Methods for Tumour Aggression Prediction in Colorectal Cancer through Virtual Immunohistochemistry, Supervised, and Self-Supervised Deep Learning

Christopher David Walsh

CRUK Scotland Institute & College of Medical, Veterinary and Life Sciences, University of Glasgow

Supervisor: Professor Robert Insall

Second Supervisor: Professor Joanne Edwards

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy, School of Cancer Sciences, College of Medical, Veterinary and Life Sciences, University of Glasgow.

Date: October 2023

Abstract

Tumour budding has emerged as a critical, independent predictor of survival and regression in colorectal cancer. It has elucidated the aggressive nature of certain tumours and the survival heterogeneity among tumours of a similar stage as determined by the TNM system. However, its clinical integration has been hampered by challenges like the absence of a standardised scoring methodology and interobserver variability when scoring in H&E. This work presents three methods to assist or automate tumour bud scoring and aggression prediction in colorectal cancer to enable consistent, reliable and cost-effective patient stratification.

The first chapter of this work explores the feasibility of translating H&E whole slide images to a virtual version of the AE1/AE3 immunohistochemical stain. This tool aims to assist pathologists in tumour bud scoring. We enhanced the loss function of the CycleGAN model to ensure accurate virtual staining while retaining histological structural detail. The modified CycleGAN model demonstrated improved structural similarity and realism in the generated images, providing a proof-of-concept model to create virtual pan-cytokeratin AE1/AE3 whole slide images that could improve the consistency and speed of manual tumour bud scoring.

Secondly, we introduced a method to train an end-to-end tumour bud segmentation system on manual annotations created using reference virtual IHC images. The virtual IHC model highlighted tumour cells or small clusters in the invasive margin of the source H&E slides, resulting in a high-quality dataset of almost 60,000 manually segmented buds. These were used to train a U-Net segmentation model that was evaluated on two distinct patient cohorts. The automated system significantly differentiated between high and low-budding populations, surpassing the manual score's hazard ratio in one cohort and proving an independent predictor of survival in univariate and multivariate Cox regression. The resulting segmentations also allowed an analysis of the tumour bud areas and distances from the tumour edge, hinting at the possibility of two distinct populations of buds in one cohort.

Finally, self-supervised learning was employed to train a feature extraction network using the VICReg architecture. The encoded representations were clustered, and the model's success was determined by running Cox regression on the observed probability distribution of clusters across the slides in the patient cohorts. These clusters were correlated with survival and demonstrated the encoding of relevant histological information. To predict tumour aggression, a custom transformer network was then used to analyse the feature vectors from the tumour tissue in the invasive margin. The model demonstrated a stronger correlation with manual budding and provided a more accurate stratification of tumour aggression, improving hazard ratios from the previous model and remaining an independent predictor of survival in univariate and multivariate Cox regression.

This research presents advancements in virtual immunohistochemistry, automated tumour bud scoring and automated tumour aggression prediction in colorectal cancer. We sought to provide a proof of concept for techniques that can pave the way for more consistent, objective, and efficient bud scoring and stratification by aggression. The methods presented, from virtual staining to self-supervised aggression prediction, highlight the synergy between technology and clinical pathology. As we move forward, the fusion of these domains promises to revolutionise diagnostic procedures and usher in an era of more personalised and effective patient care.

Contents

List of Tables

List of Figures

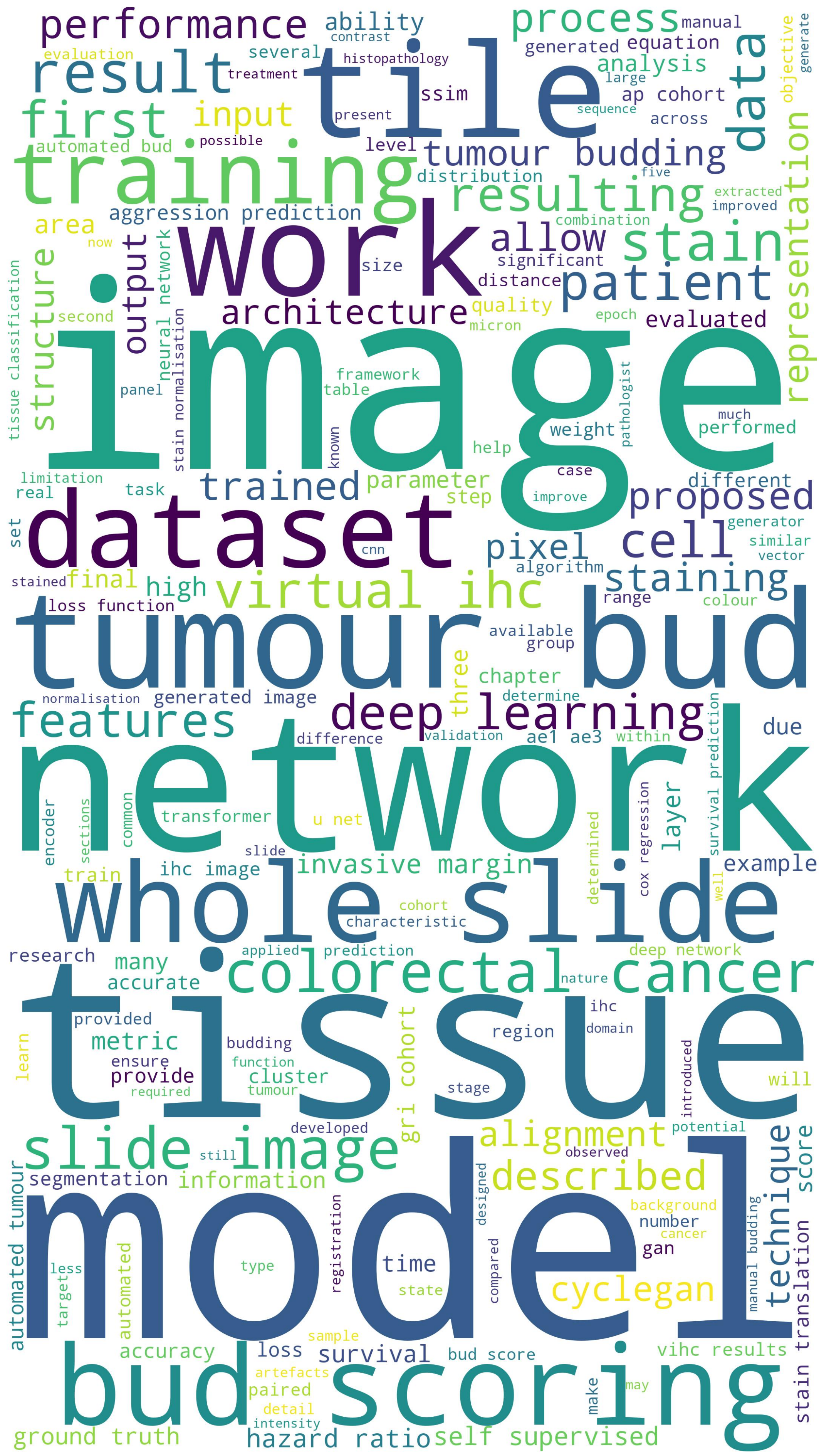
List of Algorithms

Definitions and Abbreviations

The following definitions and abbreviations are used throughout this thesis:

- **AI**: Artificial Intelligence.
- **ANN**: Artificial Neural Network.
- **CIN**: Chromosomal Instability.
- **CIMP**: CpG Island Methylator Phenotype.
- **CRC**: Colorectal Cancer.
- **CLR**: Contrastive Learning of Representations.
- **CNN**: Convolutional Neural Network.
- **CWSS**: Complex-Wavelet Structural Similarity.
- **EMT**: Epithelial-Mesenchymal Transition.
- **FAP**: Familial Adenomatous Polyposis.
- **FFPE**: Formalin Fixation and Paraffin Embedding.
- **FID**: Fréchet Inception Distance.
- **GAN**: Generative Adversarial Network.
- **GTRF**: Glasgow Tissue Research Facility.
- **H&E**: Haematoxylin and Eosin.
- **HNPCC**: Hereditary Nonpolyposis Colorectal Cancer.
- **HPF**: High Powered Field.
- **IHC**: Immunohistochemistry.
- **ITBCC**: International Tumour Budding Consensus Conference.
- **ITB**: Intratumoural Budding.
- **LMS**: Long, Medium, Short.

- **MAE**: Mean Absolute Error.
- **MET**: Mesenchymal-Epithelial Transition.
- **MSE**: Mean Squared Error.
- **MSI**: Microsatellite Instability.
- **MMR**: Mismatch Repair.
- **MSS**: Microsatellite Stable.
- **OD**: Optical Density.
- **PDC**: Poorly Differentiated Cluster.
- **PTB**: Peritumoural Budding.
- **SSIM**: Structural Similarity Index Measure.
- **SVD**: Singular Value Decomposition.
- **SSL**: Self Supervised Learning.
- **TB**: Tumour Budding.
- **TBS**: Tumour Bud Scoring.
- **TNM**: Tumour, Nodes, Metastases.
- **VIHC**: Virtual Immunohistochemistry.
- **WSI**: Whole Slide Image.



Thesis Word Count: 55478

Acknowledgements

Firstly, I would like to express my profound gratitude to my supervisor, Robert Insall, for giving me the opportunity to be part of his lab, overlooking my lack of experience in biology and giving me the chance to learn from him. Throughout my PhD, his unwavering support and encouragement have been invaluable. Five months into my PhD the pandemic began, and during those challenging times, Robert was one of my links to the outside world and his passion for science and his willingness to engage in thought-provoking discussions kept my spirits up, enriched my research and were a source of inspiration. I wish to thank him for allowing me to explore my intellectual curiosity and for his many hours supervising me and moulding my development as a scientist. The lessons I have learned will carry forward into the rest of my career. I am truly fortunate to have had him as my mentor, and I deeply appreciate all he has done for me.

I would also like to thank my second supervisor, Joanne Edwards. Her expertise in pathology has been indispensable, and I am deeply grateful for her patience in fielding my numerous questions over the years. Furthermore, the datasets provided by her lab have been the bedrock upon which this work stands. Her contributions have been pivotal in shaping the direction and depth of my research, and I am immensely thankful for her guidance and support.

I want to take this opportunity to thank Cancer Research UK for funding this research and recognising the potential of AI. Their commitment to bridging the realms of technology and biomedical research exemplifies a bold stride towards a future where we will harness the power of AI to unravel the intricacies of cancer. Thank you for providing the resources to enable this work.

I have immense respect and awe for the research taking place at the Cancer Research UK Scotland Institute. I wish to thank the entire Beatson community for welcoming me, inspiring me, and enabling my research; without them, it would not have been possible. I would particularly like to thank the current and former R06, R02 and Y90 lab members for their help, support, feedback, and camaraderie over the last four years. I owe special thanks to Crispin Miller and Ke Yuan for being members of my review panel and for providing valuable feedback and suggestions for my work. I likewise must express my deep gratitude to John Le Quesne and Hayley Morris for their invaluable input and analysis of pathology features found in this work and the sacrifice of their time to help me without asking for anything in return.

I also must acknowledge and express my gratitude to the IT team, particularly Peter McHardy and Iain White, for installing and maintaining the servers and GPUs that are the backbone and workhorse of this research. Additionally, Naveed Khan for allowing access to the High-Performance Compute Cluster that expedited so much of my analysis. Also, Colin Nixon and the members of central services for carrying out the IHC staining on my datasets.

I furthermore wish to thank the members of the Edwards lab, particularly Hester Van Wyk, for sacrificing so much time to teach me about the pathology surrounding tumour budding. I would also like to recognise Phimmada Hatthakarnkul and Kathyrn Pennel for curating and providing the patient databases to enable survival analysis and the entire team for assisting in the generation of my training datasets, including Jennifer Hay and all of the members of the Glasgow Tissue Research Facility for providing the tissue and scanning my slides.

I also must recognise the Doctors and Staff at the University Hospital Hairmyres and University Hospital Monklands, as well as the entire NHS, for providing excellence in care and the researchers who developed the treatments that enabled me to recover from my own experience with cancer and allowed me to continue my journey. Without them, none of this would have been possible.

I would also like to express my gratitude to my friends, particularly Stacey and Steven Connan-McGinty and Michael Murphy, for your unwavering support, laughter, and friendship over the years. In times of stress, you provided solace and distraction and helped me maintain a balance between work and life. I am deeply grateful to each of you for enriching my life in countless ways.

To Panupon, I must express my deepest gratitude for your patience, encouragement, support, laughter, love, and the experiences we have shared. Navigating the ups and downs of this journey would have been far more challenging without you. Thank you for being there, and I look forward to our shared adventures together.

Finally, words cannot express my gratitude towards my family, whose unwavering love and support have been the foundation on which I have built my life. I owe them all a debt of thanks. Particularly my aunts Annemarie and Christine Marshall, for their invaluable advice, love and unreserved support over the years that helped mould me and enabled me to pursue my career. Also, to all our family members that we have loved and lost. Their memories continue to inspire and drive me. Above all, I must thank my parents, Austen and Pauline Walsh. Their unconditional love, nurturing environment, and the countless opportunities they provided have been instrumental in allowing me to chase my dreams. Your sacrifices and belief in me have made all the difference. I am eternally grateful, and this work is for you.

Chapter 1

Overview

The tumour nodes metastasis (TNM) staging system is described as the most effective for many cancers (Loughrey et al. 2022), and it is the current gold standard for colorectal cancer (Weiser 2018). The TNM system is a set of procedures used to stage a patient's disease based on the pathological evaluation of collected samples, such as radiology images, resected tumour sections and lymph node biopsies (Brierley et al. 2016). However, weaknesses have been identified for its use in staging colorectal cancer, as it only utilises the current spread and location of the tumour and does not consider tumour aggression (Lea et al. 2014; Puppa et al. 2010; Dawson et al. 2019b). This can result in significant heterogeneity in patient outcomes within a stage (Maguire 2014) and leads to over and under-treatment (Lea et al. 2014). Therefore, additional prognostic factors must be considered to stratify patients within-stage to tailor the intensity of treatment to match tumour aggression and improve patient prognosis (Lugli et al. 2017).

In the search for additional prognostic factors, tumour budding (TB) emerged as a promising biomarker for colorectal cancer (Dawson et al. 2019c). A tumour bud is a single tumour cell or a cluster of up to four cancer cells in the invasive margin of colorectal cancer (Mitrovic et al. 2012). A higher density of tumour buds in the invasive margin has now been widely recognised to correlate with a higher TNM stage, higher tumour grade, to be indicative of lymph node involvement and the presence of distant metastases (Van Wyk et al. 2019; Rogers et al. 2016; Petrelli et al. 2015). Most notably, it has been accepted as an independent predictor of survival and regression in colorectal cancer (Mitrovic et al. 2021).

Tumour budding is thought to be a characteristic of tumour aggression and is described as a morphological manifestation of a tumour adopting aggressive properties through epithelial to mesenchymal transition (Redfern et al. 2018). But, this is still an ongoing area of research and debate (Grigore et al. 2016; De Smedt et al. 2017; Yamada et al. 2017). However, the idea that budding is a secondary effect of aggression is reinforced by the fact that not all aggressive tumours exhibit a high degree of budding. But the mechanisms responsible for the manifestation of aggressive growth patterns continue to be poorly understood (Pavlič et al. 2022b), which suggests the existence of still-undiscovered tumour characteristics or alterations induced in the stroma and tumour microenvironment that are closely linked to aggression. With modern image analysis techniques, it might be possible to discover and understand these features, which may lead to potential targets for treatment.

Determining tumour aggression through budding is called tumour bud scoring (Van Wyk et al. 2015). It involves counting the number of buds within 0.785mm^2 regions termed hotspots along the invasive margin of solid cancers, and the region with the highest bud count is used to determine the bud score (Lugli et al. 2017). Despite the recognised value of tumour bud scoring, its adoption into standard reporting practice has been slow due to a lack of consensus on a standardised scoring method (Van Wyk et al. 2019), poor inter-observer agreement, lack of reproducibility in haematoxylin and eosin and the time-intensive nature of the task (Lugli et al. 2017). Immunohistochemistry (IHC) targeted to highlight tumour-derived epithelial cells can alleviate the issues with poor inter-observer agreement for manual scoring (Yamadera et al. 2019; Kai et al. 2016). However, the requirement for physical IHC reduces the clinical impact of bud scoring due to the increased resource requirements. But a fully automated pipeline for the prediction of tumour aggression would alleviate the issues with standardisation, inter-observer agreement and reproducibility and allow for the inclusion of tumour bud scoring in standard reporting (Studer et al. 2021).

Over the last decade, progress in the hardware and software used for deep and machine learning has allowed significant advances in the digital analysis of medical images (Tamang et al. 2021). This includes many effective applications of deep learning to colorectal cancer, such as survival prediction using a novel stroma score calculated from features extracted from a deep neural network by Kather et al. (2019). Or the ability of deep adversarial networks to learn to generate photorealistic pathology images that retain the genetic information encoding for microsatellite instability, as shown by Krause et al. (2021). Another example is the ability of self-supervised deep networks to detect colorectal tumours using small amounts of semi-supervised data, as demonstrated by Yu et al. (2021). The success of these works reflects the richness of information present in H&E whole slide images and the potential of deep learning to utilise and extract new information from that data.

Therefore, given the need for automated independent aggression prediction in colorectal cancer and an exploration of the morphology of aggressive growth patterns at the invasive margin, the objectives of this work have been two-fold; the first was to explore methods of assisted and automated tumour bud scoring and the second was to discover new methods of direct automated aggression prediction. The second involves using state-of-the-art deep learning methods that allow the possibility of interrogation to discover morphological features correlated with aggression.

These objectives took the form of three projects. The first was to develop a tool to assist in creating ground truth for training tumour bud detection networks or assist a pathologist in manual tumour bud scoring. This project developed a deep network capable of translating from H&E to a photo-realistic virtual IHC stain that highlights tumour-derived epithelial cells. The second project used that ground truth to train a tumour bud segmentation network to directly segment tumour buds from H&E to generate an automated score. The third and final project was to use self-supervised deep networks to learn feature representations from H&E colorectal cancer tiles extracted from the invasive margin. These could be grouped through unsupervised clustering to interrogate the encoded tissue morphologies. It was then possible to demonstrate that those clusters can be used directly for survival prediction through machine learning. Alternatively, and more powerfully, this project has also shown that it is possible to use those self-supervised representations as input to train a supervised transformer network on whole slide labels that is a generalisable independent predictor of survival.

1.1 Thesis Outline

This thesis is structured as follows:

- **Chapter 1** - A high-level overview of the limitations of the TNM staging system for colorectal cancer, the need for aggression prediction methods, the arguments for why automation and digital analysis are required and how this work is placed to fill a key selection of those requirements.

- **Chapter 2** – An introduction to the necessary background of histopathology, colorectal cancer, tumour budding, tumour bud scoring, and the deep learning methods used in this work. We shall also discuss the current state of the art in virtual immunohistochemistry, automated tumour bud scoring and automated aggression prediction in colorectal cancer.

- **Chapter 3** - A description of the contribution of this work to virtual immunohistochemistry. Details are provided on the dataset and the development of the deep generative network. As well as a review of the network's performance and a discussion of the results.

- **Chapter 4** - A description of this work's techniques for automated tumour bud scoring. Details are presented on the dataset and methods used to develop the deep segmentation network. As well as the network's performance and the discussion of the results.

- **Chapter 5** - A description of this work's contribution to invasive margin detection and automated aggression prediction. Details are provided on the dataset and methods used to develop the deep self-supervised network to encode representations, the clustering process to interrogate them, and finally, the transformer network to interpret the invasive margin for aggression classification.

The performance, generalisability and future directions for the work are also assessed.

- **Chapter 6** - A discussion of the contributions and limitations of this work as a whole, the conclusions drawn from it, and the possible future directions it may take.

Chapter 2

Background

In this chapter, the background for histopathology and colorectal cancer will be introduced, as well as tumour budding and its biological context and use cases. The most relevant methods for digital analysis of pathology slides will be reviewed, including techniques such as stain normalisation and whole slide registration and alignment. The relevant theory will also be introduced for the deep learning methods used in this work, such as deep convolutional and generative networks, self-supervised learning, clustering algorithms, and transformers. Finally, we will review the state-of-the-art approaches to virtual immunohistochemistry, automated tumour bud scoring, and automated aggression prediction in colorectal cancer.

2.1 Histopathology

Histology and histopathology are often discussed together as they are interdependent (Musumeci 2014). Histology studies the structure and detail of healthy biological cells using brightfield, fluorescence or electron microscopy (O'Dowd et al. 2003). Histopathology is a field that also examines the structure and detail of biological cells through microscopy, with a key distinction; this field is dedicated to discovering the changes in cell structure that might be the cause or result of disease (Kusuma Dewi et al. 2023). This difference is critical, as understanding the distinctions between healthy and diseased tissue is a prerequisite for human or artificial intelligence-driven diagnosis through the evaluation of cell structures and their environment.

Histopathology in colorectal cancer involves analysing tissue samples obtained from the colon and rectum to detect the presence of anomalies, cancer and unusual cell morphology or structure (Compton 2003). Samples are typically acquired through surgical resection or biopsy (Arslan et al. 2020). Resection involves the removal of the tumour along with the surrounding tissue through surgery (Krbal et al. 2017), and biopsy consists of the collection of suspicious tissue, usually by needle core or endoscopic procedure (Tanaka et al. 2015). Care must be taken during sampling to avoid accidental tissue damage through crushing, haemorrhage, split or fragmentation (Taqi et al. 2018). Once a sample has been gathered, it must undergo a series of processing steps to ensure it remains as life-like as possible for inspection by microscopy.

The first of these processing steps is called fixation. It is necessary to maintain the structure of the tissue and prevent autolysis. Autolysis is when tissue undergoes self-digestion and degradation due to enzymes released after cell death (Slaoui et al. 2011). Fixation prevents this by preserving the cells through a chemical process. Formalin fixation is the most common chemical fixative in histopathology (Grigorev et al. 2018). It fixes the tissue by cross-linking the cell proteins and nucleic acids, which arrests the cell's enzymatic activity and prevents further degradation (Fattorini et al. 2020). Once the tissue is fixed and decay halted, subsequent processes can be performed to allow examination.

The next step in the histopathology pipeline is to embed the tissue. This is the process of permeating the tissue with a solid medium, typically paraffin wax, to provide support during sectioning to stop the tissue from deforming (Van Der Lem et al. 2021). These fixation and embedding steps are known as Formalin Fixation and Paraffin Embedding (FFPE) (Walsh et al. 2023). Embedding involves dehydrating the tissue by slowly replacing the water with alcohol, then clearing the tissue with a substance such as xylene to remove the alcohol and make the tissue transparent. Finally, the tissue is infiltrated with the paraffin embedding medium to make it less deformable to facilitate sectioning (Dey 2022).

Now that the tissue has been structurally protected, it can endure microtomy (Lyon et al. 1991). This is where the tissue is cut into thin slices known as sections using a microtome, a precision instrument designed to produce uniform and micron-thick tissue cuts (Slaoui et al. 2011). This is important as thinner sections allow more light to pass through the tissue when examined by brightfield microscopy. This makes it easier to visualise the cell structure and components (Mohammed et al. 2012). The sections are usually between three and five micrometres thick (Xu et al. 2019). Every section consumes a portion of the sample, and therefore, slices are a finite resource.

Embedded tissue sections in their raw state offer very little contrast (Slaoui et al. 2011). They cannot be examined by microscopy without adding a staining reagent to differentiate the tissue structures (Thajudeen et al. 2023). Therefore, after embedding, tissue staining is the next step in the histopathology workflow. Haematoxylin & eosin (H&E) is the most common stain in modern pathology (Pichat et al. 2018). It was first introduced in the late 1870s, where it quickly gained popularity as a tissue staining method and has since been in widespread use for over a century (Javaeed et al. 2021). H&E is composed of two stains. The first is haematoxylin, which is a basic dye that adheres to the basophilic structures in cells, such as the phosphate backbone of DNA in cell nuclei, furnishing them with a purple colour (Slaoui et al. 2011). The second stain is eosin, which is an acidic dye that adheres to the acidophilic structures in tissues, mainly the proteins in the cell cytoplasm and extracellular proteins

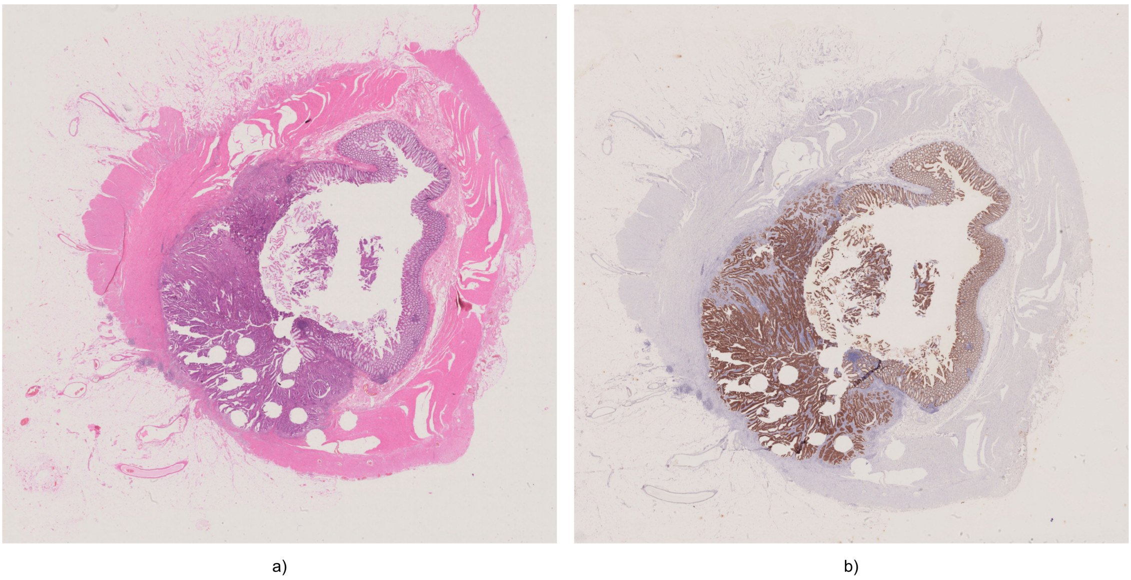


Figure 2.1: Representative whole slide images. a) Tissue stained with haematoxylin and eosin. b) The corresponding tissue stained via immunohistochemistry using a pan-cytokeratin antibody mix that targets the AE1 and AE3 antigens.

such as collagen, staining these in various shades of pink (Dey 2022). H&E helps histopathologists differentiate tissue structures and examine the tissue microenvironment and is particularly valuable for investigating cellular infiltrates (Lo et al. 2021). It is frequently used for a wide variety of pathology tasks like segmentation of glands in the prostate (Li et al. 2020), classification of early pancreatic cancer (Langer et al. 2015) and staging of colorectal cancer (Fleming et al. 2012). This suggests that a broad scope of information is held within H&E stains. However, it indiscriminately stains both tumour and non-tumour cells. Therefore, diagnosing cases with ambiguous histology or poor tumour differentiation can be challenging in H&E alone, as the human eye can have difficulty distinguishing subtle changes in colouring. In cases like this, other techniques, such as immunohistochemistry, are often applied to provide further contrast (Puppa et al. 2012). An example of a whole slide image of a tissue section stained with haematoxylin and eosin is shown in figure 2.1.

Immunohistochemistry (IHC) highlights specific proteins in a tissue section using targeted antibodies (Schacht et al. 2015). The bound antibodies are made visible by affixing chromogens of various colours (Xu et al. 2019). Selecting a chromogen colour that contrasts the counterstain (usually haematoxylin) for antibodies selectively expressed by the targeted cancer cells makes them readily distinguishable from other tissues. This process has several advantages over H&E; it provides a sharp contrast between tissue types, allowing for fast diagnosis by removing ambiguity and improving the accuracy of diagnosis and inter-observer agreement between pathologists (Bokhorst et al. 2020). However, these improvements come at a cost: the expense of the antibodies, the need for complex lab equipment, and specialised personnel to carry them out (Schacht et al. 2015). It is also a much lengthier process than H&E staining. These limitations mean that despite its value, IHC is generally reserved for complex cases, and not all labs have the resources to access it (Lugli et al. 2017). An example of a whole slide image of tissue stained with an IHC antibody targeting epithelial cells is visible in figure 2.1 b); note by comparison with the H&E stain of the same tissue shown in a) how much more distinct and identifiable the tumour and colon glands are from the surrounding tissue in the IHC stained slide.

Once the tissue samples have been prepared, they are examined by a histopathologist using brightfield microscopy (Kusuma Dewi et al. 2023). The pathologist assesses tissue architecture, cellular morphology and the tissue microenvironment to identify regions of necrosis, inflammation, and cellular alteration (Varone et al. 2012). This analysis aims to diagnose diseases like malignancy and infection to guide patient diagnosis, prognosis and treatment by elucidating the disease origin and progression. Historically, a diagnosis from a histopathology image was the gold standard for many diseases, including most cancers (Gurcan et al. 2009). But today, histopathology examinations are augmented with genomics, transcriptomics and proteomics to understand disease processes (Walsh et al. 2023). But histology images are still core in discovering new cellular morphologies associated with known or unknown diseases and their subtypes, often through new digital analysis techniques (Niehues et al. 2023).

Traditional histopathology relied on manual examination of tissue through the lens of a microscope (Titford 2006). However, over the last twenty years, there has been a cumulative shift towards digital pathology with the invention of whole slide imaging in 1994 (Pantanowitz et al. 2018). This trend has various driving forces; manual pathology was labour-intensive and time-consuming, and digital pathology has increased the efficiency and speed of reporting (Stathonikos et al. 2019). Digital pathology also allows for the standardisation of image acquisition and quality (Luchini et al. 2022). It enables immediate access to stored slides, allows for the simultaneous viewing and handling of multiple slides, and enhances the ability for pathologists to collaborate locally and across the globe (Vodovnik 2016). Perhaps the most compelling reason is that it allows a transition from qualitative and descriptive reporting in pathology to quantitative. Digital pathology enables the use of image analysis algorithms, annotations, measurement tools and deep learning. This allows the extraction of objective measurements like cell counts, tumour size, and staining intensity, which can guide diagnosis, prognosis and research (Olsen et al. 2018; Aeffner et al. 2018).

In summary, histopathology is at the heart of medical diagnosis and research. It is the cornerstone of understanding disease at the tissue level through the examination of tissue architecture, cell morphology and cytology to understand the origin, state and prognosis of disease. The modern process of fixation, embedding, staining and digitisation has provided a platform for advancement through increased efficiency of reporting and collaboration and unlocked digital analysis through deep and machine learning. The transformation from these advancements has yet to reach its full potential for progressing medical science and improving patient care.

2.2 Colorectal Cancer

Cancer is a complex disease where cells grow uncontrollably and can spread to other parts of the body. Instead of functioning normally, these cells acquire specific capabilities that allow them to evade the body's natural defences and continue to multiply. Over time, research has identified particular traits or "hallmarks" that these cells possess that help to explain their aberrant behaviour (Hanahan et al. 2000; Hanahan et al. 2011).

The hallmarks of cancer serve as a foundational framework for understanding the nature of cancer. It was introduced to distil the vast complexities of cancer's many phenotypes and genotypes into a set of core principles. This idea was presented by Hanahan et al. (2000) when they identified six biological capabilities that human cells acquire as they transition from healthy to neoplastic cells, which they proposed were essential for the formation of malignant tumours (Hanahan et al. 2000). A decade later, Hanahan et al. (2011) revisited and expanded upon their initial framework, using the latest advancements in cancer research to introduce two additional features.

There are eight hallmarks described by Hanahan et al. (2011). The first is that cancer cells can continuously signal themselves or nearby cells to keep growing and dividing. The second is that normal cells have built-in checks to prevent uncontrolled growth, but cancer cells can bypass these safety measures. The third distinguishing feature is that normal cells also have a self-destruction mechanism for programmed cell death when things go wrong, but cancer cells can resist this process. The fourth is that while most cells have a limit to how many times they can divide, cancer cells can potentially divide indefinitely, as is the case with the HeLa cell line resulting from samples taken from Henrietta Lacks, where the cells are still alive and continue to divide over 60 years after her death (Khan 2011). The fifth hallmark is that to nourish themselves, cancer cells can also stimulate the formation of new blood vessels. The sixth is one of the most dangerous features, its ability to spread and invade other parts of the body, something not observed in regular cells. The seventh feature is that cancer cells have a unique ability to change their energy production methods, allowing them to thrive in environments that challenge normal cells. The eighth and final hallmark is that our immune system is designed to detect and destroy abnormal cells, but cancer cells have found ways to evade this detection and avoid destruction. In a more recent exploration Hanahan (2022) has proposed the addition of more dimensions to the fundamental principles of cancer, emphasising the importance of future research to understand the tumour microenvironment better and suggested new avenues of exploration to refine our understanding of cancer hallmarks (Hanahan 2022).

Colorectal cancer (CRC), also known as colorectal adenocarcinoma, is a frequent digestive tract malignancy that usually arises from the glandular epithelial cells of the large intestine (Rawla et al. 2019). It is the third most common cancer worldwide and has the second highest mortality rate both worldwide (Chen et al. 2021a) and in the UK (Rua et al. 2020). The incidence of colorectal cancer increases with age, with the majority of cases occurring in individuals over 50 (Rawla et al. 2019). However, the number of cases in patients younger than 50 is increasing, and there have been recommendations to lower the screening age accordingly (Grant et al. 2022). Annually, the predicted global cases of colorectal cancer are expected to increase by 60% to over 2.2 million by 2030 (Rawla et al. 2019). This is due to an ageing population, environment and lifestyle changes, and increased exposure to risk factors (Arnold et al. 2017).

The cause of colorectal cancer can be attributed to a combination of lifestyle, genetic and environmental factors. Some known risk factors are smoking, alcohol, red meat and processed food consumption, as well as obesity and physical inactivity (Cho et al. 2019; Ramesh et al. 2018). Several inherited syndromes can lead to colorectal cancer, such as Lynch Syndrome, also known as Hereditary Nonpolyposis Colorectal Cancer (HNPCC), caused by germline mutations in DNA mismatch repair genes (Testa et al. 2018). These mutations result in microsatellite instability, a characteristic feature of HNPCC-associated colorectal cancer (Talbot et al. 2018). It is the most common inherited cause of colorectal cancer, with about 2-5% of cases caused by it (Testa et al. 2018). Another inherited cause is familial adenomatous polyposis (FAP). It is caused by mutations in the APC tumour suppressor gene, leading to the abundant formation of adenomatous polyps in the colon and rectum (Samadder et al. 2019). If these are not removed, the lifetime risk of colorectal cancer approaches 100% (Detweiler et al. 2016). Many other inherited conditions lead to an above-average lifetime risk of developing colorectal cancer (Samadder et al. 2019). But it is important to note that these environmental, genetic and lifestyle causes often interact, and the combined effect of multiple risk factors is often the origin of colorectal cancer (Cho et al. 2019).

The biological pathways that lead to colorectal cancer are diverse and complex, but three molecular and two morphological paths have been identified as fundamental routes to colorectal adenocarcinomas (Kim et al. 2018). The first of the molecular pathways is through chromosomal instability (CIN). This is indicated by changes to the number and structure of a cell's chromosomes and mutations of tumour suppressor and oncogenes (Malki et al. 2020), like the RAS oncogene, where approximately 50% of colorectal carcinomas and a similar percentage of large adenomas have this mutation (Fearon et al. 2023). It has been demonstrated that the RAS gene encodes a protein that is involved in transmitting cell signals, mutations of which can lead to the production of an always-active RAS protein that can promote uncontrolled cell division, one of the hallmarks of cancer (Fernández-Medarde et al. 2011). Another example is the loss of chromosome 17p observed in over 75% of colorectal carcinomas but infrequent in adenomas, and this region contains the p53 tumour suppressor gene (Hollstein et al. 1991). The CIN pathway accounts for 60-80% of cases (Nguyen et al. 2020). The second molecular route is the MSI pathway, which is characterised by mutations in the DNA mismatch and repair (MMR) genes (Kim et al. 2018). MMR corrects errors that randomly occur during DNA replication by removing the defective sequence and inserting the correct one (Nguyen et al. 2020). Failure to repair these defects can again lead to tumour suppressor and oncogene mutation and, as a result, colorectal cancer (Kim et al. 2018). This pathway accounts for 15-20% of all CRC cases (Malki et al. 2020). The third molecular pathway is the CpG island methylator phenotype (CIMP). DNA methylation is the process of adding a methyl group to the DNA molecule; it is a mechanism for regulating gene expression without altering the sequence itself (Elhamamsy 2016). In the CIMP pathway, there is widespread methylation of promoter sites on the DNA sequence of tumour suppressor genes (Nguyen et al. 2020). These sites should activate the transcription of the sequence into proteins. When methylated, these are effectively silenced and deactivated, leading to cancer (Kim et al. 2018). Approximately 20% of CRC tumours arise from the CIMP pathway (Malki et al. 2020). The first morphological pathway is the classical adenoma-carcinoma sequence. A premalignant tubular outgrowth of epithelial cells called an adenoma that later progresses to a malignant carcinoma characterises this route (Kim et al. 2018). The second morphological route is the serrated neoplasia pathway. These are a type of outgrowth found in the colon or rectum with a pattern resembling serrations or saw-toothed edges (Thorlacius et al. 2017). This pathway comprises three subtypes: hyperplastic polyps, sessile serrated adenomas, and traditional serrated adenomas (Kim et al. 2018). Hyperplastic polyps have narrow bases with the serrations confined to the top of the crypt; they have a low potential to turn cancerous (Nguyen et al. 2020). Sessile serrated adenomas don't have a stalk, grow flat against the colon wall, can be easily missed, and have a high malignant potential (Thorlacius et al. 2017). Traditional serrated adenomas are indicated by prominent growths and villiform projections, making them easier to detect but hard to distinguish from conventional adenomas. They are less common but have a high malignant potential (Nguyen et al. 2020). Different molecular pathways drive these morphologic routes to carcinoma. The classical pathway is driven by CIN or MSI (Kim et al. 2018). The serrated neoplastic pathway is led by CIMP and optionally MSI (Malki et al. 2020).

The gold standard staging system for colorectal cancer is the TNM staging system, which considers the attributes of the primary tumour (T), the involvement of the local lymph nodes (N) and the presence of distant metastasis (M) (Arrichiello et al. 2022). Therefore, the stage of colorectal cancer is determined based on the size and invasion of the tumour through the colon wall, the range of involvement of regional lymph nodes, and the presence of metastases in other organs of sites (Wu et al. 2022). The latest and 8th edition of the TNM system published by Brierley et al.

(2016) divides reporting of CRC into stages based on the T, N and M scores. The definition of the scores for colorectal cancer is shown in figure 2.2, and the stage is calculated based on the combination of T, N and M scores shown in table 2.1. The mortality rates for colorectal cancer vary depending on the stage of the disease.

Tumour (T):

TX Primary tumour cannot be assessed.

T0 No evidence of primary tumour.

Tis Carcinoma in situ: invasion of the lamina propria.

T1 Tumour invades submucosa.

T2 Tumour invades muscularis propria.

T3 Tumour invades subserosa or into non-peritonealised pericolic or perirectal tissues.

T4 Tumour directly invades other organs or structures and/or perforates visceral peritoneum.

- **T4a** Tumour perforates visceral peritoneum.
- **T4b** Tumour directly invades other organs or structures.

Regional Lymph Nodes (N):

NX Regional lymph nodes cannot be assessed.

N0 No regional lymph node metastasis.

N1 Metastasis in 1-3 regional lymph nodes.

- **N1a** Metastasis in 1 regional lymph node.
- **N1b** Metastasis in 2-3 regional lymph nodes.
- **N1c** Tumour deposit(s) in the subserosa, mesentery, or non-peritonealised pericolic or perirectal tissues without regional nodal metastasis.

N2 Metastasis in 4 or more regional lymph nodes.

- **N2a** Metastasis in 4-6 regional lymph nodes.
- **N2b** Metastasis in 7 or more regional lymph nodes.

Distant Metastasis (M):

M0 No distant metastasis.

M1 Distant metastasis.

- **M1a** Metastasis confined to one organ or site (liver, lung, ovary, non-regional node), without peritoneal metastases.
- **M1b** Metastasis in more than one organ/site.
- **M1c** Metastasis to the peritoneum with or without other organ involvement

Figure 2.2: The TNM System for Colorectal Cancer from (Brierley et al. 2016).

Stage	T	N	M
0	Tis	N0	M0
I	T1	N0	M0
IIA	T2	N0	M0
IIB	T3	N0	M0
IIC	T4a	N0	M0
IIIA	T1-2	N1	M0
IIIB	T3-4a	N1	M0
IIIC	T4a	N2	M0
IIID	T4b	N0-2	M0
IVA	Any T	Any N	M1a
IVB	Any T	Any N	M1b
IVC	Any T	Any N	M1c

Table 2.1: TNM Stage Reference for Colorectal Cancer from (Brierley et al. 2016).

There is a limitation with the TNM system for prognosis prediction in colorectal cancer (Karamchandani et al. 2020). As is shown in table 2.1, the TNM system leans heavily on lymph node involvement for prediction, which has been criticised for its weaknesses in accuracy for prognosis, particularly in stage II colorectal patients (Søreide et al. 2016). The TNM system does not consider the varying aggression of tumours of the same stage (Dawson et al. 2019b). This, coupled with the heterogeneity of the disease and intra and inter-pathologist variability in reporting systems, means that the accuracy of prognosis is limited for early-stage cohorts, and there can be significant variability in survival (Dimitriou et al. 2018). Patients with aggressive tumours that haven't yet notably spread can be under-treated because they are early stage as defined by the TNM system (Lea et al. 2014).

Treatment for colorectal cancer varies depending on the determined stage. The main treatments can include surgery, chemotherapy, radiotherapy, and, more recently, immunotherapy (Alaryani et al. 2022). Surgical resection is curative for stage I colorectal cancer when it is limited to the inner lining of the colon or rectum. However, adjuvant chemotherapy may be considered for high-risk patients (Akagi et al. 2020). For stage II, which has penetrated through the colon or rectum wall but has not spread to nearby lymph nodes, with no detectable distant metastases, surgical resection is also the primary treatment, and again, adjuvant chemotherapy may be recommended for high-risk cases (Zha et al. 2022). For stage III CRC, which has spread to nearby lymph nodes but not to distant sites, the standard treatment is surgical resection followed by adjuvant chemotherapy (Akagi et al. 2020). Treatment options for stage IV colorectal cancer that has spread to distant organs or tissues may include a combination of surgery, chemotherapy radiotherapy, and immunotherapy (Akagi et al. 2020). In some cases of locally advanced rectal cancer, neoadjuvant treatment in the form of a combination of chemotherapy and radiotherapy may be considered to shrink the tumour before surgery to improve the chances of a successful resection (Kim et al. 2022). Adjuvant chemotherapy and radiotherapy can harm the quality of life after treatment, sometimes temporarily, sometimes permanently. Therefore, the physician and patient must consider trade-offs of survival benefit vs the incidence of adverse events during therapy (Zha et al. 2022).

Given the lack of stratification for early-stage colorectal cancer that would allow for better decisions on what level of treatment to provide, there have been calls to include additional prognostic biomarkers in the reporting of colorectal cancer to help stratify patients within-stage to better tailor the treatment to the aggression and characteristics of the tumour, to stop over and under treatment, and improve both patient survival and quality of life (Fotheringham et al. 2019; Sun et al. 2019).

In summary, colorectal cancer or colorectal adenocarcinoma originates in the glandular epithelial cells of the large intestine. Predominantly driven by lifestyle, age and environmental factors, the global CRC cases are set to more than double by 2030. The development of colorectal cancer involves various intricate molecular pathways such as CIN, MSI and CIMP with two known morphological routes. The TNM staging system, while standard for CRC, has shown limitations in accuracy, particularly for early-stage colorectal cancer. Treatment depends on the stage and includes surgery, chemotherapy, radiotherapy and immunotherapy. Given that early-stage prognostics are inaccurate, research efforts are ongoing to include additional biomarkers to fine-tune treatment protocols and enhance patient survival and quality of life.

2.3 Tumour Budding

There is a histological phenomenon called Tumour Budding (TB), which is a promising biomarker for independent survival prediction in colorectal cancer (Mitrovic et al. 2021). Tumour budding refers to the presence of single cells or small clusters of less than five tumour cells present in the invasive margin of a tumour (Lugli et al. 2017). Tumour budding and its implications for prognosis were first recognised and discussed over sixty years ago by Imai (1960). However, the concept of tumour budding has evolved, and a greater understanding of its biological processes and significance as an independent prognostic factor has gained increasing attention in recent years (Lino-Silva et al. 2018). An example of tumour buds is shown in figure 2.3, where buds in tissue stained with H&E are visible in panel a).

The biological mechanisms behind tumour budding in colorectal cancer are still not fully understood, and there is ongoing research to comprehend its molecular processes better (Hatthakarnkul et al. 2021). However, it is thought that tumour budding is the migration of tumour cells outwards from their tumour of origin, leading to invasion of the surrounding tissue and stroma (De Smedt et al. 2016). This process is believed to be a fundamental characteristic of malignant cells and is the initial step

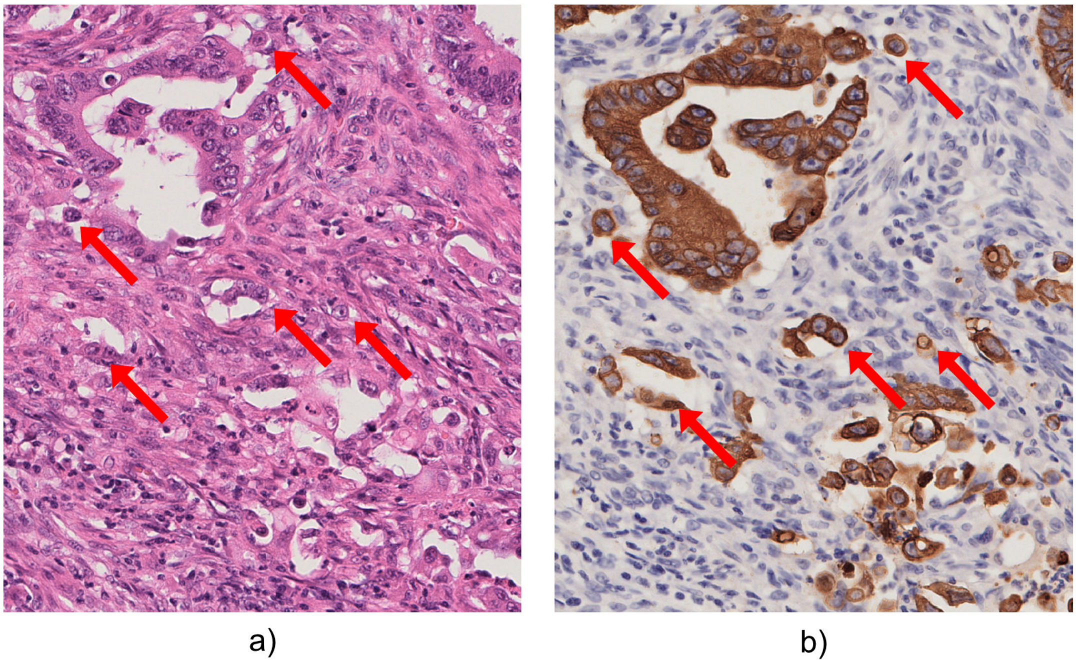


Figure 2.3: Depiction of tumour budding observed in tissue samples. a) Tissue stained with Haematoxylin and Eosin (H&E). b) The corresponding tissue highlighted with pan-cytokeratin AE1/AE3 Immunohistochemistry (IHC). Red arrows pinpoint selected tumour buds.

in tumour invasion and metastasis (Lino-Silva et al. 2018). The prevailing theory is that it is a histological manifestation of epithelial to mesenchymal transition (Grigore et al. 2016).

Epithelial-to-mesenchymal transition (EMT) is a process where cells lose their epithelial characteristics, principally their cell-to-cell adhesion properties, become more motile, contact independent and gain more migratory mesenchymal properties (Redfern et al. 2018). It is a biological process that has evolved to facilitate embryonic development (Ekblom 1989) and wound healing (Haensel et al. 2018). However, in tumours, aberrant EMT can be induced by several factors, such as interactions between the tumour cells and the microenvironment, which can lead to the secretion of cytokines, growth factors and extracellular matrix components that promote EMT (Redfern et al. 2018). In colorectal cancer, EMT is associated with acquiring aggressive behaviour, including invasion, metastasis, and resistance to therapy (De Smedt et al. 2017). EMT is thought to be a dynamic process in that tumour cells have to adopt a range of properties along the Epithelial-Mesenchymal spectrum to detach, migrate, and embed at a distant site, where they reverse this process and undergo Mesenchymal to Epithelial transition (MET) (Redfern et al. 2018). It has also been demonstrated that small groups of cells can adopt a partial epithelial/mesenchymal stage and migrate collectively (Grigore et al. 2016). The phenomenon of tumour budding is thought to be a snapshot of this process in action, early in the spectrum of EMT, partial or otherwise, where the cells are losing their tendency to stick together and, therefore, detach from the tumour core. This can happen singularly or in small clusters, and by adopting more motile properties, the cells migrate into the surrounding tissue and beyond.

The biological significance of tumour budding is now well-established for colorectal cancer, and numerous studies have shown that tumour budding is associated with adverse histopathological features and poorer survival and prognosis in colorectal cancer (De Smedt et al. 2016; Mitrovic et al. 2012; Prall et al. 2005; Zlobec et al. 2010). There are two recognised and distinct patterns of tumour budding in colorectal cancer: Peri-tumoural budding (PTB), which refers to tumour buds present in the invasive margin of the tumour and in the surrounding stroma (Lugli et al. 2017), and intra-tumoural budding (ITB), which refers to the presence of tumour buds within the tumour mass itself (Zlobec et al. 2014). Peri-tumoural budding occurs in the tumour-host interface and involves interaction between tumour cells and the surrounding stroma (Pavlič et al. 2022a). Because of this, PTB can only be assessed in endoscopic or complete surgical resection specimens where both the tumour and a margin around it are extracted (Lugli et al. 2017). Peri-tumoural budding is an independent prognostic biomarker associated with increased aggression, poor survival outcome and recurrence (Studer et al. 2021). It has been established that it is correlated with a high tumour grade, high TNM stage and is predictive of lymph node and distant metastases (Mitrovic et al. 2012; Lugli et al. 2012; Van Wyk et al. 2015; De Smedt et al. 2016; Rogers et al. 2016). Intra-tumoural budding occurs within the primary tumour mass. It reflects peri-tumoural budding in that it is also associated with more aggressive tumour behaviour and adverse clinicopathological features (Zlobec et al. 2014).

Tumour budding also has clinical significance in that it can be applied as a supplementary prognostic tool to facilitate the management of colorectal cancer patients. It can be used three-fold; the first is that budding can predict lymph node metastases in endoscopically resected tumours and help select patients that would benefit from additional surgical resection (Bosch et al. 2013; Ueno et al. 2004; Koelzer et al. 2016). The second is that budding indicates shortened disease-free survival and an increased probability of recurrence in stage II colorectal cancer (Van Wyk et al. 2015; De Smedt et al. 2016; Koelzer et al. 2016; Wang et al. 2009). Therefore, budding can help stratify patients within a stage and select those that would benefit from adjuvant therapy. Finally, intra-tumoural budding in pre-operative biopsies can help determine patients who may benefit from neo-adjuvant treatment and is predictive of tumour regression (Giger et al. 2012; Rogers et al. 2014; Zlobec et al. 2014).

In summary, tumour budding is a histological phenomenon that can serve as an independent biomarker for survival prediction in colorectal cancer. A high density of single tumour cells or small clusters of less than five at the invasive margin of a tumour is clinically prognostic of poor outcome and disease recurrence. It has been studied for over 60 years, and within the last decade or two, it has become increasingly relevant for clinical use. While the biological mechanisms behind tumour budding are still not completely understood, it is thought that it is a manifestation of tumour cells undergoing partial EMT. The biological significance of this is now well established for colorectal cancer, with several studies showing its association with adverse histological features, lymph node and distant metastases, higher tumour grade, TNM stage overall reduced survival, and increased chance of recurrence. Clinically, tumour budding can act as a supplementary prognostic tool to assist in managing colorectal cancer patients. However, despite the significant benefits it affords, clinical adoption has been slow due to a lack of standardisation and reproducibility, inter-observer variability, the time and resource-intensive nature of locating buds, and the lack of a validated method (Lugli et al. 2017). Recent years have witnessed a notable shift in how tumour budding is approached. By focusing on evidence-based quantitative reporting, standardisation and mitigation of limiting factors, progress has been made by forming international guidelines for tumour bud reporting with a scoring system that aims to address previous concerns and facilitate clinical adoption to improve patient outcomes.

2.4 Tumour Bud Scoring

The biological significance of tumour budding has been discussed for over sixty years. However, attempts to quantitatively measure it began less than twenty years ago with the work of Ueno et al. (2004). They suggested that the invasive margin should be manually surveyed for the area of most intensive budding. Once located, buds were to be counted in the viewable area of a 20x objective lens, a site covering 0.785mm^2, termed a high-powered field (HPF). A field with five or more buds was deemed as positive for budding. After the proposal of the Ueno scoring method, many studies have used similar or slightly altered techniques, confirming the utility of the system or suggesting alterations, like this research by Karamitopoulou et al. (2013), which found that the bud score was only prognostic when averaged over ten HPFs, rather than a single field. However, only 215 cases were used in this study, so it may not have been large enough for further patterns to be revealed. But remarkedly, a later study by Martinez Ciarpaglini et al. (2019) found that scoring over ten fields improved the ability of budding to stratify stage II CRC patients. Given its improved contrast and reproducibility, both of these works also suggested that buds be scored in IHC (Martinez Ciarpaglini et al. 2019; Karamitopoulou et al. 2013). Figure 2.3 shows an example of this.

In 2016, an attempt was made to internationally standardise the tumour bud scoring process. A meeting was held in Bern, Switzerland, termed the International Tumour Budding Consensus Conference (ITBCC). The primary aim of this conference was to determine if there could be an agreement on a standardised scoring system. It included participants from eleven countries (Lugli et al. 2017). The voting panel had twenty-five participants, which included twenty-two gastrointestinal pathologists, two surgeons and one translational researcher. An eBook was prepared with all relevant literature on tumour budding in colorectal cancer. The participants were asked to review the literature and vote on a series of statements regarding outstanding questions in the field to agree on the value of tumour budding, its prognostic power and how budding should be evaluated going forward (Lugli et al. 2017). The results of these votes are shown in table 2.2.

No.	**Statement**	**Grade & Recommendation**	**Evidence**
1	Tumour budding is defined as a single tumour cell or a cell cluster consisting of four tumour cells or less.	Strong Agree: 22/22 (100%)	High
2	Tumour budding is an independent predictor of lymph node metastasis in pT1 colorectal cancer.	Strong Agree: 23/23 (100%)	High
3	Tumour budding is an independent predictor of survival in stage II colorectal cancer.	Strong Agree: 23/23 (100%)	High
4	Tumour budding should be taken into account along with other clinicopathological features in a multidisciplinary setting.	Strong Agree: 23/23 (100%)	High
5	Tumour budding should be scored on H&E.	Strong Agree: 19/22 (86%)	Moderate
6	Intratumoural budding exists in colorectal cancer and has been shown to be related to lymph node metastasis.	Strong Agree: 22/22 (100%)	Low
7	Tumour budding should be assessed in one hotspot (in a field measuring 0.785 mm^2) at the invasive front.	Strong Agree: 22/22 (100%)	Moderate
8	For tumour budding assessment in colorectal cancer, the hotspot method is recommended.	Strong Agree: 22/22 (100%)	Moderate
9	A three-tier system should be used along with the budding count in order to facilitate risk stratification in colorectal cancer.	Strong Agree: 23/23 (100%)	Moderate
10	Tumour budding should be included in guidelines/ protocols for colorectal cancer reporting.	Strong Agree: 23/23 (100%)	High
11	Tumour budding and tumour grade are not the same.	Strong Agree: 23/23 (100%)	High

Table 2.2: Statements of the ITBCC 2016 based on the GRADE system from (Lugli et al. 2017).

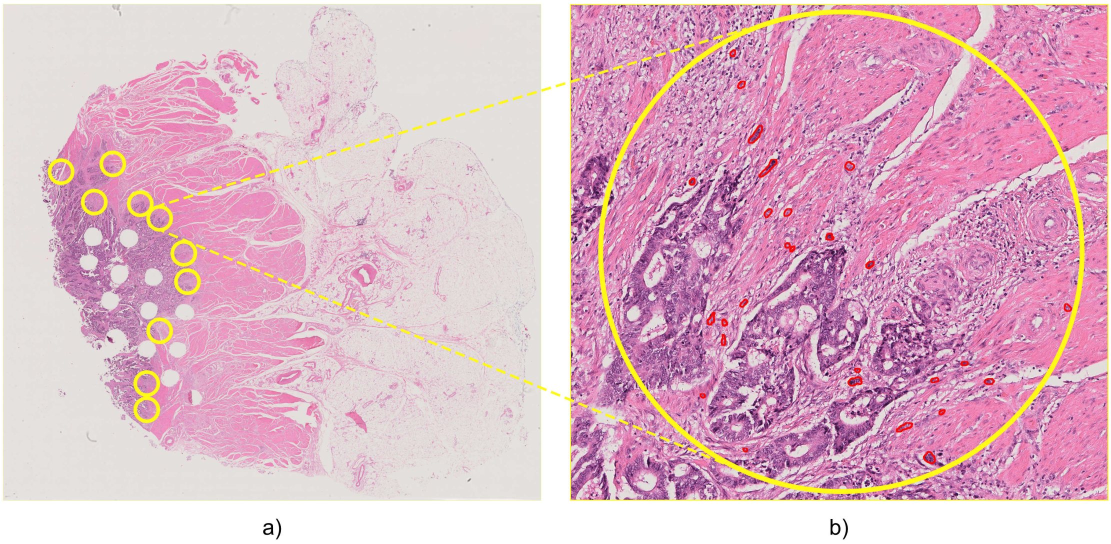


Figure 2.4: Illustration of the ITBCC tumour bud scoring methodology. a) Ten distinct 10x 0.785mm^2 hotspots located within the invasive margin (highlighted in yellow). b) Within the hotspot exhibiting the highest bud density, buds are enumerated at a 20x magnification over a 0.785mm^2 area (tumour buds indicated in red).

Statements 2, 3, 4, 6, 10 and 11 confirm the panel's agreement on the biological significance of tumour budding and its clinical value. Statements 1, 5 and 7 through 9 detail the consensus on how buds should be scored. Statement 1 creates a standard definition of a tumour bud. It differs from the original Ueno method in that a bud is defined as four cells or less; this is due to the desire for buds to be distinguished from poorly differentiated clusters (PDC)s, which had been defined as five cells or more in later studies (Ueno et al. 2012; Barresi et al. 2014). Statement 5 determines that buds should be counted on H&E-stained tissue only. This statement is the only one on which the panel did not reach a complete consensus, with only 86% agreement that it should be scored solely in H&E. The conclusion was made that even though IHC is superior to H&E for reproducibility and inter-observer agreement, budding should be scored in H&E as the majority of outcome data to that point had been based on H&E slides, and that the cost-effectiveness of H&E would allow access worldwide. However, they do acknowledge that IHC can be necessary in challenging cases. Statement 7 defines a slight alteration to how the region's size for scoring is determined. Rather than the 20x HPFs used in the Ueno method, they define that budding be scored in one "hotspot". A hotspot should be allocated the same area, 0.785mm^2, as the Ueno method, where the change is that the region to be examined is determined by area versus objective lens size. This helps overcome the varying size of the field of vision of different microscopes. A conversion table was also published to assist with this, and it can be used to calculate a normalisation factor to adjust the bud score based on eyepiece diameter. Statement 8 determines that buds be scored in one hotspot representing the densest area of buds after scanning ten hotspots along the invasive margin of the tumour. The decision to ignore the arguments of other studies on the improved reproducibility of counting in 10 HPFs was due to the concern that scoring in multiple fields would dilute the final bud count on slides that had a locally dense region of budding and that the single most dense hotspot would better reflect the maximal extent of budding in the invasive margin. Additionally, most outcome data had based their studies on the original Ueno method that used one HPF. Figure 2.4 a) provides an example of 10 hotspots along an invasive margin, and b) displays the hotspot representing the highest budding region. Finally, statement 9 defines a three-tier system for reporting bud scores. The tiers are divided into the following categories: BD1 is low, BD2 is medium, and BD3 is high. Where 0-5 buds are considered low budding, 5-9 are medium, and 10 or greater are high. A two-tier, low/high, and continuous scale was also considered. However, the three-tier system allowed better stratification of patients for pT1 and stage II colorectal cancer, and cut-offs were more useful in a clinical setting (Lugli et al. 2017). However, the panel also recommend that the absolute bud count be reported in addition to the bud score to avoid information loss. Based on the consensus statements, a standardised, evidence-based method of tumour bud scoring was outlined for future inclusion in standard reporting. This procedure is shown in algorithm 1.

Since its publication in 2017, the ITBCC guidelines for tumour bud scoring have been used and validated in several studies (Studer et al. 2021; Dawson et al. 2019c). The ITBCC scoring system has been incorporated as an additional prognostic factor in the latest TNM reference (Brierley et al. 2016) and into the reporting guidelines in the US by the American College of Pathologists (Washington et al. 2017), and here in the UK by the Royal College of Pathologists (Loughrey et al. 2018). Due to several limitations, none have included it as a core reporting item. The main concern for inclusion is still the lack of inter and intra-observer reproducibility in H&E, the resource and expertise-intensive nature of the task (Loughrey et al. 2018). Additionally, there are still consistency issues, as even though the ITBCC scoring system provides a standardised procedure, challenges remain in achieving consistent reporting across different institutions and pathologists (Studer et al. 2021). Given that the limitations of tumour budding no longer have to do with the biological significance and prognostic value but more with the technical aspects of implementing a scoring system coupled with the recent advances in deep learning, there have been calls to implement an automated system for tumour bud scoring (Haddad et al. 2021; Studer et al. 2021). The stage is now set for a computerised method to provide observational reproducibility, vastly reduce the resources required in time, cost and expertise, and provide a consistent analysis and reporting method.

Algorithm 1 ITBCC 2016 Tumour Bud Scoring System from (Lugli et al. 2017).

```
procedure Tumour Bud Scoring
    Step 1: Calculate the normalisation factor.
        $normalisation\_factor \leftarrow$ Obtain the field area for the 20x objective lens of your
        microscope based on the eyepiece field number (FN) diameter and calculate the
        normalisation factor.
    Step 2: Select the slide with the highest tumour budding density.
    Step 3: Scan the invasive margin to identify the scoring hotspot:
        $hotspot \leftarrow$ NULL, $max\_buds \leftarrow$ 0
        for $i = 1$ to $10$ do
            $field\_of\_view \leftarrow$ select medium power field (10x objective)
            $buds \leftarrow$ identify and count tumour buds in $field\_of\_view$
            if $buds > max\_buds$ then
                $hotspot \leftarrow field\_of\_view$, $max\_buds \leftarrow buds$
            end if
        end for
    Step 4: Count the number of buds in the hotspot:
        $high\_res\_hotspot \leftarrow$ obtain high power field (20x/$0.785mm^2$) of $hotspot$
        $buds \leftarrow$ identify and count tumour buds in $high\_res\_hotspot$
        $normalised\_buds \leftarrow buds * normalisation\_factor$
        if $normalised\_buds \leq 5$ then
            $score \leftarrow BD1(Low)$
        else if $5 < normalised\_buds \leq 10$ then
            $score \leftarrow BD2(Medium)$
        else
            $score \leftarrow BD3(High)$
        end if
        return $score, normalised\_buds$
end procedure
```

2.5 Stain Normalisation

There can be significant variation in tissue staining for classical and digital histopathology. These variations can arise from the staining process, such as differences in the stain preparation between batches, the amount of stain, the depth of the tissue section and the protocol itself (Ehteshami-Bejnordi et al. 2014). In digital pathology, stain colour variation can also arise from differences in illumination, focusing, camera resolution, magnification, how long after preparation the tissue is scanned and scanner model (Jose et al. 2021). These variations in stain intensity and colour can affect the manual and digital analysis of histopathology slides. It has been shown that standardising the stain colour and appearance facilitates accurate manual interpretation and diagnosis and can reduce intra and inter-observer variability (Michielli et al. 2022). Research has also demonstrated that staining variation can significantly impact the performance of digital analysis and deep learning methods, rendering them ineffective when trained and applied to data from different institutions. (Boschman et al. 2022). The variations affect deep learning techniques by impacting the texture and contrast of stained tissue, which are important features that AI algorithms focus on when making predictions (Vasiljević et al. 2020). An example illustrating selected colour and intensity variations is shown in figure 2.5; whole slide images are visible in panels a) through d), and magnified sections of their tissue are shown in panels e) through h). To address the impact of staining variation, various approaches have been developed and proposed, such as statistical matching of source and target pixel values, colour deconvolution based on Beer-Lambert's Law, and more recently GAN based style transfer techniques (Janowczyk et al. 2017; Shaban et al. 2018). This is an ongoing area of research and inquiry. Still, as discussed later in this work, GANs have a pernicious tendency to invent image features unless careful measures are taken to prevent this. Therefore, this research utilised and evaluated the more structurally robust and well-known approaches to stain normalisation, the Reinhard, Macenko and Vahadane methods. Their theory and technique, as well as strengths and limitations, will be discussed in this section.

The simplest colour normalisation technique, both in implementation and computational cost, is the method proposed by Reinhard et al. (2001). They proposed a method of colour transfer between images through statistical matching of each channel's mean and standard deviation in a decorrelated colour space. A decorrelated space is important as many methods of colour representation, such as RGB, are highly correlated where if the pixel value in one channel is high, the others tend to be similarly large in value (Reinhard et al. 2001), which is unsuitable for statistical matching. Reinhard et al. (2001) evaluated several colour spaces and chose one proposed by Ruderman et al. (1998), termed the $l\alpha\beta$ colour space. This space represents the RGB values by converting them to XYZ tristimulus values and then to a Long Medium Short (LMS) cone space, which represents colours in the human visual spectrum and separates lightness from chromaticity (colour) (Ruderman et al. 1998). The LMS axes are then decorrelated through a transform based on principal component analysis. This results in three orthogonal axes: l, α, and β. Where l represents the achromatic values in the image, the structure. α represents the chromatic values for the yellow-blue and β is the chromatic values for red-green (Ruderman et al. 1998). This decorrelation allows the channels to be manipulated independently and a source image statistically matched to a target.

The Reinhard et al. (2001) technique for colour transfer is as follows: first, the source image is standardised, the mean of each channel is subtracted, and then each channel is divided by its standard deviation. This results in an image with a mean of 0 and a standard deviation of 1. Then, the source image is scaled by the standard deviations of the target image channels, and finally, the target image channel means are added. The resulting image has the same colour statistics as the target, which can be converted back to the RGB colour space for further processing. Equations

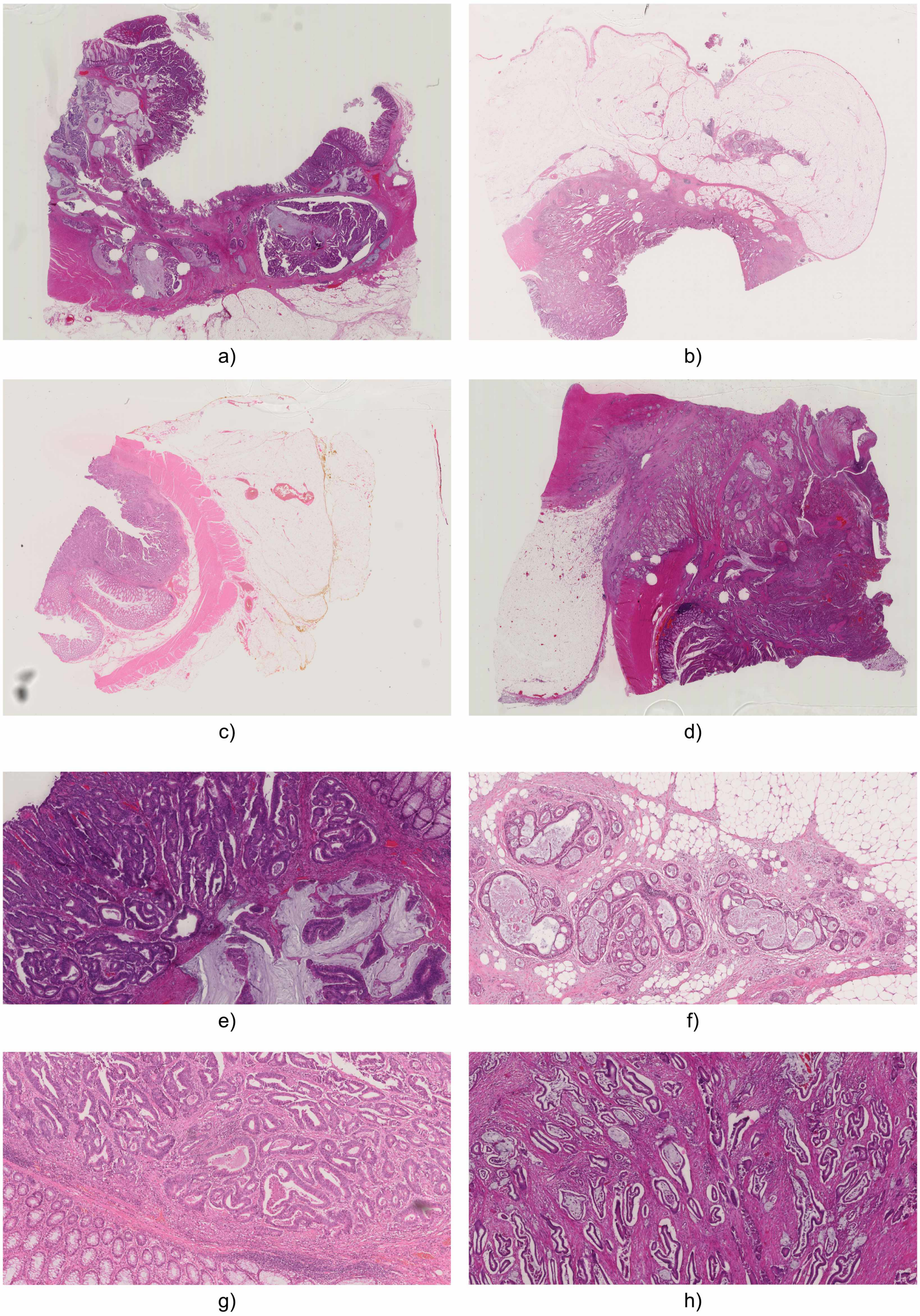


Figure 2.5: Illustration of the diverse colour and intensity variations in digitized stained tissue images. Panels a-d) present whole slide images, while panels e-h) offer magnified views of selected regions from these slides.

2.1, 2.2, and 2.3 describe how this technique is applied at the pixel level.

When applied to stain normalisation, there are limitations and potential issues with the Reinhard colour transfer method (Roy et al. 2018). Linear normalisations applied to all channel pixels in a colourspace can result in the background areas discolouring (Magee et al. 2009). An example of whole slide images and enlarged tissue regions is available in figure 2.6. Note in panels b), c), e) and g) that the background hue has been incorrectly altered due to significant differences between the source and target tissue and the linear transformation of the pixel colours at a global level. The original paper by Reinhard et al. (2001) suggested segmentation and application to regions of different colours or textures. However, in digital histopathology, this can introduce artificial boundary artefacts and intensity disparities that would affect manual or digital analysis (Roy et al. 2018).

$$l_{normalised} = (\frac{l_{source} - \bar{l}_{source}}{\hat{l}_{source}}) * \hat{l}_{target} + \bar{l}_{target} \tag{2.1}$$

$$\alpha_{normalised} = (\frac{\alpha_{source} - \bar{\alpha}_{source}}{\hat{\alpha}_{source}}) * \hat{\alpha}_{target} + \bar{\alpha}_{target} \tag{2.2}$$

$$\beta_{normalised} = (\frac{\beta_{source} - \bar{\beta}_{source}}{\hat{\beta}_{source}}) * \hat{\beta}_{target} + \bar{\beta}_{target} \tag{2.3}$$

Another commonly used stain normalisation technique is the method developed by Macenko et al. (2009). Unlike the Reinhard Colour Transfer technique, it is designed explicitly for histological image analysis to address colour variations arising from differences in the staining protocol. The Macenko method assumes that stain concentration can be determined by the amount of light absorbed by the tissue it passes through (Roy et al. 2018). It is a method for automatically determining the optimal stain vectors that represent how each stain present in a tissue sample absorbs light, and these can be substituted for those of a target sample for normalisation.

Using the Beer-Lambert law, the Macenko method first converts the RGB values into the optical density (OD) space. The Beer-Lambert law states that the optical density of a stained tissue sample is linearly proportional to the concentration of the stain and combines additively for multiple stains OD = $-log_{10}(I)$, where I is the RGB intensity. (Swinehart 1962). Then, Singular Value Decomposition (SVD) is performed to decompose the matrix of RGB OD values into orthogonal components that capture the directions of maximum variance (Macenko et al. 2009). The top two components with the most variance correspond to the two stain vector directions. This is because staining accounts for the primary source of colour variance in histology images (Roy et al. 2018). The SVD stain vector directions then provide an orthogonal basis to define a 2D plane representing the stain colour. The OD pixel values are projected onto this plane to align the pixels with the stain vectors to enable normalisation (Macenko et al. 2009). Once the idealised stain matrix is retrieved for a slide, stain separation can be performed to represent the pixels in a stain colour space. To normalise a source image to match the colour profile of a target, the Macenko method then builds a histogram of stain intensity values for each stain, finds the 99th percentile intensity

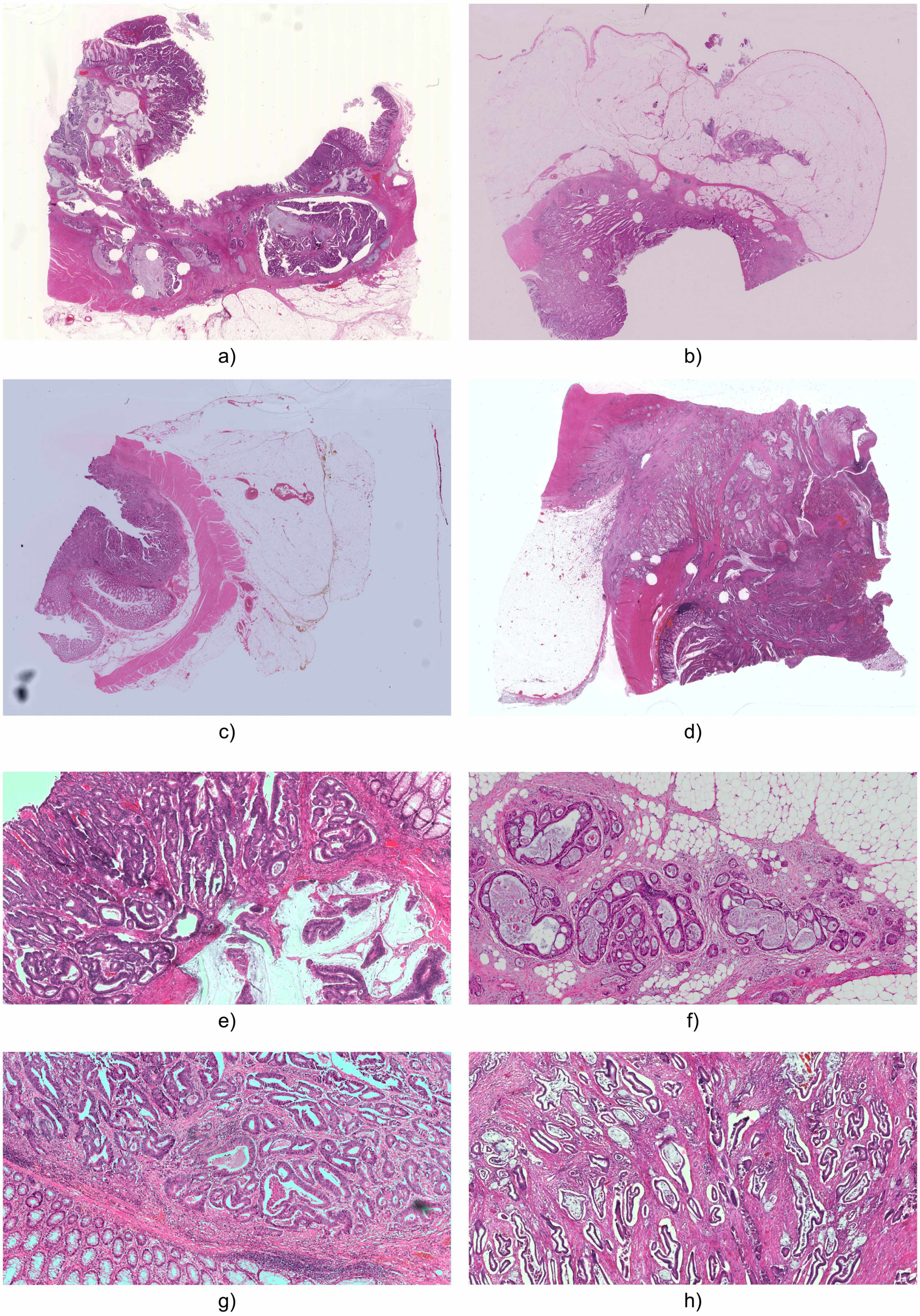


Figure 2.6: Reinhard colour transfer applied to stain normalization. Panels a-d) display whole slide images, while panels e-h) provide magnified views of specific regions from these slides. Note that while the tissue colours have been normalised, the slide backgrounds exhibit aberrant alterations, especially evident in panels b), c), e), and g)

values and uses these to scale the source image to match the target intensity in each stain channel. These rescaled values can be converted to RGB for further analysis and processing.

The Macenko method has some fundamental limitations, however. It is only designed to work with samples that have two stains. It isn't easy to extend it further, although there have been attempts to do this (Niethammer et al. 2010). Another limitation of the Macenko method is that it assumes that all stains will be present in a sample image and stain estimation will fail if sampled from an area with sparse representation of tissue types or large areas of background (Onder et al. 2014). It is computationally more complex than the Reinhard method but much less demanding than other non-negative matrix factorisation algorithms. Given the nature of the technique, it solves the issue with the Reinhard method, where background artefacts are introduced, as long as there is an adequate representation of both stains in the sampled pixels. An example of whole slide images normalised using the Macenko method is given in figure 2.7; note the significant reduction in aberrant background staining. However, in panels b) and c), a slight offset in the background colour can still be observed. If the RGB pixel values are slightly off-white, this method assumes that it results from stain absorption and assigns it to one specific stain colour. Therefore, a small amount of that stain colour will be introduced during normalisation. Note a similar effect in panel c), where black background debris has been assigned to haematoxylin and coloured purple during normalisation. Therefore, the Macenko method has an advantage over the Reinhard method because it will not globally shift the background. Still, it has other inherent limitations in that it will force the assignment of stain colour to all structures and shadows in the image, during which information loss occurs, which is sub-optimal for manual and, in particular, automated analysis.

The final stain normalisation technique we shall discuss is the Structure Preserving Colour Normalisation method by Vahadane et al. (2016). This sparse non-negative matrix factorisation (SNMF) approach models the stain intensities as sparse and non-negative in optical density space. They assume the matrices are sparse as the authors argue that this is a biologically motivated way to model the data, where a pixel should exhibit only one stain colour exclusively. The matrix values are modelled as non-negative as in brightfield RGB images converted to optical density space. The values will always be positive, as the tissue will only absorb light and not emit it (Roy et al. 2018).

The Vahadane method works as follows. First, the RGB image is converted to OD using the Beer-Lambert law as in the Macenko method above. Then, stain deconvolution is approached as a non-negative matrix factorisation problem with a sparseness constraint. They alternate between using sparse coding to estimate a stain density map H and dictionary learning to estimate a stain colour vector W. They then sort the columns of W so that the first represents haematoxylin and the second is eosin. They then scale the source density map H by a robustly chosen maximum value to match the source and target image dynamic range. Finally, they substitute the target images stain colour vector W, convert the intensity map to OD, and back to

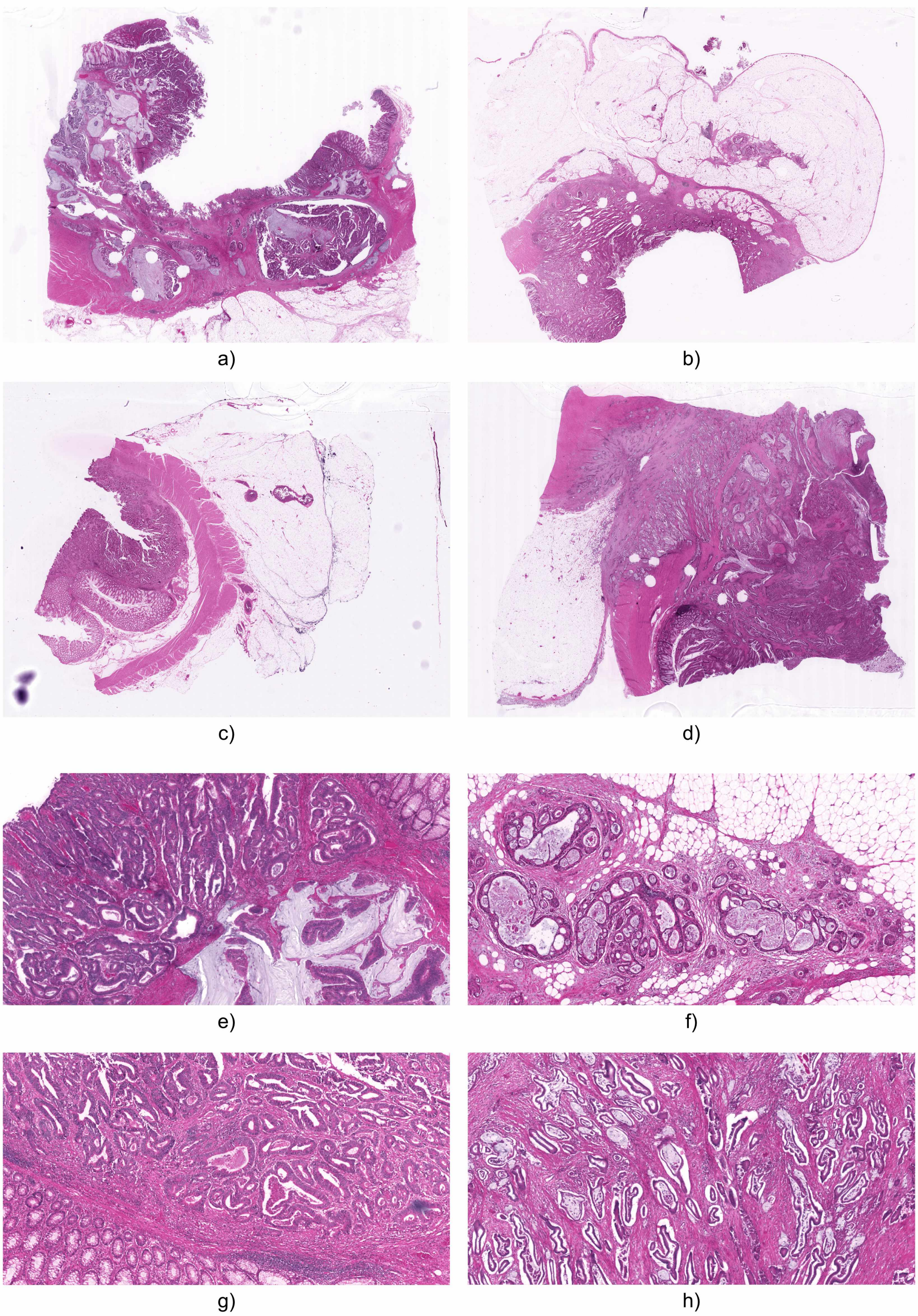


Figure 2.7: Demonstration of the Macenko stain normalisation technique. Panels a-d) display whole slide images, while panels e-h) provide magnified views of specific regions from these slides. It's worth noting that all samples exhibit accurately normalised tissue colours with minimised background artefacts.

RGB. This normalises colour while preserving structure because the stain intensities are scaled globally (Vahadane et al. 2016).

