the weighted F1-score, for example, we use the sklearn.metrics.f1_score function with the parameter average='weighted'. We do the same for the AUC, precision and recall. The balanced accuracy function in scikit-learn is defined specifically for imbalanced datasets, and we also provide the standard accuracy score for comparison.

To address the imbalance of the labels in our dataset, we upsample the complete responder slides when training our model, so that there are roughly equal samples from complete responders and not complete responders in the training dataset. We run the whole pipeline on four folds with different random data splits for training and validation.

5.4 Results

Here we provide results on the slide-level, comparing to slide-level ground truth for the response to RT and CMS4 prediction branches, but using the node-level ground truth labels for the epithelial prediction branch, as seen in Equation (5.2).

Beyond this, we present further results from experiments estimating the proportion of epithelium at the slide-level instead of individual evaluations at the node-level.

Exploring these results further, we observe that when we review the metrics within the validation set of each cohort, our trained model performs much better on one (Grampian) than the other (Aristotle). Therefore we also work on optimising the performance across cohorts, making the performance more equal.

Finally, we apply our optimised model to an unseen cohort of patients, the Salzburg dataset, to test its generalisation capabilities. The results presented here motivate our following work in the next chapter on domain adaptation in Chapter 6.

5.4.1 Node-level Epithelium

Despite the noise in our reference data used for training, using the trained DINO model as a feature extractor our approach achieves good performance in terms of mean AUC scores on all three prediction branches of our model, predicting CR to RT with 0.819 AUC, CMS4 with 0.819 AUC and epithelial tissue at the node level with 0.760 AUC across folds. Further metrics are provided in Table 5.1. Using this initial approach, the prediction performance of the model could be improved by utilising a larger training dataset and performing more exhaustive parameter searches, however the current performance of the model is sufficient to demonstrate the impact of this approach.

We then go on to explore using CTransPath as a feature extractor, spending more time optimising parameters such as the size of the GNN, using larger layers of sizes 384, 192, 96 and 48. Where the previous model does not use jumping connectivity between layers, here we add this functionality. The previous model uses loss weights of $w_1 = 1, w_2 = 1, w_3 = 0.1$ in training, though now we find reasonable results setting these all equal to 1, and training for 50 epochs. We also update our graph design as described in Sections 5.3.1 and 5.3.2, and now use mean pooling for concatenating the node activations. Finally, we apply heavier data augmentation as described in the section on Balancing Across Cohorts, Section 5.4.4. All these approaches improve our results, which can be found in Table 5.2.

Response branch	Response to RT	CMS4	Epithelial
Mean AUC (std)	0.862 (0.05)	0.846 (0.09)	0.887 (0.01)
Mean accuracy (std)	0.783 (0.03)	0.744 (0.10)	0.806 (0.01)
Mean balanced accuracy (std)	0.773 (0.04)	0.738 (0.10)	0.806 (0.01)
Mean weighted F1 (std)	0.802 (0.03)	0.777 (0.09)	0.808 (0.01)
Mean weighted precision (std)	0.844 (0.02)	0.849 (0.06)	0.815 (0.01)
Mean weighted recall (std)	0.783 (0.03)	0.744 (0.10)	0.806 (0.01)

Table 5.2: Optimised Results For each fold, we take the mean metrics for the three branch predictions from the best model on our validation data, with the best epoch chosen based on mean AUC for the three predictions. The standard deviation of the metrics across the four folds is provided in brackets. Each prediction uses an optimised threshold value determined from the validation set in order to round the output probabilities to a binary prediction. We use weighted metrics due to the class imbalance in our dataset.

5.4.2 Visualisation

This approach allows us to visualize the predictions of our model across the three prediction branches at the node level, which can be overlaid on the original WSI to show an array of intuitive heatmaps of predictions. An example of our proposed prediction maps on two slides can be seen in Figure 5.8. The predicted response to RT can now be viewed in the context of disease biology as captured by CMS4. For example, the model demonstrates that CMS4 patients are less likely to respond to RT. In addition, it is now possible to view the spatial distribution of CMS4 active regions in the tissue architecture context as shown in Figure 5.8.

A pathologist reviewing these maps assesses that the observed patterns fit the known interplay of response to therapy, CMS4 activation, and the spatial localisation of these signals. In the top slide, we observe high CMS4 activation in stromal rich regions, and interestingly also high CMS4 activation in the bottom center, dissociating from the response to RT activation map. This could be explained by the lymphocyte content, supported by the higher epithelial map activations in the same location. Expert pathologists highlight a similar pattern in certain regions of the maps for the bottom slide. Different from the slide above, the CMS4 and response to RT maps have some overlap with moderate activations here, encouraging discovery into tumour-host interactions. Ultimately, a pathologist confirmed that

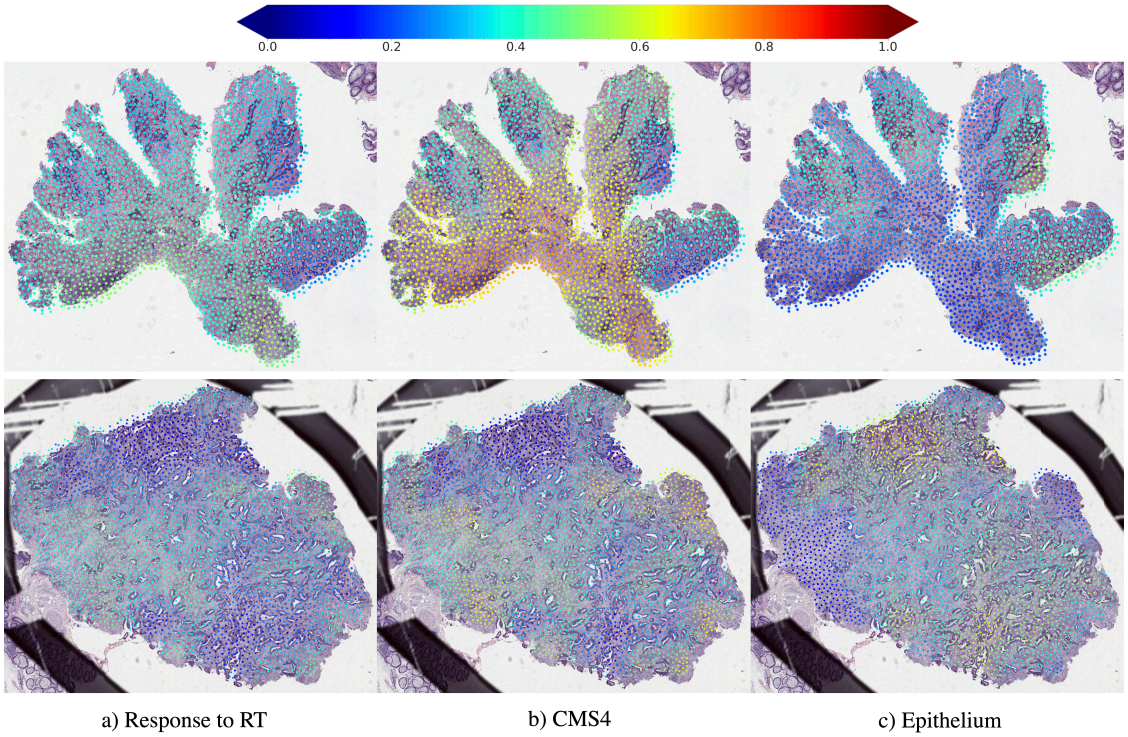


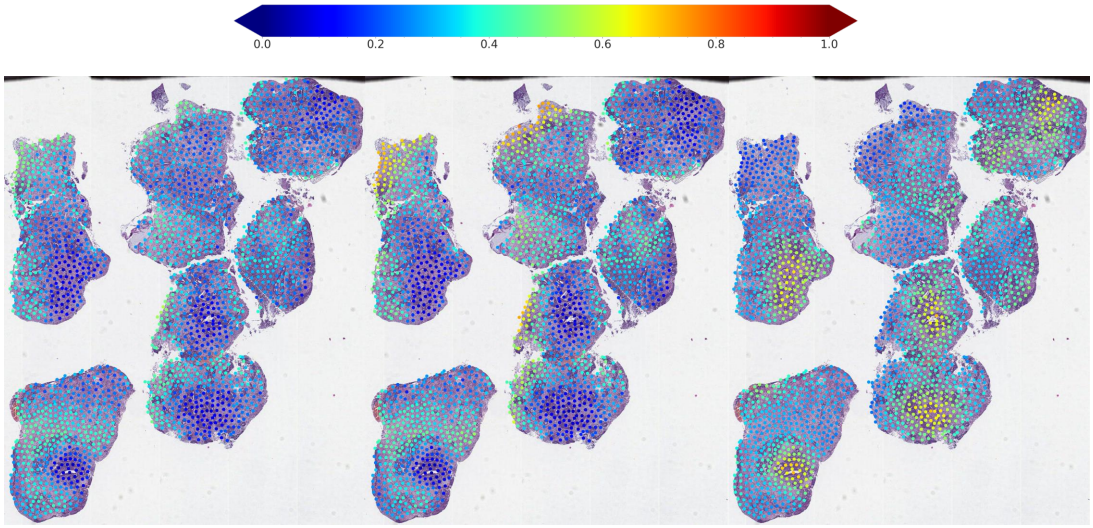
Figure 5.8: Node activation maps from the three prediction branches on two different slides, top and bottom. The nodes are coloured by their predictions. Both slides are classified as CMS4 and the patients did not have a complete response to radiotherapy.

these maps support an interpretable and trustworthy prediction in the context of response to RT. Further prediction maps and their pathologist reviews can be seen in Figures 5.9 to 5.11, providing a more extensive interpretation indicating that the proposed approach enables a level of analysis that has not been possible before.

5.4.3 Slide-level Epithelium Proportion

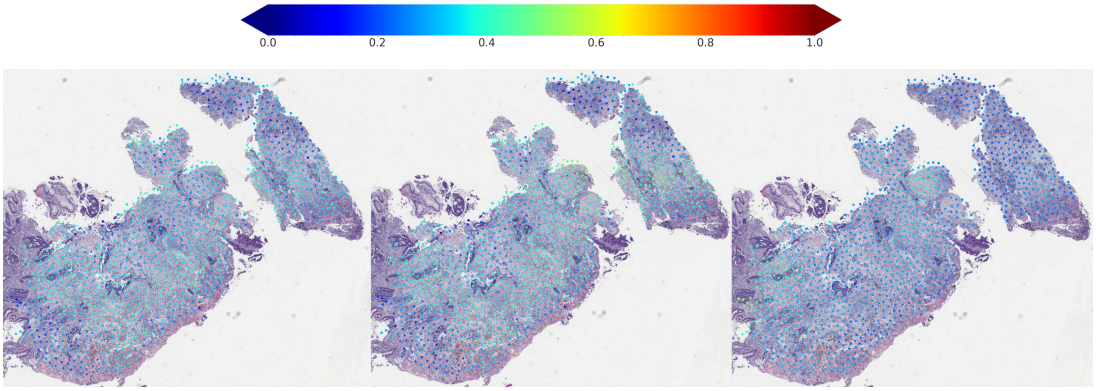
We also explore treating the epithelium label as a slide-level label instead of a node-level label, by calculating the proportion of nodes assigned as epithelial against those that are not, therefore calculating an estimate of the proportion of epithelium in the WSI.

While this does reduce the quantity of information we have regarding the epithelium label, it also makes the training of our model arguably more equitable across the three prediction branches, since now all predictions are at the slide-level and therefore all branches are trained in a weakly supervised manner. However,



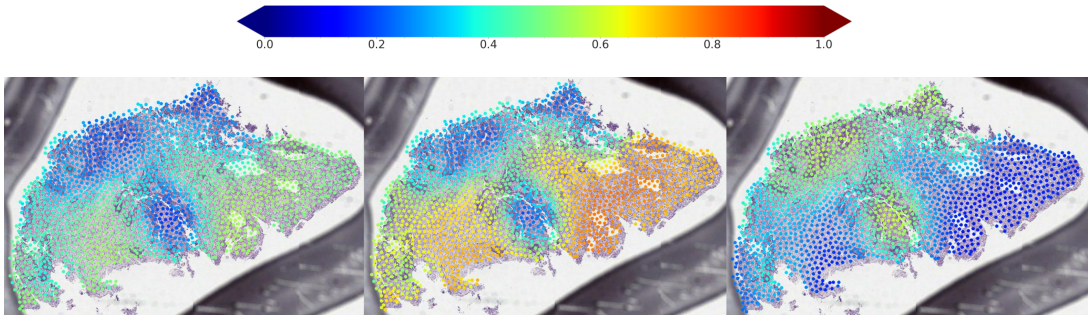
(a) Response to RT (b) CMS4 (c) Epithelium

Figure 5.9: Expert pathologist: complex glandular growth patterns are highlighted as poor response to radiotherapy (RT) regions, warranting further research for a potential biomarker; CMS4 activations highest in stromal-rich regions and lowest in epithelial regions; epithelium map shows reverse activations, lowest in stromal-rich regions and highest in epithelial-rich regions. Slide classified as CMS4; patient did not have a complete response to RT. Note any offsets of the nodes on the WSI is due to a visualization issue instead of an underlying computational issue.



(a) Response to RT (b) CMS4 (c) Epithelium

Figure 5.10: Expert pathologist: poor response to radiotherapy (RT) activations are consistent with the observed high stromal content, low epithelial content, aggressive growth patterns and immune-poor environment. Computational analysis: similarity in mid-range activation values on the left- hand side of the CMS4 and epithelium maps encourages exploration of tumour-host interactions in further research. Slide classified as CMS4; patient did not have a complete response to RT. Note any offsets of the nodes on the WSI is due to a visualization issue instead of an underlying computational issue.



(a) Response to RT (b) CMS4 (c) Epithelium

Figure 5.11: Expert pathologist: response to radiotherapy (RT) activations are enhanced towards areas with seemingly more lymphocytes in the stroma, which could explain difference from larger CMS4 activations; CMS4 activations are pronounced on areas of higher stromal content as expected; epithelium activations are larger on epithelial-rich regions as expected. Slide classified as CMS4; patient did not have a complete response to RT. Note any offsets of the nodes on the WSI is due to a visualization issue instead of an underlying computational issue.

the WSI epithelial ratio is a continuous variable $\in [0, 1]$, whereas the other two labels are binary $\in \{0, 1\}$.

We perform these experiments using CTransPath as the feature extractor, with light data augmentation and no stride across patches. Training on a portion of Grampian and Aristotle, and using the other portion of both in the validation set, for only one seed, we get the results seen in Table 5.3. The results are promising, with 0.72 Pearson correlation between the predicted and true epithelial ratios in each WSI, and a very small mean absolute error of 0.06. Furthermore, setting the epithelial branch to be a continuous slide-level label does not seem to negatively affect the results on the other two branches, compared to the results seen in Table 5.1.

5.4.4 Balancing Across Cohorts

When analysing the predictions from our initial model with the node-level epithelium predictions as well as the slide-level response to RT and CMS4 predictions, we broke our validation set down into the separate cohorts, Grampian and Aristotle. Comparing the metrics evaluated within each cohort validation set we observe that the model performed far better on images from the Grampian dataset than

Response branch	Response to RT	CMS4	Epithelial Ratio
AUC	0.859	0.884	-
Balanced accuracy	0.830	0.700	-
Weighted F1	0.879	0.855	-
Pearson correlation	-	-	0.716
MAE	-	-	0.059

Table 5.3: For a single data split fold, we provide results from training an approach to predict the continuous slide-level epithelial ratio label in parallel with the binary response to RT and CMS4 labels. Continuous metrics, Pearson correlation and mean absolute error (MAE), are provided for the continuous prediction branch, and classification metrics are provided for the binary predictions, using weighted metrics due to the class imbalance in our dataset. The metrics provided are from the best epoch, chosen based on the mean AUC and Pearson correlation across the three prediction branches in the validation set.

images from the Aristotle dataset. To solve this issue, we focus on methods to make our model more generalisable, implementing heavier data augmentation on the images before extracting the node features.