The Vahadane method has several new limitations. Because it relies on iterative optimisation techniques like LARs and coordinate descent algorithms to solve the SNMF problem, it can be very slow compared to closed-form solutions like SVD (Roy et al. 2018). However, the authors acknowledge this limitation and propose a patch-based sampling technique for whole slide images to estimate the stain matrix W on a small sample of pixels, which can then be used to compute H for the entire slide, resulting in a 20x speedup (Vahadane et al. 2016). However, it is still very computationally expensive, and this patch-based stain matrix estimation can potentially miss some tissue regions or inadvertently focus on artefacts, resulting in an inaccurate stain matrix. Another limitation is that the sparseness constraint may over-regularise and lose information when stains are colocalised within the same pixels. An example of the Vahadane structure preserving stain normalisation method is available in figure 2.8. Note the reduction in aberrant background discolouration compared to the Reinhard method. Still, it has the same limitation as the Macenko method, where all structure or shadow is assigned to a stain colour. A slight pink hue can be observed in the slide backgrounds, which have been assigned to the eosin channel in figure 2.8 panels b) and c).

In histology, there is variability in tissue staining introduced by differences in stain preparation between batches or institutions, differing protocols and techniques, and differences in illumination, scanner and camera type. These variations make it harder to compare slides stained under different conditions, and many computational pathology techniques rely on colour and stain intensity, which can affect the algorithm's performance. Stain normalisation is a tool to standardise the stain colour and appearance for robust manual or automated analysis. There are a variety of stain normalisation approaches (Roy et al. 2018), but we have reviewed three of the most common, the Reinhard et al. (2001), Macenko et al. (2009), and Vahadane et al. (2016) methods. The Reinhard method does not perform stain deconvolution. It simply matches histograms in a decorrelated colour space. It is effective but introduces background artefacts. The Macenko method addresses these limitations but introduces background artefacts of its own, but to a lesser degree at a whole slide level. It expects two stains to be present for stain matrix estimation to succeed, which can cause issues at a patch level. Still, it has a computational advantage when compared to SNMF methods. The Vahadane method incorporates domain knowledge to enforce non-negativity and sparsity, resulting in more interpretable stain intensity maps and better structure preservation than the Reinhard or Macenko methods. Still, it also suffers from the need for at least two stains present for estimation to succeed and is much more computationally intense than other methods. Each method has its advantages and disadvantages; no method dominates across all criteria, and the choice of technique must be carefully considered depending on the histopathological task and method of analysis.

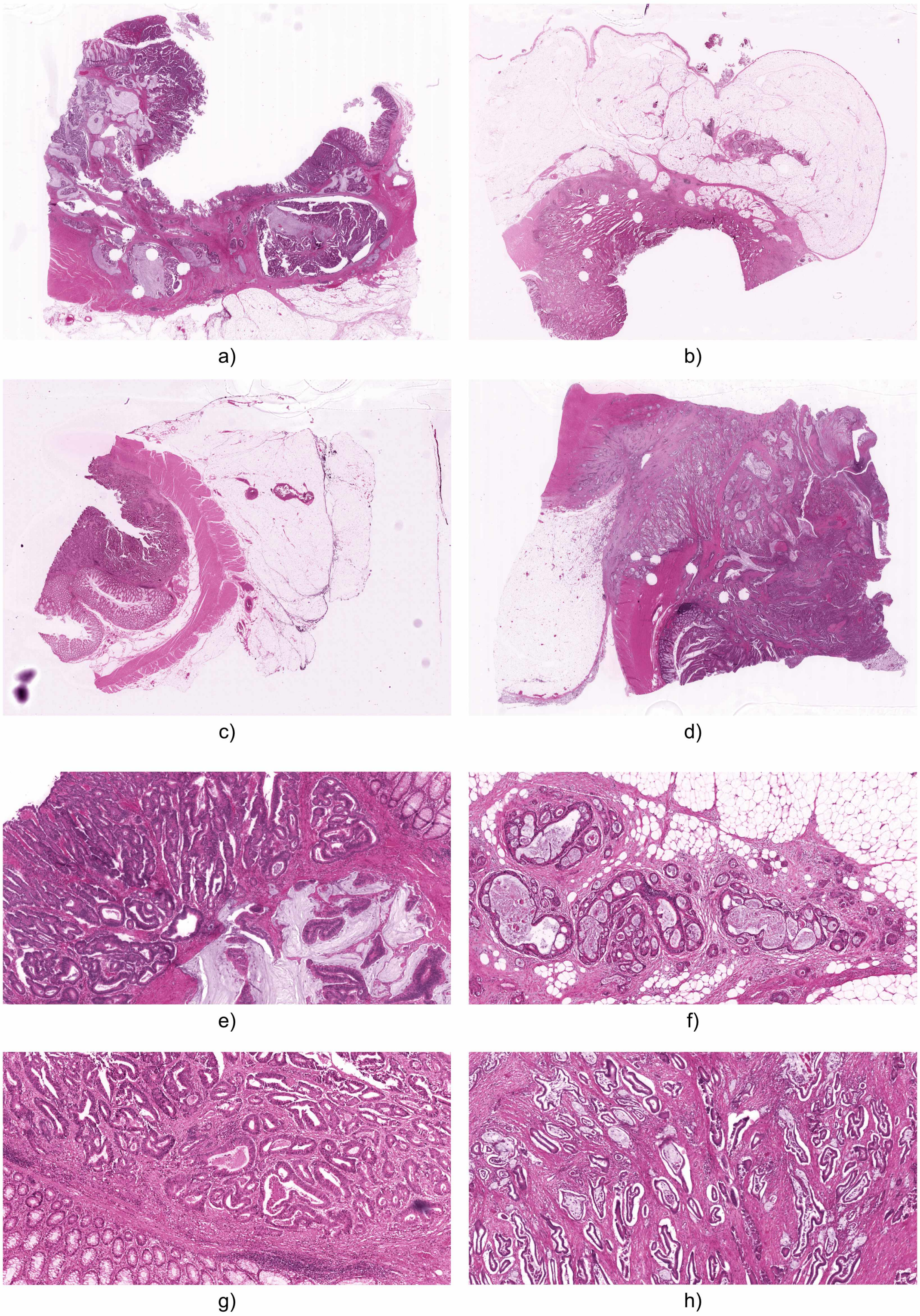


Figure 2.8: Showcase of the Vahadane structure-preserving stain normalisation technique. Panels a-d) display whole slide images, while panels e-h) provide magnified views of specific regions from these slides. It's worth highlighting that all samples exhibit accurately normalised tissue colours with minimised background artefacts.

2.6 WSI Registration and Alignment

Whole slide image registration and alignment is a crucial task in digital pathology. It involves aligning multiple images of the same tissue across serial tissue sections, images taken at different times or using other staining modalities (Zitová et al. 2003). Image registration aims to align two or more images to enable accurate comparison and analysis (Mueller et al. 2011). A vast quantity of medical image registration literature focuses on radiology images. However, as various staining and multiplex techniques are commonly used in histopathology, registration and alignment are increasingly relevant in brightfield and fluorescence microscopy. This section will focus on the theory behind the rigid, affine and deformable registration techniques used in this work and their application to register and align whole slide histopathology images with different staining modalities such as H&E and IHC. Finally, we will discuss recent and relevant advancements in the field.

Several image registration methods are commonly used: intensity-based methods, like the sum of squared differences, cross-correlation and mutual information. There are also feature-based methods, like scale-invariant feature transforms (SIFT), speeded-up robust features (SURF), or maximally stable extremal regions (MSER), as well as a mixture of both (Borovec et al. 2018). Intensity-based methods rely on similar pixel intensities between images to align them. The elementary versions of this approach will fail across different staining modalities, as the staining variation causes vast differences in pixel values (Wang et al. 2022). However, intensity-based methods are computationally efficient and can be applied to various imaging types (Gorbunova et al. 2010). A more sophisticated class of intensity-based techniques is built on ascertaining the amount of mutual information between images to establish their alignment. These work by measuring the statistical dependence or information redundancy between the intensity distributions of corresponding pixels in two images to be aligned (Pluim et al. 2003). The mutual information between the images is highest when the corresponding pixels of the image are correlated, so maximising the mutual information helps find the correct spatial alignment.

A widely used mutual information image registration technique is proposed by Mattes et al. (2001). It is an intensity-based registration method that uses mutual information as a similarity criterion; it calculates the amount of mutual information based on the joint probability distribution of the intensities of a sampled set of corresponding binned pixels in the target images. This is computed using Parzen windowing with cubic b-splines, which allows for a continuous differentiable probability distribution, a key component for the gradient descent optimisation algorithm to compute the required updates for the transform parameters to maximise the joint probabilities (Mattes et al. 2001). This type of mutual information method is much less sensitive to changes in intensity between staining modalities such as H&E and IHC. It provides a much more robust and effective approach for aligning slides of these types.

Intensity-based image registration methods perform better when the tissue is closely aligned before more complex deformable transformations take place (Gatenbee et al. 2021). Therefore, for whole slide images, hierarchical alignment is often used, where a simple rigid alignment is computed at a reduced resolution and transferred to a higher magnification to get the tissue into an approximately correct location to improve the results of more complex transformations (Chen et al. 2021c).

There are a variety of possible pixel transformation functions for digital histopathology images. Rigid transformations align and match tissue using non-deformable algorithms, which include translation, rotation and scaling, without considering local deformations or distortions (Chen et al. 2021c). There are more deformable affine transforms that, in addition to translation, rotation and scaling, also skew the image (Klein et al. 2010). However, affine registration methods may not be sufficient to align images with signification distortions or localised deformations that can occur during the fixation and embedding process (Rueckert et al. 1999). Fully deformable registration techniques allow for more flexible and non-linear transformations to align tissue that has undergone stretching, tearing and compression (Rueckert et al. 1999). There are several deformable techniques. The two tried and tested in this work were b-spline and displacement-field transforms.

B-spline transforms are a nonrigid deformable method of image registration. They are popular in medical imaging because they can model complex tissue deformations and provide smooth transformations (Borovec et al. 2020). B-spline transforms are based on the concept of B-splines, which are piecewise polynomial functions defined on a grid, which can be used to describe the deformation of an image by specifying the movement of control points (Delmon et al. 2013). Displacement field transforms are a non-rigid registration technique used in medical imaging. They model the local motion and deformation of image pixels using a grid of control points. Each control point is associated with a displacement vector representing how the point should be displaced to match the corresponding region in the target image. By iteratively adjusting the positions of the control points by updating the values of the displacement vectors, the transform can be optimised to align the moving and fixed images as an emergent property of the collective movement of the individual points. This is a critical difference between displacement field transforms and b-spline transforms. B-spline transforms use functions defined on a gride to describe the deformation and warping, while the distortion results from the collective movement of control points in the displacement field transform. B-spline transforms provide smooth warps and can model complex deformations. However, they may require more control points to represent the tissue deformation accurately. In contrast, displacement field transforms offer more flexibility in capturing local deformations as the points are independent, and the transforms are also computationally more efficient.

There have been several attempts to produce frameworks for medical image registration in radiology and histopathology. Five popular frameworks are publicly available: Elastix, Voxelmorph, MIND, VALIS and SimpleITK/ITK. All are open-source software packages with modular designs to allow the implementation of various transforms and similarity metrics. They all support multiple image modalities. Elastix, developed by Klein et al. (2010), was one of the first and offers traditional intensity-based similarity metrics. It's one of the most fully featured, offering a wide range of registration similarity metrics and optimisation options. MIND by Heinrich et al. (2012) is a framework designed to be modality independent and provides a method of capturing local image structures based on the distances from proprietary patch-based image descriptors; this allows it to operate reliably across multiple imaging types. ITK and SimpleITK by McCormick et al. (2014) and Lowekamp et al. (2013) are a diverse set of medical image analysis tools complete with extensive image registration and alignment toolset, complete with intensity and mutual information similarity metrics, and a full set of rigid, affine and deformable transforms. It is written in C++ and is highly efficient and capable of dealing with huge whole-slide images. The SimpleITK wrapper provides a Python interface that allows a high-level interface to versatile, low-level and extensible C++ code. The last two frameworks are more modern and feature deep learning approaches. Voxelmorph by Balakrishnan et al. (2019) uses convolutional neural networks to learn the registration transforms between the fixed and moving images and focuses on efficiency and speed through GPU use. It can also use cellular segmentations as auxiliary data for slide alignment (Balakrishnan et al. 2019). The last framework, VALIS by Gatenbee et al. (2021), is designed to be an open-source pipeline for the registration of histology slides. It can register any number of unsorted slides or regions of interest. It is a hierarchical registration framework that first performs a rigid whole-slide alignment using a feature-based similarity metric, then can perform further deformable transforms using feature extractors or neural networks to register landmarks and perform displacement-field transforms. In summary, while all five tools share common goals of flexible and accurate medical image registration, they have different formulations, applications, and design philosophies. Elastix and SimpleITK/ITK focus on traditional optimisation, VoxelMorph on deep learning and speed, MIND on a novel modality-invariant descriptor, and VALIS on an automated pipeline optimised for histology data.

In digital pathology, aligning and registering whole slide images is crucial for creating aligned datasets for accurate comparison and analysis. Various registration techniques, like rigid, affine and deformable transforms, are available that can be used to correct a degree of possible tissue deformations. Key to this process is the similarity metric, which can be intensity, mutual information or feature-based. The similarity metric type must be carefully measured for the task, and issues like intensity variation across staining modalities must be considered. Several frameworks exist for this in histopathology, including fully-fledged pipelines like VALIS and VoxelMorph. However, these pipelines are designed for specific use cases; further validation is required before they can be put into general use. Image registration and alignment is a powerful tool for dataset creation and analysis, but selecting the proper method for a given dataset and task is worth careful consideration.

2.7 Deep Learning

Modern deep learning can trace its origins to the quest to understand and replicate the human brain's processing and pattern recognition abilities. It has transformed numerous domains from computer vision to natural language processing, now boasting capabilities that approach and, in some cases, surpass human performance. This section will delve into the theory required for deep learning in histopathology; we will introduce the deep neural network architectures used in this work, including convolutional neural networks, U-Nets, GANs, CycleGANs, self-supervised networks, and transformers. We shall discuss the theory, strengths and limitations of each, their use cases, and specific techniques surrounding their use and training, like transfer learning and clustering based on representation learning.

2.7.1 Artificial Neural Networks

Deep learning stems from Artificial Neural Networks (ANNs). An ANN is a mathematical model that attempts to simulate the structure and function of biological neurons (Krenker et al. 2011). They consist of densely connected layers of artificial neurons, where every neuron maps a set of inputs to an output (Priddy et al. 2005). The inputs of one layer are directly connected to the next layer, and information flows in one direction from the input to the output; this type of network is commonly known as a Feedforward network (Wang et al. 2017). ANNs take advantage of the complexity that can manifest from the collective action of a few simple rules (Krenker et al. 2011).

The fundamental building blocks of classical ANNs are artificial neurons, modelled after the neurons of the human brain. The modern versions are the result of a combination of work by McCulloch et al. (1943), Rosenblatt (1958), Minsky et al. (2017). The concept was introduced in 1943 when Warren McCulloch and Walter Pitts proposed a simplified neural model in their seminal paper "A Logical Calculus of the Ideas Immanent in Nervous Activity". This work centred around demonstrating the theory that networks of artificial neurons could, in principle, compute any logical function (McCulloch et al. 1943). Then, in 1949, a mechanism was proposed for learning in biological neurons (Samarasinghe 2006). In his book, The Organisation of Behaviour, Hebb (1949) defined a method by which the weights between neurons are updated in a learning environment. This became known as Hebbian Learning (Samarasinghe 2006). The first implementation of an artificial neuron was proposed by Rosenblatt (1958) in his work, titled "The Perceptron". Rosenblatt introduced the idea of an electronic circuit with weights associated with the inputs that were summed to produce an activation, like the McCulloch-Pitts neural model, that was updatable through association in a learning environment as proposed in the Hebbian Learning method. He discussed the possibility of updating the weights through a type of backpropagation but did not propose an implementation. The following critical piece of the modern artificial neuron was introduced by Minsky et al. (2017) in 1969. They criticised the lack of the perceptron to solve problems that were not linearly separable. They suggested that multi-layer perceptrons with non-linear activation functions would allow perceptrons to solve a larger class of problems. Making the network "deeper" by adding more layers allowed the approximation of more complex functions and is the origin of the term "Deep Learning". This was a critical observation, although the impact wouldn't be realised until much later when backpropagation became popular in the late 1980s (Wang et al. 2017). Supplementing the summation junction with an activation function that squashed the output of the neuron to a known continuous range, such as -1 to 1 or 0 to 1, was necessary, as a continuous output was differentiable, which is a fundamental property for backpropagation, which is the final component of a modern ANN.

Backpropagation is a fundamental algorithm for training neural networks. It is a gradient-based optimisation algorithm that minimises the error between the predicted and actual output for a given set of examples (Haykin 2009). It propagates the input data through the network to produce a result. The error (or loss) is then calculated using a loss function defined for the task. Then, in a backward pass, the gradient of the loss concerning each weight in the network is calculated. This step is the essence of backpropagation. Working backwards from output to input, the contribution to the loss for each neuron is calculated based on the error from neurons in the subsequent layers. The gradients are then used to adjust the weights and biases in the directions that reduce the error. The magnitude of the adjustment is determined by the learning rate (Samarasinghe 2006). A frequently used update rule is gradient descent, where the weight is adjusted by subtracting the learning rate multiplied by the gradient. More sophisticated optimisation algorithms like Adam or RMSprop can also be used for weight updates, which help achieve faster and more stable convergence (Kingma et al. 2017). This process is repeated in multiple passes over the training data, called "epochs". This continues until the error reaches an acceptable minimum or stops decreasing.

We now have all the critical components of a modern artificial neural network: synaptic weights and biases associated with inputs, non-linear continuous and differentiable activation functions, and a method of updating the neural network weights to minimise prediction error, known as backpropagation. An example of a modern neuron is shown in figure 2.9 panel a), and an example of a modern artificial neural network, composed of many densely connected neurons with multiple hidden layers, which enable them to learn to approximate complex, non-linear relationships is given in panel b). However, traditional multi-layer perceptrons (MLPs) have significant limitations despite their ability to approximate complex functions. Due to full connectivity, they often have many parameters and are computationally expensive to train. For example, when run on an input image of size 100x100, they would have 1,000,000 parameters in the first layer alone, only growing in subsequent layers. This makes them limited in use for image processing and for creating deeper networks to approximate more complex functions. This brings us to the advancement in modern deep neural networks that address these limitations and sparked the revolution in medical deep learning and AI, in general, convolutional neural networks.

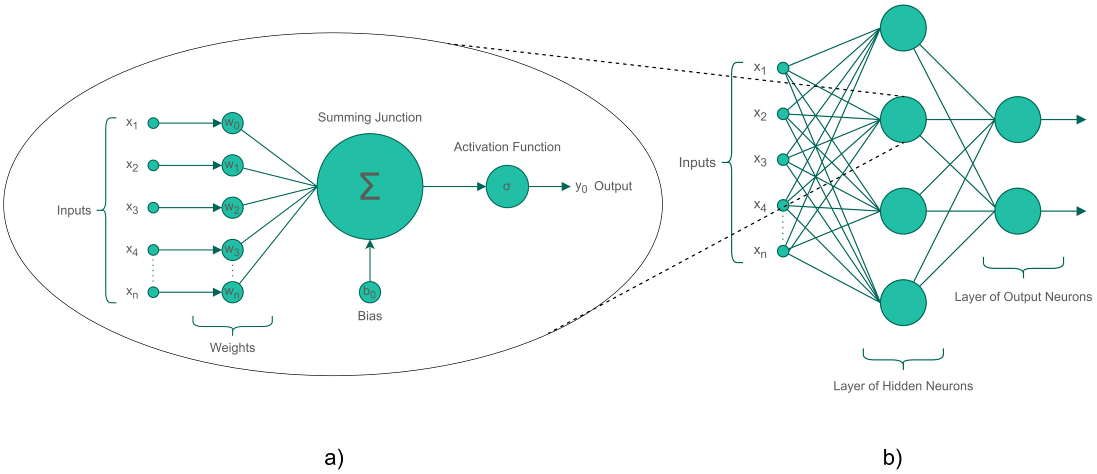


Figure 2.9: Depiction of an Artificial Neuron and Artificial Neural Network. a) The neuron receives a series of inputs, x_1 to x_n, and assigns a series of weights, w_1 to w_n, to each input along with a bias offset. These weighted inputs are subsequently aggregated and processed through a non-linear activation function, σ, culminating in an output, y. b) An ANN consists of numerous neurons. The inputs of one layer are intricately linked to every neuron in the succeeding layer, and the outputs from one layer serve as the inputs for the next.

2.7.2 Convolutional Neural Networks

The origin of the Convolutional Neural Network (CNN) dates back to the early 1960s with the work of Hubel et al. (1962), which introduced the idea of shared receptive fields in a hierarchical framework within the brain's visual system. They suggested that the visual cortex uses simple features in a location-invariant manner to build high-level abstracted representations for feature recognition. The first precursor to the modern CNN was the Noncognitron developed by Fukushima (1980), who was inspired by the hierarchical system proposed by Hubel and Wiesel. It used two basic building blocks "simple cells" and "complex" cells. Simple cells move over the image, detecting specific local patterns in their receptive field and producing a response. The complex cells build up abstract features from the input of many simple cells (Fukushima 1980). It had two landmark properties; the first was invariance because the simple cells could apply their learned weights at any point in the image. It was invariant to small translations, rotations and deformations in the image. The second was a hierarchy, in that the Noncognitron responds to more complex and abstract features with each subsequent layer. While the Noncognitron didn't become a standard computing model, it helped recognise handwritten characters and its fundamental principles served as the foundation for subsequent developments (Al-Hayani et al. 2017). In the early 1990s, the field was advanced significantly with the introduction of multi-layer convolutional neural networks by Lecun et al. (1998). Their architecture was titled LeNet-5, a network intended for handwritten and machine-printed character recognition. It had several notable features; the first was convolutional layers like the simple cells in the earlier works; these shared weights for feature recognition across the image. It also had pooling layers to "pool" information to reduce computational complexity and help group local features. It had a densely connected final layer to allow learning to take place on the output representations, and it was one of the first architectures to emphasise end-to-end learning using back-propagation (Al-Hayani et al. 2017). It set a paradigm shift in motion for the design and understanding of CNNs in that it demonstrated that deep hierarchical models could automatically learn feature representations from raw pixels, providing a change away from traditional handcrafted feature engineering at that point LeCun et al. (2015). It was later employed by the United States Postal Service for automated zip code reading, making it one of the earliest practical applications of CNNs (LeCun et al. 2015). However, after this, CNNs fell away from prominence due to the computational constraints of computers at the time and the popularity of other machine-learning methods. It wasn't until the early 2010s that interest was resurgent, driven primarily by the availability of large, labelled datasets and a dramatic increase in computing power provided by advances in GPU computing (Al-Hayani et al. 2017). Then, in 2012, the work of Krizhevsky et al. (2012) showcased the immense potential of CNNs by dramatically reducing error rates, winning the ImageNet classification competition with a deep convolutional network and subsequently steering the AI community towards deep learning and making CNNs a cornerstone of model computer vision.

A modern convolutional network has four basic building blocks: the convolutional layer, the pooling layer, the activation function, and fully connected layers (Teoh 2023). The convolutional layer is the foundation of the architecture, and it is designed to learn features from input arrays adaptively. Each convolutional layer has many filters whose kernel weights can learn to recognise features through backpropagation, and an output is produced by taking the scalar product of the kernel weights and the region connected to the input volume (O'Shea et al. 2015). The filters are small sets of matrices of kernel weights that slide or "convolve" over the input array data, such as RGB images, to produce a feature map. Using multiple filters, a CNN can learn and extract features like edges, corners, and textures and build up more abstract and complex patterns as we move deeper into the network (Mallat 2016). The strength of convolutional layers stems from their ability to preserve the spatial relationship between pixels, which makes them particularly well suited to tasks where this property is paramount, such as image processing (Sarvamangala et al. 2022). As the network is trained, the weights of these filters are automatically adjusted via backpropagation, allowing the model to optimise feature detection for specific tasks (Al-Hayani et al. 2017). An example demonstrating the operations of a convolutional layer is given in figure 2.10.

Modern CNNs exploit weight sharing to reduce the complexity of the network, speed training, and avoid overfitting to avoid the problem as described with fully connected ANNs. Pooling layers help this aim and gradually build spatial invariance by grouping similar local features (Teoh 2023). The pooling layers are usually positioned to operate on the feature output maps of the convolutional layers and reduce their dimensionality (O'Shea et al. 2015). Pooling layers pass a sliding window over the feature maps and calculate their output; the most common window size occupies a 2x2 region (Al-Hayani et al. 2017). An example of the most common pooling operations, maximum pooling and average pooling, is given in figure 2.11.

As real-world data is primarily non-linear, to approximate it, neural networks must have a method of introducing non-linearity. This is the role of activation functions in deep neural networks (Sarvamangala et al. 2022). They also aid in the process of backpropagation by providing continuous, differentiable (or sub-differentiable, in the case of ReLU) output, which is essential to allow differentiation and the computation of gradients from the error for that neuron or filter during the training process (Al-Hayani et al. 2017). There are three common types of activation functions: the sigmoid, tan-hyperbolic and rectified linear unit (Relu) functions (Sarvamangala et al. 2022). The sigmoid function accepts real-valued numbers and squashes them into the range 0 to 1. It particularly penalises significant negative and positive inputs (Al-Hayani et al. 2017). Its sigmoid equation is shown in 2.4. The hyperbolic tangent function accepts real-valued numbers and squashes them from -1 to 1. This has the advantage that because it ranges between -1 and 1, the mean of the activations tends to be closer to zero, which makes learning easier in subsequent layers. The more extensive range can strengthen the gradient during training (Michelucci 2019). The equation for the tan-hyperbolic activation function is shown in equation 2.5. Finally, the rectified linear unit (ReLU) is an activation function that has become the most prominent for deep learning use in hidden convolutional and fully connected layers. It is mathematically defined in equation 2.6, but it is a simple operation; the function returns the input if the input is greater than or equal to zero and returns zero otherwise (Sarvamangala et al. 2022). This activation function has several advantages over the others. The ReLu function is computationally inexpensive as it only requires simple

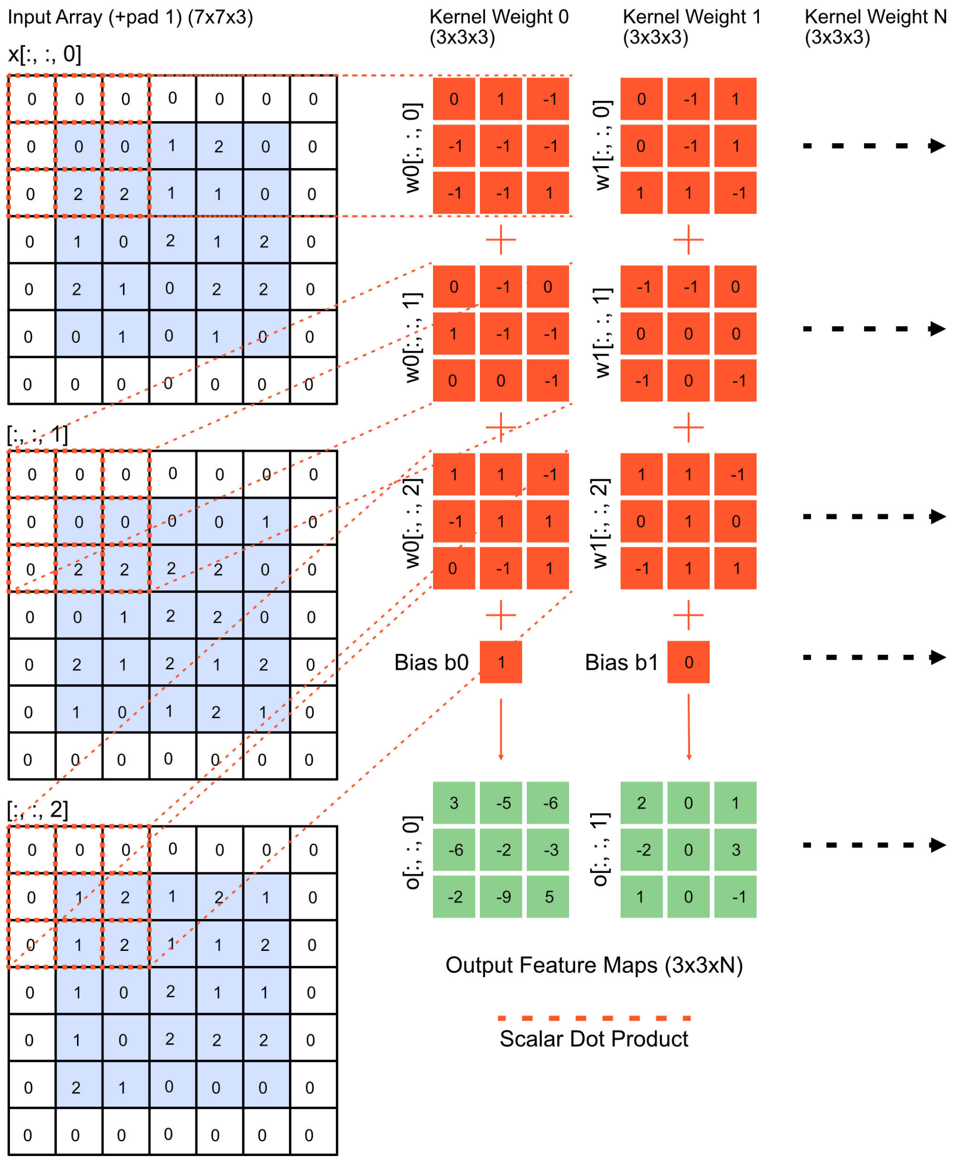


Figure 2.10: Depiction of a convolutional layer applied to a 7x7 RGB image. A convolutional layer operates by applying a collection of filter kernels to the input image. Each kernel methodically “convolves” across the input data (illustrated in blue). This movement is characterised by a stride, indicating the kernel advances by one or multiple pixels in each step. Zero padding is frequently employed (represented in white) to facilitate the kernel’s traversal across all pixels. At every position, the kernel executes a scalar dot product between its weights (portrayed in orange) and the section of the input it currently overlaps. This result is aggregated, and a bias is incorporated, yielding a singular value in the output feature map (depicted in green) for that specific location. If a layer possesses N filters operating on a 7x7 image, it will produce Nx3x3 feature maps as output.

thresholding. This allows the model to train faster and requires less computational resources than the other functions (Sarvamangala et al. 2022). Additionally, the sigmoid and tanh functions have derivatives that tend towards zero when the input is mainly positive or negative. When this happens during backpropagation, the gradient becomes vanishingly small, and the network cannot learn effectively; this is known as the vanishing gradient problem (Al-Hayani et al. 2017). Alternatively, ReLU doesn't saturate or suffer as severely from the vanishing gradient problem (Krizhevsky et al. 2012).

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.4}$$

$$f(x) = \frac{1–e^{-2x}}{1 + e^{-2x}} \tag{2.5}$$

$$f(x) = max(0, x) \tag{2.6}$$

The fully connected layer is the fourth of the most common building blocks for CNNs. These layers are like an artificial neural network (Sarvamangala et al. 2022). Each neuron in the layer is connected to all the inputs from the previous layer, and each connection has a weight associated with it. The fully connected layers are usually at the network's top, close to the output. They are necessary to aggregate and reason about the features detected by the convolutional layers. In these layers, the relationships between the observed high-level features are embedded and used for the output predictions (Teoh 2023).

Now that we have described the standard components of a modern convolutional neural network, we can describe the accepted approach to constructing a network. As described by LeCun et al. (2015) and Sarvamangala et al. (2022), a basic CNN usually consists of the following layers: the input layer, where the input array is fed into the network. That is generally followed by a convolutional layer + ReLU, then a pooling layer. This CNN+ReLu block can be repeated many times in deeper networks. This is where the main features are extracted, and the dimensionality of the input data is reduced to speed computation and build a hierarchy. Finally, one or more fully connected layers will sit at the network's top to aggregate and make predictions from the learned features (Teoh 2023). A basic example of a CNN architecture is provided in 2.12.

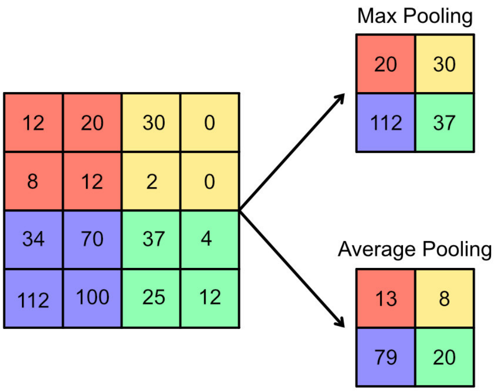


Figure 2.11: Depiction of a pooling layer within a CNN. Pooling layers serve to reduce the spatial dimensions of the input data. This is achieved by applying a specific pooling function to the data, typically within 2x2 windows (represented by distinct colours). The predominant pooling functions include max and average pooling. Max pooling delivers the peak value from the window to the downsampled feature map. Conversely, average pooling calculates the mean value within the window.

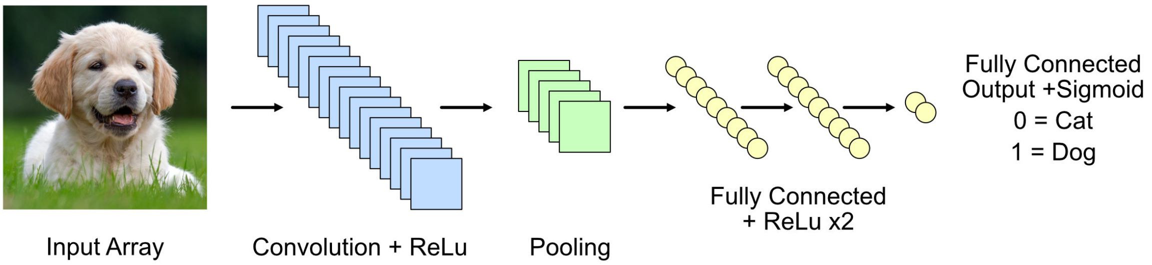


Figure 2.12: Displayed is a representative CNN architecture comprising six layers. The first layer serves as the input, accommodating either 2D or 3D array structures. Following this is the convolutional layer, adept at discerning features within the input array, and a rectified linear unit (ReLU) activation function, follows the convolutional layer. Subsequently, the third layer is a pooling layer, designed to condense the data's dimensionality and group similar spatial features. The second last component group consists of two fully connected layers, which learn to represent the input features, also utilising ReLU activation functions. The architecture concludes with a two-neuron output layer, which, through a sigmoid activation function, provides the probabilities for the predicted output class.

ResNet

Since AlexNet won the ImageNet challenge in 2012, numerous landmark CNN architectures have been released that improve classification performance on the ImageNet dataset. One of these advances, and one of the most notable, is a family of models known as ResNets (Teoh 2023). Initially proposed by He et al. (2015), the architecture introduces the idea of "residual learning", where blocks of convolutional layers are implemented with a skip connection that bypasses them. These are known as "residual blocks" (Sarvamangala et al. 2022). This allows these layers to learn the difference between their input and output, the "residual", rather than learning the underlying mapping directly. He et al. (2015) show that deep networks suffer from the vanishing gradient problem where adding more layers negatively impacts the training error when the gradients tend to zero in deep networks. ResNet architectures overcome this by residual learning, as the skip connections allow improved gradient flow through the network. He et al. (2015) also demonstrated that ResNets could be successfully trained with over 1000 layers, with the release of a family of models with a range of repetitions of the residual blocks from ResNet18 to ResNet1202. These can achieve state-of-the-art performance on ImageNet classification as well as other downstream tasks, including those in medical imaging (Anwar et al. 2018). An example of the foundational building block of residual convolutional networks, the "residual block", is provided in figure 2.13.

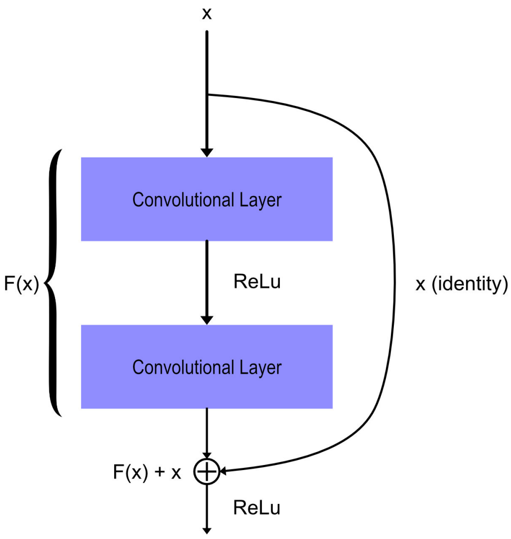


Figure 2.13: An example of a residual block as described by He et al. (2015). The residual block is a fundamental component of ResNet architectures. It is composed of several convolutional layers with ReLU activation functions and a skip connection. The skip connection allows the convolutional layers to learn the residual mapping between the input and output of the block or the minimal correction to the input signal to achieve the desired output. The identity mapping facilitates the flow of gradients through the network, which helps mitigate the vanishing gradient problem as network depth increases.

UNet

The U-Net is another landmark convolutional neural network used extensively in medical image analysis. It was proposed by Ronneberger et al. (2015). Initially designed for biomedical image segmentation, it has become a typical architecture for image generation of all types (Anwar et al. 2018). The architecture consists of a contracting path to capture context, often called the encoder. The encoder follows a typical CNN design and consists of repeated blocks of convolutions, ReLU activations and max pooling for downsampling. At each step, the image width and height are halved, doubling the number of feature channels. This allows the network to capture more contextual information and hierarchical feature representations in deeper layers and the spatial context of the input. The encoder outputs feature representations in what is often called the latent space (Al-Hayani et al. 2017). These are then used in an expansive pathway known as the decoder. This part of the network upsamples the feature maps, doubles the width and height at each stage, and halves the number of feature channels at each step. There are skip connections passing information from the encoder to the decoder that allows the combination of high-resolution features from the encoder to enable precise localisation of features in the generated image. The decoder allows the network to draw from the compressed feature representation and the detailed features in the skip connections to output full-resolution pixel labels for a segmentation or generated image Ronneberger et al. (2015). The U-Net is a fully convolutional encoder-decoder architecture designed to learn the most pertinent image features through compression and expansion. Figure 2.14 shows an example of the U-Net architecture.

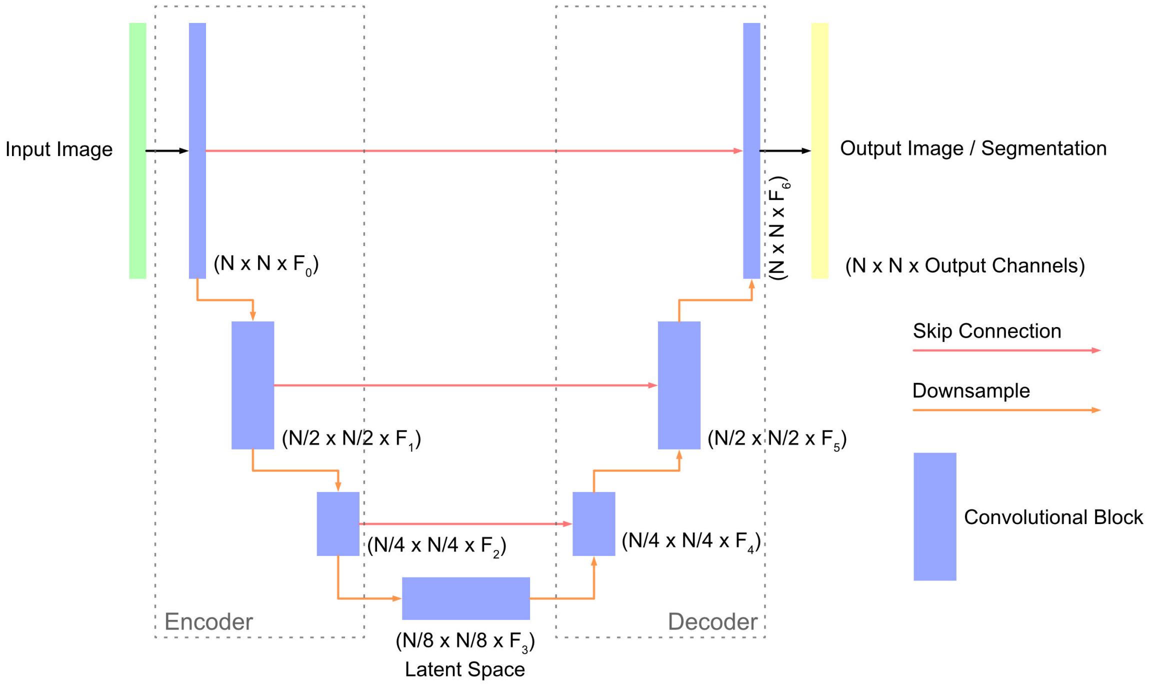


Figure 2.14: An example of a U-Net architecture as described by Ronneberger et al. (2015). This network has a symmetrical U-shape which characterises its design. It comprises a contracting path, typically a convolutional network that sequentially downsamples the input image to extract features known as the encoder. The encoder is followed by an expansive path, a series of convolutional layers that progressively upsample the features to produce a generated image or segmentation known as the decoder. The decoder is also connected to the encoder via skip connections, which allow the network to pass pertinent information across the bridge to facilitate learning. This contraction and expansion help compress the input image's information into a latent space containing the most important features, which can then be drawn from to recreate the output image.

2.7.3 Generative Deep Networks

Generative deep networks are a type of network that focuses on the reproduction of data from a target distribution through unsupervised deep learning. There is a subset of these networks called Generative Adversarial Networks (GANs) that were first proposed by Goodfellow et al. (2014), which ushered in a novel paradigm in the way deep networks could be trained to generate realistic data (Gui et al. 2023). GANs consist of two models, typically neural networks, but they could be any differentiable function approximator. The models are known as the generator, which seeks to produce realistic synthetic data and a discriminator, endeavouring to distinguish between authentic and generated data (Yi et al. 2019). Drawing inspiration from game theory, the objective of this configuration is to have the generated data be convincing enough that the discriminator can no longer distinguish between the generated and authentic data (Jose et al. 2021). As the training progresses, the generator continually refines its output in response to feedback from the discriminator, aiming to perfect its data reproduction. The discriminator constantly strives to perfect its detection of fake data. The adversarial nature of the training drives both networks to improve (Gui et al. 2023). The GAN objective is represented by the expression given in equation 2.7. Where $\mathbb{E}$ denotes the expectation, x are samples from the real data distribution p_{data}, z are random noise samples from the distribution p_z, $D(x)$ is the discriminator's estimate of the probability that real data instance x is genuine, $G(z)$ is the data generated by the generator from noise z, and finally, $D(G(z)$ is the discriminator's estimate of the probability that a fake instance is genuine. Over the last several years, GANs have seen a surge in popularity, particularly in image generation. A diagram of a GAN model configured to use a CNN to generate images, often called a DCGAN, as proposed by Radford et al. (2016), is displayed in figure 2.15.

$$\min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.7}$$

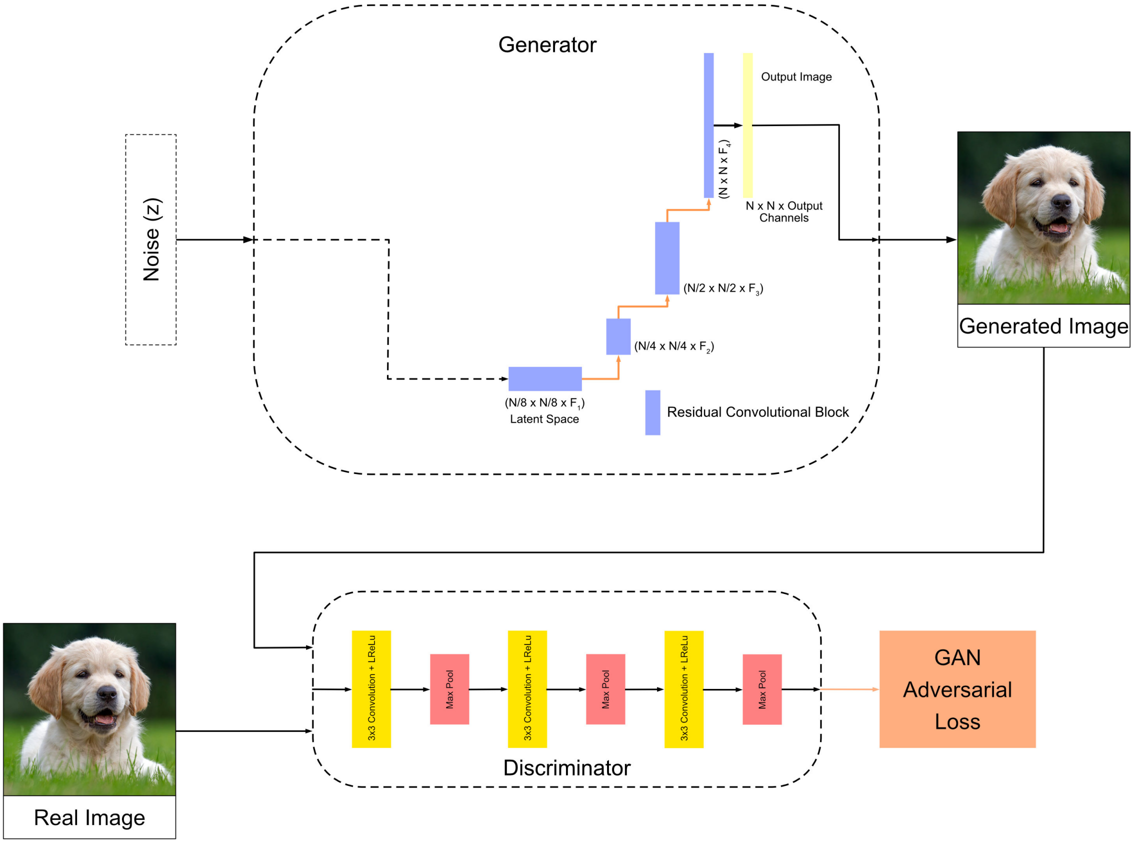


Figure 2.15: Depiction of a GAN architecture as described by Goodfellow et al. (2014), with a CNN generator as described by Radford et al. (2016). The generator network takes a random noise vector as input and produces an image. The discriminator network takes an image as input and outputs a probability that the image is real or fake. The generator and discriminator are trained simultaneously in an adversarial manner. The generator is trained to fool the discriminator by producing realistic images, and the discriminator is trained to distinguish between real and fake images. This process continues until the generator can produce realistic images that fool the discriminator.

Pix2Pix GAN

The Pix2Pix GAN, short for “pixel-to-pixel”, is a specialised conditional generative adversarial network designed for image translation tasks. It is a supervised model proposed by Isola et al. (2018). The Pix2Pix GAN conditions both its generator and discriminator on input images, enabling the translation of images from one domain to another, such as turning sketchiness into coloured photos or black-and-white images into coloured versions (Patgiri et al. 2021). The architecture usually includes a U-Net-based generator, as described in section 2.7.2. The discriminator is different to the standard GAN, where the discriminator tries to classify entire images as real or fake. A novel feature of the Pix2Pix GAN is that the discriminator is modified to accept the source image as a condition and classify patches of the generated image as real or fake. It is described as a PatchGAN discriminator (Isola et al. 2018). Another distinctive feature of the Pix2Pix GAN is that it merges the traditional adversarial GAN loss with an L1 loss. This ensures that the generated images are realistic and similar in content to the target images. However, this configuration now requires paired source and target images for successful training. The updated Pix2Pix GAN adversarial loss expression is given in equation 2.8, where G is the generator, D is the discriminator, x and y are the source and target images, and G(x) is an image output by the generator in the target domain y. The second component of the Pix2Pix objective is the L1 loss term, which is given in equation 2.9. Finally, the full model objective is given in equation 2.10, which is a weighted sum of the adversarial loss and the L1 loss, where λ is a weight factor that determines the importance of the L1 loss relative to the adversarial loss. In the original Pix2Pix paper, λ is set to 100 (Isola et al. 2018).

$$\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{x,y}[\text{log}\, D(x, y)] + \mathbb{E}_{x}[\text{log}(1 - D(x, G(x)))] \tag{2.8}$$

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y}[||y - G(x)||_1] \tag{2.9}$$

$$\mathcal{L}(G, D) = \mathcal{L}_{\text{GAN}}(G, D) + \lambda\mathcal{L}_{L1}(G) \tag{2.10}$$

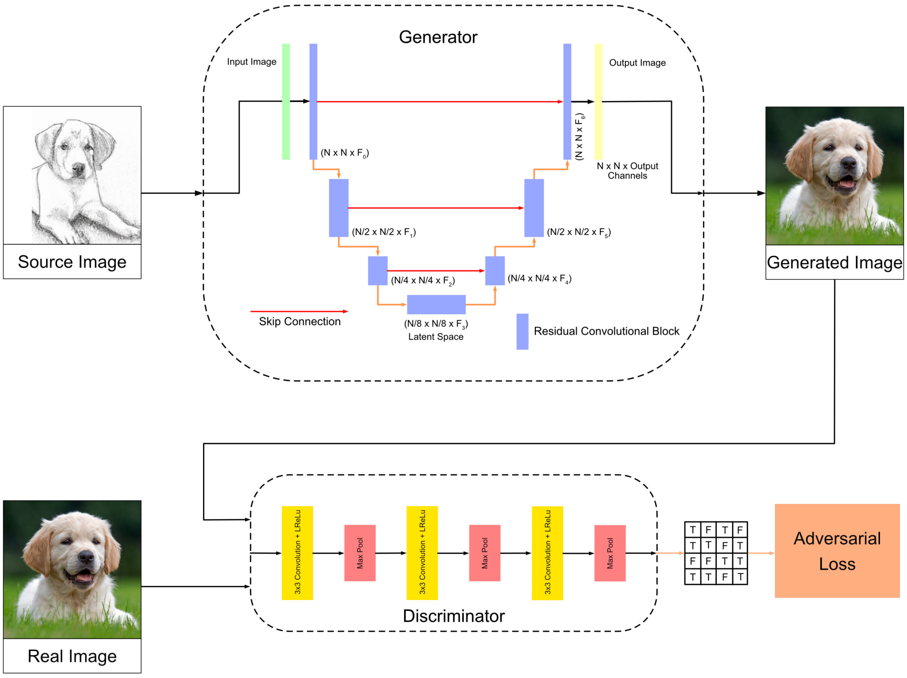


Figure 2.16: An example of a Pix2Pix GAN architecture as described by Isola et al. (2018). The generator network takes an image as input and produces an image as output. The discriminator network takes a pair of images as input and outputs a probability that the image pair is real or fake. The generator and discriminator are trained simultaneously in an adversarial manner. The generator is trained to fool the discriminator by producing realistic images, and the discriminator is trained to distinguish between real and fake images. This process continues until the generator can produce realistic images that fool the discriminator. This allows the generator to learn a mapping between the input and output images, which can be used to translate images from one domain to another, as shown here between a drawing and a photo.

2.7.4 The CycleGAN

The task of style translation has become widespread in many domains, medical imaging included. However, with previous GAN architectures, paired images that are difficult to gather and align were required. The CycleGAN network architecture, as proposed by Zhu et al. (2020), was a breakthrough in the realm of image-to-image style translation, as it can learn to translate between two image domains using unpaired data (Zhu et al. 2019). The architecture consists of two main components: generator-discriminator pairs. The pairs translate from the source to the target domain and back again (Lei et al. 2020). To ensure that the translation process retains the content of the original images, the CycleGAN introduces "cycle consistency" loss as expressed in equation 2.13. After an image is translated from domain X to domain Y and then back to domain X, this loss ensures that the reconstructed image is like the original (Zhu et al. 2020). The PatchGAN discriminators function as in traditional GANs, determining the realism of the generated images. Their adversarial loss objectives are expressed in equations 2.11 and 2.12. Finally, to ensure the colour and tint of the generated images are retained, an identity loss is introduced, where a real image from the target domain is passed through the generator assigned to that domain, and the generated image is constrained to be as similar as possible to the input. This is known as "identity loss" (Xu et al. 2019); it is expressed in equation 2.14. A diagram depicting the configuration of the network and its loss functions is displayed in figure 2.17. The combined loss function is given in equation 2.15, where λ is a value to control the weight of the cycle and identity losses relative to the adversarial loss. The full objective of the network is given in equation 2.16. This goal of generating realistic images while maintaining their structure and colour through the cycle is a powerful combination that results in an unsupervised learning method to translate between style domains without requiring expensive paired data.

$$\begin{aligned}\mathcal{L}_{G_Y}(G_Y, D_Y, X, Y) = {} & \mathbb{E}_{y\ p_{data}(y)}[logD_y(y)] \\ & + \mathbb{E}_{x\ p_{data}(x)}[log(1 - D_Y(G_Y(x)))]\end{aligned} \tag{2.11}$$

$$\begin{aligned}\mathcal{L}_{G_X}(G_X, D_X, X, Y) = {} & \mathbb{E}_{x\ p_{data}(x)}[logD_X(x)] \\ & + \mathbb{E}_{y\ p_{data}(y)}[log(1 - D_X(G_X(y)))]\end{aligned} \tag{2.12}$$

$$\begin{aligned}\mathcal{L}_{cyc}(G_X, G_Y, X, Y) = {} & \mathbb{E}_{x\ p_{data}(x)}[\|G_X(G_Y(x)) - x\|_2] \\ & + \mathbb{E}_{y\ p_{data}(y)}[\|G_Y(G_X(y)) - y\|_2]\end{aligned} \tag{2.13}$$

$$\begin{aligned}\mathcal{L}_{\text{idt}}(G_X, G_Y) = {} & \mathbb{E}_x[||G_X(x) - x||_1] \\ & + \mathbb{E}_y[||G_Y(y) - y||_1]\end{aligned} \tag{2.14}$$

$$\begin{aligned}\mathcal{L}(G_X, G_Y, D_X, X, Y) = {} & \mathcal{L}_{G_X}(G_X, D_X, X, Y) \\ & + \mathcal{L}_{G_Y}(G_Y, D_Y, X, Y) \\ & + \lambda \cdot \mathcal{L}_{cyc}(G_X, G_Y, X, Y) \\ & + (\lambda \cdot 0.5) \cdot \mathcal{L}_{\text{idt}}(G_X, G_Y)\end{aligned} \tag{2.15}$$

$$G_X^*, G_Y^* = \text{arg} \min_{G_X, G_Y} \max_{D_X, D_Y} \mathcal{L}(G_X, G_Y, D_X, X, Y) \tag{2.16}$$

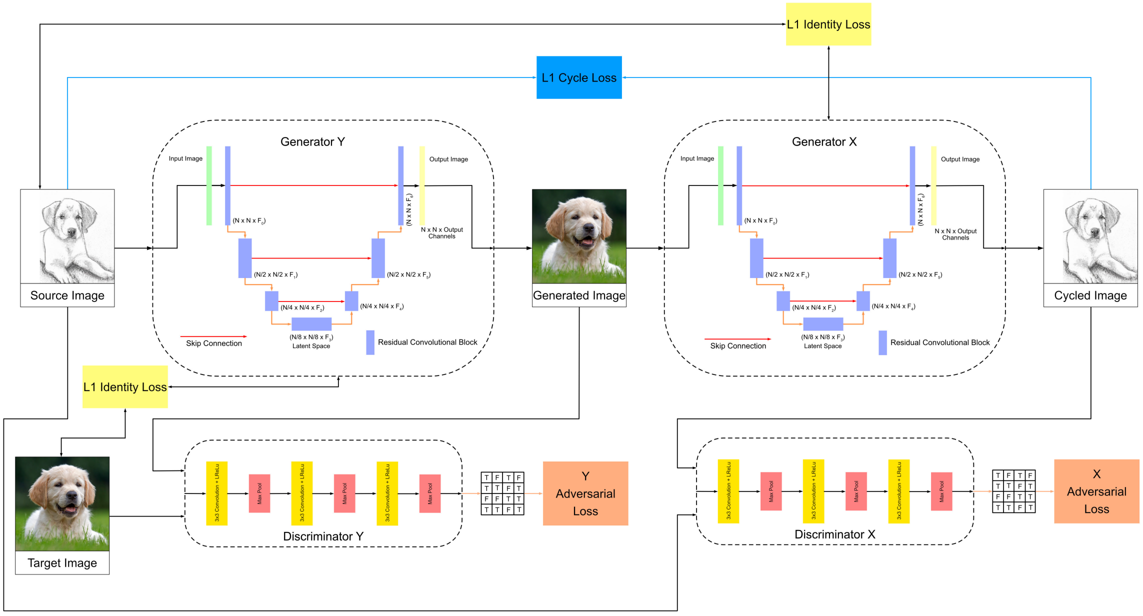


Figure 2.17: An example of a CycleGAN architecture as described by Zhu et al. (2020). It consists of twin generator-discriminator pairs. The first pair is tasked with creating realistic images in the target domain, and the second pair uses those generated images to recreate the images from the source domain. The pairs have an adversarial loss to ensure that images are realistic. The cycled images are constrained to be similar to teach the network to maintain the structure. Finally, a target image of each domain is passed through its respective generator and constrained to be identical, allowing the model to learn colour mapping. This is called identity loss. The resulting combination of this configuration allows the model to learn a mapping between the two domains without paired data.

2.7.5 Self-Supervised Learning

Self-supervised deep learning (SSL) has emerged as a promising technique to reduce the requirement for labelled data in deep learning (Poon et al. 2021). Supervised deep learning methods require large numbers of annotated examples, which can be time-consuming and expensive to obtain (Thiam et al. 2021). Self-supervised deep learning provides new protocols to pre-train a neural network on unlabelled data. The critical aspect is that they allow the models to learn valuable representations without explicit supervision (Sikorsky et al. 2021). In recent years, SSL has become very popular due to its ability to use large-scale unlabelled datasets (Jaiswal et al. 2021). The application of SSL has now become widespread given its benefits, and it has been successfully applied across various domains, including computer vision, natural language processing and medical image analysis (Poon et al. 2021; Thiam et al. 2021; Jing et al. 2019). By training on the readily available and vast quantities of accessible unlabelled data, self-supervised models can learn more generalisable and robust representations, improving performance on traditional downstream tasks like classification (Poon et al. 2021). Overall, SSL offers an efficient and powerful new approach to training deep neural networks with little or no manual annotation. In recent years, it has been one of the most valuable tools to emerge in deep learning. Contrastive Representation Learning is one of the most popular and practical SSL approaches proposed in this field.

Contrastive Representation Learning

Contrastive self-supervised deep learning, also known as contrastive learning of representations (CLR), is an exciting technique recently gaining traction. It focuses on learning feature representations from unlabelled data by distinguishing between similar or dissimilar sample pairs. This is often achieved by maximising the agreement of feature representations between differently augmented views of the same sample (positive pairs) while minimising the agreement between different samples of data (negative pairs) (Wang et al. 2023; Liu et al. 2023).

The typical implementation pipeline of CLR involves several stages. The first is pre-training, where data augmentations are applied to generate positive and negative sample pairs from the unlabelled data. These pairs are then input to a selected encoder network to obtain feature representations. A contrastive loss function is used to maximise the agreement between embeddings of positive pairs and minimise the agreement between negative pairs, thereby training the encoder to learn features in the data that can discriminate between the positive and negative pairs, without requiring supervision (Huang et al. 2023; Liu et al. 2023). Once pre-training has converged to a minimum, a classification or prediction layer is attached to the pre-trained encoder. This combined model can then be fine-tuned on a smaller labelled dataset to use the learned representations for a specific downstream task (Chowdhury et al.

2021). The pre-trained weights from the encoder provide a strong foundation, enabling the model to achieve better performance while being more robust and often generalising better due to the comparatively diverse and sizable quantities of unlabelled data used in pretraining (Liu et al. 2023).

Contrastive self-supervised learning was introduced in computer vision around 2014; however, the introduction of models like SimCLR by Chen et al. (2020) have resulted in a surge in popularity within the last few years due to the demonstration of significant performance gains (Wang et al. 2023). It has been widely applied to images, where data augmentations like flipping, mirroring, rotation, cropping and colour distortion are routinely used. These augmentations force the model to pay attention to the content of the images rather than superficial details, resulting in more robust feature extraction (Wang et al. 2023; Liu et al. 2023). Given the shortage of supervised data in medical imaging, CLR has been adapted to this domain, and there have been several studies where researchers have explored contrastive pre-training on unlabelled medical images and then performed fine-tuning on smaller supervised datasets (Chowdhury et al. 2021; Spathis et al. 2022; Huang et al. 2023). There have been a variety of new architecture flavours surrounding contrastive learning like VICReg by Bardes et al. (2022), which only requires positive augmented examples and uses a novel loss function to ensure diverse, invariant features are learned over a much smaller batch size. In summary, contrastive representation learning has emerged as a powerful tool, especially in domains like medical imaging with limited labelled data; by distinguishing between augmented views of data samples, it learns invariant feature representations that can be used to improve the performance and robustness of downstream tasks. To illustrate the implementation of these techniques, we shall highlight and discuss two of the most influential network architectures below, SimCLR introduced by Chen et al. (2020), and VICReg proposed by Bardes et al. (2022).

SimCLR

SimCLR (Chen et al. 2020) is a contrastive representation learning architecture that uses a neural network encoder (typically a ResNet) to convert an input image into a representation vector. After the base encoder, there is a projection head that maps the representation to a space where the contrastive loss can be applied, and the projection head is usually several layers of dense neurons arranged like a small multilayer perceptron (MLP). A key component of SimCLR is the use of extensive data augmentation. Two augmented views of the same input image create a positive pair, while augmentations of a different randomly sampled image form negative pairs. In the paper, they use random cropping, resizing, colour jitter and blur. A diagram depicting the implementation of a SimCLR network is given in figure 2.18. The loss function for the SimCLR architecture is shown below in equations 2.17 and 2.18.

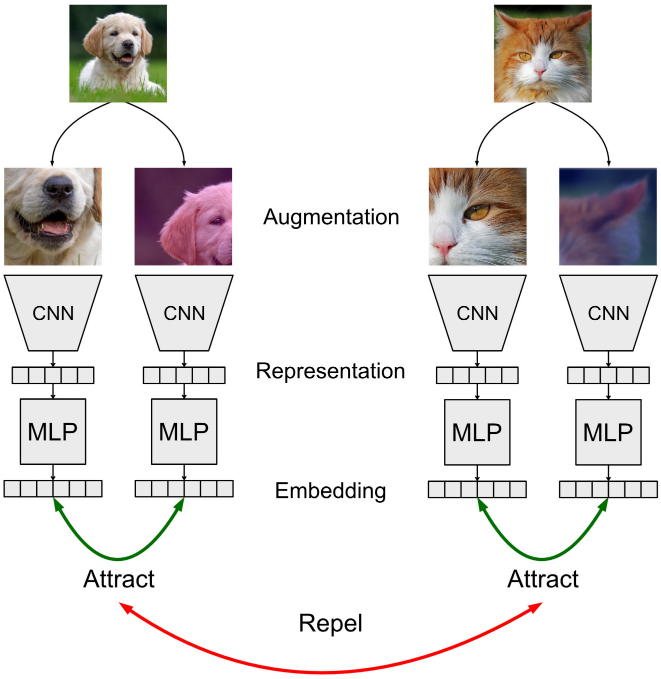


Figure 2.18: An example of a SimCLR architecture as described by Chen et al. (2020). The model consists of a base encoder network, a projection head and a contrastive loss function. The encoder network is a convolutional neural network that encodes the input image into a feature vector. The projection head is a small neural network that projects the feature vector into a latent embedding space. The model is trained to maximise the similarity of the two augmented versions of the same image and minimise the similarity of two augmented versions of different images. This process allows the model to learn a representation of the input images that capture the salient image features.

$$\mathcal{L}(i,j) = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{2.17}$$

Where: z_i and z_j are the representations of two augmented views of the same image (a positive pair) after being passed through the projection head and sim(a, b) is the cosine similarity between vectors a and b, defined in equation 2.18.

$$\text{sim}(a,b) = \frac{a \cdot b}{\|a\|_2 \|b\|_2} \tag{2.18}$$

τ is the temperature parameter, a hyperparameter that scales the similarity values in the loss function. $\mathbb{1}_{[k \neq i]}$ is an indicator function that is 1 when $k \neq i$ and 0 otherwise. This ensures that the fraction's denominator doesn't include the similarity of the representation with itself. The denominator sums over all negative pairs in the batch for the given positive pair.

When SimCLR was introduced, it achieved state-of-the-art performance on several ImageNet benchmarks without using any labelled data during the pre-training phase. It has become a popular choice for self-supervised contrastive representation learning. However, it is limited by being computationally expensive due to the requirement of enormous batch sizes during pre-training (Jaiswal et al. 2021).

VICReg

VICReg (Bardes et al. 2022) is an architecture designed to regularise the features in a representation learning network. It uses a neural network encoder (typically a ResNet) to convert an input image into a representation vector. After encoding, a projection head composed of several layers of dense neurons maps the representations into an embedding space. The key feature of the VICReg architecture is that it does not require negative samples for contrast. Instead, the network creates two augmented views of a batch of images. It then relies on clever regularisation of the learned representations to ensure they conform with the target objective of encoding the most salient image features. This is a much more efficient approach than SimCLR, as the need for negative examples is removed. A diagram of the VICReg architecture is given in figure 2.19.

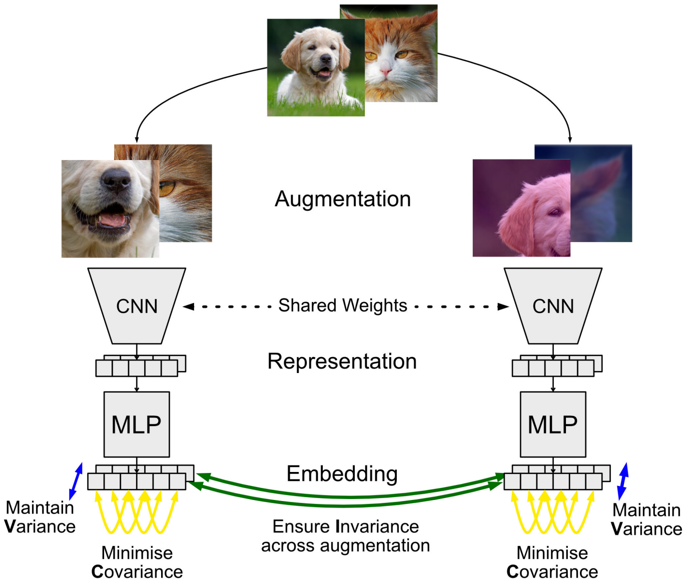


Figure 2.19: An example of a VICReg architecture as proposed by Bardes et al. (2022). An input image is first processed by a chosen encoder network into a feature representation vector. The representation is then passed through a projection head, which maps it into an embedding space. The VICReg loss function ensures the **V**ariance of the representations is maintained across the batch, that the representations are **I**nvariant across augmentations, and that the **C**ovariance is minimised so the representations do not encode redundant information.

The VICReg loss function has three terms that ensure robust representations are learned from the unlabelled data, as expressed in equation 2.19.

$$\mathcal{L}(Z, Z') = \lambda s(Z, Z') + \mu[v(Z) + v(Z')] + \nu[c(Z) + c(Z')] \tag{2.19}$$

The first term $s(Z, Z')$ represents the **I**nvariance loss, which minimises the mean squared error between embedding vectors Z and Z' from two different augmented views of the same image and helps to learn invariance to distortions. $v(Z)$ is the **V**ariance term that maintains the standard deviation of each dimension of the embeddings Z above a threshold using a hinge loss and helps to prevent collapse. $c(Z)$ is the **C**ovariance term, which minimises the off-diagonal entries of the covariance matrix of Z and decorrelates the embedding dimensions, helping to reduce the encoding of redundant information. λ, μ and ν are hyperparameters controlling the relative importance of the terms. The key idea is to regularise both embedding branches to preserve information content by maintaining variance and decorrelating dimensions. The invariance term brings the embeddings of different views closer and ensures the related features in the augmented images are maintained in the learned embeddings.

The combination of regularisation components and lack of negative examples makes VICReg much more stable and efficient to train than other approaches. It has also achieved state-of-the-art results on the standard ImageNet dataset, exceeding that of SimCLR, while training with vastly reduced batch sizes, making it one of the architectures of choice for self-supervised representation learning.

Clustering

Clustering is an unsupervised learning technique that groups similar data points into clusters based on selected measures of similarity, where the intent is to discover the inherent groupings within a dataset (Ezugwu et al. 2022). When dealing with vast amounts of unlabelled data, such as in the medical imaging domain, this process is pivotal as it aids in extracting meaningful information from such data where explicit labels don't exist. Clustering can be instrumental in revealing patterns, anomalies or structures that might otherwise go unnoticed.

There exists an entire taxonomy of clustering algorithms, each with a unique targeted approach and application, but there is a core group of techniques that we shall introduce. The first of these is partitional clustering methods, which include methods like k-means that divide data into distinct clusters by iteratively reassigning points to optimise cluster cohesion (Ezugwu et al. 2022; Singh et al. 2020). Density-based clustering algorithms, such as DBSCAN, identify clusters by selecting regions separated by lower density (Wegmann et al. 2021; Ezugwu et al. 2022). Grid-based clustering methods quantise the data into cells and then perform density clustering on these cells (Ezugwu et al. 2022). Another type is termed fuzzy clustering, which allows grouping data points with memberships in multiple clusters to varying degrees (Singh et al. 2020). There are also a group of methods based on nature-inspired processes like neural network clustering that employ ANNs like self-organising maps, and there are metaheuristic clustering methods that use optimisation algorithms such as genetic algorithms to assign data points to groups based on a selected measure of fitness (Singh et al. 2020).

There are also exciting techniques based on community detection within networks. The Louvain (Blondel et al. 2008) and Leiden (Traag et al. 2019) algorithms are two protocols for this type of clustering. They are commonly used with graph-based representations of data, where the data points are embedded in k-nearest neighbour graphs, with the neighbour distances assigned to the edge weights in the graph. The Louvain algorithm works by optimising the modularity of the network. It does this by measuring the quality of a network partitioning as it is split into communities and iteratively merges nodes into communities to build a hierarchy (Hairol Anuar et al. 2021). It is fast and straightforward but can get stuck in poor local optima (Traag et al. 2019). The Leiden algorithm (Traag et al. 2019) improves upon this by using a more granular approach. After an initial Louvain pass, it recursively optimises the detected communities. This is slower than the Louvain algorithm but helps avoid poor local optima, finds higher quality partitions, and works well on a range of network sizes (Hairol Anuar et al. 2021). They are both agglomerative approaches and identify densely connected groups of notes as communities. They are particularly applicable to self-supervised representations, as clustering the similarity graph of embeddings can discover meaningful groups aligned with the SSL training objective of finding semantically similar clusters, and the Louvain and Leiden algorithms efficiently extract these patterns without supervision.

In conclusion, clustering plays a pivotal role in modern data science, as is evidenced by the diverse variety of available clustering techniques. It is especially significant in the analysis of medical images, where when applied to the representations extracted from self-supervised representation learning, it can reveal valuable insights into underlying patterns or associations without manual supervision.

2.7.6 Transformers

There is one final type of deep network model that is relevant to this work: transformers. Transformers were proposed by Vaswani et al. (2017) and have developed over the last few years into one of the most powerful tools ever to be developed in deep learning, revolutionising several fields, including natural language processing and computer vision. While CNNs have long been the dominant architecture for image processing, recent research has shown that transformers offer several advantages over CNNs (Bai et al. 2021). Unlike CNNs, which rely on convolutional layers to extract local features, transformers utilise self-attention mechanisms to capture global dependencies within an input sequence (Touvron et al. 2021). This allows transformers to model long-range dependencies and capture contextual information effectively, making them well-suited for tasks that require understanding complex relationships and patterns in visual data (Lin et al. 2021).

Transformers comprise an encoder and a decoder, each containing multiple identical stacked layers. Input tokens, words or small image patches are first converted into embeddings using linear dense layers as described in section 2.7.1. Because transformers don't have an in-built ability to determine position, a positional encoding value is added to the embedding vectors to provide the model detail on the position of each token in the input sequence. The core idea behind transformers is the attention mechanism. Figure 2.20 shows a diagram depicting it. The attention mechanism allows the model to focus on different parts of the input sequence when producing an output. It consists of a scaled dot-product attention module that comprises a multi-head attention layer. Scaled dot-product attention works by repeating the input and passing it through three linear lays that learn to convert the information to query Q, key K and value vectors V. Given the queries and keys, the attention scores are computed as the dot product of Q and K and scaled by the root of their depth. The resulting matrix is then passed through a SoftMax function to scale it to a range between 0 and 1. These scores determine the weight of each value in the final output or how much attention should be paid to each. These scaled dot-product attention modules are run in multiple instances known as heads. These allow the transformer model to learn different mappings between inputs and output sequences and pay attention to parts in parallel to construct its output.

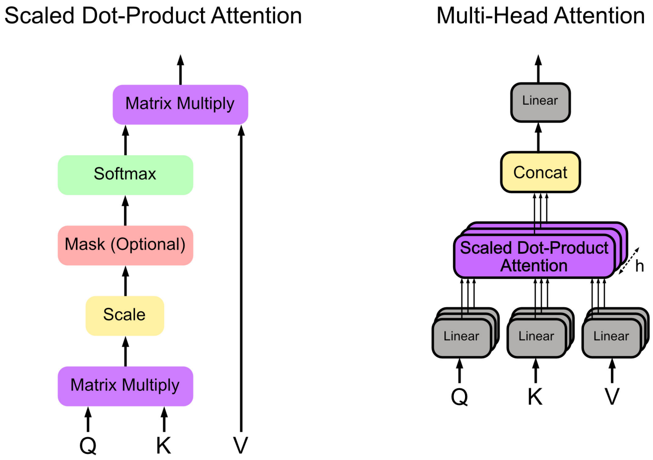


Figure 2.20: An example of multi-head self-attention as proposed by Vaswani et al. (2017). On the left is the key component of the multi-head attention module. The input is repeated and passed through a linear layer to produce a query, key and value. The attention score between a query and key is computed as the dot product between the two, and the matrix can be masked in the decoder to obscure future outputs from consideration. A softmax function is applied to the scores, producing attention weights, which are used to take a weighted sum of the values, producing the output of the attention mechanism. On the right, the multi-head attention module is shown. This is the key building block of a transformer. The original input is repeated over many heads, which allows the model to learn different relationships between the input and output. The outputs of the heads are concatenated and passed through a linear layer to produce the final output.

After the attention mechanism, the output is channelled into a feed-forward neural network as described in section 2.7.1, and each sub-layer in the network, the transformer has a residual connection to skip it, followed by a normalisation layer;

this helps in training deeper models and to mitigate the vanishing gradient problem (Touvron et al. 2021). This output is then passed from the encoder to the decoder or into the final layer in the network. The final layer of the transformer model is a linear layer followed by a SoftMax layer for tasks like classification or token prediction. Figure 2.21 gives a diagram depicting the entire transformer architecture.

In essence, the transformer's power lies in its attention mechanism, allowing it to assess the significance of different input parts dynamically. Because of this, a key benefit of transformers over CNNs is that they have demonstrated increased resistance to adversarial attacks (Bai et al. 2021) and demonstrated better generalisation capabilities, allowing them to perform well on various datasets and tasks with minimal finetuning (Guo et al. 2022). This, combined with their ability to capture long-range dependencies, makes them well-suited to a wide range of image-processing tasks.

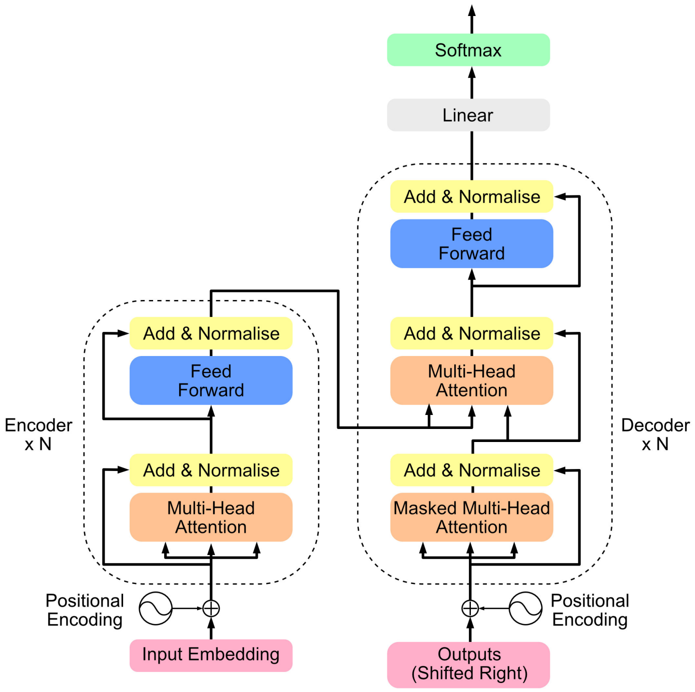


Figure 2.21: An example of a the transformer architecture as proposed by Vaswani et al. (2017). The model consists of an encoder and a decoder. The encoder comprises a stack of identical layers containing a multi-head self-attention module and a feed-forward network. The decoder is also composed of a stack of identical layers. However, each layer also contains a multi-head self-attention module that is masked to prevent it from attending to future outputs. The encoder and decoder are connected by an attention module that allows the decoder to focus on relevant input parts. The decoder output is passed through a linear layer and a softmax function to produce the final output sequence.

2.8 Virtual Immunohistochemistry

Histopathology, as described in section 2.1, is the examination of biopsied or resected tissue under a microscope, and it has been the cornerstone of medical diagnosis since its inception. The most used staining technique is haematoxylin and eosin, which highlights general cellular structures. However, while H&E provides a wealth of information, there are instances where it is insufficient, particularly when molecular-level differentiation and characterisation are required (Burlingame et al. 2020; Fujitani et al. 2019; Haan et al. 2021). Immunohistochemistry can help visualise these difficult regions. The IHC process is detailed in section 2.1, but in summary, it is a staining technique based on antigen-antibody binding, which can highlight features by attaching a chromogen at a molecular level (Liu et al. 2021a). Despite the invaluable insights that IHC offers, it is not universally accessible. The cost, equipment and expertise involved in the process can limit its use, particularly in low-resource settings, which can lead to disparities in the quality of diagnosis and patient care (Burlingame et al. 2020; Liu et al. 2021a). Physical staining methods, including IHC, are also difficult or impossible to reverse and time-consuming. Once tissue is stained with one technique, it typically cannot undergo another staining process, restricting the amount of information derived from a single sample. (Fujitani et al. 2019; Haan et al. 2021).

Virtual Immunohistochemistry (VIHC) has emerged as a promising solution to address the challenges and limitations associated with traditional histopathological staining techniques. The concept of virtual staining revolves around using automated algorithms to artificially replicate the effect of chemical and molecular staining without altering the slide physically (Xu et al. 2019). VIHC can leverage advanced computational techniques like Generative Adversarial Networks (as described in section 2.7.3) or CycleGANs (as defined in section 2.7.4) to generate virtually stained whole slide images from other modalities, thereby maximising the information extracted from a single biopsy or resection, reducing the cost, and resources associated with IHC and expediting the diagnostic process (Levy et al. 2021; Liu et al. 2021a). In essence, the evolution of Virtual Immunohistochemistry underscores the broader shift towards integrating advanced computational methodologies in histopathology, aiming to enhance diagnostic accuracy, efficiency and accessibility.

Despite their advanced nature, the deep-learning techniques used in virtual immunohistochemistry research have significant limitations. These limitations must be carefully considered when used for VIHC in a medical setting. The first concerns the preservation of fine structural detail during the stain translation process. This is especially important for diagnosis when minute tissue structure and cellular details are essential. GANs and CycleGANs, when not appropriately trained or constrained, might introduce artefacts or fail to capture these intricate details, which are crucial for accurate disease assessment (Liu et al. 2021a). Additionally, GANs can sometimes generate features in the output that do not exist in the input, a phenomenon referred to as “hallucinations” (Haan et al. 2021). In the context of VIHC, this could lead to false positives or misdiagnosis, which must be addressed before use. Finally, quantitively evaluating the performance of GANs is non-trivial. Traditional image accuracy metrics are not directly applicable, and specialised metrics like the Fréchet Inception Distance (FID) (Heusel et al. 2018) or Structural Similarity Index (SSIM) (Wang et al. 2004) are often used to evaluate generated images. However, these might not always align with human perceptual judgements (Liu et al. 2021a).

The remainder of this section will provide an overview of the state of the art in virtual staining. To do this, much of the publicly available research discussing the subject has been assessed. This review included twenty-two publications dating from 2017 onwards. Sixteen papers concern virtual stain-to-stain translation, and six discuss label-free staining. Given that H&E is applied to nearly all clinical cases, covering 80% of all human tissue staining performed globally (Haan et al. 2021), we chose to focus on developing and improving methods of stain-to-stain translation for virtual IHC. A summary of the published work on this area is provided below in table 2.3. We will now evaluate the latest trends in the field and the state of the art, including what has been done to address the propensity of generative deep networks to introduce artefacts and hallucinations into translated images and how generated virtual stains are evaluated for accuracy and perceptual quality.

To conduct this review, sixteen published papers with the highest number of citations for virtual staining were analysed; see table 2.3. Each piece of work was evaluated, and the following was recorded: the type of stains used in translation, the type of tissue, how the dataset was prepared in terms of the generation of training data, if the images were paired or unpaired, if they originated from consecutive serial sections, or if the tissue was washed and restained. We also evaluated the deep learning architecture used, if there were any novel modifications to its loss function, and finally, how the quality of the generated images was assessed.

In the available research, a diverse range of stains were used as targets for stain translation, including special stains like Masson’s Trichrome, which highlights collagen fibres, and PAS, period acid-Schiff, which highlights glycoproteins (Bai et al. 2023) and many IHC stains like cytokeratin CK8/18/19 and others like Ki67 and HER2. This suggests that virtual staining technology could apply to various stains and that there is enough information in other imaging and staining modalities to recreate them using artificial intelligence. It is worth noting that thirteen out of sixteen papers (81.25%) used H&E as the source stain, offering robust evidence that it provides enough histological detail to enable VIHC. Virtual staining was possible on the following tissue: colorectal liver metastases, pancreas, kidney, breast, prostate, brain and lung and was feasible on both human and mouse tumours. This also suggests that virtual staining is possible for an extensive range of tumour types. To train a deep neural network to translate between staining styles, a set of images is required

Author	IHC Stains Translated	Tissue Type	Paired / Unpaired	Deep Learning Architecture	Loss Function	Image Evaluation Metrics	Quantitative Scores
(Xu et al. 2019)	HE → CK18/19	Colorectal Liver Metastases	Serial Sections (4um)	CycleGAN (ResNet Generator)	Novel - Addition of Photorealism + SSIM + Classification Performance	Visual Assessment	-
(Lahiani et al. 2019)	KI67-CD8 → FAP-CK	Colorectal Liver Metastases	Serial Sections	CycleGAN (Resnet Generator)	Standard	Visual Assessment + Positive Cell Density	-
(Burlingame et al. 2020)	HE → IF (DAPI + PAN-CK + aSMA)	PDAC	Same Section (Restained)	cGAN (UNet Generator)	Standard	SSIM	Not Reported
(Haan et al. 2021)	HE → PAS, MT, JS	Kidney	Same Section (Restained)	cGAN (Unet Generator)	Novel - Addition of Total Variation Loss	Visual Assessment	-
(Mercan et al. 2020)	HE → PHH3	Breast	Same Section (Restained)	cGAN + CycleGAN (Resnet Generators)	Standard	Downstream Task Assessment	-
(Levy et al. 2021)	HE → MT	Liver	Serial Sections (5um)	CycleGAN (UNet Generator)	Standard	Downstream Task Assessment	-
(Lo et al. 2021)	HE → PAS, MT	Kidney	Unpaired	CycleGAN (Resnet Generator)	Standard	SSIM + Visual Assessment	0.8478 (0.82-0.87
(Lahiani et al. 2021)	HE → FAP-CK	Colorectal Liver Metastases	Serial Sections	CycleGAN (Resnet Generator)	Novel - Addition of Perceptual Consistency Loss	CWSSIM	0.83 +/- 0.17
(Liu et al. 2021a)	HE → Ki-67	Neuroendocrine Breast	Unpaired (Training) Serial Sections (Test)	CycleGAN (UNet Generator)	Novel - Pathology Consistency Loss + SSIM + Minimising Encoder Feature Space	CSS	0.8743+/-0.263
(Xie et al. 2022)	HE → CK8	Prostate	Serial Sections	Vid2Vid (Cascaded Refinement Network)	Standard	3D SSIM	0.41
(Fujitani et al. 2019)	HE → MT	Mouse Pancreas	Serial Section s (4um)	(FCNN Generator)	MSE	MSE - Chromatic Distribution + Visual Assessment	-
(Chen et al. 2021b)	MUSE → HE	Mouse Brain + Mouse Liver	Same Section (Restained)	cGAN + CycleGAN (UNet Generators)	Novel - Addition of SSIM Loss	SSIM, PSNR PCC, MMD + Visual Assessment	SSIM - 0.30 +/- 0.09 (training) 0.73 +/- 0.08 (validation)
(Zhang et al. 2022)	HE → CC10, Ki67, proSPC, ER, HER2, PR, ASMA, OilRedO	Lung, Breast and Atherosclerotic Lesions	Serial Sections	Modified CycleGAN	Novel - Addition of Style + Content Loss	PSNR + Histogram Error	-
(Liu et al. 2022)	HE → HER2	Breast	Same Section (Restained)	cGAN - Pyramidal (UNet Generator)	Novel - L1 Pyramidal Loss	PSNR + SSIM	PSNR - 21.160 SSIM - 0.477
(Hong et al. 2021)	HE → PanCK	Colorectal	Same Section (Restained)	cGAN (UNet Generator)	Novel L1 loss in HED colour space.	Visual Assessment + Downstream Task Assessment	-
(Bouteldja et al. 2022)	CD31 aSMA, Col3, NGAL → PAS	Kidney	Serial Sections (1-2um)	CycleGAN (UNet Generator)	Novel - Addition of a Segmentation Loss	Segmentation Accuracy	-

Table 2.3: A summary of the current literature on virtual immunohistochemistry.

in the style of the source stain and another set in the target stain. These must be structurally paired and aligned at the pixel level for supervised networks like a GAN, or they can be unpaired for an unsupervised network like a CycleGAN. There are benefits and challenges to both approaches. Eight of sixteen (50.00%) used paired consecutive serial sections cut between 1 and 5 microns. Six out of sixteen (35.50%) used the same tissue that was washed and restained or otherwise where the process allowed dual staining. Finally, two of sixteen (12.50%) used unpaired images in the source and target stain styles. The most frequent arguments for consecutive serial sections are that they are easier to produce than restained sections, they are more common in clinical practice, and because it's a clean piece of tissue for each stain, it is not deformed or damaged by the washing process (Zhang et al. 2022; Chen et al. 2021b; Fujitani et al. 2019). The most common arguments against serial sections were that morphological differences could arise between sections, leading to challenges with registration and alignment and that it consumes more tissue that is a finite resource (Hong et al. 2021; Xu et al. 2019; Burlingame et al. 2020). The most common arguments for restaining the same tissue were that because the tissue is, in theory, identical, it can provide more accurate ground truth and guarantee pixel-level registration with minimal deformation of the tissue and that it uses less tissue than consecutive sections (Hong et al. 2021; Mercan et al. 2020; Burlingame et al. 2020). The most common arguments against restaining the same tissue were that it is difficult or not possible in certain circumstances and that, in practice, the tissue can be distorted. Additionally, some of the same challenges as consecutive sections can be encountered with tissue artefacts and deformations, although usually to a lesser degree (Fujitani et al. 2019; Hong et al. 2021; Zhang et al. 2022).

A variety of deep learning architectures were used across the literature; the two most common were the Conditional GAN, usually based on the Pix2Pix architecture by Isola et al. (2018), and the CycleGAN architecture by Zhu et al. (2020). In the surveyed work, there were ten CycleGAN architectures (62.50%) and four Conditional GANs (25.00%). The remaining architectures were custom deep convolutional neural networks. The most common arguments for the use of the CycleGAN in the context of virtual staining were that it represents the state of the art in unsupervised domain adaption technology in that it has repeatedly demonstrated its ability to generate realistic style translations in other domains and, given its nature, it can handle unpaired images, that is highly beneficial for stain translation because of the nature of histology images there may be subtle histomorphological variations between sections or due to tissue deformation (Bouteldja et al. 2022; Lo et al. 2021; Xu et al. 2019). The arguments against CycleGAN were mainly that it requires substantially more resources than the Pix2Pix network to train and that, given that it can rely on unpaired data, there exists the possibility that it can provide less specific style transformation than a Pix2Pix network (Mercan et al. 2020; Hong et al. 2021). The most common arguments for using the Pix2Pix network were that, given that it is a supervised network, it could learn the direct mapping concerning context and structure between different types of images, ensuring high semantic correctness and that it has a proven track record for image-to-image translation in a medical setting (Hong et al. 2021; Burlingame et al. 2020). The most common argument against the Pix2Pix network was the requirement that it be trained on paired training data, where the tissue registration must be accurate at the pixel level to ensure acceptable results (Mercan et al. 2020).

The loss functions in the evaluated works were either direct implementations of the standard GAN loss as described in section 2.7.3, the standard CycleGAN loss as defined in section 2.7.4 or novel variations of those functions. Nine of the sixteen works (56.25%) were novel additions or modifications to the loss of the proposed architecture; the remainder (45.75%) used the customary approach. We shall now discuss the novel additions to the loss functions and the author's arguments for their intended effects on the stain translation process. Three of the works implemented the Structural Similarity Index (SSIM) by Wang et al. (2004) as a measure of loss; in these works, it was added in place of, or as an addition to the cycle loss term, as measured on the CycleGANs reconstructed source images. The authors argued that SSIM was an improvement over the L1 loss proposed in the original CycleGAN paper by Zhu et al. (2020), as L1 loss has a tendency to blur the image as it is just a mean average error between the pixel values (Zhu et al. 2020; Liu et al. 2021a; Chen et al. 2021b). In contrast, the SSIM metric is designed to provide a more accurate and meaningful assessment of perceived image quality by comparing the luminance, contrast and structure of two images (Heusel et al. 2018). The second most common addition to the loss functions was including a custom term to minimise the distance between the latent space representations of the CycleGAN generators. The authors argued that this would force the networks to ensure the semantic content was shared across the generated images and that only the style should vary across images (Lahiani et al. 2021; Liu et al. 2021a). The three final novel alterations were the addition of a photorealism loss term, as defined by Luan et al. (2017) to constrain the structure across the input and stain translated images in a CycleGAN proposed by Xu et al. (2019). Including an L1 loss to the GAN function in a stain-deconvolved haematoxylin, eosin, and DAB colour space to emphasise the difference in stain intensity proposed by Hong et al. (2021). Finally, the last novel loss addition was the inclusion of segmentation networks that share the encoder parameters in a CycleGAN to force the encoders to understand the semantic content as proposed by Bouteldja et al. (2022) and Liu et al. (2021a).

Unfortunately, no standard method of comparison for stain-translation performance has yet emerged in the field, and a wide variety of metrics are used to evaluate performance in the literature. The most common method was a visual assessment, either by the authors or expert pathologists, in seven of the sixteen works (43.75%); however, given that this is a subjective measure, it is not directly comparable. The second most common method, and the most likely emerging candidate for a standard quantitative evaluation metric, is the structural similarity index measure (SSIM), used to evaluate five of the sixteen works. The runner-up is Peak Signal to Noise Ratio (PSNR), which measures the quality of a reconstructed image by quantifying the difference between it and the original in decibels (Horé et al. 2010). PSNR was used in three of the methods (18.75%). Other works used variations on the SSIM method; one only used the contrast term of the SSIM equation and called it CSS to avoid the loss of information resulting from the difference in luminance between stains (Liu et al. 2021a). While this idea is promising, it isn't easy to directly compare it to the other methods. The final evaluation metric used was Complex-Wavelet Structural Similarity (CWSS) (Sampat et al. 2009). Lahiani et al. (2021) argues that because CWSSIM is invariant to small translation and rotations, it will be a more useful metric in the evaluation of stain translation due to the deformations of this nature that can occur in histology slides again, this is a good argument but is more complex to implement and harder to compare to standard SSIM.

Given the lack of agreement on a standard quantitative measure and any quantitative measure of image quality in many works, it is difficult to conclude the best-performing method. Of the five that used SSIM, one did not report their final value, one only reported the value on the reconstructed images and not the stain-translated image, and another reported it across a 3D reconstruction of a block and is therefore not comparable to the other 2D generated images. The two remaining works differed in implementation, with the work of Chen et al. (2021b) providing an SSIM value on the stain translated image produced by a CycleGAN of 0.30 and Liu et al. (2022) reporting the SSIM value of 0.477 for the stain-translated images created by a pyramidal Pix2Pix conditional GAN. The difference in performance between the CycleGAN and Pix2Pix networks is interesting, given that, due to its design, the CycleGAN network should provide improved structural reliability in the generated images. However, what must be considered is the difference in staining modalities. In work by Chen et al. (2021b), the generated images are in the H&E domain, which offers little contrast between tissue, whereas in the work of Liu et al. (2022), the generated stain domain is HER2, an IHC stain with a chromogen providing a stark contrast between target cells and background tissue. This improved contrast would likely result in a higher SSIM score by default; therefore, it is not a fair comparison. This again highlights the non-trivial nature of assessing the quality of generated images, especially when the evaluation is across differing stain types.

After a review of the current research in virtual staining, it is clear that the CycleGAN architecture is the state-of-the-art deep network for stain translation. It is the most utilised architecture in the published works, and there is a robust set of arguments towards its competence for producing structurally accurate photo-realistic images and for its capability in dealing with the semi-paired nature of consecutive serial sections (Xu et al. 2019; Lahiani et al. 2019; Levy et al. 2021; Lo et al. 2021; Lahiani et al. 2021; Liu et al. 2021a; Chen et al. 2021b; Bouteldja et al. 2022). The most common alteration to its loss function was through the addition of an SSIM component between the source and generated image to reinforce the retention of structure or between the source and recovered source images to bolster the structure retention throughout the cycle. Interestingly, in the reviewed work, most modifications to the CycleGAN loss function concerned the recovered image or the generator parameters, most likely led by the intended unsupervised, unpaired nature of the CycleGAN. What is missing from the literature is an examination of how an attempt to train the CycleGAN in a supervised manner with constraints to the generated stain-translated image and its “semi-paired” counterpart from a consecutive slice or restained tissue affects the performance of these networks for stain-translation, this gap will be the focus of chapter 3 in this work.

2.9 Automated Tumour Bud Scoring

Tumour budding is a histopathological feature observed in colorectal cancer. It is characterised by the presence of individual tumour cells or small clusters of up to five cells, typically located at the invasive margin of a solid tumour (Fauzi et al. 2020; Tavolara et al. 2022; Weis et al. 2018). It is described in detail in section 2.3, but in brief, it has now become an established prognostic factor in colorectal cancer, with its presence and density being predictive of nodal metastatic disease, which is a useful marker in early-stage colorectal cancer (Fisher et al. 2021; Jepsen et al. 2018). A high density of tumour buds has been associated with various adverse pathological features such as lympho-vascular invasion, venous invasion, regional and distant lymph node metastases, local recurrence and reduced overall and disease-free survival (Jepsen et al. 2018; Takamatsu et al. 2019). These associations make it a valuable prognostic factor for risk stratification to distinguish patients that require more aggressive treatment, especially in early-stage and endoscopically resected tumours (Fauzi et al. 2020; Banaeeyan et al. 2020). Despite its clinical significance, the routine implementation of tumour bud scoring in clinical practice has been hindered by inconsistent criteria, definitions, and assessment methods (Fisher et al. 2021; Tavolara et al. 2022). However, efforts towards standardisation and implementation in clinical practice are underway, such as the guidelines proposed during the International Tumour Budding Consensus Conference in 2016 and the introduction of computer-aided detection techniques, which aim to standardise and facilitate its use (Bergler et al. 2019; Bokhorst et al. 2023).

The idea of automated tumour bud scoring in colorectal cancer has been gaining increasing attention in recent years due to its potential to standardise, expedite and enhance the accuracy of the assessment of tumour budding. Manual tumour bud scoring is recognised as time-consuming, subjective, and prone to inter-observer variability, which can lead to inconsistencies in results (Tavolara et al. 2022; Weis et al. 2018; Lu et al. 2022). Given that colorectal cancer is the third most common cancer, with incidence rates set to increase, the manual evaluation of tumour budding would be unfeasible at scale with the existing resources, emphasising the need for automation (Fauzi et al. 2020). Automated methods have demonstrated the ability to be more sensitive in detecting buds than manual methods, and their results can strongly correlate with manual counts, indicating their reliability (Fisher et al. 2021; Takamatsu et al. 2019). Furthermore, as pathology increasingly digitises, integrating digital image analysis tools becomes more feasible. This should negate any prior objections to the resource-intensive nature of the image acquisition process (Jepsen et al. 2018). Given the importance of tumour budding as a prognostic factor, its accurate assessment is crucial. Automated systems can reduce the workload of pathologists, especially in high-volume settings and mitigate the risk of missed diagnoses due to visual fatigue (Lu et al. 2022). Additionally, automated bud scoring can provide more information at scale, such as spatial statistics, which might offer additional insights and could be particularly useful in large-scale studies or clinical settings where time and consistency are paramount (Weis et al. 2018; Caie et al. 2014). In summary, the interest in automated tumour bud scoring is not only a technical curiosity but a necessary evolution to enable the use of an important biomarker and ensure consistent, accurate and efficient patient diagnosis and stratification to enhance care and survival outcomes.

The remainder of this section will provide an overview of the state of the art in automated tumour bud scoring. Much of the publicly available research discussing the subject has been assessed to carry this out. This review was conducted on twelve publications dating from 2014 onwards. Seven papers concern automated budding on IHC pan-cytokeratin-stained slides. We shall discuss all works but focus mainly on the remaining five pieces of research that focus on deep learning-based approaches to automated bud scoring on H&E-stained slides, given that nearly 80% of all human tissue is stained with H&E by default (Haan et al. 2021). A summary of the published work in automated bud scoring is given below in table 2.4.

Author	Staining	Dataset	Method	Evaluation Metric	Result
(Caie et al. 2014)	IF - Pan-CK, D2-40, DAPI	50 CRC patient slides	Commercial software: segmentation + nuclei detection	Survival analysis of high vs low budding	HR = 5.7 (2.38-13.8)
(Bokhorst et al. 2018)	H&E	60 CRC patient slides	CNN: segmentation on H&E, IHC ground truth	ROC with manual annotation	F1 = 0.36 Recall = 0.72
(Jepsen et al. 2018)	Pan-CK	126 CRC patient slides	Visopharm app: segmentation based on Pan-CK + size constraints	Statistical correlation of manual and automated counts	R = 0.84 p<0.001
(Weis et al. 2018)	Pan-CK	381 patients - 20 slides + TMAs	MATLAB CNN: trained on 6292 tiles labelled for bud presence.	Statistical correlation of manual and automated counts	R = 0.86 p=0.003. No correlation with survival outcome.
(Bergler et al. 2019)	Pan-CK	87 CRC patient slides	Thresholding and morphological operations, filtered with AlexNet	ROC with manual annotations	Precision = 0.977 Sensitivity = 0.934
(Takamatsu et al. 2019)	Pan-CK	517 CRC patient slides	Manual field selection + ImageJ macro analysis of binary image.	Inter-observer agreement manual vs semi-automated	Manual Kappa = 0.463 Semi-Automated Kappa = 0.781
(Banaeeyan et al. 2020)	H&E	5 CRC patient slides	CNN segmentation trained on manual annotations	IOU between automated and manual segmentations	IOU = 0.49
(Fauzi et al. 2020)	Pan-CK	5 CRC patient slides	Thresholding and morphological operations	ROC with manual annotations.	Sensitivity = 0.70 Specificity = 0.82
(Fisher et al. 2021)	Pan-CK	186 CRC TMA cores	Open source software QuPath: binary classifier (stain deconvolution)	Survival analysis of high vs low budding	HR = 1.06 (1.01-1.11), p <0.05
(Lu et al. 2022)	H&E	100 CRC patient slides	Fast R-CNN object detection	ROC with manual annotations	AUC = 0.96
(Tavolara et al. 2022)	H&E	120 CRC patient slides	Segmentation with a Swin transformer	ROC with manual annotations	Precision = 0.3856 Recall = 0.3254
(Bokhorst et al. 2023)	H&E	981 CRC patient slides	UNet (tissue segmentation), UNet (epithelial segmentation) and Hovernet nuclei segmentation	ROC + Survival Analysis of high vs low budding	F1 = 0.58 Recall = 0.95 HR = 1.8 (1.2-2.6)

Table 2.4: A summary of the literature on automated tumour bud scoring.

To conduct this review, each piece of work was evaluated, and the following was recorded: the dataset scale and type of tissue stain, the method used for automated tumour bud scoring, how the results were evaluated, and the quantitative metrics logged. If made, the arguments for and against the choice of stain, scoring method and evaluation metrics were also recorded for later comparative analysis. We will now review the latest trends in the field, the state of the art for H&E-based scoring, and how these methods are evaluated for reproducibility and accuracy.

The most common tissue stains in the evaluated works were H&E and IHC pan-cytokeratin, which appeared in eleven of the twelve papers (91.67%). The remaining paper used immunofluorescence pan-ck, D2-40, and DAPI stains to highlight the cell nuclei for segmentation. Five of the twelve works (41.66%) developed methods for automated bud scoring using H&E slides. The most common arguments for H&E over IHC were that H&E is more routinely applicable than immunofluorescence of IHC, as it is recognised as the gold standard histological stain for diagnosing CRC (Bokhorst et al. 2018; Tavolara et al. 2022). Notably, professional bodies like the College of American Pathologists and the ITBCC advocate for tumour bud counts on H&E sections, reserving IHC for cases with challenging inflammation only due to the resource requirements of its use (Tavolara et al. 2022). Additionally, while IHC offers enhanced clarity in certain areas, H&E remains superior in evaluating the microenvironment surrounding the buds (Fauzi et al. 2020; Fisher et al. 2021). Arguments were also made that while automated methods can benefit from having IHC ground truth, deployed models should rely solely on H&E, ensuring streamlined and efficient use on already available clinical data (Bokhorst et al. 2018; Tavolara et al. 2022). Three out of five (60%) models trained on H&E used IHC to assist with manual annotations. For the remaining work, seven out of twelve (58.33%) reviewed papers used IHC or IF pan-cytokeratin as the primary stain in the models. The most common arguments for the use of IHC/IF over H&E in the automated assessment of tumour budding is that IHC enhances the contrast of the budding cells, making them easier to detect, which leads to improved reproducibility of assessments and a reduction in the workload associated with manual annotation, thereby minimising interobserver variations in ground truth data (Jepsen et al. 2018). Also, cytokeratin IHC has been shown to provide more accurate identification of tumour budding in cases with significant inflammation (Takamatsu et al. 2019). Studies have also consistently shown that higher tumour budding counts result from IHC, with some indicating counts three to six times greater than H&E, which should vastly improve the accuracy of manually annotated ground truth (Fauzi et al. 2020; Fisher et al. 2021). Pan-cytokeratin is also particularly adept at differentiating tumour cells from background stromal and inflammatory cells, making it invaluable in highlighting single-cell buds, which can be challenging to differentiate in H&E (Fauzi et al. 2020; Fisher et al. 2021).

In the evaluated work, various methods are employed for automated tumour bud detection. Five out of twelve (41.67%) used commercial or open-source software packages for segmentation using prepackaged neural networks or machine learning algorithms trained on manual annotations or thresholding of the cytokeratin slides (Caie et al. 2014; Jepsen et al. 2018; Weis et al. 2018; Takamatsu et al. 2019; Fisher et al. 2021). Two used basic thresholding and morphological operations to segment the tumour buds from the raw pan-cytokeratin images (Bergler et al. 2019; Fauzi et al. 2020). The remaining work used more complex methodologies based on deep learning, notably because they all detected buds on H&E, where the more straightforward methods are not applicable.

We shall now discuss the deep learning techniques used on the H&E-based

slides. The first of these works by Bokhorst et al. (2018) use a VGG (Simonyan et al. 2015) network to act as a classifier on patches where the whole slide images have been tessellated, and then the patches labelled as containing tumour buds, tumour glands and background. This was an efficient approach as it reduced the need for any manual pixel-level segmentation, which is resource-intensive to generate. However, the resulting F1 score of 0.36 falls just above the baseline three-class random expectation of 0.33. The next of these works by Banaeeyan et al. (2020) use an encoder-decoder network based on SegNet (Badrinarayanan et al. 2016), which is similar to the U-Net architecture discussed in section 2.7.2, except that rather than transferring the actual feature map values in the skip connections, only the index of the highest value in the pooling layers is transferred to place emphasis on what to pay attention to during upsampling. The SegNet-based architecture was trained on manual pixel-level segmentation created by the pathologist. The resulting segmentations performed well with a weighted IOU of 0.83. The next of the works based on H&E was that proposed by Lu et al. (2022), which uses a Fast RCNN network (Girshick 2015) for object detection, where bounding boxes containing tumour buds have been manually marked by pathologists for training. This is a type of network that proposes classified bounding boxes for objects in an image rather than pixel-level segmentation. This achieved a notable AUC score of 0.96 in identifying regions with buds. The penultimate technique based on H&E was proposed by Tavolara et al. (2022). The architecture is based on a Swin transformer (Liu et al. 2021b). This state-of-the-art encoder model differentiates itself from a traditional CNN by using a natural language processing-based "attention" mechanism, described in section 2.7.6, at multiple scales of an input image through a shifted window approach. Here, a decoder module is attached to recreate the segmentation maps. This model performed sub-optimally on the measured precision and recall metrics of 0.3856 and 0.3254, respectively. Most likely because transformers require a large amount of data to train. The final H&E-based technique was proposed by (Bokhorst et al. 2023). This method was based on the U-Net architecture described in section 2.7.2. It uses the U-Net architecture for tissue segmentation to detect the tumour, core, and epithelial tissue segmentation based on IHC ground truth, and finally, Hovernet (Graham et al. 2019) nuclei segmentation. Buds can be found by overlaying the nuclei segmentations with the IHC-positive predictions. It is evaluated using ROC characteristics, where it achieved a better-than-random F1 score of 0.58. Most notably it was also assessed using survival analysis over 981 patient slides, where it has provided a baseline cox-regression hazard ratio of 1.8 CI(1.2-2.6).

There was a variety of evaluation metrics used across the literature. The most common, being used in six of the twelve (50%) works, was analysis using receiver operator characteristics (ROC) to compare the automated vs. manual annotations, with the best-performing method producing a value of 0.977 precision and 0.934 sensitivity by Bergler et al. (2019). The next most common evaluation metric was survival analysis, using cox-regression and Kaplan-Meier plots used in three out of twelve (25%) works. The best-performing method resulted in a hazard ratio of 5.7 CI(2.38-13.8) for high budding. However, this was only over 50 patients and might not reflect an achievable score over a significantly larger cohort. The remainder of the work was evaluated using statistical correlation between the predicted and actual bud counts, three out of twelve (25%), and intersection over union analysis, one out of twelve (8.33%).

After a review of the current research in automated tumour bud scoring, classical CNN-based segmentation with a U-Net-like network is the most popular. Based on the evaluated statistics, this architecture is currently state of the art for automated bud scoring. This type of network is used in three of the five H&E methods (60%). They tend to be trained on manually annotated ground truth or ground truth inferred from physical IHC. Given the promise of our virtual IHC methods proposed in chapter 3 and the success of using IHC ground truth for assisting in H&E-based training, as was done in three of five (60%) evaluated H&E papers, there exists a gap in the literature to determine if it is possible to train a deep convolutional U-Net-based segmentation network to detect buds in H&E based on virtual immunohistochemistry ground truth. This will be the focus of our work as presented in chapter 4.

2.10 Automated Survival Prediction

Survival prediction in oncology refers to estimating the probability of survival for patients based on various factors, including histopathological images. In the context of colorectal cancer, this ability to predict survival is of paramount importance as it aids in tailoring treatments, guiding clinical decisions, and achieving more targeted treatment plans (Dawson et al. 2019c; Bychkov et al. 2018; Kather et al. 2019). This is of critical importance as while chemotherapy can be very effective in treating colorectal cancer, it can also have significant side effects (Ascierto et al. 2020). Therefore, finding the set of patients for which it is necessary and beneficial is paramount, and this is why survival prediction is key (Wulczyn et al. 2021). However, there is a shortage of trained specialists in histopathology, with a census report from The Royal College of Pathologists census stating that less than 3% of pathology departments in the UK have enough staff to cover their workload (Martin 2018). This suggests a clinical need for automated methods to assist in this area.

Deep learning has emerged as a powerful tool in this domain. It can use deep neural networks to predict time-to-event outcomes such as overall survival or disease-free survival (Mobadersany et al. 2018; Jiang et al. 2020). Several studies have demonstrated the potential of deep learning in survival prediction. For instance, Bychkov et al. (2018) investigated using a deep learning model to predict the five-year disease-specific survival of CRC patients using H&E images alone. Similarly, Kather et al. (2019) highlighted the potential of deep learning in extracting prognostic biomarkers directly from histology slides and demonstrated the ability of deep learning to extract previously unknown prognostic information in the stromal component of the CRC tissue.

However, there are challenges to consider. The vast amount of image data, combined with morphological heterogeneity and unknown discriminative patterns, makes survival prediction a challenging task (Wulczyn et al. 2020). Furthermore, while deep learning models can provide insights into prediction mechanisms, such as recognising structures related to prognosis, the interpretability of these models remains a topic of ongoing research (Mobadersany et al. 2018). Acknowledging the grounding role of traditional manual methods in oncology is essential, as histology has been a cornerstone in cancer diagnosis and prognostics for over a century. But manual evaluations, while foundational, can sometimes be subjective and may not capture the intricate details that deep learning can (Mobadersany et al. 2018; Bychkov et al. 2018).

In summary, survival prediction in colorectal cancer using deep learning on H&E WSI offers a promising avenue for enhancing the accuracy and precision of prognostic estimates. However, the lack of interpretability of deep learning and the proven nature of manual methods suggest that a compromise must be reached between the black-box nature of deep learning and the subjective nature of manual evaluation.

This section will provide an overview of the published research in automated survival prediction. Much of the publicly available work surrounding colorectal cancer was assessed to ascertain the field's current state. This review included ten publications dating from 2018 onwards. A summary is provided in table 2.5. We will now evaluate the latest trends in the field and the state of the art, including what work has been done to make automated survival prediction more interpretable.

Author	Cancer	Dataset	Objective	Method	Evaluation Metric	Result
(Mobadersany et al. 2018)	Glioma	1061 H&E WSI from 769 patients	Survival Prediction	CNN used to directly output a risk score from input tiles.	C-Index + Hazard Ratio	C-Index=0.801 HR=3.05 $p<0.05$
(Bychkov et al. 2018)	Colorectal Cancer	H&E TMAs from 641 Patients	Survival Prediction	CNN used for feature extraction, LSTM used to read tile vectors over WSI and make prediction.	AUC Hazard Ratio	AUC=0.69 HR=2.3 (1.79–3.03)
(Kather et al. 2019)	Colorectal Cancer	86+25 training and validation WSI and 862 H&E TCGA WSI	Deep Stroma Score	CNN used to predict a deep stroma score over WSI tiles.	HR	HR=1.99 (1.27-3.12) $p<0.05$
(Skrede et al. 2020)	Colorectal Cancer	828 training H&E WSI and 1122 Test WSIs	Survival Prediction	CNNs used to directly predict survival outcome.	HR	HR=3.04 (2.07-4.47) $p<0.05$
(Wulczyn et al. 2020)	Pan-Cancer	6075 training, 3011 validation and 3009 Test H&E WSI	Survival Prediction	CNN used to predict the probability of an event occurring over binned time periods from sampled WSI tiles.	HR	HR=1.58 (1.28-1.70) $p<0.05$
(Jiang et al. 2020)	Colorectal Cancer	168 H&E WSI	Survival and Recurrence Prediction	CNNs used to predict tissue type and gradient boosting decision tree used to predict survival and recurrence from tissue proportions.	HR	HR=10.687 (2.908-39.272) $p<0.05$
(Brockmoeller et al. 2022)	Colorectal Cancer	514 H&E WSI	Lymph Node Metastasis Prediction	CNN was trained to predict LNM status per tile, and proportion of tiles per WSI was used to classify LNM status.	AUC	0.733 (0.67-0.758)
(Tsai et al. 2023)	Colorectal Cancer	1888 H&E WSI	Survival Prediction	ResNet50 used in combination with vision transformers to cluster tiles in WSI, multiple instance learning used to make prediction from clusters.	Logrank Test	$p<0.05$
(Li et al. 2022)	Colorectal Cancer	1713 H&E WSI	Survival Prediction	Feature extraction using Xception CNN, prediction using multiple instance learning network.	C-Index	0.611 (0.58-0.63)
(Laleh et al. 2021)	Colorectal and Gastrointestinal Cancer	413 CRC WSI and 362 gastric H&E	Survival Prediction	ResNet50 used to generate a risk score for tiles across WSI, this is then averaged to produce a score per WSI.	Logrank Test + HR	LR $p<0.05$ HR=0.50 (0.3227-0.7864)

Abbreviations: WSI= Whole Slide Image, TMA= Tissue Micro Array, HR= Hazard Ratio, AUC= Area under ROC Curve.

Table 2.5: A summary of the literature on automated survival prediction.

In the reviewed work, eight of ten projects (80%) involved colorectal cancer, all of which used H&E images as input to their survival prediction algorithms. Nine of ten projects (90%) utilised whole slide images, and one operated on tissue microarrays. Eight of ten works (80%) had survival prediction as their primary objective. Two had the prediction of other biomarkers as their primary goal, with Kather et al. (2019) generating a Deep Stroma Score and Brockmoeller et al. (2022) predicting distant lymph node metastasis.

The various survival prediction methods were primarily evaluated using survival analysis, quoting a Cox regression hazard ratio in six of the ten works (60%), two of ten used receiver-operator characteristics for analysis of their accuracy to real-survival predictions, another two used a correlation index score for comparison with real survival scores, and the remaining two used the log-rank statistical test to determine that the predicted survival of the groups had a statistically significantly different distribution.

The methods for survival prediction were diverse and ranged in complexity. Four of the ten works (40%) generated a score by assigning the patient-level label to each tile in the whole slide image and then training the network to match that score, either by subsampling tiles and directly predicting survival probability or outputting tile-level predictions across the whole slide. These tile-level predictions were then aggregated by averaging or machine learning (Mobadersany et al. 2018; Skrede et al. 2020; Wulczyn et al. 2020; Laleh et al. 2021). Two of the ten projects (20%) directly predicted a prognostic biomarker to infer survival. Kather et al. (2019), predicted a deep stroma score, determined using the neuron activations of a network trained on tissue decomposition over the WSI, and Brockmoeller et al. (2022) assigned patient level LNM scores to each tile in their WSI and trained the network to reproduce the scores, the proportion of the tile predictions per WSI was then used to make the final biomarker prediction. The research of Jiang et al. (2020) performed classification tissue type per tile in the cohort WSI and then used gradient-boosted decision trees to predict survival.

The three remaining works are the most interesting and promising in terms of interpretability. The first is the work by Bychkov et al. (2018), which used a pre-trained convolutional neural network to extract feature representations from the tiles of the H&E whole slide images. Then, they used a long short-term memory network (LSTM) to operate on those vectors, representing the whole slide image as a sequence of them to make their final survival prediction. This is novel as it allows the sequence of the observed features to be considered, unlike the previously mentioned methods where this information is not considered and is lost. However, one downside of this approach is that the relationships learned by an LSTM cannot be directly queried, so its final output predictions are not directly interpretable. The following method by Li et al. (2022) uses the Xception CNN network to perform feature extraction on the H&E tiles. Then, multiple instance learning (MIL) is used to assign patient-level labels to the tiles from whole slide images as a bag; the MIL network can then be used to predict survival from the extracted features. The benefits of this approach are that the MIL network can learn to distinguish which tiles contribute most to the prediction and is, therefore, much more interpretable. Although only one attention weight per tile can be extracted and visualised, it provides no information on how they relate. The final work by Tsai et al. (2023) was the most interpretable. It used a ResNet50 network in combination with a vision transformer to extract features and generate tile clusters. These clusters were then used in a multiple instance learning network to make the final prediction. This is ripe for analysis and interpretability as clusters can be assigned to each input tile and correlated with poor outcomes, the most prognostic of which can be examined for prominent histological features that correlate with survival, known or unknown.

The work of Tsai et al. (2023) used transformers to assign tiles to individual clusters by reading the ResNet50 representation vector, but it missed the primary benefit of transformers: their ability to consider how each item in a sequence globally relates to every other to make the final prediction, this is described in section 2.7.6. This is ideal for whole slide images, where the global positioning of features and how they relate are significant. We proposed that a CNN network could be trained to extract feature vector representations of WSI tiles, as their strength lies in recognising local features that are semantically relevant. These representations could then be presented to a transformer network as a sequence representing the whole slide or a target portion, building on the work of Bychkov et al. (2018). As transformers have an attention matrix that can be queried, with weights of how each item in a sequence attends with every other, it could understand the global significance of how tissue types and histological features relate with each other and how they are important to the final prediction. These weights can be extracted and visualised to highlight histological features relevant to prognosis, potentially revealing new information. The aim was that this method would result in more accurate and interpretable results. This implementation and results of this proposed approach shall be explored in chapter 5.

Chapter 3

Ensuring Accurate Stain Reproduction in Deep Generative Networks for Virtual Immunohistochemistry

3.1 Introduction

This chapter presents the contributions of this work toward ensuring accurate staining and structural reproduction in virtual stain translation using deep generative networks. The overarching aim of this project was to investigate methods for automated aggression prediction in colorectal cancer or otherwise assist pathologists in assessing colorectal cancer biopsies. As discussed in section 2.3 and 2.4, there is a histological phenomenon called tumour budding, which is defined as the migration of single cells or small tumour clusters outwards from the core into the surrounding tissue (Mitrovic et al. 2021). The density of these cells in hotspots within the invasive margin of colorectal cancers is a recognised biomarker for independent survival prediction (Mitrovic et al. 2021). The existence of an independent biomarker for survival is significant as there is heterogeneity in the survival of stage II colorectal cancer patients due to the camouflaging of aggressive tumour phenotypes during staging and diagnosis resulting from the current TNM guidelines (Lugli et al. 2017). Patient outcomes could be improved by including tumour bud scoring in standard reporting (Koelzer et al. 2016). However, assessment of tumour budding on standard H&E stained slides can be challenging as it indiscriminately stains both tumour and non-tumour cells (Koelzer et al. 2015). This is particularly true in cases with ambiguous histology or poor tumour differentiation as the pattern-matching ability of humans can have difficulty distinguishing the subtle changes in colouring (Lugli et al. 2017). In cases like this, other techniques, such as immunohistochemistry, are often applied to provide further contrast (Puppa et al. 2012). The enhanced contrast of IHC, usually a pan-cytokeratin antibody that binds to tumour-derived epithelial cells like AE1/AE3, vastly improves the reproducibility, accuracy and speed of manual tumour bud scoring by pathologists (Martinez Ciarpaglini et al. 2019; Karamitopoulou et al. 2013). However, physical IHC requires complex equipment, expert personnel and increased time to produce (Kai et al. 2016). This suggests that the field would benefit from the development of a virtual method of translating H&E to IHC pan-cytokeratin AE1/AE3. No such method existed, but as discussed in 2.8, other work in virtual staining suggested it was possible. Therefore, the first goal of this project was to create a method to restain whole slide H&E colorectal cancer images to IHC virtually. It was to serve a dual purpose in being a useful independent tool for diagnosis and as a technique for producing a reliable ground truth dataset of tumour buds for later use in training automated scoring systems.

The proposed virtual staining method had to fulfil specific requirements to be valid. The utilised deep network had to be structurally accurate and promote faithful reproduction of the desired stain. This meant that an architecture and loss function had to be developed that upheld these standards, and a training dataset of paired H&E and IHC whole slide images had to be created that were aligned down to the pixel level. This was a non-trivial task. This chapter presents the methods, results, and a discussion of the developed approach to fulfil these requirements and how it compares to other methodologies in the field.

3.2 Materials and Methods

3.2.1 Dataset

We required a ground truth dataset of structurally paired H&E and IHC AE1/AE3 slides to train a neural network for stain translation. No dataset was publicly available with the desired characteristics; therefore, we had to generate one in-house. This was done by preparing consecutive serial sections with the minimum feasible separation of 2.5 microns to ensure the tissue structure was consistent between slices. The first of the paired sections was stained in H&E, and the second with our target IHC stain, pan-cytokeratin AE1/AE3.

The Glasgow Tissue Research Facility (GTRF) and the NHSGCC Biorepository (ethics no 22/ed/0207) supplied the source tissue. The slides were fully anonymised, and the metadata was removed before use in this research. The Edwards group in the Institute of Cancer Sciences at the University of Glasgow cut pairs of serial sections at 2.5-micron intervals from eight CRC resections, resulting in sixteen total slides. The Edwards group and histology services at the Cancer Research UK Scotland Institute carried out the H&E and IHC staining, respectively. Prepared slides were scanned with a Hamamatsu S60 slide scanner with an LED light source at 20x magnification.

To test and compare the performance of the model, separate test data was also required. Therefore, two additional slide pairs were cut and stained using the same process as above but performed on a different day and in a different lab to ensure the stain preparation was unique. This time, the Glasgow Tissue Research Facility cut the sections and carried out H&E staining, and members of the Edwards group performed the AE1/AE3 IHC staining. This was done to simulate real variations in stain preparation. These slide pairs were reserved during training and only used to generate the final performance metrics for the trained models.

The raw dataset had to meet three standards to be suitable for training a deep network for virtual stain translation. The first was that the paired H&E and AE1/AE3 slides had to have the tissue structure matched between sections, accurate to the cellular level. The second requirement was to implement a protocol to ensure that the training, validation, test and future inference data belong to the same colour domain as the training data. This was required to assist model generalisability. The third and final requirement was to filter the training data to remove artefacts resulting from the staining process, namely deformed, folded or missing tissue and areas of overstaining. A figure demonstrating the paired slides and some of their challenges is available in figure 3.1. The methods and protocols developed to meet these requirements are provided in sections 3.2.2 and 3.2.3.

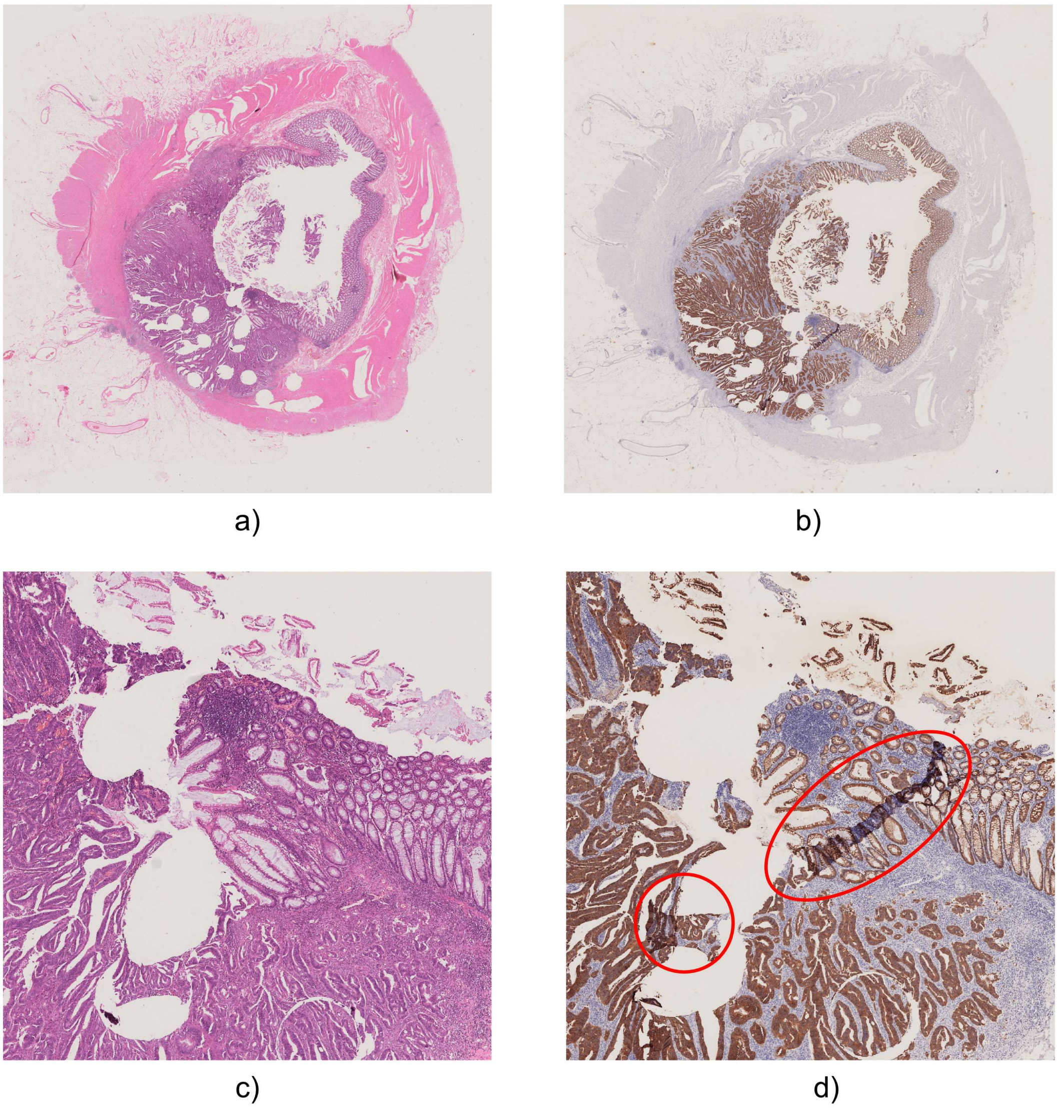


Figure 3.1: This figure depicts a representative whole slide image stained with haematoxylin and eosin in panel a), juxtaposed with its serial section stained using immunohistochemistry for AE1/AE3 in panel b). Both images originate from the same patient and represent consecutive serial sections of a tissue block. Panels c) and d) magnify specific regions of the slides where tissue exhibits folding artefacts during the staining and imaging process (demarcated in red). Such regions are not alignable and necessitate exclusion from the training dataset.

3.2.2 Stain Normalisation

Stain normalisation is a common requirement of deep learning on histopathological whole slide images, as discussed in section 2.5. Several open-source tools exist for stain normalisation but tend to work at the tile level only and often lack stain augmentation techniques and the implementation of data structures that support whole slide images and efficient parallel processing. Therefore, we decided to create a toolkit to fulfil these requirements for use across this project, in later internal projects and for subsequent open-source release and publication.

The most popular and best-performing stain normalisation techniques were reviewed. Their theory is discussed in section 2.5. The toolkit was developed to be modular to allow the later inclusion of new normalisation techniques by extending an abstract base class. As it exists now, it implements methods for whole slide and tile normalisation with the Reinhard et al. (2001), Macenko et al. (2009), and Vahadane et al. (2016) normalisation techniques, it is known as the Beatson Augmentation and Stain Normalisation Toolkit, or “BEAST”.

Modular functions were designed for whole slide file reading to enable later replacement or extension with libraries for new or custom formats. The default function attempts to read slides using the OpenSlide-Python toolkit by Goode et al. (2013), an excellent and versatile library for working with whole slide images. It has a limitation, however, that read images are returned as Python image library data structures. This means that they are limited by Python array addressing, which is 32-bit. No image can contain more than 2^{32} pixels. Images larger than this can be expected in colorectal slide scans, as they can be substantial multi-gigapixel images. The BEAST toolkit automatically circumvents this limitation when unmanageable dimensions are detected. The toolkit will tessellate the slide into a set of smaller coordinates and then incrementally load those into a NumPy array (which uses 64-bit addressing). If there is no desire to work with the entire slide image, the user can specify that the tessellated tiles can be directly normalised or augmented and written out independently. This allows the toolkit to work in environments with memory constraints. The toolkit can also read slides using the Tifffile library by Gohlke (2022), which can directly return whole slide images as NumPy arrays. This alternative is specifiable by command line option or automatically used as a failover when the required system libraries aren’t available for OpenSlide.

An extensible class was written to output whole slide images as Open Microscopy Environment tiff (.ome.tiff) files (Besson et al. 2019). This function queries the input OpenSlide or Tifffile metadata tiff tags to find the X and Y pixel resolution. It creates a pyramidal ome.tiff image that includes the necessary metadata for the resulting file to be displayed accurately with the correct pixel resolution and scale in common whole slide image viewers.

Some of the stain normalisation protocols can be computationally intensive. Therefore, the toolkit was written to use parallel processing by default. The number of threads can be stated as a command line option, and the toolkit will use the Python standard library's multiprocessing module to process the whole slide tiles in parallel for quick and efficient stain normalisation.

Finally, a set of utilities for stain normalisation and the training of deep neural networks were included in the toolkit; these tools can augment the stain colour profile through random perturbations to the stain intensity when translated into its stain colour space. This allows for the creation of a more diverse set of stain representations to train deep networks. It also includes a utility to estimate a stain matrix over a group of whole slide images or tiles, allowing users to compute an idealised target normalisation stain component matrix using a user-specified heuristic.

The code for the toolkit is available here: Github – Beatson Augmentation and Stain Normalisation Toolkit for viewing or extension. It is fully documented using the Sphinx documentation system and can be easily installed as a Python package for general use.

3.2.3 Image Registration, Alignment and Post-Processing

To prepare a dataset for training a deep network, as discussed above, the input data had to have paired tissue in the source and target stains that were suitable to train a deep network in a supervised manner for stain translation. Structural inconsistencies had to be minimised in the dataset, as the deep network could learn to reproduce those if not otherwise constrained. This was a complex task as the slides were serial sections stained on different days in different labs. This meant that the slides had to be stain normalised to ensure consistency and later generalisation of the deep network and be registered and aligned to ensure the tissue structure matched at the pixel level. There were many open-source pathology registration and alignment tools, but most focused on MRI and CT scans. However, there were several like Elastix (Klein et al. 2010), QuPath (Bankhead et al. 2017) and VALIS (Gatenbee et al. 2021) that were designed to work with whole slide histopathology images. We evaluated several of the existing end-to-end packages, and they were very successful at approximately aligning the slides. However, none could effectively align the tissue to the pixel level, mainly due to large deformations or missing tissue between slides. Therefore, we had to develop an automated pipeline to account for this. We proposed that several levels of registration alignment and filtering would resolve this by reducing the alignment area and, therefore, the scale of deformation that must be addressed in each local area as the magnification level increases. Additionally, at each iteration, the quality of the fit could be incrementally improved for the initialisation of the subsequent round. Our proposed process is shown in figure 3.2, and a detailed description of each step is given below.

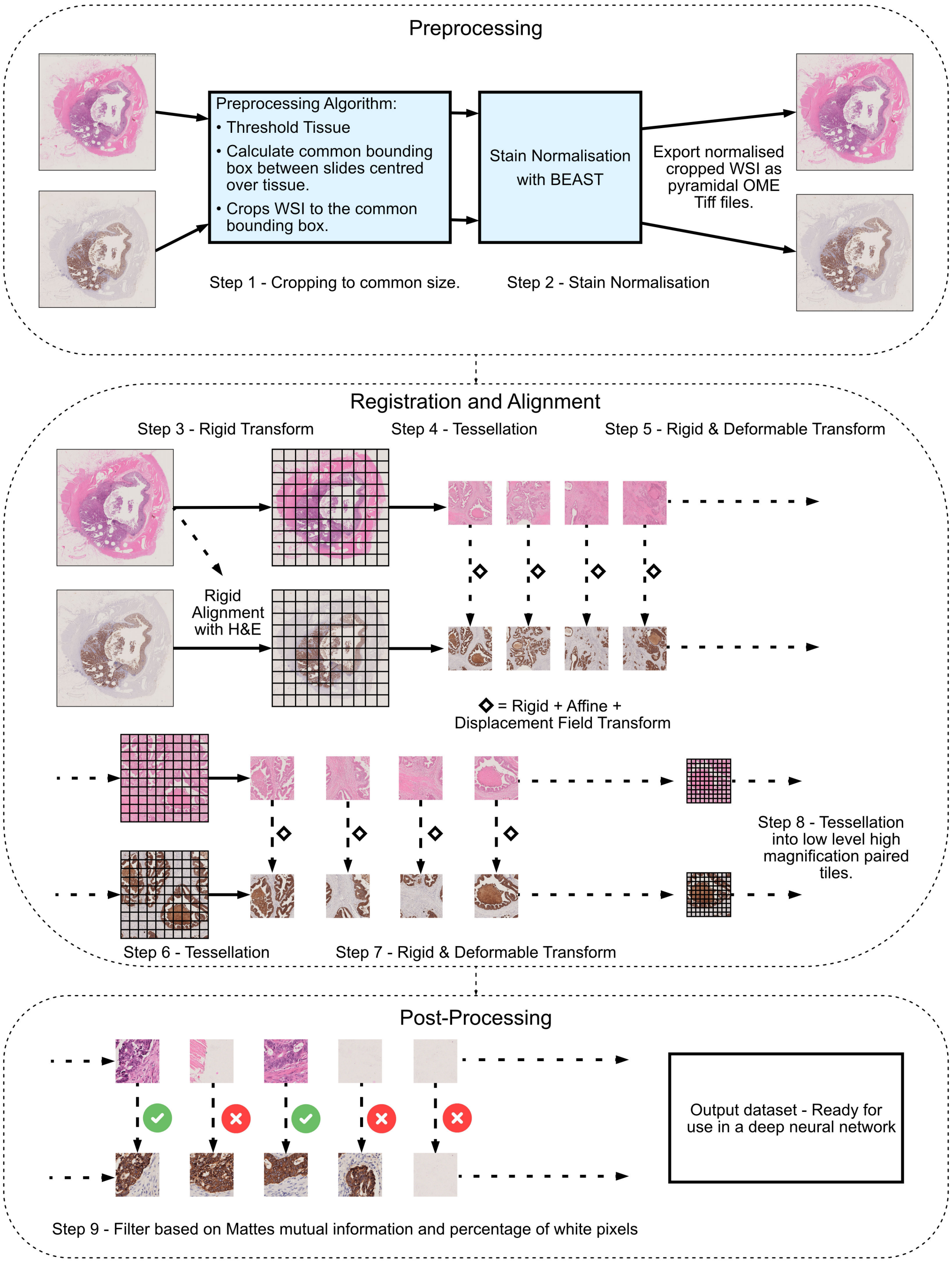

Preprocessing
Preprocessing Algorithm:
• Threshold Tissue
• Calculate common bounding box between slides centred over tissue.
• Crops WSI to the common bounding box.
Stain Normalisation with BEAST
Export normalised cropped WSI as pyramidal OME Tiff files.
Step 1 - Cropping to common size.
Step 2 - Stain Normalisation
Registration and Alignment
Step 3 - Rigid Transform
Step 4 - Tessellation
Step 5 - Rigid & Deformable Transform
Rigid Alignment with H&E
◇ = Rigid + Affine + Displacement Field Transform
Step 8 - Tessellation into low level high magnification paired tiles.
Step 6 - Tessellation
Step 7 - Rigid & Deformable Transform
Post-Processing
Output dataset - Ready for use in a deep neural network
Step 9 - Filter based on Mattes mutual information and percentage of white pixels

Figure 3.2: Our proposed method for creating a paired H&E-IHC dataset. The initial step is to cut serial sections from the same tissue block. The first section is stained in H&E, and the second is stained in IHC AE1/AE3. The sections are then scanned, cropped, stain normalised, rigidly registered and resampled to align with each other. The scanned whole slide image is then tessellated into smaller tiles then rigid, affine and deformable displacement field transforms are used to align them further. This tessellation and alignment process is repeated a second time on the aligned tiles to further align the highest magnification to the pixel level. The resulting tiles are then filtered using the Mattes mutual information metric to remove artefacts, areas of overstaining and tiles that do not align. Tiles are also filtered if they contain too much background in one or both slides. A few thousand background tiles are reintroduced to the dataset to ensure the network can handle background translation. The resulting set of tiles was then ready for use in training a deep network for virtual stain translation.

After evaluating the effectiveness of common registration libraries, we chose a low-level one called the Image Registration Toolkit (ITK) (McCormick et al. 2014) to build our registration framework. This was because it was equipped with mutual information metrics, the only image similarity metrics we found that could successfully align the different staining modalities. It was also written in C++ and ran natively without sandboxing or virtual machines. This allowed it full access to the system memory and 64-bit array addressing, and it could handle large whole slide images efficiently. Additionally, it had a neatly documented Python interface called SimpleITK (Lowekamp et al. 2013) that allowed integration using Python glue code with other common deep learning frameworks used in this work. This toolkit was used to create our multi-scale image registration method.

The first stage in our pipeline was to preprocess the whole slide images. In step one, we wrote a program using the OpenCV Python library (Culjak et al. 2012) to threshold the image in the LAB colour space to separate the tissue from the slide background. The SimpleITK toolkit uses a proprietary image coordinate system that requires input registration images to be identical in their dimensions. Therefore, we wrote an algorithm to calculate the dimensions of a common bounding box for the tissue in both slides. Our algorithm would then centre the bounding box on the centroid of the detected tissue polygon on each slide to ensure that all tissue was visible. Finally, our algorithm would crop the whole slide image to the dimensions of the calculated bounding box so both would be identically sized. If desired, the whole slide images could then be written out to file ready for step two. However, we were fortunate to have access to a system that could simultaneously hold both uncompressed images in RAM. Therefore, for efficiency, in step two, we directly stain normalised the cropped images using our previously developed Beatson Augmentation and Stain Normalisation Toolkit (BEAST). We had previously calculated idealised stain colour component matrix values for H&E slides and IHC AE1/AE3 slides. This was done using the BEAST toolkit to retrieve the median values for the stain components in their respective stain intensity channels from a set of 10 target slides from an unrelated dataset. Each slide was selected as having favourable staining by a translational cancer pathologist. Once the slides were normalised, they would be output to a full-resolution uncompressed OME.tiff file following the format and metadata standards set by Besson et al. (2019). This concluded the preprocessing stage.

The next stage in our pipeline was multi-resolution registration and alignment of the whole slide images. In step three of our process, we used simple ITK to rigidly align the full-resolution whole slide images through a Similarity 2D transform with the Mattes mutual information similarity metric. We used a random pixel sampling strategy, sampling 35% of the pixels and using 75 bins for the histograms to calculate the metric. We used a learning rate of 1 over 1000 iterations with a downsampling factor of 4 and a smoothing factor of 4. The gradient tolerance was set to 1e-4 to halt alignment when the calculated steps dropped below that tolerance level. These settings were arrived at through a parameter sweep using a custom testing framework. We used this framework on a few selected examples of whole slide images. We set up a series of brute force tests to alter the settings over a range of predefined values iteratively. It would record the resulting image similarity score and time taken to align. The optimal settings were balanced between peak similarity scores and unreasonable alignment times. Any settings that resulted in alignments that took more than several hours were deemed unsuccessful. Indeed, many would run for multiple days, even on a 128-core Intel Xeon server with 512GB of RAM. We will provide the performance outcomes from a variety of parameter combinations in the results section of this chapter.

Once the whole slide images were aligned using the rigid method in the previous step, we split the slide into smaller tiles for further registration and alignment in step four. The size of these tiles had to be very carefully chosen. Careful consideration was necessary for several reasons; the main concern was that the edge of the image tends to become distorted or otherwise have artefacts introduced during alignment. Therefore, for every tile, we had to include a boundary that would be excluded from the final dataset. This boundary also meant that the extracted tiles had to overlap to ensure that all the pixels would be available in the resulting dataset and none excluded in the border. We also had to think about how the lowest-level and highest-magnification tiles would be extracted from preceding tiles and how their dimensions would affect the amount of overlap and size of the final output tiles for training. Figure 3.3 gives an overview of our proposed custom algorithm for determining the tile dimensions for extraction. It was known that the tiles were destined for use in the Pix2Pix and CycleGAN networks that would have a residual backbone in the encoder. The ResNet family of models has architectures designed to work with tiles of size 224^2. Therefore, this was our base resolution. For the second highest magnification of tiles, we decided to make the tile size a multiple of ten. So, the base tile size was 2240 plus a buffer of 1000 on each edge, making the resulting tile 4240^2. During the tessellation of the lowest magnification tiles, we decided to quadruple the interior dimension extracted in the highest magnification. This made the interior window four times as large. Therefore, the tile had an internal window size of 8960 plus the buffer of 1000 pixels on each edge, resulting in a starting tessellation size of 10960 with a stride of 8960. This configuration and choice of stride ensured that the buffer zones never overlapped, so no tissue was excluded from training, and that there was always a region of 1000 pixels around an aligned tile that was not included in the training data, where alignment artefacts could occur without impact. The relationship between output tile size and the previous level is shown in 3.1. The stride was 224 or 448 at the lowest level to output dataset tiles, or the stride was equal to the previous useable tissue window, which is given by the previous stride multiplied by the number of tiles extracted along a dimension. The buffer size can be variable, but 1000 worked well in our initial evaluation and was taken forward for the rest of the project. Our choice to only have three levels of alignment, with an upper tile size of 10960, stems from the fact that in a colorectal whole slide image, this is about 1/5 to 1/10 of the dimension of the average image. This was a good size for local registration and alignment to combat tissue deformation. The 1000-pixel gap at the edge of the tiles at this size also tended to lie in the background at the edge of the slides.

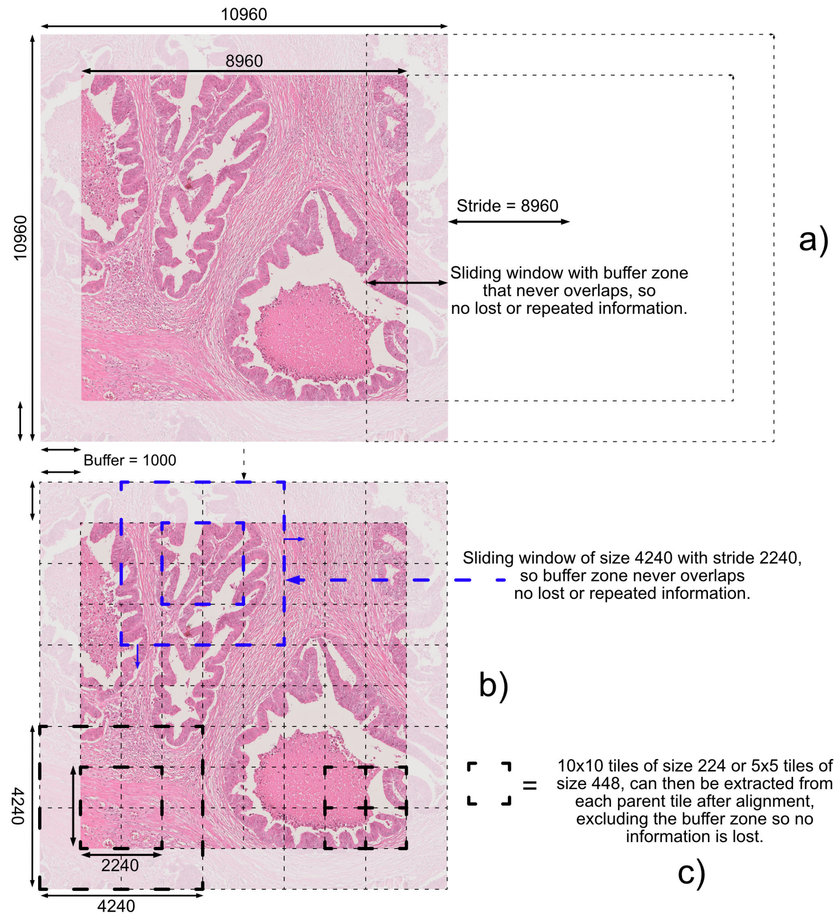

10960
8960
10960
Stride = 8960
Sliding window with buffer zone that never overlaps, so no lost or repeated information.
a)
Buffer = 1000
Sliding window of size 4240 with stride 2240, so buffer zone never overlaps no lost or repeated information.
b)
10x10 tiles of size 224 or 5x5 tiles of size 448, can then be extracted from each parent tile after alignment, excluding the buffer zone so no information is lost.
=
c)
4240
2240
4240

Figure 3.3: Our proposed method of WSI tessellation, first as shown in a) the WSI is tessellated into tiles of size 10960, using a sliding window with a stride of 8960. This leaves a non-overlapping buffer of 1000 pixels at the edge of the tiles to exclude alignment artefacts. Then after a round of registration and deformable alignment, in b) the tiles from the previous step are tessellated into tiles of size 4240 using a sliding window with stride 2240, again leaving a non-overlapping buffer of 1000 pixels at the edge of the tiles to exclude alignment artefacts. Finally, in c), the tiles from the previous step are tessellated into tiles of size 224 or 448 as required by the downstream task.

$$\text{tile dimensions} = (n \times \text{stride dimensions}) + 2 \times \text{buffer dimensions} \tag{3.1}$$

Once the whole slide image had been tessellated, the first round of deformable image registration and alignment was carried out on the tiles of size 10960 in step five. This was once again performed with the SimpleITK framework. This time, there were three rounds of registration and alignment on each tile. We configured a chained transform. The first step was a rigid alignment that was first performed as before. Then, an affine transform scaled and skewed the images and was initialised from the rigid alignment. Then, a fully deformable displacement field transform was initialised from the affine alignment. This allowed the tissue pixels to be transformed to match all but the most severe deformations. We again used Mattes mutual information as the image similarity metric, with the same parameters as above for the pixel sampling method, sample percentage, and number of histogram bins. This time, we also set a learning rate of 1 and a gradient tolerance of 1e-4 over 1000 iterations. This was superior as determined by our testing framework.

In step six, we repeated the tessellation process on the tiles of size 10960^2, resulting in tiles of size 4240^2. Then, in step seven, we repeated the rigid, affine and displacement field transform on the tiles of size 4240^2. This time however, we increased the domain mesh size to 20^2, which meant there were 400 points overlayed on the image that moved to deform pixels and tissue of the AE1/AE3 patch to conform to the H&E. We also made use of the shrink factor setting, first to perform the rigid and affine transforms at a downsampling factor of 4, then 2, and 1 to push the tissue into the correct position iteratively. This was done on images smoothed with a Gaussian kernel of size two at each level. Then, in the last part of the composite transform, we again used a displacement field transform, random pixel sampling, and 75 histogram bins, sampling 35% of the pixels with a learning rate of 1, a run length of 1000 iterations and a gradient tolerance of 1e-4. This was done pyramidally with shrink factors 4, 2 and 1. But this time, so that the highest detail was preserved, the smoothing kernel size was set to 2 at shrink factor 4, then 1 at shrink factor 2 and finally, no smoothing was done on level 1. This was run for 1000 iterations with a gradient tolerance of 1e-4. These parameters were determined by a series of brute force tests, sweeping over a range of values using our test framework and a subset of the tiles. The results of this sweep will be provided in a later section.

Then, in step eight, we split our aligned tiles into the final size as training requires. The last step, number nine, and our post-processing step was a filtering process. We first determined the percentage of white pixels in a tile pair. That tile was dropped if the percentage in either domain was over 90%. We copied back 2000 tiles with more than 99.9% white pixels to train the network to translate the background. Finally, we used the Mattes mutual information metric to compare all pairs of output tiles. From manual observation, we determined a minimal cutoff for a Mattes metric score that resulted in an acceptable alignment between the tiles; a cutoff of 0.15 was used. We also calculated the percentage of pixels with the desired IHC chromogen stain on each tile pair. This was written to a CSV file with the tile identifier and IHC pixel stain percentage. This allowed us to restrict the balance of tiles with an IHC stain shown to the network later to improve training performance.

Once these steps were carried out, we had a paired collection of H&E and IHC tiles containing 96,719 tiles of size 224^2 and 19,791 tiles of size 448^2. These had the tissue aligned at the pixel level filtered to remove artefacts and extraneous amounts of background pixels. The calculated IHC stain metadata allowed versatility in choosing tiles shown to a network and the minimum percentage of IHC staining they should contain. Our dataset now met the required parameters for training a deep neural network for stain translation in a supervised manner.

3.2.4 Deep Network Architectures and Training

U-Net

As an initial proof of concept, we decided to test if it was possible to infer a pan-cytokeratin AE1/AE3 IHC marker from H&E using a simple U-Net. The architecture was implemented as described in section 2.7.2. We developed the model using Python in the TensorFlow library, version 2 (Abadi et al. 2016). We used a learning rate of 2e-4, the Adam optimiser (Kingma et al. 2017). We normalised the input patches to a range of -1 to 1 and used the tanh activation function as described in equation 2.5. The loss criterion was a simple mean squared error. We used the MobileNetV2 (Sandler et al. 2019) network as the backbone of the U-Net. It had five downsampling and upsampling layers, reducing the feature output map to a size of 14^2 from the input size of 224^2 before recreating an RGB image of the same dimension.

To evaluate the utility of the staining in the generated images, we inferred virtual AE1/AE3 slides from 186 slides in a test dataset with known tumour bud scores generated in H&E. Then we had a pathologist produce a tumour bud score again using our virtual pan-cytokeratin slides and the methodology as described in algorithm 1.

Pix2Pix GAN

Once we established proof of concept, we attempted to generate more realistic images with a Generative Adversarial Network (GAN). This was to be used as a baseline performance measure for later more advanced methods and as validation that photo-realistic image generation was possible using our prepared dataset. A Pix2Pix GAN was chosen as the architecture as it was the second most popular network in the virtual staining literature and comparatively quick to train compared to the CycleGAN. The architecture was implemented as described in section 2.7.3. We designed a custom testing framework using the Pytorch (Paszke et al. 2019) Python library. Our framework accepts command line parameters using the Python standard library Argparse module. It allows the user to specify values such as learning, rate, momentum, optimiser type, lambda value, Mattes threshold value, weight initialisation type, and Boolean values to perform stain and colour augmentation, spatial augmentation, and whether to use dropout.

All experiments were evaluated by training for 50 epochs with a batch size 16. Our models were trained on Nvidia A6000 GPUs. Batch size 16 was the largest they could support with a ResNet or U-Net backbone and a maximum input tile size of 448^2. We performed colour and spatial augmentation of the input tiles using the Albumentations Python library (Buslaev et al. 2020). We performed spatial and colour augmentations to improve the generalisability of our model. The chosen spatial augmentations were one of a vertical flip with a probability of 0.5, a horizontal flip with a probability of 0.5, and a random rotation of 90 degrees with a probability of 0.5. Then, a random resize and crop to our target tile size. The colour augmentations we performed were a composition of colour jitter with a probability of 0.8, brightness range of 0.4, contrast range of 0.2, saturation range of 0.4 and hue range of 0.1. This was followed by a random gamma adjustment with a probability of 0.2 and simulated image compression with a probability of 0.2 and a compression ratio of 60%. This was to simulate the jpeg artefacts that can be introduced from the compression during image storage on certain slide scanners. The final use of the Albumentations library was to normalise the resulting tensors to a range of -1 to 1. We used a learning rate of 2e-4 as default, with the Adam optimiser, with a beta1 value of 0.5. We had a cosine decay learning rate schedule and used batch normalisation on the convolutional layers.

The loss function of the generator network was a weighted combination of L1 loss on the generated and real target image, along with the network's ability to fool the discriminator, as described in detail in section 2.7.3. The loss function of the discriminator network was based on its ability to detect the actual versus generated target images, as described in 2.7.3.

The structural similarity index measure (SSIM) was the most popular image evaluation metric in the virtual staining literature. Therefore, using this metric, our framework was designed to evaluate and record the generated validation images automatically. The details of SSIM are described below in section 3.2.5. Many virtual staining papers discuss the need for an image evaluation metric more in line with human visual perception. Interestingly, none mentioned the Fréchet Inception Distance (FID) (Heusel et al. 2018). This is the most common metric for evaluating the quality of GAN-generated images in other deep-learning settings and works by calculating the distance between the distributions of features extracted from real and generated images. This method of feature-based comparison is more perceptual than SSIM (Benny et al. 2021). Therefore, this was included as an additional default evaluation metric in our testing framework. The details of this metric are described below in section 3.2.5. All reported metrics result from at least three repeats with identical architecture and hyperparameter settings.

CycleGAN

One of our primary concerns regarding the use of generative networks for stain translation was their tendency to hallucinate visual details in their output. The CycleGAN architecture was the most prevalent in the virtual staining literature due to its ability to encourage the preservation of structural elements during style translation tasks. Therefore, this architecture was also chosen for evaluation and was implemented as described in section 2.7.4. For the assessment of model performance, we used our testing framework as designed for the Pix2Pix GAN above, with the same adjustable parameters, but extended to include modifiable weights for the lambda1 and lambda2 values for the identity and cycle loss terms in the CycleGAN loss function.

All experiments were evaluated by training for 50 epochs with a batch size of 8. Our models were trained on Nvidia A6000 GPUs as before. A batch size of 8 was the largest they could support with a ResNet or U-Net backbone and a maximum input tile size of 448^2 due to the duplication of the generator and discriminator architectures in the CycleGAN design. We performed colour and spatial augmentation of the input tiles using the Albumentations Python library (Buslaev et al. 2020), with the same parameters we applied to the Pix2Pix GAN. We also normalised our input images to a range of -1 to 1 to make the data zero-centred to help with the vanishing gradient problem. We used a learning rate of 2e-4 as default, with an Adam optimiser with a beta1 of 0.5. The optimiser operated on chained parameters from both generators so they would be trained in lockstep, as in the original CycleGAN paper by Zhu et al. (2020). We had a cosine decay learning rate schedule and used batch normalisation on the convolutional layers.

This implementation of the CycleGAN was to provide a baseline of the performance of the original network design. Therefore, the loss is as described by Zhu et al. (2020). It is based on coupled generator networks and is a weighted linear combination of L1 error on the identity of a source and target image passed through their respective generators, along with the ability of each generator to fool its discriminator, and finally, the L1 error on the cyclic recovered image from each staining type. This loss is detailed in section 2.7.4. The loss function of each discriminator network was based on its ability to detect the actual versus generated target images, as described in 2.7.4.

The structural similarity index measure (SSIM) and Fréchet Inception Distance (FID) metrics were again used to evaluate the quality of generated images. The details of this metric are described below in section 3.2.5. All reported metrics result from at least three repeats with identical architectures and hyperparameter settings.

Proposed VIHC Network Architecture

Our proposed network architecture for virtually translating H&E to pan-cytokeratin AE1/AE3 is a CycleGAN with several modifications. The base architecture is similar to that introduced by Zhu et al. (2020), with two generator-discriminator pairs connected in a cycle. However, in our architecture, the generators have ResNet backbones and no skip connections. They are custom models with two downsampling layers, with 128 and 256 convolutional filters, kernel size of three and stride two, followed by nine residual convolutional blocks, kernel size three, stride one, and reflection padding, each with 256 filters, with residual connections described in section 2.7.2. Two upsampling layers follow, with the same configuration as the downsampling layers. All layers use batch normalisation. This is a modification to the original CycleGAN design. The batch size is increased to eight from one. This is used over instance normalisation to combat colour artefacts introduced during inference of whole slide images with a batch size of one. A ReLU activation function follows each layer; the final output activation function is tanh. The discriminator design is unchanged from the original CycleGAN paper. It is a three-layer PatchGAN discriminator with 64, 256 and 512 filters. Each layer has a kernel size of four and a stride of two with batch normalisation and a LeakyReLU function. Spectral normalisation, as described by Miyato et al. (2018), is applied to each convolutional layer in the discriminator to assist in network regularisation and to stabilise GAN training.

The CycleGAN network was intended to maintain structural detail by pairing two GANs, one to translate from one style to another and the next to recover the original image for structural comparison. Our proposed architecture follows this idea with a forward and backward stain translation cycle. The forward cycle, which we will refer to as state 1, is displayed in figure 3.4. The cycle translation process of state one is as follows: a batch of real H&E images is passed to the first generator-discriminator pair, the purpose of this pair is to create IHC images. The IHC generator accepts the H&E images as a condition and attempts to translate them to virtual IHC photorealistic images in our target stain. The IHC discriminator accepts a batch of real and generated IHC images. Its purpose is to distinguish the real images from the fake. The pair are trained adversarially and aim to improve at their chosen task. The generated images fed to the discriminator are randomly sampled from an image buffer that retains a history of the last fifty generated image batches. This is done to stabilise training. The batch of generated IHC images is passed to the next generator-discriminator pair, the purpose of which is to translate IHC images to H&E. This allows the recovery of an H&E image batch from the generated IHC that can then be compared to the original H&E batch and constrained by a chosen loss function to maintain the structure of the image throughout the cycle.

We shall refer to the backward cycle as state two, a diagram of which can be viewed in figure 3.5. It is configured as follows: a batch of real IHC images is passed to the first generator-discriminator pair, and this pair aims to create H&E images.

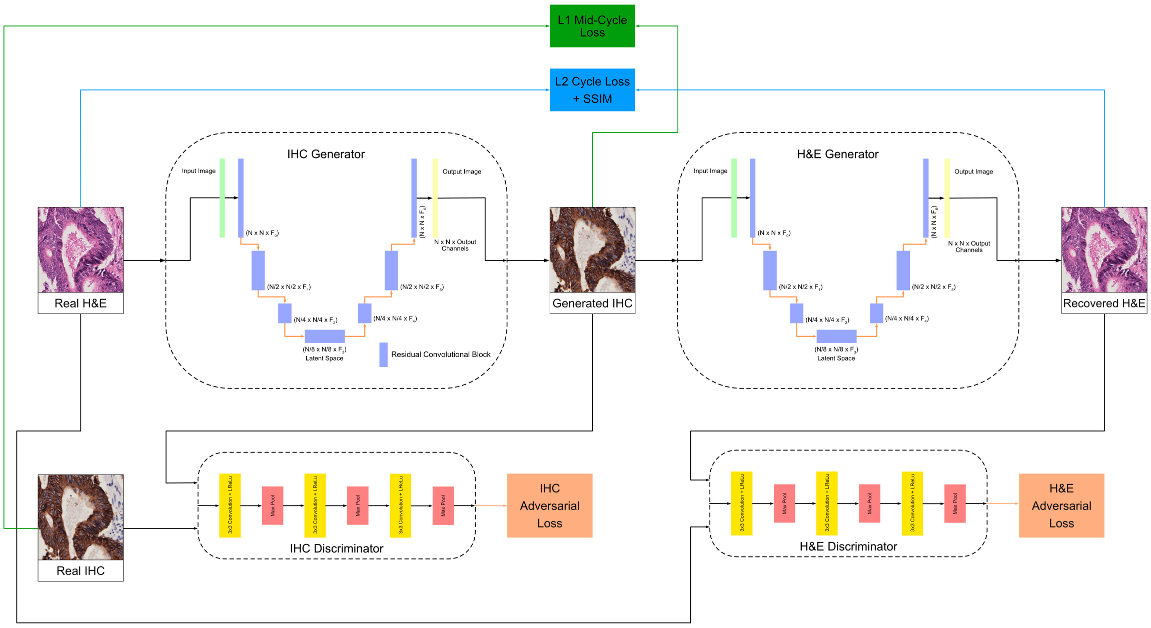


Figure 3.4: State 1 of our modified CycleGAN network. In this state, the network is configured to translate from H&E to IHC and recover the original H&E from the generated image. The original CycleGAN adversarial loss and cycle losses are retained. Two additional loss terms are included: the first of these is the structural similarity index measure (SSIM) loss between the real H&E and recovered image, called the cycle SSIM loss. The second is our proposed addition, the L1 norm of the error between the real and generated IHC images called the mid-cycle loss.

The H&E generator accepts the IHC images as a condition and attempts to translate them to photorealistic virtual H&E images. The H&E discriminator accepts a batch of real and generated H&E images. The generator-discriminator pair are trained as above. The generated images fed to the discriminator are again randomly sampled from an image buffer that retains a history of the last fifty generated image batches. The batch of generated H&E images is passed to the next generator-discriminator pair to translate H&E images to IHC. This allows the recovery of an IHC image batch from the generated H&E that can then be compared to the original IHC batch and constrained through a chosen loss function to maintain the structure of the image throughout the cycle.

A critical difference between the protocols for training our architecture and the CycleGAN architecture of Zhu et al. (2020) is that rather than a single optimiser operating on both generators and a single optimiser operating on both discriminators, we employ four, one for each sub-network of the model. This allowed the Adam optimiser to adjust its learning rate separately for each stain domain, as they are visually very distinct, and to stabilise training. In addition to decoupling optimiser parameters, our proposed architecture significantly differs in the loss function. A diagram depicting the loss applications can be viewed in 3.6. The adversarial loss of each generator-discriminator pair is only fed to the optimisers of that pair. We add SSIM loss to the recovered images in the cycle, which is applied to the optimisers of both generators. This allows them to work together to maintain the structure, as with the cycle loss.

The idea that drove this section of the project was to investigate how introducing paired data and a loss term to train the CycleGAN in a supervised manner would affect its capability for stain translation. The idea was to constrain the generated image to its paired real image. We proposed that this could be achieved by adding what we term, "mid-cycle" loss. It is an L1 loss term, in the middle of the cycle in each state, that operates between the generated and real image pairs. How this is implemented can be viewed in the diagrams of network configuration in figures 3.4 and 3.5. We shall now discuss the loss functions of the model in detail.

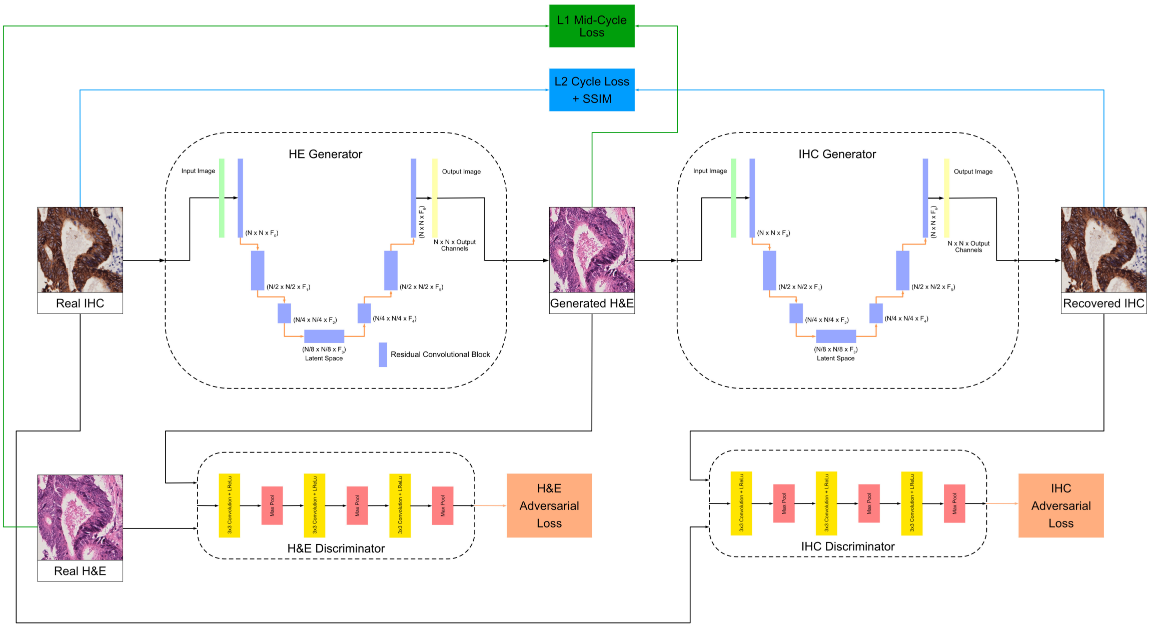


Figure 3.5: State 2 of our modified CycleGAN network. In this state, the network is configured to translate from IHC to H&E and recover the original IHC from the generated image. The original CycleGAN adversarial loss and cycle losses are retained. Two additional loss terms are included: the first of these is the structural similarity index measure (SSIM) loss between the real H&E and recovered image, called the cycle SSIM loss. The second is our proposed addition, the L1 norm of the error between the real and generated H&E images called the mid-cycle loss.

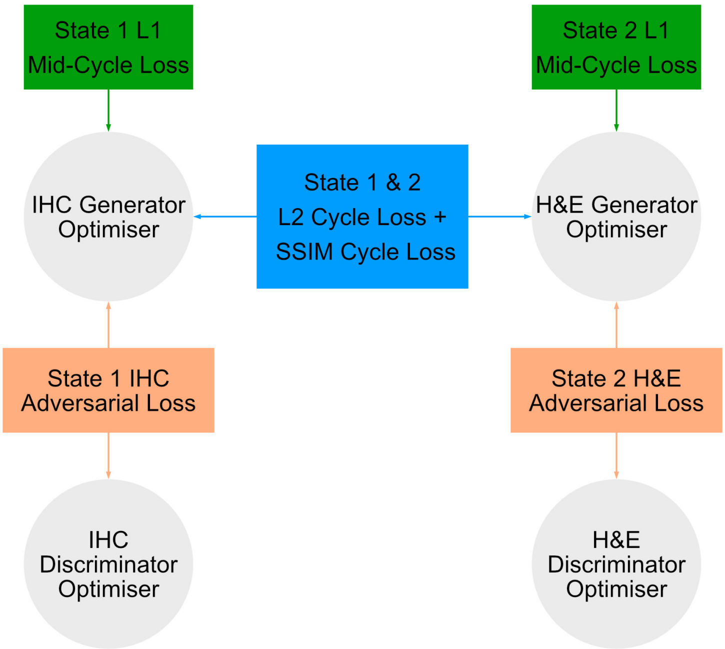


Figure 3.6: A critical element in the configuration of our network is that separate optimisers exist for each generator and discriminator. This was found to be key for the stability of the network during training. The adversarial and mid-cycle losses are specific to each generator-discriminator pair and are supplied only to the optimisers for those networks. However, the cycle loss and cycle SSIM loss are included in the loss calculation for both generator optimisers. This is crucial for their cooperation in learning to maintain the structure throughout the cycle.