In the initial work we use lighter data augmentation which included the following applied on the training dataset: resize images to 224 x 224, randomly flip vertically with probability 0.5, randomly flip horizontally with probability 0.5, add colour jitter with the parameters 0.1 for brightness, 0.05 for contrast, 0 for saturation and 0.1 for hue, randomly rotate at right angles, and normalise with respect to the ImageNet dataset mean and standard deviation. The heavier augmentations we apply to balance results better across cohorts includes all of the above with stronger variation in the colour jitter, plus additional augmentations. Specifically, the colour jitter parameters are now either maintained or increased to 0.1 for brightness, 0.25 for contrast, 0.5 for saturation and 0.25 for hue. We also apply the following to the training dataset: Gaussian blur the image using a kernel size of 9 x 9 pixels, randomly adjust the sharpness with factor of 2 and a probability of 0.2, and randomly auto-contrast the image with probability of 0.5.

The results from this approach can be seen in Table 5.5, using CTransPath as a pre-trained feature extractor, with metrics broken down by cohort. Balancing the results across prediction branches is another challenge in this approach, and we choose the best model as the one which maximises the AUC across all three

Cohort	# CR to RT	# CMS4
Grampian	43	20
Aristotle	16	13

Table 5.4: The number of positive labels for each slide-level binary class, in one of our random seed training datasets, broken down by cohort.

Evaluation Cohort	Response to RT AUC	CMS4 AUC	Epithelial AUC
Overall	0.862 (0.05)	0.846 (0.09)	0.887 (0.01)
Grampian	0.950 (0.04)	0.923 (0.08)	0.888 (0.00)
Aristotle	0.636 (0.09)	0.705 (0.13)	0.888 (0.01)

Table 5.5: Results from using heavier data augmentations on the images, given on the validation set of each patient data cohort (Grampian or Aristotle) across four random data split folds. The results are from the best epoch in each fold, chosen based on the mean AUC for the three predictions across the combined cohorts. The mean of the metrics across the four folds is given and the standard deviation is provided in brackets. Each prediction uses an optimised threshold value determined from the validation set in order to round the output probabilities to a binary prediction. We use weighted metrics due to the class imbalance in our dataset.

prediction branches. Using heavier data augmentation increases our performance in all three prediction branches, in the validation sets across the two cohorts.

Our model still performs far better on the Grampian cohort than on Aristotle, despite efforts to help the model generalise across cohorts. There are almost twice as many images from Grampian than Aristotle in our training and validation datasets, so this cohort imbalance could partially explain the bias here. The number of positive labels in terms of the binary slide-level labels (CR to RT and presence of CMS4) in each cohort can be found in Table 5.4, for the training dataset in one fold (i.e. random seed split). The number of positive complete responders in the training set from Aristotle makes up under a third of the complete responders, potentially explaining why our model performs less well on the Aristotle validation set for this classification branch.

5.4.5 External Test Set

Using the model described in Section 5.4.4, we apply the model from the best performing fold to an unseen test dataset. The Salzburg dataset is taken from a cohort of patients (n=55) in a different country (Austria) to the country where the patient cohorts in the training and validation data originate from (UK), and therefore makes a good external test set for our model.

The Salzburg data is visually rather different from the Grampian and Aristotle data, with more biopsies per WSI and fainter staining, making it a challenging test set. These differences are discussed more in Chapter 6.

The results of applying our best model onto the Salzburg data can be seen in Table 5.6. For comparison, we also provide the metrics on the validation set from training, for this specific fold of model. For all metrics excluding AUC, we use the optimal thresholds determined from the validation set for rounding the prediction to a binary one.

The model achieves 0.65 AUC on the response to RT prediction on the test set, and only 0.55 on the CMS4 prediction. Comparing to the model on the validation set, it can also be observed that with AUC scores of 0.94 for response to RT and 0.73 for CMS4, this model naturally performs better on the former prediction branch over the latter, as is reflected in the test set prediction results. The epithelial prediction is very reasonable on the test set, with an AUC score of 0.85, not far from the performance observed on the validation set, of 0.89 AUC.

While clearly some meaningful signals are being detected by our model in the test set, these metrics would not suffice for trustworthiness or application of our model. Therefore, in the next chapter we focus on developing domain adaptation methods to improve the performance of our existing model on a new domain, in an unsupervised manner. It should be noted that initially, the previous version of this model generalised even worse to the Salzburg cohort, achieving metrics of 0.544 AUC, 0.500 balanced accuracy and 0.840 weighted F1. It's this version of the model that we use for the work in the next chapter, Chapter 6, and hence those metrics and the model that generated them are used there as a baseline. However,

Cohort	Response branch	Response to RT	CMS4	Epithelial
Grampian and Aristotle (validation set)	AUC	0.936	0.730	0.894
	Accuracy	0.829	0.622	0.813
	Balanced accuracy	0.824	0.621	0.813
	Weighted F1	0.836	0.662	0.816
	Weighted precision	0.855	0.762	0.822
	Weighted recall	0.829	0.622	0.813
Salzburg (test set)	AUC	0.650	0.548	0.848
	Accuracy	0.800	0.618	0.724
	Balanced accuracy	0.595	0.496	0.764
	Weighted F1	0.817	0.585	0.727
	Weighted precision	0.838	0.569	0.798
	Weighted recall	0.800	0.618	0.724

Table 5.6: External Test Cohort Results We apply the best model from validation on an external test cohort of patients, the Salzburg dataset (n=55). The results for this model are also provided on the original validation dataset (validation subsets of Grampian and Aristotle) for comparison. For both datasets we use thresholds to optimise the balanced accuracy in the validation set, in order to round the output probabilities to a binary prediction (where applicable). We use weighted metrics due to the class imbalance in our dataset.

we then returned to this work and managed to improve the generalisability of our model to the metrics stated above, which still leaves room for improvement.

5.4.6 Ablation Studies

With the initial version of this work we also provide ablation studies on the original model. Using ablation studies, we prove our model and the prediction maps it produces are robust. Changing the dropout, loss weights, loss function, and message passing aggregation methods only changes prediction AUC scores by absolute values up to 0.03. The node activation maps are also very visually similar across ablation study models.

We find that predicting these outcomes individually in a single branch model, particularly with response to RT, can result in slightly higher AUC scores, but we consciously make this trade-off in order to provide better interpretability of the model predictions. The focus of this research is not to achieve the best possible

Ablation Study	Response to RT AUC	CMS4 AUC	Epithelial AUC
Including Unmatched CMS4	0.819	0.819	0.760
Excluding Unmatched CMS4	0.803	0.874	0.754

Table 5.7: Results from an ablation study on the effect of excluding the unmatched CMS4 WSIs from the dataset, instead of including them and defining them as 'not CMS4'. Excluding the noisy labels increases the AUC score of the CMS4 prediction on the validation set, at the detriment of the AUC scores for the two other branch predictions.

metrics, but to develop robust methods which can add context and explanation to clinical black box deep learning model predictions, with the view to ease clinical translation of such models.

To explore the effects of the noisy CMS4 ground truth labels, we remove from our dataset any WSIs classified as 'Unmatched' for the CMS call, which for the main results of this work we defined as 'Not CMS4'. Removing this data and rerunning our analysis improved our predictions for CMS4 by +0.06 AUC, and reduced our response to RT and epithelial predictions by −0.02 and −0.01 respectively. The results can be found in Table 5.7. These small changes indicate that the noise in our data does not degrade the performance of our classifier, reinforcing it as a robust and accurate model.

5.5 Analysis of Gradients

Once we have the node predictions for each prediction branch (CR to RT, CMS4 and epithelium), we can explore whether we can quantify variability in the outcomes within the WSI. Specifically, we aim to explore how the predicted values of CR and CMS4 change across the epithelial tissue boundaries, to explore the 'flow' or 'gradients' of the outcomes at tissue boundaries.

We use the epithelial ground truth labels to provide locations of the tissue boundaries. We use the existing WSI graphs to find connected nodes, and then compare the epithelial label ground truth values at neighbouring nodes to find

node pairs which cross the epithelial tissue boundaries, avoiding reverse duplicates i.e. only measuring the flow in one direction, from non-epithelial to epithelial tissue. Then, we evaluate the changes in the predicted CR and CMS4 values for each node pair at the boundaries.

To compare whether the gradients of the predicted outcome values are different across boundaries to within boundaries, we also collect all non-boundary pairs of nodes (again, avoiding duplicates but with arbitrary direction this time). Similarly, we calculate the change in predicted CR and CMS4 values across nodes within the epithelial and non-epithelial tissue regions, without distinguishing between these groups. This allows us to quantify the expected homogeneity of predictions within similar tissue sections, providing a suitable comparison for predictions at tissue boundaries.

To test whether the outcome predictions across tissue boundaries are different to those within tissue boundaries, we run a two-sided, two sample t-test for the two outcomes CR and CMS4, assuming equal population variances and normal underlying population distributions for each hypothesis test. Formally, we test the hypotheses:

$\boldsymbol{H_0}$**:** CMS4 changes across boundaries = CMS4 changes within boundaries,

$\boldsymbol{H_1}$**:** CMS4 changes across boundaries $\neq$ CMS4 changes within boundaries

for CMS4, and similarly for CR,

$\boldsymbol{H_0}$**:** CR changes across boundaries = CR changes within boundaries,

$\boldsymbol{H_1}$**:** CR changes across boundaries $\neq$ CR changes within boundaries.

We run these tests individually for each slide in our validation set, using model predictions for our best fold model from Section 5.4. We use a critical level of 5% to test for significance, meaning if the hypothesis test returns a p-value less than 0.05 we call this a significant p-value and can reject the null hypothesis that the means are equivalent across between- and within- boundary samples.

Out of 111 slides in our validation set, we find that 109 have significant p-values for the CMS hypothesis test, and 110 have significant p-values for the CR hypothesis test, meaning that in the vast majority of cases we can reject the hypothesis that

the outcome flows across boundaries are the same as the flows within boundaries. From this we can interpret that there is indeed a different flow across the tissue boundaries in terms of outcome values within a slide, for both predicting the CR to RT and presence of the CMS4 molecular subtype.

To visualise the outcome differences across epithelial boundaries, we plot histograms of the raw differences in predictions. For a single slide within the validation set, we provide these histograms in Figure 5.12, for both the CR (top) and CMS4 (bottom) outcomes. Each figure contains two plots, with the histogram on the left being the differences across the epithelial boundaries, and the histogram on the right being the differences within boundaries for comparison. We observe, as expected, that the differences in predicted CR and CMS4 values on nodes within boundaries are approximately normally distributed around zero.

After observing the trends in the histograms of the boundary flows for multiple slides, we run a one-sided t-test to test whether the size of flows across the boundaries are larger than flows within the boundaries. The hypothesis tests are now formally defined as follows:

$\boldsymbol{H}_0$**:** CMS4 changes across boundaries = CMS4 changes within boundaries, vs.

$\boldsymbol{H}_1$**:** CMS4 across boundaries > CMS4 within boundaries,

and similarly,

$\boldsymbol{H}_0$**:** CR changes across boundaries = CR changes within boundaries, vs.

$\boldsymbol{H}_1$**:** CR changes across boundaries > CR changes within boundaries.

These one-sided tests give the same results as the two-sided tests, with 109/111 slides having significant p-values for the CMS hypothesis test, and 110/111 having significant p-values for the CR hypothesis test. From this we can now reasonably hypothesise that the values of both CR and CMS4 are greater in the epithelial sections of tissue compared to immediate neighbouring nodes in the non-epithelial tissue sections. However, this does not necessarily mean that there are generally-speaking higher CR and CMS4 prediction values in the epithelial than in the non-epithelial tissue, as we are just evaluating flows at the boundaries compared to within boundaries.

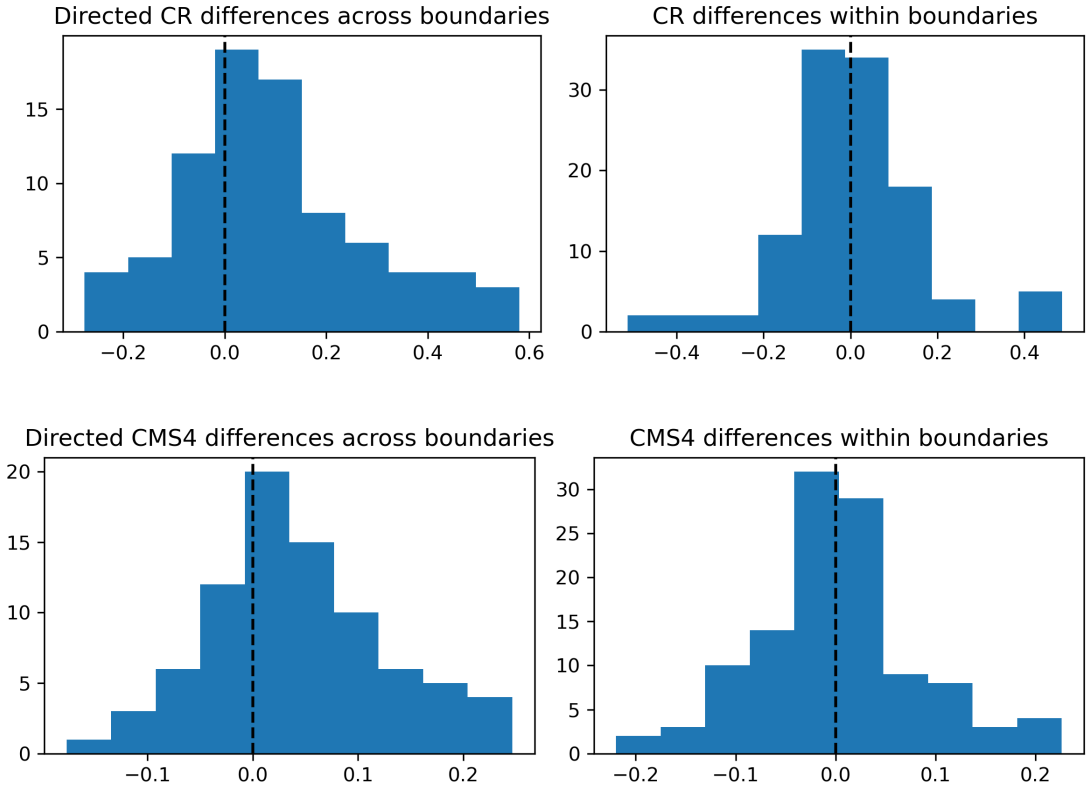


Figure 5.12: Histograms of the differences in CR (top) and CMS4 (bottom) predicted values between neighbouring nodes, across epithelial tissue boundaries (left) and within these boundaries for both epithelial and non-epithelial tissue (right). The differences across boundaries are taken going from the non-epithelial tissue to the epithelial tissue, and therefore have a positive or negative sign associated. The differences within boundaries have an arbitrary positive or negative sign as there is no direction associated with the neighbouring nodes here. The dashed vertical line represents a theoretical zero difference between node values, provided for reference.

This analysis demonstrates significant differences between both response to RT and CMS4 predictions within the WSI across tissue boundaries, clearly signifying that both predictions are heterogeneous dependent on the tissue regions. Here we have demonstrated feasibility of this idea, as there is no existing study which tests this specifically, but it would be interesting to further explore the gradients of molecular traits and therapy responses around these boundaries.

5.6 Conclusion

By setting the prediction of response to therapy in context with disease biology and spatial organisation of the tissue we are providing a novel approach for enhancing the interpretability of complex prediction tasks. These results do not only enhance the interpretability, they also provide new ways to utilise large retrospective clinical trial cohorts for which no additional molecular data is available. In future, these methods could be used to help better characterise tumour-stromal interactions of the tissue.

We argue that this work also advances the state of the art in feature representation and analysis. By using a multi-task learning approach, our prediction maps derive from the same graph model, and hence they share underlying graph features. The prediction branches only diverge at the final stage of translating these graph features into outcome predictions for our three clinically relevant outcomes. Importantly, this level of visualisation is not only accessible to pathologists, but this joint prediction model also enhances the communication between pathologists and oncologists which is critical for patient management. By cross-referencing these prediction maps with our prior understanding of cancer biology, this approach can help to establish trust in the prediction model and also help to identify potential failure cases.