The first component of our loss function is adversarial loss, as in the original CycleGAN architecture, which is applied to both generator-discriminator pairs separately. In the following, we express the IHC domain as domain Y and the H&E domain as domain X. The adversarial objective for the IHC generator, G_Y, is described in equation 3.2. Here, G_Y tries to generate photorealistic images in domain Y, using images from domain X as a condition. At the same time, the IHC discriminator D_Y is trained to distinguish between translated samples $G(x)$ and real samples from domain Y. Generator G_Y aims to minimise this objective against its adversary D_Y that tries to maximise it: $\min_{G_Y} \max_{D_Y} \mathcal{L}_{IHC}(G_Y, D_Y, X, Y)$. We similarly apply adversarial loss with a similar objective to the H&E generator and discriminator as expressed in equation 3.3, with the similar min-max objective: $\min_{G_X} \max_{D_X} \mathcal{L}_{H\&E}(G_X, D_X, X, Y)$.

$$\mathcal{L}_{G_Y}(G_Y, D_Y, X, Y) = \mathbb{E}_{y\ p_{data}(y)}[logD_y(y)] + \mathbb{E}_{x\ p_{data}(x)}[log(1 - D_Y(G_Y(x)))] \quad (3.2)$$

$$\mathcal{L}_{G_X}(G_X, D_X, X, Y) = \mathbb{E}_{x\ p_{data}(x)}[logD_X(x)] + \mathbb{E}_{y\ p_{data}(y)}[log(1 - D_X(G_X(y)))] \quad (3.3)$$

As in the original CycleGAN architecture, we constrain our stain translation to be cycle-consistent. For each image x from domain X, the cycle should be able to map it to domain Y and back to the original image x: $x \rightarrow G_Y(x) \rightarrow G_X(G_Y(x)) \approxeq x$. Zhu et al. (2020) called this forward cycle consistency; in our model, we refer to it as state one. Similarly, for each image y from domain Y, the cycle should map it to domain X and back to the original image y: $y \rightarrow G_X(y) \rightarrow G_Y(G_Y(y)) \approxeq y$, which was called backward cycle consistency, we refer to this as state two in our model. We alter the original cycle-consistency loss from an L1 loss to an L2 loss. Zhu et al. (2020) argued that L1 loss was superior to L2 because L2 encouraged blurring. However, our evaluation found that when combined with a cycle SSIM loss, L2 loss produced superior results. The cycle consistency loss was applied to both generator-discriminator pairs so they could work together to maintain structure throughout the cycle. The cycle-consistency objective of our model is expressed in equation 3.4.

$$\begin{aligned}\mathcal{L}_{cyc}(G_X, G_Y, X, Y) = \mathbb{E}_{x\ p_{data}(x)}[\|G_X(G_Y(x)) - x\|_2] \\ + \mathbb{E}_{y\ p_{data}(y)}[\|G_Y(G_X(y)) - y\|_2]\end{aligned} \tag{3.4}$$

After a review of the state of the art in virtual staining, we found that the best performing CycleGAN-based models included an SSIM term in their loss function, either across the mapping from $x \rightarrow y$ or on the cycled image $x \rightarrow G_Y(x) \rightarrow G_X(G_Y(x))$. Given that this part of the project aimed to investigate the training of GANs with paired data for virtual IHC and that we would constrain the generated IHC images, it was decided to apply the SSIM loss across the cycle. We used the TorchMetrics Python library (Detlefsen et al. 2022), which has an SSIM metric module and calculates the SSIM value as described below in section 3.2.5. The SSIM value ranges from -1 to 1, with increasing image similarity resulting in a score toward one. To modify it for loss optimisation, we subtracted it from one. The SSIM cycle consistency loss was applied to both generator-discriminator pairs to encourage them to work together to maintain structure throughout the cycle. The SSIM cycle-consistency objective of our model is expressed in equation 3.5.

$$\begin{aligned}L_{cycssim}(G_X, G_Y, X, Y) = (1 - \mathbb{E}_{x\ p_{data}(x)}[SSIM(G_X(G_Y(x)), x)]) \\ + (1 - \mathbb{E}_{y\ p_{data}(y)}[SSIM(G_Y(G_X(y)), y)])\end{aligned} \tag{3.5}$$

The final addition that we contribute to the CycleGAN model for virtual staining, trained in a supervised manner, was to include a term that compares the generated images in the middle of the cycle to their paired counterpart. To constrain $G_Y(x)$ and y, and $G_X(y)$ and x to be as similar as possible. The idea behind this is that when using consecutive paired serial sections, there will be small changes to morphological details between the sections, even when cut at 2.5 microns. This causes standard GANs like the Pix2Pix model to hallucinate structural details when trained using paired data. However, the CycleGAN is designed to maintain the structural detail from the cycle-consistency losses. Therefore, we can apply a supervised loss term to ensure the stain colour is reproduced accurately in the generated image without compromising the structure. We call this "mid-cycle" loss in the context of the CycleGAN. This loss is calculated and applied independently to each stain generator so they can freely learn to reproduce their target stain colour accurately. The expression of the mid-cycle loss objective is given for G_Y in equation 3.6, and for G_X in equation 3.7.

$$\mathcal{L}_{midcycY}(G_Y, X, Y) = \mathbb{E}_{(x,y)\ p_{data}(x,y)}[\|G_Y(x) - y\|_1] \tag{3.6}$$

$$\mathcal{L}_{midcycX}(G_X, X, Y) = \mathbb{E}_{(x,y)\ p_{data}(x,y)}[\|G_X(y) - x\|_1] \tag{3.7}$$

The full objective of our model is achieved by a linear combination of the adversarial loss of the IHC and H&E generator-discriminator pairs, the cycle-loss and SSIM cycle loss applied to both pairs and the mid-cycle loss of each pair. The full objective for the IHC generator-discriminator pair is expressed in equation 3.8, and the full objective for the H&E generator-discriminator pair is expression in equation 3.9. The parameters λ, α, and β in the loss function regulate the importance of different losses to the overall objective. In our best-performing model, $\lambda_1 and \lambda_2 = 10$, $\alpha_1 and \alpha_2 = 5$, and $\beta_1 and \beta_2 = 50$.

$$\begin{aligned}\mathcal{L}_{IHC}(G_X, G_Y, D_Y, X, Y) = {} & \mathcal{L}_{G_Y}(G_Y, D_Y, X, Y) \\ & + \lambda_1 \cdot \mathcal{L}_{cyc}(G_X, G_Y, X, Y) \\ & + \alpha_1 \cdot \mathcal{L}_{cycssim}(G_X, G_Y, X, Y) \\ & + \beta_1 \cdot \mathcal{L}_{midcycY}(G_Y, X, Y)\end{aligned} \tag{3.8}$$

$$\begin{aligned}\mathcal{L}_{H\&E}(G_X, G_Y, D_X, X, Y) = {} & \mathcal{L}_{G_X}(G_X, D_X, X, Y) \\ & + \lambda_2 \cdot \mathcal{L}_{cyc}(G_X, G_Y, X, Y) \\ & + \alpha_2 \cdot \mathcal{L}_{cycssim}(G_X, G_Y, X, Y) \\ & + \beta_2 \cdot \mathcal{L}_{midcycX}(G_X, X, Y)\end{aligned} \tag{3.9}$$

The optimal models for G_Y and G_X can be found by solving equation 3.10.

$$G_X^*, G_Y^* = \text{arg} \min_{G_X, G_Y} \max_{D_X, D_Y} \mathcal{L}_{IHC}(G_X, G_Y, D_Y, X, Y) + \mathcal{L}_{H\&E}(G_X, G_Y, D_X, X, Y) \tag{3.10}$$

3.2.5 Image Evaluation Metrics

The following metrics were used to evaluate the validation and test images produced by all models during this part of the project. We describe their implementation below.

Structural Similarity Index Measure

The Structural Similarity Index Measure (SSIM) was proposed by Wang et al. (2004). It is a popular metric for image quality assessment in deep learning (Horé et al. 2010). It quantifies the similarity between a reference and target image by considering luminance, contrast and structure. SSIM considers local and global differences in these aspects and aims to emulate human perception. It computes a value between -1 and 1, with higher values indicating greater similarity, making it a valuable tool in evaluating the perceptual quality of images, particularly in style translation scenarios (Lo et al. 2021; Chen et al. 2021b; Liu et al. 2022).

An expression representing how the SSIM value is computed between images is given in equation 3.11. Here, x and y are the two compared images, and μ_x and μ_y are the mean values of the images x and y, respectively. σ_x^2 and σ_y^2 are the variances of the images, and σ_{xy} is the covariance between images x and y. C_1 and C_2 are constants to avoid instability when the denominator is near zero.

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1) \cdot (2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1) \cdot (\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.11}$$

Fréchet Inception Distance

The Fréchet Inception Distance (FID), introduced by Heusel et al. (2018), is a popular metric to assess the quality and diversity of images generated by deep learning models (Zingman et al. 2023). It quantifies the similarity between the statistical distributions of real images and those produced by a generative model. FID combines two components: feature extraction using an Inception network to capture high-level image features and calculation of the Fréchet distance between them. The Fréchet distance is a measure of similarity between two Gaussian distributions. In this case, between the feature representations of real and generated images. A lower FID score indicates that the generated images are more like the real images in terms of their visual appearance and overall distribution of features. The FID score focuses on high-level features by the nature of its hierarchical Inception CNN used for feature extraction. This makes it a more perceptually comprehensive metric for assessing the quality and diversity of deep-learning images. The SSIM score primarily evaluates low-level structural similarities; therefore, paired with FID, it gives a comprehensive overview of the global structural and perceptual differences between images. The FID expression is given in equation 3.12. Where μ_1 and μ_2 represent the mean vectors of the Gaussian distributions of the feature embeddings of the real and generated images, and σ_1 and σ_2 are the covariance matrices of the Gaussian distributions corresponding to the real and generated image feature vectors. Tr refers to the linear algebra trace operation, which is the sum of elements along the main diagonal of a square matrix.

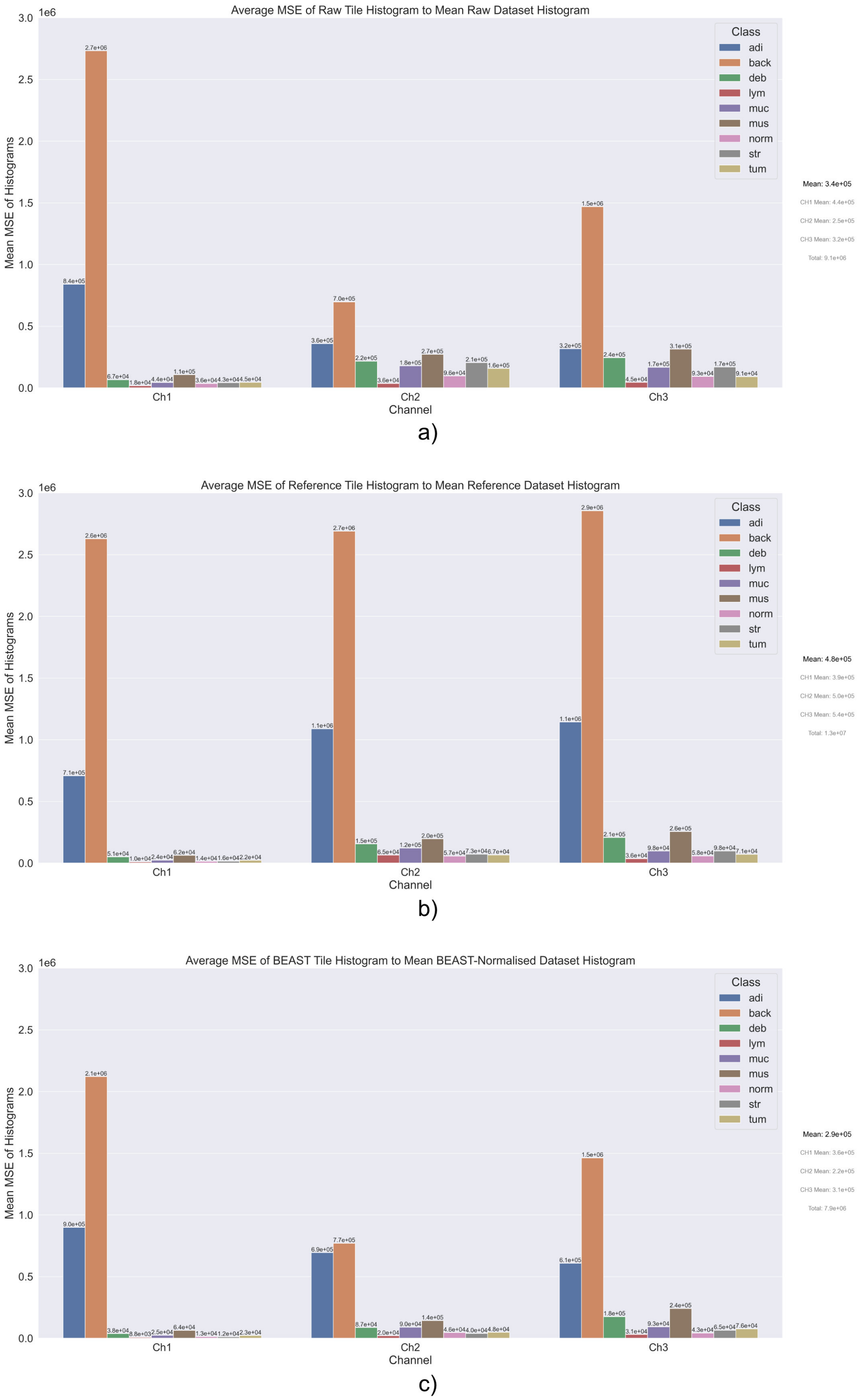
$$FID = \|\mu_1 - \mu_2\|^2 + \mathrm{Tr}(\sigma_1 + \sigma_2 - 2(\sigma_1\sigma_2)^{1/2}) \tag{3.12}$$

3.3 Results

In digital histopathology, the quality of a sample, in many terms but particularly in its staining and the representation of the tissue structure, plays a pivotal role in the precise interpretation and diagnosis of disease. If virtual stain translation is ever to have a place in a clinical setting, the translated images must reproduce both the desired staining and original structure accurately. The results presented in this section delve into the efficacy of using generative deep networks for virtual stain translation. We offer our findings on the structural accuracy and realism of the translated images and our method for improving their accuracy. We also present the results from our efforts to develop tools to normalise, register and align serial whole-slide images as required to create the necessary datasets to train deep networks for stain translation. The subsequent figures and tables will describe the performance metrics, compare our method with traditional approaches, and highlight the limitations and successes of our system.

3.3.1 Stain Normalisation

The first step in this project section was to develop a toolkit to normalise pathology images for stain translation. This was necessary as common normalisation toolkits do not carefully mitigate the introduction of background artefacts during normalisation, most likely because these artefacts are not as impactful when the tiles are used in everyday downstream tasks such as classification or prediction. However, when such tiles are used in image generation, the introduction of artefacts will significantly affect the quality and accuracy of the network output. The method of toolkit development was described in section 3.2.2. The results of the toolkit development are presented below.



Average MSE of Raw Tile Histogram to Mean Raw Dataset Histogram

a)

Average MSE of Reference Tile Histogram to Mean Reference Dataset Histogram

b)

Average MSE of BEAST Tile Histogram to Mean BEAST-Normalised Dataset Histogram

c)

Figure 3.7: Average MSE between a tile histogram and the mean histogram of the dataset for each channel in the LAB colour space. Panel a) shows the metrics for the raw dataset. Panel b) shows those of the same dataset, normalised with a common pathology toolkit, and panel c) shows the metrics for the same dataset, normalised with the BEAST toolkit.

The first result compares the average mean square error between a tile and the mean histogram of the entire dataset; this is shown in figure 3.7. Deep networks require that unseen data have a similar distribution to the training data to generalise well. Therefore, this figure intends to assess how similar the tiles are to the mean histogram of the dataset after normalisation to a target distribution in order to highlight significant discrepancies. It should provide a course description of the effectiveness of normalisation. The metrics were calculated on the NCT 100K dataset provided by Kather et al. (2018). The raw tiles were normalised using the well-regarded pathology toolkit: TIA toolbox by Pocock et al. (2022); this was our reference for normalisation performance. In the figure, we compare the average squared difference between a tile and the mean of the dataset for the raw data, shown in panel a), the identical tiles as normalised by the reference toolkit, and again as normalised by our Beatson Augmentation and Stain Normalisation Toolkit, BEAST. The main findings were that the reference toolkit increased the average mean squared difference in several classes, notably the background and adipose classes, as is displayed in panel b, compared to a. This was a result of the propensity of Vahadane and Macenko method to introduce background artefacts when little tissue is present from which to estimate an accurate stain component matrix. Due to careful filtering and selective normalisation of only pixel luminosity in the background pixels without staining, the BEAST toolkit vastly reduced the number and intensity of these background artefacts, as observed when comparing panels c to b.

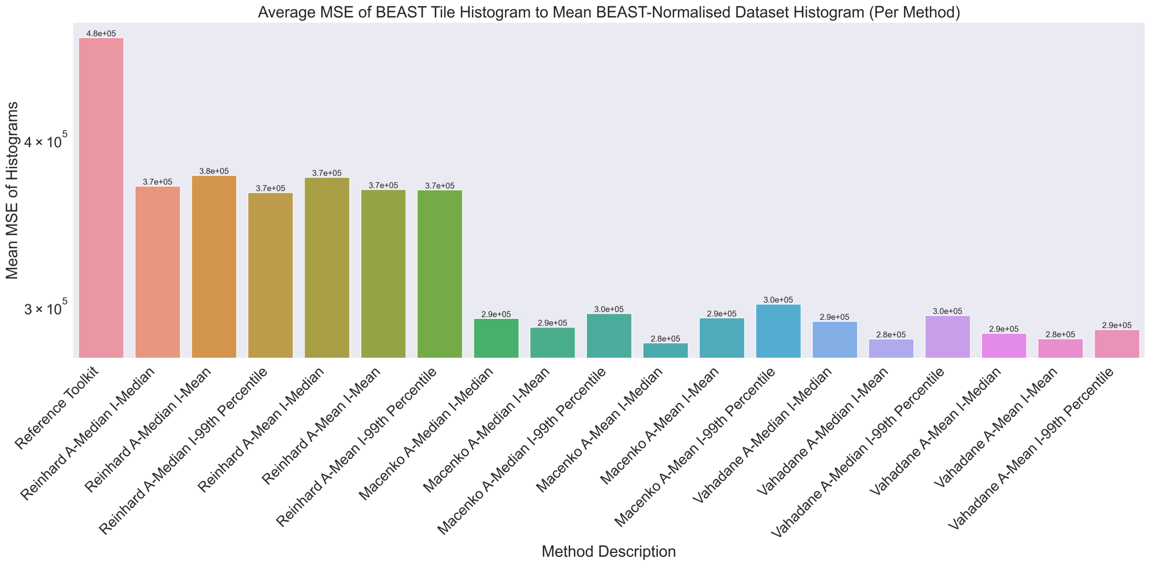


Figure 3.8: A comparison of the reference toolkit, and BEAST normalisation methods for various parameter settings and their effectiveness at normalising a validation dataset. The comparison metric is the average MSE between a tile histogram and the mean histogram of the dataset for each channel in the LAB colour space. The prefix "A" refers to the aggregation method. Prefix "I" refers to the intensity normalisation method.

The following result is shown in figure 3.8. It is a comparison of the toolkit's implementation of the normalisation techniques by Reinhard et al. (2001), Macenko et al. (2009) and Vahadane et al. (2016). During stain metadata estimation over a group of images, the stain metadata must be aggregated into one target matrix; shown in the figure is the aggregation method, as prefixed by "A". The other critical choice in normalisation was by which statistical measure the reference point for stain intensity scaling was chosen. Also shown in the figure is the difference in performance based on the choice of statistical measure of intensity, prefixed by an "I". The best-performing method was the Macenko method, with mean metadata aggregation and median-based intensity normalisation.

The final result from our work on stain normalisation was a visual comparison of the tiles from the raw, reference normalised, and BEAST normalised datasets shown in figure 3.9. Note the background artefacts in the reference dataset as indicated by the red box and their absence in the BEAST normalised tiles, shown in the green box. This is a clear example of the importance of careful normalisation when training deep networks for image generation.

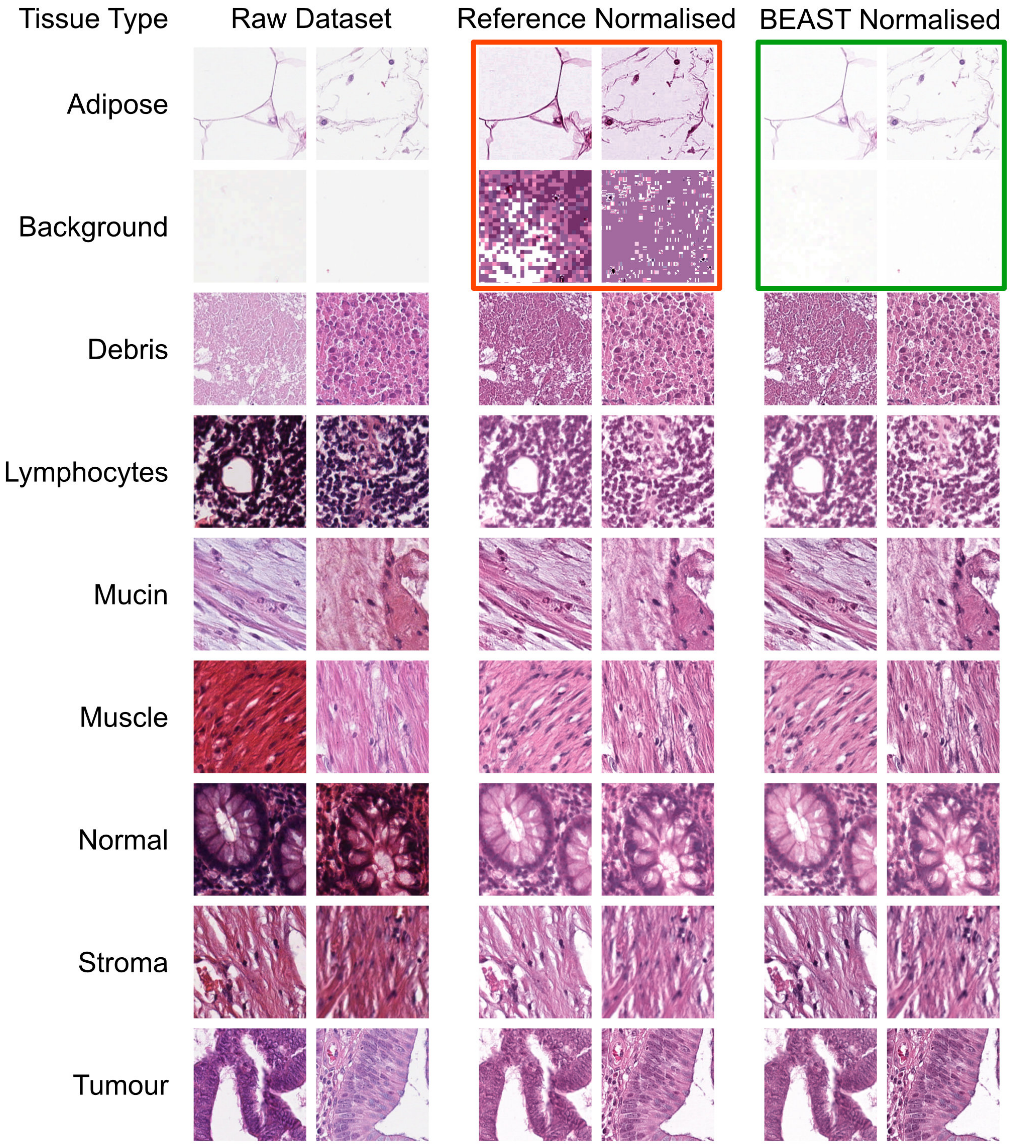
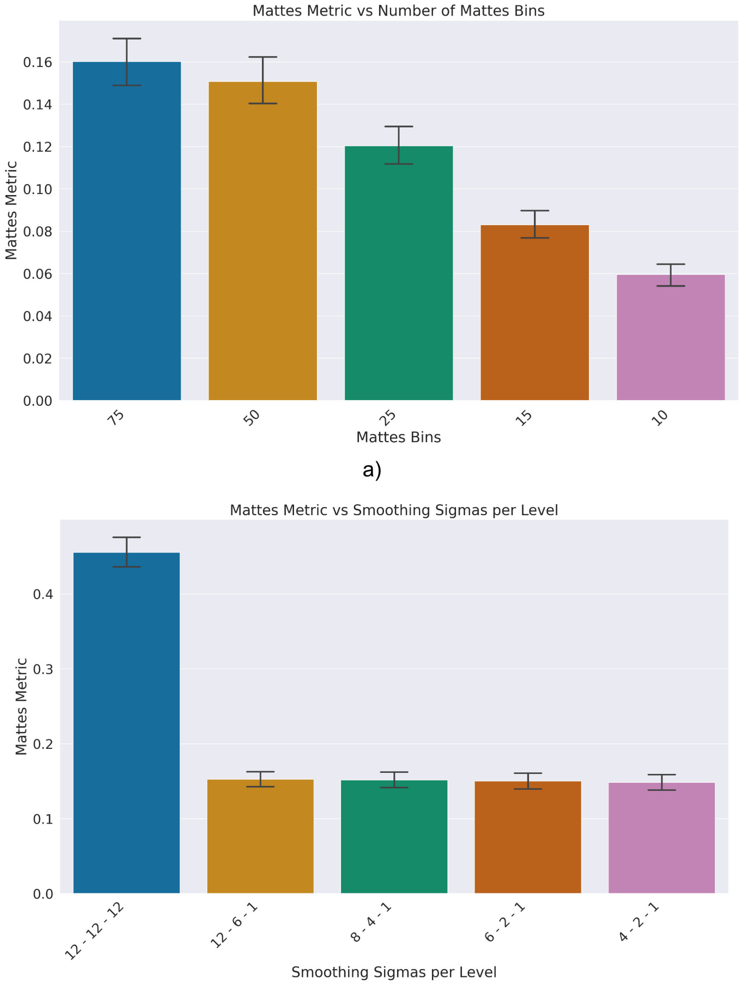


Figure 3.9: A visual comparison of normalisation methods, comparing a selection of tiles from the raw dataset, those normalised with the reference toolkit, and those normalised with the BEAST toolkit. Note the background artefacts highlighted by the red box and their absence in the BEAST normalised tiles, shown in the green box.

3.3.2 Image Registration

To train deep networks for virtual stain translation, creating paired datasets is the cornerstone of supervised learning. Central to this is whole-slide image registration and alignment. Ensuring that corresponding regions in paired images align perfectly is paramount to the learned retention of structural details in supervised networks. Even minor misalignments can introduce errors, leading to inaccurate stain translations. Image registration by aligning images from different stains ensures that each pixel corresponds to its rightful counterpart, laying the foundation for a reliable and robust virtual staining process. The results of our quest to find a set of protocols to do this for haematoxylin and eosin and pan-cytokeratin AE1/AE3 stained whole slide images are presented in the figures and tables below.



a)

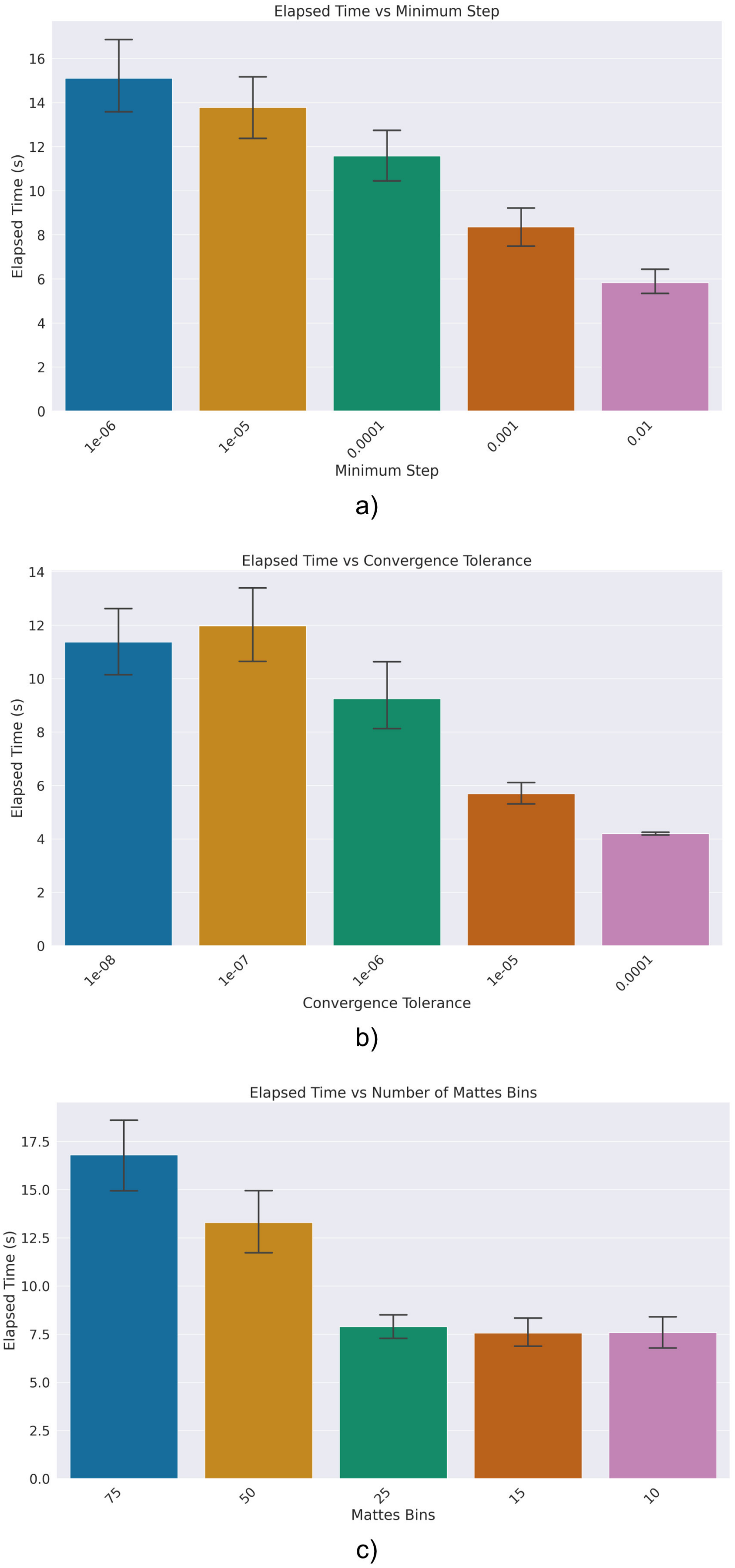
b)

Figure 3.10: Rigid Alignment: Effect of parameters on Mattes metric. Panel a) shows Mattes metric versus the number of bins used in the histogram. Panel b) shows the size of the Gaussian kernel used to smooth the image at each downsample level.

The first step in our whole slide alignment process was a rigid alignment. This was a translation and rotation of the moving image to align it with the fixed image. As described in section 3.2.3, we used a brute force approach to optimise the hyperparameters of the chosen SimpleITK image transforms. Select results of the parameter optimisation for the rigid alignment are shown in figure 3.10. Here, the most impactful parameters on the resulting Mattes metric are displayed. Panel a) shows the effect of altering the number of bins used in the Mattes histogram. The Mattes histogram is a joint histogram of the two images, and the number of bins is the number of discrete values the histogram can take. Panel b) shows the effect of altering the smoothing sigma value for each down-sampled level of the image. The smoothing sigma is the standard deviation of the Gaussian kernel used to smooth the image at each down-sample level. The optimal value for the number of bins was 75. The results of the smoothing sigma sweep are deceiving as smoothing the image also smooths the joint histogram, artificially increasing the Mattes metric, which is important to note. However, a value of 4, 2 and 1 produced the best visual results despite the seemingly lower Mattes metric. The results of the sweep for the other parameters on the Mattes metric are provided in full in the appendix, section A.1.

The alignment transforms had to be run over thousands of tiles during dataset creation. Therefore, they also had to complete within a practical period of time. A perfect alignment is not useful if it takes excessive computational power and time to achieve. A balance had to be struck between the quality of alignment and the elapsed time per tile. The following results show the three parameters that most affected the time to align a tile rigidly. The first of these is the minimum step of the optimisation algorithm. This is the smallest allowable step size in the optimiser parameter space. The second is the convergence tolerance, which is the value at which the optimisation will stop if the change in the metric is below this value. The third is the number of bins used in the Mattes histogram. The minimum step had little effect on the Mattes value, but smaller values significantly increased the elapsed time. However, too small a value might result in early stopping; therefore, 1e^{-4} was chosen as a balance. A convergence tolerance value below 1e^{-4} also did not affect the resulting Mattes metric, so this value was chosen. The choice of the number of Mattes bins did substantially affect the time. However, given the vast improvement to the resulting alignment and the fact it did not add an infeasible amount of time to alignment, the setting of 75 bins was chosen for use. The results of the sweeps for the other parameters on the elapsed time are given in full in the appendix, section A.2.

Once the whole slide image, or a tile resulting from tessellation, had undergone rigid alignment, the next step was a deformable composite affine and displacement-field transform. The affine transform allowed scaling and skewing of the images, and the displacement-field transform allowed the images to freely deform by the alteration of a grid of control points called a displacement field that is overlayed onto the image, where the pixels were “dragged” to follow the points to better align with the target image. These transforms are detailed in section 2.6. The most impactful results of the parameter sweep for the deformable alignment are shown in figures 3.12 to 3.17.



a)

b)

c)

Figure 3.11: Rigid Alignment: Effect of parameters on elapsed time. Panel a) shows the effect of limiting the minimum step in optimisation. Panel b) shows the effect of limiting the convergence tolerance, and panel c) shows the effect of altering the number of Mattes histogram bins.

The first of these was the convergence tolerance, which is the value at which the optimisation will stop if the change in the metric is below. The effect of this parameter on the Mattes metric and the elapsed time of the affine-displacement-field transform is given in figure 3.12. Again, it had little effect on the Mattes metric but substantially impacted the elapsed time. So, a convergence tolerance of 1e-4 was selected.

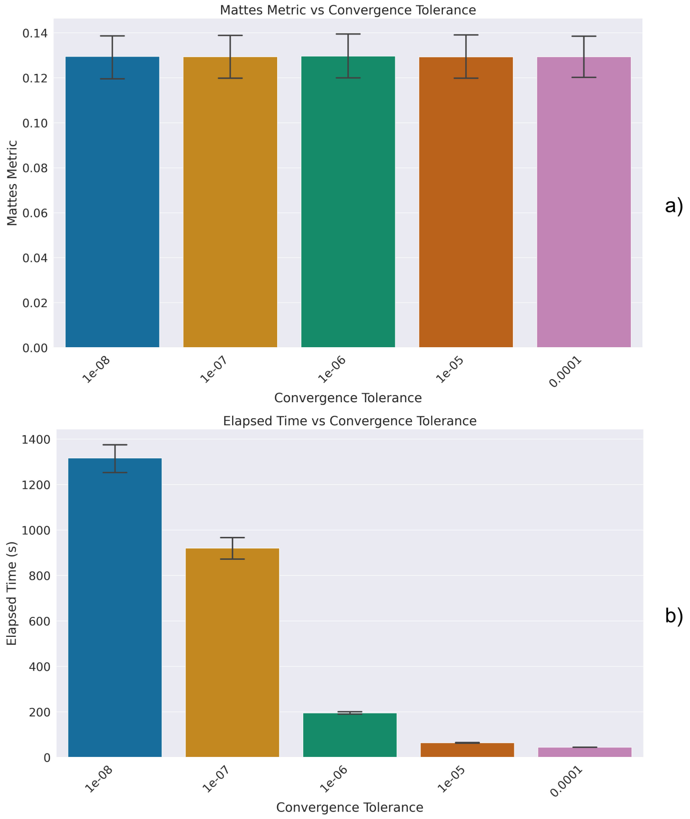


Figure 3.12: Displacement-Field Alignment: Effect of convergence tolerance. Panel a) shows the effect of altering the optimisation convergence tolerance on the Mattes metric. Panel b) shows the effect on elapsed time.

The second was the learning rate estimation, the frequency at which the learning rate is updated. The effect of this parameter on the Mattes metric and the elapsed time of the affine-displacement-field transform is given in figure 3.13. There was a slight improvement in Mattes metric by estimating the necessary optimisation learning rate once at the start of alignment. Estimating every step resulted in a decreased score and vastly increased the elapsed time. Therefore, the learning rate was estimated once at the optimisation's start for subsequent use.

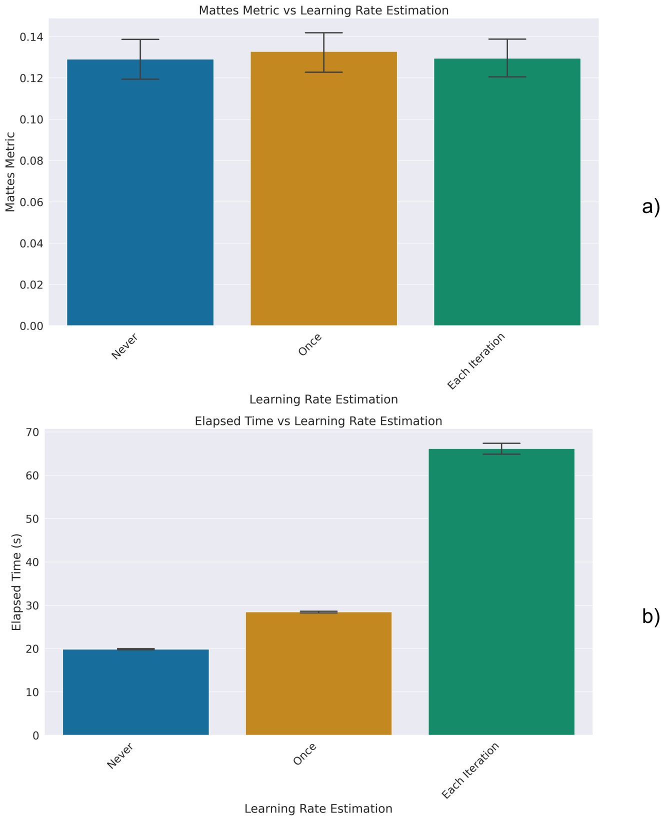


Figure 3.13: Displacement-Field Alignment: Effect of learning rate estimation. Panel a) shows how altering the learning rate estimation frequency affects the Mattes metric. Panel b) shows the effect on elapsed time.

The third was the number of bins used in the Mattes histogram. This is the number of discrete values that the histogram can take. The effect of this parameter on the Mattes metric and the elapsed time of the affine-displacement-field transform is given in figure 3.14. There was a marked improvement in Mattes metric by increasing the number of bins. However, this came at the cost of an increase in elapsed time. Still, once again, given the marked visual and quantitative improvement in alignment, a value of 75 bins was chosen as a balance between the two.

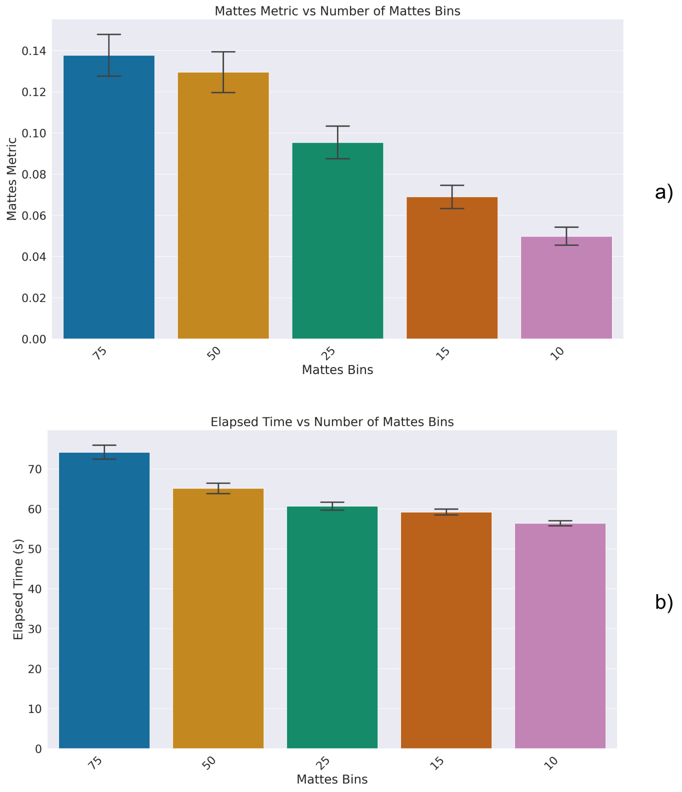


Figure 3.14: Displacement-Field Alignment: Effect of the number of Mattes bins. Panel a) shows how altering the number of Mattes histogram bins affects the metric. Panel b) shows the effect on elapsed time.

The fourth was the percentage of pixels sampled from the image to calculate the Mattes metric. The optimisation could be sped up by randomly sampling a subset of the image pixels to calculate the metric. The effect of this parameter on the Mattes metric and the elapsed time of the affine-displacement-field transform is given in figure 3.15. The percentage sample did not have a significant impact on the alignment score. However, in a counter-intuitive result, by sampling too few pixels, it took the optimisation longer to converge. A value of 35% was chosen to balance time lost to unnecessary computation and time lost to lack of information to complete the alignment.

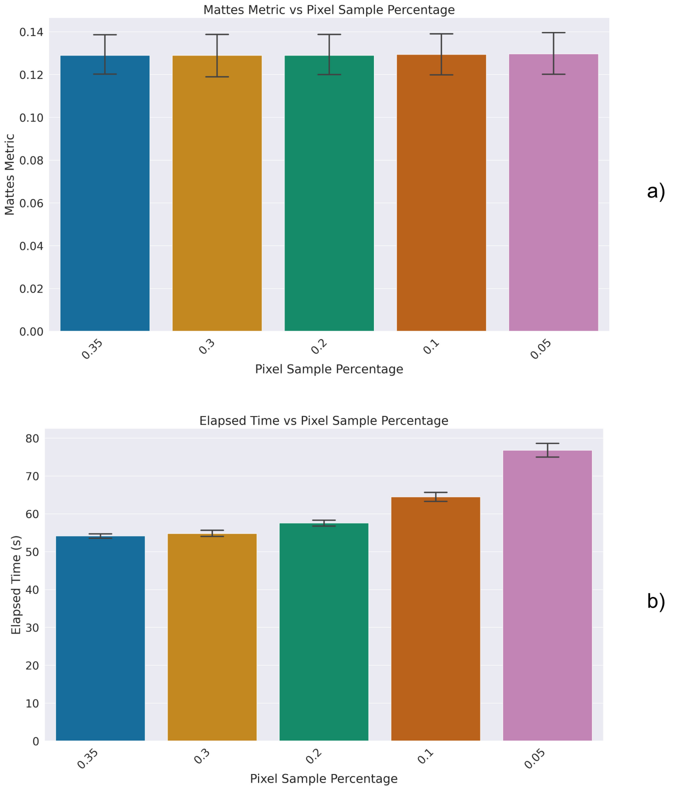


Figure 3.15: Displacement-Field Alignment: Effect of pixel sample percentage. Panel a) shows how altering the percentage of pixels sampled from the image affects the metric. Panel b) shows the effect on elapsed time.

The fifth was the shrink factor, which is the factor by which the image is downsampled at each optimisation level. The effect of this parameter on the Mattes metric and the elapsed time of the affine-displacement-field transform is given in figure 3.16. The shrink factor impacted the alignment score, where larger shrink factors at the top of the pyramid resulted in better alignment at lower levels. Lower shrink factors increased time, as expected with the increased pixel number. The best compromise between speed and alignment quality was starting at a large value and quickly returning to full resolution. A value of 8-4-1 was chosen as a balance between the two.

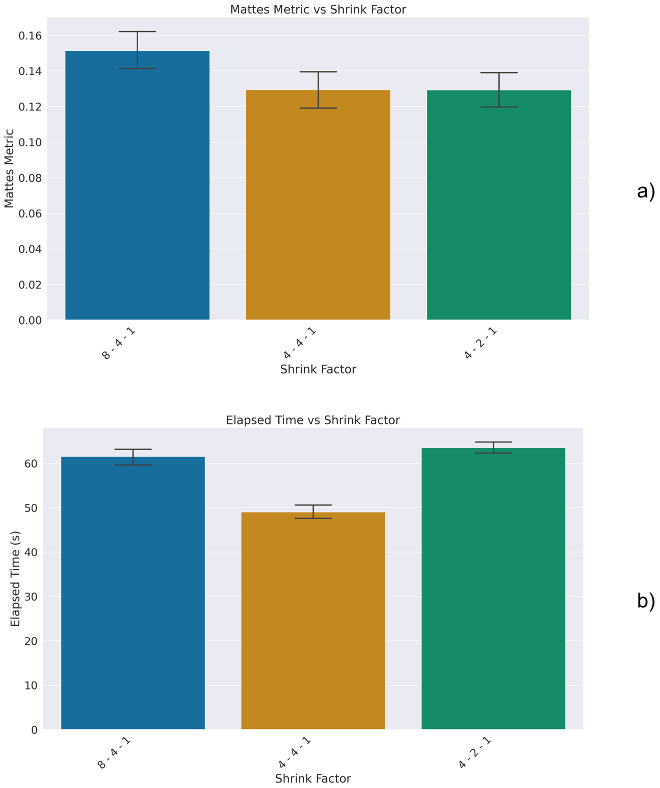


Figure 3.16: Displacement-Field Alignment: Effect of shrink factor. Panel a) shows how altering the shrink factor affects the metric. Panel b) shows the effect on elapsed time.

The sixth was the smoothing sigma, which is the standard deviation of the Gaussian kernel used to smooth the image at each downsample level. The effect of this parameter on the Mattes metric and the elapsed time of the affine-displacement-field transform is given in figure 3.17. The smoothing sigma again artificially increased the Mattes score by smoothing the joint histogram. However, a value of 2-1-0 was found to improve the visual quality of the alignment and was taken forward. A lower smoothing score resulted in increased alignment time due to the increased detail and complexity in alignment. The results of the sweeps for the other parameters on the Mattes metric and elapsed time are given in full in the appendix, in sections A.5 and A.6.

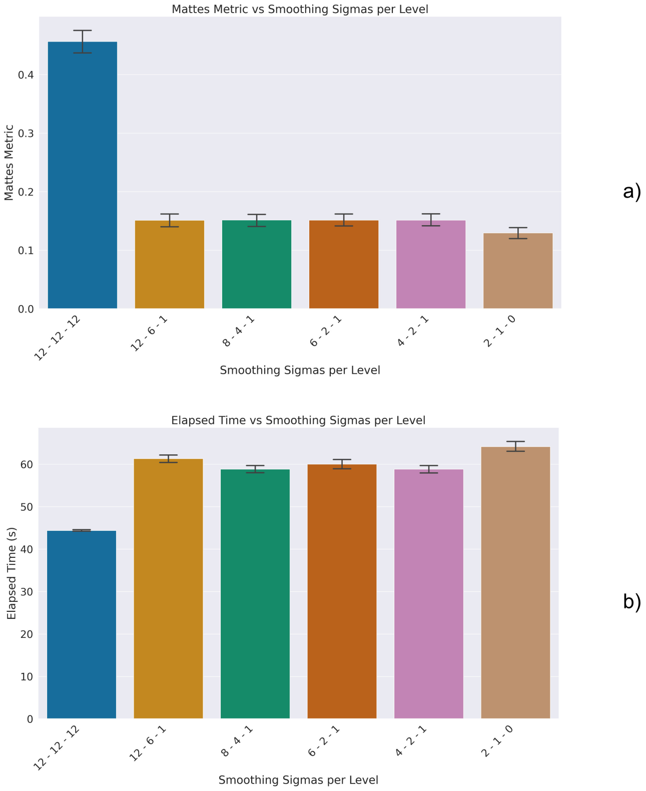


Figure 3.17: Displacement-Field Alignment: Effect of smoothing. Panel a) shows how altering the smoothing sigma affects the metric. Panel b) shows the effect on elapsed time.

The final optimal parameter configurations for our framework built using SimpleITK to align whole slide images rigidly, followed by tessellation subsequent multiple rounds of rigid and deformable tile alignment, are given in tables 3.1 and 3.2, respectively.

Rigid Alignment: Optimal Parameter Values	
Parameter	**Value**
Convergence Tolerance	1e-4
Initial Learning Rate	1
Learning Rate Estimation	Once
Minimum Step	1e-4
Number of Iterations	1000
Number of Mattes Bins	75
Pixel Sample Percentage	35%
Shrink Factor	4-2-1
Smoothing Sigma	4-2-1

Table 3.1: The optimal parameter settings for rigid alignment of paired H&E and IHC whole slide images using a SimpleITK Similarity 2D Transform, with the Mattes mutual information similarity metric.

Displacement-Field Alignment: Optimal Parameter Values	
Parameter	**Value**
Convergence Tolerance	1e-4
Initial Learning Rate	1
Learning Rate Estimation	Once
Number of Iterations	1000
Number of Mattes Bins	75
Pixel Sample Percentage	35%
Shrink Factor	8-4-1
Smoothing Sigma	2-1-0

Table 3.2: The optimal parameter settings for a composite affine and deformable displacement-field alignment of paired H&E and IHC whole slide images using a SimpleITK Displacement-Field transform, with the Mattes mutual information similarity metric.

The final results from our research into whole slide image registration and alignment is an example of the resulting whole slide alignment that is now possible on consecutive serial H&E and IHC AE1/AE3 images. This is available in figure 3.18. An example of the mid-level alignments that are the intermediary step to refine alignment further are given in the appendix, section A.7. Finally, an example of the highest-resolution tile-based deformable alignment is shown in figure 3.19. This critical step allows cellular-level alignment and enables supervised virtual stain translation.

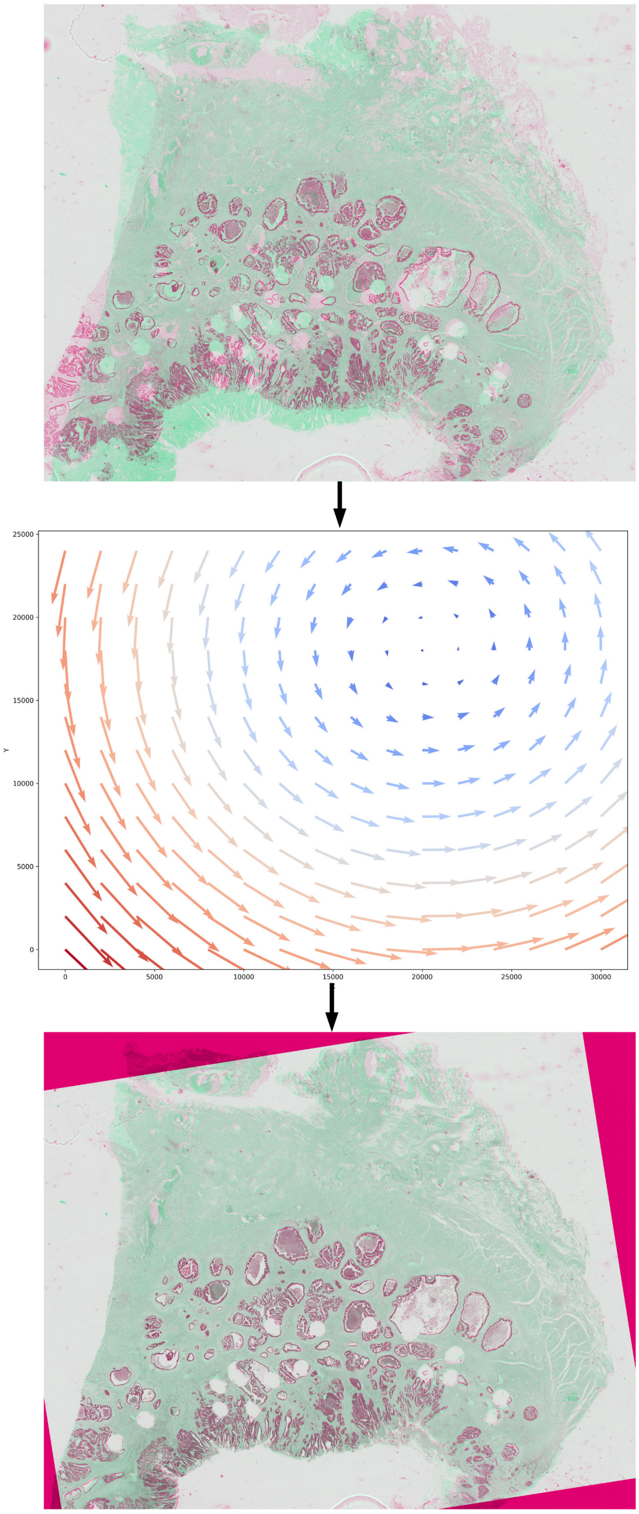


Figure 3.18: An example of whole-slide image registration and alignment using a rigid transform with rotation, scaling and translation. The top image displays a composite overlay of an H&E and AE1/AE3 WSI. The middle image displays a visual representation of the pixel changes to realign the image after registration. The bottom image displays the same composite overlay after rigid alignment.

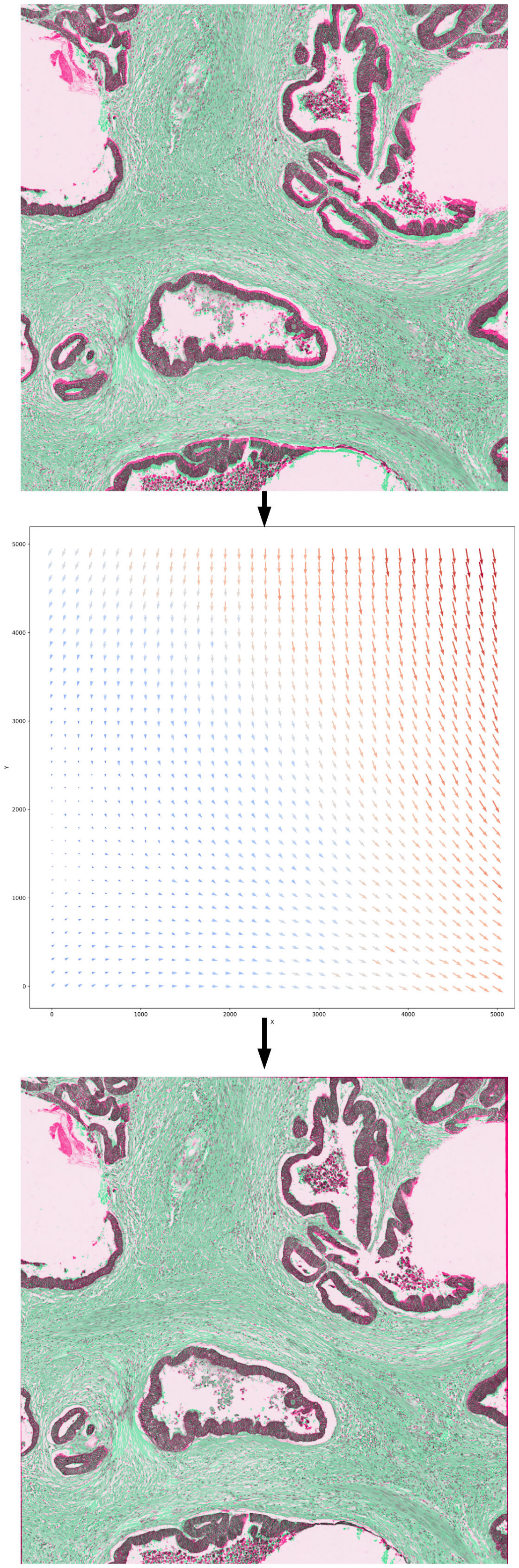


Figure 3.19: An example of a deformable displacement field transform, initialised by a rigid and affine transform, with rotation, scaling, skewing and translation. The top image displays a composite overlay of an H&E and AE1/AE3 tile from the tessellated whole slide image, at the highest resolution at which alignment is performed. The middle image displays a visual representation of the pixel changes to realign the image after registration. The bottom image displays the same composite overlay after rigid alignment.

3.3.3 Virtual IHC: U-Net

Once a set of registered and aligned whole slide images in the desired source and target staining had been successfully created at an appropriate standard as described above. This dataset could be deployed to train and evaluate a series of deep networks for stain translation. Our first proof of concept was to use a basic U-Net architecture, as described in section 3.2.4, to translate H&E stained images to AE1/AE3 stained images. A visual example of the U-Net stain translation is available in figure 3.20. The real H&E and AE1/AE3 images are available in the appendix for comparison, in sections A.8 and A.9 respectively. Note that while blurred in the enlarged areas, the virtual staining approximately represents the original AE1/AE3 staining. This result suggested that H&E to AE1/AE3 stain translation was possible with deep learning and led us to develop the project further.

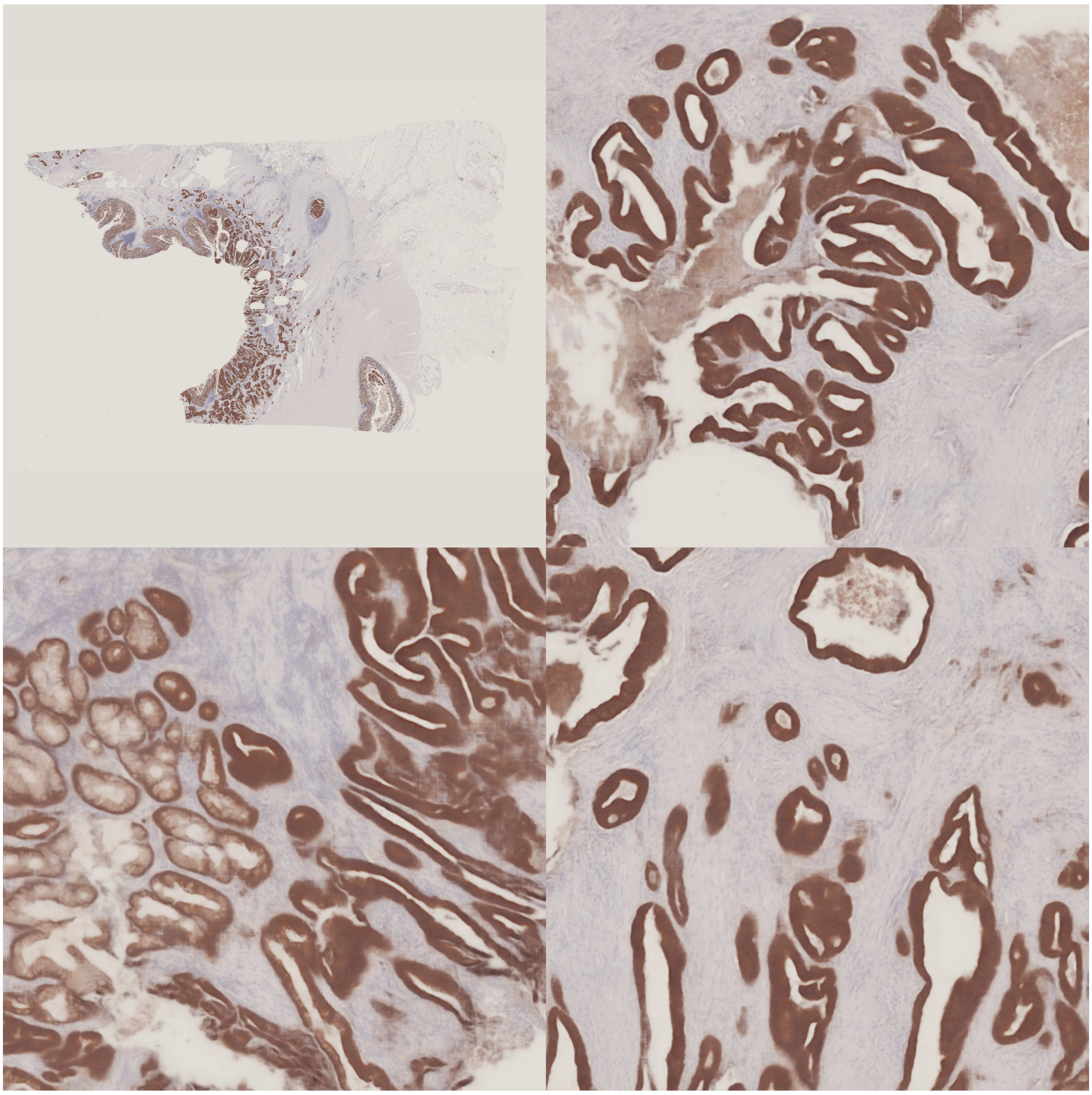


Figure 3.20: An example of **virtual IHC staining** as generated by a **U-Net** model. The top left quadrant contains a virtual IHC AE1/AE3 whole slide image translated from the unseen H&E test dataset, and the other three quadrants contain select magnified regions. The real H&E and AE1/AE3 images are available in the appendix for comparison, in sections A.8 and A.9 respectively.

We developed an evaluation framework to compare virtual IHC models using quantitative metrics. These metrics were generated on an unseen test dataset to evaluate the performance of the models on two whole slide images that were sourced from a different lab under different conditions to the training data. The first of the quantitative metrics was the structural similarity index measure (SSIM) (Wang et al. 2004). This measures the structural similarity between two images, resulting in a value between -1 and 1, with 1 being a perfect match. The second was the Fréchet inception distance (FID) (Heusel et al. 2018). This is a measure of the distance between two distributions of images, and it is a value between 0 and infinity, with 0 being a perfect match. It was utilised to compare the distributions of the real and the stain-translated IHC images to determine the authenticity of the generated features. The training for each network and hyperparameter configuration was repeated at least three times for 50 epochs, where a fold of the training dataset was held out to generate validation data. The epoch used for evaluation on the test dataset was chosen closest to the period before the loss started to increase on the holdout validation dataset (indicating overfitting) and which had acceptable visual results on the validation dataset. The results of the quantitative metrics from the U-Net model are given in table 3.3.

U-Net - Quantitative metrics from test dataset

Run Number	SSIM Value	FID Value
1	0.4685	218.60
2	0.4614	215.41
3	0.4617	219.32
Mean	0.4638 +/- 0.0040	217.77 +/- 2.08

Table 3.3: The SSIM and Fréchet inception distance of the U-Net model on virtually stained images from the test dataset.

To establish if the virtual IHC AE1/AE3 images had any clinical utility, a practising colorectal cancer surgeon, who was also an expert on tumour budding, scored a dataset of 186 test CRC slides, for which we had full access to the patient history. The tumour bud scores were generated following the algorithm established in the 2016 international standards as published by Lugli et al. (2017). The bud scoring procedure is available in algorithm 1. A Kaplan Meier plot was generated to compare the survival of patients with a high tumour bud score (score 3) and a low tumour bud score (score 1 or 2). The results of this are shown in figure 3.21. A log-rank test was used to compare the survival curves, and the p-value was less than 0.05, indicating a statistically significant difference in survival between the two groups. The hazard ratio was 2.59 with a p-value of less than 0.05, indicating that patients with a high tumour bud score had a 2.59 times higher risk of death than those with a low tumour bud score. This result suggests that the virtual IHC AE1/AE3 images have clinical utility, providing a proof of concept for using virtual IHC in clinical practice for tumour bud scoring.

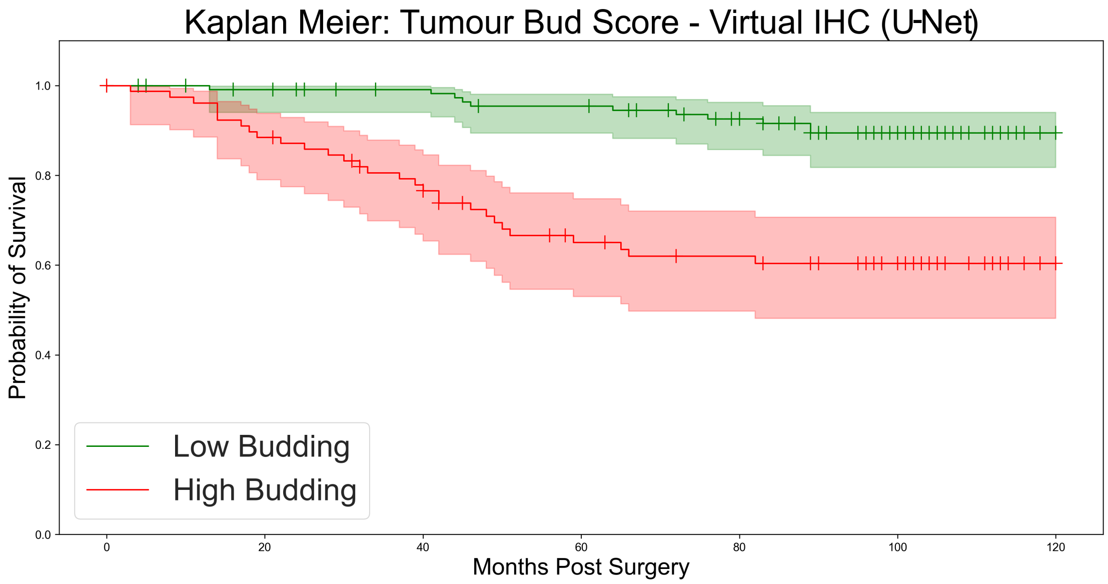


Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
25.8	<0.05 (3.8e-07)	21.33

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
121	79	11	29

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
2.59	1.75	3.85	<0.05 (2.07e-06)	18.88

Figure 3.21: A Kaplan Meier plot of tumour bud scores generated by a colorectal cancer surgeon on 186 CRC slides translated to virtual AE1/AE3. The survival of patients with a high tumour bud score (score 3) and a low tumour bud score (score 1 or 2) are compared. Table 1 shows the results of a log-rank test comparing the survival curves. Table 2 shows the predicted grade and the number of cancer-specific deaths within that grade. Table 3 shows the results of a Cox proportional hazards model comparing the survival curves.

3.3.4 Pix2Pix GANs

Once we had an established proof of concept that our dataset was of sufficient quality and size to train a deep model for virtual stain translation. We started researching methods to produce more realistic images. The Pix2Pix GAN was the second most popular network in the virtual IHC field, capable of photorealistic image generation and designed for supervised style translation tasks like that presented by our dataset. Several different flavours of the Pix2Pix model were implemented as described in section 3.2.4, then evaluated using our framework as described above in section 3.3.3. An example of Pix2Pix-based stain translation is available in figure 3.22.

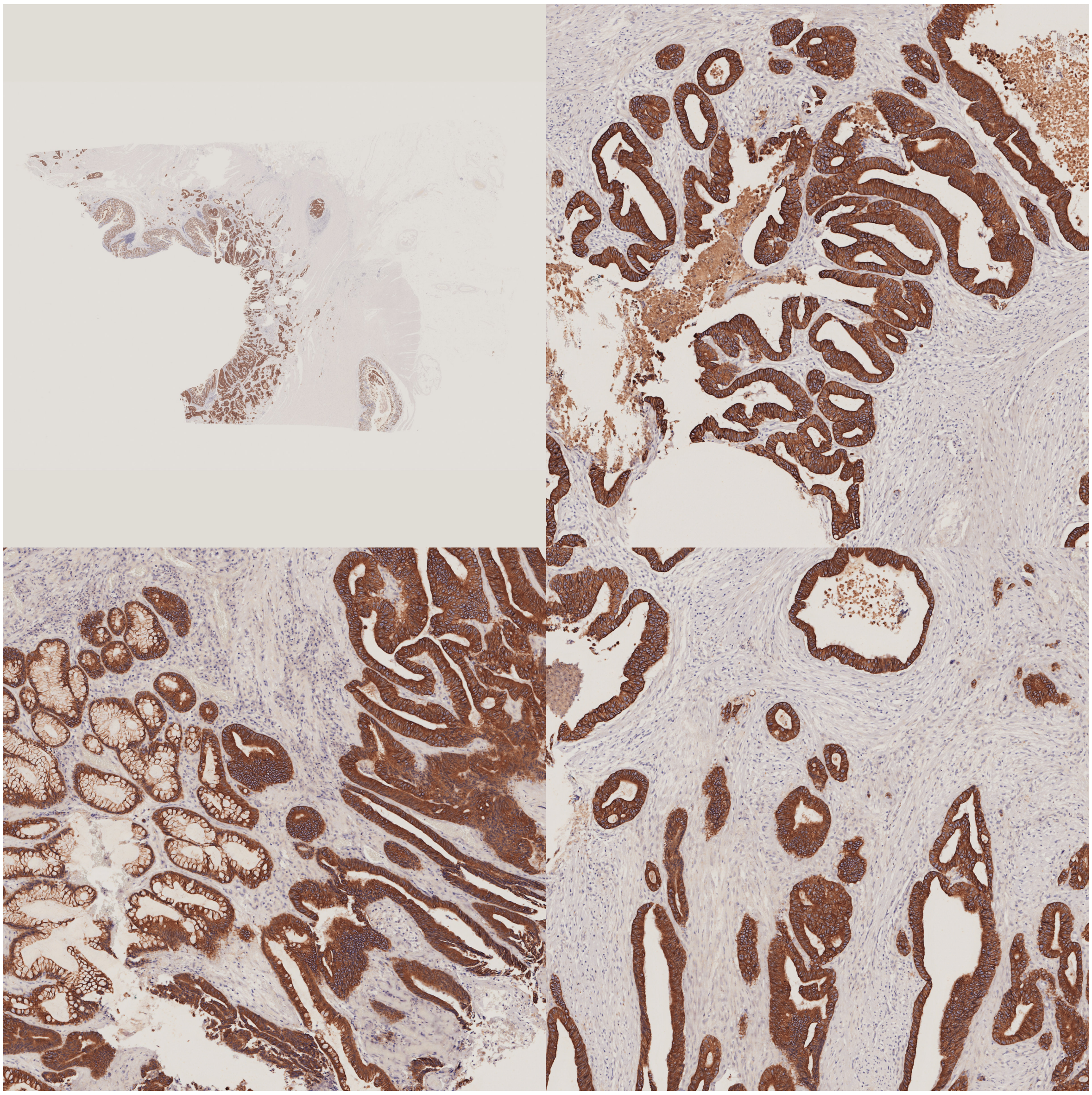


Figure 3.22: An example of **virtual IHC staining** as generated by a **Pix2Pix GAN**. The top left quadrant contains a virtual IHC AE1/AE3 whole slide image translated from the unseen H&E test dataset, and the other three quadrants contain select magnified regions. Note the improved realism when compared to the U-Net model. The real H&E and AE1/AE3 images are available in the appendix for comparison, in sections A.8 and A.9 respectively.

When compared to the real H&E and AE1/AE3 images, which are available in the appendix for comparison, in sections A.8 and A.9 respectively, these models produced photorealistic virtual IHC images, with a marked improvement in realism compared to the U-Net model. However, it has a fatal flaw for virtual stain translation, as shown in column c of figure 3.24. The Pix2Pix models hallucinate features not present in the original H&E image. This was unacceptable for our use case, as it could result in the generation of false tumour buds in the virtual IHC images or, in a potentially worse scenario, result in their removal. This would significantly impact the clinical utility of the virtual IHC images, which invalidates the Pix2Pix model for use in a medical setting. However, the vastly improved realism suggested that GANs were a valid avenue of research for virtual IHC images if properly constrained to maintain the original image structure. As a baseline for later comparison, the results of the quantitative metrics from the Pix2Pix models are given in table 3.4. Three flavours of models were evaluated, each with a different encoder. The first encoder was a ResNet as described in section 2.7.2, with six residual blocks, the second was a ResNet with nine residual blocks, and the third was a U-Net as described in 2.7.2.

The results of the quantitative metrics from the Pix2Pix models are given in table 3.4. The SSIM values for all Pix2Pix models were less than the U-Net, likely due to the artificial increase in SSIM score due to the blurring in U-Net-generated images. The FID values for all Pix2Pix models were substantially less than the U-Net, reflecting their vastly improved realism. The Resnet9-based model had the best balance between SSIM and FID; therefore, it was taken forward as the encoder in subsequently implemented generator networks.

A) Pix2Pix GAN (Resnet6 Encoder) - Quantitative metrics from test dataset

Run Number	SSIM Value	FID Value
1	0.4265	22.64
2	0.4150	22.19
3	0.4027	10.08
Mean	0.4148 +/- 0.0119	18.30 +/- 7.13

B) Pix2Pix GAN (Resnet9 Encoder) - Quantitative metrics from test dataset

Run Number	SSIM Value	FID Value
1	0.4271	23.07
2	0.4208	33.68
3	0.4184	15.79
Mean	0.4221 +/- 0.0045	24.18 +/- 8.99

C) Pix2Pix GAN (U-Net Encoder) - Quantitative metrics from test dataset

Run Number	SSIM Value	FID Value
1	0.3933	47.55
2	0.3871	57.82
3	0.4026	31.55
Mean	0.3943 +/- 0.0078	45.64 +/- 13.24

Table 3.4: The SSIM and Fréchet inception distance of a Pix2Pix GAN with a Resnet6 A), Resnet9 B) and U-Net encoder C) on virtually stained images from the test dataset.

3.3.5 CycleGAN

The most popular generative deep network in the virtual IHC field is the CycleGAN. The original CycleGAN is designed to work with unpaired data to translate style while retaining structure. This characteristic is ideal for stain-translation tasks, where the training source and target whole slide images can become deformed during the staining process. Deformation can occur both during restaining of tissue and across serial sections. Structural differences can also arise in serial sections as tissue varies across the slices. Therefore, the CycleGAN was our primary target for evaluation to discover if it could generate structurally accurate virtual IHC AE1/AE3 images from H&E colorectal cancer slides. The CycleGAN was implemented exactly as in the original paper by Zhu et al. (2020) and trained using our dataset as described in section 3.2.4. An example stain-translation on a test whole slide image using the best performing original CycleGAN model is available in figure 3.23.

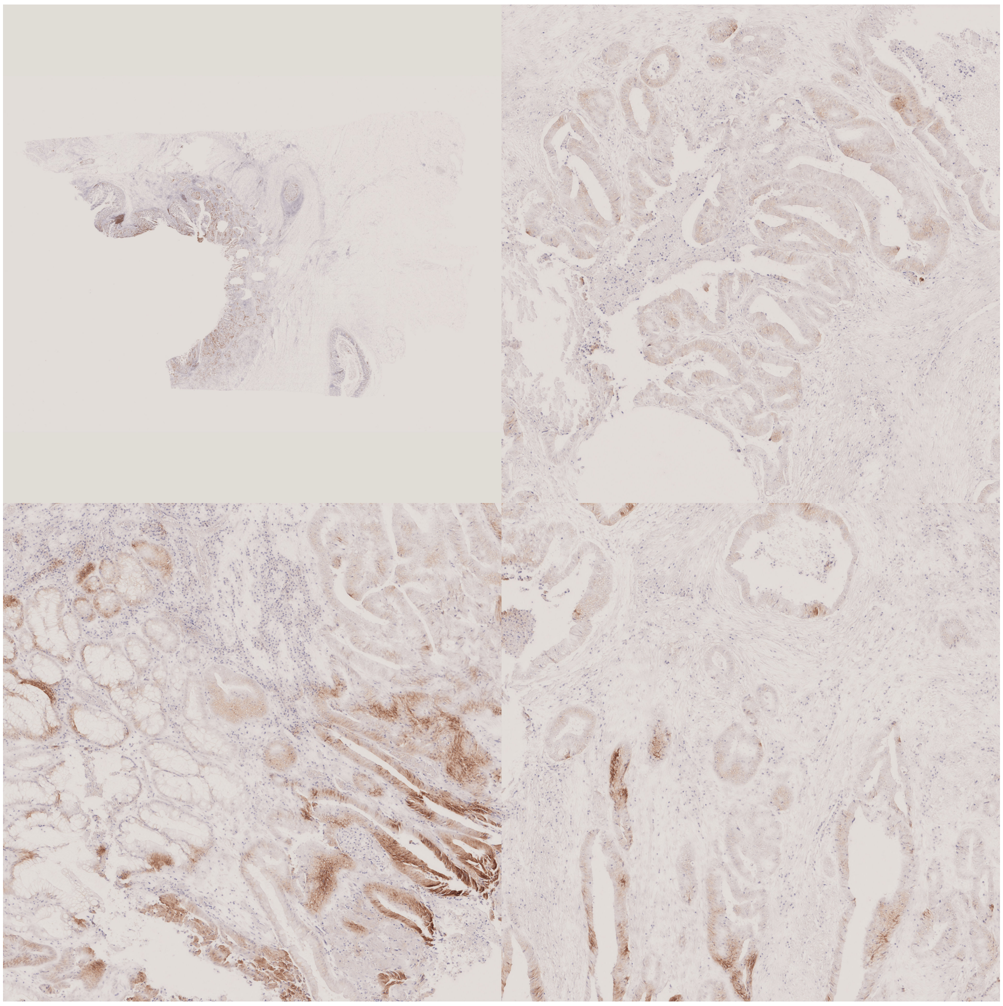


Figure 3.23: An example of **virtual IHC staining** as generated by the original **CycleGAN** model. The top left quadrant contains a virtual IHC AE1/AE3 whole slide image translated from the unseen H&E test dataset, and the other three quadrants contain select magnified regions. The real H&E and AE1/AE3 images are available in the appendix for comparison, in sections A.8 and A.9 respectively. Note the perfect retention of structural details but the poor stain reproduction.

When compared to the real H&E and IHC images, available in the appendix for comparison, in sections A.8 and A.9. The stain translations from the original CycleGAN have near-perfect reproduction of the input tissue structure. However, their stain reproduction is poor, with the AE1/AE3 stain appearing washed out and lacking contrast in the test dataset. In the training dataset, we also saw areas of false staining. We experimented with a variety of hyperparameter combinations and selected the best of those for evaluation. The best configuration was trained 3 times for 50 epochs as with the other models and evaluated on the test dataset using our framework. The SSIM and FID values of the original CycleGAN are available in table 3.5.

CycleGAN (Resnet9 Encoder) - Quantitative metrics from test dataset

Run Number	SSIM Value	FID Value
1	0.4075	76.29
2	0.3924	89.29
3	0.3950	62.73
Mean	0.3983 +/- 0.0081	76.10 +/- 13.28

Table 3.5: The SSIM and Fréchet inception distance of a CycleGAN with a Resnet9 encoder on virtually stained images from the test dataset.

Despite the near-ideal structural reproduction, the CycleGAN SSIM values were worse than the Pix2Pix models. SSIM also has a contrast and luminance component, which is affected by the lack of accurate staining. Predictably, the Fréchet inception distance values were significantly worse than the Pix2Pix models due to the inaccurate stain reproduction, resulting in features that would not occur in natural images.

We had now evaluated several generative deep networks used in style translation. Each had a different deficiency in the task. Figure 3.24 overviews these failings. Columns a) and b) show the real H&E and IHC images. Column c) shows a Pix2Pix GAN stain translation on the training dataset. Highlighted in red are the possible structural hallucinations in unconstrained generative deep networks. Shown in column d) are the same training tiles translated by an unmodified CycleGAN. Highlighted in orange are some select areas where the structure is retained but also have significant issues in stain reproduction. Despite the issues with stain reproduction in virtual AE1/AE3 images, the near-perfect structural reproduction made the CycleGAN the most promising of the evaluated models. We, therefore, investigated methods to improve the stain reproduction while retaining the structural accuracy of the CycleGAN.

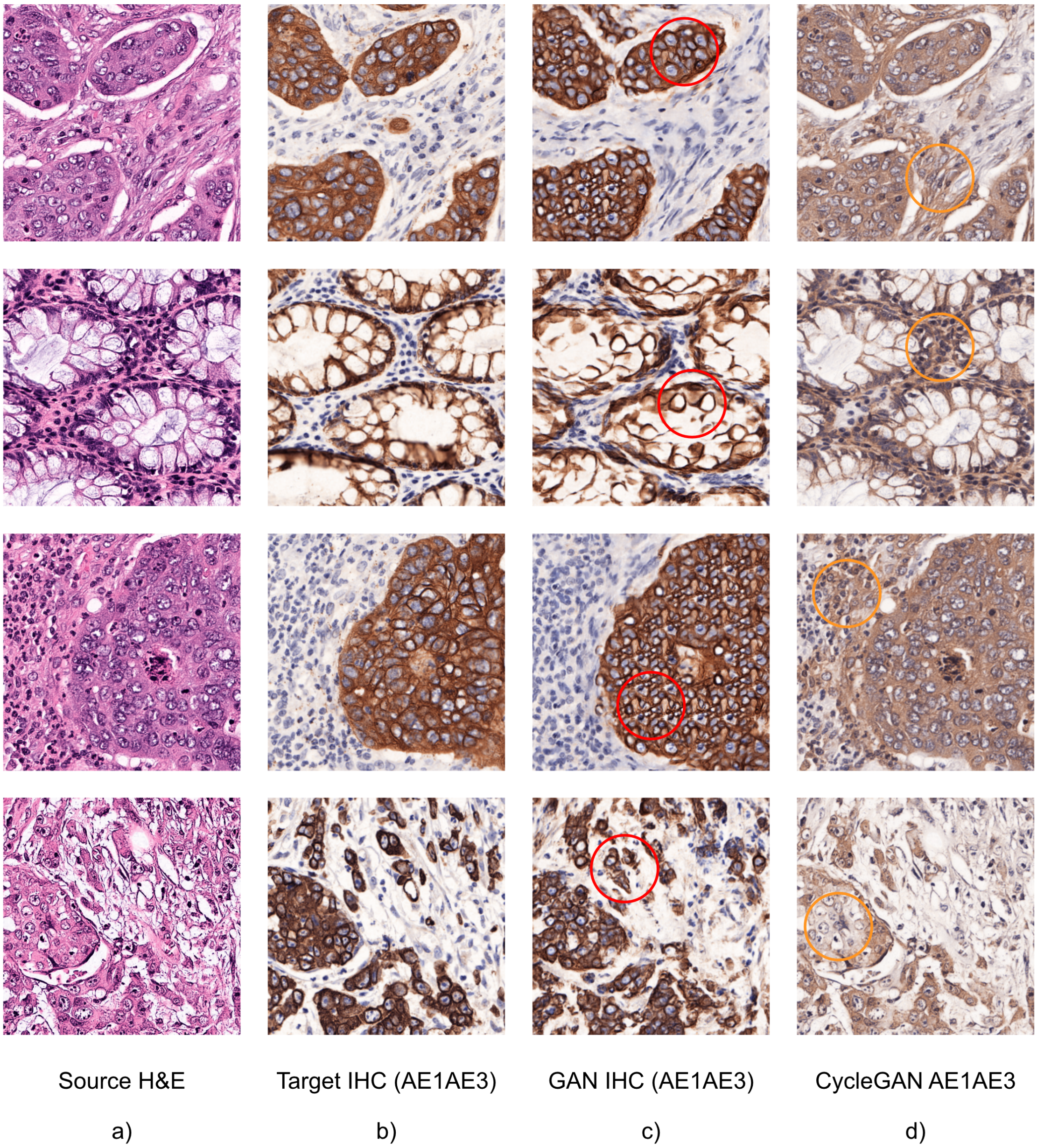


Figure 3.24: Shown are examples of the structural hallucinations possible in GANs. Columns a) and b) show the real H&E and IHC images. Column c) shows a Pix2Pix GAN stain translation on the training dataset. Note in red the structural hallucinations. Shown in column d) are the same training tiles translated by an unmodified CycleGAN. Note that, as highlighted in orange, while the structure is retained significant issues can arise in stain reproduction when trained in an unpaired manner.

3.3.6 Our Proposed Model

We hypothesised that modifying the CycleGAN to learn to reproduce the staining in a supervised manner while maintaining the structure by unsupervised cycle consistency loss would allow the model to learn to generate virtual IHC images accurately. To achieve this, we added a novel L1 loss term to constrain the CycleGAN to make the generated images as similar as feasibly possible to their real counterparts, to teach it to accurately stain the generated images in a supervised manner. We call this “mid-cycle” loss. This was paired with an L2 cycle loss and the addition of cycle SSIM loss, as is common in other virtual IHC implementations. Finally, an adversarial loss term was added, as is common in GANS. The network architecture and loss are described in detail in section 3.2.4. An example stain-translation on a test whole slide image using our best performing mid-cycle CycleGAN model is available in figure 3.25.

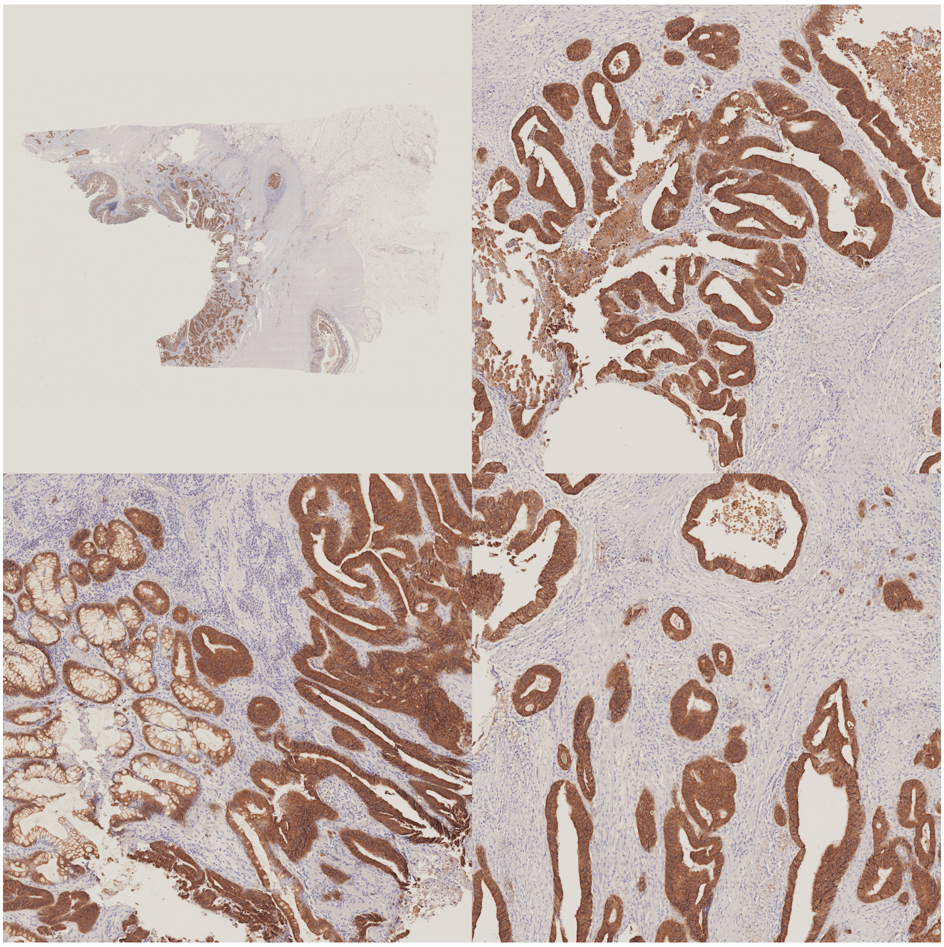


Figure 3.25: An example of **virtual IHC staining** as generated by our proposed modifications to the **CycleGAN** model. The top left quadrant contains a virtual IHC AE1/AE3 whole slide image translated from the unseen H&E test dataset, and the other three quadrants contain select magnified regions. The real H&E and AE1/AE3 images are available in the appendix for comparison, in sections A.8 and A.9 respectively. Note the absence of hallucinations, retention of structural detail and the much-improved staining accuracy.

Our mid-cycle CycleGAN model underwent an extensive number of design iterations, including hyperparameter configurations. Each model took about three weeks to train for 50 epochs. The model was split over two Nvidia A6000 GPUs, each holding a generator-discriminator pair; one GPU was tasked with computing the network loss, and the other held the Inception network utilised for FID validation loss. In this configuration, we managed to fit the entire training framework on only two GPUs. The framework and accompanying models occupied approximately eighty gigabytes of GPU memory during training. Once an acceptable network had been trained, it was evaluated on the test dataset using our evaluation framework. The SSIM and FID values of the CycleGAN with mid-cycle loss are available in table 3.6. The SSIM values were improved when compared to the original CycleGAN. Similarly, the FID values were significantly improved, indicating a more realistic image. These quantitative improvements, most importantly, translate to a notable improvement in the staining quality while retaining structural detail. A comparative example of the images generated by a CycleGAN with mid-cycle loss in how they compare to the real paired IHC images is available in figure 3.26.

Our Proposed CycleGAN (Resnet9 Encoder) - Quantitative metrics from test dataset

Run Number	SSIM Value	FID Value
1	0.4168	64.42
2	0.4273	54.86
3	0.4114	55.57
Mean	0.4185 +/- 0.0081	58.29 +/- 5.33

Table 3.6: The SSIM and Fréchet inception distance values of our proposed CycleGAN with mid-cycle, cycle-ssim loss and a Resnet9 encoder on virtually stained images from the test dataset.

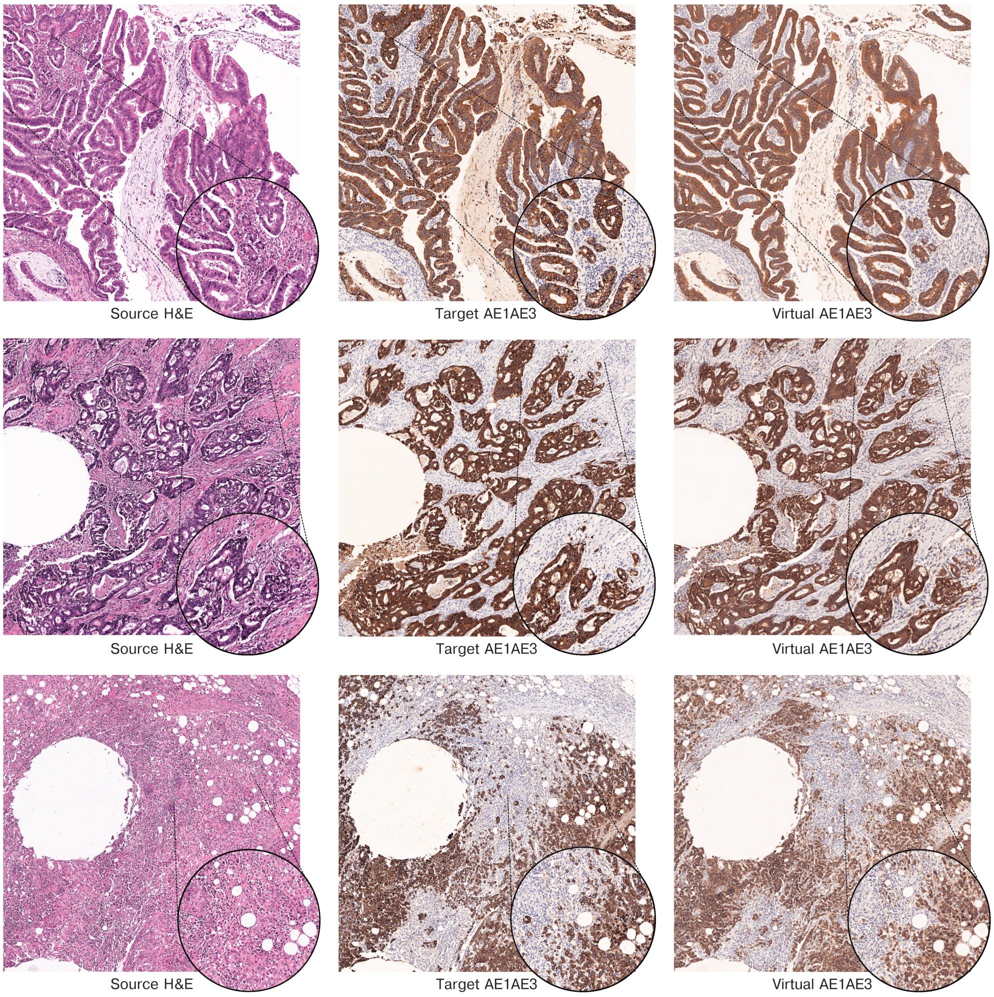


Figure 3.26: Shown in the first column is the source H&E image, in column two is the target AE1/AE3 image. Column three displays the stain translation from a CycleGAN modified with mid-cycle loss. Note the improved stain quality compared to an unmodified CycleGAN, while structural details are retained and hallucinations are eliminated.

The most important metric for the evaluation of a stain-translation network as it relates to the task of tumour bud scoring in this project was its accuracy in reproducing the stain and structure of individual tumour cells in the invasive margin of a tumour. In figure 3.27 below, the ability of the network to recognise and stain single cells is demonstrated on a test dataset. By the nature of the CycleGAN, it can perform stain translation in both directions from H&E to IHC and from IHC to H&E. A real H&E image is shown in panel a), and its translated IHC AE1/AE3 image is shown in panel b). When this is compared to the real IHC image in panel d), it can be observed that the network performs well at correctly identifying and staining the single cells from the source image. Note that the visible single cells in the original H&E and translated IHC are different from those in the real H&E and translated H&E due to the histological differences between the serial sections of the tissue block. When the real IHC image is translated to H&E, as shown in panel c), the structure is maintained, and correct staining is produced. This is a promising result for the use of virtual IHC in tumour bud scoring, as it demonstrates that the network can correctly identify and stain single cells, which was the primary goal of our aim to develop a virtual stain to assist in manual tumour bud scoring.

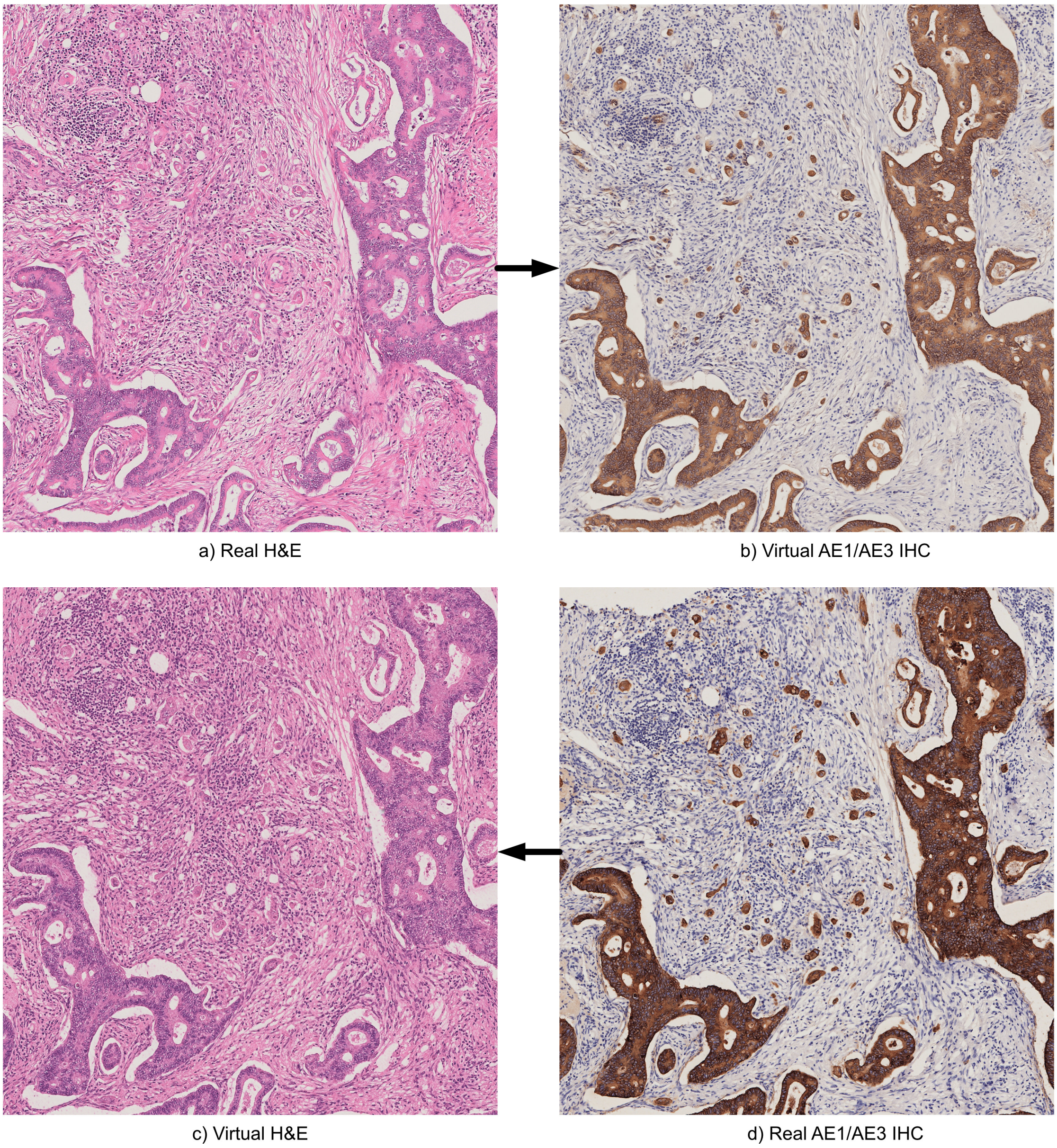


Figure 3.27: An example of the ability of our proposed model to translate between H&E and IHC, and IHC back to H&E, while retaining the structure, and also correctly virtually staining single or small clusters of tumour cells from our test IHC images. Panel a) displays the real H&E image used as input for stain translation. Panel b) displays the virtually stained AE1/AE3 image. Panel d) displays the real AE1/AE3 image for comparison, and panel c) displays the virtually stained H&E image.

3.3.7 Comparison of VIHC Model Performance

With the gathered evaluation results, we could now compare the performance of common virtual IHC model architectures and our proposed model. A boxplot of the SSIM values for the models based on their performance over approximately twenty-three thousand generated VICH tiles on the test dataset can be viewed in figure 3.28. The U-Net scores are artificially high due to the effect of image blurring resulting from its mean-squared error loss. The blurring increases the SSIM value by preserving large-scale structures but reducing the high-frequency details. SSIM evaluates local patterns and textures, and if one of the images is similar but blurred, it tends to reduce the minor differences, resulting in an artificially high score. Therefore, the Pix2Pix models provided the baseline value for the SSIM performance of photo-realistic virtual IHC images from this dataset. The Resnet9-based model had the most consistently high-scoring performance of the Pix2Pix models, validating it for use in later generators. The Resnet6-based model had a much wider range of SSIM scores. The Pix2Pix U-Net model was the worst performing. This is likely because U-Net models have skip connections that enforce the structure to be maintained across the generator; this forces H&E features to appear in the IHC images and reduces the SSIM score as compared to the real IHC image. The unmodified CycleGAN model had the second-worst score. This was a result of its poor ability to stain the output tissue correctly. The CycleGAN with mid-cycle loss restored the ability of the model to stain the output tissue while maintaining structure, resulting in a much-improved SSIM that is similar in distribution to the best-performing Pix2Pix models.

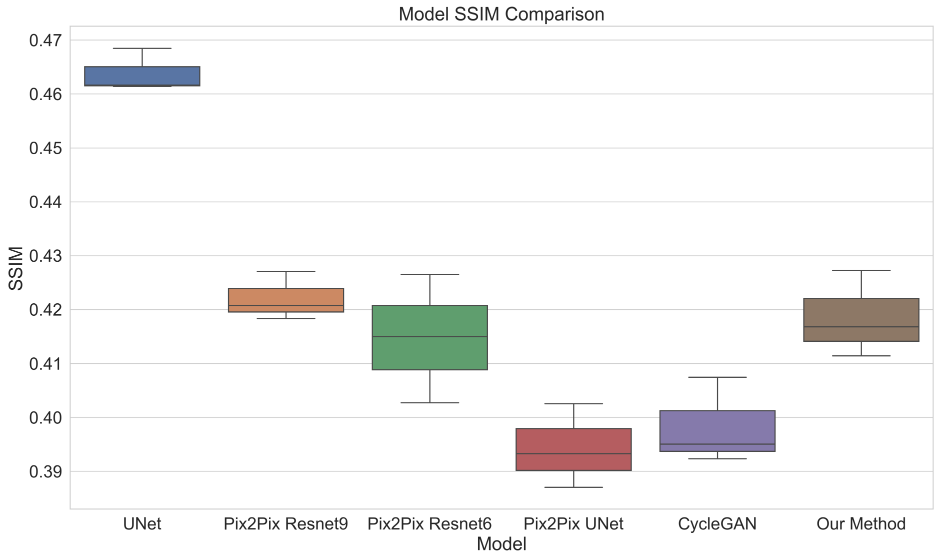


Figure 3.28: A comparison of the structural similarity index measure (SSIM) values for the evaluated models on the test dataset. A higher SSIM value indicates greater similarity between the generated and target images. Note that our model with mid-cycle loss restores the CycleGAN performance and it has a similar distribution to the best-performing Pix2Pix models.

A boxplot of the FID values for the models based on their performance over approximately twenty-three thousand generated virtual IHC tiles on the test dataset is given in figure 3.29. The U-Net model had the highest FID value, indicating that the generated images are the least similar to the real IHC images due to the blurring effect. The Pix2Pix models had the best FID scores, indicating that they are the most like the real IHC images. This was expected because they are not structurally constrained to be similar to the input H&E. They were free to recreate the features of the real IHC, like lack of cytoplastic detail compared to H&E. The unmodified CycleGAN had the worst FID score of the GAN models, indicating that it is the least realistic. This can likely be attributed to poor stain reproduction. The CycleGAN with mid-cycle loss was superior to the original CycleGAN FID score and approached the performance of the Pix2Pix models, indicating that it was also capable of generating realistic IHC images, striking a balance between maintaining the structural details of the H&E and reproducing more realistic staining and features than the unmodified CycleGAN.

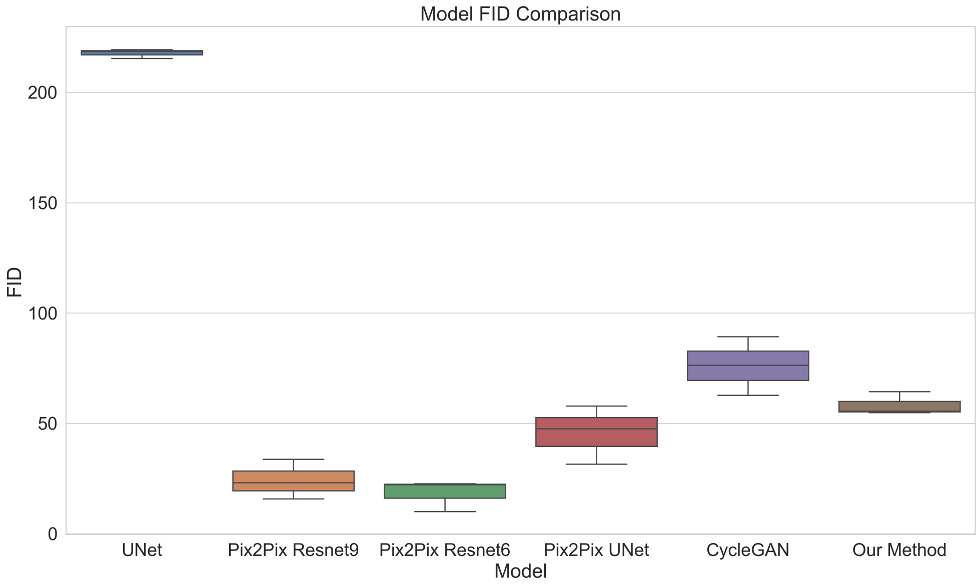


Figure 3.29: A comparison of the Fréchet Inception Distances (FID) for the evaluated models on the test dataset. A lower FID value indicates greater similarity between the generated and target images. Note that our model with mid-cycle loss restores some of the CycleGAN performance and it has a similar distribution to the Pix2Pix U-Net model.

3.4 Discussion

This chapter aimed to investigate the feasibility of translating colorectal cancer H&E whole slide images to AE1/AE3 pan-cytokeratin IHC, first, as a tool to assist pathologists in manually scoring tumour budding and second, as a method to develop a high-quality ground truth dataset of segmented tumour buds for later automated processes using a large dataset of existing H&E whole slide images. To achieve this goal, the developed virtual stain translation model had to maintain the histological features during translation and accurately reproduce the desired stain. Therefore, we evaluated a set of models for structural accuracy and the realism of the generated image features and developed a novel loss function for the CycleGAN (Zhu et al. 2020) model to ensure it meets these requirements.

An essential step in many deep learning image processing workflows is first to normalise the input images. This moves the distribution of pixel values to a range the network expects to observe, which will help it generalise. Without this step, many deep learning models will fail at their given task unless they have been diligently trained using data augmentation to allow the model to observe the bulk of possible inputs. Careful normalisation is critical in image generation and style translation tasks like virtual immunohistochemistry, as any artefacts introduced during normalisation will be carried into the generated data. This makes it more sensitive to these artefacts than many classification, prediction or segmentation tasks. After trialling several common stain normalisation toolkits, they were found to pay little attention to the introduction of background artefacts, most likely because in most histopathology tasks in deep learning, the background is discarded and unimportant. Therefore, our first step was to develop a stain normalisation Python library called the Beatson Augmentation and Stain Normalisation Toolkit (BEAST). In it, we implemented standard normalisation methods and introduced careful preprocessing steps to detect, filter and correctly normalise only the luminance of the background pixels without altering their colour. We also implemented stain augmentation and deconvolution tools that assist in training dataset preparation and evaluation. Without this toolkit, the trained virtual IHC models would learn to introduce the same background artefacts in generated images. Compared to common normalisation packages, ours had a lower average mean squared error between the tile histograms and the mean histogram of the dataset, particularly for the background and adipose tissue classes, indicating better normalisation and improved visual results.

We had a dataset of eight pairs of H&E and AE1/AE3 whole slide images prepared as training and validation datasets and another two prepared as a test dataset. To make these suitable for training, they first had to be aligned with cellular-level accuracy. After the evaluation of several software packages, like Halo (Horai et al. 2019), QuPath (Bankhead et al. 2017) and VALIS (Gatenbee et al. 2021), we found that they were able to align the whole slide images approximately. However, due to the multiple staining domains and large areas of distortion in our slides, they seemed to struggle to overcome many of the regions of tissue deformations introduced during the staining process, and large regions of cells would be out of alignment. Therefore, we also had to develop our own protocols for the alignment of H&E and AE1/AE3 whole slide images at the cellular level. To overcome the areas of significant deformation, we developed a multi-step process that first approximately registered and aligned the whole slide images, then repeatedly tessellates and aligned the slide at higher levels of magnification, using affine and deformable displacement-field transformations from the SimpleITK (Lowekamp et al. 2013) Python library. We performed a detailed brute force search of the transform parameters to ensure we got the best possible alignment within a practical period so it could feasibly be scaled to thousands of whole slide images if necessary. A key finding from this section of work was that displacement-field transforms were superior to B-spline transforms as they allowed for the correction of more radical tissue deformations. This results from the nature of the transforms. In a displacement field transform, every point in the image grid can move independently to realign the tissue, capturing local deformations more effectively than a B-spline transform. This is typically smoother as every control point is interconnected, and the movement of one affects the others, resulting in smoother global transforms. An optimal set of parameters was found that utilised the Mattes mutual information metric, affine, and displacement-field transforms to pyramidally register and align H&E and AE1/AE3 whole slide images, accurate to the cellular level. This enabled the creation of paired data to train deep stain-translation networks.

To determine the feasibility of virtual stain translation between colorectal H&E and pan-cytokeratin AE1/AE3 whole slide images, we first trained a proof-of-concept U-Net model with a simple MSE loss on the paired dataset. This model produced blurred images and lacked histological detail, but it faithfully reproduced the translated stain. This resulted in artificially high SSIM scores and predictably poor FID scores. However, the model was evaluated by using it to translate 186 previously unseen whole slide images from H&E to virtual AE1/AE3, and these were then used to locate and score the density of tumour buds in the invasive margin. The resulting survival analysis, as shown in figure 3.21 was promising; using a log-rank test on the survival distributions of the high and low bud score groups, they were found to be significantly separated, with a p-value less than 0.05. Additionally, as determined by Cox regression, the hazard ratio of the high bud score group was 2.59 (1.75-3.85). Indicating that this group was 2.59 times more likely to have a cancer-specific death than the low-grade group. This was an encouraging result that suggests that virtual immunohistochemistry has clinical utility.

We then evaluated GAN-based translation models to develop a network that could generate photo-realistic versions of these IHC images with a similar or improved level of staining accuracy but retained the structural detail. The first of these models were

Pix2Pix GANs. We evaluated three flavours of this model, with Resnet6, Resnet9 and U-Net-based encoders. We found that the U-Net-based model had the poorest SSIM and FID scores; we hypothesise that this is because the skip connections force the duplication of structural details from the H&E. While this seems ideal for our goal, it makes network training unstable, as the real IHC images have different histological features due to the nature of their chemistry, like the details in the cytoplasm present in H&E compared to AE1/AE3 which only focuses on the cytokeratin in the cytoplasm of epithelial cells. The inclusion of such features may make it easy for the discriminator to spot the fakes and destabilise training. In contrast, a model capable of generating any features but constrained to reproduce the correct structural information present in both, like the cell nuclei, will have superior performance. While we found that all of the Pix2Pix ResNet-based networks produced admirable SSIM and FID scores, these models had the propensity to hallucinate structural features in the output images, as shown in figure 3.24. This removed them from consideration for future experiments, as retention of structural detail is essential for diagnosis in a clinical setting. However, their ability to generate photo-realistic pathology images led us to look for other GAN-based models with this characteristic.

The most popular network in the virtual IHC literature is the CycleGAN model. It is known for its ability to perform style translation on unpaired data while maintaining the structure. Therefore, this was the next logical model to be evaluated. We found that it performed exceptionally well at recreating the histological details of the input image. However, the accuracy of the generated staining was the poorest of the evaluated models.

Spurred by the structural detail of the CycleGAN, we launched a series of experiments to alter its loss function to promote staining accuracy. We found that the identity loss of the original CycleGAN was redundant in this setting and had little effect on the resulting ability to generate appropriately stained images. We, therefore, removed it and, in its place, introduced a novel term to the loss, which we call "mid-cycle" loss. This is an L1 loss term that constrains the generated images at the mid-point of the training cycle to be as similar as possible to their counterpart real IHC image, migrating the network from an unsupervised to a semi-supervised paradigm. We also introduced an SSIM loss to augment the existing cycle loss and reduce any blurring the L1 loss may promote. We found that to train successfully, we had to make several significant modifications to the network design. We first had to uncouple the generator parameters by providing a separate optimiser for each network. This allowed tailored independent updates of the learning rate by the Adam algorithm and improved stability. It also reduced the complexity of backpropagation as chaining parameters of both generators might be more stable on images like the popular horse-to-zebra dataset, as the image components are similar: grass, sky, etc. However, H&E and IHC images have very different luminosity and chrominance characteristics, and chained parameters might result in suboptimal updates as the networks try to match both. Separate optimisers also reduce the risk of error propagation if there is an issue with one of the networks, like the exploding gradient issue. It might also adversely affect the other network when parameters are chained. We also switched from a batch size of one to eight and subsequently changed from instance normalisation to batch normalisation. This was a pivotal step to allow consistent staining in the output images, as with instance normalisation, the staining would be patchy and highly variable. The improvement in the quality of the generated virtual IHC images that result from our modifications to the design can be viewed in figure 3.26, the ability of the CycleGAN network to stain the images accurately was restored, while not affecting the retention of structural detail. The accuracy of the proposed virtual IHC model extends to single-cell tumour cells as well as clusters and beyond, as can be observed in figure 3.27. This is a critical characteristic of any viable virtual IHC network, but it is essential in the context of tumour bud scoring on colorectal cancer.

In future experiments, we would like to test several different configurations of the loss function. We want to investigate how SSIM loss at the mid-cycle point would affect the metrics. We would also like to experiment with converting the images into various colour spaces, such as the LAB colour space, which has separate luminance and chrominance channels to represent the structure and colour. Using the BEAST toolkit, we could also convert the three-channel RGB images to two-channel haematoxylin and eosin, and haematoxylin and DAB to provide the raw stain intensity values to research the effect of deep network performance when presented with cleaner information about the structure and staining. At points during this work, we also experimented with a metric called the staining dice coefficient, which was the dice coefficient calculated from the raw and virtually stain-translated images deconvolved into the stain colour space using the BEAST toolkit. These intensity values were then thresholded to provide a binary mask representing the area of stained pixels, which was then used to compute a dice score. This allowed a quantitative metric for how many pixels in a virtually stained image agreed with its real counterpart. We want to explore and validate it as a loss function and quantitative evaluation metric for virtual IHC. Finally, deep learning is advancing at a breakneck speed, and many new network architectures have been proposed throughout this work that have promising applications for this task. We ran some preliminary experiments with a stable diffusion network configured for stain translation and saw promising results that we would like to explore fully. There are also new generative networks that use state-of-the-art encoders, such as transformers, that may perform even better than the current designs.

This work has underscored the importance of stain normalisation and the precise alignment of whole slide images to ensure accurate structural and stain reproduction in virtual immunohistochemistry. Through the development of a dedicated Python toolkit, we've laid the groundwork for more consistent and reliable stain normalisation for virtual IHC. The rigorous framework developed for registering and aligning whole slide images has offered a robust base for generating paired training data essential for stain translation. The evaluation of a selection of the most popular virtual IHC networks in the literature has provided valuable insights into the respective strengths and weaknesses in this context. Notably, our novel approach of integrating a mid-cycle loss into the CycleGAN model and altering its design to make it more suited to the image characteristics of the task has resulted in a promising protocol, striking a balance between accurate staining and structural reproduction. Compared to the other evaluated models, our proposed technique returns the SSIM scores of the CycleGAN to a similar range of the other generative models and similarly improves the realism of the generated images. This has many potential applications and benefits. If made available, a virtual pan-cytokeratin AE1/AE3 model would allow for the possible generation of a virtual IHC image within minutes of an H&E image being stained and scanned. This would allow for faster and more accurate manual tumour bud scoring at a fraction of the cost and complexity of physical immunohistochemistry, which should speed diagnosis, allow for the stratification of stage II colorectal patients and improve survival outcomes.

Chapter 4

Automated Tumour Bud Scoring by Deep Learning Trained with Virtual IHC Ground Truth

4.1 Introduction

The predominant goal of this project was to investigate methods for automated aggression prediction in colorectal cancer to develop tools to enable the stratification of colorectal cancer patients and improve patient care. This chapter outlines our research in developing deep convolutional neural networks to segment tumour buds in H&E. We used our model from the previous chapter to create virtual immunohistochemistry pan-cytokeratin versions of H&E whole slide images for use as our ground truth. These were referenced to generate a high-quality dataset of example H&E tumour bud segmentations. As discussed in sections 2.3, and 2.4, a histological phenomenon called tumour budding is observed at the invasive front of colorectal cancer tumours. It involves individual or small clusters of up to five cells detaching from the primary tumour and migrating outwards, representing a shift in cell behaviour towards a more aggressive and migratory tumour (Fisher et al. 2021; Jepsen et al. 2018; Tavolara et al. 2022). A high density of tumour buds is now a recognised adverse prognostic factor in CRC, linked to aggressive tumour behaviour and poorer survival rates (Jepsen et al. 2018; Fauzi et al. 2020). Despite the significance of tumour budding, its routine use is limited in the clinic by inconsistent scoring methods, time-intensive nature and inter-observer variability, which results in reproducibility issues when scored on H&E(Fisher et al. 2021; Tavolara et al. 2022). Automated tumour bud scoring in colorectal cancer is now being given serious attention due to its potential for standardisation, efficiency, accuracy, and addressing the challenging, time-consuming and subjective nature of manual assessment (Fisher et al. 2021; Fauzi et al. 2020; Jepsen et al. 2018). Many attempts have been made in the literature to automate bud scoring on physical IHC slides as they provide high contrast for the tumour cells. However, these are resource-intensive, and not every lab has the necessary expertise, equipment and resources to produce them (Kai et al. 2016). Additionally, they are not routinely done as part of clinical diagnosis for colorectal cancer, so an automated method using a physical pan-cytokeratin stain would introduce extra overhead. As a result, the field has been shifting towards H&E-based methods of automated tumour bud scoring (Bokhorst et al. 2018; Banaeeyan et al. 2020; Lu et al. 2022; Tavolara et al. 2022; Bokhorst et al. 2023). When generating data to train these networks, manually created H&E-based ground truth annotations are often inconsistent. While physical IHC can increase agreement, it comes with added costs (Tavolara et al. 2022), and it restricts the creation of ground truth to existing or newly created paired H&E and IHC slides. Training tumour bud segmentation networks using ground truth created by referencing images virtually restained with pan-cytokeratin would make dataset generation more efficient and accurate and allow the reliable creation of ground truth from large H&E cohorts, either existing or prospective. This chapter presents the methods, results and discussion of the developed approach to create a dataset from reference virtual IHC whole slide images and its effectiveness in training a deep convolutional U-Net for automated tumour bud segmentation on H&E slides.

4.2 Materials and Methods

4.2.1 Dataset

To train a deep learning model in automated tumour bud scoring and for later analysis of efficacy, we required numerous training and test images, enough to train a deep network and a sizeable enough patient cohort that we could perform survival analysis on the trained model. The Glasgow Tissue Research Facility (GTRF) and the NHS-GCC Biorepository (ethics no 22/ed/0207) had datasets with patient histories dating back over ten years. The first of these was named the “GRI” cohort after the hospital at which was collected at the Glasgow Royal Infirmary. A colorectal cancer surgeon and tumour budding expert, Dr Hester Van Wyk, conducted manual tumour bud scoring on the entire cohort. There were 785 patients in the dataset, and each had a corresponding whole slide image from a tumour resection. An SPSS database was maintained and provided by Kathryn Pennel from the Edwards lab, containing related pathological characteristics, such as patient metadata, tumour location, stage, level of differentiation, and cancer-specific survival, among others. Table 4.2.1 provides a breakdown of the available information and its representation by the percentage of the cohort. All slides and patient data were anonymised before use in this study. The dataset had to be filtered to remove certain patients with problematic features or missing data. A patient was removed from the cohort if they had neo-adjuvant chemotherapy, as this can alter the pathology of a tumour. They were also removed if they died within one month of the surgery, as this suggests they died from other risk factors like surgical complications rather than as a result of the disease. Finally, any patients were removed if they did not have data for cancer-specific survival in months, a follow-up cancer-specific death status or a manual tumour bud score. To ensure that the resulting 689 patient subset was still representative of the larger population, a Chi-Squared test was run on the categorical variables between the whole cohort and the subgroup, and a Kolmogorov-Smirnov test was run on continuous variables to compare if the distributions were similar. The resulting cohort and subset characteristics breakdown and the distribution p-values are given in table 4.2.1. Tumour location was the only characteristic whose distribution was not preserved after filtering due to a significant reduction in the percentage of rectal tumours. However, this cohort was now ready for use in the preparation of ground truth to train a deep network, the methods for which shall be described in the following sections.

We also required a patient cohort from a different institution to ensure that any trained model could generalise to distinct staining and scanning conditions. The Glasgow Tissue Research Facility (GTRF) and NHSGCC Biorepository (ethics no 22/ed/0207) also provided a second patient cohort, named the “AP” Cohort. This was gathered from the Glasgow Infirmary and Western Infirmary. It comprised 1030 patients, with over 2000 whole slide images, ranging from 1 to 3 slides per patient. A member of the Edwards lab, Phimmada Hatthakarnkul, performed tumour bud scoring over the entire cohort, and again, the corresponding patient metadata and clinicopathological characteristics were provided by Kathyrn Pennel, also from the Edwards lab. All whole slide images and the patient databases were anonymised before being received for use in this study. Filtering based on neoadjuvant treatment, thirty-day mortality and missing values was performed as in the GRI cohort, resulting in a subset of 609 patients. The distributions were analysed for similarity as before, and again, only the tumour location was significantly different this time due to the significant overall change in patient numbers between the entire cohort and subset and the drop in the percentage of rectal tumours. The cohort and subset characteristics breakdown and the distribution p-values are given in table 4.2.1.

Forty slides were randomly selected from the GRI cohort to serve as ground truth in generating training and validation data for the model, and the remainder of the was held out as test slides. The entire AP cohort was held out to evaluate the ability of the model to generalise across institutes. With these patient cohorts, we now had the necessary datasets for the training and evaluation of an automated tumour bud scoring model using deep learning. We shall discuss the methods and techniques used for ground truth creation, training and evaluation below in section 4.2.2.

GRI Cohort & Subset Characteristics

Characteristics (n = Whole Cohort/Subset)	Whole Cohort {n (%)}	Subset {n (%)}	p-value
Age (n=785/689)	68.34 (21-98)	68.72 (21-98)	1.0000
Sex (n=785/689)			0.9062
Female	352 (44.84%)	312 (45.28%)	
Male	433 (55.16%)	377 (54.72%)	
Location (n=785/689)			0.0473
Right	302 (38.47%)	286 (41.51%)	
Left	244 (31.08%)	233 (33.82%)	
Rectum	239 (30.45%)	170 (24.67%)	
TNM Stage (n=785/689)			0.9978
I	239 (30.45%)	170 (24.67%)	
II	96 (12.23%)	84 (12.19%)	
III	377 (48.03%)	330 (47.90%)	
IV	312 (39.75%)	275 (39.91%)	
Tumour Differentiation (n=779/685)			0.9244
Well	696 (89.35%)	614 (89.64%)	
Poor	83 (10.65%)	71 (10.36%)	
Tumour Perforation (n=785/689)			0.8747
Absent	761 (96.94%)	666 (96.66%)	
Present	24 (3.06%)	23 (3.34%)	
Margin Involvement (n=785/689)			0.4432
Absent	727 (92.61%)	646 (93.76%)	
Involved	58 (7.39%)	43 (6.24%)	
Venous Invasion (n=785/689)			0.7931
Absent	379 (48.28%)	327 (47.46%)	
Present	406 (51.72%)	362 (52.54%)	
Peritoneal Involvement (n=785/689)			0.4267
Absent	594 (75.67%)	508 (73.73%)	
Involved	191 (24.33%)	181 (26.27%)	
MMR Status (n=761/671)			0.9596
Proficient	191 (25.10%)	118 (17.59%)	
Deficient	136 (17.87%)	449 (66.92%)	
MSI-Low	511 (67.15%)	104 (15.50%)	
GMS (n=756/668)			0.9177
0	119 (15.74%)	110 (16.47%)	
1	480 (63.49%)	423 (63.32%)	
2	157 (20.77%)	135 (20.21%)	
mGPS (n=785/689)			0.8671
0	479 (61.02%)	412 (59.80%)	
1	176 (22.42%)	162 (23.51%)	
2	130 (16.56%)	115 (16.69%)	
Tumour Budding (n=757/689)			1.0000
Low	532 (70.28%)	485 (70.39%)	
High	225 (29.72%)	204 (29.61%)	
Overall Survival (n=784/689)			0.9719
Alive	275 (35.08%)	240 (34.83%)	
Dead	509 (64.92%)	449 (65.17%)	
Cancer Specific Survival (n=784/689)			0.9314
Alive	552 (70.41%)	491 (71.26%)	
Dead	232 (29.59%)	198 (28.74%)	

Table 4.1: Clinicopathological characteristics of the GRI patient cohort, compared with the subset of patients with available data for all variables used in the study. A Chi-squared test was performed for categorical variables to evaluate the quality of the subsampled data to ensure that the distributions were similar. A Kolmogorov-Smirnov test was run for continuous variables.

AP Cohort & Subset Characteristics

Characteristics (n = Whole Cohort/Subset)	Whole Cohort {n (%)}	Subset {n (%)}	p-value
Age (n=1030/609)	69.89 (27-94)	70.27 (27-94)	0.9974
Sex (n=1030/609)			0.5675
Female	491 (47.67%)	300 (49.26%)	
Male	539 (52.33%)	309 (50.74%)	
Location (n=1023/602)			0.0211
Right	430 (42.03%)	259 (43.02%)	
Left	340 (33.24%)	228 (37.87%)	
Rectum	253 (24.73%)	115 (19.10%)	
TNM Stage (n=1030/609)			0.1642
I	253 (24.56%)	115 (18.88%)	
II	138 (13.40%)	97 (15.93%)	
III	483 (46.89%)	287 (47.13%)	
IV	388 (37.67%)	206 (33.83%)	
Tumour Differentiation (n=1030/609)			0.6114
Well	919 (89.22%)	549 (90.15%)	
Poor	111 (10.78%)	60 (9.85%)	
Tumour Perforation (n=1030/609)			0.5914
Absent	964 (93.59%)	565 (92.78%)	
Present	66 (6.41%)	44 (7.22%)	
Margin Involvement (n=1030/609)			0.3149
Absent	960 (93.20%)	576 (94.58%)	
Involved	70 (6.80%)	33 (5.42%)	
Venous Invasion (n=1030/609)			0.6068
Absent	681 (66.12%)	411 (67.49%)	
Present	349 (33.88%)	198 (32.51%)	
Peritoneal Involvement (n=1030/609)			0.8189
Absent	734 (71.26%)	438 (71.92%)	
Involved	296 (28.74%)	171 (28.08%)	
MMR Status (n=1000/605)			0.9987
Proficient	822 (82.20%)	498 (82.31%)	
Deficient	38 (3.80%)	22 (3.64%)	
MSI-Low	112 (11.20%)	68 (11.24%)	
GMS (n=989/604)			0.2958
0	313 (31.65%)	189 (31.29%)	
1	491 (49.65%)	319 (52.81%)	
2	185 (18.71%)	96 (15.89%)	
mGPS (n=836/444)			0.7274
0	437 (52.27%)	234 (52.70%)	
1	243 (29.07%)	121 (27.25%)	
2	156 (18.66%)	89 (20.05%)	
Tumour Budding (n=953/609)			0.4750
Low	684 (71.77%)	448 (73.56%)	
High	269 (28.23%)	161 (26.44%)	
Overall Survival (n=1011/609)			0.3879
Alive	355 (35.11%)	219 (35.96%)	
Dead	656 (64.89%)	390 (64.04%)	
Cancer Specific Survival (n=1011/609)			0.2031
Alive	687 (67.95%)	433 (71.10%)	
Dead	324 (32.05%)	176 (28.90%)	

Table 4.2: Clinicopathological characteristics of the AP patient cohort, compared with the subset of patients with available data for all variables used in the study. A Chi-squared test was performed for categorical variables to evaluate the quality of the subsampled data to ensure that the distributions were similar. A Kolmogorov-Smirnov test was run for continuous variables.

4.2.2 Methods

Dataset Creation

The primary aim of the work in this chapter was to determine the utility of virtual immunohistochemistry whole slide images in highlighting tumour-derived epithelial cells to create a high-quality dataset of ground truth segmentations. The overall efficacy was determined by using this dataset in the downstream task of training an automated tumour bud scoring model to operate directly on H&E images. A dataset of paired H&E tiles and binary masks representing the pixels containing tumour buds was required to train the final model. The bud segmentations had to be generated manually on H&E while referencing a paired IHC slide. We were taught to segment them in H&E and IHC over one year by a colorectal cancer surgeon and tumour bud expert, Doctor Hester Van Wyk. A random subsample of forty H&E slides were taken from the GRI cohort, and paired virtual AE1/AE3 slides were inferred from them using the best virtual IHC colorectal cancer model as developed in chapter 3. The slides were first normalised using the BEAST toolkit as described in section 3.2.2. Then, over a few months, all cells in the invasive margin of each slide were scored using the ITBCC guidelines and protocols as described in algorithm 1. The buds were manually segmented, resulting in a dataset of 59,717 buds. This was done in QuPath (Bankhead et al. 2017) with the H&E and virtual IHC images open side by side and synchronised. An example of one hotspot with the manual positive and negative annotations can be observed in figure 4.1, panel a), the corresponding reference virtual IHC image that was used to ensure the cells were epithelial-derived through the expression of pan-cytokeratin is visible in panel b), note how much more distinguishable the buds are in the reference virtual IHC image. An initial model was trained, as we shall describe below. We used this model to segment the cells in the invasive margin of the same slides and used the incorrectly identified cells as negative examples. We iteratively repeated this process sixteen times until we had a suitable dataset that resulted in satisfactory performance. The cell segmentations were exported to JSON files containing the line coordinates of the annotations. A program was designed to load and represent these annotations as polygons using the Shapely (Gillies et al. 2007) Python library. The corresponding whole slide images were loaded using OpenSlide-Python (Goode et al. 2013), and the region around each polygon was cropped. The H&E tile and a rasterised version of the polygons were output to size 128x128x3 and 128x128x1 tiles, respectively. Each tile was focused on one segmented bud. However, the buds were positioned with a random offset from the tile centroid to stop the model from being biased by segmentations consistently in the centre. However, it was ensured that the entire segmentation was contained within the tile boundary and not cropped accidentally. The final dataset consisted of 119,267 tiles comprised of positive and negative segmented cells. A few select example H&E tiles and their paired binary segmentation masks are visible in figure 4.2

Deep Network Architecture and Training

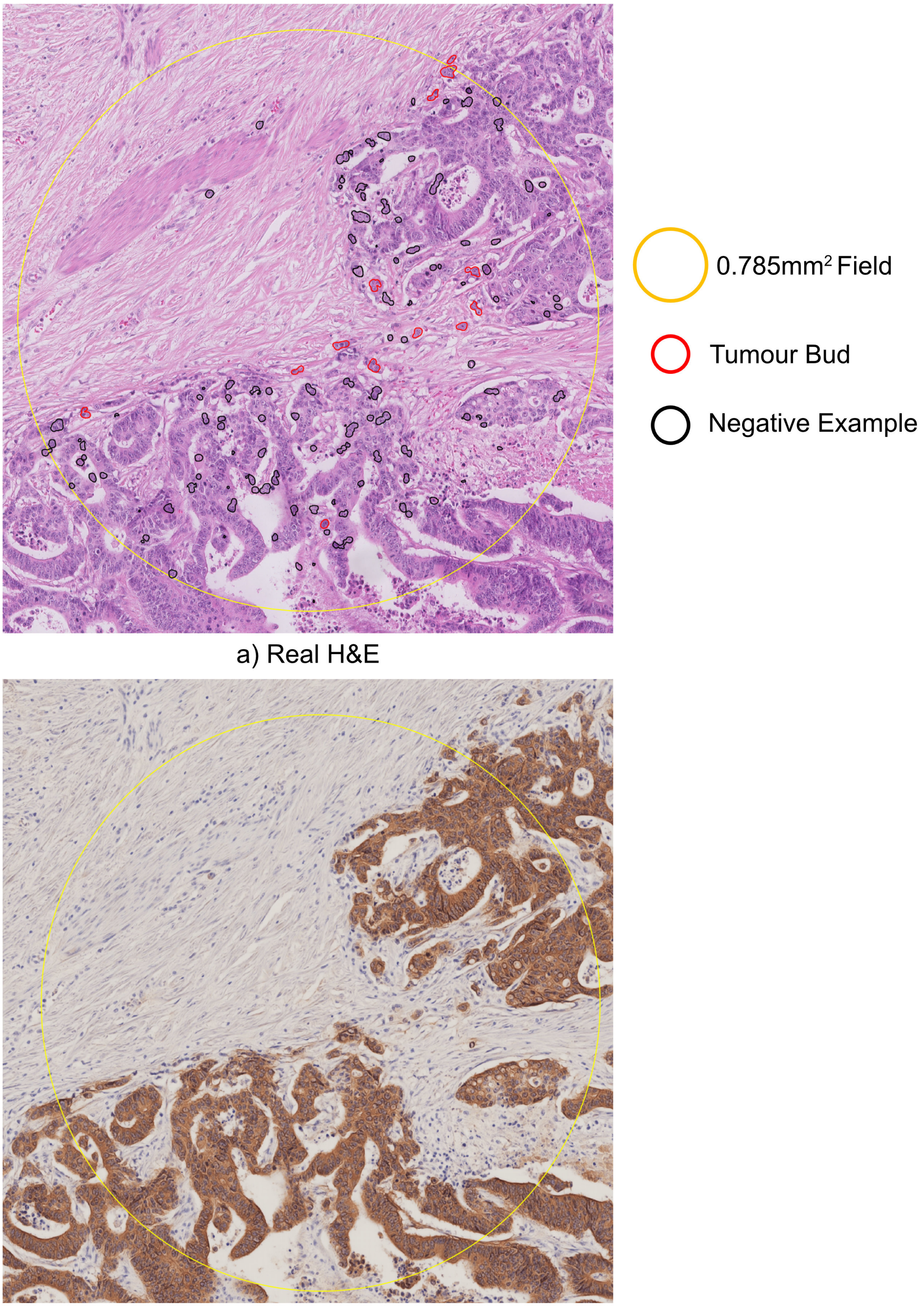
As the literature review in section 2.9 revealed, the state-of-the-art segmentation model for automated tumour bud scoring was a U-Net as described in section 2.7.2. We experimented with a variety of different backbone networks as the encoder within the U-Net, including the Inception V3 (Szegedy et al. 2015), the Mobile-NetV2 (Sandler et al. 2019), and Resnet50 (He et al. 2015) networks. The use of transfer learning compared to training from the initial random weights was also evaluated. The final network architecture that resulted from the search was a U-Net with a ResNet50 backbone as the encoder, starting with 64 filters in the first convolutional layer, then 128, 256, 512, 1024, and finally culminating in a latent space with 2048 filters in the bridge between the encoder and decoder. The input image was down-sampled five times from 128^2 to 4^2 before being restored to the original dimensions in the output segmentation mask. Skip connections were maintained after each downsampling operation and concatenated with the output from the bridge or previous layer to be input to the next upsampling convolutional layer. Figure 4.3 shows a diagram depicting our network architecture.

The model architecture was implemented in Python with the Pytorch library (Paszke et al. 2019). A learning rate of 2e^{-4} and the ADAM optimiser (Kingma et al. 2017) were used. The input tiles were normalised to a range of -1 to 1. This helps combat the vanishing gradient problem and allows for the use of the tanh activation function after the final layer. ReLU activations were also applied to the other convolutional layers to help with this issue. Each convolutional layer was also regularised with batch normalisation, and there were three dropout layers in the first three upsampling operations on the decoder side. The cross-entropy dice loss was used as the criterion for network optimisation. The model was trained for at least 1000 epochs with a batch size of 32 and a weight decay of 1e^{-4}. The training was carried out in a data-parallel fashion distributed over four A6000 GPUs. The best-performing epoch was chosen based on the lowest cross-entropy dice loss and an evaluation of the visual results of the segmentations.

Objective Field Placement

The ability to detect buds in H&E is only one side of an automated scoring process. The other is the ability to target the correct slide region to score for tumour buds. Following the ITBCC guidelines, tumour buds should only be scored in the first five hundred microns around the leading edge of a solid tumour, known as the invasive margin. However, regions that exhibit pseudo buds, such as areas of glandular fragmentation or mucin, need to be avoided.

A framework was created in Python to score buds using our developed deep



a) Real H&E

b) Virtual Pan-Cytokeratin AE1/AE3

Figure 4.1: An example of how the ground truth dataset of manual tumour bud segmentations was created. For each slide in our training dataset, a virtual IHC image was generated from the H&E whole slide image, shown in b) and used as a manual reference for the ground truth segmentation masks shown in a). The ITBCC scoring guidelines were used to create the dataset. The segmentations were drawn manually using QuPath (Bankhead et al. 2017).

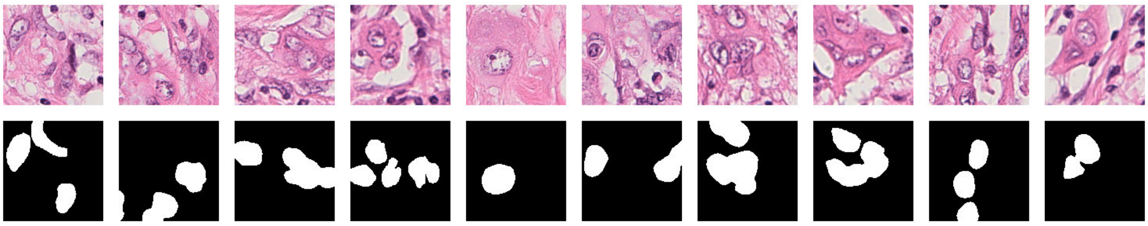


Figure 4.2: Selected examples of the resulting binary segmentation masks generated from the manual tumour bud annotations. Each bud is randomly located within a tile; other visible buds are also included. The corresponding region is cropped from the H&E whole slide image and output to create a dataset of paired H&E and ground truth segmentation masks.

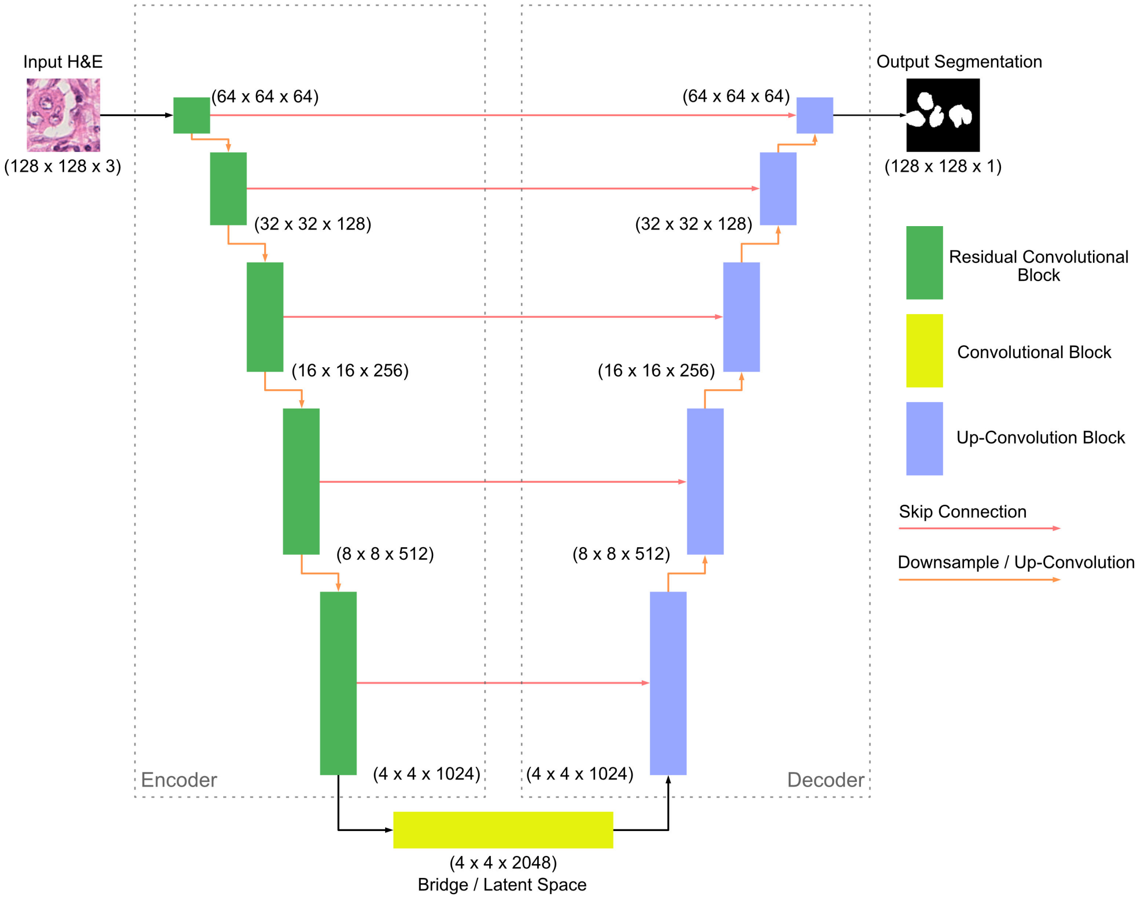


Figure 4.3: The architecture of the Resnet50 U-Net model used for automated tumour bud scoring. The Resnet50 encoder backbone is pre-trained on ImageNet and updated during training. The U-Net decoder is initialised with random weights. An input H&E tile is passed through the network, and the output is a segmentation mask of the same size as the input.

network. The following process, shown in figure 4.4, was implemented to automatically position the required 0.785mm^2 fields along the entire invasive margin of a whole slide image to allow bud scoring in the correct regions. The first step in this process, shown in panel a), was to convert the input H&E whole slide image to a virtual AE1/AE3 version using our virtual IHC model from chapter 3. This was then converted to the LAB colour space using the OpenCV (Culjak et al. 2012) Python library.

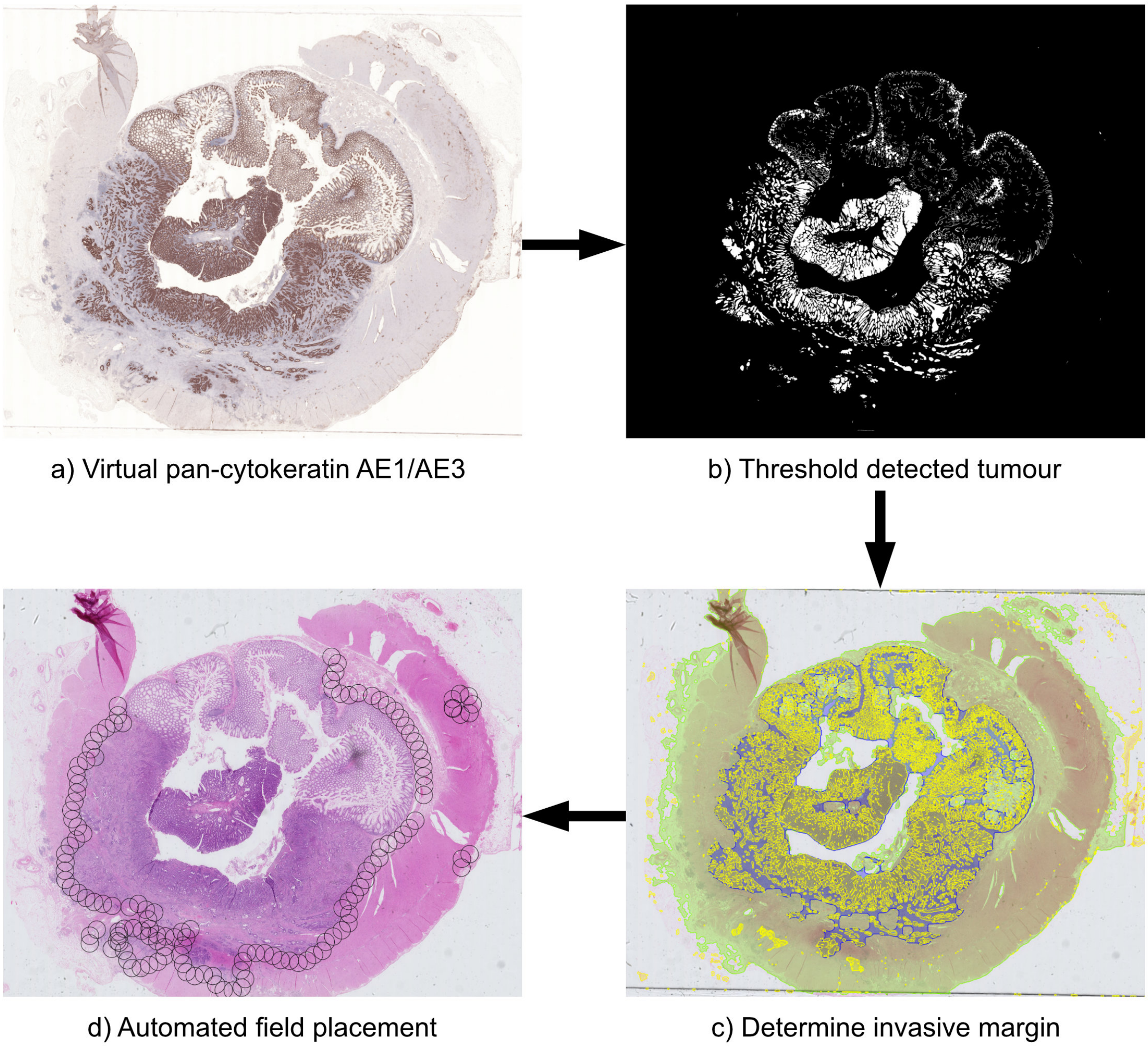

a) Virtual pan-cytokeratin AE1/AE3
b) Threshold detected tumour
d) Automated field placement
c) Determine invasive margin

Figure 4.4: Diagrammatic representation of the automated procedure for objective field placement in tumour bud scoring. a) Depicts the input whole slide image, which is subsequently transformed into a virtual pan-cytokeratin image. b) Illustrates the resultant thresholding of cytokeratin within the virtual IHC image. c) Demonstrates the generated masks for the tumour core, invasive margin, and surrounding tissue, achieved through morphological operations on the cytokeratin threshold. d) Highlights the culmination of the process, where masks of the tumour core and invasive margin are amalgamated to determine the placement of the objective fields.

The brown pixels were thresholded in the virtual AE1/AE3 images to detect the stained regions representing the tumour-derived epithelial tissue. This binary mask was subjected to blurring, dilation, and erosion to get a cohesive representation of the tumour core. A threshold was then applied to the blue range of the colour channels using a similar process to retrieve a tissue mask. A visualisation of the retrieved binary mask representing the tumour core is given in panel b). The polygon representing the tumour core was then buffered outwards by 250 microns to determine the centre of the invasive margin. An intersection between this and the region of detected tissue was performed. The resulting coordinates represent a line encircling the leading edge of the tumour core that lies only on tissue. This line was then used to guide the placement of 0.785mm^2 fields. The fields were positioned every half field size and overlapped to ensure all tissue in the invasive margin was present in at least one field. The automatically placed fields are visible in panel d).

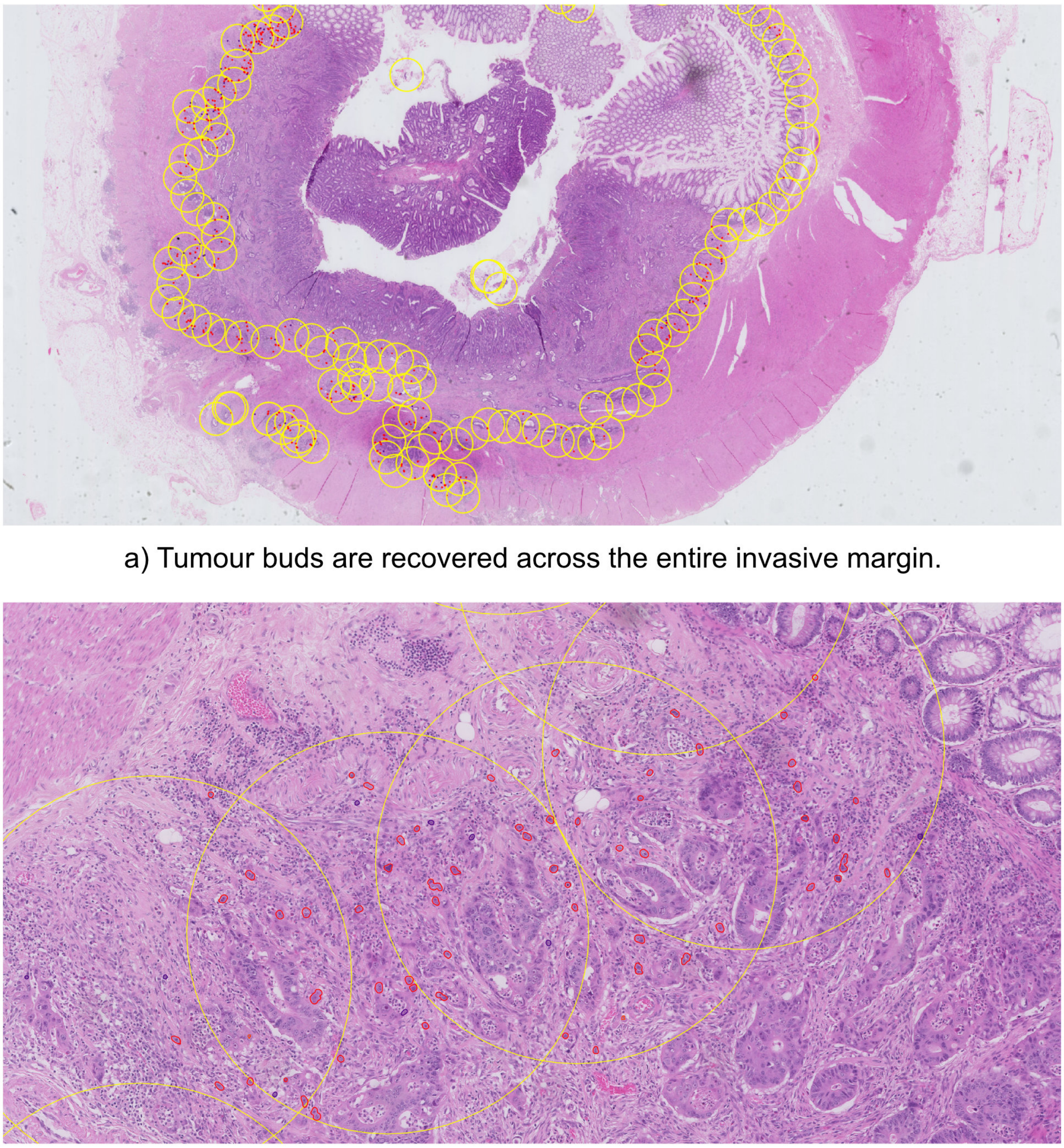
The fields were tessellated into tiles of size 128^2, with a stride of 64. A half-tile stride was employed so buds that span tiles were not missed. The tiles were then passed through the trained U-Net. OpenCV (Culjak et al. 2012) was used to recover the polygons from the binary segmentation masks generated by the deep network. The coordinates of these polygons were then translated back into the reference frame of the whole slide image. An example of a whole slide image with the recovered buds shown in red can be observed in figure 4.5 panel a). Finally, each field in the invasive margin was looped over, and the number of buds within its boundary was tracked and counted as shown in panel b) of figure 4.5. The field with the highest number of buds determines the bud score for that slide. The proposed protocols for our end-to-end automated tumour bud scoring process are summarised in algorithm 2 on page 144.

Network Evaluation and Tissue Classification

Candidate deep segmentation models were trained and selected for competence using the dice score as the loss measure. The final performance evaluation was a combination of classification metrics calculated for each cohort on the predicted versus manual tumour bud score, including accuracy, precision, recall, specificity and f1 score. The litmus test for model success was its performance during survival analysis on the GRI and AP cohorts of 689 and 609 patients with survival histories dating back over ten years and the independence of automated tumour bud score in univariate and multivariate Cox proportional hazard analysis.

Our study also investigated the influence of microenvironment tissue type on the prognostic ability of tumour buds. To accomplish this task, we utilised a Resnet50 neural network, as described in section 2.7.2, which was trained to classify colorectal tissue. We used the default and architecture described in the paper, except we removed the final dense layer and replaced it with a 9-neuron dense layer for supervised prediction of tissue class using the public colorectal tissue dataset, NCT-100K, published by Kather et al. (2018). The model was trained for 1000 epochs, with a ground truth training and validation split of 80% and 20%, respectively, a batch size 256 and a learning rate 2e-4 with an Adam optimiser.

This combination of techniques allowed us to do a comparative analysis of classification accuracy at the patient level for automated versus manual tumour bud scoring, investigate how microenvironment tissue type affected the predictive power of the bud score and analyse the overall survival at the cohort level.



a) Tumour buds are recovered across the entire invasive margin.

b) The field with highest density is located to determine the score.

Figure 4.5: An illustration of the automated tumour bud scoring methodology. a) Displays the input H&E whole slide image, where buds are identified across the entire invasive margin utilizing the trained U-Net. b) Depicts the procedure of enumerating the buds within a 0.785mm^2 objective field to pinpoint the area with the highest bud density. The bud count within this designated field subsequently informs the final tumour bud score.

Algorithm 2 Protocols for Automated Tumour Bud Scoring.

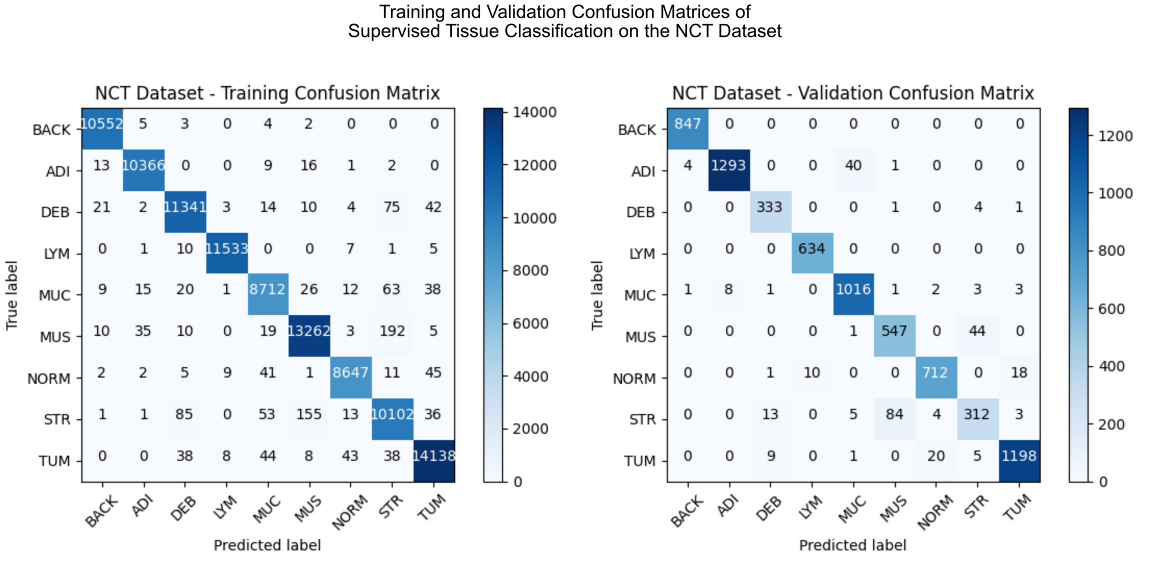
```
procedure Automated Tumour Bud Scoring on H&E
    Step 1: Perform stain normalisation on the H&E WSI pathology image:
        Use the BEAST toolkit to obtain a normalised image.
    Step 2: Convert the normalised image to a virtual IHC AE1/AE3 image:
        Use the virtual stain translation model from chapter 3.
    Step 3: Convert the virtual IHC WSI to the LAB colour space:
        Threshold to retrieve tissue pixels as a binary mask.
        Threshold to retrieve brown-coloured pixels representing the epithelial-derived
        tumour tissue as another mask.
    Step 4: Predict the centre of the invasive margin:
        Buffer the binary tumour mask from step 3 by 250 microns.
    Step 5: Place objective fields around the tumour:
        Use the invasive margin centre line determined in step 4 as a guide to place
        objective fields.
        Overlap another field every half field size.
    Step 6: Tessellate regions in the H&E image corresponding to the fields, and
infer tumour buds:
        Pass the tile through the bud segmentation U-Net to generate the tumour bud
        masks.
        Use OpenCV python to recover polygons from segmentations.
        Convert coordinates back to the reference frame of the WSI.
    Step 7: Score the slide based on the number of detected buds:
        Initialize $max\_buds \leftarrow 0$ and $score \leftarrow$ NULL
        for each $field$ do
            $buds \leftarrow$ count detected buds in $field$
            if $buds < 10$ then
                $current\_score \leftarrow$ Low Budding
            else
                $current\_score \leftarrow$ High Budding
            end if
            if $buds > max\_buds$ then
                $max\_buds \leftarrow buds$
                $score \leftarrow current\_score$
            end if
        end for
        return $score$
end procedure
```

4.3 Results

This section outlines the results of our proposed method for automated tumour bud scoring, including the performance of the tissue classification model, the tumour bud segmentation model, classification statistics on the automated and manual budding scores, Cohen's kappa metrics for correlation with manual budding and the survival analysis results of automated budding scores determined by buds detected in many combinations of colorectal tissue classes. We also analysed the utility of the most performant configurations as an independent biomarker over the GRI and AP patient cohorts described in section 4.2.1.

4.3.1 Tissue Classification Model

The architecture of our tissue classification model was based on the ResNet50 network as described in section 2.7.2 and trained as described in section 4.2.2. The epoch with the highest level of performance was 854 of 1000, where it reached a validation accuracy of 96.12%, with a precision of 94.62% and a recall of 94.75%. The confusion matrices for the training and validation sets are shown in figure 4.6, and the performance metrics for the model are summarised in table 4.3.1.



a) b)

Figure 4.6: The confusion matrices for the tissue classification model, trained on the NCT-100K dataset. Panel a) shows the performance on the training set, and panel b) shows the performance on the validation set.

Metric	Value
Training Accuracy	0.9867
Validation Accuracy	0.9612
Training Precision	0.9864
Validation Precision	0.9462
Training Recall	0.9865
Validation Recall	0.9475

Table 4.3: The performance metrics of the tissue classification model on the NCT-100K dataset.

4.3.2 Automated Tumour Bud Scoring Model

The automated tumour bud scoring model was trained on the dataset described in section 4.2.2 and evaluated on a 20% split held out as the validation set. The model was trained for 1000 epochs, and the epoch with the lowest dice loss, epoch 760, was chosen as the best-performing model. The performance metrics for the model are summarised in table 4.3.2. The model achieved a pixel accuracy of 99.87%, a dice score of 0.9628 on the training set, a pixel accuracy of 99.63%, and a dice score of 0.9249 on the validation set. The model was then used to recover tumour bud segmentation masks across all tissue within the whole slide images of the GRI and AP cohorts. This allowed us to analyse the predictive ability of the automated tumour bud score using buds with various areas located at varying distances from the tumour edge, residing across multiple tissue types as determined by the tissue classification model. Figure 4.7 shows some examples of the output segmentations on the validation set and how they correspond to the ground truth. Each group displays a stain normalised and augmented H&E input tile, followed by the ground truth manual and resulting automated segmentation. We recovered buds from all tissue in the slides to allow us to determine the distributions of area and distance from the tumour edge to define a search space of the parameters to use in our final automated bud scoring protocols. These could be combined with tissue type to allow us to determine the most predictive cut-offs that define the area of a bud, the distance from the tumour edge, and the tissue type that maximises the predictive ability of the automated tumour bud score. The results of this analysis are presented in the following sections.

Metric	Value
Training Pixel Accuracy	0.9987
Validation Pixel Accuracy	0.9963
Training Dice Score	0.9628
Validation Dice Score	0.9249

Table 4.4: The performance metrics of the H&E tumour bud segmentation model on the validation dataset.

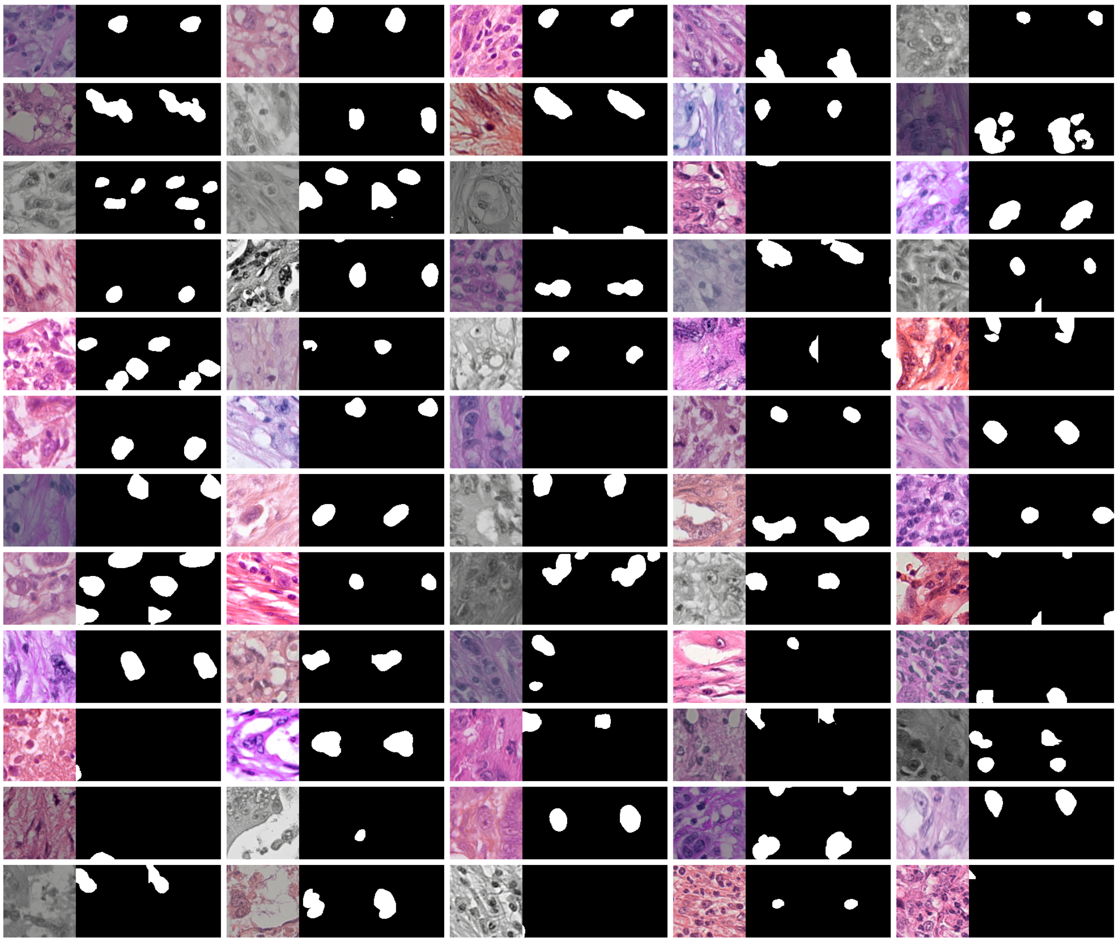
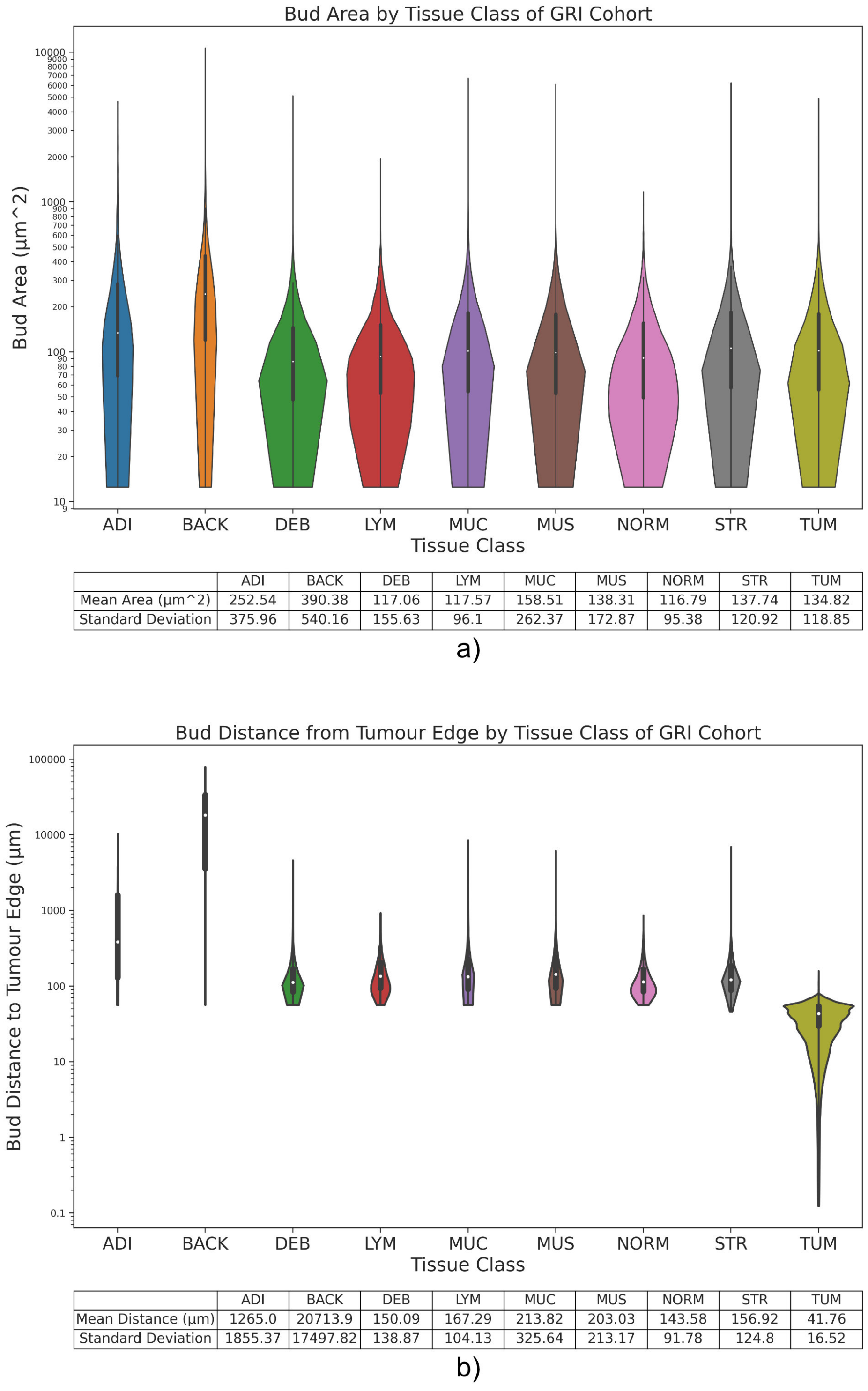


Figure 4.7: Displayed are representative tumour bud segmentations from the validation dataset. Each set has three images. The leftmost image presents the normalised and stain-augmented H&E tile. The central image showcases the ground truth segmentation mask. The rightmost image reveals the segmentation mask produced by the trained U-Net.

4.3.3 GRI Cohort

We plotted the distributions of all detected buds in the GRI cohort to determine the optimal filter parameters to remove false detections, pseudo buds and non-relevant cells based on area and distance from the tumour core. This plot can be viewed in figure 4.8. Panel a) displays the area distribution of the detected cells, and panel b) displays the distance distribution. The area is measured in microns squared, and the distance in microns. The distributions are categorised by the tissue class of the microenvironment in which the cells reside. The area distribution shows that most buds were less than 1,000 microns squared, with outliers ranging up to 10,000. The distance distribution shows that most detected buds were within 500 microns of the tumour edge. We used these distributions to define the search space for the optimal filter parameters for our automated tumour bud scoring protocols. We performed a parameter search over every combination of minimum and maximum thresholds for defining a tumour bud and several varieties of the tissue classes in which the detected buds reside. As a baseline, we show the results for all tissue classes. After this, the adipose, background, debris, and mucin classes were removed per the 2016 ITBCC guidelines. An analysis was then performed of the Cox regression hazard ratio and Cohen's kappa values of the automated bud scores determined over the entire cohort as calculated by each parameter and tissue combination to evaluate that configuration's utility. A sweep of the required threshold for high-grade budding per objective field size was also performed to ascertain the optimal setting for each combination of area, distance and tissue type. The Cohen's kappa and Cox regression hazard ratios resulting from the parameter sweep over selected tissue class combinations are shown in figure 4.9.

Panel a) of figure 4.9 shows Cohen's kappa values of the sweep. The highest values, which correlate with manual bud scores, result from buds detected in the lymphocyte-muscle-stroma-tumour tissue combination, with the lowest correlation resulting from buds residing in muscle tissue alone. In panel b) of figure 4.9, the hazard ratio values of the sweep can be observed. The most significant hazard ratios result from buds detected in the muscle-tumour tissue combination, with some of the lowest hazard ratios resulting from buds residing across lymphocyte-muscle-normal-stroma-tumour tissue. A closer analysis of the utility of budding in all tissue, lymphocyte-muscle-normal-stroma-tumour, and muscle-tumour tissue combinations was performed. The results of this analysis are given in the sections below.



	ADI	BACK	DEB	LYM	MUC	MUS	NORM	STR	TUM
Mean Area (µm^2)	252.54	390.38	117.06	117.57	158.51	138.31	116.79	137.74	134.82
Standard Deviation	375.96	540.16	155.63	96.1	262.37	172.87	95.38	120.92	118.85

a)

	ADI	BACK	DEB	LYM	MUC	MUS	NORM	STR	TUM
Mean Distance (µm)	1265.0	20713.9	150.09	167.29	213.82	203.03	143.58	156.92	41.76
Standard Deviation	1855.37	17497.82	138.87	104.13	325.64	213.17	91.78	124.8	16.52

b)

Figure 4.8: Shown are the distributions of bud area and distance from the tumour edge of the detected buds in the GRI cohort categorised by the tissue class of the microenvironment in which the bud resides. Panel a) shows the area distribution measured in microns squared, and panel b) shows the distance distribution measured in microns.