Naturally, extending the amount of training data and improving model training could improve model performance, which is already impressive. However, this work relies on access to well annotated clinical trial samples which limits the ability to include more data for training and testing. Our datasets are limited in size, particularly the Salzburg dataset, used here as an external test set. The performance of our method on this limited test set was not up to a standard that could be useful in the clinical setting, even though our method should be able to work on any size of patient cohort. This motivates our next work in Chapter 6, where we develop a domain adaptation method for this model which allows us to be able to better predict on the Salzburg dataset, and can work effectively for any size of dataset.

6

Domain Adaptation

Contents

6.1 Introduction 116

6.2 Related Work 118

6.2.1 Unsupervised Domain Adaptation 118

6.2.2 Histology Domain Adaptation 120

6.3 Methods 122

6.3.1 Source Model 123

6.3.2 Clustering 124

6.3.3 Cluster Triplet Loss 125

6.4 Experiments 128

6.4.1 Data 128

6.4.2 Results 129

6.4.3 Comparison with State-of-the-Art 133
6.4.4 Ablation Studies 134
6.5 Discussion 137
6.5.1 Advantages 137
6.5.2 Limitations 137
6.5.3 Conclusion 139

Contributions

Deep learning models that predict cancer patient treatment response from medical images need to be generalisable across different patient cohorts. However, this can be difficult due to heterogeneity across patient populations. Here we focus on the problem of predicting colorectal cancer patients' response to neoadjuvant radiotherapy from digital histology images scanned from tumour biopsies, and we adapt this prediction model onto a new, visibly different, target cohort of patients.

We present a novel unsupervised domain adaptation method with a Cluster Triplet Loss function, using minimal information from the source domain, resulting in an improvement in AUC from 0.544 to 0.818 on the target cohort. We avoid the use of pseudo-labels and class feature centres to avoid adding noise and bias to the adapted model, and perform experiments to verify the preferable performance of our model over such state-of-the-art methods. Our proposed approach can be applied in many complex medical imaging cases, including prediction on large whole slide images, based on combining predictions from smaller, memory-feasible representations of the image extracted from graph neural networks.

Sections of this work have been published in the proceedings for the workshop on Domain adaptation, Explainability, Fairness in AI for Medical Image Analysis (DEF-AI-MIA) at the Computer Vision and Pattern Recognition (CVPR) 2024 conference [157].

As shown in the previous chapter, our developed graph-based approach to predicting response to radiotherapy (RT) was tested on an unseen dataset, Salzburg, in order to test the translatability and generalisability of our trained model. This small dataset contains a single histology WSI from each of 55 patients, all of whom had capecatibine with RT (CapRT) treatment prior to surgery. Some of these patients went on to have further adjuvant treatment including CapRT or

other chemotherapy treatments such as CAPOX, but analysing the effect of these adjuvant treatments is beyond the scope of this work, yet could be interesting to explore in future. When applying our previously trained model to this new dataset of digital images of the pre-treatment biopsies, to predict whether the patient had a complete response (CR) to the neoadjuvant treatment, our model performed poorly, achieving only 0.54 AUC.

This motivates our work on domain adaptation in the following chapter. We develop an unsupervised domain adaptation (UDA) technique that can help our multi-task graph model to translate onto a new, different, cohort of patients, with the primary aim of predicting only response to RT treatment.

6.1 Introduction

Adapting a deep learning model in the field of medical imaging from one group of patients to another can be challenging, due to the wide variability that can occur between patients. In this work we focus on using deep learning to predict colorectal cancer (CRC) patients' response to RT from a digital histology image of the pre-treatment tumour tissue, and we attempt to adapt this model to a completely unseen cohort of patients from a different geographic region. In this work we focus on UDA, since for this prediction model to be useful in clinical practice we would need to adapt the model without knowledge of the patient outcomes at time of use.

While much research has been done on using domain adaptation in other fields, application to histology images is more challenging due to complications arising from the size and heterogeneity of this imaging modality [158].

Histology slides are the haematoxylin and eosin (H&E) stained, digitally scanned, tumour tissue slices cut from a biopsy sample. These slices are scanned at very high resolution, resulting in extremely large file sizes. Images must be split into smaller sections to fit into computer memory, and a MIL method is then required to combine the predictions into one prediction per slide. Here we present a domain adaptation method which can circumnavigate MIL frameworks by focusing only

on the intermediate feature representation within a model, preserving any optional MIL methods on the resulting features for final outputs. Specifically, we make predictions from naturally segmented tissue regions using a graph neural network (GNN) approach, using the features within the GNN to help adapt our model to a new domain.

While socioeconomic factors could influence patients' experience with cancer in different regions or countries [159], batch effects in histology images can commonly develop from the processing of the tumour biopsy once it is removed from the patient. The process of slicing, staining and scanning the tissue sample is performed slightly differently across medical centres, which introduces an inherent domain shift into the data [160]. Here we train and evaluate our method using three cohorts of patients from different medical centres, all of which use different tissue processing practices.

We approach our binary prediction problem with a generalised view, avoiding pseudo-labels by focusing only on adapting the underlying features to a new domain, and preserving the original classification branches. By avoiding the use of pseudo-labels, unlike many other UDA approaches which may use current model predictions as fixed pseudo-labels or initialisations of learnable ones [161–168], we avoid adding bias and noise from our source model into our predictions.

Furthermore, we avoid the use of class-based clustering to find a cluster representative for each class label, as many in the literature have done [161, 162, 167, 169, 170], to allow for more variance within each class label by clustering on the whole feature set at once, allowing for a natural number of clusters that is not constrained by the number of class labels in the dataset. This approach works much better particularly for binary outcome data since it allows for more than two clusters to represent the entire source dataset.

In this paper we develop a feature-alignment UDA technique to transfer our trained clinical model onto an unseen target cohort without the use of any target labels. We propose a novel approach, defining a loss function to be used in a 'source-supervised' training manner for domain adaptation. This loss only requires a lightweight representation of the source data to guide the learning of a new,

domain-adapted, target model. Our method allows for distributed training of a cohort-tuned model without requiring any training or updating of the original model, therefore providing a secure federated learning technique that can protect patient confidentiality between locations. Rather than confusing the results with all the dataset permutations, we focus on the dataset which is most dissimilar as our target dataset, as this is the biggest challenge. This also mimics application in clinical practice where we would need to transfer frozen pre-trained models onto to new cohort domains to better predict patient outcomes, without advance knowledge of a patient's response to treatment. This method requires no assumptions on the size of the batch or cohort and can be applied to even a single new data point.

6.2 Related Work

6.2.1 Unsupervised Domain Adaptation

Clustering

While many papers have explored the use of clustering for domain adaptation, with various methods of aligning source and domain distributions using contrastive or adversarial loss approaches [167, 171–173], to the best of our knowledge none have used the lightweight clustering approach we suggest here.

The intuition behind our domain adaptation approach builds on the idea of Attracting and Dispersing [163], where the authors aim to bring similar features together and dissimilar features apart in the feature space. This unsupervised method uses k-nearest neighbours and pseudo-labels to maximise consistency of predictions between neighbours, and minimise similarity of dissimilar feature predictions. A similar method, Structurally Regularized Deep Clustering (SRDC) [162], uses K-Means to cluster intermediate network features, but clusters on the target data instead of the source data, though the method is unsupervised so the target data is unlabelled. This method minimises the Kullback-Leibler (KL) divergence between the distributions of the predicted target labels and the true source labels, as well as

the KL divergence between the learnable source and target cluster centres. The loss therefore focuses on both predictions and the intermediate feature representations, since the authors claim that just explicitly aligning the features could lead to worse underlying target discrimination, but this task is made more feasible since this approach uses the target labels in a supervised manner. Another approach using K-Means is the Source Hypothesis Transfer (SHOT) method proposed by Liang *et al.* [161], who freeze the final classifier layer of a source model and use the rest as initialisation for a target model. Their unsupervised approach predicts pseudo-labels and minimises entropy, finding target class centroids in a manner similar to weighted K-Means, and then defining a target sample's pseudo-label by its nearest neighbour class centroid, measured using cosine distance.

Pseudo-labels

Most UDA approaches use pseudo-labels to train their model [161–168], which can provide more information in the multi-class classification setting than the binary one. These pseudo-labels are commonly used for masking or as an indicator method to calculate some further statistic for use in a loss function [163, 166]. Methods using pseudo-labels depend heavily on the teacher model having a prior reasonable accuracy on the target domain, which is not always the case, as pointed out by Li *et al.* [174]. Crucially, they also observe that there are no common methods to evaluate the quality of these pseudo-labels. While many papers acknowledge this caveat and propose methods to counteract it [162, 165, 167, 174], it is a clear inherent design flaw that can add unnecessary bias and noise. Zhang *et al.* acknowledge this and regularize their pseudo-labels with weights during training, by measuring distances to feature centroids of classes [165]. The Divide and Contrast method divides the target data into source-like or not, and makes the reasonable assumption that pseudo-labels from source-like target data are more accurate than those from target-specific samples [167]. The SRDC authors initially try an approach that adopts the information maximisation loss, but admit that using this loss alone, the unreliability of the source model on the target data could lead to some wrong target predictions.

To counter this they add an extra term onto their loss function using pseudo-labels as an indicator on the predicted labels from the training target model [162].

Triplet loss

The idea of a triplet loss using central features was first introduced by [169] for object retrieval, where they propose a Triplet Centre Loss (TCL) to align features of the same class to a learnable class centre, and repel features from different classes. They use Euclidean distance to measure the difference between the class centre and sample features, as we do here, though for their negative sample in their triplet loss they choose the closest negative centre. They also use class labels to identify the corresponding class centre, so the method is not unsupervised. Other works have used a similar approach using a triplet loss on feature centres [170, 175–177], across different fields. Most focus on calculating the feature centres from pseudo-labels to find a centre representing each class in a classification problem [170, 175, 176].

The Centroid Triplet Loss proposed by Wieczorek *et al.* for image retrieval [175] uses a traditional triplet loss on the target features with the positive example as the centroid of the class of that target example, and the negative example as the centroid of a negative class, which is similar to what we propose here, but differing in our exclusion of any assumed or known class information. Lagunes-Fortiz *et al.* [170] use a different negative sample in their triplet loss, using a sample from the domain itself instead of a feature centre. The triplet loss has also been used to define target and source clusters as class guided constraints [164], for better class alignment between the domains.

6.2.2 Histology Domain Adaptation

Staining

In the field of deep learning on histopathology, tissue staining and processing can vary heavily across hospitals and laboratories, and efforts have been made to counter these cohort staining effects [178–181] beyond traditional colour normalisation methods

[182, 183]. However, sometimes this approach alone is not enough to guarantee domain generalisability of a model. Lafarge *et al.* [178] propose a domain-adversarial neural network (DANN) to predict the probability that a sample comes from a particular domain, allowing removal of domain-specific features while maintaining those features which are useful for prediction. They also experiment with traditional staining domain adaptation methods, and their best results are achieved when the DANN is used in addition to colour augmentation or stain normalization.

Feature alignment

In this work we focus on feature alignment between the source and target domains. Of the feature alignment approaches in the field of histopathology that use a cluster-based approach, most use pseudo-labels to find a class prediction which can help to update class-wise feature centres [184, 185]. Distill-SODA [185] is one such source-free UDA method that performs Monte Carlo simulations of its clustering for robustness. Similar to our method, they calculate a cluster centroid to compare with target features in their loss function; however their centroids are not label-agnostic but are constrained to one per class, instead of naturally deriving them from the source domain. Another feature-alignment approach introduced by Jian *et al.* [186] trains a CNN to map target images into the source model feature space, minimising the difference between domains. This method goes further to introduce a Siamese model to encourage patches from the same WSI to be classified with the same label, but this approach does not account for naturally occurring heterogeneity within the tissue sample. Wang *et al.* [187] focus on using GNN node features for alignment of CRC histology images for nuclei detection using an adversarial loss. Abbet *et al.* [188] use few source labels to train a model for CRC tissue classification.

Binary classification

Most research focuses on multi-class classification or segmentation problems, where pseudo-labels or class-centres can provide a higher quantity of information. Some works focus on binary classification problems such as epithelium-stroma classification,

with one paper training a single model on source and target at once and adapting the kernels of a CNN to the target domain using a simple vector multiplication of the eigenvectors corresponding to the largest eigenvalues from the target and source domains [189]. Qi *et al.* [190] also work on epithelium-stroma classification and apply a curriculum learning approach, measuring cosine similarity between samples and class centroids to avoid samples that are more likely to give false pseudo-labels, selecting initial training samples based on maximum distance to source domain.

Li *et al.* [191] focus on classifying tumour as benign or malignant on breast, lung and colon cancer histology slides. Despite the lack of outcome classes they do, however, have multiple dataset cohorts, and so their UDA approach trains a separate feature extractor on each source and target domain, and uses the source labels to learn alignment of the feature distributions. Optimal transport has also been used to penalize domain prediction in a binary classification of tumour vs normal tissue [192]. We found no previous research on UDA for models which predict patient treatment response from histology images.

Triplet loss on histology

Very little research has applied triplet loss for domain adaptation on histology, and even less for unsupervised approaches. Sikaroudi *et al.* [193] use triplet loss in their efforts to learn hospital-agnostic histology representations, again focusing on the class-conditional shift across domains. They take a supervised approach with a cross entropy loss on the target predictions, as well as KL divergence to align feature domains, and a metric loss to separate classes.

6.3 Methods

This work assumes we already have a pre-trained source model which we wish to adapt to a new domain. We describe the source model below, which is building on a similar previous model in this field [148] that was introduced in Chapter 5, and then explain how we train a new model (using the weights of the source model at

initialisation) to adapt the prediction to a new domain. We explain the clustering approach used on the source data to extract a lightweight representation of the source data, which is then used in our proposed Cluster Triplet Loss function to train and adapt the new model.

We first introduce some terminology. The source data x_s and source model $\mathcal{M}_s = \mathcal{H}_s(\mathcal{F}_s)$ define the data on which the corresponding original model was trained and validated on, where $\mathcal{H}_s$ is the classifier part of the model and $\mathcal{F}_s$ is the feature part of the model which we use in this work. The target model $\mathcal{M}_t = \mathcal{H}_s(\mathcal{F}_t)$ is an updated version of the source model that is trained here on the target data x_t, the new unseen dataset whose domain we are trying to adapt to, where we use the same classifier from the source model $\mathcal{H}_s$ but update the feature part of the model for $\mathcal{F}_t$.

6.3.1 Source Model

Our source model is a GNN with three Graph Isomorphism Network layers [87] of feature sizes 64, 32 and 16. Instead of feeding our WSI straight into this GNN, we first apply a superpixel method on the WSI and then calculate superpixel features from patch features in the same region (size $[1, 768]$) [148], extracted using the self-supervised pre-trained large histology model CTransPath [116]. From these superpixel features we construct a graph representation of each WSI, where the nodes and node features are defined from the superpixels and the edges of the graph are defined by nearest neighbours using Delaunay triangulation. These graphs are then used as input to the GNN, which is trained in a weakly supervised manner to predict a patient's response to RT.

On the source validation dataset the source model achieved metrics of 0.93 AUC, 0.80 balanced accuracy and 0.89 weighted F1, as seen in Table 6.6. Evidently our source model can perform well on the source cohorts, and while efforts were made to generalise this model in training, the application of this model on an unseen test cohort demonstrates the inadequate generalisability of the model with metrics of 0.54 AUC, 0.50 balanced accuracy and 0.84 weighted F1, as seen in Table 6.2. Efforts made to avoid overfitting on the training cohorts include extensive data

augmentation on the training images prior to extracting features, heavy dropout in the GNN and classification branches ($p = 0.5$), training on more than one geographic cohort of patients, and applying a multi-task learning approach to ensure the final feature set includes information on molecular traits and spatial tissue architecture as well [148].

For this work we are only concerned with the intermediate feature representation, not the final prediction stage of the model. When training our new domain-adapted model we freeze the classification branches on our target model (of which there are multiple due to a multi-task learning approach with the source model, where one of these branches predicts the patient's response to RT), and we train only on the GNN layers before this. Hence in this work we focus on the node-level features of our dataset, rather than the slide-level features. We refer to the node feature extractor part of the source model as $\mathcal{F}_s$, and the classifiers after this remain fixed across the source and target models.

6.3.2 Clustering

We use clustering on the source data to extract a lightweight, high-level representation of the source data feature set. GNNs provide us with the node-level predictions from the superpixel nodes, providing an intuitive, natural representation of tissue segments within the tumour. We extract the features of these nodes from the final layer in our GNN before it splits into three prediction branches for the multi-task learning approach.

We apply our clustering approach to the normalised concatenated set of node feature vectors from the source data cohorts seen in training. The concatenated feature vectors are of size $[N, 16]$, where $N = 134,132$ is the total number of nodes and 16 is the number of features per node. To find the optimal number of clusters, k_{opt}, we calculate the silhouette width [194] of the clustering for the number of clusters $k = 2, \ldots, 20$. We select the number of clusters as the cluster in this range with the highest silhouette width and Calinski-Harabasz index [195], and lowest David Bouldin score [196] for the most distinct clusters in an unsupervised

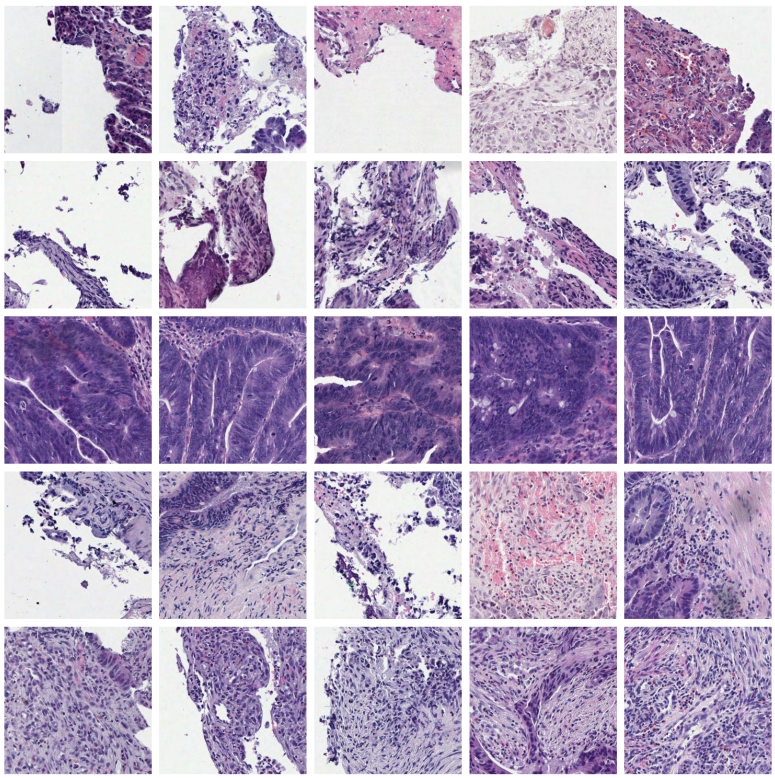


Figure 6.1: Nearest neighbour tissue segments for each of the five optimal clusters found on the source data features. Each row represents a single cluster centre, containing the five nearest neighbours when comparing the optimal cluster centres C to the extracted features of the source data x_s.

setting. Due to the large sample size we use the K-Means MiniBatch approach, implemented in the Python library sklearn.cluster (version 1.1.3) [197]. We fit the MiniBatch K-Means on a subsample ($n = 10,000$ node features) of our source dataset for efficiency, using the optimal number of clusters. We extract the resulting cluster centres C of size $[k_{opt}, 16]$.

6.3.3 Cluster Triplet Loss

To train and adapt our model onto the target dataset, we propose the Cluster Triplet Loss, which makes use of the source clustering from the previous section.

Our proposed Cluster Triplet Loss works on a per-sample basis, meaning it can be used to adapt a model to any size of cohort. For each feature vector provided, it calculates the mean squared error loss between the feature vector and the fixed

Algorithm 2: Training with Cluster Triplet Loss

Input : source feature model $\mathcal{F}_s$, source data x_s, target data x_t

Extract source features $\mathcal{F}_s(x_s)$ from final layer of GNN before classification;
Run K-Means on $\mathcal{F}_s(x_s)$ for $k = 2, ..., 20$ clusters and calculate optimal k_{opt} using silhouette width;
From best K-Means extract k_{opt} cluster centres C;
Initialise target model $\mathcal{F}_t$ with weights from source model $\mathcal{F}_s$;
while *Training* **do**
    Extract target features from target model, $\mathcal{F}_t(x_t)$;
    Calculate Euclidean distance from $\mathcal{F}_t(x_t)$ to each cluster centre in C with Eq. (6.1);
    Find closest (C_{pos}) and furthest (C_{neg}) clusters to target features using distances with Eq. (6.2);
    Calculate mean triplet loss for $\mathcal{F}_t(x_t)$ with Eqs. (6.3) and (6.4) over the batch and backpropagate
end

Output : adapted target feature model $\mathcal{F}_t$

source cluster centres, akin to one iteration of the traditional K-Means algorithm. From this we select the closest and furthest cluster centres to our input feature vector, and give these as the positive and negative samples in the calculation of the triplet loss, with the input feature vector as the anchor, to move the feature vector onto the cluster domain while simultaneously clustering the sample. We vectorize and apply this method simultaneously on all feature vectors from the model training batch. In our triplet loss implementation we use a margin of 1 and we swap the distance between the input and the negative cluster centre with the distance between the positive and negative cluster centres, as proposed by Balntas *et al.* [198].

We first define the source model $\mathcal{M}_s = \mathcal{H}_s(\mathcal{F}_s)$, where $\mathcal{H}_s$ is the classifier part of the model and $\mathcal{F}_s$ is the feature part of the model which we adapt to a new domain. We define the target model $\mathcal{M}_t = \mathcal{H}_s(\mathcal{F}_t)$, where we use the same classifier from the source model, $\mathcal{H}_s$, but update the feature part of the source model to get $\mathcal{F}_t$. Hence the source model and target model have the exact same model architecture but different model weights.

In our proposed Cluster Triplet Loss function, we start with the cluster centres, C, from the optimal clustering of the source data. Taking our input target data,

x_t, in a batch of size b, we calculate the Euclidean distance d_{ij} between the input passed through the model and each cluster centre,

$$d_{ij} = \|\mathcal{F}_t(x_{t_i}) - C_j\|^2, \tag{6.1}$$

where $i = 1, \ldots, b$ denotes each node input within the batch.

We use these distances to find the closest ($C_{j_{pos}}$) and furthest ($C_{j_{neg}}$) cluster centres, using

$$j_{pos_i} = \arg\min_j d_{ij}, \quad j_{neg_i} = \arg\max_j d_{ij}. \tag{6.2}$$

We use these positive and negative cluster centres in our adjusted triplet loss function, as defined by

$$L_i(x_{t_i}) = \max\{\|\mathcal{F}_t(x_{t_i}) - C_{j_{pos_i}}\|^2 - \|C_{j_{pos_i}} - C_{j_{neg_i}}\|^2 + \mu, 0\} \tag{6.3}$$

using the margin $\mu = 1$.

Finally we reduce the output by taking the mean over our batch, and back-propagate through the model $\mathcal{F}_t$ with the batch loss

$$L_b(x_t; C, \mu) = \frac{1}{b} \sum_i L_i(x_{t_i}, j_{pos_i}, j_{neg_i}; C, \mu), \tag{6.4}$$

where the cluster centres C and margin μ are fixed, but j_{pos_i} and j_{neg_i} vary depending on Equations (6.1) and (6.2).

The algorithm for our whole method can be found in Algorithm 2. Steps 1-3 need only be performed once, and then, given the source data representation C, steps 4 onwards can be used to train any number of target models on different domains.

Cohort	CR	NoCR	% CR/Total	Total
Aristotle	24	97	20%	121
Grampian	61	186	25%	247
Salzburg	6	49	11%	55

Table 6.1: Slide counts split by outcome (CR - positive, complete response to radiotherapy, NoCR - negative, no complete response to radiotherapy) across patient cohorts.

6.4 Experiments

6.4.1 Data

For our experiments we have three private CRC histology datasets, Grampian, Aristotle and Salzburg, all from different geographic locations in Europe. For all datasets we have the digital WSIs of the H&E stained tumour tissue taken from pre-treatment biopsies. For Grampian and Aristotle we have the patients' recorded response to neoadjuvant RT treatment, categorised as pathological CR if there are no tumour cells remaining after the treatment course is completed, or defined as no complete response (NoCR) if any number of tumour cells remain post-treatment. For the Salzburg data we define CR to RT as having a Dworak tumour regression grade of 4, post-treatment. See Chapter 3 for more details. In this work we aim to predict the response to RT as our primary binary outcome. The outcome response counts across cohorts are given in Table 6.1, where we can see that the ratio of positive to negative outcomes (% CR/Total) is similarly imbalanced across all cohorts.

Two of these cohorts, Grampian and Aristotle, were used for training our original source model, with the WSIs from roughly 30% patients in each cohort used for validation, and the rest used for weakly supervised training. The third cohort of patients, Salzburg, is introduced for this work as our target dataset, previously unseen by our model in training and validation. Hence we refer to Grampian and Aristotle as our source data, and Salzburg as our target data.

The differences between the cohorts can be visualised in the reduced dimensionality UMAP projection [199] in Figure 6.2. For each WSI in the cohorts we extract the unsupervised CTransPath features [116], which we use as input to our models.

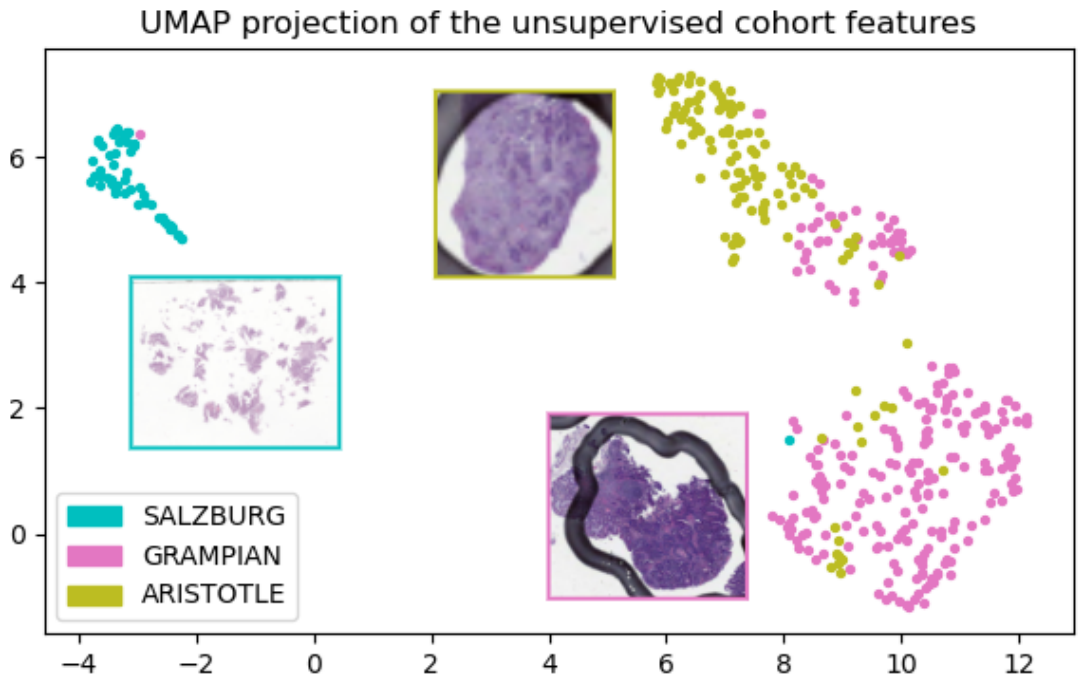


Figure 6.2: UMAP projections of unsupervised CTransPath features from our different patient cohorts, using the mean features per WSI. Our target dataset in this work, Salzburg, is clearly very different from our source cohorts, Grampian and Aristotle. For each cohort we overlay a region of a randomly sampled WSI in that cohort, shown in a box of the same colour, to help visualise the cohort differences.

We fit a UMAP on the mean features per WSI, and plot the resulting embeddings, colouring by cohort. Our target cohort, Salzburg, is clearly very different to our two source cohorts, Grampian and Aristotle, and we observe the trend of sparse biopsy specimens across the Salzburg data.

6.4.2 Results

Clustering

Applying our clustering method to our source data, we find $k_{opt} = 5$ optimal cluster centres in the feature space with the highest silhouette width of 0.28. These clusters can be visualised in Figure 6.1, where for each of the optimal five clusters we have plotted the five nearest neighbours to the cluster centres from the source data.

Training target model

We use the weights from our source model to initialise a new target model, as described in Section 6.3.1. In training the target model we use heavy training data augmentations using the Pytorch torchvision.transforms library (version 0.13.1) as follows: resize, random vertical flip ($p = 0.5$), random horizontal flip ($p = 0.5$), colour jitter (brightness 0.1, contrast 0.25, saturation 0.5 and hue 0.25), Gaussian blur over a kernel of size 9, random adjust sharpness ($p = 0.2$), random auto contrast ($p = 0.5$), rotation by multiples of 90 degrees and normalizing the colour channels. We use the Adam optimiser with a learning rate of $1e{-}3$ with weight decay $1e{-}4$. We use a batch size of 32 and train for 30 epochs to avoid overfitting to the new domain.

Method results

The results from our proposed method can be seen in Table 6.2, which shows the mean and standard deviation of metrics from five separate seed rounds of training, each initialised with a different random seed. Where required, we use the unoptimised threshold of 0.5 for metric calculations for fair comparison across experiments. Our domain adapted model achieves an AUC of 0.82 and a balanced accuracy of 0.62 on the target dataset, improving over the source model by +0.27 AUC and +0.12 balanced accuracy, demonstrating the effectiveness of our proposed method on this complex real world dataset.

Visualising domain shift

The differences in the intermediate model features before and after domain adaptation can be visualised by plotting a UMAP embedding of the node features in Figure 6.3. We randomly subsampled the source data for balanced outcomes to better visualise the shift. The feature embeddings are coloured by both domain and outcome, specifying whether the data is from the source or target domain, and whether the patient outcome is a positive CR or a negative NoCR. We have plotted the UMAP embedding of the fixed source cluster centres in black, which can be

Models	AUC	BAcc	F1
Source	0.544	0.500	0.840
Distill-SODA	0.511±0.00	0.461±0.00	0.736±0.00
TCL	0.684±0.01	0.605±0.04	0.860±0.02
SHOT	0.578±0.00	0.543±0.00	0.830±0.00
SRDC	0.498±0.11	0.500±0.00	0.840±0.00
Ours	**0.818**±0.04	**0.619**±0.04	**0.878**±0.02

Table 6.2: Results for our methods: 'Source' model with no domain adaptation, comparison state-of-the-art unsupervised domain adaptation methods on the target dataset, and 'Ours' applying the Cluster Triplet Loss proposed in this paper. Metrics provided are the mean and standard deviation of the AUC, balanced accuracy (BAcc) and weighted F1 score (F1) over five seed rounds.

useful as fiducial markers across the two plots since they aren't updated after domain adaptation. The top scatter plot shows features extracted from the source model, and the bottom scatter plot shows features extracted from our adapted target model.

The top plot in Figure 6.3 shows the Source CR (purple) and Source NoCR (red) classes are reasonably separated, demonstrating the competency of our source model on the source data. Before adaptation, the target outcomes (orange and green) are mixed in with each other, and the Target CR (green) shows no overlap with the Source CR (purple). However, after domain adaptation, the Target CR (green) features have moved towards the Source CR (purple) domain, better aligning the features for this minority class across domains.