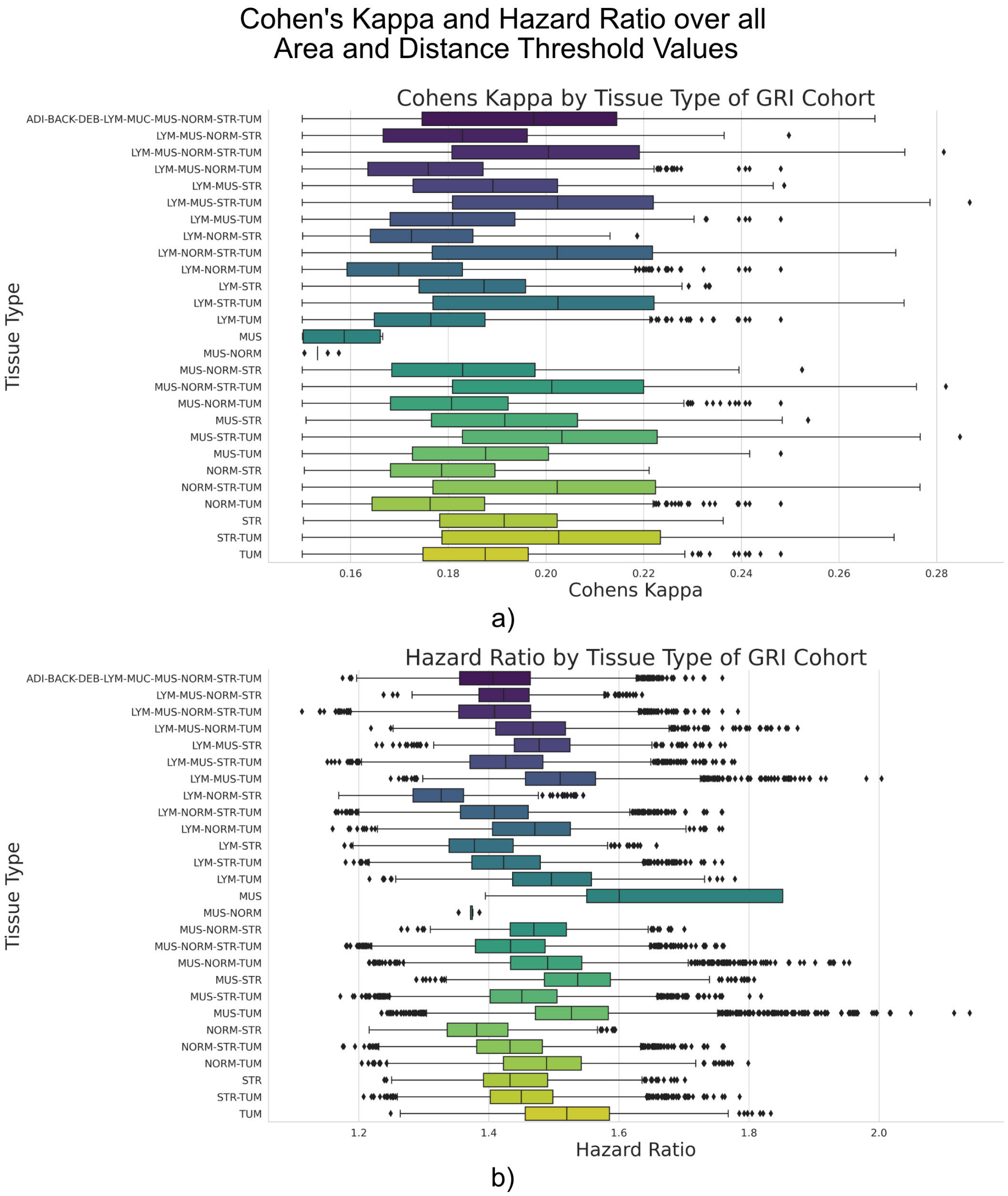
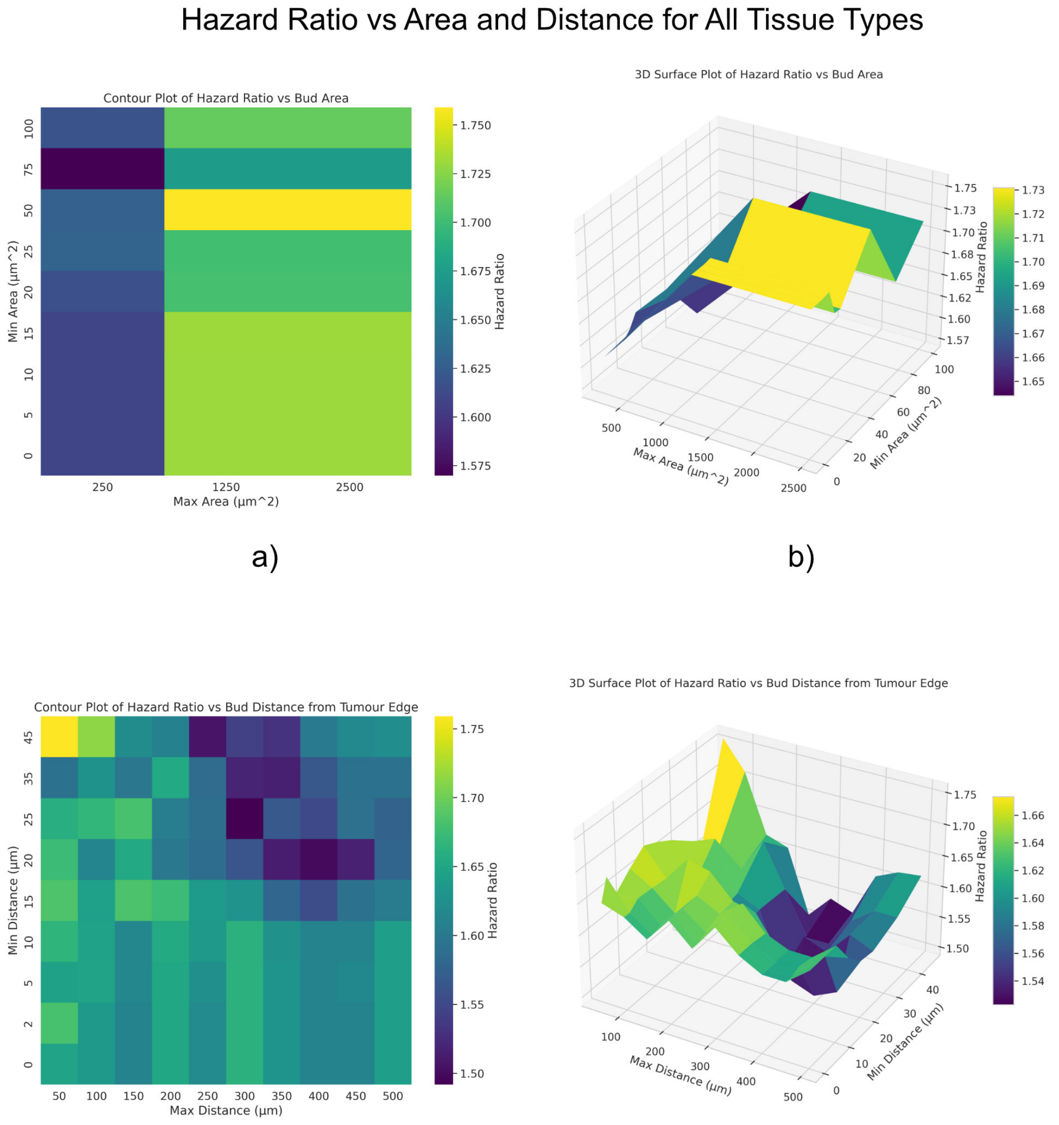


Figure 4.9: Displayed are the results from the parameter sweep over the area, distance, and high-grade budding threshold parameters for the GRI cohort. Values are shown for each combination of parameters and tissue type. Panel a) shows the Cohen's kappa values, and panel b) shows the hazard ratio values.

We analysed the utility of buds residing in all tissue classes initially. To highlight the trends in the parameter space, we plotted a 3D surface plot of the Cox regression hazard ratio over all combinations of minimum and maximum area and distance values. We also plotted a 2D heatmap of the values to visualise the results when the 3D plot has obscured values. These plots can be viewed in figure 4.10. The most prognostic buds have a maximum area larger than 1000 microns, with the minimum area only marginally affecting the outcome. The distance sweep was the most revealing in this tissue combination, with the most prognostic buds having a maximum distance of no more than 200 microns from the tumour front and a minimum distance of more than 30 microns.

The Kaplan Meier plots and survival analysis statistics resulting from detected buds residing in all tissue types are displayed in figure 4.11. Panel a) shows the Kaplan Meier plot of patients without filtering and the determined cutoff of 24 buds per 0.786mm^2 objective field. Panel b) displays the Kaplan Meier plot of bud scores over the GRI cohort determined after the filtering of detected buds based on the most performant values of a distance minimum of 45μm, area minimum of 50μm^2, a distance maximum of 50μm and an unlimited area maximum.

The 2016 ITBCC guidelines state that mucinous, adipose tissue and debris should be avoided due to the presence of pseudobuds as described by Haddad et al. (2023). Therefore, we repeated the analysis on a subset of tissue combinations with these regions removed. The results of this analysis are shown in figure 4.12. The most prognostic buds once again have a maximum area larger than 1000 microns, with the minimum area only negligibly affecting the result. The distance sweep was also the most notable in this tissue combination. The most prognostic buds were again close to the tumour core, having a maximum distance of no more than 200 microns from the tumour front and a minimum distance of more than 30 microns. Buds further out were even less prognostic than in all tissue types. Kaplan Meier plots and the survival analysis results performed on detected buds in lymphocyte, muscle, normal, stroma, and tumour tissue are displayed in figure 4.13.



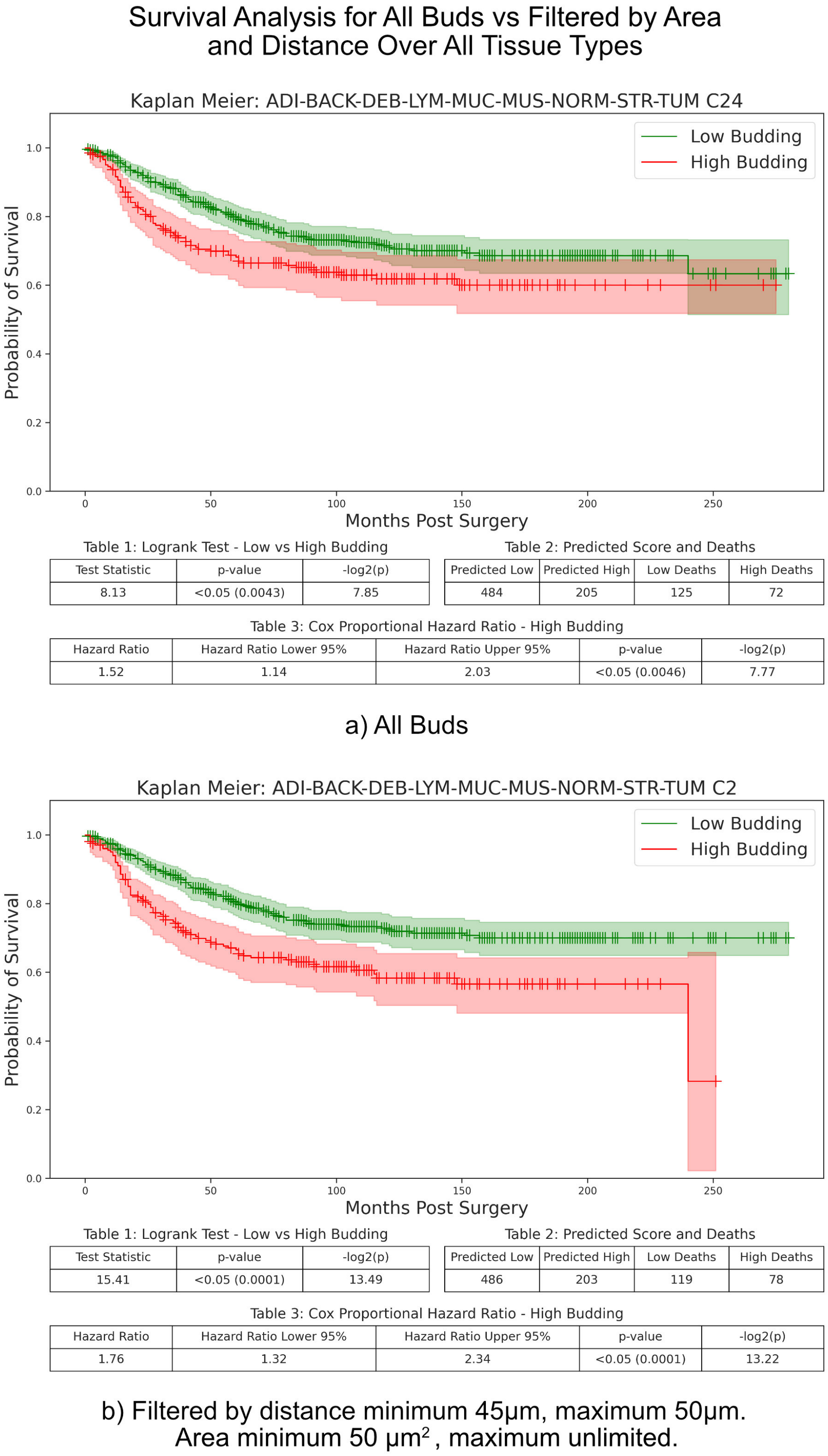
a)

b)

c)

d)

Figure 4.10: Shown is the hazard ratio of the area and distance threshold parameters over all tissue classes in the GRI cohort. Panel a) displays a 2D heatmap of the hazard ratios resulting from the minimum and maximum area thresholds evaluated. Panel b) displays a 3D contour plot of the values to better highlight change over the parameter space. Panel c) shows a 2D heatmap of the hazard ratios resulting from the minimum and maximum distance thresholds evaluated. Panel d) displays a 3D contour plot of the values.



Survival Analysis for All Buds vs Filtered by Area and Distance Over All Tissue Types

Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
8.13	<0.05 (0.0043)	7.85

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
484	205	125	72

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
1.52	1.14	2.03	<0.05 (0.0046)	7.77

a) All Buds

Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
15.41	<0.05 (0.0001)	13.49

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
486	203	119	78

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
1.76	1.32	2.34	<0.05 (0.0001)	13.22

b) Filtered by distance minimum 45µm, maximum 50µm.
Area minimum 50 µm^2, maximum unlimited.

Figure 4.11: Kaplan Meier plot of automated bud score with no filtering and the best-determined cutoff of 24 buds per 0.785mm^2 objective field. Panel b) displays the Kaplan Meier plot of automated bud scores over the GRI cohort after filtering of detected buds based on the most performant values of a distance minimum of 45μm, area minimum of 50μm^2, a distance maximum of 50μm and an unlimited area maximum, with the best cutoff of 2 buds per 0.785mm^2 objective field.

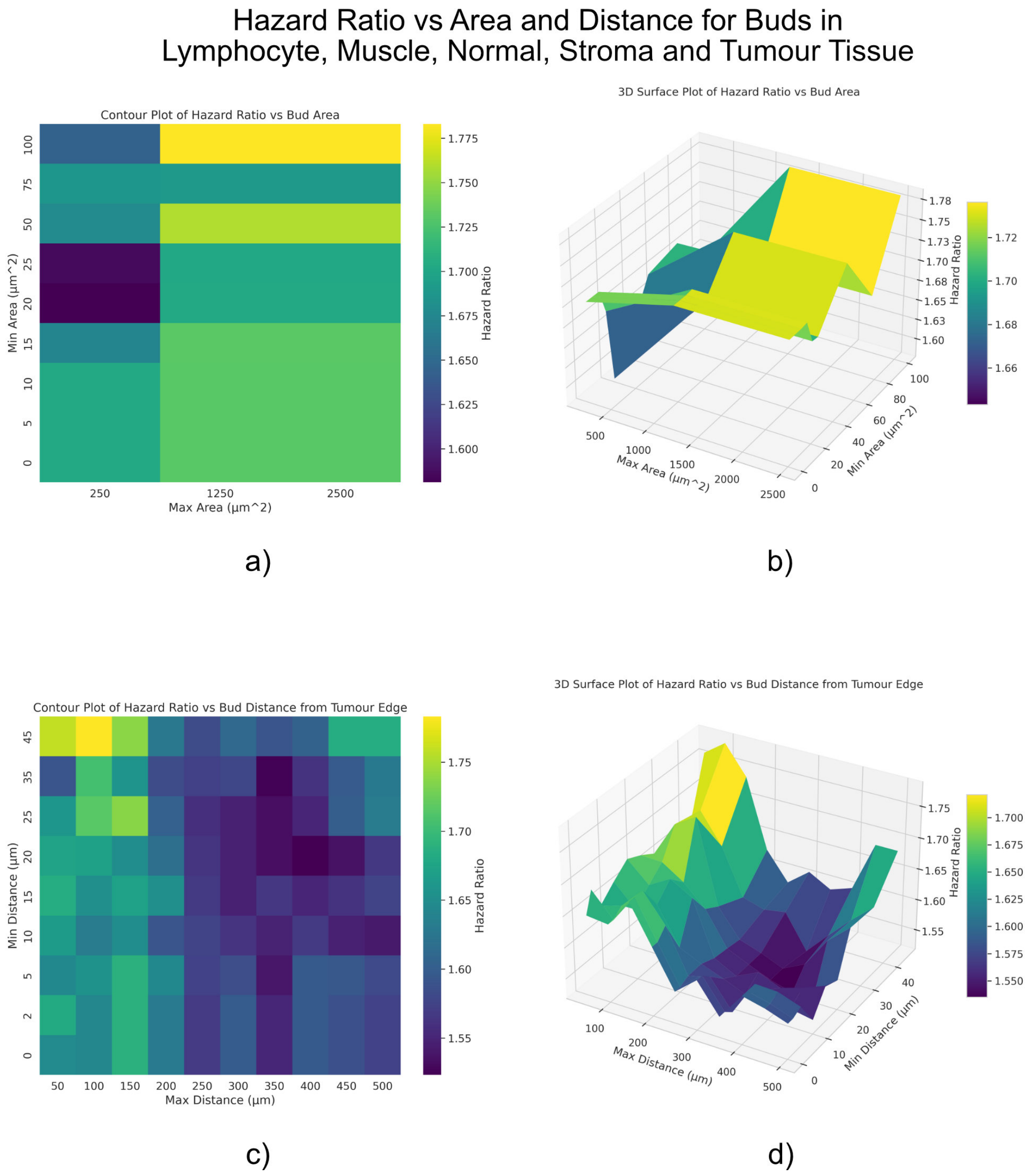
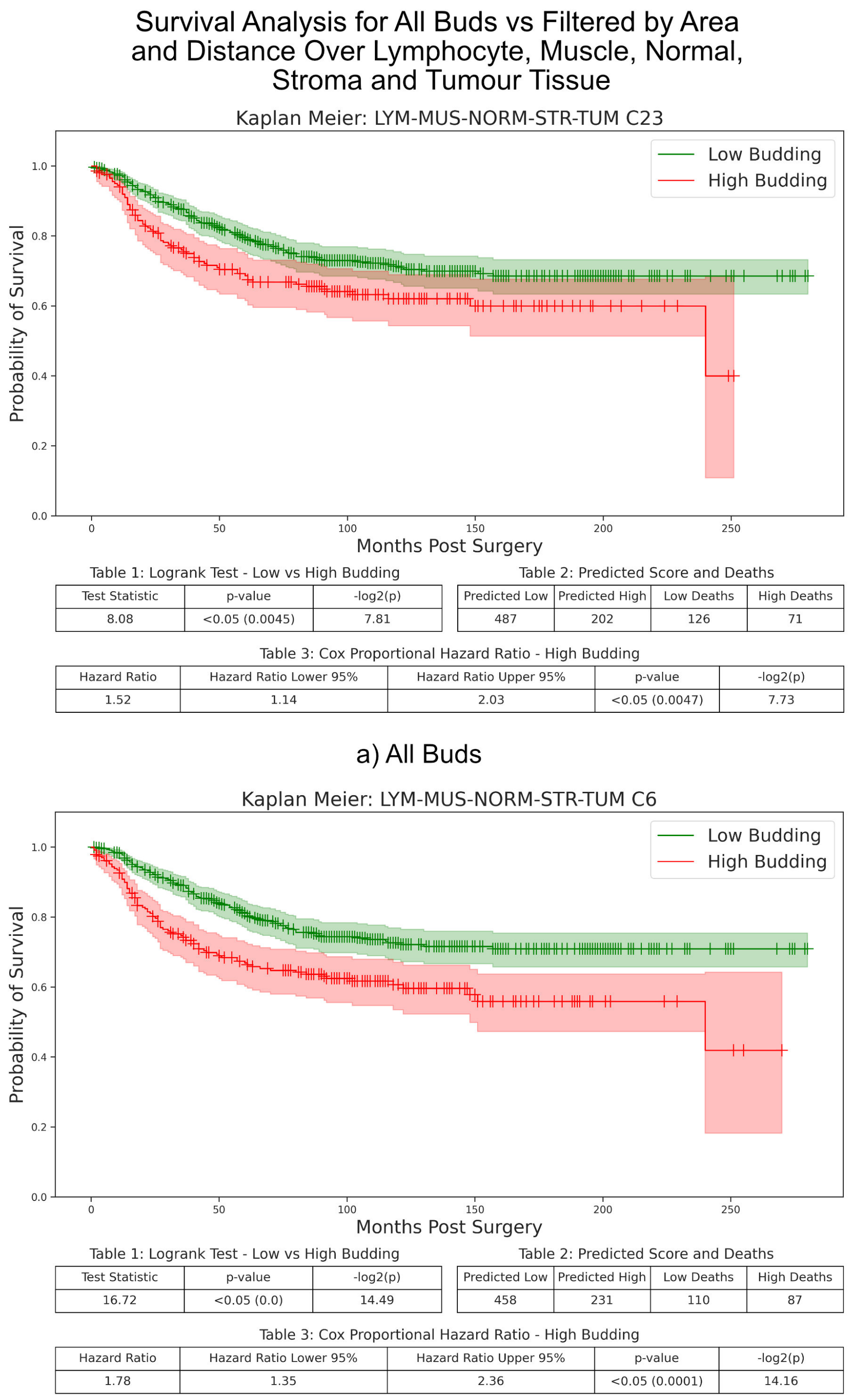


Figure 4.12: Shown is the hazard ratio of the area and distance threshold parameters over the lymphocyte, muscle, normal, stroma, and tumour tissue classes in the GRI cohort. Panel a) displays a 2D heatmap of the hazard ratios resulting from the minimum and maximum area thresholds evaluated. Panel b) displays a 3D contour plot of the values to better highlight change over the parameter space. Panel c) shows a 2D heatmap of the hazard ratios resulting from the minimum and maximum distance thresholds evaluated. Panel d) displays a 3D contour plot of the values.



Survival Analysis for All Buds vs Filtered by Area and Distance Over Lymphocyte, Muscle, Normal, Stroma and Tumour Tissue

Kaplan Meier: LYM-MUS-NORM-STR-TUM C23

Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
8.08	<0.05 (0.0045)	7.81

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
487	202	126	71

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
1.52	1.14	2.03	<0.05 (0.0047)	7.73

a) All Buds

Kaplan Meier: LYM-MUS-NORM-STR-TUM C6

Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
16.72	<0.05 (0.0)	14.49

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
458	231	110	87

Table 3: Cox Proportional Hazard Ratio - High Budding

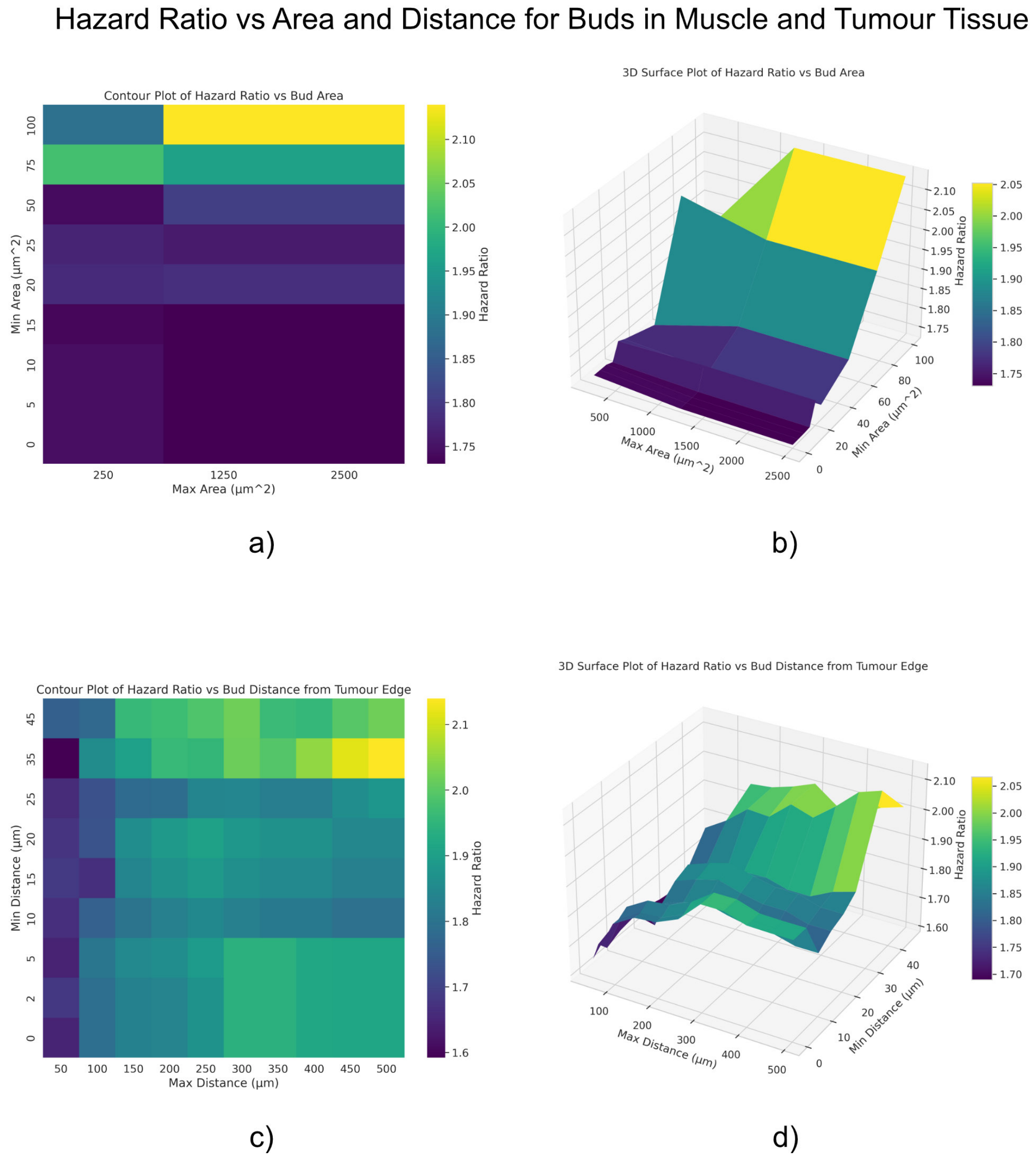
Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
1.78	1.35	2.36	<0.05 (0.0001)	14.16

b) Filtered by distance minimum 45μm, maximum 100μm.
Area minimum 100 μm^2, maximum unlimited.

Figure 4.13: Kaplan Meier plot of automated bud score with no filtering and the best-determined cutoff of 23 buds per 0.785mm^2 objective field. Panel b) displays the Kaplan Meier plot of automated bud scores over lymphocyte, muscle, normal, stroma and tumour tissue in the GRI cohort after the filtering of detected buds based on the most performant values of a distance minimum of 45μm, area minimum of 100μm^2, a distance maximum of 100μm and an unlimited area maximum, with the best cutoff of 6 buds per 0.785mm^2 objective field.

Next, we analysed the prognostic ability of the most performant tissue combination: muscle and tumour. The results of this analysis are shown in figure 4.14. The most prognostic buds again have a maximum area larger than 1000 microns, but the minimum area has become significant, with values over 80 μm^2 being the most predictive. The distance sweep was again the most notable in this tissue combination. The most prognostic buds were located far from the tumour core, having a maximum distance of more than 450 microns from the tumour front and a minimum distance of more than 40 microns. In a reversal of the other tissue combinations, large, distant buds were now the most predictive. Kaplan Meier plots and the survival analysis results performed on detected buds residing in muscle and tumour tissue are displayed in figure 4.15.

A comparison of the survival curves of the GRI cohort using manual and automated tumour bud scoring is given in figure 4.16. The automated bud scoring results in a higher hazard ratio with greater deaths in the predicted high aggression population and a smaller number in the low group when compared to manual budding. The classification metrics for the best-performing automated bud scoring model on the GRI cohort using manual tumour bud scoring as ground truth are shown in table 4.3.3. The model achieved an accuracy of 65.60%, a precision of 42.39%, a recall of 45.10%, an F1 score of 43.71%, and a Cohen's kappa of 0.1898, which shows moderate agreement with manual budding. The hazard ratio of the model was larger than manual budding at 2.14 (1.62-2.83).



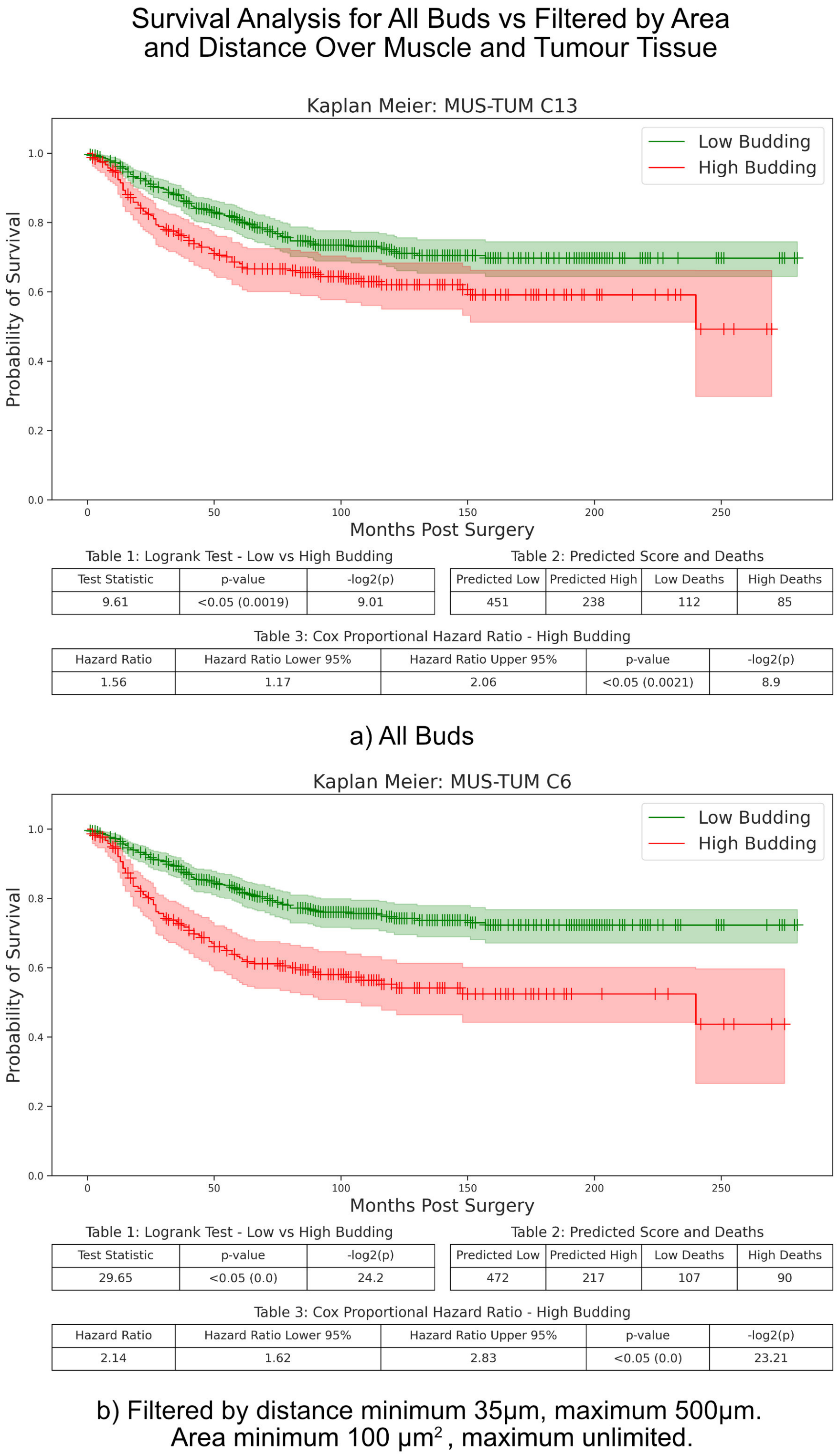
a)

b)

c)

d)

Figure 4.14: Shown is the hazard ratio of the area and distance threshold parameters over muscle and tumour tissue classes in the GRI cohort. Panel a) displays a 2D heatmap of the hazard ratios resulting from the minimum and maximum area thresholds evaluated. Panel b) displays a 3D contour plot of the values to better highlight change over the parameter space. Panel c) shows a 2D heatmap of the hazard ratios resulting from the minimum and maximum distance thresholds evaluated. Panel d) displays a 3D contour plot of the values.



Survival Analysis for All Buds vs Filtered by Area and Distance Over Muscle and Tumour Tissue

Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
9.61	<0.05 (0.0019)	9.01

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
451	238	112	85

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
1.56	1.17	2.06	<0.05 (0.0021)	8.9

a) All Buds

Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
29.65	<0.05 (0.0)	24.2

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
472	217	107	90

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
2.14	1.62	2.83	<0.05 (0.0)	23.21

b) Filtered by distance minimum 35µm, maximum 500µm.
Area minimum 100 µm^2 , maximum unlimited.

Figure 4.15: Kaplan Meier plot of automated bud score with no filtering and the best-determined cutoff of 13 buds per 0.785mm^2 objective field. Panel b) displays the Kaplan Meier plot of automated bud scores over muscle and tumour tissue in the GRI cohort determined after the filtering of detected buds based on the most performant values of a distance minimum of 35μm, area minimum of 100μm^2, a distance maximum of 500μm and an unlimited area maximum, with the best cutoff of 6 buds per 0.785mm^2 objective field.

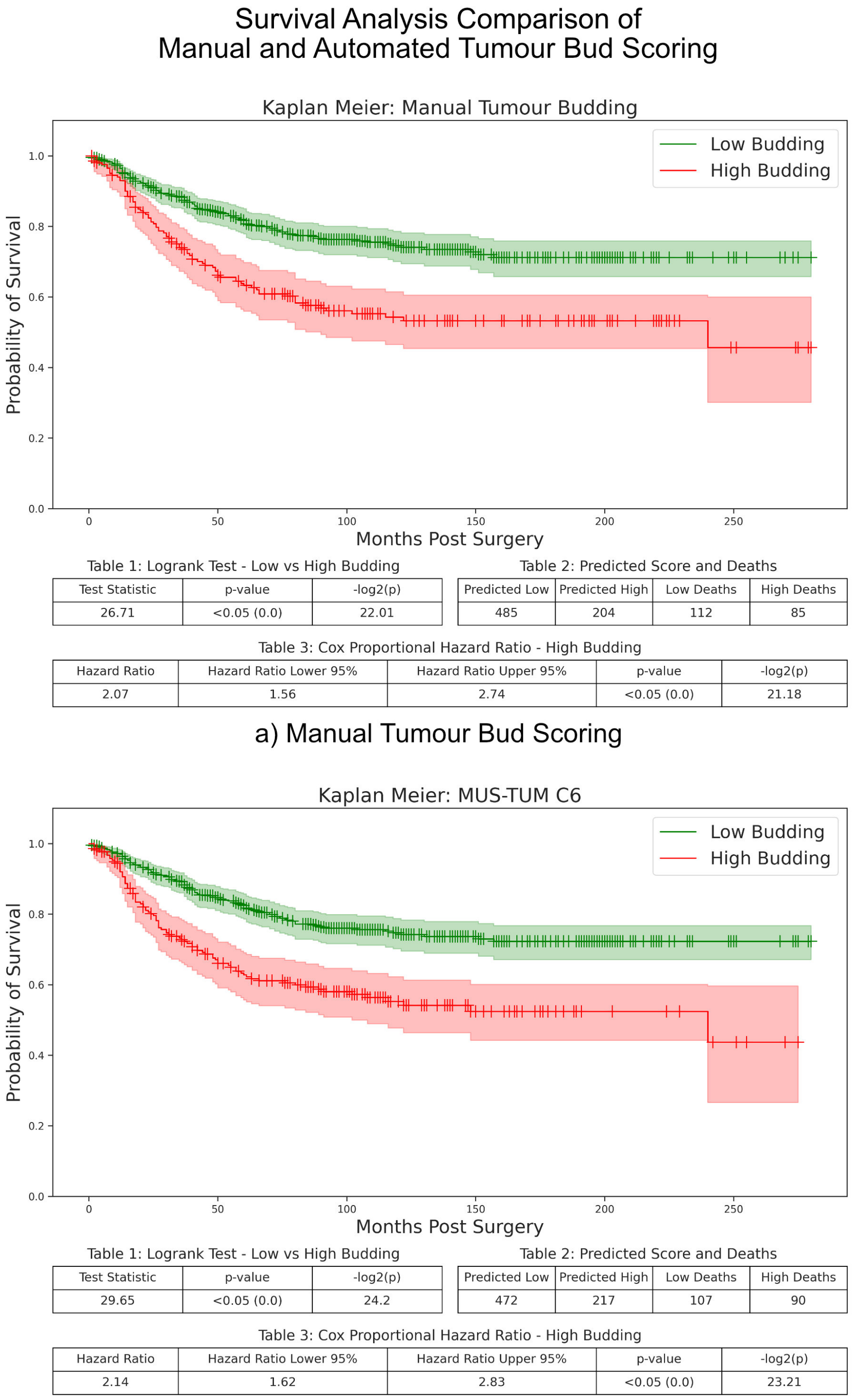


Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
26.71	<0.05 (0.0)	22.01

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
485	204	112	85

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
2.07	1.56	2.74	<0.05 (0.0)	21.18

a) Manual Tumour Bud Scoring

Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
29.65	<0.05 (0.0)	24.2

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
472	217	107	90

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
2.14	1.62	2.83	<0.05 (0.0)	23.21

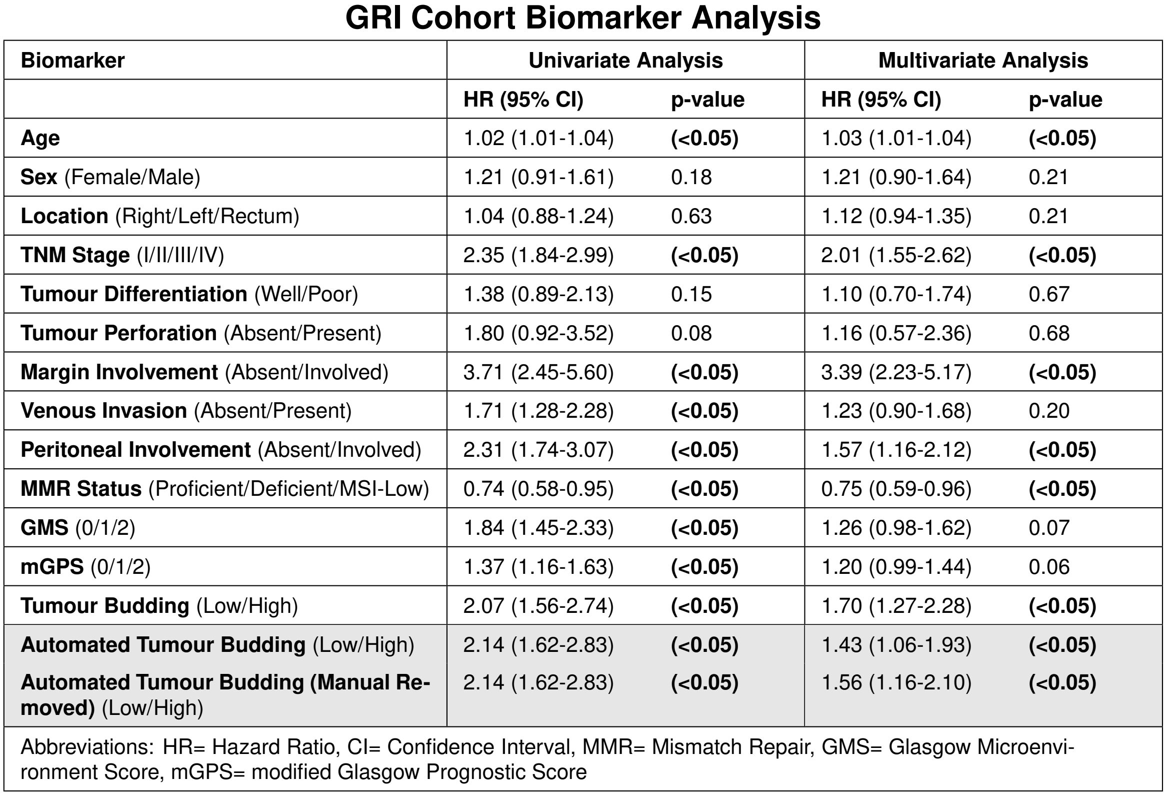
b) Best Performing Automated Tumour Bud Scoring
Model on Muscle and Tumour Tissue

Figure 4.16: Shown here is a comparison of the survival curves of the GRI cohort using manual and automated tumour bud scoring. Panel a) shows the Kaplan Meier plot of the survival of patients with manual tumour bud scores. Panel b) shows the Kaplan Meier plot of the survival of patients with automated tumour bud scores. Tables 2 and 3 note that the automated bud scoring results in a higher hazard ratio with a greater number of deaths in the predicted high aggression population and a smaller number in the low group.

Metric	Accuracy	Precision	Recall	F1	Cohen's Kappa	Hazard Ratio
Value	0.6560	0.4239	0.4510	0.4371	0.1898	2.14 (1.62-2.83)

Table 4.5: Best model performance on the GRI cohort using manual tumour bud scoring as ground truth.

This chapter aimed to develop a concept protocol for automated tumour bud scoring trained on virtual IHC ground truth that could stratify patients within a stage by tumour aggression. Therefore, we had to determine the utility of the most performant model configuration as an independent biomarker over the GRI patient cohort using Cox regression survival analysis. The results of this analysis are shown in table 4.3.3. The automated tumour bud score was a significant prognostic biomarker in univariate analysis, with hazard ratios of 2.14 (1.62-2.83). In multivariate analysis, with a backward conditional elimination model, the hazard ratio was 1.43 (1.06-1.93) with a p-value of less than 0.05, demonstrating that the automated tumour bud score was a significant prognostic biomarker independent of other clinicopathological variables.



GRI Cohort Biomarker Analysis

Biomarker	**Univariate Analysis**		**Multivariate Analysis**	
	HR (95% CI)	**p-value**	**HR (95% CI)**	**p-value**
Age	1.02 (1.01-1.04)	**(<0.05)**	1.03 (1.01-1.04)	**(<0.05)**
Sex (Female/Male)	1.21 (0.91-1.61)	0.18	1.21 (0.90-1.64)	0.21
Location (Right/Left/Rectum)	1.04 (0.88-1.24)	0.63	1.12 (0.94-1.35)	0.21
TNM Stage (I/II/III/IV)	2.35 (1.84-2.99)	**(<0.05)**	2.01 (1.55-2.62)	**(<0.05)**
Tumour Differentiation (Well/Poor)	1.38 (0.89-2.13)	0.15	1.10 (0.70-1.74)	0.67
Tumour Perforation (Absent/Present)	1.80 (0.92-3.52)	0.08	1.16 (0.57-2.36)	0.68
Margin Involvement (Absent/Involved)	3.71 (2.45-5.60)	**(<0.05)**	3.39 (2.23-5.17)	**(<0.05)**
Venous Invasion (Absent/Present)	1.71 (1.28-2.28)	**(<0.05)**	1.23 (0.90-1.68)	0.20
Peritoneal Involvement (Absent/Involved)	2.31 (1.74-3.07)	**(<0.05)**	1.57 (1.16-2.12)	**(<0.05)**
MMR Status (Proficient/Deficient/MSI-Low)	0.74 (0.58-0.95)	**(<0.05)**	0.75 (0.59-0.96)	**(<0.05)**
GMS (0/1/2)	1.84 (1.45-2.33)	**(<0.05)**	1.26 (0.98-1.62)	0.07
mGPS (0/1/2)	1.37 (1.16-1.63)	**(<0.05)**	1.20 (0.99-1.44)	0.06
Tumour Budding (Low/High)	2.07 (1.56-2.74)	**(<0.05)**	1.70 (1.27-2.28)	**(<0.05)**
Automated Tumour Budding (Low/High)	2.14 (1.62-2.83)	**(<0.05)**	1.43 (1.06-1.93)	**(<0.05)**
Automated Tumour Budding (Manual Removed) (Low/High)	2.14 (1.62-2.83)	**(<0.05)**	1.56 (1.16-2.10)	**(<0.05)**

Abbreviations: HR= Hazard Ratio, CI= Confidence Interval, MMR= Mismatch Repair, GMS= Glasgow Microenvironment Score, mGPS= modified Glasgow Prognostic Score

Table 4.6: Univariate Cox regression survival analysis was used to determine hazard ratios (HR) and 95% Confidence Intervals (CI) for available GRI cohort biomarkers. Multivariable Cox regression survival analysis using a backward conditional elimination model and a statistical significance threshold of 0.05 was performed to identify independent prognostic biomarkers. Note that the automated tumour bud score is an independent prognostic biomarker.

4.3.4 AP Cohort

To confirm the ability of our model to generalise across a dataset prepared and scanned at a different institution under different conditions, we were provided with a second patient cohort called the AP cohort as described in section 4.2.1. This cohort was scanned on older scanners at a different hospital from an earlier period to the GRI cohort. Unfortunately, the scanned images had many artefacts, with the majority consisting of out-of-focus areas scanning errors, and the most pervasive was pen drawn around the tumour edge. This made them less than ideal for digital analysis. However, the cohort contained over two thousand slides and an extensive clinicopathological database, so they still had value in allowing us to determine if our model could generalise and perform on sub-optimal whole slide images.

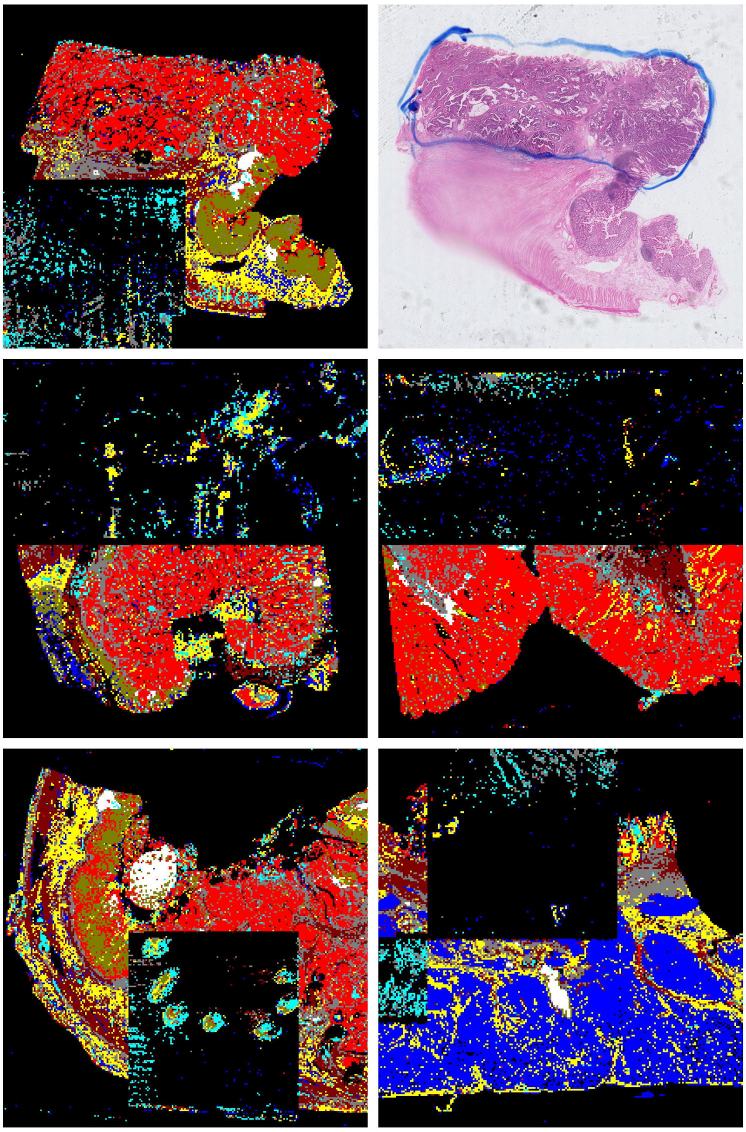


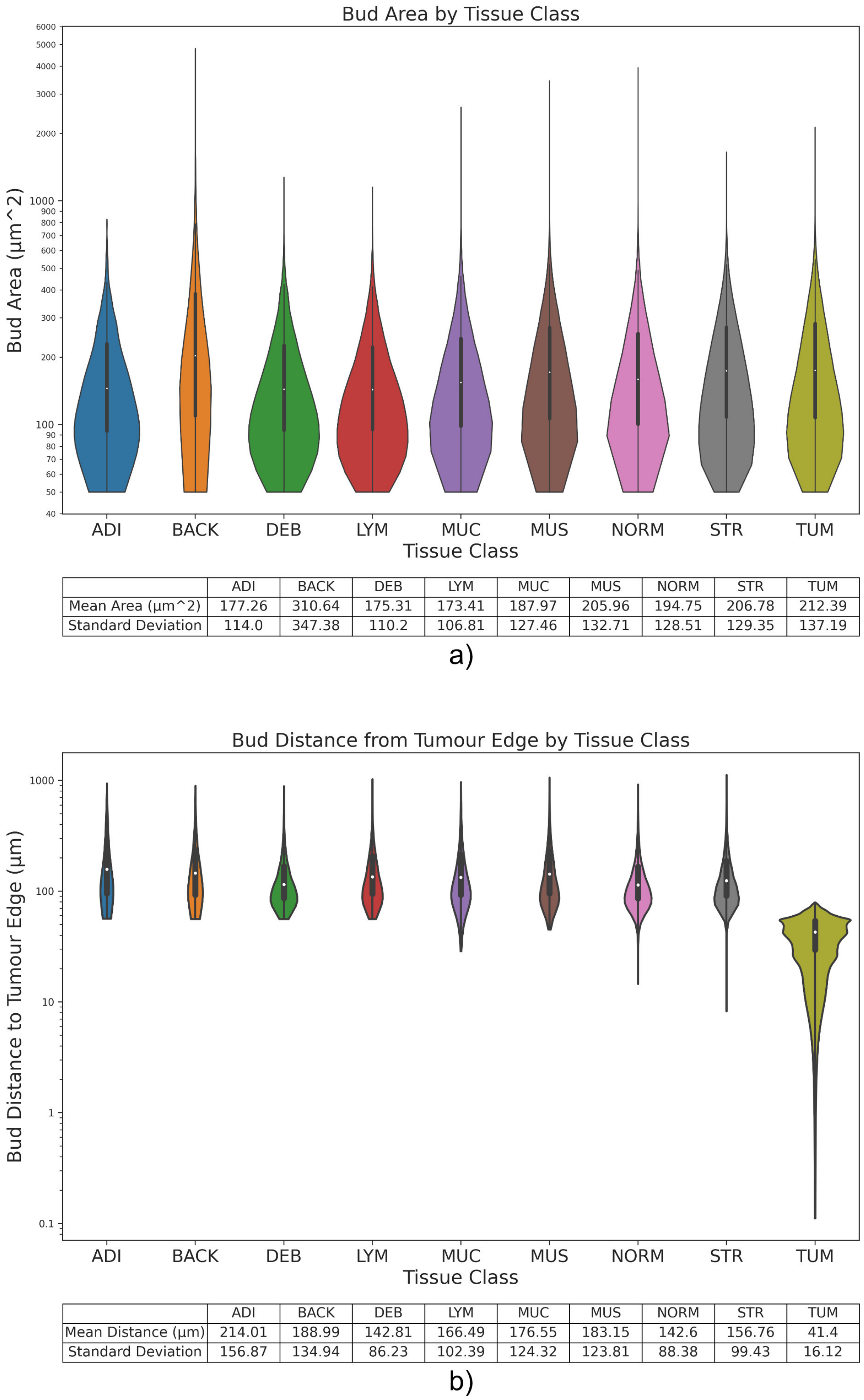
Figure 4.17: Due to the older nature of the AP cohort, it had a litany of scanning artefacts ranging from unfocused regions to pen drawn on the slide. These artefacts caused the tissue classification model to fail over large regions of the whole slide images.

The tissue classification stage experienced the worst performance issues of the pipeline, given that it was trained in a supervised manner on a dataset that specifically excluded these types of artefacts. Figure 4.17 shows examples of the issues encountered in the AP cohort. Unfocused areas were entirely misclassified, which occurred most often in the regions closest to the areas with ink on the glass. These were usually the areas around the tumour core, the most pertinent areas for this task. However, the model could still classify the tissue successfully in the regions without focus, scanning or ink artefacts.

To analyse the results of our model on the AP cohort, it was first tested with the parameter settings determined from the GRI cohort. For comparison, the best available performance of the deep segmentation model was determined by repeating the parameter sweep over filtering and cut-off thresholds of the segmentations retrieved over the AP cohort. The distributions of the area and distance from the tumour edge of all detected buds in the AP cohort were plotted to determine a range for the sweep. Figure 4.18 shows the distributions of the area and distance from the tumour edge of the detected buds in the AP cohort categorised by the tissue class of the microenvironment in which the bud resides. The area distribution shows that most buds were less than 1000 microns squared, with outliers ranging up to 5,000. The distance distribution shows that most detected buds were within 500 microns of the tumour edge, with outliers ranging up to 1000 microns. We used these distributions to define the search space for the optimal filter parameters for our automated tumour bud scoring protocols.

We performed a parameter search over every combination of minimum and maximum thresholds and many varieties of the tissue classes in which the detected buds reside. The baseline hazard ratios over all tissue classes can be observed in the appendix B, in figure B.1, and the Kaplan Meier survival curves are displayed in figure B.2. To obtain a baseline for the ITBCC guidelines, the adipose, background, debris, and mucin classes were removed, and the analysis was repeated. The resulting hazard ratios over area and distance can be observed in the appendix in figure B.3, and the Kaplan Meier survival curves are given in figure B.4.

To ascertain the optimal setting for each combination, a sweep of the required threshold for high-grade budding per objective field size was also performed for each configuration of tissue, area and distance filter parameters. The Cohen's kappa and Cox regression hazard ratios resulting from the parameter sweep over selected tissue class combinations are shown in figure 4.19.



	ADI	BACK	DEB	LYM	MUC	MUS	NORM	STR	TUM
Mean Area (µm^2)	177.26	310.64	175.31	173.41	187.97	205.96	194.75	206.78	212.39
Standard Deviation	114.0	347.38	110.2	106.81	127.46	132.71	128.51	129.35	137.19

a)

	ADI	BACK	DEB	LYM	MUC	MUS	NORM	STR	TUM
Mean Distance (µm)	214.01	188.99	142.81	166.49	176.55	183.15	142.6	156.76	41.4
Standard Deviation	156.87	134.94	86.23	102.39	124.32	123.81	88.38	99.43	16.12

b)

Figure 4.18: Shown are the distributions of bud area and distance from the tumour edge of the detected buds in the AP cohort categorised by the tissue class of the microenvironment in which the bud resides. Panel a) shows the area distribution measured in microns squared, and panel b) shows the distance distribution measured in microns.

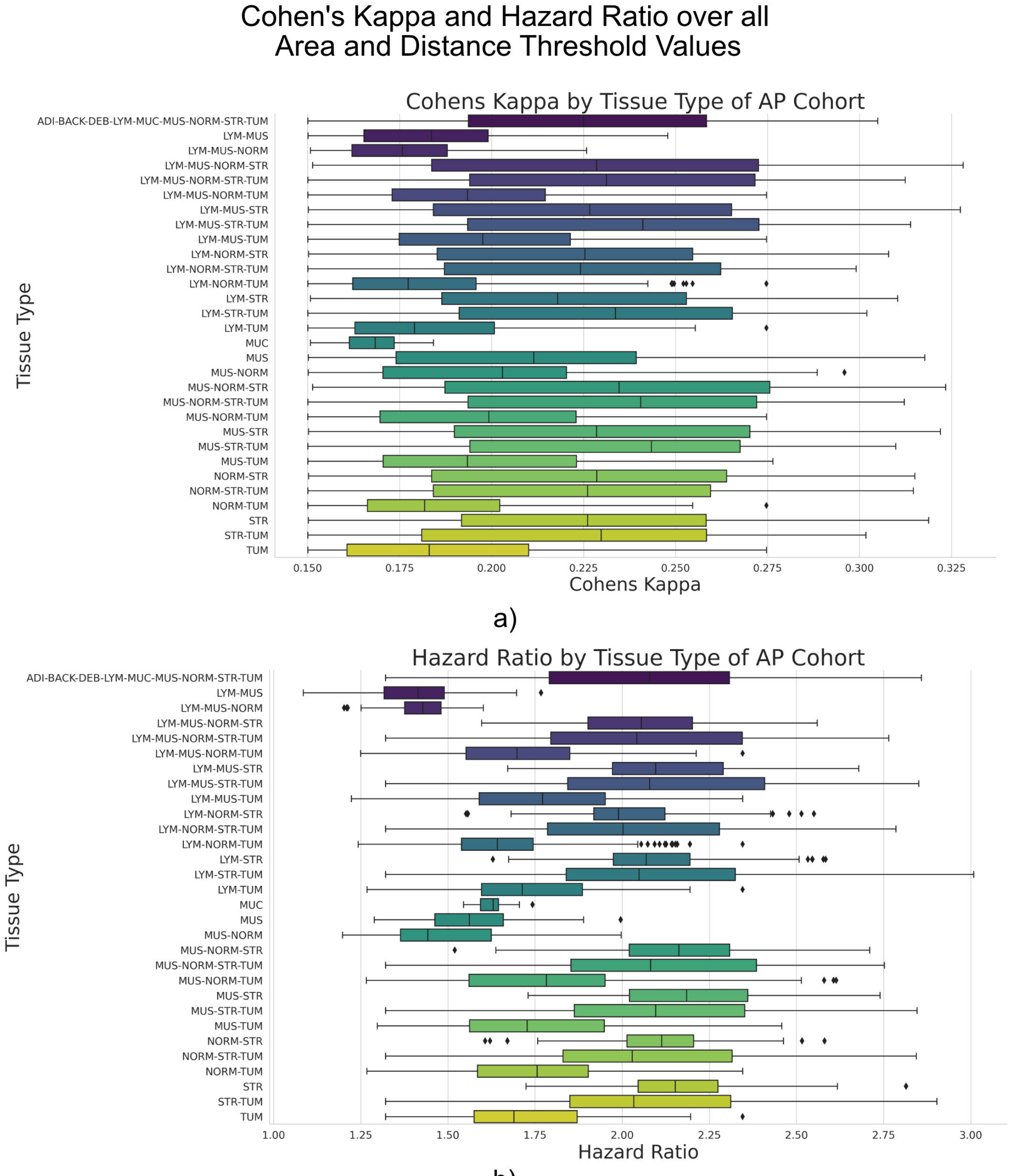


Figure 4.19: Displayed are the results from the parameter sweep over the area, distance, and high-grade budding threshold parameters for the AP cohort. Values are shown for each combination of parameters and tissue type. Panel a) shows the Cohen's kappa values, and panel b) shows the hazard ratio values.

Once the baseline values were determined using the parameter sweep, we ran the bud scoring pipeline with the best configuration selected over the GRI cohort. Again, to highlight the trends in the parameter space, we plotted a 3D contour plot of the Cox regression hazard ratio over all combinations of minimum and maximum area and distance values. We also plotted a 2D heatmap of the values to help visualise the results when the 3D contour plot has obscured values. These plots can be viewed in figure 4.10. Note that inverse to the GRI Cohort, the most prognostic buds in this tissue in the AP cohort have a small maximum area of less than 1000 microns, with small minimum areas also having superior performance. Like the GRI tissue, the most prognostic buds for this tissue type were at the outer range of the 500-micron sweep.

Kaplan Meier plots and the survival analysis results performed on detected buds residing in the muscle and tumour tissue types are displayed in figure 4.21. For comparison between cohorts, the performance metrics of this model on the AP cohort are given in table 4.3.4.

Metric	Accuracy	Precision	Recall	F1	Cohen's Kappa	Hazard Ratio
Value	0.7028	0.4451	0.5031	0.4723	0.2665	2.45 (1.82-3.31)

Table 4.7: Best model performance on the AP cohort using manual tumour bud scoring as ground truth.

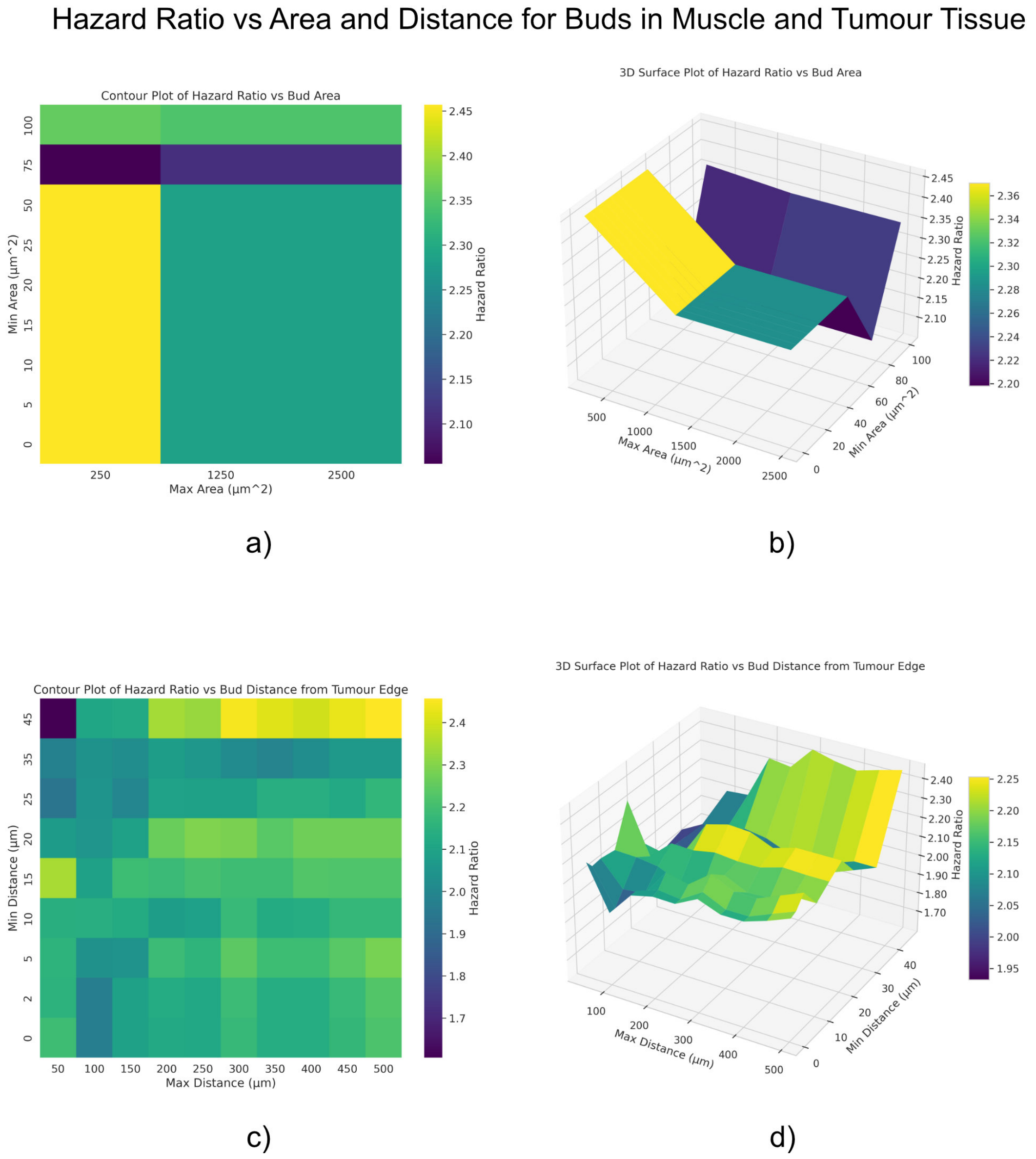
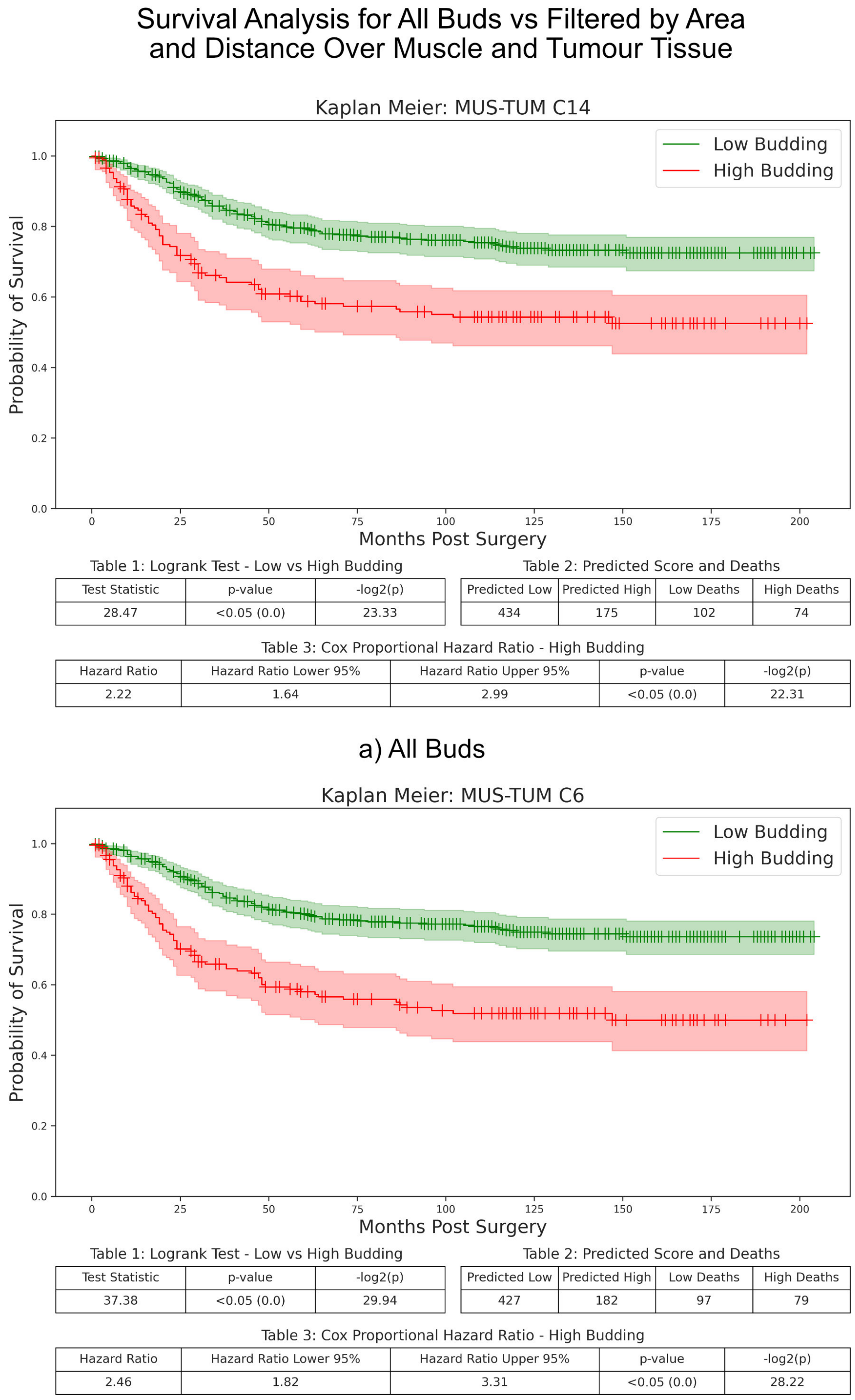


Figure 4.20: Shown is the hazard ratio of the area and distance threshold parameters over muscle and tumour tissue classes in the AP cohort. Panel a) displays a 2D heatmap of the hazard ratios resulting from the minimum and maximum area thresholds evaluated. Panel b) displays a 3D contour plot of the values to better highlight change over the parameter space. Panel c) shows a 2D heatmap of the hazard ratios resulting from the minimum and maximum distance thresholds evaluated. Panel d) displays a 3D contour plot of the values.



Survival Analysis for All Buds vs Filtered by Area and Distance Over Muscle and Tumour Tissue

Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
28.47	<0.05 (0.0)	23.33

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
434	175	102	74

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
2.22	1.64	2.99	<0.05 (0.0)	22.31

a) All Buds

Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
37.38	<0.05 (0.0)	29.94

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
427	182	97	79

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
2.46	1.82	3.31	<0.05 (0.0)	28.22

b) Filtered by distance minimum 35µm, maximum 500µm.
Area minimum 100 µm^2 , maximum unlimited.

Figure 4.21: Kaplan Meier plot of automated bud score with no filtering and the best-determined cutoff of 14 buds per 0.785mm^2 objective field. Panel b) displays the Kaplan Meier plot of automated bud scores over muscle and tumour tissue in the GRI cohort determined after the filtering of detected buds based on the most performant values of a distance minimum of 35μm, area minimum of 100μm^2, a distance maximum of 500μm and an unlimited area maximum, with the best cutoff of 6 buds per 0.785mm^2 objective field.

The prognostic ability of the most performant tissue combination of the AP cohort was also evaluated. This combination was of lymphocyte, stroma, muscle and tumour tissue. The hazard ratios resulting from the sweep over minimum and maximum thresholds for bud area and distance from the tumour core are given in figure 4.22. Panels a) and b) display the evaluated minimum and maximum area threshold values. Panel c) and d) show the results of the distance parameter sweep. The most prognostic buds once again had a small maximum area, less than 1000 microns. However, the minimum area became important, with values less than 60 μm^2, resulting in higher hazard ratios. The maximum distance of the most prognostic buds was greater than 100 microns from the tumour front, and a minimum distance of up to 30 microns had superior performance. Kaplan Meier plots and the survival analysis results performed on detected buds residing in lymphocyte, stroma, muscle and tumour tissue are displayed in figure 4.23.

A comparison of the survival curves of the AP cohort using manual and automated tumour bud scoring is given in figure 4.24. The automated bud scoring system resulted in a lower hazard ratio with fewer deaths in the predicted high aggression population and a larger number in the low group. The classification metrics for the best-performing automated bud scoring model on the AP cohort using manual tumour bud scoring as ground truth are shown in table 4.3.4. The model achieved an accuracy of 73.07%, a precision of 48.99%, a recall of 45.34%, an f1 score of 47.10%, and a Cohen's kappa of 0.2907, which shows moderate agreement with manual budding. The hazard ratio of the model was less than manual budding at 3.01 (2.23-4.06).

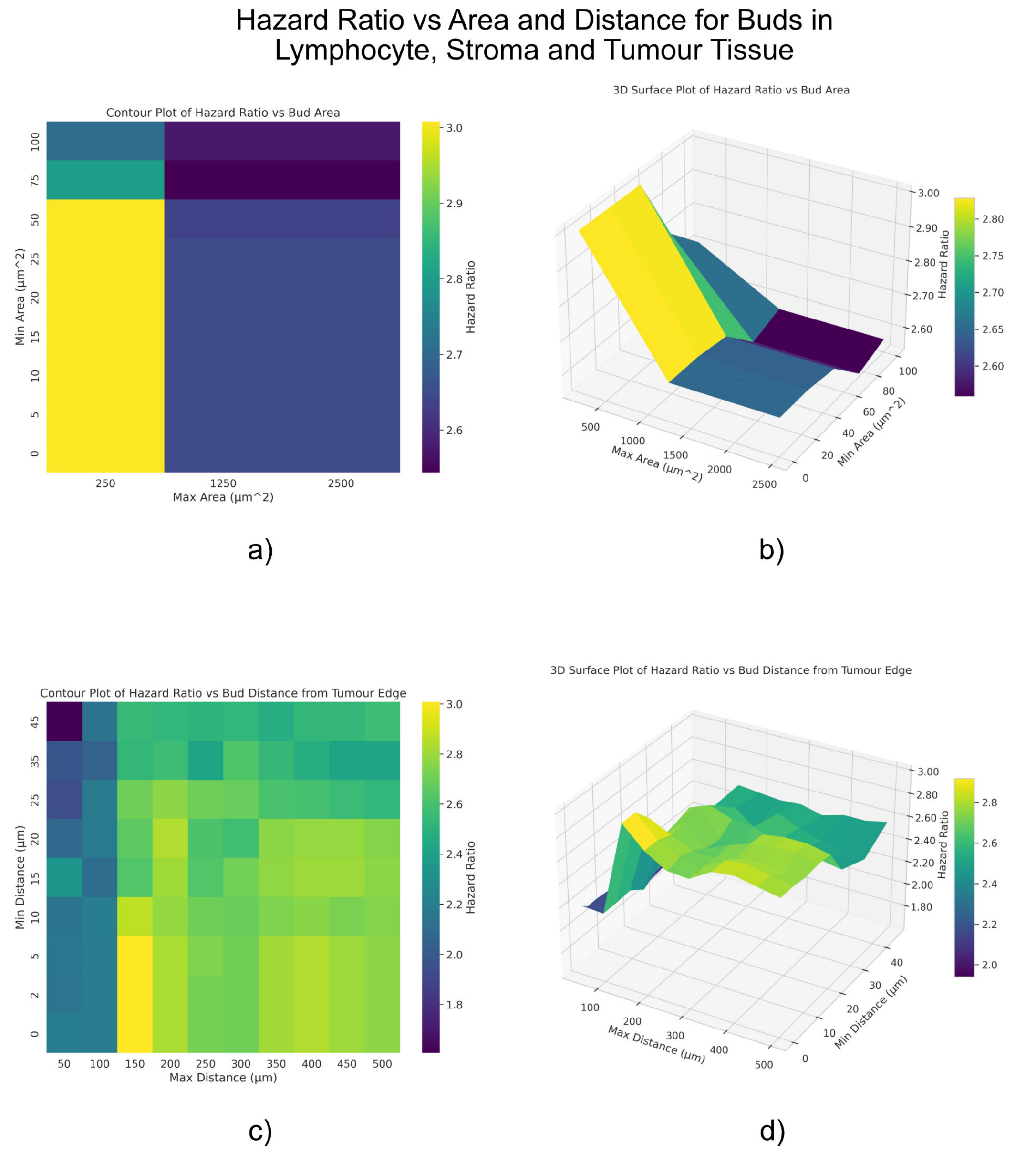


Figure 4.22: Shown is the hazard ratio of the area and distance threshold parameters over lymphocyte, stroma, muscle and tumour tissue classes in the AP cohort. Panel a) displays a 2D heatmap of the hazard ratios resulting from the minimum and maximum area thresholds evaluated. Panel b) displays a 3D contour plot of the values to better highlight change over the parameter space. Panel c) shows a 2D heatmap of the hazard ratios resulting from the minimum and maximum distance thresholds evaluated. Panel d) displays a 3D contour plot of the values.

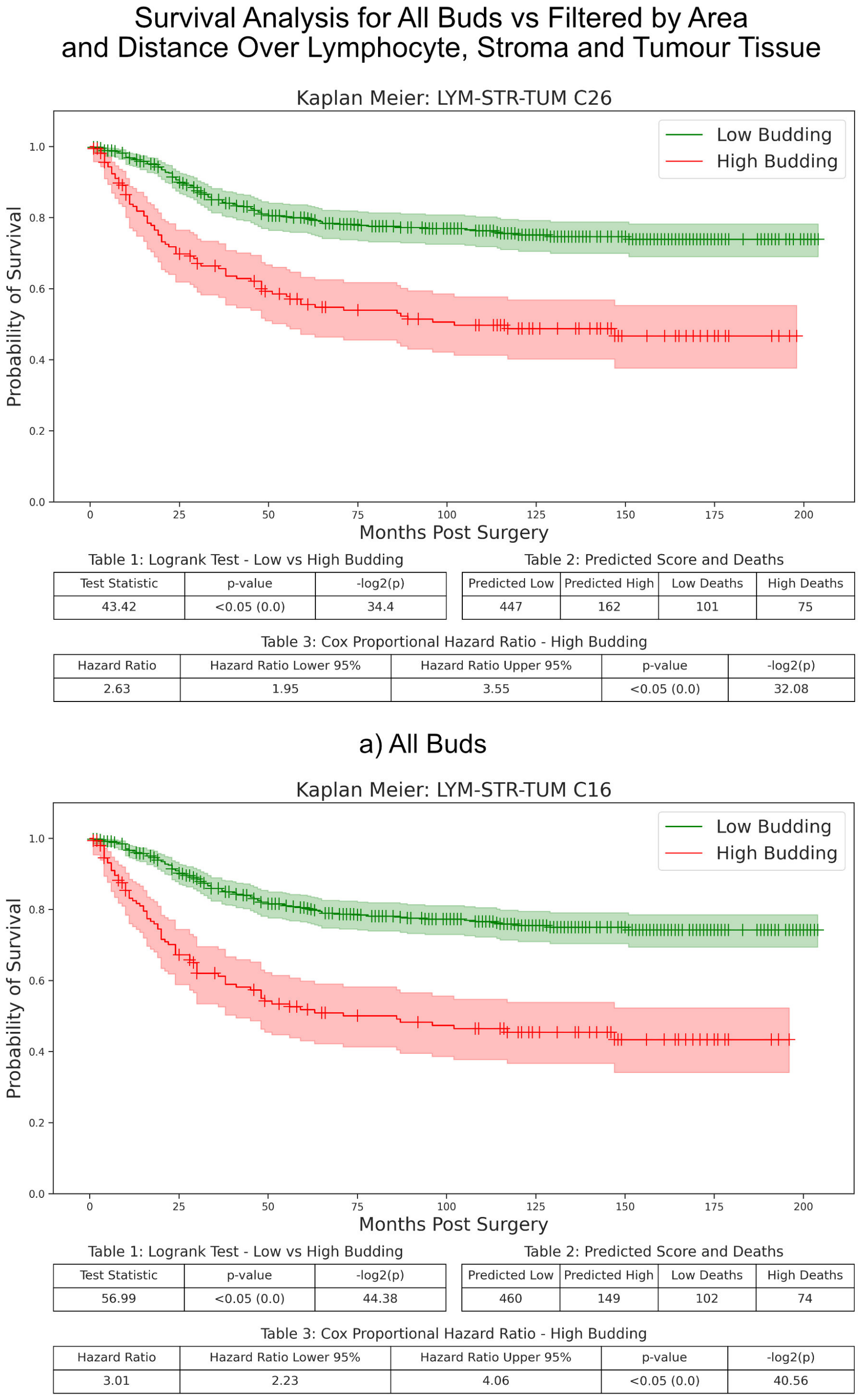


Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
43.42	<0.05 (0.0)	34.4

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
447	162	101	75

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
2.63	1.95	3.55	<0.05 (0.0)	32.08

a) All Buds

Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
56.99	<0.05 (0.0)	44.38

Table 2: Predicted Score and Deaths

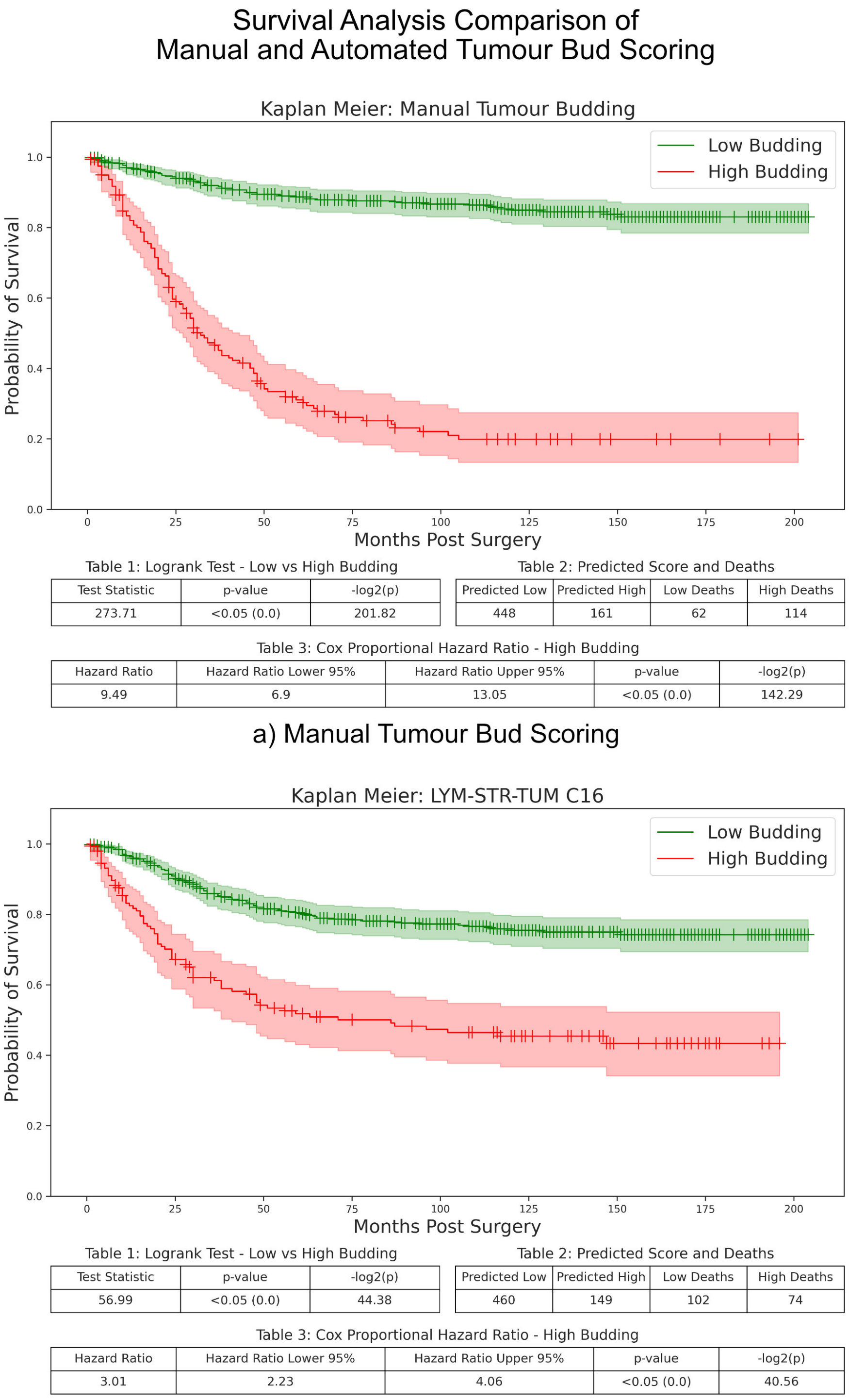
Predicted Low	Predicted High	Low Deaths	High Deaths
460	149	102	74

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
3.01	2.23	4.06	<0.05 (0.0)	40.56

b) Filtered by distance: no minimum, maximum 150 μm.
Area minimum 50 μm^2 , maximum 250 μm^2.

Figure 4.23: Kaplan Meier plot of automated bud score with no filtering and the best-determined cutoff of 14 buds per 0.785mm^2 objective field. Panel b) displays the Kaplan Meier plot of automated bud scores over muscle and tumour tissue in the GRI cohort determined after the filtering of detected buds based on the most performant values of a distance minimum of 35μm, area minimum of 100μm^2, a distance maximum of 500μm and an unlimited area maximum, with the best cutoff of 6 buds per 0.785mm^2 objective field.



Survival Analysis Comparison of Manual and Automated Tumour Bud Scoring

Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
273.71	<0.05 (0.0)	201.82

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
448	161	62	114

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
9.49	6.9	13.05	<0.05 (0.0)	142.29

a) Manual Tumour Bud Scoring

Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
56.99	<0.05 (0.0)	44.38

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
460	149	102	74

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
3.01	2.23	4.06	<0.05 (0.0)	40.56

b) Best Performing Automated Tumour Bud Scoring Model on Muscle and Tumour Tissue

Figure 4.24: Shown here is a comparison of the survival curves of the AP cohort using manual and automated tumour bud scoring. Panel a) shows the Kaplan Meier plot of the survival of patients with manual tumour bud scores. Panel b) shows the Kaplan Meier plot of the survival of patients with automated tumour bud scores. Tables 2 and 3 note that the automated bud scoring results in a lower hazard ratio.

Metric	Accuracy	Precision	Recall	F1	Cohen's Kappa	Hazard Ratio
Value	0.7307	0.4899	0.4534	0.4710	0.2907	3.01 (2.23-4.06)

Table 4.8: Best model performance on the AP cohort using manual tumour bud scoring as ground truth.

The final evaluation of performance was the utility of the most performant model as an independent biomarker over the AP patient cohort using Cox regression. The results are shown in table 4.3.4. The automated tumour bud score was a significant prognostic biomarker in univariate analysis, with a hazard ratio of 3.01 (2.23-4.06). In multivariate analysis, with a backward conditional elimination model, the hazard ratio was 1.05 (0.68-1.62) with a p-value of 0.83. It was not independent due to its correlation with manual budding. When manual budding was removed, the automated hazard ratio was 1.77 (1.19-2.62) with a p-value of less than 0.05, demonstrating that the automated tumour bud score was a significant prognostic biomarker independent of clinicopathological variables other than manual tumour bud scoring.