Quantifying domain shift

We measured the distance between our target features and our source cluster centres before and after domain adaptation. Measuring the distance from the target data to the *closest* cluster centre, the mean distance over the target data decreased from 0.146 to 0.137 after our domain adaptation method (−0.009). However, measuring the distance from the target data to *all* cluster centres, the mean distance increased from 1.100 to 1.141 (+0.041). This highlights how our loss function is designed to both pull the nearest cluster centre closer, but also push other cluster centres

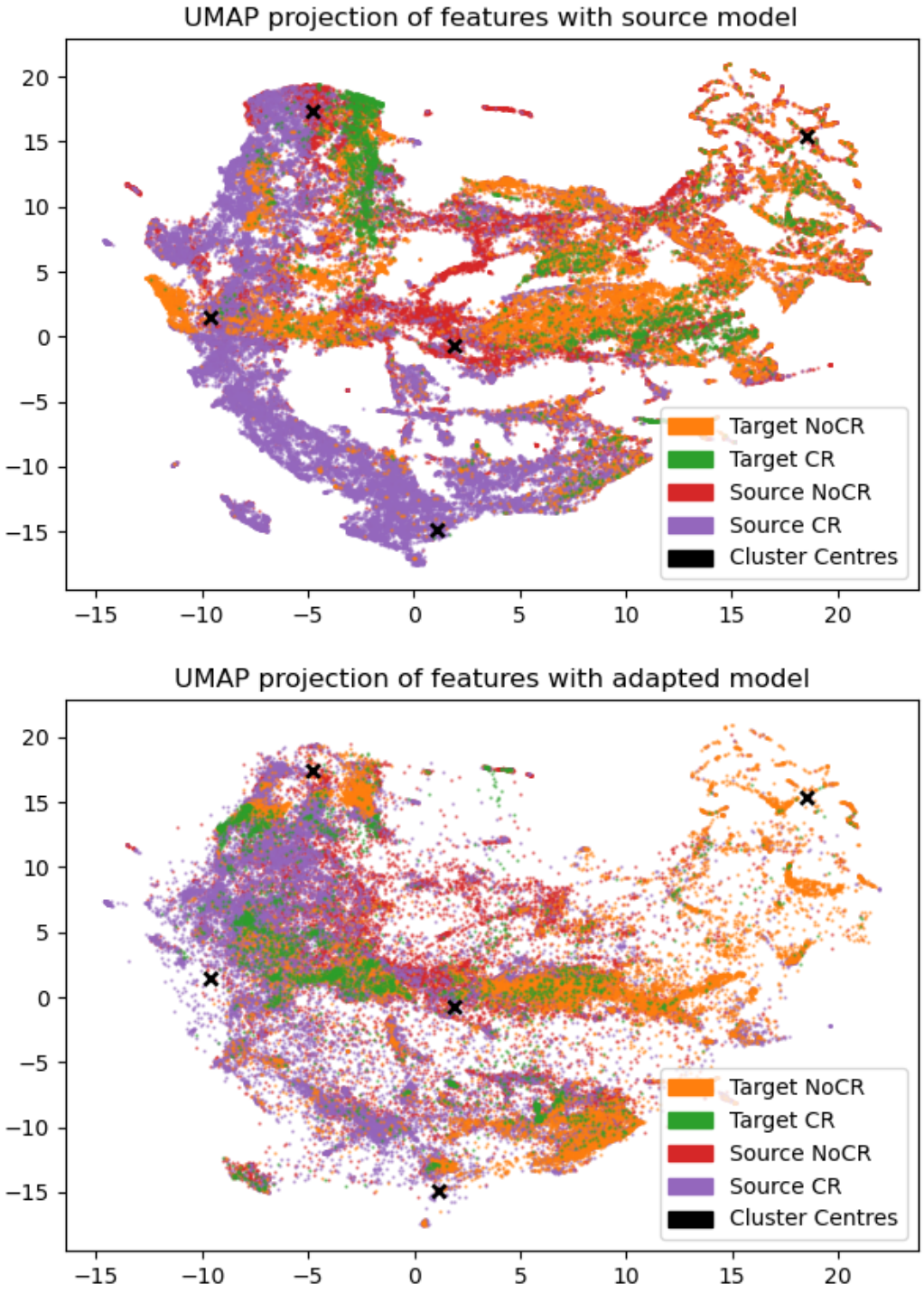


Figure 6.3: UMAP projections of our intermediate model features before (top) and after (bottom) applying our UDA method. Features are coloured by the source or target domain and positive (CR) or negative outcome (NoCR). The five stationary source cluster centres are overlaid in black. The target features across domains are better aligned after domain adaptation.

further away, helping to create more distinct clusters in the feature set with the idea of guiding the model to an easier classification decision.

6.4.3 Comparison with State-of-the-Art

As well as comparing our proposed method to our baseline source model, we also implemented select state-of-the-art (SOTA) UDA methods for further comparison. Due to the intricate nature of most published methods, we chose to implement only those which had their code publicly available online.

We implemented four UDA methods, Distill-SODA [185], SHOT [161], TCL [169] and SRDC [162]. The results can be found in Table 6.2. Distill-SODA is the only method here which was specifically introduced for histology images, whereas the other methods are for general computer vision or other fields. When choosing the best epochs to evaluate results for each model, we either chose the final epoch as defined in the papers or the epoch with the lowest training loss if unspecified.

In all of our SOTA implementations we had to adapt the method to our prediction problem. Firstly, these methods were implemented for multi-class classification or segmentation, so we had to adapt the code for a binary prediction problem, which in some cases meant there was less information from class pseudo-labels (since there are maximum two classes here). Secondly, most of these methods use either the original image or a pixel-level representation as input, so we had to adapt the methods to work on features of segmented tissue regions within each image.

Though the SRDC method is unsupervised [162], it requires the full source dataset, including source labels, during training, and the target labels are used to validate the model during training. Hence for this method we use the WSI label as the label for each individual segmented tissue region within the WSI.

In our implementation of Distill-SODA [185], we were unable to use their proposed adversarial data augmentation method, AdvStyle, since our source model is frozen for this research, and we aim to adapt it as is. It's possible this method could work better on our data if we were to use this pre-training method on our source data.

In the implementation of TCL [169], this method is originally introduced in a supervised setting, using the target labels to identify which class centre it should be using as the positive sample in the triplet loss function. We implement an unsupervised variant of TCL, using our fixed source cluster centres instead of learnable class centres, but keeping the idea that the negative example in the triplet loss function is the nearest negative cluster centre i.e. the second closest cluster centre, where the closest is our positive example.

Overall our method has the best metrics compared to all other UDA methods implemented here.

6.4.4 Ablation Studies

Number of clusters

For the following ablation studies we trained each model variation over five different random seeds and averaged the results. We experimented with the number of clusters and scaling the cluster centres before use in the loss function. The results from using different numbers of clusters can be found in Table 6.3. We found that the optimal number of clusters from our clustering analysis achieved the best results compared to other numbers of clusters. Scaling the cluster centres $\in [0, 1]$ didn't improve results either, achieving 0.747 AUC over five rounds with the optimal number of $k_{opt} = 5$ clusters.

Clustering methods

To find the representative cluster centres from our source dataset we also tried a consensus clustering approach with hierarchical clustering. Despite the theory behind this method we found it not to be robust, with unrealistic cluster distributions. Furthermore, due to the large sample size of the node-level data, we had to drastically reduce the size of the dataset before applying the consensus clustering, which could lead to loss of information. The number of features was small enough to ensure that dimensionality reduction on the number of features was not required. We tried

Clusters k	AUC	BAcc	F1
$k = 3$	0.665	0.514	0.809
$k = 4$	0.748	0.593	0.845
$k = 5$	**0.818**	**0.619**	**0.878**
$k = 6$	0.601	0.583	0.818
$k = 7$	0.745	0.611	0.868

Table 6.3: Results from an ablation study on changing the number of clusters used for the cluster centres in our Cluster Triplet Loss function, giving average AUC, balanced accuracy (BAcc) and weighted F1 score (F1) over five seed rounds. The optimal number of clusters found from our clustering analysis, $k = 5$, demonstrates the best results compared to values of $k \pm 2$ from the optimal k.

a self-organizing map (SOM) to reduce the dimension of the data, with different map sizes, and extracted the resulting cluster labels from the consensus clustering on the reduced SOM sample set. To calculate the cluster centres, for each cluster label we then collected all the SOM samples with that label, and calculated the mean of their feature set. However, we found the traditional K-Means approach more robust, explainable and efficient for finding the cluster centres of the source data. The only downside of K-Means compared to hierarchical consensus clustering is that the user must choose the number of clusters before applying the model, but this can be selected in a methodical way using clustering metrics.

Removing a source cohort

We ran experiments where we removed one of the two source cohorts before calculating the source cluster centres, and then trained our model on the target cohort using the reduced cohort clusters in our loss function. The results averaged over five rounds can be shown in Table 6.4, excluding Grampian and Aristotle in turn. These results demonstrate the importance of including both source datasets in the training set of the source model. We would expect the source model to be more generalisable when trained on more than one cohort domain, and these results show that such a model can be better adapted to new domains using our domain adaptation method.

Excluded Cohort	k	AUC	BAcc	F1
None	5	**0.818**	**0.619**	**0.878**
Grampian	2	0.650	0.490	0.830
Aristotle	5	0.573	0.527	0.768

Table 6.4: Results from an ablation study on removing a source cohort for the calculation of the cluster centres used in our Cluster Triplet loss function, averaging results over five seed rounds. Metrics provided are AUC, balanced accuracy (BAcc) and weighted F1 score (F1). We also provide the optimal number of clusters k found from clustering without the specified cohorts.

Models	k	AUC	BAcc	F1
Source	-	0.544	0.500	0.840
Source with Stain Norm	-	0.646	0.573	0.866
Ours	5	**0.818**±0.04	**0.619**±0.04	**0.878**±0.02
Ours with Stain Norm	5	0.757±0.04	0.540±0.04	0.852±0.02

Table 6.5: Results from an ablation study using Vahadane stain normalisation. Source with Stain Norm shows the source model trained on data with Vahadane stain normalisation [183], applied on the target data. Ours with Stain Norm shows our UDA approach with Vahadane stain normalisation [183] on the target data. We show the Source model with no domain adaptation and our best model using the Cluster Triplet Loss proposed in this paper with the optimal number of clusters $k_{opt} = 5$, Ours, for comparison. Metrics provided are the mean and standard deviation of the AUC, balanced accuracy (BAcc) and weighted F1 score (F1) over five seed rounds.

Staining

For comparison, we used Vahadane stain normalisation [183] on the target data, which has been show to be an effective technique in histology domain adaptation [178]. The source model predictions were better with stain normalisation than without (see *Source with Stain Norm* results in Table 6.5), but still do not match the results from our proposed UDA method. We applied our UDA method on the stain-normalised target data (see *Ours with Stain Norm* results in Table 6.5), but it did not show any improvement over the standard implementation.

6.5 Discussion

6.5.1 Advantages

This approach was developed with clinical application in mind, and works particularly well for situations where no labels are available in the target dataset. For example, if we have an existing trained and validated model that can predict a patient's response to RT, and we want to use that on completely new cohort of patients to predict their respective response to treatment, we can do so in real time without knowing in advance what the responses will be.

In the case where another domain adaptation method may need labels from the target domain for training, they may also require a substantially large subset of labelled data in order to do so. In contrast, our approach can adapt our model to a new domain consisting of only a single sample, and so does not need a large cohort with corresponding labels in order to be applicable.

6.5.2 Limitations

Following on from the advantages, where we highlight that we can apply our method to a new, unlabelled, cohort consisting of only a single sample, it's important to mention that in the case where enough new labels are available, it would be worth exploring the approach of re-training the original model instead of applying our domain adaptation technique.

We acknowledge that our adapted model is only trained up to the point of feature extraction, meaning the classification branches for the prediction of outcomes from these domain-shifted features are not updated. Since we are shifting the feature domain onto that of the original source features, on which the existing classification branches were trained to predict from, this part of the model should adapt without further training. However, there could be some useful cohort-specific information being missed in this final step.

Original Validation Data	AUC	BAcc	F1
Original model (teacher)	**0.931**	**0.803**	**0.885**
Adapted model (student)	0.736	0.686	0.799

Table 6.6: Results from applying the adapted model onto the original source domain data. Metrics provided are AUC, balanced accuracy (BAcc) and weighted F1 score (F1), with the metrics for the adapted model averaged over the five seed rounds for which those models were trained.

Applying adapted model to original domain

Using our adapted model, which was trained to shift the feature set of the Salzburg data onto the domain of the original training and validation datasets, Grampian and Aristotle, we retested how our adapted model performed on the original source data. Since the weights of the final MLP layer were not considered during this domain adaptation training, we use the weights for the final classification layer from the original teacher model, and the rest of the weights from the adapted student model.

Testing only on the same validation set as the original teacher model, the adapted model achieved reasonable metrics with the default thresholds of 0.5, however it performed worse on the data than the original model did. The results can be seen in Table 6.6, showing that the performance in terms of AUC dropped from 0.93 to 0.74, now also worse than the adapted model performs on the target dataset.

This demonstrates a downside of our method, that when implementing the user may have to use different models depending on the domain. However, this could equally be seen as an advantage in that one can fine-tune a model that can work well on a new domain in an unsupervised manner. It's also possible that the adapted model has been overfitted to the new domain, demonstrated by the fact that the adapted model performs better on the new domain than the original one. This could potentially be counteracted by early stopping in training or a further constraint on the proposed cluster triplet loss function.

6.5.3 Conclusion

We propose a novel method that uses graph node features and source cluster centres in a Cluster Triplet Loss function for UDA of a histology deep learning model. Our approach allows for local domain adaptation within the WSI so that different tissue sections in one target image do not have to be 'shifted' by the same amount.

Whilst our proposed method is not entirely source-free, we require only a dense representation of the original source data, which avoids having to store the memory intensive source dataset and would preserve patient data anonymity if implemented in different hospital settings. This cluster centre summary of the source data could easily be transferred between hospitals or academic institutions, since it is relatively a very small representation and could even be shared via email. This method is generalisable across any number of outcome classes and any number of samples, and can be applied to multiple different deep learning and MIL approaches.

7

Conclusion

Contents

7.1 Contributions 141

7.2 Translation to Clinic 141

7.2.1 Patient Perspectives 141

7.2.2 Biomarker Validation 143

7.3 Limitations 144

7.3.1 Data 144

7.3.2 Research 144

7.4 Future Work 145

7.4.1 Tumour Microenvironment 145

7.4.2 Temporal Modelling 145

7.4.3 Domain Adaptation 146

7.4.4 Patient Treatments 146

7.1 Contributions

In this research we have tackled the problem of predicting how CRC patients respond to radiotherapy (RT) treatment, in order to make better individualised patient treatment decisions. We have use deep learning tools to do so, as these can be efficiently applied on readily available H&E stained digital biopsy slides, taken as part of routine clinical care. We have thoughtfully considered how we can make contextual predictions, tailoring our methods to the nature of the histology slide. Our work has focused on the practicalities of implementing such a prediction in a clinical setting, developing novel interpretability methods and domain adaptation techniques.

In Chapter 4 we proposed novel variations of the ViT model to enhance it for this modality, incorporating spatial information and tissue morphology to achieve the first deep learning model that can predict response to RT from histology slides of CRC. In Chapter 5 we brought in additional context to the therapy response prediction and developed a novel pipeline for interpretability, using multi-task learning on a meaningful graph representation of the WSI to predict molecular traits and spatial distributions of the tissue in parallel to response. Finally, in Chapter 6 we address the generalisability of our model and developed an unsupervised domain adaptation technique to drastically improve performance of our model on an unseen test set.

7.2 Translation to Clinic

7.2.1 Patient Perspectives

While traditionally we think about the clinician as the user of the medical imaging AI tools we develop, particularly in this work, it is also important to consider how the patients themselves will interpret such tools. With this in mind, I presented my research to a representative cancer Patient & Public Involvement (PPI) group.

In a short presentation, I explained the methods I developed in Chapters 4 and 5 to predict a CRC patient's response to RT treatment. I explained the problem of stratifying patients for treatment, why deep learning could be useful in solving this problem, and how my research specifically aims to develop deep learning methods in context of the histology images and cancer biology. I also explained how my methods have been made more interpretable to clinicians, in the first case by visualising the attention scores of the proposed ViT models, and in the second case by implementing multi-task learning approaches and providing intuitive graph node visualisations. Finally, I stressed the importance of thorough testing of trained models on unseen, external patient cohorts, in order to ensure generalisability of the approaches across a population.

I explained my methods in a way that could be understood by a non-technical group, and got feedback on whether the visualisations produced from my work could be useful to patients as well as pathologists. The PPI group provided very positive feedback on my methods, and complimented my ability to explain everything in plain and accessible language. I also took the opportunity at the end of my presentation to raise discussion questions, including the following:

- Do you think such visualizations could ever be useful to patients?
- How do you feel about AI being used to determine treatment decisions?
- Are the explainability features of the model valuable or would you be fine with a black box model?

Overall the response was positive, giving the impression that patients would be happy for AI to be used in their patient treatment decisions if it has been tested and shown to be accurate. They wisely commented that they don't fully understand other common processes such as blood tests, but are content to trust the clinician's interpretation and knowledge of such results. Therefore, an important focus of clinical translation should be to ensure clinicians are well trained in the understanding and application of clinical AI models.

7.2.2 Biomarker Validation

When considering translating this work into the clinic, we can evaluate it with regards to the stages of biomarker development outlined in the Literature Review in Chapter 2. The research done here has shown early feasibility for an imaging approach which can stratify CRC patients for a particular chemoradiotherapy treatment.

Our method is practical with regards to the resources required and the cost, since our approach is evaluated on a histology slide of the pre-treatment tissue biopsy, which is taken as part of the existing patient pathway. These histology slides are not always scanned digitally, which prohibits the use of our method, though active parties are encouraging hospitals to introduce this scanning stage into their standard clinical practice routine. Other similar approaches which try to quantify response to treatment are often times based off RNA-sequencing, which is far more costly and time-consuming.

We have also worked towards the requirement of generalisability across demographic and geographic patient domains, which can have an effect on both the tumour and the histology slide. Though our method did not generalise well to an unseen cohort in Chapter 5, in Chapter 6 we developed a method which proved effective at adapting our method to a new unseen patient domain, a method which can be applied to any new domain regardless of the number of patients in a cohort.

Finally, our approach offers innovative insight into the model decision-making process, which can inform clinician's decisions in context of the usual factors which inform patient treatment. We argue that these contributions could help clinicians both understand and adapt to using our imaging biomarker in clinic.

All this being said, further validation, and potentially development, of the proposed methods on multiple and extensive patient cohorts is crucial before any prospective clinical trials should be considered. It should also be considered exactly how the binary prediction should be used in a clinical setting, alongside the graph visualisations, and the choice of the threshold for the binary outcome should be optimised and statistically validated, whilst considering the impact of false positives and false negatives on patient outcomes. Finally, we would have to consider the

temporal aspect of this prediction of complete response (CR) to RT, and whether recurrence of cancer should be considered and at what point in time, to ensure that the prediction is being used with transparency.

7.3 Limitations

Here we discuss limitations of this work, with regards to the imbalanced and limited datasets, and the scope of the research.

7.3.1 Data

This research is very much constrained by the size of the datasets. Though there are some CRC histology slides publicly available which have not been used here, this is because these slides do not have recorded corresponding patient treatment data, which is crucial for this work. We have optimised the use of the three datasets available to us, using two for training and development of the model and the third for testing and domain adaptation.

Furthermore, the data we had was imbalanced with regards to the outcome, since most CRC patients considered here do not have a CR to RT. As touched on briefly above, it would be crucial to avoid predicting false positives here, predicting that a patient has a CR to RT and therefore requires no further treatment, when in fact they will not have a pCR and will require further attention to treat the tumour. It follows that correct prediction of the minority CR class is extremely important, and therefore to gather enough power for any statistically significant results regarding CR prediction we would need many more data points with this label.

7.3.2 Research

It could be said that a limitation of the research done here is the lack of a clinical trial, whether that be in the planning phase or in the active phase. However, this is considered outside the scope of this research project, and regardless, we

have already discussed how further development and validation of our approach could be performed prior to a trial.

7.4 Future Work

7.4.1 Tumour Microenvironment

Building on the work in Chapter 5 where, with multi-task learning, we predicted the response to RT and presence of CMS4 and epithelial tissue in parallel, we could further explore the tumour microenvironment by predicting other molecular traits here. In their recently published work, Domingo *et al.* found that stromal cells, as captured by CMS4, and immune response, as indicated by CMS1, are predictors associated with CR in CRC. Hence, further work could explore predicting CMS1 in this pipeline, which would demonstrate presence of cytotoxic lymphocytes in the tumour microenvironment, another known prognostic feature that can be predicted from histology slides [20, 46, 57].

7.4.2 Temporal Modelling

Thus far in this research we have focused on static prediction problems. In future work, it could be interesting to explore the temporal evolution of the tumour tissue and how this could affect cancer patient treatment decisions, since we do not today have good models of temporal progression in colorectal cancer histology.

In Chapter 5 we introduced some analysis of the gradients of our predicted variables across epithelial tissue boundaries. These gradient flows currently just demonstrate the differences of predicted CMS4 and response to RT values across tissue boundaries. However, we want to explore this idea further, inspiring research that can imagine a movement or a flow, introducing a temporal aspect to potentially provide insight on how the tissue may evolve over time. Furthermore, we could introduce another dimension by considering these gradients on 3D histology WSIs, dependent on data availability.

Another angle from which we could introduce a temporal aspect to future work is by using generative models. With models such as VQ-VAE or diffusion models we could demonstrate how digital histology images of cancer tissue could evolve over time. The CMS molecular subtypes of CRC could be used as a representative latent distribution of the images, used to generate the next time step of the tissue distribution. Furthermore, we could incorporate prior knowledge of the cell evolution into our framework, combining expertise from both pathologists and biological mathematicians, to constrain how the images can change to ensure the tissue is naturally evolving in our output. Using the existing datasets, we could measure the effectiveness of generated images using automated counts of tumour cells and the ground truth of how the patient responded to RT treatment, indicating whether any tumour cells were remaining post-treatment.

7.4.3 Domain Adaptation

As we demonstrated in our ablation studies on our domain adaptation method in Section 6.4.4, finding an optimal clustering of the source data is the key to getting the best results from this method. It may be possible to extend this work to test how this approach can generalise onto multiple target cohorts. To imitate real life application, a cumulative approach should be considered to recalculate the cluster centres over each new target domain, to measure how this affects the model adaptability.

The power of this approach depends to some degree on how much the disease space is covered by the disease variation in the source data. If we are confident our source model has seen a particular disease variation before, we could be far more aggressive in shifting features, and similarly less aggressive for outliers, introducing some sort of weighted outlier detection approach.

7.4.4 Patient Treatments

Finally, we need to make sure our research is relevant to the current state of clinical treatment pathways. Since the clinical studies which provided our datasets

were completed, treatment for CRC has evolved somewhat, meaning the pathway considered here consisting of neoadjuvant CapRT is no longer the most commonly used. A Professor of Colorectal Surgery at the University of Oxford informed us that long-course CAPOX (capecitabine in combination with oxaliplatin) is now a more popular form of total neoadjuvant therapy (TNT) than just capecitabine with short-course RT in the UK [200]. On a specific subset of patients with locally advanced rectal cancer, a study comparing the effect of neoadjuvant chemotherapy with CAPOX (without RT) against chemoradiotherapy with capecitabine concluded that CAPOX could be an effective alternative treatment [201]. Hence, in future work we could consider how we can adapt our deep learning methods to predict outcomes for different forms of therapy.

References

[1] Hildebrand LA et al. "Artificial Intelligence for Histology-Based Detection of Microsatellite Instability and Prediction of Response to Immunotherapy in Colorectal Cancer". In: *Cancers (Basel)* 13.3 (2021), p. 391. URL: doi:10.3390/cancers13030391.

[2] Korsuk Sirinukunwattana et al. "Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning". In: *Gut* 70.3 (2021). Ed. by, pp. 544–554. eprint: https://gut.bmj.com/content/70/3/544.full.pdf. URL: https://gut.bmj.com/content/70/3/544.

[3] *CR UK: Bowel Cancer.* https://www.cancerresearchuk.org/about-cancer/bowel-cancer. Accessed: 19-08-2024.

[4] *CR UK: Bowel Cancer Statistics.* https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer. Accessed: 19-08-2024.

[5] Alastair J Morton et al. "Long-term adverse effects and healthcare burden of rectal cancer radiotherapy: systematic review and meta-analysis". In: *ANZ journal of surgery* 93.1-2 (2023), pp. 42–53.

[6] *Leeds University: Rectal Cancer Treatment.* https://www.leeds.ac.uk/news-health/news/article/4736/new-treatment-could-spare-early-stage-

rectal-cancer-patients-life-altering-side-effects. Accessed: 19-08-2024.

[7] *CR UK: Bowel Cancer Treatment*. https://www.cancerresearchuk.org/about-cancer/bowel-cancer/treatment. Accessed: 19-08-2024.

[8] *NHS: Bowel Cancer Treatment*. https://www.nhs.uk/conditions/bowel-cancer/treatment/. Accessed: 26-05-2022.

[9] A. Alkan et al. "Biomarkers and Cell-based Models to Predict the Outcome of Neoadjuvant Therapy for Rectal Cancer Patients". In: *Biomarker Research* 9.1 (2021).

[10] W.K. Chatila, J.K. Kim, and H. Walch. "Genomic and Transcriptomic Determinants of Response to Neoadjuvant Therapy in Rectal Cancer". In: *Nature Medicine* 28 (2022), pp. 1646–1655.

[11] T J Jr George, C J Allegra, and G Yothers. "Neoadjuvant Rectal (NAR) Score: a New Surrogate Endpoint in Rectal Cancer Clinical Trials". In: *Current Colorectal Cancer Reports* 11.5 (2015), pp. 275–280. URL: doi:10.1007/s11888-015-0285-2.

[12] J Joshua Smith et al. "Assessment of a watch-and-wait strategy for rectal cancer in patients with a complete response after neoadjuvant therapy". In: *JAMA oncology* 5.4 (2019), e185896–e185896.

[13] M. Lai and B. Lü. "3.04 - Tissue Preparation for Microscopy and Histology". In: *Comprehensive Sampling and Sample Preparation* 3 (2012), pp. 53–93. URL: doi.org/10.1016/B978-0-12-381373-2.00070-3.

[14] Rathore S et al. "Segmentation and Grade Prediction of Colon Cancer Digital Pathology Images Across Multiple Institutions". In: *Cancers (Basel)* 11.11 (2019), p. 1700. URL: doi:10.3390/cancers11111700.

[15] Enric Domingo et al. "Identification and validation of a machine learning model of complete response to radiation in rectal cancer reveals immune infiltrate and TGFβ as key predictors". In: *EBioMedicine* 106 (2024).

[16] *DICOM Whole Slide Imaging (WSI).* `https://dicom.nema.org/dicom/dicomwsi/`. Accessed: 26-05-2022.

[17] Xueke Shi et al. "TGF-β signaling in the tumor metabolic microenvironment and targeted therapies". In: *Journal of Hematology & Oncology* 15.1 (2022), p. 135.

[18] Robert A Weinberg and Robert A Weinberg. *The biology of cancer*. WW Norton & Company, 2006.

[19] Yoshiro Itatani, Kenji Kawada, and Yoshiharu Sakai. "Transforming growth factor-β signaling pathway in colorectal cancer and its tumor microenvironment". In: *International journal of molecular sciences* 20.23 (2019), p. 5822.

[20] Hao Xu, Qianhui Xu, and Lu Yin. "Prognostic value of tumor immune cell infiltration patterns in colon adenocarcinoma based on systematic bioinformatics analysis". In: *Cancer cell international* 21 (2021), pp. 1–13.

[21] Alexandra M Zaborowski, Des C Winter, and Lydia Lynch. "The therapeutic and prognostic implications of immunobiology in colorectal cancer: a review". In: *British Journal of Cancer* 125.10 (2021), pp. 1341–1349.

[22] Jafar Nouri et al. Nojadeh. "Microsatellite instability in colorectal cancer". In: *EXCLI Journal* 17 (2018), pp. 159–168.

[23] Ben Tran et al. "Impact of BRAF mutation and microsatellite instability on the pattern of metastatic spread and prognosis in metastatic colorectal cancer". In: *Cancer* 117.20 (2011), pp. 4623–4632.

[24] Magdalena C Liebl and Thomas G Hofmann. "The role of p53 signaling in colorectal cancer". In: *Cancers* 13.9 (2021), p. 2125.

[25] Kexin Ding et al. "Spatially aware graph neural networks and cross-level molecular profile prediction in colon cancer histopathology: a retrospective multi-cohort study". In: *The Lancet Digital Health* 4.11 (2022), e787–e795.

[26] J. Guinney et al. “The Consensus Molecular Subtypes of Colorectal Cancer”. In: *Nature Medicine* 21 (2015), pp. 1350–1356.

[27] J. Guinney et al. “The consensus molecular subtypes of colorectal cancer”. In: *Nature Medicine* 21 (2015), pp. 1350–1356. URL: https://doi.org/10.1038/nm.3967.

[28] Lin Qi et al. “Identification of Prognostic Spatial Organization Features in Colorectal Cancer Microenvironment using Deep Learning on Histopathology images”. In: *Medicine in Omics* 2 (2021), p. 100008.

[29] Annie M. Young (Editor), Richard Hobbs (Editor), and David J. Kerr (Editor). *ABC of Colorectal Cancer, 2nd Edition*. BMJ Books, 2011.

[30] P. Snaebjornsson et al. “pT4 stage II and III colon cancers carry the worst prognosis in a nationwide survival analysis. Shepherd’s local peritoneal involvement revisited”. In: *International Journal of Cancer* 135.2 (2014), pp. 467–478. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijc.28676. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.28676.

[31] Swamy R. “Histopathological reporting of pT4 tumour stage in colorectal carcinomas: dotting the ’i’s and crossing the ’t’s”. In: *Journal of Clinical Pathology* 63.2 (2010), pp. 110–115. URL: doi:10.1136/jcp.2009.069658.

[32] Maria-Gabriela Anitei et al. “Prognostic and Predictive Values of the Immunoscore in Patients with Rectal Cancer”. In: *Clinical Cancer Research* 20.7 (Apr. 2014), pp. 1891–1899. URL: https://doi.org/10.1158/1078-0432.CCR-13-2830.

[33] H J S Jones et al. “Stromal composition predicts recurrence of early rectal cancer after local excision”. In: *Histopathology* 79 (2021), pp. 947–956. URL: https://doi.org/10.1111/his.14438.

[34] Jana Lipkova et al. “Artificial Intelligence for Multimodal Data Integration in Oncology”. In: *Cancer Cell* 40.10 (2022), pp. 1095–1110.

[35] Jun Xu et al. “A Deep Convolutional Neural Network for Segmenting and Classifying Epithelial and Stromal Regions in Histopathological Images”. In: *Neurocomputing* 191 (2016), pp. 214–223.

[36] V.H. Koelzer et al. “CD8/CD45RO T-cell infiltration in endoscopic biopsies of colorectal cancer predicts nodal metastasis and survival”. In: *Journal of Translational Medicine* 12.81 (2014). URL: https://doi.org/10.1186/1479-5876-12-81.

[37] Gloria Alfonsín et al. “Stratification of Colorectal Patients Based on Survival Analysis Shows the Value of Consensus Molecular Subtypes and Reveals the CBLL1 Gene as a Biomarker of CMS2 Tumours”. In: *International Journal of Molecular Sciences* 25.3 (2024), p. 1919.