AP Cohort Biomarker Analysis

Biomarker	**Univariate Analysis**		**Multivariate Analysis**	
	HR (95% CI)	**p-value**	**HR (95% CI)**	**p-value**
Age	1.00 (0.99-1.01)	0.89	1.01 (0.99-1.02)	0.29
Sex (Female/Male)	1.19 (0.88-1.60)	0.26	1.10 (0.76-1.60)	0.62
Location (Right/Left/Rectum)	0.93 (0.76-1.13)	0.46	1.31 (1.01-1.68)	**(<0.05)**
TNM Stage (I/II/III/IV)	3.12 (2.49-3.92)	**(<0.05)**	2.37 (1.76-3.19)	**(<0.05)**
Tumour Differentiation (Well/Poor)	2.50 (1.68-3.73)	**(<0.05)**	1.88 (1.15-3.09)	**(<0.05)**
Tumour Perforation (Absent/Present)	2.42 (1.56-3.75)	**(<0.05)**	1.47 (0.86-2.50)	0.16
Margin Involvement (Absent/Involved)	4.30 (2.75-6.74)	**(<0.05)**	1.21 (0.65-2.25)	0.55
Venous Invasion (Absent/Present)	2.22 (1.65-2.99)	**(<0.05)**	0.92 (0.61-1.39)	0.69
Peritoneal Involvement (Absent/Involved)	2.90 (2.16-3.91)	**(<0.05)**	0.90 (0.59-1.38)	0.63
MMR Status (Proficient/Deficient/MSI-Low)	0.77 (0.62-0.97)	**(<0.05)**	0.96 (0.71-1.30)	0.79
GMS (0/1/2)	2.08 (1.67-2.59)	**(<0.05)**	1.52 (1.16-2.00)	**(<0.05)**
mGPS (0/1/2)	1.92 (1.57-2.36)	**(<0.05)**	1.68 (1.34-2.09)	**(<0.05)**
Tumour Budding (Low/High)	9.49 (6.90-13.05)	**(<0.05)**	6.98 (4.68-10.39)	**(<0.05)**
Automated Tumour Budding (Low/High)	3.01 (2.23-4.06)	**(<0.05)**	1.05 (0.68-1.62)	0.83
Automated Tumour Budding (Manual Removed) (Low/High)	3.01 (2.23-4.06)	**(<0.05)**	1.77 (1.19-2.62)	**(<0.05)**

Abbreviations: HR= Hazard Ratio, CI= Confidence Interval, MMR= Mismatch Repair, GMS= Glasgow Microenvironment Score, mGPS= modified Glasgow Prognostic Score

Table 4.9: Univariate Cox regression survival analysis was used to determine hazard ratios (HR) and 95% Confidence Intervals (CI) for available AP cohort biomarkers. Multivariable Cox regression survival analysis using a backward conditional elimination model and a statistical significance threshold of 0.05 was performed to identify independent prognostic biomarkers. Automated tumour bud scoring was correlated with manual budding in the AP cohort.

4.4 Discussion

Tumour budding has emerged as a pivotal prognostic indicator in colorectal cancer. Manual assessment of budding, however, is often subject to inter-observer variability, is time-consuming and can be influenced by the subjectivity of human interpretation. Automated bud scoring, leveraging advancements in digital pathology and deep learning, has emerged as a promising solution to these challenges as it has the potential to provide a consistent, objective, and rapid evaluation of a patient's budding score. This chapter aimed to develop and validate an automated tumour bud scoring model using virtual IHC pan-cytokeratin images as the ground truth for a deep learning segmentation model and to explore its prognostic significance in patient stratification.

To achieve this goal, we used the virtual immunohistochemistry model from chapter 3 to infer virtual IHC whole slide images over forty patients of the GRI cohort. We utilised them to highlight single tumour cells or small clusters of up to four cells that expressed IHC staining in the invasive margin of those slides. These were used to develop a dataset of 59,717 manually segmented buds, which were converted to a dataset of paired H&E and binary segmentation masks. This dataset was deployed to train a deep-learning model in detecting buds directly in H&E. The virtual IHC images proved to be an indispensable resource. A mask of the tumour core was mined from them to allow the automated determination of the tumour edge and, therefore, the coordinates of the invasive margin to enable the algorithmic placement of 0.785mm^2 fields. These fields were used to determine budding density, as advised in the ITBCC recommendations for bud scoring. These techniques allowed us to develop an end-to-end automated tumour bud scoring pipeline. The process of this workflow is described in algorithm 2.

A tissue classification model was trained to allow the differentiation of the tissue microenvironment surrounding the buds to determine if there was a prognostic difference from scoring in different tissue classes. The tissue classification model was trained in a supervised manner on the NCT dataset (Kather et al. 2018), then applied to the newer GRI and older AP cohorts, which were prepared and scanned at different institutions. The tissue classification model performed very well on the training dataset, with high training and validation accuracy. The resulting confusion matrices showed excellent performance for nine-class classification. Only tending to misclassify adipose tissue as mucin and stroma as muscle. The classification model performed well during testing on the GRI cohort, with few errors, as the slides only had a small number of artefacts or scanning errors and were normalised successfully using the BEAST toolkit. However, in the AP cohort, many slides had artefacts that reduced the focus or obscured large tissue regions that affected both the deep learning models and the normalisation process. The tissue classification model handled areas of deposited ink reasonably well, classifying most as background or debris.

However, the unfocused regions confused it drastically, with many significantly misclassified. This was difficult to mitigate using the supervised approach but can be tackled with more robust techniques, like self-supervised learning, which will be described in the next chapter.

Analysing model performance on the GRI cohort provides intriguing insights into the behaviour and potential of automated tumour bud scoring. Notably, the highest positive Cohens kappa values, which indicated a correlation with manual bud scores, were observed for buds located in lymphocytes, muscle, stroma, normal, and tumour tissue. This is likely because looking for tumour cells in this combination of tissue is the convention when pathologists manually score tumour buds in the invasive margin. The parameter search uncovered the increased predictive ability of small buds close to the tumour core in these tissues. In stark contrast, when buds were filtered to only include those in muscle and tumour tissue, a population emerged that was more prognostic. These were large and distant from the tumour core. This observation hints at the potential existence of two distinct bud populations: small cells and clusters still associated with the tumour core and a second group comprised of larger cells and clusters that have migrated a significant distance away from the core, which is supported by the idea that buds are small clusters that have undergone partial EMT and remain together but have gained a migratory capacity. However, more work is required to validate this hypothesis with larger and more diverse patient cohorts. However, it suggests that while effective, the current guidelines for tumour budding may have room for improvement. The litmus test of model performance was the analysis of prognostic independence as a biomarker to stratify patients. Notably, the automated bud score surpassed manual budding regarding the hazard ratio in univariate analysis and the number of adverse events in the high bud score groups. Most importantly, it was a significant independent predictor of survival in multivariate Cox regression.

The analysis of the AP cohort serves as a critical testbed for evaluating the generalisability of our model. Distinctly, the distance distributions in the AP cohort diverge from those in the GRI cohort, a discrepancy that may be attributed to the slide artefacts. Despite these challenges, the parameter settings determined initially on the GRI cohort continue to significantly delineate high and low budding populations as determined by the log-rank test of the distribution difference. Remarkedly, smaller buds emerge as more prognostic within the same tissue type in this cohort, contrasting to the split population observed in the GRI cohort. The final hazard ratio of automated budding was diminished in the AP cohort, a decline potentially influenced by the scanning artefacts that impede accurate tissue classification and bud segmentation. Nevertheless, the ability of the automated system to significantly separate high and low-grade groups remain intact. Notably, only manual tumour budding correlates with the automatic bud score and outperforms it in multivariate analysis. Had this not been the case, the automated budding would have stood as an independent predictor of survival. These findings underscore several pivotal future directions. Firstly, there is a pressing need to source a subsequent cohort that is characterised by fewer scanning artefacts and is equipped with patient survival data. Such a cohort would facilitate the further refinement and validation of our model. Secondly, the challenges posed by the artefacts highlight the necessity for more resilient methodologies. As discussed in subsequent chapters, incorporating the artefacts into the training data and exploring techniques like self-supervised learning can offer a pathway to more robust models in the face of such consistencies.

The research presented in this chapter highlights the potential of using virtual IHC ground truth as a foundational basis for training an automated tumour bud scoring model on H&E slides. This approach provides a more objective method for generating tumour bud annotations in H&E, addressing the difficulties inherent in manual annotations. Our proposed scoring model, developed using an H&E segmentation dataset informed by virtual IHC, showed prognostic value and proof of concept for later systems that may be deployed for use in the clinic. Interestingly, our analysis of the segmented cells hints at two unique tumour bud populations: one with smaller buds closer to the tumour core and another with larger clusters that seem to have migrated far from the tumour's edge. However, it's crucial to mention that this distinction was primarily observed in only one of the two cohorts studied, underscoring the need for further validation with a broader and more diverse patient cohort. Despite this, the success of automated tumour bud scoring rooted in virtual IHC ground truth offers many potential benefits, including the potential to quickly generate large and reliable datasets to train models to identify tumour cells in H&E. If systems like our proposed automated bud scoring model were further developed and implemented in a clinical setting, it would mitigate the challenges of inter-observer variability, ensure consistent and objective scores, and significantly reduce reliance on manual resources for bud scoring. This would reduce both the time and costs associated with providing it in standard reporting, which has the potential to positively impact patient care by assisting pathologists in providing a precise diagnosis, paving the way for treatments tailored to tumour aggressiveness and ultimately improving patient survival.

Chapter 5

Colorectal Tumour Aggression Prediction by Self-Supervised Deep Learning and Transformer Networks

5.1 Introduction

The mandate for this project was to seek out methods for automated tumour bud scoring in colorectal cancer for use as a biomarker to predict survival and regression to allow for the consistent and objective stratification of patients into high and low-risk groups. This would help address the heterogeneity that currently exists in the survival of patients within stages as grouped by the TNM system. As reviewed in section 2.3 and chapter 4, a high density of tumour buds observed in the invasive margin of colorectal cancer is now recognised as an adverse prognostic factor linked to aggressive tumour behaviour and poorer cancer-specific survival (Jepsen et al. 2018; Fauzi et al. 2020). However, its routine use in the clinic has been held back by a lack of consistent scoring methods, resource constraints due to the time-intensive nature of bud scoring, and inter-observer variability in H&E, leading to issues with reproducibility (Fisher et al. 2021; Tavolara et al. 2022). As discussed in chapter 4, several attempts have been made to develop an automated system to score in H&E with varying levels of success. While prognostically significant, our proposed method was still limited by the supervised nature of the training process and the artefacts that can be encountered in real-world data.

As examined in section 2.10, there have been many attempts to predict survival from H&E images directly. This end-to-end process makes it more robust but is limited in clinical usefulness by a lack of interpretability (Mobadersany et al. 2018). However, the benefits of the survival prediction approaches were that they work with patient-level labels that can be applied at the slide level to generate large quantities of data to use in semi-supervised training.

We propose that aggression prediction with patient-level labels based on tumour budding offers a more interpretable and histologically verifiable alternative. Tumour aggression in colorectal cancer refers to the characteristics of the tumour cells that contribute to growth, invasion and metastasis, and poor patient outcomes (Dawson et al. 2019a; Park et al. 2017). Tumour budding focuses on a specific biological phenomenon linked to aggressive tumour behaviour. This specificity should result in more accurate and targeted predictions than the broader and more intricate task of predicting survival. Many external factors, such as healthcare quality and patient adherence to treatment, can influence the ground truth in survival prediction. In contrast, high-density budding as ground truth for aggression prediction is a quantifiable histomorphological phenomenon focusing more on the tumour's intrinsic characteristics.

In our review of the current state-of-the-art survival prediction models in section 2.10, we saw that appropriate feature extraction was crucial for most models, often based on the slide-level labels or manual annotations. Self-supervised learning, as discussed in section 2.7.5, has emerged as a modern technique that can learn to extract feature representations from vast amounts of unlabelled data. It can learn semantically meaningful representations on input tiles over as much training data as can be provided. Therefore, the resulting feature extractor can be significantly more robust and generalisable. The review in section 2.10 also highlighted a gap in the literature to apply transformer models to interpret a sequence of feature representations that represent the whole slide image, or a targeted portion of it, and use patient-level labels as the ground truth for training in aggression prediction for colorectal cancer whole slide images.

These observations lead us to the work presented in this chapter of applying self-supervised learning and transformers to predict aggression in H&E colorectal cancer images. There are many currently recognised aggressive histological tumour characteristics, such as how well differentiated the tumour glands are, the depth of invasion, the lymph node involvement, distant metastasis, and tumour size (Jepsen et al. 2018; Fauzi et al. 2020; Dawson et al. 2019a). The characteristics of the tumour microenvironment are now recognised to play a role in tumour aggression as demonstrated by Park et al. (2017). Additionally, molecular mechanisms correlated with aggression may present features distinguishable by deep learning (Krause et al. 2021; Malki et al. 2020). Therefore, many features are present in H&E whole slide images related to aggression that a deep learning model could utilise to make predictions. We proposed that using the patient bud scores as the ground truth to develop a deep learning-based aggression biomarker would enable the model to learn to recognise correlated aggressive tumour characteristics.

While this work is ongoing, we expect that the interpretable nature of the transformer will help us understand the relationship between histological features revealed as being correlated with aggression and cancer-specific survival. While survival prediction has relevance, we hope that a shift towards aggression prediction using tumour budding as ground truth offers a more targeted, efficient, and biologically relevant approach. Such a method promises to stratify patients as manual or automated bud scoring would, resulting in better patient outcomes with more personalised treatments. But it may also provide valuable insights into tumour behaviour, with interpretable predictions that may bridge the gap between the black-box nature of deep learning and clinical applicability.

5.2 Materials and Methods

5.2.1 Dataset

The GRI and AP cohorts were reused as the raw source datasets for this project section. This was efficient as we already had access and ethical approval and were sizeable enough datasets for the self-supervised learning we wished to perform. The nature and characteristics of the whole slide images and patient cohorts are described in section 4.2.1. The GRI cohort comprised 785 patients, each with clinicopathological data and a corresponding whole slide image picked by a pathologist to score budding. In the AP cohort, there were 1030 patients, accompanied by related data and 2224 whole slide images. The corresponding patient-level bud scores were known, but the exact slide used to score was not. The AP cohort had issues with focus, ink and scanning artefacts but was used as a baseline to check for generalisation and allowed comparison with the results of our automated tumour bud scoring system from chapter 4. The work in this chapter is still ongoing, and we also hope to evaluate it on The Cancer Genome Atlas colorectal adenocarcinoma dataset. The following sections will provide the processes for tissue classification, invasive margin detection, and slide tessellation used to create our training and testing datasets.

To finetune our self-supervised network to perform tissue classification, we required a dataset of labelled colorectal tissue tiles, and the NCT-100K dataset by Kather et al. (2018) met those needs. This is a publicly available dataset of over one hundred thousand RGB colorectal tissue tiles of size 224^2 pixels, each with a label of the corresponding tissue type from nine classes: background, adipose, lymphocytes, debris, mucin, muscle, stroma, normal glands, and tumour tissue.

5.2.2 Self-Supervised Learning for Feature Extraction

We had to perform some preprocessing steps to tessellate our whole slide images into usable input tiles for deep learning. The first of these was to detect the regions of tissue, extract them at the desired level of magnification, discard the background and write out the tiles for later use. Interestingly, the first use case of the self-supervised network was as a more robust tissue classifier to improve the invasive margin segmentation. But to train this, we had to perform a bootstrapping process and use our supervised tissue classification framework, developed in chapter 4, to classify the tissue initially. This resulted in full-resolution coordinates for the tissue across the whole slide images. These could then be scaled to the desired magnification level, and tiles output containing only tissue at 2.5x, 5x, 10x, and 20x; an example of the output tissue coordinates is given in figure 5.1. This was an essential step as it allowed the creation of a dataset of over 14 million tiles from tissue in the GRI cohort, where each tile also had a corresponding tissue classification of the nine available types. This information was critical in the subsequent successful training of the self-supervised network, as it allowed us to sample a balanced number of tiles from each class during self-supervised training and improved the quality of the dataset and training stability.

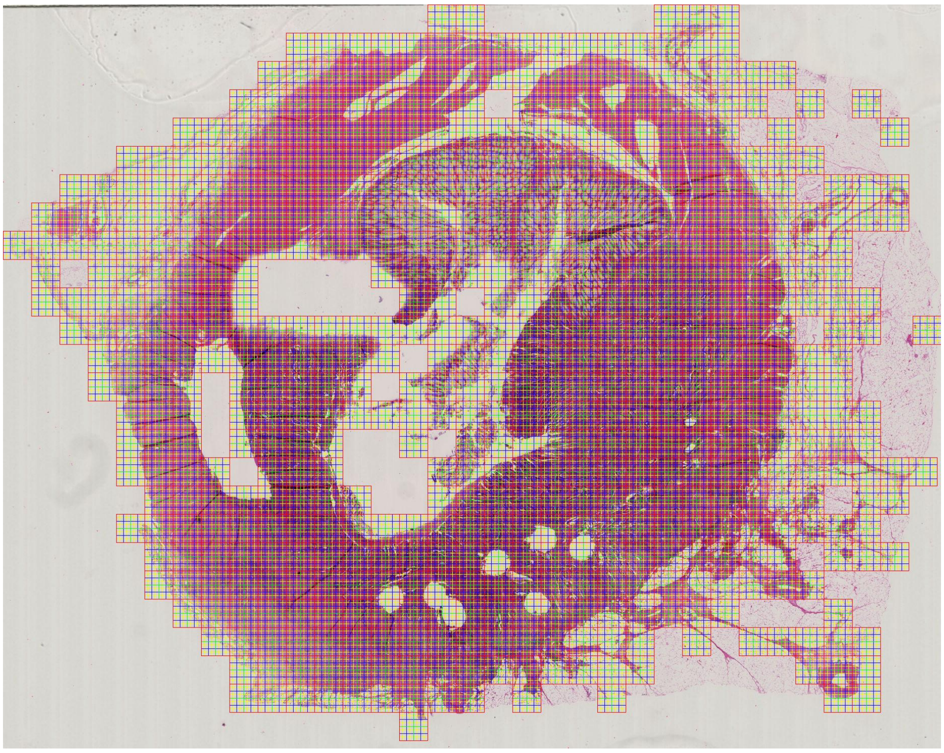


Figure 5.1: Shown is an example of supervised tissue detection and whole slide tessellation. The red boxes highlight regions extracted and downsampled to 224 pixels square at 2.5x, green 5x, blue 10x and yellow 20x magnification.

As discussed in section 2.7.5, many successful self-supervised representation learning architectures have been proposed. We chose to use the VICReg architecture by Bardes et al. (2022), as it was designed to generate semantically meaningful feature representations by its novel loss function that reduces redundancy and maintains the variance of the representations while ensuring those augmented versions of the input patch produce similar representation outputs. It does not need negative examples and can be successfully trained at a comparatively small batch size, so it was more efficient and suited to our hardware constraints. Figure 5.2 shows our architecture configuration. It consists of a pre-trained ResNet50 (He et al. 2015) encoder that outputs feature representations with a vector length of 2048. These were then fed into a projector head that uses multiple linear layers to expand the representations to a vector of length 8192 for the loss calculations.

The model was implemented in Python with the Pytorch library (Paszke et al. 2019). A learning rate of 2e^{-4} was used for training with an ADAM optimiser (Kingma et al. 2017). The input tiles were stain augmented during training using the BEAST toolkit described in section 3.2.2 and normalised to a range of -1 to 1. Only augmentation was done on the input tiles as stain normalisation provided prominent image features for the self-supervised models to recognise and had a detrimental effect on the quality of the learned representations during initial tests. The model was trained for 1000 epochs with a batch size of 32 and a weight decay of 1e^{-4}. The training was carried out in a data-parallel fashion and distributed over four A6000 GPUs. The number of high-grade and low-grade budding slides shown to the model at each epoch was balanced, and ten per cent were held out for validation. The same number of tiles for each tissue class were randomly extracted from each patient slide in the dataset at each epoch. This ensured the model saw a balanced set of tissue classes and aggression levels at each epoch to encourage training stability and more robust representations. The best-performing epoch was chosen with the lowest validation loss before the training and validation losses diverged. After training, the projection head is removed from the ResNet50 network, and the resulting model can extract feature representations from input tiles for utilisation in downstream tasks.

To evaluate the quality of the learned representations, we performed clustering using the Leiden community detection algorithm proposed by Traag et al. (2019). The benefits of this method are discussed in more detail in section 2.7.5. However, in summary, it is an algorithm that works by determining densely connected communities in a graph and iteratively partitioning them into smaller groups to select the optimal collections of instances. This is a promising algorithm to evaluate the learned representations, as by the nature of the loss function, similar image features should be densely clustered in the representation space. Each candidate model was evaluated by performing principal component analysis to reduce the dimensions of the output features. A KNN graph of the feature representations was constructed, and the Nvidia rapids GPU library was used to perform Leiden clustering. The resulting Leiden graph modularity score, a visual assessment of the partitioned clusters, and a Cox proportional hazard regression of the cluster distributions at a slide level in the GRI cohort were used to select the final model for feature extraction.

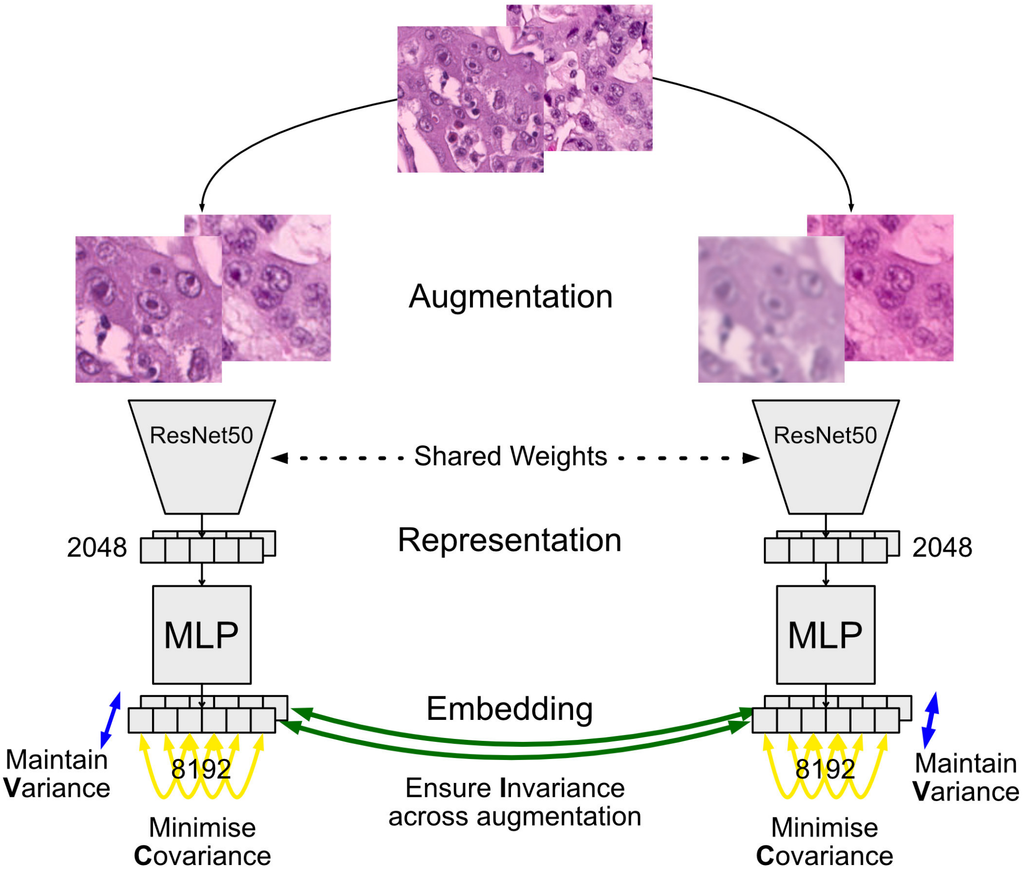


Figure 5.2: This is the architecture of the self-supervised learning network used to extract tissue representations. It is VICReg-based (Bardes et al. 2022) with a ResNet-50 (He et al. 2015) encoder that outputs representation vectors of length 2048, followed by a projection head consisting of three fully connected dense layers, each with 8192 hidden units. The architecture works by augmenting the input tiles and then using the projector head to project the augmented representations into a higher-dimensional embedding space. The loss is then calculated between the augmented embeddings and the original embeddings. The loss function aims to reduce the redundancy of the embeddings while maintaining their variance and minimising the distance between the augmented representations and the original representations.

5.2.3 SSL Tissue Classification and Invasive Margin Segmentation

This section of the project aimed to improve the targeting ability of our automated invasive margin segmentation framework. We proposed to do this by training a more robust tissue classification model using self-supervised deep learning to train a feature extraction network over a vast amount of unlabelled data. Then, finetune it using the NCT dataset by Kather et al. (2018) to classify the tissue over our cohorts. The best network was employed from the self-supervised feature extraction process described above. A nine-neuron dense layer was appended to the top of the trained ResNet50 feature encoder, and then the resulting network was trained to classify the tissue in the NCT dataset. We used the same hyperparameter settings as the initial self-supervised network for subsequent fine-tuning. The final network selection was made using the classification performance on a validation set of tiles from the NCT dataset, also provided separately by Kather et al. (2018).

When an optimal tissue classification model was found, we performed classification over all slides in the GRI and AP patient cohorts. This resulted in a map of the tissue classes for each slide. The tumour tiles were separated into their own group. The stroma, muscle, normal glands and lymphocytes were grouped into a “tissue” class, and the adipose, mucin, and debris classes were marked as exclusion zones. A binary mask of each grouping was created for each slide in the dataset, and then the tumour mask was buffered by 500 microns. The intersection of this mask and the “tissue” mask consisting of tiles from stroma, muscle, normal glands, and areas of inflammatory infiltrate was then computed. This intersecting region was our newly determined invasive margin. This has several advantages over the virtual IHC margin segmentation method, as it can specifically recognise and avoid areas of mucin and debris that are classified as pseudobuds (Haddad et al. 2023) and allows for more accurate and finer control of the invasive margin segmentation. Figure 5.3 provides a diagram depicting the segmentation process.

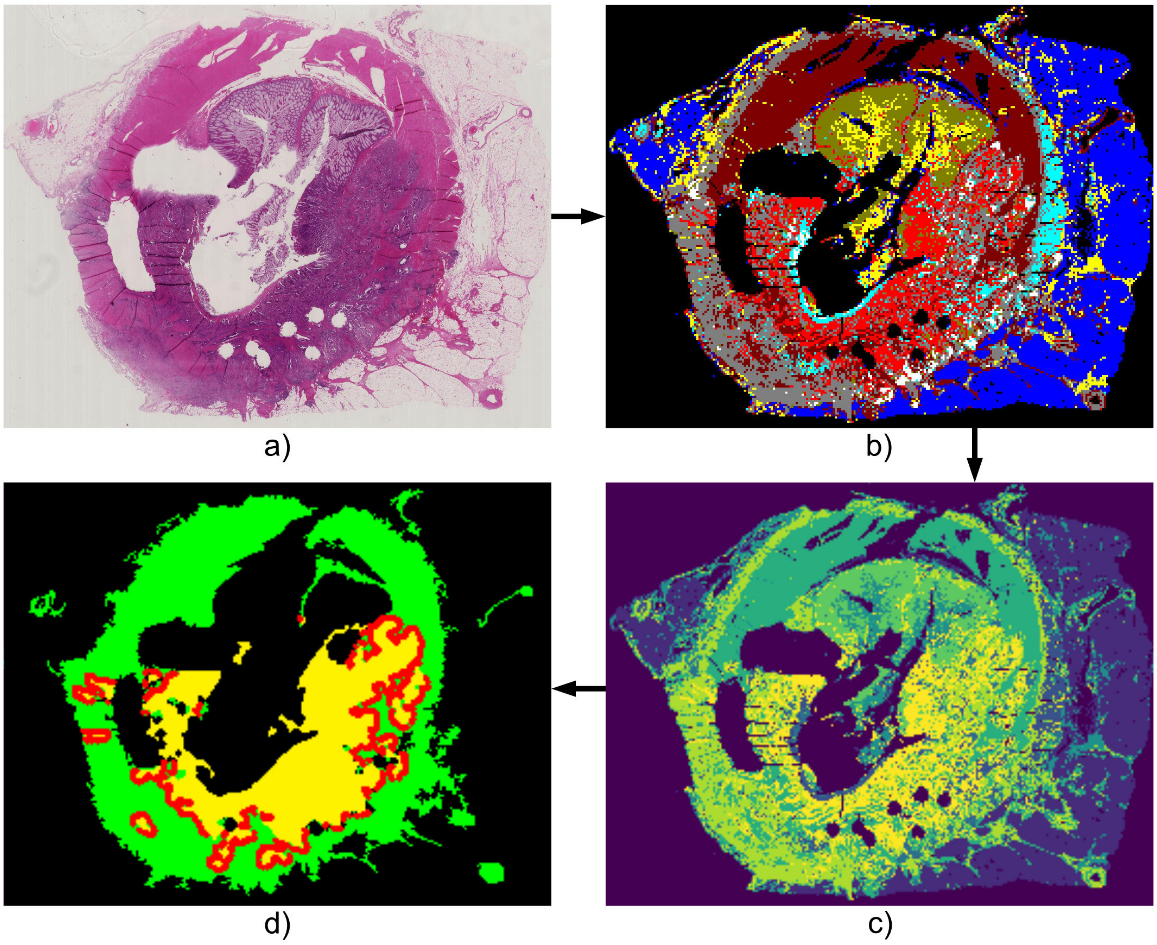


Figure 5.3: This is our proposed invasive margin segmentation process. The input whole slide image, shown in panel a), is tessellated, and the tissue is classified using the self-supervised tissue classification network, resulting in the tissue class mask shown in panel b). A mask is extracted for the tumour tissue, and the stroma, muscle, normal glands and lymphocyte tissue are combined into one tissue mask. The adipose, background, debris and mucin tiles are combined into an exclusion mask. Panel c) displays the combination of masks, tumour tissue in yellow, other tissue in green and background and excluded tissue in purple and blue. The tumour mask is buffered outwards by 500 microns, and the intersection between it and the tissue mask is taken to create the invasive margin mask shown in red in panel d).

5.2.4 Transformer Network for Aggression Prediction

This section aimed to use the extracted feature representations from the self-supervised encoder network to train a transformer network for aggression prediction in colorectal tumours. The driving theory behind our proposal was that the self-supervised convolutional network could learn to recognise and extract semantically relevant local features within the input H&E tiles. These feature representations were extracted from a whole slide image and assembled into a sequence to represent it. This sequence was then used as input to a transformer network, the strength of which is being able to consider how elements relate globally in the input data. The hope was that this combination of local feature extraction and the capability to learn global relationships would result in an improved ability to predict the aggression of colorectal tumours. The ground truth for training was the patient-level tumour budding score for each slide. We hoped the transformer network would learn the relationships between extracted feature representations that correlate with high aggression. The ideal setup would have been to extract the feature representations from all tissue tiles and feed the entire tissue sequence to the transformer. But given that, as described in section 2.7.6, within each layer of the multi-head attention modules in the transformer are parameter matrices for the attention between each item in the sequence. These are square matrices, with the dimension of a side equal to the length of the input sequence. This means that the longer the input sequence, the larger the set of matrices in each module. This quickly becomes a memory concern in even the largest GPU but is easily parallelised between GPUs. However, given that we only had access to four A6000s on one machine, we could only deal with a maximum sequence length of less than 10,000 tile representations. Therefore, we decided to limit our initial tests to tiles classified as belonging to the tumour, as these would contain the most information about the tumour features and level of aggression. For each slide in the GRI cohort, the features were extracted for the tumour tiles in the invasive margin for each virtual high-power field of size 0.785mm^2 in the invasive margin, and the entire set unravelled into a sequence representing the margin. The training was carried out on 100 high and low-budding slides from the GRI cohort, with the rest being held out for validation.

Figure 5.4 shows our transformer network architecture for aggression prediction. It consists of two layers of multi-head attention and feedforward modules. A transformer network often consists of an encoder and a decoder module for sequence translation, but to reduce memory constraints, we only use the encoder architecture and feed the output sequence to an adaptive averaging layer and then a final dense linear layer before applying a SoftMax function over the two-class output. Before the input sequence reaches the first layer of the model, it undergoes a process called positional encoding. This process adds a unique encoding to each value, which provides the transformer with information about its global location in the sequence. The model was implemented in Python with the Pytorch library (Paszke et al. 2019). A

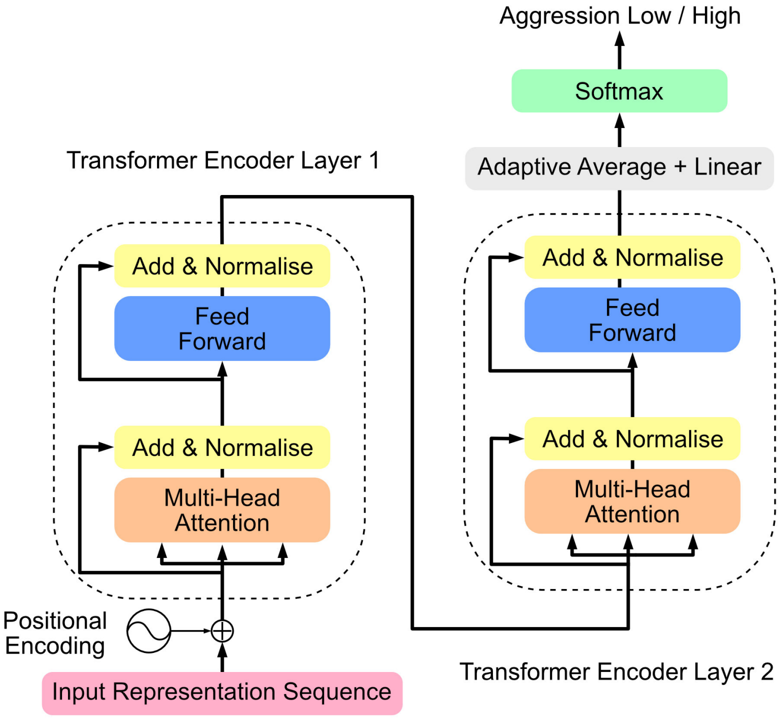


Figure 5.4: Displayed is our proposed transformer architecture for aggression prediction. The input tiles are passed through a ResNet50 (He et al. 2015) encoder to extract feature representations. These are then passed through a two-layer transformer encoder (Vaswani et al. 2017). The final representation sequence is passed through an adaptive averaging layer and a linear layer to predict the patient-level aggression score.

learning rate of 1e^{-5} was used for training, with dropout before the final linear layer, a weight decay rate of 2e^{-2}, a batch size of 1 with sequence length varying by the slide, and an ADAM optimiser (Kingma et al. 2017). Models were trained for 100 epochs, and the classification metrics were used to select the best-performing epoch compared to the manual budding scores. The final model performance was evaluated using the Cox regression hazard ratios of the output aggression predictions on the remainder of the GRI cohort and the subsequent performance on the test dataset of features extracted from the held-out AP cohort slides. This is still a work in progress, and we want to extend our model to more extensive sequence lengths so we can include a more diverse set of tissue classes and analyse the relationships between them as the transformer affords the ability to do. However, we shall now discuss the results of our preliminary work with this method.

5.3 Results

This section outlines the results of our proposed method for automated colorectal tumour aggression prediction, including the quality of the learned self-supervised H&E representations, the improved tissue classification model trained using self-supervised learning, and the survival performance of the resulting automated aggression score.

5.3.1 Self-Supervised Learning for Feature Extraction

Our best-performing self-supervised learning model had a ResNet50 encoder as proposed by He et al. (2015) and trained with the VICReg loss parameters as described by Bardes et al. (2022), with training hyperparameters listed in section 5.2.2. The validation loss had reached a minimum and started to diverge from the training loss by epoch 145, and this model, along with several epochs on either side, were evaluated. However, epoch 145 was the best-performing network, with a VICReg training loss of 18.39 and validation loss of 42.34.

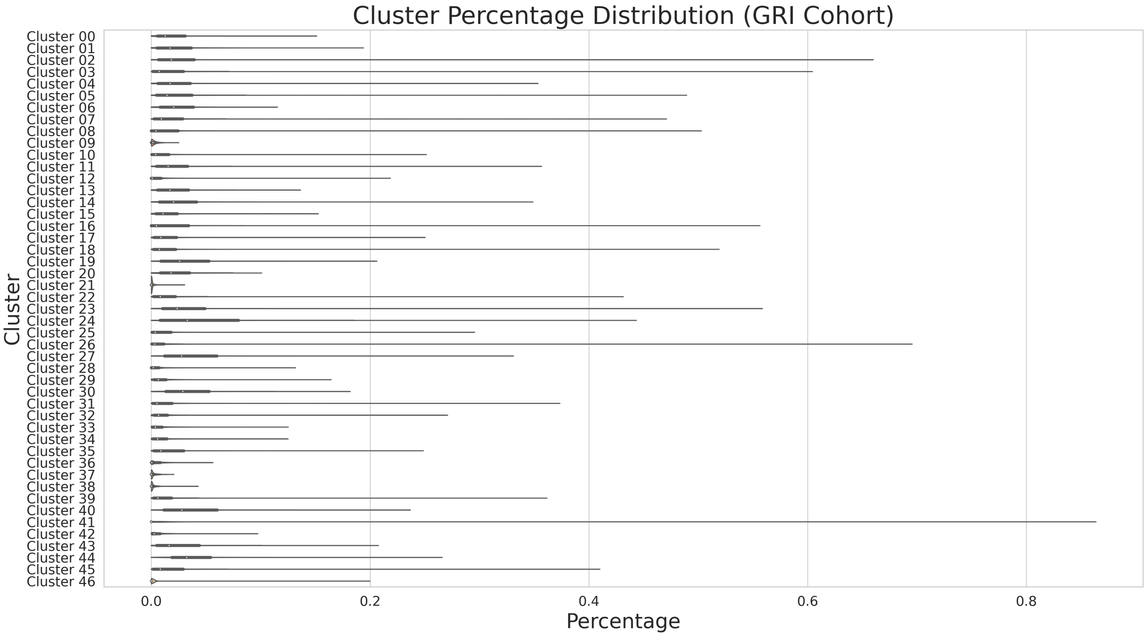


Figure 5.5: Shown is a violin plot of the distribution of the Leiden clustering of the GRI cohort feature representations. The y-axis represents the cluster number, and the x-axis represents the percentage of that cluster in the slides. The median and interquartile range are plotted as white dots and thicker grey lines.

The first verification check performed on trained models was to cluster the representations and check their distribution across the dataset to ensure they were not becoming too specific to single slides. Figure 5.5 displays a violin plot of the distribution of cluster percentages across the slides in the GRI cohort. Clusters 9, 21, 36, 37 and 38 were only exhibited over a shallow range of percentages in slides in the

GRI cohort. Upon investigation, they were clusters identifying out-of-focus areas and folded tissue. Tiles belonging to these clusters were subsequently discarded from further analysis.

To further evaluate the quality of the feature representations extracted using self-supervised learning, we performed Cox regression on the clusters to determine how each relates to the survival of the patients in the cohort. This was a valuable test as initial epochs would see poor clustering performance and a wide range of hazard ratio values with no clusters significantly correlated with outcome. Figure 5.6 displays the corresponding Cox hazard ratio for each cluster produced by our best-performing model and the 95% confidence intervals. Some clusters, notably 14 and 19, had high hazard ratio values with a tighter range of confidence intervals. These clusters are visible in figure 5.7 and 5.8 respectively. This demonstrated that meaningful information related to survival had been encoded in the learned representations and validated the model for use in downstream tasks.

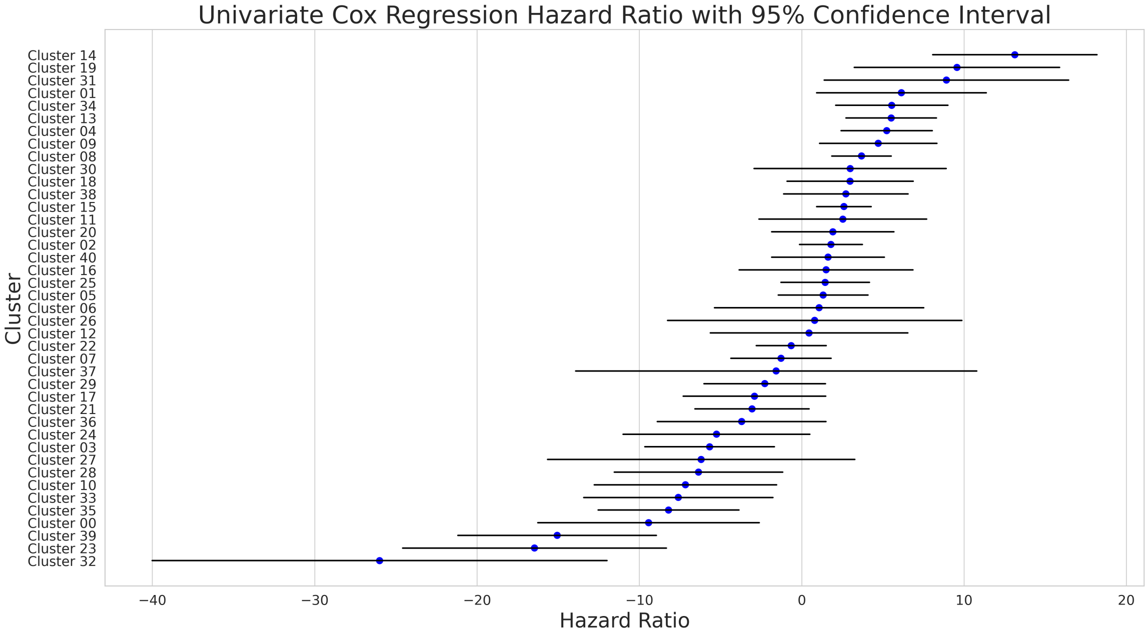


Figure 5.6: Shown here are the hazard ratios of the Leiden clusters of the GRI cohort feature representations. The y-axis represents the cluster number, and the x-axis represents the hazard ratio. The error bars represent the 95% confidence intervals.

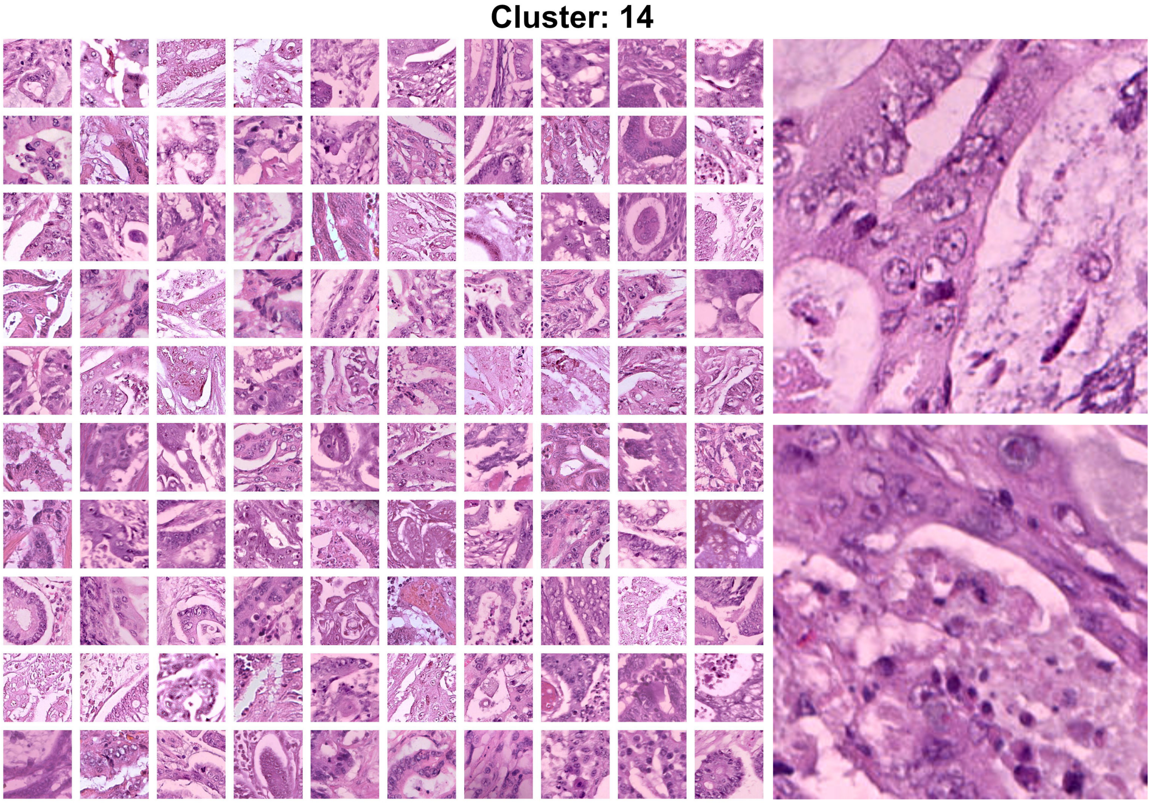


Figure 5.7: Cluster 14: The most hazardous set of tissue features as determined by self-supervised learning on the GRI cohort.

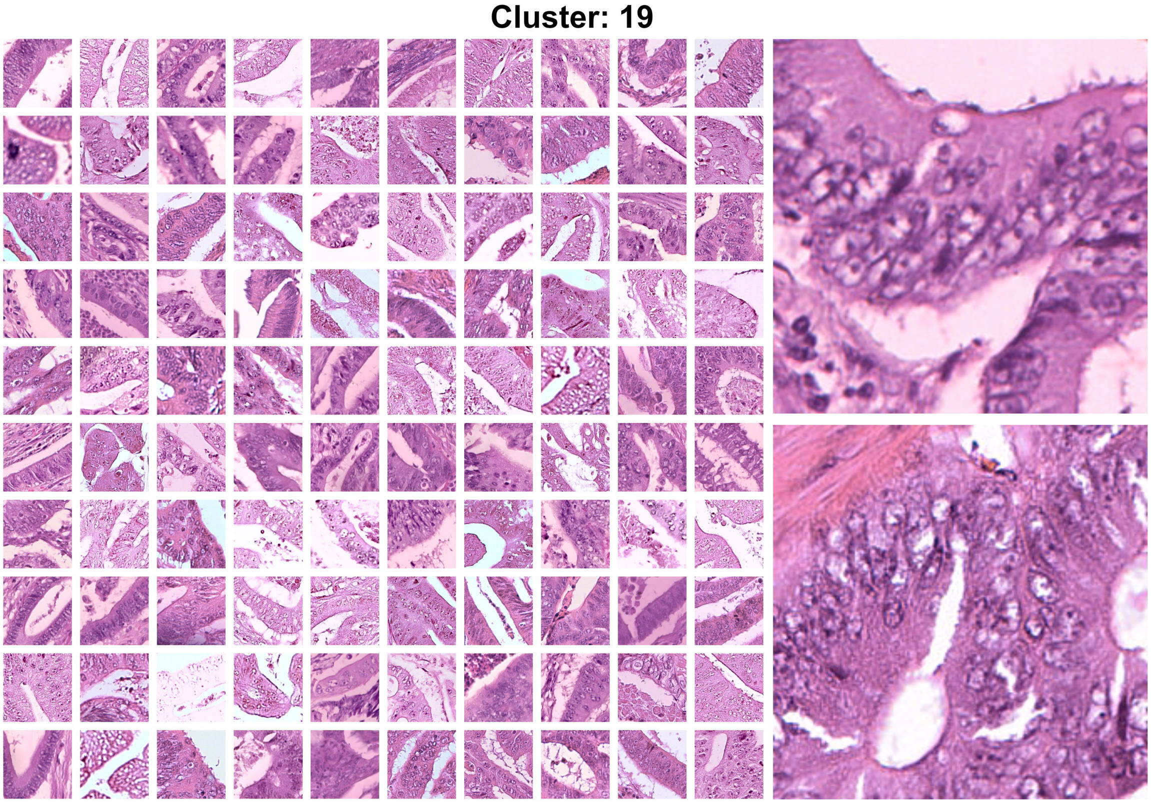


Figure 5.8: Cluster 19: The second most hazardous set of tissue features as determined by self-supervised learning on the GRI cohort.

5.3.2 SSL Tissue Classification and Invasive Margin Segmentation

The tissue classification model used the ResNet50 encoder described above was pre-trained using self-supervised learning on the GRI cohort, with a nine-class dense layer appended to the top. We trained it in two ways. The performance metrics of both models and how they compare to the supervised model trained in chapter 4 are given in table 5.1. The first model was trained by allowing all network weights to be trained on the NCT dataset by Kather et al. (2018). This resulted in an excellent training and validation accuracy of 98.74% and 96.44%, respectively. However, as can be observed in the bottom right quadrant of figure 5.9, it performed poorly when applied to the real GRI and AP datasets, misclassifying a lot of the tumour as stroma and background as debris. The second model was trained by finetuning only the dense layer on the NCT dataset and fixing the weights learned by self-supervised learning in the ResNet50 encoder. Interestingly, the optimal epoch of this model had reduced performance on the NCT dataset, with a training and validation accuracy of 90.42% and 90.03%, respectively. However, as can be observed in the top right quadrant of figure 5.9, when assessed on the GRI and AP cohorts, it had far superior real-world tissue classification performance that was much more robust to artefacts and scanning errors than the supervised model trained on the NCT dataset alone (visible in the bottom left quadrant of the same figure).

The improved tissue classification model was then used to classify the tissue in the GRI and AP cohorts. The resulting tissue maps were used to segment the invasive margin, as described in section 5.2.3. Figure 5.10 displays some of the resulting invasive margin masks. The model detected the invasive margin with a high degree of accuracy, and the resulting masks were much more accurate than those produced by our previous method. The centre line of the resulting invasive margin polygon was then used to determine the virtual high-power fields for feature extraction for the transformer network.

Model	Accuracy	Precision	Recall
Supervised (Training)	0.9867	0.9864	0.9865
Supervised (Validation)	0.9612	0.9462	0.9475
Self-Supervised Free (Training)	0.9874	0.9865	0.9866
Self-Supervised Free (Validation)	0.9644	0.9484	0.9522
Self-Supervised Fixed (Training)	0.9042	0.9004	0.9003
Self-Supervised Fixed (Validation)	0.9003	0.8559	0.8704

Table 5.1: Performance Metrics of our tissue classification models on the nine-class NCT Dataset (Kather et al. 2018).

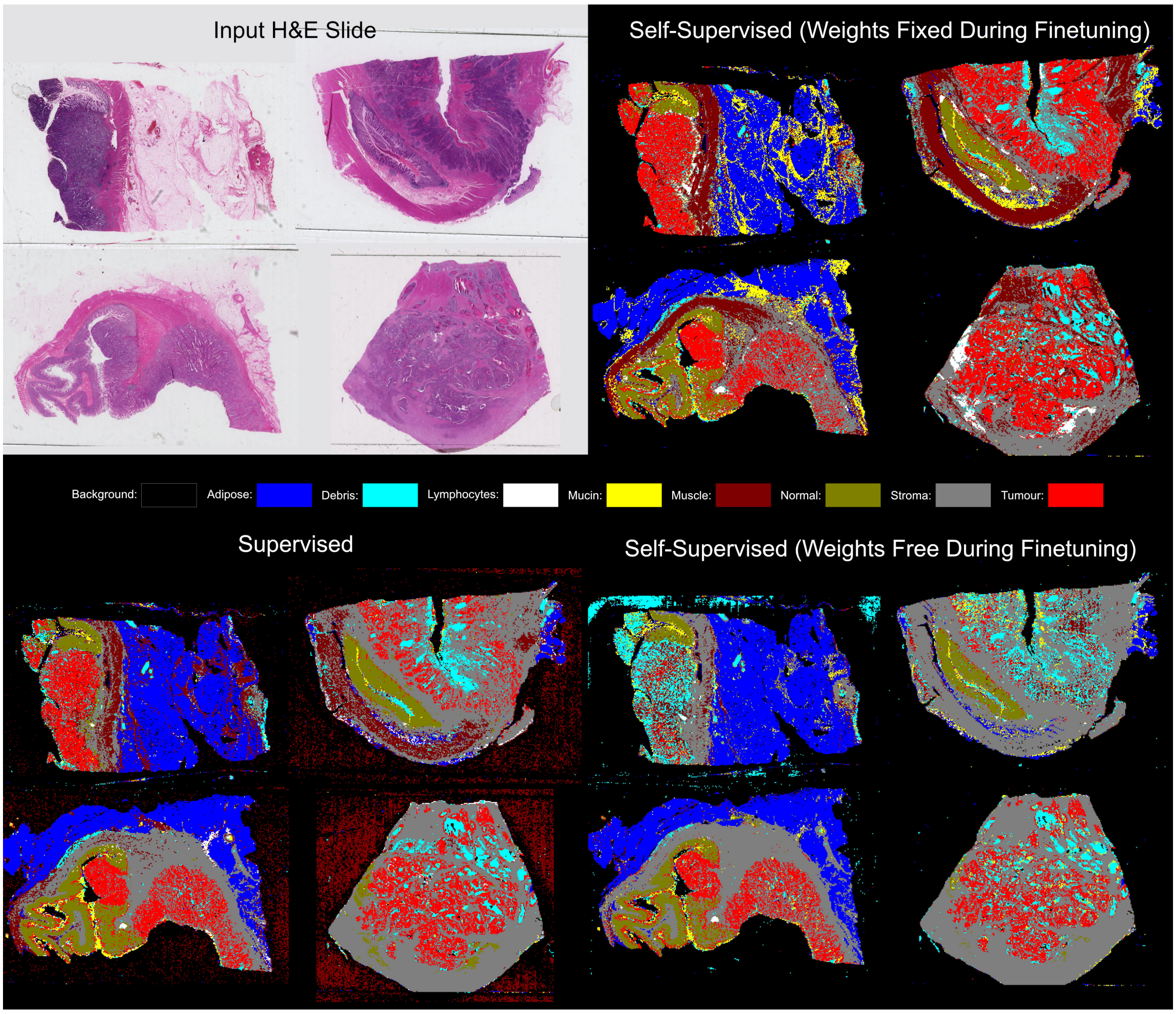

Input H&E Slide
Self-Supervised (Weights Fixed During Finetuning)
Background:
Adipose:
Debris:
Lymphocytes:
Mucin:
Muscle:
Normal:
Stroma:
Tumour:
Supervised
Self-Supervised (Weights Free During Finetuning)

Figure 5.9: A visual depiction of the tissue classes. The top left quadrant displays some selected input H&E slides that the supervised model from chapter 4 struggled with, as displayed in the bottom left quadrant. The bottom right quadrant displays the results of the self-supervised model with unlocked weights in all layers, effectively resuming supervised learning. The top right quadrant displays the results of the self-supervised model with locked weights in all layers but the final dense layer, effectively fine-tuning the model. Note the improved classification ability.

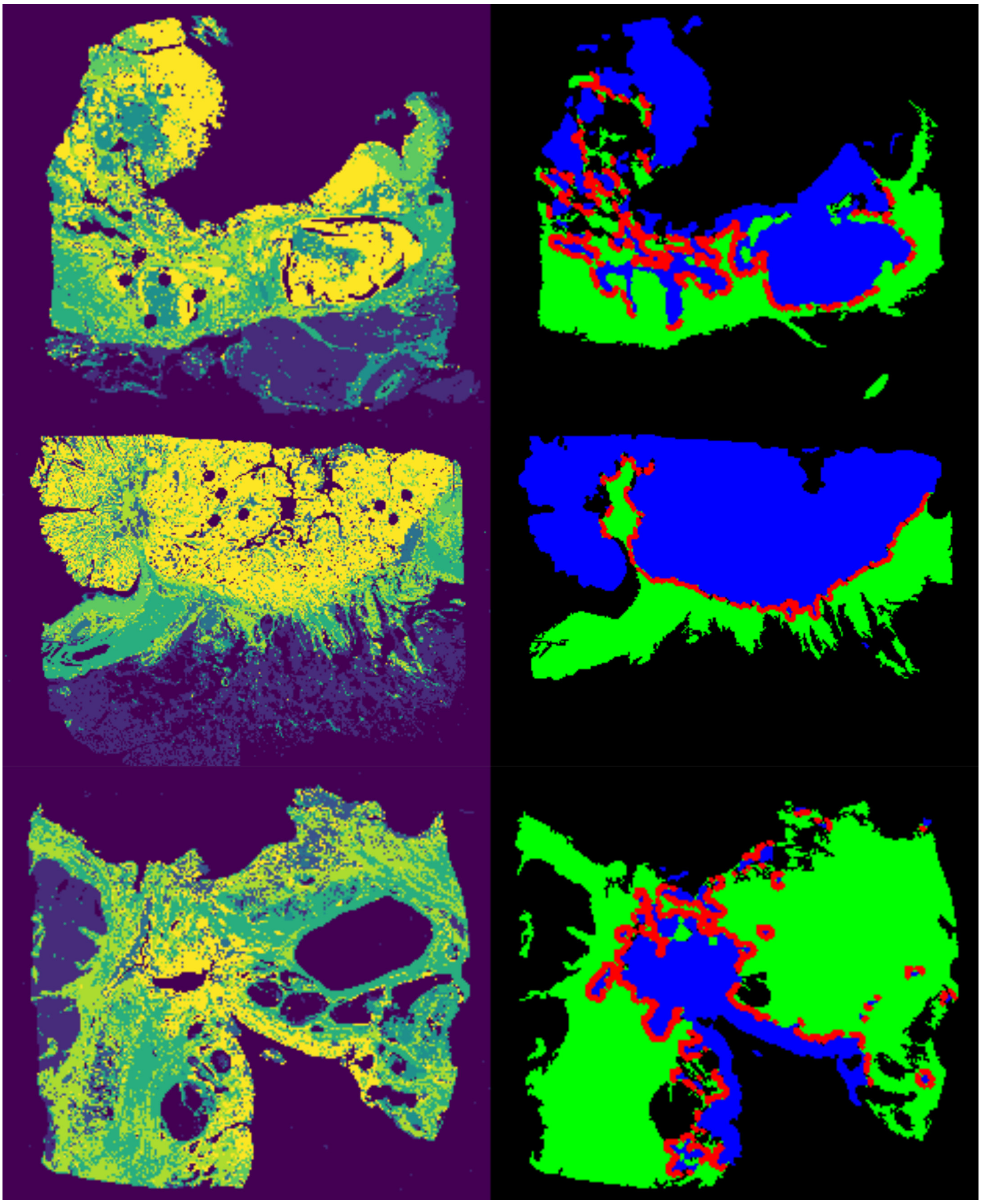


Figure 5.10: Displayed here are some select examples of the performance of our invasive margin segmentation algorithm based on the slide tissue classification masks. On the left, the aggregated tumour mask is visible in yellow, and the tissue mask is shown in green. On the right, the resulting overlayed invasive margin mask is visible in red.

5.3.3 Transformer Network for Aggression Prediction

A performant tissue classification model allowed for accurate invasive margin segmentation across all slides in the GRI and AP cohorts. The detected margins were then tessellated into tiles of size 224^2. All tissue except tumour tissue was discarded to allow the invasive margin sequences to meet the memory constraints of our GPUs. The remaining tiles were run through the self-supervised feature extraction model, resulting in a vector of 2048 values representing the tissue features for each tile. These were then unravelled into a sequence representing the invasive margin. This varied in length from approximately 2,000 to 12,000 tiles. The maximum memory capacity of the GPUs allowed for maximum sequence lengths of approximately 10,000 tiles. In cases with more than this, the sequences were truncated. Our transformer network was designed for automated aggression prediction from these margin sequences and was trained as described in section 5.2.4. It was trained over 200 slides selected from the GRI cohort, evenly split between high and low-budding patients. The remainder of the GRI cohort was held out for validation. The best training epoch was chosen based on the classification performance of the network. The final model was evaluated for classification accuracy and Cox-regression survival analysis over the GRI and AP cohorts, the results of which we shall present below.

GRI Cohort

The best-performing aggression prediction model resulted from epoch 89. On the remainder of the GRI cohort, it was used to predict high and low aggression, and when compared to the high and low budding scores, it achieved an accuracy of 63.47%, with a precision and recall of 0.4383 and 0.8308, respectively. It had an area under the ROC curve of 0.7501 and a Cohen's Kappa of 0.3042, indicating moderate agreement with the budding scores. The complete metric list is provided in table 5.2.

Figure 5.11 presents the most informative evaluation, the Cox-regression survival analysis. Panel a) displays the Kaplan Meier plot of the groups determined by manual budding, which are significantly separated as determined by the log-rank test given in Table 1 of that panel, and the hazard ratio of the high budding group is 2.14 (1.62-2.85) provided by table 3. Panel b) displays the Kaplan Meier plot of the low and high aggression groups determined by our automated aggression prediction model, the distributions of which are significantly separated as determined by the log-rank test. The hazard ratio from the two-class automated aggression score was 2.52 (1.84-3.47); this surpasses the manual budding scores. The ITBCC recommendations suggest splitting patients into a low, medium and high budding group to select those most at risk. Our aggression prediction model outputs a continuous probability score, which allows arbitrary segregation and the separation of the output probabilities into three classes. Panel c) provides the Kaplan Meier plot of these groups, where the medium aggression group has a hazard ratio of 2.31 (1.65-3.24) and the high aggression group has a hazard ratio of 3.1 (2.09-4.06).

Finally, the independence of the automated aggression score was evaluated over the GRI cohort to determine if the model had learned to detect a biomarker already present in our database. Univariate and multivariate cox-regression survival analysis was performed. When manual budding was considered, it had a hazard ratio of 1.54 (1.09-2.19), and with manual budding removed, it had an improved hazard ratio of 1.79 (1.29-2.51), demonstrating a correlation with manual budding. However, it was still statistically independent of all other biomarkers in both cases. The full cohort biomarker analysis results are provided in figure 5.3.3.

Metric	**Value**
Accuracy	0.6347
Precision	0.4383
Recall	0.8308
ROC AUC	0.7501
F1	0.5738
Cohens Kappa	0.3042
Hazard Ratio	3.1 (2.09-4.6)

Table 5.2: These are the performance metrics for our automated aggression prediction model over the GRI cohort compared to the manual budding scores.

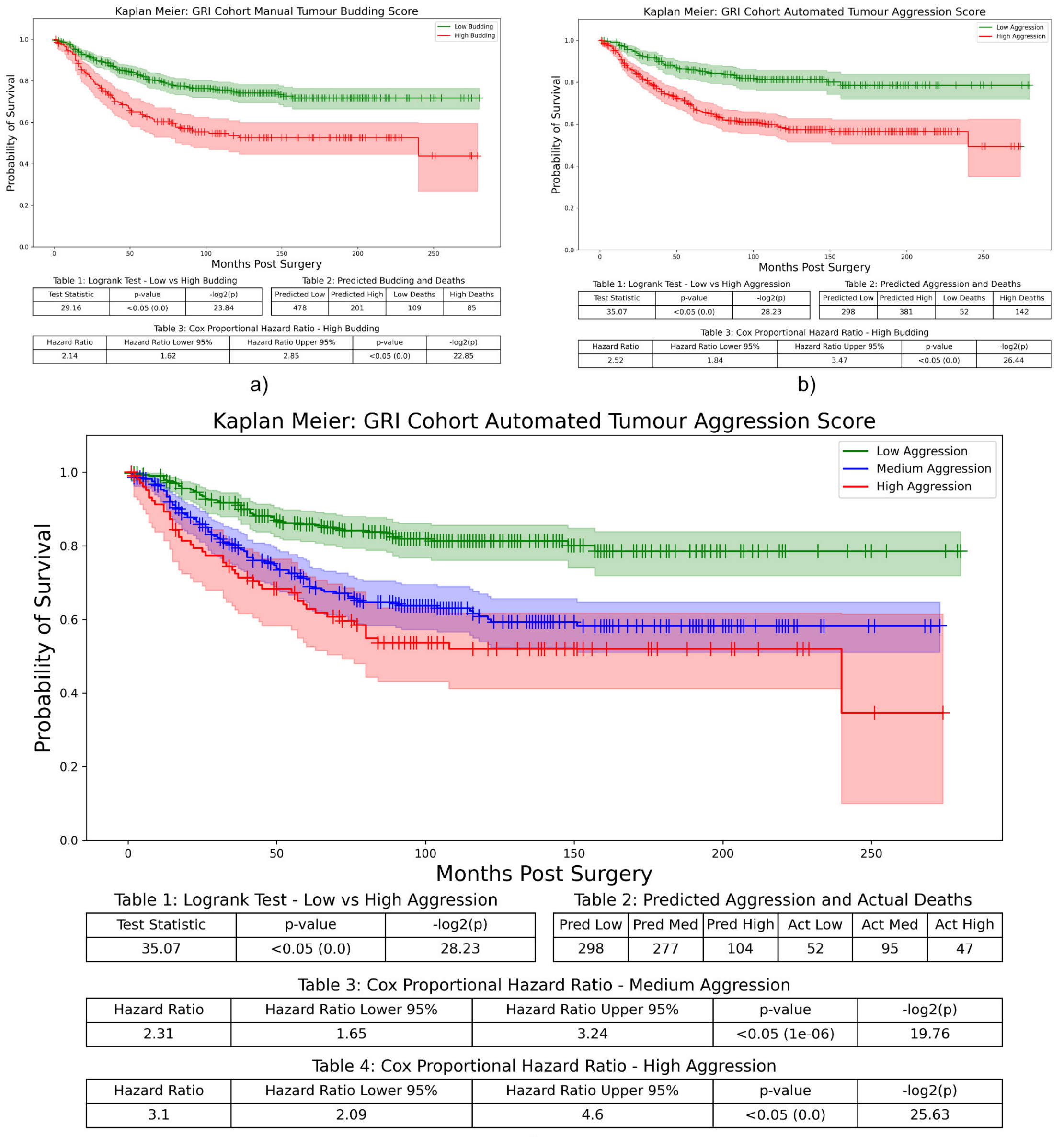


Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
29.16	<0.05 (0.0)	23.84

Table 2: Predicted Budding and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
478	201	109	85

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
2.14	1.62	2.85	<0.05 (0.0)	22.85

a)

Table 1: Logrank Test - Low vs High Aggression

Test Statistic	p-value	-log2(p)
35.07	<0.05 (0.0)	28.23

Table 2: Predicted Aggression and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
298	381	52	142

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
2.52	1.84	3.47	<0.05 (0.0)	26.44

b)

Table 1: Logrank Test - Low vs High Aggression

Test Statistic	p-value	-log2(p)
35.07	<0.05 (0.0)	28.23

Table 2: Predicted Aggression and Actual Deaths

Pred Low	Pred Med	Pred High	Act Low	Act Med	Act High
298	277	104	52	95	47

Table 3: Cox Proportional Hazard Ratio - Medium Aggression

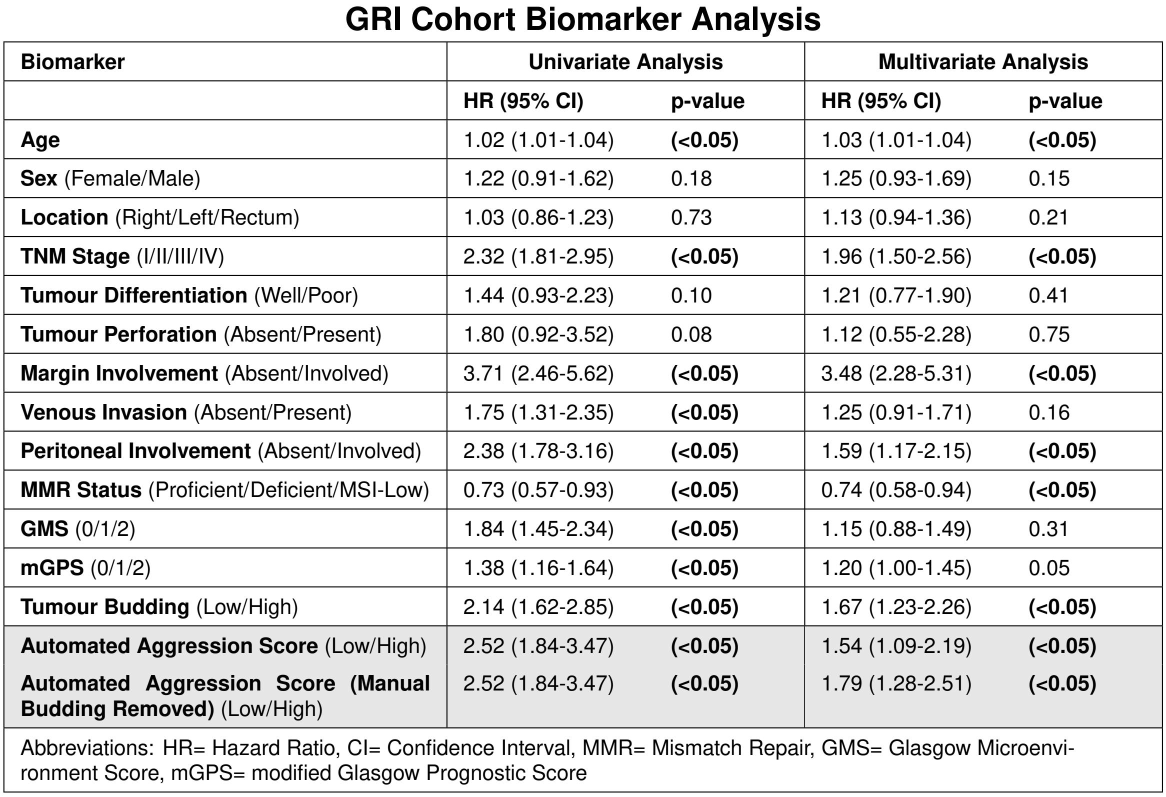
Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
2.31	1.65	3.24	<0.05 (1e-06)	19.76

Table 4: Cox Proportional Hazard Ratio - High Aggression

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
3.1	2.09	4.6	<0.05 (0.0)	25.63

c)

Figure 5.11: Shown here are the results of the survival analysis of the automated aggression scores on the GRI cohort. Panel a) displays the Kaplan Meier plots, log-rank statistics, and cox-regression hazard ratios on the manual budding scores, panel b) displays the same information based on our automated aggression score, and panel c) displays the same information based on the ITBCC recommended three-class automated aggression score.



GRI Cohort Biomarker Analysis

Biomarker	**Univariate Analysis**		**Multivariate Analysis**	
	HR (95% CI)	**p-value**	**HR (95% CI)**	**p-value**
Age	1.02 (1.01-1.04)	**(<0.05)**	1.03 (1.01-1.04)	**(<0.05)**
Sex (Female/Male)	1.22 (0.91-1.62)	0.18	1.25 (0.93-1.69)	0.15
Location (Right/Left/Rectum)	1.03 (0.86-1.23)	0.73	1.13 (0.94-1.36)	0.21
TNM Stage (I/II/III/IV)	2.32 (1.81-2.95)	**(<0.05)**	1.96 (1.50-2.56)	**(<0.05)**
Tumour Differentiation (Well/Poor)	1.44 (0.93-2.23)	0.10	1.21 (0.77-1.90)	0.41
Tumour Perforation (Absent/Present)	1.80 (0.92-3.52)	0.08	1.12 (0.55-2.28)	0.75
Margin Involvement (Absent/Involved)	3.71 (2.46-5.62)	**(<0.05)**	3.48 (2.28-5.31)	**(<0.05)**
Venous Invasion (Absent/Present)	1.75 (1.31-2.35)	**(<0.05)**	1.25 (0.91-1.71)	0.16
Peritoneal Involvement (Absent/Involved)	2.38 (1.78-3.16)	**(<0.05)**	1.59 (1.17-2.15)	**(<0.05)**
MMR Status (Proficient/Deficient/MSI-Low)	0.73 (0.57-0.93)	**(<0.05)**	0.74 (0.58-0.94)	**(<0.05)**
GMS (0/1/2)	1.84 (1.45-2.34)	**(<0.05)**	1.15 (0.88-1.49)	0.31
mGPS (0/1/2)	1.38 (1.16-1.64)	**(<0.05)**	1.20 (1.00-1.45)	0.05
Tumour Budding (Low/High)	2.14 (1.62-2.85)	**(<0.05)**	1.67 (1.23-2.26)	**(<0.05)**
Automated Aggression Score (Low/High)	2.52 (1.84-3.47)	**(<0.05)**	1.54 (1.09-2.19)	**(<0.05)**
Automated Aggression Score (Manual Budding Removed) (Low/High)	2.52 (1.84-3.47)	**(<0.05)**	1.79 (1.28-2.51)	**(<0.05)**

Abbreviations: HR= Hazard Ratio, CI= Confidence Interval, MMR= Mismatch Repair, GMS= Glasgow Microenvironment Score, mGPS= modified Glasgow Prognostic Score

Table 5.3: Univariate Cox regression survival analysis was used to determine hazard ratios (HR) and 95% Confidence Intervals (CI) for available GRI cohort biomarkers. Multivariable Cox regression survival analysis using a backward conditional elimination model and a statistical significance threshold of 0.05 was performed to identify independent prognostic biomarkers. Automated aggression scoring was an independent predictor of survival in both univariate and multivariate analyses.

AP Cohort

The best-performing aggression prediction model was also validated on the AP cohort of entirely held-out slides from a different institution. Two-class aggression scores were retrieved for all two thousand slides, and the probabilities were averaged over all available slides for a patient. The resulting aggression scores were compared to the high and low budding scores. It achieved an accuracy of 71.26%, with a precision and recall of 0.4735 and 0.7862, respectively. It had an area under the ROC curve of 0.8101 and a Cohen's Kappa of 0.3898, indicating moderate agreement with the budding scores. The complete metric list is provided in table 5.4.

Figure 5.11 presents the Kaplan Meier plots and Cox-regression survival analysis of the manual budding, two-class and three-class automated aggression scores. Panel a) displays the Kaplan Meier plot of the groups determined by manual budding, which are significantly separated as determined by the log-rank test given in Table 1 of that panel, and the hazard ratio of the high budding group is 9.47 (6.87-13.06) provided by Table 3. Panel b) displays the Kaplan Meier plot of the low and high aggression groups determined by our automated aggression prediction model, the distributions of which are significantly separated as determined by the log-rank test. The hazard ratio from the two-class automated aggression score was 3.16 (2.31-4.32). The ITBCC recommendations suggest splitting patients into a low, medium and high budding group to select those most at risk. Our aggression prediction model outputs a continuous probability score, which allows arbitrary segregation and the separation of the output probabilities into three classes. Panel c) provides the Kaplan Meier plot of these groups, where the medium aggression group has a hazard ratio of 1.92 (1.3-2.86) and the high aggression group has a hazard ratio of 3.89 (2.7-5.6).

Finally, the independence of the automated aggression score was evaluated over the AP cohort to determine if the model had learned to detect a biomarker already present in our database. Univariate and multivariate cox-regression survival analysis was performed. When manual budding was considered, it had a hazard ratio of 0.92 (0.60-1.42), and with manual budding removed, it had a hazard ratio of 1.92 (1.31-2.82) in multivariate analysis, showing a correlation with manual budding. The automated aggression score was not independent with manual budding included in the cohort, but it is independent with it removed. The full cohort biomarker analysis results are provided in figure 5.3.3.

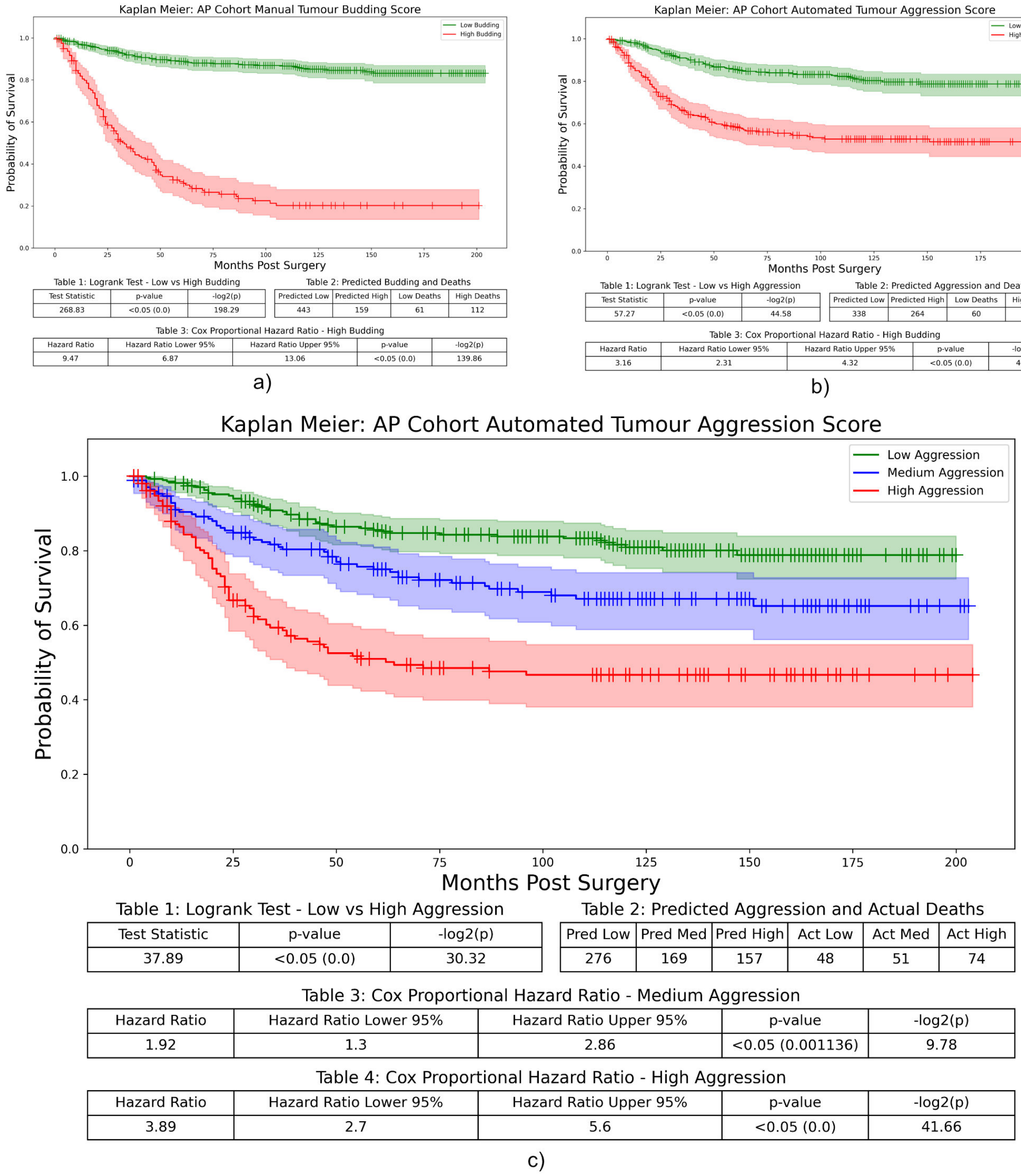


Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
268.83	<0.05 (0.0)	198.29

Table 2: Predicted Budding and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
443	159	61	112

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
9.47	6.87	13.06	<0.05 (0.0)	139.86

a)

Table 1: Logrank Test - Low vs High Aggression

Test Statistic	p-value	-log2(p)
57.27	<0.05 (0.0)	44.58

Table 2: Predicted Aggression and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
338	264	60	113

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
3.16	2.31	4.32	<0.05 (0.0)	40.39

b)

Table 1: Logrank Test - Low vs High Aggression

Test Statistic	p-value	-log2(p)
37.89	<0.05 (0.0)	30.32

Table 2: Predicted Aggression and Actual Deaths

Pred Low	Pred Med	Pred High	Act Low	Act Med	Act High
276	169	157	48	51	74

Table 3: Cox Proportional Hazard Ratio - Medium Aggression

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
1.92	1.3	2.86	<0.05 (0.001136)	9.78

Table 4: Cox Proportional Hazard Ratio - High Aggression

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
3.89	2.7	5.6	<0.05 (0.0)	41.66

c)

Figure 5.12: Shown here are the results of the survival analysis of the automated aggression scores on the AP cohort. Panel a) displays the Kaplan Meier plots, log-rank statistics, and cox-regression hazard ratios on the manual budding scores, panel b) displays the same information based on our automated aggression score, and panel c) displays the same information based on the ITBCC recommended three-class automated aggression score.

Metric	**Value**
Accuracy	0.7126
Precision	0.4735
Recall	0.7862
ROC AUC	0.8101
F1	0.5910
Cohens Kappa	0.3898
Hazard Ratio	3.89 (2.7-5.6)

Table 5.4: These are the performance metrics for our automated aggression prediction model over the AP cohort compared to the manual budding scores.

AP Cohort Biomarker Analysis

Biomarker	**Univariate Analysis**		**Multivariate Analysis**	
	HR (95% CI)	**p-value**	**HR (95% CI)**	**p-value**
Age	1.00 (0.99-1.01)	0.93	1.01 (0.99-1.02)	0.29
Sex (Female/Male)	1.18 (0.88-1.59)	0.27	1.10 (0.76-1.60)	0.61
Location (Right/Left/Rectum)	0.93 (0.76-1.13)	0.46	1.30 (1.01-1.68)	**(<0.05)**
TNM Stage (I/II/III/IV)	3.13 (2.48-3.94)	**(<0.05)**	2.37 (1.76-3.19)	**(<0.05)**
Tumour Differentiation (Well/Poor)	2.53 (1.70-3.77)	**(<0.05)**	1.88 (1.14-3.09)	**(<0.05)**
Tumour Perforation (Absent/Present)	2.44 (1.57-3.79)	**(<0.05)**	1.47 (0.86-2.51)	0.16
Margin Involvement (Absent/Involved)	4.33 (2.76-6.78)	**(<0.05)**	1.21 (0.65-2.25)	0.55
Venous Invasion (Absent/Present)	2.22 (1.65-2.99)	**(<0.05)**	0.92 (0.61-1.38)	0.68
Peritoneal Involvement (Absent/Involved)	2.91 (2.16-3.92)	**(<0.05)**	0.90 (0.58-1.38)	0.63
MMR Status (Proficient/Deficient/MSI-Low)	0.78 (0.62-0.98)	**(<0.05)**	0.97 (0.71-1.31)	0.82
GMS (0/1/2)	2.08 (1.67-2.59)	**(<0.05)**	1.52 (1.15-2.00)	**(<0.05)**
mGPS (0/1/2)	1.92 (1.56-2.36)	**(<0.05)**	1.67 (1.34-2.09)	**(<0.05)**
Tumour Budding (Low/High)	9.47 (6.87-13.06)	**(<0.05)**	6.96 (4.67-10.38)	**(<0.05)**
Automated Aggression Score (Low/High)	3.16 (2.31-4.32)	**(<0.05)**	0.92 (0.60-1.42)	0.71
Automated Aggression Score (Manual Budding Removed) (Low/High)	3.16 (2.31-4.32)	**(<0.05)**	1.92 (1.31-2.82)	**(<0.05)**
Abbreviations: HR= Hazard Ratio, CI= Confidence Interval, MMR= Mismatch Repair, GMS= Glasgow Microenvironment Score, mGPS= modified Glasgow Prognostic Score				

Table 5.5: Univariate Cox regression survival analysis was used to determine hazard ratios (HR) and 95% Confidence Intervals (CI) for available AP cohort biomarkers. Multivariable Cox regression survival analysis using a backward conditional elimination model and a statistical significance threshold of 0.05 was performed to identify independent prognostic biomarkers. Automated tumour bud scoring was correlated with manual budding in the AP cohort.

5.4 Discussion

There is heterogeneity in the survival of stage II colorectal cancer patients, with some patients having as poor an outcome as stage III or IV patients (Maguire 2014). This is attributable to tumours with a range of aggressiveness being treated similarly because they are assigned the same stage in the current system (Lea et al. 2014). Tumour budding has now been recognised as a hallmark of aggression that correlates with a higher TNM stage and tumour grade and indicates a higher probability of lymph node involvement and distant metastases (Rogers et al. 2016; Petrelli et al. 2015). However, despite this known correlation with aggression and the fact that budding is an independent predictor of survival and prognosis that can stratify these patients, it has yet to be adopted into routine clinical reporting in colorectal cancer (Loughrey et al. 2018). This is due to a lack of consensus on a standardised scoring method, the time-intensive nature of the task, and the poor inter-observer agreement that results from H&E scoring (Lugli et al. 2017). An automated method of bud scoring would help address these challenges. As discussed in sections 2.9, several recent publications have attempted to develop strategies for this in H&E. However, generating the manual segmentations required as ground truth to train these models is difficult at scale, a limitation we help to address in chapter 4. An end-to-end method of survival prediction is the most feasible method of training over a large enough set of whole slide images and patients that would allow the model to be robust and sufficiently generalisable for clinical use. In section 2.10, we reviewed the published research on existing survival prediction methods, and while many were successful, all used disease-specific survival in months or years or the current survival status as their ground truth. These factors can be influenced by several variables like quality of healthcare received, general health, and adherence to treatment, among many others. Therefore, we proposed that patient-level labels for tumour budding and, by proxy, aggression were used, as they focus on a more specific biological phenomenon linked to aggressive tumour behaviour, which should result in more accurate and targeted predictions than the broader task of survival prediction where a myriad of factors influences the ground truth. We shall now discuss the results of our work to apply self-supervised learning and transformer networks to predict aggression directly from H&E whole slide images.