[38] Paul W Sweeney et al. “Modelling the transport of fluid through heterogeneous, whole tumours in silico”. In: *PLoS computational biology* 15.6 (2019), e1006751.

[39] Narmin Ghaffari Laleh et al. “Classical mathematical models for prediction of response to chemotherapy and immunotherapy”. In: *PLoS computational biology* 18.2 (2022), e1009822.

[40] Barbara D Wichtmann et al. “Are we there yet? The value of deep learning in a multicenter setting for response prediction of locally advanced rectal cancer to neoadjuvant chemoradiotherapy”. In: *Diagnostics* 12.7 (2022), p. 1601.

[41] Cheng Jin et al. “Predicting treatment response from longitudinal images using multi-task deep learning”. In: *Nature communications* 12.1 (2021), p. 1851.

[42] A. Rogers et al. “Prognostic significance of tumor budding in rectal cancer biopsies before neoadjuvant therapy”. In: *Modern Pathology* 27 (2014), pp. 156–162. URL: https://doi.org/10.1038/modpathol.2013.124.

[43] F Zhang et al. “Predicting Treatment Response to Neoadjuvant Chemoradiotherapy in Local Advanced Rectal Cancer by Biopsy Digital Pathology Image Features”. In: *Clinical and Translational Medicine* 10.2 (2020), e110.

[44] Namjoo Kim et al. "Detection of Microsatellite Instability in Colorectal Cancer Patients With a Plasma-Based Real-Time PCR Analysis". In: *Frontiers in Pharmacology* 12 (2021). URL: https://www.frontiersin.org/article/10.3389/fphar.2021.758830.

[45] T André et al. "Pembrolizumab in Microsatellite-Instability-High Advanced Colorectal Cancer". In: *The New England Journal of Medicine* 383.23 (2020), pp. 2207–2218. URL: doi:10.1056/NEJMoa2017699.

[46] Bilal M et al. "Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study". In: *Lancet Digit Health* 3.12 (2021), e763–e772.

[47] Echle A et al. "Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning". In: *Gastroenterology* 159.4 (2020), pp. 1406–1416.

[48] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[49] Bangwei Guo et al. "Predicting microsatellite instability and key biomarkers in colorectal cancer from H&E-stained images: achieving state-of-the-art predictive performance with fewer data using swin transformer". In: *The Journal of Pathology: Clinical Research* 9.3 (2023), pp. 223–235.

[50] Christian Szegedy et al. "Rethinking the Inception Architecture for Computer Vision". In: *CoRR* abs/1512.00567 (2015). arXiv: 1512.00567. URL: http://arxiv.org/abs/1512.00567.

[51] Maxime W Lafarge et al. "Image-based consensus molecular subtyping in rectal cancer biopsies and response to neoadjuvant chemoradiotherapy". In: *NPJ precision oncology* 8.1 (2024), p. 89.

[52] Pei-Chen Tsai et al. “Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients”. In: *Nature communications* 14.1 (2023), p. 2102.

[53] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *CoRR* (2020). arXiv: 2010.11929. URL: https://arxiv.org/abs/2010.11929.

[54] Joel Saltz et al. “Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images”. In: *Cell reports* 23.1 (2018), pp. 181–193.

[55] Lennard Kiehl et al. “Deep learning can predict lymph node status directly from histology in colorectal cancer”. In: *European Journal of Cancer* 157 (2021), pp. 464–473.

[56] Elena Martínez-Fernandez et al. “Computer aided classifier of colorectal cancer on histopatological whole slide images analyzing deep learning architecture parameters”. In: *Applied Sciences* 13.7 (2023), p. 4594.

[57] Juha P Väyrynen et al. “Prognostic significance of immune cell populations identified by machine learning in colorectal cancer using routine hematoxylin and eosin–stained sections”. In: *Clinical Cancer Research* 26.16 (2020), pp. 4326–4338.

[58] Skrede OJ et al. “Deep learning for prediction of colorectal cancer outcome: a discovery and validation study”. In: *Lancet* 395.10221 (2020), pp. 350–360. URL: doi:10.1016/S0140-6736(19)32998-8.

[59] G. Campanella et al. “Clinical-grade computational pathology using weakly supervised deep learning on whole slide images”. In: *Nature Medicine* 25 (2019), pp. 1301–1309. URL: https://doi.org/10.1038/s41591-019-0508-1.

[60] O. Iizuka et al. “Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours”. In: *Scientific Reports* 10 (2020), p. 1504. URL: https://doi.org/10.1038/s41598-020-58467-9.

[61] F. Li et al. “Deep learning-based predictive biomarker of pathological complete response to neoadjuvant chemotherapy from histological images in breast cancer”. In: *Journal of Translational Medicine* 19.348 (2021). URL: https://doi.org/10.1186/s12967-021-03020-z.

[62] F. Kanavati et al. “A deep learning model for the classification of indeterminate lung carcinoma in biopsy whole slide images”. In: *Scientific Reports* 11 (2021), p. 8110. URL: https://doi.org/10.1038/s41598-021-87644-7.

[63] Hang Li et al. “DT-MIL: Deformable Transformer for Multi-instance Learning on Histopathological Image”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Springer International Publishing, 2021, pp. 206–216.

[64] D. Bychkov et al. “Deep learning based tissue analysis predicts outcome in colorectal cancer”. In: *Scientific Reports* 8 (2018), p. 3395. URL: https://doi.org/10.1038/s41598-018-21758-3.

[65] Runyu Hong et al. “Predicting endometrial cancer subtypes and molecular features from histopathology images using multi-resolution deep learning models”. In: *Cell Reports Medicine* 2.9 (2021).

[66] M. Y. Lu et al. “Data-efficient and weakly supervised computational pathology on whole-slide images”. In: *Nature Biomedical Engineering* 5.6 (2021), pp. 555–570. URL: doi:10.1038/s41551-020-00682-w.

[67] Maximilian Ilse, Jakub Tomczak, and Max Welling. “Attention-based Deep Multiple Instance Learning”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2127–2136. URL: https://proceedings.mlr.press/v80/ilse18a.html.

[68] Yash Sharma et al. *Cluster-to-Conquer: A Framework for End-to-End Multi-Instance Learning for Whole Slide Image Classification*. 2021. URL: doi.org/10.48550/arxiv.2103.10626.

[69] Jangho Kwon and Kihwan Choi. "Weakly supervised attention map training for histological localization of colonoscopy images". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 3725–3728.

[70] Andrew Broad et al. "Attention-guided sampling for colorectal cancer analysis with digital pathology". In: *Journal of Pathology Informatics* 13 (2022), p. 100110.

[71] Omar SM El Nahhas et al. "Regression-based Deep-Learning predicts molecular biomarkers from pathology slides". In: *nature communications* 15.1 (2024), p. 1253.

[72] Olga Fourkioti, Matt De Vries, and Chris Bakal. "CAMIL: Context-Aware Multiple Instance Learning for Cancer Detection and Subtyping in Whole Slide Images". In: *arXiv preprint arXiv:2305.05314* (2023).

[73] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[74] Richard J Chen et al. "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16144–16155.

[75] Manuel Tran et al. "B-Cos Aligned Transformers Learn Human-Interpretable Features". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 514–524.

[76] Ziwang Huang et al. "Integration of Patch Features Through Self-supervised Learning and Transformer for Survival Analysis on Whole Slide Images". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Ed. by Marleen de Bruijne et al. Cham: Springer International Publishing, 2021, pp. 561–570.

[77] Yi Zheng et al. "A graph-transformer for whole slide image classification". In: *IEEE Transactions on Medical Imaging* (2022), pp. 1–1.

[78] Zhuchen Shao et al. “TransMIL: Transformer based Correlated Multiple. Instance Learning for Whole Slide Image Classication”. In: *CoRR* abs/2106.00908 (2021). arXiv: 2106.00908. URL: https://arxiv.org/abs/2106.00908.

[79] Xiangxiang Chu et al. “Do We Really Need Explicit Position Encodings for Vision Transformers?” In: *CoRR* abs/2102.10882 (2021). arXiv: 2102.10882. URL: https://arxiv.org/abs/2102.10882.

[80] Md. Amirul Islam, Sen Jia, and Neil D. B. Bruce. “How Much Position Information Do Convolutional Neural Networks Encode?” In: *CoRR* abs/2001.08248 (2020). arXiv: 2001.08248. URL: https://arxiv.org/abs/2001.08248.

[81] Michael M. Bronstein et al. “Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges”. In: *arXiv preprint arXiv:2104.13478* (2021).

[82] Pierre Besson et al. “Geometric deep learning on brain shape predicts sex and age”. In: *Computerized Medical Imaging and Graphics* 91 (2021), p. 101939.

[83] Devanshu Arya and Marcel Worring. “Exploiting relational information in social networks using geometric deep learning on hypergraphs”. In: *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. 2018, pp. 117–125.

[84] Kenneth Atz, Francesca Grisoni, and Gisbert Schneider. “Geometric deep learning on molecular representations”. In: *Nature Machine Intelligence* 3.12 (2021), pp. 1023–1032.

[85] Boris Weisfeiler and Andrei Leman. “The reduction of a graph to canonical form and the algebra which appears therein”. In: *nti, Series* 2.9 (1968), pp. 12–16.

[86] Petar Velickovic et al. “Graph attention networks”. In: *stat* 1050.20 (2017), pp. 10–48550.

[87] Keyulu Xu et al. “How Powerful are Graph Neural Networks?” In: *CoRR* abs/1810.00826 (2018). arXiv: 1810.00826. URL: http://arxiv.org/abs/1810.00826.

[88] Yue Wang et al. “Dynamic graph cnn for learning on point clouds”. In: *ACM Transactions on Graphics (tog)* 38.5 (2019), pp. 1–12.

[89] Harshita Sharma et al. “A review of graph-based methods for image analysis in digital histopathology”. In: *Diagnostic pathology* 1.1 (2015).

[90] Mayar Allam et al. “Spatially variant immune infiltration scoring in human cancer tissues”. In: *NPJ precision oncology* 6.1 (2022), p. 60.

[91] Korsuk Sirinukunwattana et al. “Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer”. In: *Scientific reports* 8.1 (2018), p. 13692.

[92] Ruoyu Li et al. “Graph CNN for survival analysis on whole slide pathological images”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 174–182.

[93] Y. Lee, J.H. Park, and S. Oh. “Derivation of Prognostic Contextual Histopathological Features from Whole-Slide Images of Tumours via Graph Deep Learning”. In: *Nature Biomedical Engineering* (2022).

[94] Milan Aryal and Nasim Yahya Soltani. “Position-Aware Graph-Based Learning of Whole Slide Images”. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[95] Sara Arabyarmohammadi et al. “Triangular Analysis of Geographical Interplay of Lymphocytes (TriAnGIL): Predicting Immunotherapy Response in Lung Cancer”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 797–807.

[96] Oscar Pina and Verónica Vilaplana. “Self-supervised Graph Representations of WSIs”. In: *Proceedings of the First International Workshop on Geometric Deep Learning in Medical Image Analysis*. Ed. by Erik Bekkers, Jelmer M. Wolterink, and Angelica Aviles-Rivero. Vol. 194. Proceedings of Machine Learning Research. PMLR, Nov. 2022, pp. 107–117.

[97] Wenqi Lu et al. “SlideGraph+: Whole Slide Image Level Graphs to Predict HER2 Status in Breast Cancer”. In: *Medical Image Analysis* 80 (2022), p. 102486.

[98] Wenqi Lu et al. “Capturing Cellular Topology in Multi-Gigapixel Pathology Images”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2020, pp. 1049–1058.

[99] Yawen Wu et al. “Transfer learning-assisted survival analysis of breast cancer relying on the spatial interaction between tumor-infiltrating lymphocytes and tumors”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 612–621.

[100] Ashwin Raju et al. “Graph Attention Multi-instance Learning for Accurate Colorectal Cancer Staging”. In: *MICCAI (5)*. 2020, pp. 529–539. URL: https://doi.org/10.1007/978-3-030-59722-1_51.

[101] Yanning Zhou et al. “CGC-Net: Cell Graph Convolutional Network for Grading of Colorectal Cancer Histology Images”. In: *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*. IEEE. IEEE, 2019, 388–398. URL: https://doi.org/10.1109/ICCVW.2019.00050.

[102] Kexin Ding et al. “Feature-Enhanced Graph Networks for Genetic Mutational Prediction Using Histopathological Images in Colon Cancer”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Ed. by Anne L. Martel et al. Cham: Springer International Publishing, 2020, pp. 294–304.

[103] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. “Self-attention with relative position representations”. In: *arXiv preprint arXiv:1803.02155* (2018).

[104] Esther Danenberg et al. “Breast tumor microenvironment structures are associated with genomic features and clinical outcome”. In: *Nature genetics* 54.5 (2022), pp. 660–669.

[105] Yuzhou Feng et al. “Spatial analysis with SPIAT and spaSim to characterize and simulate tissue microenvironments”. In: *Nature Communications* 14.1 (2023), p. 2697.

[106] Tong Fu et al. “Spatial architecture of the immune microenvironment orchestrates tumor immunity and therapeutic response”. In: *Journal of hematology & oncology* 14.1 (2021), p. 98.

[107] Jovan Tanevski et al. “Learning tissue representation by identification of persistent local patterns in spatial omics data”. In: *bioRxiv* (2024), pp. 2024–03.

[108] Ren Yuan Lee et al. “The promise and challenge of spatial omics in dissecting tumour microenvironment and the role of AI”. In: *Frontiers in Oncology* 13 (2023), p. 1172314.

[109] Denis Schapiro et al. “histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data”. In: *Nature methods* 14.9 (2017), pp. 873–876.

[110] Korsuk Sirinukunwattana et al. “Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images”. In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1196–1206.

[111] Olivier De Wever and Marc Mareel. “Role of tissue stroma in cancer cell invasion”. In: *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 200.4 (2003), pp. 429–447.

[112] Yiping Jiao et al. “Deep learning-based tumor microenvironment analysis in colon adenocarcinoma histopathological whole-slide images”. In: *Computer Methods and Programs in Biomedicine* 204 (2021), p. 106047.

[113] Hangbo Bao et al. “Beit: Bert pre-training of image transformers”. In: *arXiv preprint arXiv:2106.08254* (2021).

[114] Mathilde Caron et al. “Emerging Properties in Self-Supervised Vision Transformers”. In: *Proceedings of the International Conference on Computer Vision (ICCV)*. 2021.

[115] Ting Chen et al. “A simple framework for contrastive learning of visual representations”. In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[116] Xiyue Wang et al. “Transformer-based unsupervised contrastive learning for histopathological image classification”. In: *Medical Image Analysis* 81 (2022), p. 102559. URL: `https://www.sciencedirect.com/science/article/pii/S1361841522002043`.

[117] Richard J Chen and Rahul G Krishnan. “Self-supervised vision transformers learn visual concepts in histopathology”. In: *arXiv preprint arXiv:2203.00585* (2022).

[118] Richard J Chen et al. “Towards a General-Purpose Foundation Model for Computational Pathology”. In: *Nature Medicine* (2024).

[119] Maxime Oquab et al. “Dinov2: Learning robust visual features without supervision”. In: *arXiv preprint arXiv:2304.07193* (2023).

[120] Zachary C Lipton. “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.” In: *Queue* 16.3 (2018), pp. 31–57.

[121] Arun Das and Paul Rad. “Opportunities and challenges in explainable artificial intelligence (xai): A survey”. In: *arXiv preprint arXiv:2006.11371* (2020).

[122] GLEASON DF. “Histologic grading and clinical staging of prostatic carcinoma”. In: *Urologic Pathology* (1977).

[123] Seon-Kyu Kim et al. “A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients”. In: *Molecular oncology* 8.8 (2014), pp. 1653–1666.

[124] Steven A Buechler et al. “ColoType: a forty gene signature for consensus molecular subtyping of colorectal cancer tumors using whole-genome assay or targeted RNA-sequencing”. In: *Scientific reports* 10.1 (2020), p. 12123.

[125] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[126] Ramprasaath R Selvaraju et al. "Grad-CAM: Why did you say that?" In: *arXiv preprint arXiv:1611.07450* (2016).

[127] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. "The false hope of current approaches to explainable artificial intelligence in health care". In: *The Lancet Digital Health* 3.11 (2021), e745–e750.

[128] Harry B Burke. "Predicting clinical outcomes using molecular biomarkers". In: *Biomarkers in cancer* 8 (2016), BIC–S33380.

[129] Lawrence H Schwartz et al. "RECIST 1.1—Update and clarification: From the RECIST committee". In: *European journal of cancer* 62 (2016), pp. 132–137.

[130] James PB O'connor et al. "Imaging biomarker roadmap for cancer studies". In: *Nature reviews Clinical oncology* 14.3 (2017), pp. 169–186.

[131] *NICE: Immunoscore for predicting risk of colon cancer relapse.* `https://www.nice.org.uk/advice/mib269`. Accessed: 31-08-2024.

[132] Enric Domingo et al. "Prognostic and Predictive Value of Immunoscore in Stage III Colorectal Cancer: Pooled Analysis of Cases From the SCOT and IDEA-HORG Studies". In: *Journal of Clinical Oncology* (2024), JCO–23.

[133] *MSIntuit® CRC: Optimize MSI testing for colorectal cancer.* `https://www.owkin.com/msintuit-crc`. Accessed: 04-09-2024.

[134] *NICE: Molecular testing strategies for Lynch syndrome in people with colorectal cancer.* `https://www.nice.org.uk/guidance/dg27/chapter/1-Recommendations`. Accessed: 04-09-2024.

[135] Charlie Saillard et al. "Validation of MSIntuit as an AI-based pre-screening tool for MSI detection from colorectal cancer histology slides". In: *Nature Communications* 14.1 (2023), p. 6695.

[136] Daniel E Spratt et al. "Artificial intelligence predictive model for hormone therapy use in prostate cancer". In: *NEJM evidence* 2.8 (2023), EVIDoa2300023.

[137] *ArteraAI: The science behind the ArteraAI Prostate Test.* `https://artera.ai/arteraai-prostate-cancer-test`. Accessed: 31-08-2024.

[138] *Paige.AI: Paige Receives First Ever FDA Approval for AI Product in Digital Pathology.* `https://paige.ai/paige-receives-first-ever-fda-approval-for-ai-product-in-digital-pathology/`. Published: 22-09-2021; Accessed: 31-08-2024.

[139] Catarina Eloy et al. "Artificial intelligence–assisted cancer diagnosis improves the efficiency of pathologists in prostatic biopsies". In: *Virchows Archiv* 482.3 (2023), pp. 595–604.

[140] *NICE: Paige Prostate for prostate cancer.* `https://www.nice.org.uk/advice/mib280`. Accessed: 31-08-2024.

[141] Goode A et al. "OpenSlide: A vendor-neutral software foundation for digital pathology". In: *J Pathol Inform* 4.27 (2014).

[142] Ruby Wood et al. "Enhancing Local Context of Histology Features in Vision Transformers". In: *Artificial Intelligence over Infrared Images for Medical Applications and Medical Image Assisted Biomarker Discovery*. Cham: Springer Nature Switzerland, 2022, pp. 154–163.

[143] Zeyu Gao et al. "Instance-based Vision Transformer for Subtyping of Papillary Renal Cell Carcinoma in Histopathological Image". In: *CoRR* abs/2106.12265 (2021). arXiv: `2106.12265`. URL: `https://arxiv.org/abs/2106.12265`.

[144] B. Schmauch et al. "A deep learning model to predict RNA-Seq expression of tumours from whole slide images". In: *Nature Communications* 11 (2020), p. 3877. URL: `https://doi.org/10.1038/s41467-020-17678-4`.

[145] Xiankai Lu et al. “Deep regression tracking with shrinkage loss”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 353–369.

[146] Natalie C Fisher et al. “Biological misinterpretation of transcriptional signatures in tumor samples can unknowingly undermine mechanistic understanding and faithful alignment with preclinical data”. In: *Clinical Cancer Research* 28.18 (2022), pp. 4056–4069.

[147] Charles X. Ling, Jin Huang, and Harry Zhang. “AUC: A Better Measure than Accuracy in Comparing Learning Algorithms”. In: *Advances in Artificial Intelligence*. Ed. by Yang Xiang and Brahim Chaib-draa. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 329–341.

[148] Ruby Wood et al. “Joint Prediction of Response to Therapy, Molecular Traits, and Spatial Organisation in Colorectal Cancer Biopsies”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan et al. Cham: Springer Nature Switzerland, 2023, pp. 758–767.

[149] R. Achanta et al. “SLIC Superpixels Compared to State-of-the-Art Superpixel Methods”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (2012), pp. 2274–2282.

[150] Richard J. Chen and Rahul G. Krishnan. *Self-Supervised Vision Transformers Learn Visual Concepts in Histopathology*. 2022.

[151] Hanwen Xu et al. “A whole-slide foundation model for digital pathology from real-world data”. In: *Nature* (2024), pp. 1–8.

[152] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[153] Radhakrishna Achanta et al. “SLIC superpixels compared to state-of-the-art superpixel methods”. In: *IEEE transactions on pattern analysis and machine intelligence* 34.11 (2012), pp. 2274–2282.

[154] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015.

[155] Korsuk Sirinukunwattana and et al. "Image-based Consensus Molecular Subtype (imCMS) Classification of Colorectal Cancer using Deep Learning". In: *Gut* 70.3 (2021). Ed. by, pp. 544–554.

[156] Stéfan van der Walt et al. "scikit-image: image processing in Python". In: *PeerJ* 2 (June 2014), e453. URL: https://doi.org/10.7717/peerj.453.

[157] Ruby Wood et al. "Cluster Triplet Loss for Unsupervised Domain Adaptation on Histology Images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 5122–5131.

[158] Mostafa Jahanifar et al. *Domain Generalization in Computational Pathology: Survey and Guidelines*. 2023. arXiv: 2310.19656 [eess.IV].

[159] John M. Carethers and Chyke A. Doubeni. "Causes of Socioeconomic Disparities in Colorectal Cancer and Intervention Framework and Strategies". In: *Gastroenterology* 158.2 (2020). Colorectal Cancer: Recent Advances & Future Challenges, pp. 354–367. URL: https://www.sciencedirect.com/science/article/pii/S0016508519414820.

[160] Karin Stacke et al. "Measuring Domain Shift for Deep Learning in Histopathology". In: *IEEE Journal of Biomedical and Health Informatics* PP (Oct. 2020), pp. 1–1.

[161] Jian Liang, Dapeng Hu, and Jiashi Feng. "Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 6028–6039. URL: https://proceedings.mlr.press/v119/liang20a.html.

[162] Hui Tang, Ke Chen, and Kui Jia. "Unsupervised Domain Adaptation via Structurally Regularized Deep Clustering". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 8722–8732.

[163] Shiqi Yang et al. "Attracting and Dispersing: A Simple Approach for Source-free Domain Adaptation". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 5802–5815. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/26300457961c3e056ea61c9d3ebec2a4-Paper-Conference.pdf.

[164] Xiaoshun Wang, Yunhan Li, and Xiangliang Zhang. "Improved triplet loss for domain adaptation". In: *IET Computer Vision* 18.1 (2024), pp. 84–96. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/cvi2.12226. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cvi2.12226.

[165] Pan Zhang et al. "Prototypical Pseudo Label Denoising and Target Structure Learning for Domain Adaptive Semantic Segmentation". In: *CoRR* abs/2101.10979 (2021). arXiv: 2101.10979. URL: https://arxiv.org/abs/2101.10979.

[166] Meng Zhou, Zhe Xu, and Raymond Kai-yu Tong. "Superpixel-guided class-level denoising for unsupervised domain adaptive fundus image segmentation without source data". In: *Computers in Biology and Medicine* 162 (2023), p. 107061. URL: https://www.sciencedirect.com/science/article/pii/S0010482523005267.

[167] Ziyi Zhang et al. "Divide and Contrast: Source-free Domain Adaptation via Adaptive Contrastive Learning". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 5137–5149. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/215aeb07b5996c969c0123c3c6ee8f54-Paper-Conference.pdf.

[168] Yexun Zhang et al. “Domain-Invariant Adversarial Learning for Unsupervised Domain Adaption”. In: *CoRR* abs/1811.12751 (2018). arXiv: 1811.12751. URL: http://arxiv.org/abs/1811.12751.

[169] Xinwei He et al. “Triplet-Center Loss for Multi-view 3D Object Retrieval”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1945–1954.

[170] Miguel Lagunes-Fortiz, Dima Damen, and Walterio Mayol-Cuevas. “Centroids Triplet Network and Temporally-Consistent Embeddings for In-Situ Object Recognition”. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2020, pp. 10796–10802.

[171] Saeid Motiian et al. “Unified Deep Supervised Domain Adaptation and Generalization”. In: *CoRR* abs/1709.10190 (2017). arXiv: 1709.10190. URL: http://arxiv.org/abs/1709.10190.

[172] Daehee Kim et al. “SelfReg: Self-Supervised Contrastive Regularization for Domain Generalization”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 9619–9628.

[173] Issam H. Laradji and Reza Babanezhad. “M-ADDA: Unsupervised Domain Adaptation with Deep Metric Learning”. In: *CoRR* abs/1807.02552 (2018). arXiv: 1807.02552. URL: http://arxiv.org/abs/1807.02552.

[174] Yundong Li, Longxia Guo, and Yizheng Ge. “Pseudo Labels for Unsupervised Domain Adaptation: A Review”. In: *Electronics* 12.15 (2023). URL: https://www.mdpi.com/2079-9292/12/15/3325.

[175] Mikołaj Wieczorek, Barbara Rychalska, and Jacek Dąbrowski. “On the Unreasonable Effectiveness of Centroids in Image Retrieval”. In: *Neural Information Processing*. Ed. by Teddy Mantoro et al. Cham: Springer International Publishing, 2021, pp. 212–223.

[176] Alaa Alnissany and Yazan Dayoub. “Modified centroid triplet loss for person re-identification”. In: *Journal of Big Data* 10.74 (2023).

[177] Xiaodong Wang and Feng Liu. "Triplet Loss Guided Adversarial Domain Adaptation for Bearing Fault Diagnosis". In: *Sensors* 20.1 (2020). URL: https://www.mdpi.com/1424-8220/20/1/320.

[178] Maxime W. Lafarge et al. "Learning Domain-Invariant Representations of Histological Images". In: *Frontiers in Medicine* 6 (2019). URL: https://www.frontiersin.org/articles/10.3389/fmed.2019.00162.

[179] William Dee, Rana Alaaeldin Ibrahim, and Eirini Marouli. "Histopathological Domain Adaptation with Generative Adversarial Networks Bridging the Domain Gap Between Thyroid Cancer Histopathology Datasets". In: *bioRxiv* (2023). eprint: https://www.biorxiv.org/content/early/2023/05/24/2023.05.22.541691.full.pdf. URL: https://www.biorxiv.org/content/early/2023/05/24/2023.05.22.541691.

[180] Huihui Zhou et al. "Unsupervised domain adaptation for histopathology image segmentation with incomplete labels". In: *Computers in Biology and Medicine* 171 (2024), p. 108226. URL: https://www.sciencedirect.com/science/article/pii/S001048252400310X.

[181] Geetank Raipuria, Anu Shrivastava, and Nitin Singhal. "Stain-AgLr: Stain Agnostic Learning for Computational Histopathology Using Domain Consistency and Stain Regeneration Loss". In: *Domain Adaptation and Representation Transfer: 4th MICCAI Workshop, DART 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*. Singapore, Singapore: Springer-Verlag, 2022, pp. 33–44. URL: https://doi.org/10.1007/978-3-031-16852-9_4.

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

[183] Abhishek Vahadane et al. “Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images”. In: *IEEE Transactions on Medical Imaging* 35.8 (2016), pp. 1962–1971.

[184] Kuo-Sheng Cheng et al. “Domain-Centroid-Guided Progressive Teacher-based Knowledge Distillation for Source-Free Domain Adaptation of Histopathological Images”. In: *IEEE Transactions on Artificial Intelligence* (2023), pp. 1–14.

[185] Guillaume Vray et al. “Distill-SODA: Distilling Self-Supervised Vision Transformer for Source-Free Open-Set Domain Adaptation in Computational Pathology”. In: *IEEE Transactions on Medical Imaging* (2024), pp. 1–1.

[186] Jian Ren et al. “Unsupervised Domain Adaptation for Classification of Histopathology Whole-Slide Images”. In: *Frontiers in Bioengineering and Biotechnology* 7 (2019). URL: `https://www.frontiersin.org/articles/10.3389/fbioe.2019.00102`.

[187] Zhi Wang et al. “Cross-Domain Nuclei Detection in Histopathology Images Using Graph-Based Nuclei Feature Alignment”. In: *IEEE Journal of Biomedical and Health Informatics* 28.1 (2024), pp. 78–88.

[188] Christian Abbet et al. “Self-rule to multi-adapt: Generalized multi-source feature learning using unsupervised domain adaptation for colorectal cancer tissue detection”. In: *Medical Image Analysis* 79 (2022), p. 102473. URL: `https://www.sciencedirect.com/science/article/pii/S1361841522001207`.

[189] Yue Huang et al. “Epithelium-Stroma Classification via Convolutional Neural Networks and Unsupervised Domain Adaptation in Histopathological Images”. In: *IEEE Journal of Biomedical and Health Informatics* 21.6 (2017), pp. 1625–1632.

[190] Qi Qi et al. “Curriculum Feature Alignment Domain Adaptation for Epithelium-Stroma Classification in Histopathological Images”. In: *IEEE Journal of Biomedical and Health Informatics* 25.4 (2021), pp. 1163–1172.

[191] Xiangning Li et al. “Unsupervised Domain Adaptation for Cross-domain Histopathology Image Classification”. In: *Multimedia Tools and Applications* 83 (Aug. 2023), pp. 1–21.

[192] Kianoush Falahkheirkhah et al. “Domain adaptation using optimal transport for invariant learning using histopathology datasets”. In: *Medical Imaging with Deep Learning*. Ed. by Ipek Oguz et al. Vol. 227. Proceedings of Machine Learning Research. PMLR, July 2024, pp. 1765–1782. URL: https://proceedings.mlr.press/v227/falahkheirkhah24a.html.

[193] Milad Sikaroudi, Shahryar Rahnamayan, and H. R. Tizhoosh. *Hospital-Agnostic Image Representation Learning in Digital Pathology*. 2022. arXiv: 2204.02404 [eess.IV].

[194] Peter J. Rousseeuw. “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. URL: https://www.sciencedirect.com/science/article/pii/0377042787901257.

[195] T. Caliński and J Harabasz. “A dendrite method for cluster analysis”. In: *Communications in Statistics* 3.1 (1974), pp. 1–27. URL: https://www.tandfonline.com/doi/abs/10.1080/03610927408827101.

[196] David L. Davies and Donald W. Bouldin. “A Cluster Separation Measure”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2 (1979), pp. 224–227.

[197] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[198] Vassileios Balntas et al. “Learning local feature descriptors with triplets and shallow convolutional neural networks”. In: *British Machine Vision Conference*. 2016. URL: https://api.semanticscholar.org/CorpusID:27938870.

[199] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. arXiv: 1802.03426 [stat.ML].

[200] *Greater Manchester Colorectal Pathway Board. Guideline: Neoadjuvant and Non-operative Management of Rectal Cancer*. https://gmcancer.org.uk/wp-content/uploads/2024/05/Neoadjuvant-and-Non-operative-Management-of-Rectal-Cancer-Guidelines-v4.pdf. Accessed: 05-09-2024.

[201] Wei-Jian Mei et al. "Neoadjuvant chemotherapy with CAPOX versus chemoradiation for locally advanced rectal cancer with uninvolved mesorectal fascia (CONVERT): initial results of a phase III trial". In: *Annals of Surgery* 277,4 (2023), pp. 557–564.