We first trained a feature extraction network using self-supervised learning from over 12 million tiles extracted from the GRI patient cohort to achieve this goal. Evaluating the semantic relevance of the encoded representations is difficult, given that the training data was entirely unlabelled. However, we sought to determine how meaningful the encoded representations were by utilising Leiden clustering on the k-nearest neighbour graph of the representations and then determining the Cox-regression hazard ratio of the clusters. We hoped this would reveal how well the different tiles were grouped in the representation space by sets containing tiles with tumour or mucin having a higher hazard ratio than muscle or adipose tissue. The large hazard ratios observed in many of the resulting clusters underscore the potential of self-supervised learning to extract meaningful information directly from unlabelled whole slide images. This result also suggested that the learned representations captured pertinent histological information and were aptly tailored for downstream tasks. This chapter is ongoing work, and we hope to delve deeper into the tissue types identified in the clusters and attempt to discern known from any potentially unknown histological features. This would allow for more interpretability in the decisions of the downstream models and potentially reveal new pathology.

Once a suitable performance had been achieved in the feature extraction network, this was used to improve our tissue classification framework. Finetuning the model weights learned by self-supervised learning on a small, labelled dataset resulted in a much more robust and generalisable tissue classification model that could handle the slides from different institutions and scanners as present in the GRI and AP cohorts. Freezing the learned encoder weights and only allowing the final dense classification layer resulted in the best performance, far superior to the supervised model or the model where the weights learned through self-supervision were allowed to be updated. There may be several reasons for this, such that the frozen weights might help to prevent overfitting, ensuring the model doesn't overly adapt to the small, labelled dataset. It may also be that freezing the weights helps to preserve the general features learned from the vast unlabelled dataset, maintaining their broad applicability. Or it might even have a regularisation effect and mitigate overfitting risks. It could also be that the self-supervised weights better encode the dataset distribution, and allowing the weights to be updated on the small, labelled dataset introduces bias. From a computational perspective, the frozen weights also had several advantages. The training was much more efficient and stable, which may help avoid potential pitfalls from drastic weight updates.

The resulting segmentation of the invasive margin was much improved due to the superior quality of the tissue classification. This allowed better targeting of regions around the tumour core and excluding areas like mucin and debris. However, large parts of the slides in the AP cohort were unreadable due to being drastically out of focus, obscured entirely by pen or folded tissue. No level of generalisation ability will help if the information is degraded to this level. However, the tissue classification in the AP cohort was much improved compared to the supervised model.

The ability to extract histologically information-rich feature vectors in a targeted manner around the tumour core unlocked the ability to feed a deep neural network with tile sequences representing the invasive margin. We trained a transformer network to learn to recognise the configuration of encoded histological features and predict a tumour aggression score that correlates with tumour budding. We cannot constrain the features the network recognises as relevant to prediction, so we cannot guarantee it learns to recognise budding; therefore, we deem the network an aggression prediction model. However, the output classifications of the GRI cohort had higher Cohen's kappa values when compared to the automated bud scoring model designed to directly identify budding, with the transformer model score being 0.3042 versus 0.1898 for the manual model on the GRI cohort and 0.3898 versus 0.2665 for the AP cohort, suggesting that the transformer model is learning to encode information that allows it to correlate better with the human pathologist's determination of aggression. The improved accuracy, precision, recall and F1 scores of the transformer-based classifications over the GRI and AP cohorts also support this idea.

The most salient evaluation of the network performance was the cox-regression survival analysis of the transformer classifications on the two cohorts. We saw that the high-aggression patient groups determined by the transformer had a hazard ratio of 2.52 for two-class aggression versus 2.14 in the automated bud-scoring model from chapter 4. This improvement generalised to the held-out AP cohort with a transformer-based hazard ratio of 3.16 for two-class aggression versus 3.01 for the automated bud scoring model from chapter 4. Most importantly, in univariate and multivariate analysis with all available biomarkers, the automated deep aggression score was an independent biomarker for survival in the GRI cohort. It is independent in univariate analysis and was also independent in multivariate analysis when manual budding was removed from consideration in the AP cohort.

Despite the success of this model, we are currently limited by available commercial hardware to the sequence length of tissue we can examine in the invasive margin, and we hope as technology improves or if we gain access to more computational resources, we can expand the representation vector length and consider more tissue classes during aggression prediction. The transformer model also allows for examining the weights it assigns to the importance of the different elements in the input sequence as they relate to the prediction. This allows for a high level of model interpretability. As we make the process more efficient and gain access to more resources, we will seek to examine how the interplay between tissue types determines the aggression score and hope to extract new information regarding how the tumour microenvironment influences or is influenced by tumour aggression in the process.

In this work, a transformer network, operating on self-supervised feature representations extracted from unlabelled H&E tiles, demonstrated superior classification capability and also showed increased Cohen's kappa values, highlighting an increased correlation with human pathologists as compared to the automated bud scoring model from chapter 4. It also had a superior ability to stratify patients with more prognostic Cox-regression hazard ratios in the high-aggression groups. The availability of a continuous probability also enabled a more nuanced stratification of the colorectal cancer patients into the low, medium, and high aggression groups better fitting with the ITBCC recommendations, which could be pivotal in informing clinical decisions. The transformer network not only enhances the precision of aggression scoring but also holds the potential to unveil novel pathological features through model interpretation. Given adequate computational resources, it would be possible to scale this method to whole slide images, revealing a comprehensive overview of how the tumour and microenvironment interrelate during prediction. If further developed, this model could facilitate a more accurate and detailed stratification of patients and unlock unknown information about the tumour microenvironment. This should translate into tailored treatment strategies and improved patient survival. This underscores the rapid progress of deep learning and the transformative potential of integrating advanced computational techniques into research and clinical pathology.

Chapter 6

Conclusion

Tumour budding has emerged as a pivotal prognostic indicator in colorectal cancer, shedding light on the aggressive nature of certain tumours and the reason for the heterogeneity in survival in similarly staged colorectal tumours as determined by the current TNM staging system (Maguire 2014). Despite its recognised importance, integration into routine clinical reporting has been hindered due to challenges such as the lack of a standardised scoring methodology, the time-intensive nature of manual assessment, and the variability in the interpretations between pathologists when scoring in H&E (Lea et al. 2014; Rogers et al. 2016; Petrelli et al. 2015; Loughrey et al. 2018; Lugli et al. 2017). The subjectivity inherent in human interpretation underscores the need for more consistent and objective methods for evaluating tumour aggression. The development of automated approaches offers a promising avenue to address these challenges. In this work, we have presented three methods to assist or automate the process of tumour bud scoring and aggression prediction in colorectal cancer.

The first chapter of this work explored the feasibility of translating H&E whole slide images to virtual versions of the AE1/AE3 IHC stain that highlights tumour buds and can subsequently reduce the time required for manual evaluation and inter-observer variability. The primary objective was to develop a tool to assist pathologists in tumour bud scoring and create a high-quality ground truth dataset for further automated methods. To ensure the accuracy of this translation, it was imperative that the model retained the intricate histological features present in the original images. The CycleGAN model proposed by Zhu et al. (2020) emerged as a promising architecture for the retention of structural detail. However, given its unpaired training objective, it did not maintain accuracy in the generated virtual staining. We proposed an enhancement to the loss function through the addition of a mid-cycle loss term that constrained the generated virtual staining to be similar to real IHC ground truth. This ensured that the resulting model reproduced precise histological details and encouraged the generation of accurate virtual staining.

A significant challenge in this process was the need for precise stain normalisation and alignment of whole slide images that was accurate to the cellular level. We developed the Beatson Augmentation and Stain Normalisation Toolkit to address the issue of accurate whole slide stain normalisation. The toolkit improved the consistency and reliability of stain normalisation while avoiding the introduction of artefacts that would affect the quality of the generated virtual immunohistochemistry images. Additionally, a framework was established for meticulously aligning whole slide images, which was essential for creating accurate paired H&E and IHC training data.

This resulted in an architecture and associated protocols to train a more accurate and reliable virtual pan-cytokeratin AE1/AE3 model. This approach, when compared to other models, improved scores for the structural similarity index measure, indicating higher structural agreement with the ground truth, and reduced Fréchet inception distance scores, indicating improved realism in the generated images. This model could assist with and improve the diagnostic process of colorectal cancer, enabling the generation of a virtual IHC image shortly after the H&E image is processed as part of standard reporting. This rapid turnaround could expedite the manual tumour bud scoring process, reducing the costs and complexities associated with manual tumour bud scoring.

The second chapter of this work aimed to develop a method of training an end-to-end tumour bud segmentation and scoring model using manually generated H&E segmentations created by referencing virtual IHC pan-cytokeratin versions of the whole slide images. The virtual IHC model was instrumental in highlighting the tumour cells or small clusters in the invasive margin of the input H&E slides. This resulted in a dataset of almost 60,000 manually segmented buds. These were then used to train a U-Net segmentation model with a ResNet50 encoder to detect buds directly in H&E whole slide images. The virtual IHC slides were also used to detect the tumour core and automatically determine the invasive margin. However, this had limitations as the tissue in the margin could not be filtered for areas of mucin or glandular fragmentation known to exhibit pseudo-budding.

The model performance was evaluated on two distinct cohorts, revealing intriguing insights. In the GRI cohort, the model hinted at the existence of two distinct bud populations: smaller clusters close to the tumour core and larger ones that had migrated a distance away from the core. This separate population was not observed in the AP patient cohort. However, slides in this group had many scanning artefacts that potentially impacted the segmentation accuracy of the model. Despite these challenges, the automated system effectively differentiated between high and low-budding populations in both cohorts. In the GRI cohort, the automated bud score surpassed manual budding in its prognostic significance and was an independent predictor of survival in univariate and multivariate Cox regression. However, the results on the AP cohort underscored the need for a more resilient model, especially in the presence of slide artefacts. Nevertheless, the model retained the ability to be prognostically significant in univariate analysis and was an independent predictor of survival if manual budding was removed, as it was highly correlated with it.

The results demonstrated that the slides produced by our virtual IHC AE1/AE3 model offer accurate ground truth in training an automated tumour bud scoring model. Thus, this is a promising avenue to enable objective and consistent manual bud scoring. The bud area and distance analysis demonstrated that the current recommendations for bud scoring may not fully capture the information available in the histological features and that more research is required. While the automated tumour bud scoring model demonstrated potential, further refinements and validations are necessary to ensure robustness and generalisability in diverse clinical settings. Future directions for this work include the verification of the disparate budding populations on a newly sourced cohort with fewer artefacts and the exploration of techniques like self-supervised learning for more robust model training. However, the proposed approach has provided a proof of concept for an end-to-end automated bud scoring model. If fully realised, such a system could revolutionise how tumour budding is assessed, leading to more accurate diagnoses and stratification of patients.

In the final chapter of this work, we employed self-supervised learning to train a feature extraction network using over 12 million tiles extracted from the GRI patient cohort using the VICReg architecture proposed by Bardes et al. (2022). The encoded representations were clustered using community detection on a k-nearest neighbour graph. The success of the model in encoding histologically relevant features was determined by performing Cox regression on an observed probability distribution of the clusters in slides across the GRI cohort. The results indicated that the self-supervised representations effectively captured pertinent histological information relevant to survival. Leveraging these learned representations, we also improved our tissue classification framework, resulting in a more robust model capable of handling diverse slide conditions. We then used this tissue classification framework to segment the invasive margin in a targeted manner, excluding mucin and debris to avoid areas of pseudo-budding, resulting in a more accurate and relevant segmentation of the invasive margin.

The final advancement proposed in this research was to use a custom transformer network to analyse the feature vectors extracted from the tumour tissue in the invasive margins of the GRI and AP patient cohorts. This model has a novel objective as we tasked it with predicting tumour aggression by utilising the patient-level budding scores as the ground truth during training. Interestingly, the transformer model demonstrated a stronger correlation with the manual budding assessments by pathologists when not constrained to only consider tumour buds as in the previously developed automated bud scoring model. Furthermore, survival analysis revealed that the transformer model classifications provided a more accurate stratification of a patient's tumour aggression, with larger Cox regression hazard ratios than the previous model.

The design of the transformer model also offered interpretability, which should allow for future insights into how different tissue types influence the aggression score.

As technology advances, we anticipate examining more tissue in the invasive margin and across whole slides, potentially revealing new insights about the tumour microenvironment and its role in aggression. This chapter provided a proof of concept that the combination of self-supervised learning and transformer networks holds immense potential in the realm of digital pathology. It offered more precise predictions and unlocked the potential to uncover novel pathological features without human intervention. The compelling result from this approach underscores the urgency of integrating deep learning into clinical pathology. The time to develop and introduce this technology has arrived. It is evident that with proper refinement, interpretability, and verification, deep learning can assist pathologists and clinicians and pave the way for more tailored treatments and improved patient survival.

References

Abadi, M. et al. (May 31, 2016). *TensorFlow: A system for large-scale machine learning*. arXiv: 1605.08695[cs].

Aeffner, F. et al. (Dec. 1, 2018). “Digital Microscopy, Image Analysis, and Virtual Slide Repository”. In: *ILAR Journal* 59.1, pp. 66–79.

Akagi, T. and M. Inomata (2020). “Essential Updates 2018/2019: Essential advances in surgical and adjuvant therapies for colorectal cancer”. In: *Annals of Gastroenterological Surgery* 4.1, pp. 39–46.

Alaryani, F. S. and S. S. T. Alrdahe (Jan. 7, 2022). “A Review of Treatment, Risk Factors, and Incidence of Colorectal Cancer”. In: *International Journal of Applied Pharmaceutics*, pp. 1–6.

Anwar, S. M., M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. Khan (Nov. 2018). “Medical Image Analysis using Convolutional Neural Networks: A Review”. In: *Journal of Medical Systems* 42.11, p. 226. issn: 0148-5598, 1573-689X.

Arnold, M., M. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray (Apr. 1, 2017). “Global patterns and trends in colorectal cancer incidence and mortality”. In: *Gut* 66.4, pp. 683–691.

Arrichiello, G. et al. (2022). “Beyond N staging in colorectal cancer: Current approaches and future perspectives”. In: *Frontiers in Oncology* 12.

Arslan, E., T. Aksoy, R. Gürsu, N. Dursun, E. Çakar, and T. Çermik (Feb. 17, 2020). “The Prognostic Value of ^{18}F-FDG PET/CT and KRAS Mutation in Colorectal Cancers”. In: *Molecular Imaging and Radionuclide Therapy* 29.1, pp. 17–24.

Ascierto, P. A., F. M. Marincola, B. A. Fox, and J. Galon (Sept. 2020). “No time to die: the consensus immunoscore for predicting survival and response to chemotherapy of locally advanced colon cancer patients in a multicenter international study”. In: *Oncoimmunology* 9.1, p. 1826132.

Badrinarayanan, V., A. Kendall, and R. Cipolla (Oct. 10, 2016). *SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation*. arXiv: 1511.00561[cs].

Bai, B., X. Yang, Y. Li, Y. Zhang, N. Pillar, and A. Ozcan (Mar. 3, 2023). “Deep learning-enabled virtual histological staining of biological samples”. In: *Light: Science & Applications* 12.1, p. 57.

Bai, Y., J. Mei, A. Yuille, and C. Xie (Nov. 9, 2021). *Are Transformers More Robust Than CNNs?* arXiv: 2111.05464[cs].

Balakrishnan, G., A. Zhao, M. R. Sabuncu, J. V. Guttag, and A. V. Dalca (Feb. 4, 2019). “VoxelMorph: A Learning Framework for Deformable Medical Image Registration”. In: *IEEE Transactions on Medical Imaging* 38.8, pp. 1788–1800.

Banaeeyan, R. et al. (Nov. 2020). “Tumor Budding Detection in H&E-Stained Images Using Deep Semantic Learning”. In: *2020 IEEE REGION 10 CONFERENCE (TENCON)*. 2020 IEEE REGION 10 CONFERENCE (TENCON), pp. 52–56.

Bankhead, P. et al. (Dec. 4, 2017). “QuPath: Open source software for digital pathology image analysis”. In: *Scientific Reports* 7, p. 16878. doi: 10.1038/s41598-017-17204-5.

Bardes, A., J. Ponce, and Y. LeCun (Jan. 28, 2022). *VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning*. arXiv: 2105.04906[cs].

Barresi, V. et al. (June 2014). “Poorly differentiated clusters (PDCs) as a novel histological predictor of nodal metastases in pT1 colorectal cancer”. In: *Virchows Archiv* 464.6, pp. 655–662.

Benny, Y., T. Galanti, S. Benaim, and L. Wolf (May 2021). “Evaluation Metrics for Conditional Image Generation”. In: *International Journal of Computer Vision* 129.5, pp. 1712–1731.

Bergler, M. et al. (2019). “Automatic Detection of Tumor Buds in Pan-Cytokeratin Stained Colorectal Cancer Sections by a Hybrid Image Analysis Approach”. In: *Digital Pathology*. Ed. by C. C. Reyes-Aldasoro, A. Janowczyk, M. Veta, P. Bankhead, and K. Sirinukunwattana. Vol. 11435. Cham: Springer International Publishing, pp. 83–90.

Besson, S. et al. (2019). “Bringing Open Data to Whole Slide Imaging”. In: *Digital Pathology*. Ed. by C. Reyes-Aldasoro, A. Janowczyk, M. Veta, P. Bankhead, and K. Sirinukunwattana. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 3–10.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre (Oct. 9, 2008). “Fast unfolding of communities in large networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008. arXiv: 0803.0476[cond-mat,physics:physics].

Bokhorst, J. M., L. Rijstenberg, D. Goudkade, I. Nagtegaal, J. van der Laak, and F. Ciompi (2018). “Automatic Detection of Tumor Budding in Colorectal Carcinoma with Deep Learning”. In: *Computational Pathology and Ophthalmic Medical Image Analysis*. Ed. by D. Stoyanov et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 130–138.

Bokhorst, J. M. et al. (May 1, 2020). “Assessment of individual tumor buds using keratin immunohistochemistry: moderate interobserver agreement suggests a role for machine learning”. In: *Modern Pathology* 33.5, pp. 825–833.

Bokhorst, J. et al. (Sept. 2023). “Fully Automated Tumor Bud Assessment in Hematoxylin and Eosin-Stained Whole Slide Images of Colorectal Cancer”. In: *Modern Pathology* 36.9, p. 100233.

Borovec, J., A. Muñoz-Barrutia, and J. Kybic (Oct. 1, 2018). “Benchmarking of Image Registration Methods for Differently Stained Histological Slides”. In: pp. 3368–3372.

Borovec, J. et al. (Apr. 7, 2020). “ANHIR: Automatic Non-Rigid Histological Image Registration Challenge”. In: *IEEE Transactions on Medical Imaging* 39.10. 29, pp. 3042–3052.

Bosch, S., S. Teerenstra, J. De Wilt, C. Cunningham, and I. Nagtegaal (July 24, 2013). “Predicting lymph node metastasis in pT1 colorectal cancer: a systematic review of risk factors providing rationale for therapy decisions”. In: *Endoscopy* 45.10, pp. 827–841.

Boschman, J. et al. (2022). “The utility of color normalization for AI-based diagnosis of hematoxylin and eosin-stained pathology images”. In: *The Journal of Pathology* 256.1, pp. 15–24.

Bouteldja, N., B. M. Klinkhammer, T. Schlaich, P. Boor, and D. Merhof (Jan. 1, 2022). “Improving unsupervised stain-to-stain translation using self-supervision and meta-learning”. In: *Journal of Pathology Informatics* 13, p. 100107.

Brierley, J., M. Gospodarowicz, and C. Wittekind (Dec. 2016). *TNM Classification of Malignant Tumours, 8th Edition*. 8th. Wiley-Blackwell.

Brockmoeller, S. et al. (2022). “Deep learning identifies inflamed fat as a risk factor for lymph node metastasis in early colorectal cancer”. In: *The Journal of Pathology* 256.3, pp. 269–281.

Burlingame, E. A. et al. (Oct. 16, 2020). “SHIFT: speedy histological-to-immunofluorescent translation of a tumor signature enabled by deep learning”. In: *Scientific Reports* 10.1, p. 17507.

Buslaev, A., A. Parinov, E. Khvedchenya, V. Iglovikov, and A. A. Kalinin (Feb. 24, 2020). “Albumentations: fast and flexible image augmentations”. In: *Information* 11.2, p. 125. arXiv: `1809.06839[cs]`.

Bychkov, D. et al. (Feb. 21, 2018). “Deep learning based tissue analysis predicts outcome in colorectal cancer”. In: *Scientific Reports* 8.1, p. 3395.

Caie, P. D., A. K. Turnbull, S. M. Farrington, A. Oniscu, and D. J. Harrison (2014). “Quantification of tumour budding, lymphatic vessel density and invasion through image analysis in colorectal cancer”. In: *Journal of Translational Medicine* 12.1, p. 156. issn: 1479-5876. (Visited on 08/03/2021).

Chen, T., S. Kornblith, M. Norouzi, and G. Hinton (June 30, 2020). *A Simple Framework for Contrastive Learning of Visual Representations*. arXiv: `2002.05709[cs, stat]`.

Chen, Y., X. Zheng, and C. Wu (2021a). “The Role of the Tumor Microenvironment and Treatment Strategies in Colorectal Cancer”. In: *Frontiers in Immunology* 12.

Chen, Z., W. Yu, I. H. M. Wong, and T. T. W. Wong (Sept. 1, 2021b). “Deep-learning-assisted microscopy with ultraviolet surface excitation for rapid slide-free histological imaging”. In: *Biomedical Optics Express* 12.9, pp. 5920–5938.

Chen, Z. et al. (Dec. 1, 2021c). “A hierarchical and multi-view registration of serial histopathological images”. In: *Pattern Recognition Letters* 152, pp. 210–217.

Cho, Y. et al. (July 15, 2019). “Genetic Risk Score, Combined Lifestyle Factors and Risk of Colorectal Cancer”. In: *Cancer Research and Treatment* 51.3, pp. 1033–1040.

Chowdhury, A., J. Rosenthal, J. Waring, and R. Umeton (Sept. 10, 2021). “Applying Self-Supervised Learning to Medicine: Review of the State of the Art and Medical Implementations”. In: *Informatics* 8.3, p. 59.

Compton, C. (Apr. 1, 2003). “Colorectal Carcinoma: Diagnostic, Prognostic, and Molecular Features”. In: *Modern Pathology* 16.4, pp. 376–388.

Culjak, I., D. Abram, T. Pribanic, H. Dzapo, and M. Cifrek (May 2012). “A brief introduction to OpenCV”. In: *2012 Proceedings of the 35th International Convention MIPRO*. 2012 Proceedings of the 35th International Convention MIPRO, pp. 1725–1730.

Dawson, H., A. Blank, I. Zlobec, and A. Lugli (Dec. 1, 2019a). “Potential clinical scenarios of tumour budding in colorectal cancer.” In: *Acta Gastro-enterologica Belgica* 82.4, pp. 515–518.

Dawson, H., R. Kirsch, D. Messenger, and D. Driman (July 1, 2019b). “A Review of Current Challenges in Colorectal Cancer Reporting”. In: *Archives of Pathology & Laboratory Medicine* 143, pp. 869–882.

Dawson, H. et al. (Mar. 1, 2019c). “Validation of the International Tumor Budding Consensus Conference 2016 recommendations on tumor budding in stage I-IV colorectal cancer”. In: *Human Pathology* 85, pp. 145–151.

De Smedt, L., S. Palmans, and X. Sagaert (Apr. 1, 2016). “Tumour budding in colorectal cancer: what do we know and what can we do?” In: *Virchows Archiv* 468.4, pp. 397–408.

De Smedt, L. et al. (Jan. 2017). “Expression profiling of budding cells in colorectal cancer reveals an EMT-like phenotype and molecular subtype switching”. In: *British Journal of Cancer* 116.1, pp. 58–65.

Delmon, V., S. Rit, R. Pinho, and D. Sarrut (Mar. 7, 2013). “Registration of sliding objects using direction dependent B-splines decomposition”. In: *Physics in Medicine and Biology* 58.5. 67, pp. 1303–1314.

Detlefsen, N. S. et al. (Feb. 11, 2022). “TorchMetrics - Measuring Reproducibility in PyTorch”. In: *Journal of Open Source Software* 7.70, p. 4101. issn: 2475-9066.

Detweiler, C. J., D. M. Cardona, D. S. Hsu, and S. J. McCall (Feb. 16, 2016). “Primary high-grade neuroendocrine carcinoma emerging from an adenomatous polyp in the setting of familial adenomatous polyposis”. In: *Case Reports* 2016, bcr2015214206.

Dey, P. (2022). *Basic and Advanced Laboratory Techniques in Histopathology and Cytology*. 2nd. Singapore: Springer Nature Singapore.

Dimitriou, N., O. Arandjelović, D. J. Harrison, and P. D. Caie (Oct. 2, 2018). "A principled machine learning framework improves accuracy of stage II colorectal cancer prognosis". In: *npj Digital Medicine* 1.1, pp. 1–9.

Ehteshami-Bejnordi, B., N. Timofeeva, I. Otte-Höller, N. Karssemeijer, and J. A. W. M. Van Der Laak (Mar. 20, 2014). "Quantitative analysis of stain variability in histology slides and an algorithm for standardization". In: SPIE Medical Imaging. Ed. by M. N. Gurcan and A. Madabhushi. San Diego, California, USA.

Ekblom, P. (1989). "Developmentally regulated conversion of mesenchyme to epithelium". In: *The FASEB Journal* 3.10, pp. 2141–2150.

Elhamamsy, A. R. (July 2016). "DNA methylation dynamics in plants and mammals: overview of regulation and dysregulation: DNA Methylation Dynamics in Plants and Mammals". In: *Cell Biochemistry and Function* 34.5, pp. 289–298.

Ezugwu, A. E. et al. (Apr. 1, 2022). "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects". In: *Engineering Applications of Artificial Intelligence* 110, p. 104743.

Fattorini, P. et al. (Jan. 2020). "A Novel HPLC-Based Method to Investigate on RNA after Fixation". In: *International Journal of Molecular Sciences* 21.20, p. 7540.

Fauzi, M. F. A., W. Chen, D. Knight, H. Hampel, W. L. Frankel, and M. N. Gurcan (Feb. 2020). "Tumor Budding Detection System in Whole Slide Pathology Images". In: *Journal of Medical Systems* 44.2, p. 38.

Fearon, E. R. and B. Vogelstein (2023). "A genetic model for colorectal tumorigenesis". In: *Cell* 61.5 (), pp. 759–767. (Visited on 09/23/2023).

Fernández-Medarde, A. and E. Santos (Mar. 2011). "Ras in Cancer and Developmental Diseases". In: *Genes & Cancer* 2.3, pp. 344–358.

Fisher, N., M. Loughrey, H. Coleman, M. Gelbard, P. Bankhead, and P. Dunne (June 17, 2021). *Development of a semi-automated method for tumor budding assessment in colorectal cancer and comparison with manual methods*.

Fleming, M., S. Ravula, S. Tatishchev, and H. Wang (2012). "Colorectal carcinoma: Pathologic aspects". In: *Journal of Gastrointestinal Oncology* 3.3, p. 21.

Fotheringham, S., G. A. Mozolowski, E. M. A. Murray, and D. J. Kerr (June 1, 2019). "Challenges and solutions in patient treatment strategies for stage II colon cancer". In: *Gastroenterology Report* 7.3, pp. 151–161.

Fujitani, M. et al. (May 1, 2019). "Re-staining Pathology Images by FCNN". In: *International Conference on Machine Vision Applications*, p. 8757875.

Fukushima, K. (Apr. 1980). "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". In: *Biological Cybernetics* 36.4, pp. 193–202.

Gatenbee, C. et al. (Nov. 10, 2021). "VALIS: Virtual Alignment of pathoLogy Image Series". In: *bioRxiv*.

Giger, O., S. Comtesse, A. Lugli, I. Zlobec, and M. Kurrer (July 2012). “Intra-tumoral budding in preoperative biopsy specimens predicts lymph node and distant metastasis in patients with colorectal cancer”. In: *Modern Pathology* 25.7, pp. 1048–1053.

Gillies, S. et al. (2007). *Shapely: manipulation and analysis of geometric objects*. toblerity.org. url: https://github.com/Toblerity/Shapely.

Girshick, R. (Sept. 27, 2015). *Fast R-CNN*. arXiv: 1504.08083[cs].

Gohlke, C. (July 4, 2022). *cgohlke/tifffile: v2022.5.4*. url: https://zenodo.org/record/6795861.

Goode, A., B. Gilbert, J. Harkes, D. Jukic, and M. Satyanarayanan (Sept. 27, 2013). “OpenSlide: A vendor-neutral software foundation for digital pathology”. In: *Journal of Pathology Informatics* 4, p. 27. issn: 2229-5089.

Goodfellow, I. J. et al. (June 10, 2014). *Generative Adversarial Networks*. arXiv: 1406.2661[cs,stat].

Gorbunova, V., S. Durrleman, P. Lo, X. Pennec, and M. de Bruijne (Apr. 2010). “Lung CT registration combining intensity, curves and surfaces”. In: *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. 2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 340–343.

Graham, S. et al. (Nov. 13, 2019). *HoVer-Net: Simultaneous Segmentation and Classification of Nuclei in Multi-Tissue Histology Images*. arXiv: 1812.06499[cs].

Grant, R. R. C. et al. (2022). “Adjuvant chemotherapy is associated with improved overall survival in select patients with Stage II colon cancer: A National Cancer Database analysis”. In: *Journal of Surgical Oncology* 126.4, pp. 748–756.

Grigore, A., M. Jolly, D. Jia, M. Farach-Carson, and H. Levine (Apr. 29, 2016). “Tumor Budding: The Name is EMT. Partial EMT.” In: *Journal of Clinical Medicine* 5.5, p. 51.

Grigorev, I. and D. Korzhevskii (June 2018). “Current Technologies for Fixation of Biological Material for Immunohistochemical Analysis (Review)”. In: *Sovremennye tehnologii v medicine* 10.2, p. 156.

Gui, J., Z. Sun, Y. Wen, D. Tao, and J. Ye (Apr. 2023). “A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications”. In: *IEEE Transactions on Knowledge and Data Engineering* 35.4, pp. 3313–3332.

Guo, J. et al. (June 14, 2022). *CMT: Convolutional Neural Networks Meet Vision Transformers*. arXiv: 2107.06263[cs].

Gurcan, M., L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener (Oct. 30, 2009). “Histopathological Image Analysis: A Review”. In: *IEEE Reviews in Biomedical Engineering* 2.2, pp. 147–171.

Haan, K. de et al. (Aug. 12, 2021). “Deep learning-based transformation of H&E stained tissues into special stains”. In: *Nature Communications* 12, p. 4884.

Haddad, T. S. et al. (2023). “Pseudobudding: ruptured glands do not represent true tumor buds”. In: *The Journal of Pathology* 261.1, pp. 19–27.

Haddad, T. et al. (Sept. 2021). “Improving tumor budding reporting in colorectal cancer: a Delphi consensus study”. In: *Virchows Archiv* 479.3, pp. 459–469.

Haensel, D. and X. Dai (Mar. 2018). “Epithelial-to-mesenchymal transition in cutaneous wound healing: Where we are and where we are heading: EMT in Cutaneous Wound Healing”. In: *Developmental Dynamics* 247.3, pp. 473–480.

Hairol Anuar, S. H. et al. (Dec. 1, 2021). “Comparison between Louvain and Leiden Algorithm for Network Structure: A Review”. In: *Journal of Physics: Conference Series* 2129.1, p. 012028.

Hanahan, D. (Jan. 12, 2022). “Hallmarks of Cancer: New Dimensions”. In: *Cancer Discovery* 12.1, pp. 31–46.

Hanahan, D. and R. A. Weinberg (Jan. 2000). “The Hallmarks of Cancer”. In: *Cell* 100.1, pp. 57–70.

— (Mar. 2011). “Hallmarks of Cancer: The Next Generation”. In: *Cell* 144.5, pp. 646–674.

Hatthakarnkul, P. et al. (Nov. 2021). “Systematic review of tumour budding and association with common mutations in patients with colorectal cancer”. In: *Critical Reviews in Oncology/Hematology* 167, p. 103490.

Al-Hayani, W. and Z. Wang (Sept. 2017). “Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review”. In: *Neural Computation* 29.9, pp. 2352–2449.

Haykin, S. (2009). *Neural networks and learning machines*. 3rd ed. New York: Prentice Hall. 906 pp.

He, K., X. Zhang, S. Ren, and J. Sun (Dec. 10, 2015). *Deep Residual Learning for Image Recognition*. arXiv: 1512.03385[cs].

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. Mahwah, United States: Taylor & Francis Group.

Heinrich, M. P. et al. (Oct. 1, 2012). “MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration”. In: *Medical Image Analysis* 16.7, pp. 1423–1435.

Heusel, M., H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter (Jan. 12, 2018). *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium*. arXiv: 1706.08500[cs,stat].

Hollstein, M., D. Sidransky, B. Vogelstein, and C. C. Harris (July 5, 1991). “p53 Mutations in Human Cancers”. In: *Science* 253.5015, pp. 49–53.

Hong, Y. et al. (Sept. 28, 2021). “Deep learning-based virtual cytokeratin staining of gastric carcinomas to measure tumor–stroma ratio”. In: *Scientific Reports* 11, p. 19255.

Horai, Y. et al. (Oct. 2019). “Quantification of histopathological findings using a novel image analysis platform”. In: *Journal of Toxicologic Pathology* 32.4, pp. 319–327.

Horé, A. and D. Ziou (Aug. 2010). “Image Quality Metrics: PSNR vs. SSIM”. In: *2010 20th International Conference on Pattern Recognition*, pp. 2366–2369.

Huang, S. C., A. Pareek, M. Jensen, M. P. Lungren, S. Yeung, and A. S. Chaudhari (Apr. 26, 2023). “Self-supervised learning for medical image classification: a systematic review and implementation guidelines”. In: *npj Digital Medicine* 6.1, p. 74.

Hubel, D. H. and T. N. Wiesel (Jan. 1, 1962). “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex”. In: *The Journal of Physiology* 160.1, pp. 106–154.

Imai, T. (Sept. 1960). “Growth Patterns in Human Carcinoma: Their classification and relation to prognosis”. In: *Obstetrics & Gynecology* 16.3, pp. 296–308.

Isola, P., J. Y. Zhu, T. Zhou, and A. A. Efros (Nov. 26, 2018). *Image-to-Image Translation with Conditional Adversarial Networks*.

Jaiswal, A., A. R. Babu, M. Zadeh, D. Banerjee, and F. Makedon (Mar. 2021). “A Survey on Contrastive Self-Supervised Learning”. In: *Technologies* 9.1, p. 2.

Janowczyk, A., A. Basavanhally, and A. Madabhushi (Apr. 1, 2017). “Stain Normalization using Sparse AutoEncoders (StaNoSA): Application to digital pathology”. In: *Computerized Medical Imaging and Graphics*. Recent Developments in Machine Learning for Medical Imaging Applications 57, pp. 50–61. issn: 0895-6111.

Javaeed, A., S. Qamar, S. Ali, M. Mustafa, A. Nusrat, and S. Ghauri (Oct. 4, 2021). “Histological Stains in the Past, Present, and Future”. In: *Cureus* 13 (10), e18486.

Jepsen, R. K. et al. (Sept. 2018). “Digital image analysis of pan-cytokeratin stained tumor slides for evaluation of tumor budding in pT1/pT2 colorectal cancer: Results of a feasibility study”. In: *Pathology - Research and Practice* 214.9, pp. 1273–1281.

Jiang, D. et al. (June 25, 2020). “A machine learning-based prognostic predictor for stage III colon cancer”. In: *Scientific Reports* 10.1, p. 10333.

Jing, L. and Y. Tian (Feb. 16, 2019). *Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey*. arXiv: 1902.06162[cs].

Jose, L., S. Liu, C. Russo, A. Nadort, and A. Di Ieva (Jan. 1, 2021). “Generative Adversarial Networks in Digital Pathology and Histopathological Image Processing: A Review”. In: *Journal of Pathology Informatics* 12.1, p. 43.

Kai, K. et al. (Feb. 2016). “Cytokeratin immunohistochemistry improves interobserver variability between unskilled pathologists in the evaluation of tumor budding in T1 colorectal cancer: Interobserver variability in budding”. In: *Pathology International* 66.2, pp. 75–82.

Karamchandani, D. et al. (Jan. 2020). “Challenges with colorectal cancer staging: results of an international study”. In: *Modern Pathology* 33.1, pp. 153–163.

Karamitopoulou, E. et al. (Feb. 2013). “Proposal for a 10-high-power-fields scoring method for the assessment of tumor budding in colorectal cancer”. In: *Modern Pathology* 26.2, pp. 295–301.

Kather, J. N., N. Halama, and A. Marx (Apr. 7, 2018). *100,000 histological images of human colorectal cancer and healthy tissue*. Version v0.1. doi: 10.5281/zenodo.1214456.

Kather, J. et al. (Jan. 24, 2019). “Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study”. In: *PLOS Medicine* 16.1, e1002730.

Khan, F. A. (July 2011). “The Immortal Life of Henrietta Lacks”. In: *The Journal of IMA* 43.2, pp. 93–94.

Kim, K. E., Y. J. Lee, J. Y. Lee, W. K. Jeong, S. K. Baek, and S. U. Bae (June 30, 2022). “Minimally invasive treatments for early colorectal cancer: comparison of endoscopic resection and laparoscopic surgery”. In: *Korean Journal of Clinical Oncology* 18.1, pp. 47–55.

Kim, N. K., K. Sugihara, and J. T. Liang, eds. (2018). *Surgical Treatment of Colorectal Cancer*. Singapore: Springer Singapore.

Kingma, D. P. and J. Ba (Jan. 29, 2017). *Adam: A Method for Stochastic Optimization*.

Klein, S., M. Staring, K. Murphy, M. A. Viergever, and J. W. Pluim (Jan. 2010). “elastix: A Toolbox for Intensity-Based Medical Image Registration”. In: *IEEE Transactions on Medical Imaging* 29.1. 2787, pp. 196–205.

Koelzer, V. H., I. Zlobec, and A. Lugli (Jan. 2016). “Tumor budding in colorectal cancer—ready for diagnostic practice?” In: *Human Pathology* 47.1, pp. 4–19.

Koelzer, V. et al. (May 1, 2015). “Tumor budding in colorectal cancer revisited: results of a multicenter interobserver study”. In: *Virchows Archiv* 466.5, pp. 485–493.

Krause, J. et al. (Feb. 9, 2021). “Deep learning detects genetic alterations in cancer histology generated by adversarial networks”. In: *The Journal of Pathology*, p. 5638.

Krbal, L., J. Soukup, S. John, and V. Hanusova (Dec. 24, 2017). “Derivation and basic characterization of colorectal carcinoma primary cell lines”. In: *Biomedical Papers* 161.4, pp. 360–368.

Krenker, A., J. Bester, and A. Kos (Apr. 11, 2011). “An Introduction to Artificial Neural Networks”. In: *Artificial Neural Networks - Methodological Advances and Biomedical Applications*. Ed. by K. Suzuki. InTech.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger. Vol. 25. Curran Associates, Inc.

Kusuma Dewi, A., B. Purwanto, and Widjiati (June 14, 2023). “Introduction to Histopathology”. In: *Molecular Histopathology and Cytopathology*. Ed. by A. Kara, V. Gelen, and H. Kara. IntechOpen.

Lahiani, A., J. Gildenblat, I. Klaman, S. Albarqouni, N. Navab, and E. Klaiman (2019). “Virtualization of Tissue Staining in Digital Pathology Using an Unsupervised Deep Learning Approach”. In: *Digital Pathology*. Ed. by C. Reyes-Aldasoro, A. Janowczyk,

M. Veta, P. Bankhead, and K. Sirinukunwattana. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 47–55.

Lahiani, A., I. Klaman, N. Navab, S. Albarqouni, and E. Klaiman (Feb. 2021). "Seamless Virtual Whole Slide Image Synthesis and Validation Using Perceptual Embedding Consistency". In: *IEEE Journal of Biomedical and Health Informatics* 25.2, pp. 403–411.

Laleh, N. G., A. Echle, H. S. Muti, K. J. Hewitt, V. Schulz, and J. N. Kather (2021). "Deep learning for interpretable end-to-end survival (E-E Surv) prediction in gastrointestinal cancer histopathology". In: *Journal of Machine Learning Research* 156, pp. 1–13.

Langer, L., Y. Binenbaum, L. Gugel, M. Amit, Z. Gil, and S. Dekel (July 1, 2015). "Computer-aided diagnostics in digital pathology: automated evaluation of early-phase pancreatic cancer in mice". In: *International Journal of Computer Assisted Radiology and Surgery* 10.7, pp. 1043–1054.

Lea, D., S. Håland, H. R. Hagland, and K. Søreide (Oct. 1, 2014). "Accuracy of TNM staging in colorectal cancer: a review of current culprits, the modern role of morphology and stepping-stones for improvements in the molecular era". In: *Scandinavian Journal of Gastroenterology* 49, pp. 1153–1163.

LeCun, Y., Y. Bengio, and G. Hinton (May 28, 2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444. (Visited on 08/03/2021).

Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner (Nov. 1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

Lei, Y., R. L. J. Qiu, T. Wang, W. J. Curran, T. Liu, and X. Yang (Dec. 30, 2020). *Generative Adversarial Network for Image Synthesis*. arXiv: 2012.15446[physics].

Levy, J. J. et al. (Apr. 2021). "A large-scale internal validation study of unsupervised virtual trichrome staining technologies on nonalcoholic steatohepatitis liver biopsies". In: *Modern Pathology* 34.4, pp. 808–822.

Li, X., J. Jonnagaddala, M. Cen, H. Zhang, and S. Xu (Nov. 15, 2022). "Colorectal Cancer Survival Prediction Using Deep Distribution Based Multiple-Instance Learning". In: *Entropy* 24.11, p. 1669.

Li, Y. et al. (2020). "Automated Gleason Grading and Gleason Pattern Region Segmentation Based on Deep Learning for Pathological Images of Prostate Cancer". In: *IEEE Access* 8, pp. 117714–117725.

Lin, T., Y. Wang, X. Liu, and X. Qiu (June 15, 2021). *A Survey of Transformers*. arXiv: 2106.04554[cs].

Lino-Silva, L. S., R. A. Salcedo-Hernández, and A. Gamboa-Domínguez (2018). "Tumour budding in rectal cancer. A comprehensive review". In: *Współczesna Onkologia* 22.2, pp. 61–74.

Liu, S., C. Zhu, F. Xu, X. Jia, Z. Shi, and M. Jin (June 2022). "BCI: Breast Cancer Immunohistochemical Image Generation through Pyramid Pix2pix". In: *2022*

IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). New Orleans, LA, USA: IEEE, pp. 1814–1823.

Liu, S. et al. (Aug. 2021a). “Unpaired Stain Transfer Using Pathology-Consistent Constrained Generative Adversarial Networks”. In: *IEEE Transactions on Medical Imaging* 40.8, pp. 1977–1989.

Liu, Z., A. Alavi, M. Li, and X. Zhang (Jan. 2023). “Self-Supervised Contrastive Learning for Medical Time Series: A Systematic Review”. In: *Sensors* 23.9, p. 4221.

Liu, Z. et al. (Aug. 17, 2021b). *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. arXiv: 2103.14030[cs].

Lo, Y., I. Chung, S. Guo, M. Wen, and C. Juang (Jan. 1, 2021). “Cycle-consistent GAN-based stain translation of renal pathology images with glomerulus detection application”. In: *Applied Soft Computing* 98, p. 106822.

Loughrey, M. B., P. Quirke, and N. A. Shepherd (Sept. 2018). *Dataset for Histopathological Reporting of Colorectal Cancer*. rcpath.org. url: https://www.rcpath.org/static/c8b61ba0-ae3f-43f1-85ffd3ab9f17cfe6/G049-Dataset-for-histopathological-reporting-of-colorectal-cancer.pdf (visited on 07/27/2023).

Loughrey, M. et al. (Mar. 2022). “Dataset for Pathology Reporting of Colorectal Cancer: Recommendations From the International Collaboration on Cancer Reporting (ICCR)”. In: *Annals of Surgery* 275, e549–e561.

Lowekamp, B. C., D. T. Chen, L. Ibáñez, and D. Blezek (2013). “The Design of SimpleITK”. In: *Frontiers in Neuroinformatics* 7.

Lu, J. et al. (May 1, 2022). “Development and application of a detection platform for colorectal cancer tumor sprouting pathological characteristics based on artificial intelligence”. In: *Intelligent Medicine* 2.2, pp. 82–87.

Luan, F., S. Paris, E. Shechtman, and K. Bala (July 2017). “Deep Photo Style Transfer”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, pp. 6997–7005. isbn: 978-1-5386-0457-1.

Luchini, C. et al. (June 1, 2022). “Ki-67 assessment of pancreatic neuroendocrine neoplasms: Systematic review and meta-analysis of manual vs. digital pathology scoring”. In: *Modern Pathology* 35.6, pp. 712–720.

Lugli, A., E. Karamitopoulou, and I. Zlobec (May 22, 2012). “Tumour budding: a promising parameter in colorectal cancer”. In: *British Journal of Cancer* 106.11, pp. 1713–1717.

Lugli, A. et al. (Sept. 2017). “Recommendations for reporting tumor budding in colorectal cancer based on the International Tumor Budding Consensus Conference (ITBCC) 2016”. In: *Modern Pathology* 30.9, pp. 1299–1311.

Lyon, H., B. van Deurs, P. E. Høyer, P. Prentø, and M. Møller (1991). “Tissue Processing: V. Embedding”. In: *Theory and Strategy in Histochemistry: A Guide to the Selection and Understanding of Techniques*. Ed. by H. Lyon. Berlin, Heidelberg: Springer, pp. 197–205. (Visited on 07/17/2023).

Macenko, M. et al. (June 2009). “A method for normalizing histology slides for quantitative analysis”. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI). Boston, MA, USA: IEEE, pp. 1107–1110.

Magee, D. et al. (Jan. 2009). “Colour Normalisation in Digital Histopathology Images”. In: *Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*.

Maguire, A. (2014). “Controversies in the pathological assessment of colorectal cancer”. In: *World Journal of Gastroenterology* 20, pp. 9850–9861.

Malki, A., R. A. ElRuz, I. Gupta, A. Allouch, S. Vranic, and A. E. Al Moustafa (Dec. 24, 2020). “Molecular Mechanisms of Colon Cancer Progression and Metastasis: Recent Insights and Advancements”. In: *International Journal of Molecular Sciences* 22.1, p. 130.

Mallat, S. (Apr. 13, 2016). “Understanding deep convolutional networks”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065, p. 20150203.

Martin, J. (Aug. 2018). *Meeting Pathology Demand - Histopathology Workforce Census*. Royal College of Pathologists, p. 20. url: https://www.rcpath.org/static/952a934d-2ec3-48c9-a8e6e00fcdca700f/Meeting-Pathology-Demand-Histopathology-Workforce-Census-2018.pdf (visited on 07/23/2023).

Martinez Ciarpaglini, C. et al. (Oct. 2019). “Improving tumour budding evaluation in colon cancer by extending the assessment area in colectomy specimens”. In: *Histopathology* 75.4, pp. 517–525.

Mattes, D., D. R. Haynor, H. Vesselle, T. K. Lewellyn, and W. Eubank (July 3, 2001). “Nonrigid multimodality image registration”. In: Medical Imaging 2001. Ed. by M. Sonka and K. M. Hanson. San Diego, CA, pp. 1609–1620.

McCormick, M., X. Liu, J. Jomier, C. Marion, and L. Ibanez (2014). “ITK: enabling reproducible research and open science”. In: *Frontiers in Neuroinformatics* 8.

McCulloch, W. S. and W. Pitts (Dec. 1, 1943). “A logical calculus of the ideas immanent in nervous activity”. In: *The bulletin of mathematical biophysics* 5.4, pp. 115–133.

Mercan, E. et al. (Apr. 3, 2020). “Virtual Staining for Mitosis Detection in Breast Histopathology”. In: pp. 1770–1774.

Michelucci, U. (2019). *Advanced Applied Deep Learning: Convolutional Neural Networks and Object Detection*. Berkeley, CA: Apress.

Michielli, N. et al. (Jan. 1, 2022). “Stain normalization in digital pathology: Clinical multi-center evaluation of image quality”. In: *Journal of Pathology Informatics* 13, p. 100145.

Minsky, M. and S. Papert (Sept. 22, 2017). *Perceptrons: An Introduction to Computational Geometry*. The MIT Press.

Mitrovic, B., D. Schaeffer, R. H. Riddell, and R. Kirsch (Oct. 2012). “Tumor budding in colorectal carcinoma: time to take notice”. In: *Modern Pathology* 25.10, pp. 1315–1325.

Mitrovic, B. et al. (Sept. 2021). “Prognostic and Predictive Value of Tumor Budding in Colorectal Cancer”. In: *Clinical Colorectal Cancer* 20, pp. 256–264.

Miyato, T., T. Kataoka, M. Koyama, and Y. Yoshida (Feb. 16, 2018). *Spectral Normalization for Generative Adversarial Networks*. arXiv: 1802.05957[cs,stat].

Mobadersany, P. et al. (Mar. 27, 2018). “Predicting cancer outcomes from histology and genomics using convolutional networks”. In: *Proceedings of the National Academy of Sciences of the United States of America* 115.13, E2970–E2979.

Mohammed, F., T. F. Arishiya, and S. Mohamed (Dec. 31, 2012). “Microtomes and microtome knives”. In: *Annals of Dentistry University of Malaya* 19.2, pp. 43–50.

Mueller, D., D. L. J. Vossen, and B. Hulsken (Oct. 1, 2011). “Real-time deformable registration of multi-modal whole slides for digital pathology”. In: *Computerized Medical Imaging and Graphics* 35.7, pp. 542–556.

Musumeci, G. (2014). “Past, present and future: overview on histology and histopathology”. In: *Journal of Histology and Histopathology* 1.1, p. 5.

Nguyen, L. H., A. Goel, and D. C. Chung (Jan. 2020). “Pathways of Colorectal Carcinogenesis”. In: *Gastroenterology* 158.2, pp. 291–302.

Niehues, J. et al. (Apr. 2023). “Generalizable biomarker prediction from cancer pathology slides with self-supervised deep learning: A retrospective multi-centric study”. In: *Cell Reports Medicine* 4.4, p. 100980.

Niethammer, M., D. Borland, J. S. Marron, J. Woosley, and N. E. Thomas (2010). “Appearance Normalization of Histology Slides”. In: *Machine learning in medical imaging. MLMI (Workshop), author* 6357, pp. 58–66.

O’Dowd, G., S. Bell, and S. Wright (May 7, 2003). *Wheater’s Functional Histology*. 7th. Elsevier. 480 pp.

O’Shea, K. and R. Nash (Dec. 2, 2015). *An Introduction to Convolutional Neural Networks*. arXiv: 1511.08458[cs].

Olsen, T. G. et al. (Jan. 1, 2018). “Diagnostic Performance of Deep Learning Algorithms Applied to Three Common Diagnoses in Dermatopathology”. In: *Journal of Pathology Informatics* 9.1, p. 32.

Onder, D., S. Zengin, and S. Sarioglu (Nov. 2014). “A Review on Color Normalization and Color Deconvolution Methods in Histopathology”. In: *Applied Immunohistochemistry & Molecular Morphology* 22.10, pp. 713–719.

Pantanowitz, L., A. Sharma, A. B. Carter, T. Kurc, A. Sussman, and J. Saltz (Nov. 21, 2018). “Twenty Years of Digital Pathology: An Overview of the Road Travelled, What is on the Horizon, and the Emergence of Vendor-Neutral Archives”. In: *Journal of Pathology Informatics* 9, p. 40.

Park, J. H., H. Van Wyk, C. S. D. Roxburgh, P. G. Horgan, J. Edwards, and D. C. McMillan (May 2017). “Tumour invasiveness, the local and systemic environment and the basis of staging systems in colorectal cancer”. In: *British Journal of Cancer* 116.11, pp. 1444–1450.

Paszke, A. et al. (Dec. 3, 2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. arXiv: 1912.01703[cs,stat].

Patgiri, R., A. Biswas, and P. Roy, eds. (2021). *Health Informatics: A Computational Perspective in Healthcare*. Vol. 932. Studies in Computational Intelligence. Singapore: Springer Singapore.

Pavlič, A. et al. (2022a). “Tumour budding and poorly differentiated clusters in colon cancer – different manifestations of partial epithelial–mesenchymal transition”. In: *The Journal of Pathology* 258.3, pp. 278–288.

— (2022b). “Tumour budding and poorly differentiated clusters in colon cancer different manifestations of partial epithelial–mesenchymal transition”. In: *The Journal of Pathology* 258.3, pp. 278–288.

Petrelli, F. et al. (Sept. 2015). “Tumour Budding and Survival in Stage II Colorectal Cancer: a Systematic Review and Pooled Analysis”. In: *Journal of Gastrointestinal Cancer* 46.3, pp. 212–218.

Pichat, J., J. Iglesias, T. Yousry, S. Ourselin, and M. Modat (May 1, 2018). “A Survey of Methods for 3D Histology Reconstruction”. In: *Medical Image Analysis* 46, pp. 73–105.

Pluim, J., J. Maintz, and M. Viergever (Aug. 2003). “Mutual-information-based registration of medical images: a survey”. In: *IEEE Transactions on Medical Imaging* 22.8. 2024, pp. 986–1004.

Pocock, J. et al. (Sept. 24, 2022). “TIAToolbox as an end-to-end library for advanced tissue image analytics”. In: *Communications Medicine* 2.1, pp. 1–14. doi: 10.1038/s43856-022-00186-5.

Poon, H., H. Wang, and H. Lang (July 27, 2021). *Combining Probabilistic Logic and Deep Learning for Self-Supervised Learning*. arXiv: 2107.12591[cs].

Prall, F., H. Nizze, and M. Barten (July 2005). “Tumour budding as prognostic factor in stage I/II colorectal carcinoma”. In: *Histopathology* 47.1, pp. 17–24.

Priddy, K. L. and P. E. Keller (2005). *Artificial Neural Networks: An Introduction*. SPIE Press. 184 pp.

Puppa, G. et al. (2012). “Diagnostic reproducibility of tumour budding in colorectal cancer: a multicentre, multinational study using virtual microscopy”. In: *Histopathology* 61.4, pp. 562–575.

Puppa, G., A. Sonzogni, R. Colombari, and G. Pelosi (June 1, 2010). “TNM Staging System of Colorectal Carcinoma: A Critical Appraisal of Challenging Issues”. In: *Archives of Pathology & Laboratory Medicine* 134, pp. 837–852.

Radford, A., L. Metz, and S. Chintala (Jan. 7, 2016). *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. arXiv: 1511.06434[cs].

Ramesh, R., D. Dsouza, Y. S, and M. M. S (Mar. 2018). “Colorectal Cancer: A Review of Disease Diagnosis, Surgical Intervention and Treatment Procedures”. In: *Indian Research Journal of Pharmacy and Science* 5.1, pp. 1260–1279.

Rawla, P., T. Sunkara, and A. Barsouk (2019). “Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors”. In: *Gastroenterology Review* 14.2, pp. 89–103.

Redfern, A., L. Spalding, and E. Thompson (Apr. 2018). “The Kraken Wakes: induced EMT as a driver of tumour aggression and poor outcome”. In: *Clinical & Experimental Metastasis* 35.4, pp. 285–308.

Reinhard, E., M. Ashikhmin, B. Gooch, and P. Shirley (2001). “Color Transfer between Images”. In: *IEEE Computer Graphics and Applications*, p. 8.

Rogers, A. C. et al. (Jan. 2014). “Prognostic significance of tumor budding in rectal cancer biopsies before neoadjuvant therapy”. In: *Modern Pathology* 27.1, pp. 156–162.

Rogers, A. C. et al. (Sept. 2016). “Systematic review and meta-analysis of the impact of tumour budding in colorectal cancer”. In: *British Journal of Cancer* 115.7, pp. 831–840.

Ronneberger, O., P. Fischer, and T. Brox (May 18, 2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. arXiv: `1505.04597[cs]`.

Rosenblatt, F. (1958). “The perceptron: A probabilistic model for information storage and organization in the brain.” In: *Psychological Review* 65.6, pp. 386–408.

Roy, S., A. kumar Jain, S. Lal, and J. Kini (Nov. 1, 2018). “A study about color normalization methods for histopathology images”. In: *Micron* 114, pp. 42–61.

Rua, T. et al. (Sept. 1, 2020). “An observational study to compare the utilisation of computed tomography colonography with optical colonoscopy as the first diagnostic imaging tool in patients with suspected colorectal cancer”. In: *Clinical Radiology* 75.9, 712.e23–712.e31.

Ruderman, D. L., T. W. Cronin, and C. C. Chiao (Aug. 1, 1998). “Statistics of cone responses to natural images: implications for visual coding”. In: *Journal of the Optical Society of America A* 15.8, p. 2036.

Rueckert, D., L. Sonoda, C. Hayes, D. Hill, M. Leach, and D. Hawkes (Aug. 1999). “Nonrigid registration using free-form deformations: application to breast MR images”. In: *IEEE Transactions on Medical Imaging* 18.8, pp. 712–721.

Samadder, N. J., N. Baffy, K. V. Giridhar, F. J. Couch, and D. Riegert-Johnson (June 1, 2019). “Hereditary Cancer Syndromes—A Primer on Diagnosis and Management, Part 2: Gastrointestinal Cancer Syndromes”. In: *Mayo Clinic Proceedings* 94.6, pp. 1099–1116.

Samarasinghe, S. (2006). *Neural Networks for Applied Sciences and Engineering*. USA: Auerbach Publications.

Sampat, M. P., Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey (Nov. 2009). “Complex Wavelet Structural Similarity: A New Image Similarity Index”. In: *IEEE Transactions on Image Processing* 18.11, pp. 2385–2401.

Sandler, M., A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen (Mar. 21, 2019). *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. arXiv: 1801.04381[cs].

Sarvamangala, D. R. and R. V. Kulkarni (Mar. 2022). “Convolutional neural networks in medical image understanding: a survey”. In: *Evolutionary Intelligence* 15.1, pp. 1–22.

Schacht, V. and J. Kern (Mar. 2015). “Basics of Immunohistochemistry”. In: *Journal of Investigative Dermatology* 135.3, pp. 1–4.

Shaban, M. T., C. Baur, N. Navab, and S. Albarqouni (Apr. 4, 2018). *StainGAN: Stain Style Transfer for Digital Histological Images*.

Sikorsky, I. and J. Xu (Aug. 8, 2021). “A Review of Self-supervised Learning Methods in the Field of Medical Image Analysis”. In: *International Journal of Image, Graphics and Signal Processing* 13.4, pp. 33–46. issn: 20749074, 20749082.

Simonyan, K. and A. Zisserman (Apr. 10, 2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv: 1409.1556[cs].

Singh, S. and S. Srivastava (Jan. 1, 2020). “Review of Clustering Techniques in Control System: Review of Clustering Techniques in Control System”. In: *Procedia Computer Science*. International Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020 173, pp. 272–280.

Skrede, O.-J. et al. (Feb. 2020). “Deep learning for prediction of colorectal cancer outcome: a discovery and validation study”. In: *The Lancet* 395.10221, pp. 350–360.

Slaoui, M. and L. Fiette (Jan. 1, 2011). “Histopathology procedures: from tissue sampling to histopathological evaluation.” In: *Methods of Molecular Biology* 691, pp. 69–82.

Søreide, K., M. M. Watson, and H. R. Hagland (Apr. 1, 2016). “Deciphering the Molecular Code to Colorectal Liver Metastasis Biology Through Microsatellite Alterations and Allelic Loss: The Good, the Bad, and the Ugly”. In: *Gastroenterology* 150.4, pp. 811–814.

Spathis, D., I. Perez-Pozuelo, L. Marques-Fernandez, and C. Mascolo (Feb. 11, 2022). “Breaking away from labels: The promise of self-supervised machine learning in intelligent health”. In: *Patterns* 3.2, p. 100410.

Stathonikos, N., T. Q. Nguyen, C. P. Spoto, M. A. M. Verdaasdonk, and P. J. van Diest (2019). “Being fully digital: perspective of a Dutch academic pathology laboratory”. In: *Histopathology* 75.5, pp. 621–635.

Studer, L. et al. (2021). “Taking tumour budding to the next frontier — a post International Tumour Budding Consensus Conference (ITBCC) 2016 review”. In: *Histopathology* 78.4, pp. 476–484.

Sun, G., X. Dong, X. Tang, H. Qu, H. Zhang, and E. Zhao (2019). “The prognostic value of immunoscore in patients with colorectal cancer: A systematic review and meta-analysis”. In: *Cancer Medicine* 8.1, pp. 182–189.

Swinehart, D. F. (July 1, 1962). “The Beer-Lambert Law”. In: *Journal of Chemical Education* 39.7, p. 333.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna (Dec. 11, 2015). *Rethinking the Inception Architecture for Computer Vision*. arXiv: 1512.00567[cs].

Takamatsu, M. et al. (May 2019). “Immunohistochemical evaluation of tumor budding for stratifying T1 colorectal cancer: optimal cut-off value and a novel computer-assisted semiautomatic method”. In: *Modern Pathology* 32.5, pp. 675–683.

Talbot, A. and D. J. Gallagher (Feb. 2018). “Contribution of Lynch syndrome to early onset malignancy in Ireland.” In: *Journal of Clinical Oncology* 36.4, pp. 586–586.

Tamang, L. and B. Kim (Nov. 19, 2021). “Deep Learning Approaches to Colorectal Cancer Diagnosis: A Review”. In: *Applied Sciences* 11.22, p. 10982.

Tanaka, S., N. Asayama, K. Shigita, N. Hayashi, S. Oka, and K. Chayama (2015). “Towards safer and appropriate application of endoscopic submucosal dissection for T1 colorectal carcinoma as total excisional biopsy: Future perspectives”. In: *Digestive Endoscopy* 27.2, pp. 216–222.

Taqi, S., S. Sami, L. Sami, and S. Zaki (2018). “A review of artifacts in histopathology”. In: *Journal of Oral and Maxillofacial Pathology : JOMFP* 22.2, p. 279.

Tavolara, T. E. et al. (Apr. 4, 2022). “Automatic generation of the ground truth for tumor budding using H&E stained slides”. In: *Medical Imaging 2022: Digital and Computational Pathology*. Digital and Computational Pathology. Ed. by R. M. Levenson, J. E. Tomaszewski, and A. D. Ward. San Diego, United States: SPIE, p. 11. (Visited on 07/03/2023).

Teoh, T. T. (2023). *Convolutional Neural Networks for Medical Applications*. SpringerBriefs in Computer Science. Singapore: Springer Nature Singapore.

Testa, U., E. Pelosi, and G. Castelli (June 2018). “Colorectal Cancer: Genetic Abnormalities, Tumor Progression, Tumor Heterogeneity, Clonal Evolution and Tumor-Initiating Cells”. In: *Medical Sciences* 6.2, p. 31.

Thajudeen, A., R. Begum, S. Nivetha, and J. Jayapriya (2023). “Efficacy of Kumkum as a Surrogate for Eosin in Routine Histological Sections: An Observational Study”. In: *Journal of Clinical and Diagnostic Research*.

Thiam, P., H. Hihn, D. A. Braun, H. A. Kestler, and F. Schwenker (2021). “Multi-Modal Pain Intensity Assessment Based on Physiological Signals: A Deep Learning Perspective”. In: *Frontiers in Physiology* 12.

Thorlacius, H., Y. Takeuchi, T. Kanesaka, O. Ljungberg, N. Uedo, and E. Toth (July 3, 2017). “Serrated polyps – a concealed but prevalent precursor of colorectal cancer”. In: *Scandinavian Journal of Gastroenterology* 52.6, pp. 654–661.

Titford, M. (June 1, 2006). “A Short History of Histopathology Technique”. In: *Journal of Histotechnology* 29.2, pp. 99–110.

Touvron, H., M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou (Apr. 7, 2021). *Going deeper with Image Transformers*. (Visited on 09/16/2023).

Traag, V. A., L. Waltman, and N. J. van Eck (Mar. 26, 2019). "From Louvain to Leiden: guaranteeing well-connected communities". In: *Scientific Reports* 9.1, p. 5233.

Tsai, P.-C. et al. (Apr. 13, 2023). "Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients". In: *Nature Communications* 14, p. 2102.

Ueno, H. et al. (Aug. 2004). "Risk factors for an adverse outcome in early invasive colorectal carcinoma". In: *Gastroenterology* 127.2, pp. 385–394.

Ueno, H. et al. (Feb. 2012). "New Criteria for Histologic Grading of Colorectal Cancer". In: *American Journal of Surgical Pathology* 36.2, pp. 193–201.

Vahadane, A. et al. (Aug. 2016). "Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images". In: *IEEE Transactions on Medical Imaging* 35.8, pp. 1962–1971.

Van Der Lem, T., M. De Bakker, G. Keuck, and M. Richardson (Nov. 2021). "Wilhelm His Sr. and the development of paraffin embedding". In: *Der Pathologe* 42 (S1), pp. 55–61.

Van Wyk, H., J. Park, C. Roxburgh, P. Horgan, A. Foulis, and D. McMillan (Feb. 2015). "The role of tumour budding in predicting survival in patients with primary operable colorectal cancer: A systematic review". In: *Cancer Treatment Reviews* 41.2, pp. 151–159.

Van Wyk, H. et al. (Dec. 2019). "The Relationship Between Tumor Budding, Tumor Microenvironment, and Survival in Patients with Primary Operable Colorectal Cancer". In: *Annals of Surgical Oncology* 26.13, pp. 4397–4404.

Varone, V., C. Bellevicine, and G. Troncone (Jan. 1, 2012). "1.22 - Biopsic Sampling (Cancer)". In: *Comprehensive Sampling and Sample Preparation*. Ed. by J. Pawliszyn. Oxford: Academic Press, pp. 413–439.

Vasiljević, J., F. Feuerhake, C. Wemmert, and T. Lampert (Dec. 22, 2020). *Towards Histopathological Stain Invariance by Unsupervised Domain Augmentation using Generative Adversarial Networks*. arXiv: `2012.12413[cs,eess]`.

Vaswani, A. et al. (Aug. 1, 2017). *Attention Is All You Need*. arXiv: `1706.03762[cs]`.

Vodovnik, A. (Jan. 1, 2016). "Diagnostic time in digital pathology: A comparative study on 400 cases". In: *Journal of Pathology Informatics* 7.1, p. 4.

Walsh, E. and M. Halushka (June 6, 2023). "A Comparison of Tissue Dissection Techniques for Diagnostic, Prognostic, and Theragnostic Analysis of Human Disease". In: *Pathobiology* 90.3, pp. 199–208.

Wang, C. W., Y. C. Lee, M. A. Khalil, K. Y. Lin, C. P. Yu, and H. C. Lien (July 8, 2022). "Fast cross-staining alignment of gigapixel whole slide images with application to prostate cancer and breast cancer analysis". In: *Scientific Reports* 12.1.

Wang, H. and B. Raj (Mar. 2, 2017). *On the Origin of Deep Learning*.

Wang, L. M. et al. (Jan. 2009). “Tumor Budding is a Strong and Reproducible Prognostic Marker in T3N0 Colorectal Cancer”. In: *American Journal of Surgical Pathology* 33.1, pp. 134–141.

Wang, W., E. Ahn, D. Feng, and J. Kim (Aug. 1, 2023). “A Review of Predictive and Contrastive Self-supervised Learning for Medical Images”. In: *Machine Intelligence Research* 20.4, pp. 483–513.

Wang, Z., A. Bovik, H. Sheikh, and E. Simoncelli (Apr. 2004). “Image quality assessment: from error visibility to structural similarity”. In: *IEEE Transactions on Image Processing* 13.4. Conference Name: IEEE Transactions on Image Processing, pp. 600–612.

Washington, M. et al. (June 2017). *Protocol for the Examination of Specimens From Patients With Primary Carcinoma of the Colon and Rectum*. College of American Pathologists. url: https://documents.cap.org/protocols/cp-gilower-colonrectum-17protocol-4010.pdf (visited on 07/27/2023).

Wegmann, M., D. Zipperling, J. Hillenbrand, and J. Fleischer (June 24, 2021). *A review of systematic selection of clustering algorithms and their evaluation*. arXiv: 2106.12792[cs].

Weis, C. A. et al. (Dec. 2018). “Automatic evaluation of tumor budding in immunohistochemically stained colorectal carcinomas and correlation to clinical outcome”. In: *Diagnostic Pathology* 13.1, p. 64.

Weiser, M. (June 2018). “AJCC 8th Edition: Colorectal Cancer”. In: *Annals of Surgical Oncology* 25, pp. 1454–1455.

Wu, W. et al. (Feb. 21, 2022). “The value of tumor deposits in evaluating colorectal cancer survival and metastasis: a population-based retrospective cohort study”. In: *World Journal of Surgical Oncology* 20.1, p. 41.

Wulczyn, E. et al. (Apr. 19, 2021). “Interpretable Survival Prediction for Colorectal Cancer using Deep Learning”. In: *npj Digital Medicine* 4.1, p. 71.

Wulczyn, E. et al. (June 17, 2020). “Deep learning-based survival prediction for multiple cancer types using histopathology images”. In: *PLOS ONE* 15.6, e0233678.

Xie, W. et al. (Jan. 15, 2022). “Prostate Cancer Risk Stratification via Nondestructive 3D Pathology with Deep Learning–Assisted Gland Analysis”. In: *Cancer Research* 82.2, pp. 334–345.

Xu, Z., C. Moro, B. Bozóky, and Q. Zhang (Jan. 13, 2019). *GAN-based Virtual Re-Staining: A Promising Solution for Whole Slide Image Analysis*.

Yamada, N. et al. (Feb. 1, 2017). “Tumor budding at the invasive front of colorectal cancer may not be associated with the epithelial-mesenchymal transition”. In: *Human Pathology* 60, pp. 151–159.

Yamadera, M. et al. (June 2019). “Differential clinical impacts of tumour budding evaluated by the use of immunohistochemical and haematoxylin and eosin staining in stage II colorectal cancer”. In: *Histopathology* 74.7, pp. 1005–1013.

Yi, X., E. Walia, and P. Babyn (Dec. 2019). “Generative adversarial network in medical imaging: A review”. In: *Medical Image Analysis* 58, p. 101552.

Yu, G. et al. (Nov. 2, 2021). “Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images”. In: *Nature Communications* 12.1, p. 6311.

Zha, S., T. Li, Q. Zheng, and L. Li (2022). “Whether Patients With Stage / Colorectal Cancer Benefit From Adjuvant Chemotherapy: A Modeling Analysis of Literature Aggregate Data”. In: *Frontiers in Pharmacology* 13.

Zhang, R. et al. (Aug. 2022). “MVFStain: Multiple virtual functional stain histopathology images generation based on specific domain mapping”. In: *Medical Image Analysis* 80, p. 102520.

Zhu, J. Y., T. Park, P. Isola, and A. A. Efros (Aug. 24, 2020). *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*. arXiv: `1703.10593[cs]`.

Zhu, M., S. Gong, Z. Qian, and L. Zhang (2019). “A Brief Review on Cycle Generative Adversarial Networks”. In: *Proceedings of the 7th IIAE International Conference on Intelligent Systems and Image Processings 2019*. 7th IIAE International Conference on Intelligent Systems and Image Processings 2019. The Institute of Industrial Applications Engineers, Japan.

Zingman, I., S. Frayle, I. Tankoyeu, S. Sukhanov, and F. Heinemann (Apr. 6, 2023). *A comparative evaluation of image-to-image translation methods for stain transfer in histopathology*. arXiv: `2303.17009[cs,eess]`.

Zitová, B. and J. Flusser (Oct. 2003). “Image registration methods: a survey”. In: *Image and Vision Computing* 21.11, pp. 977–1000.

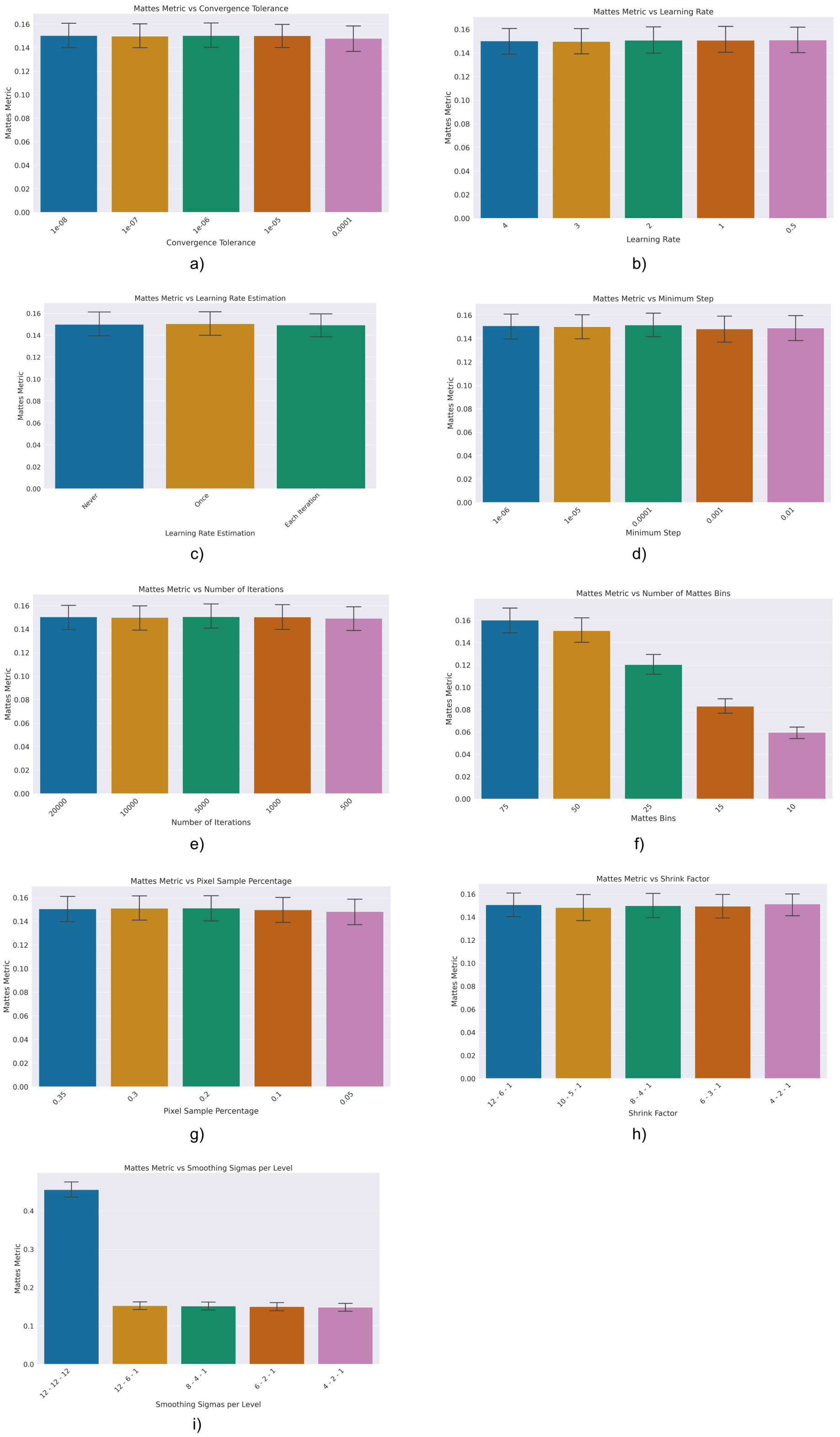
Zlobec, I. and A. Lugli (Nov. 30, 2010). “Epithelial mesenchymal transition and tumor budding in aggressive colorectal cancer: Tumor budding as oncotarget”. In: *Oncotarget* 1.7, pp. 651–661.

Zlobec, I. et al. (Feb. 2014). “Intratumoural budding (ITB) in preoperative biopsies predicts the presence of lymph node and distant metastases in colon and rectal cancer patients”. In: *British Journal of Cancer* 110.4, pp. 1008–1013.

Appendix A

Chapter 3 - Ensuring Accurate Virtual Staining

A.1 Image Registration



a) b)

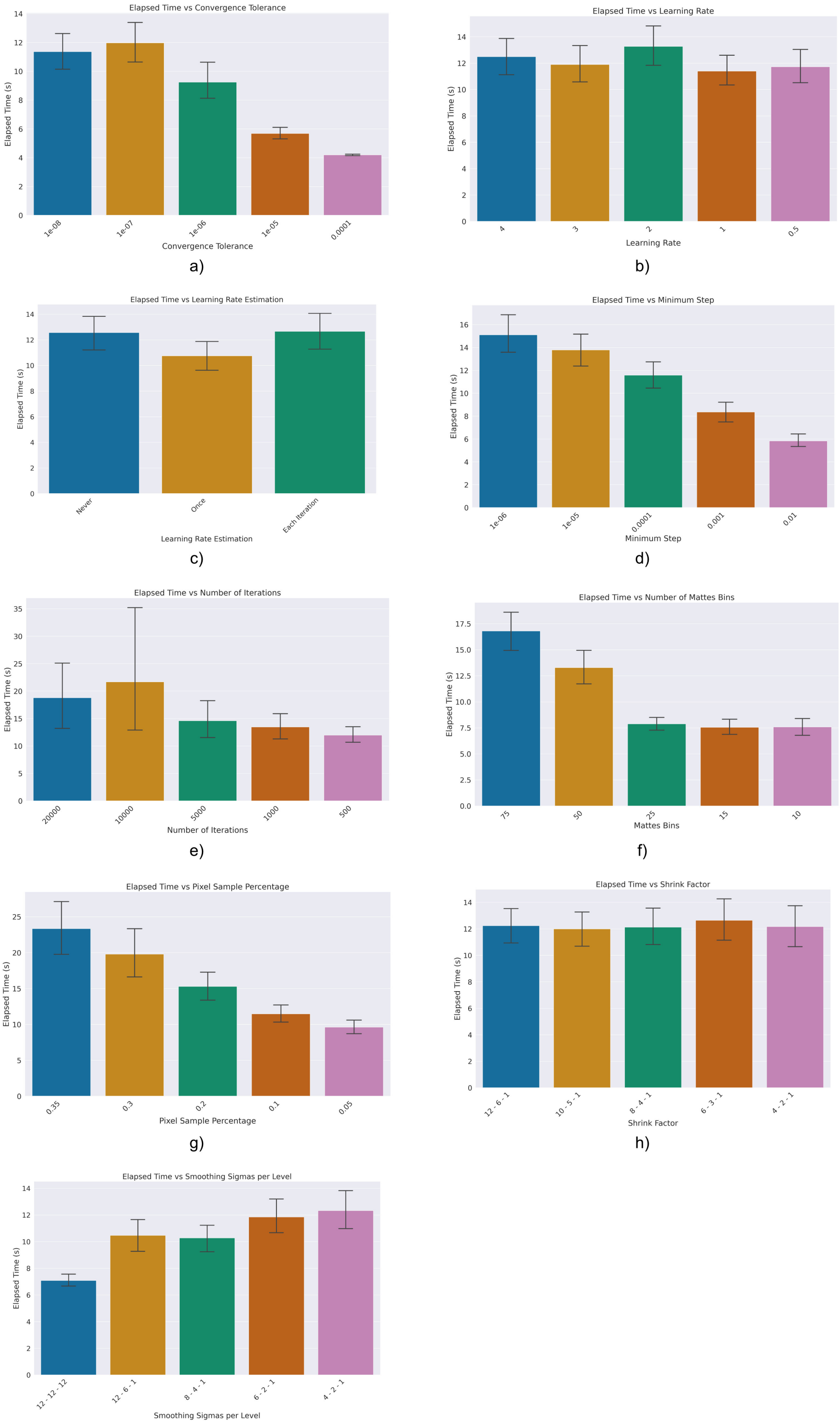
c) d)

e) f)

g) h)

i)

Figure A.1: The results of the parameter sweep for the rigid-only alignment. The y-axis of all graphs displays the Mattes metric. The x-axis displays the specified parameters of the alignment algorithm. The following panels display the effects of their specified parameter: a) convergence tolerance, b) learning rate, c) learning rate estimation, d) minimum step, e) number of iterations, f) number of Mattes bins, g) number of pixel samples as a percentage of the image, h) the configured shrink factors, and i) the configured smoothing sigmas per level.



a)

b)

c)

d)

e)

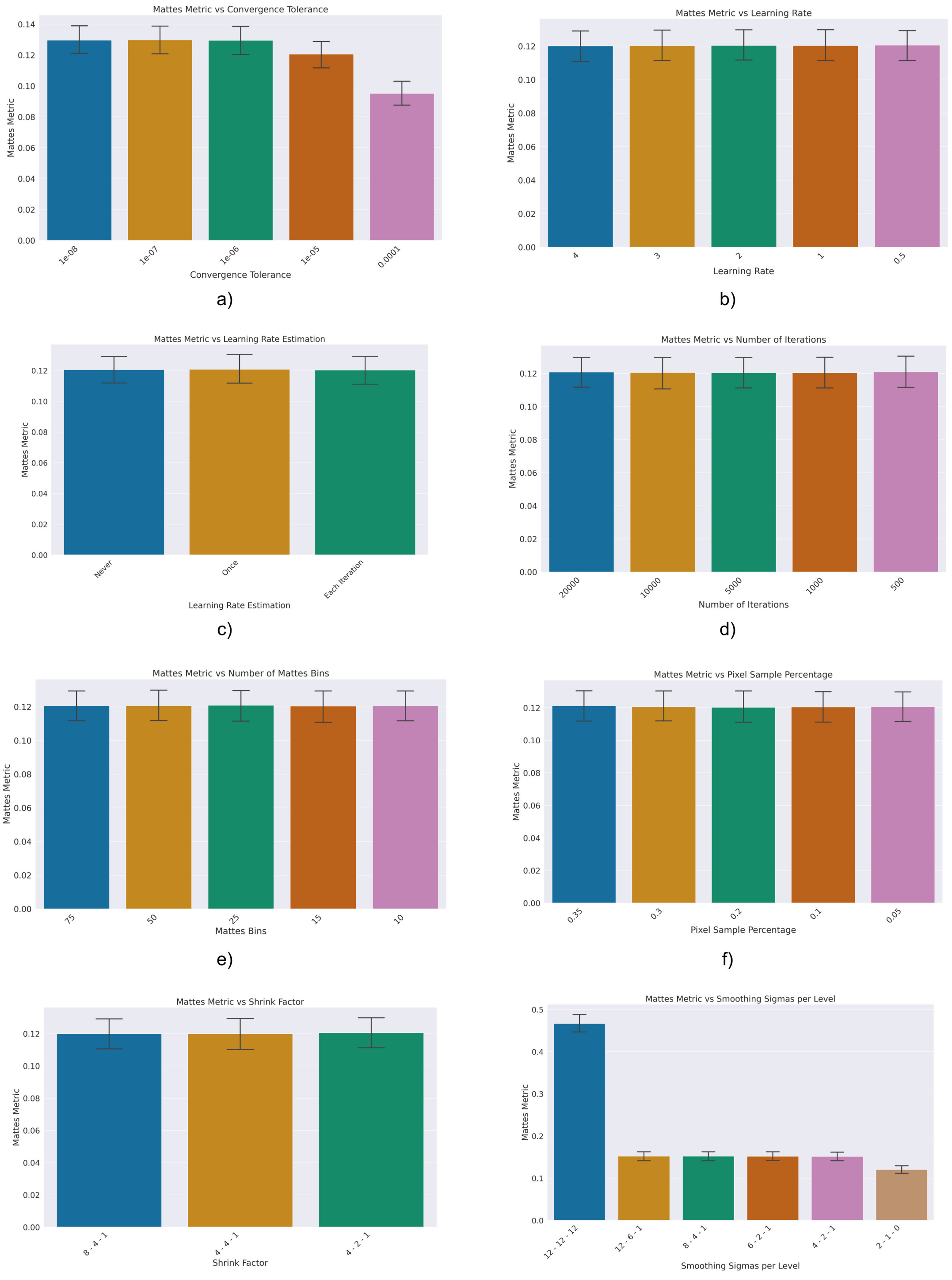
f)

g)

h)

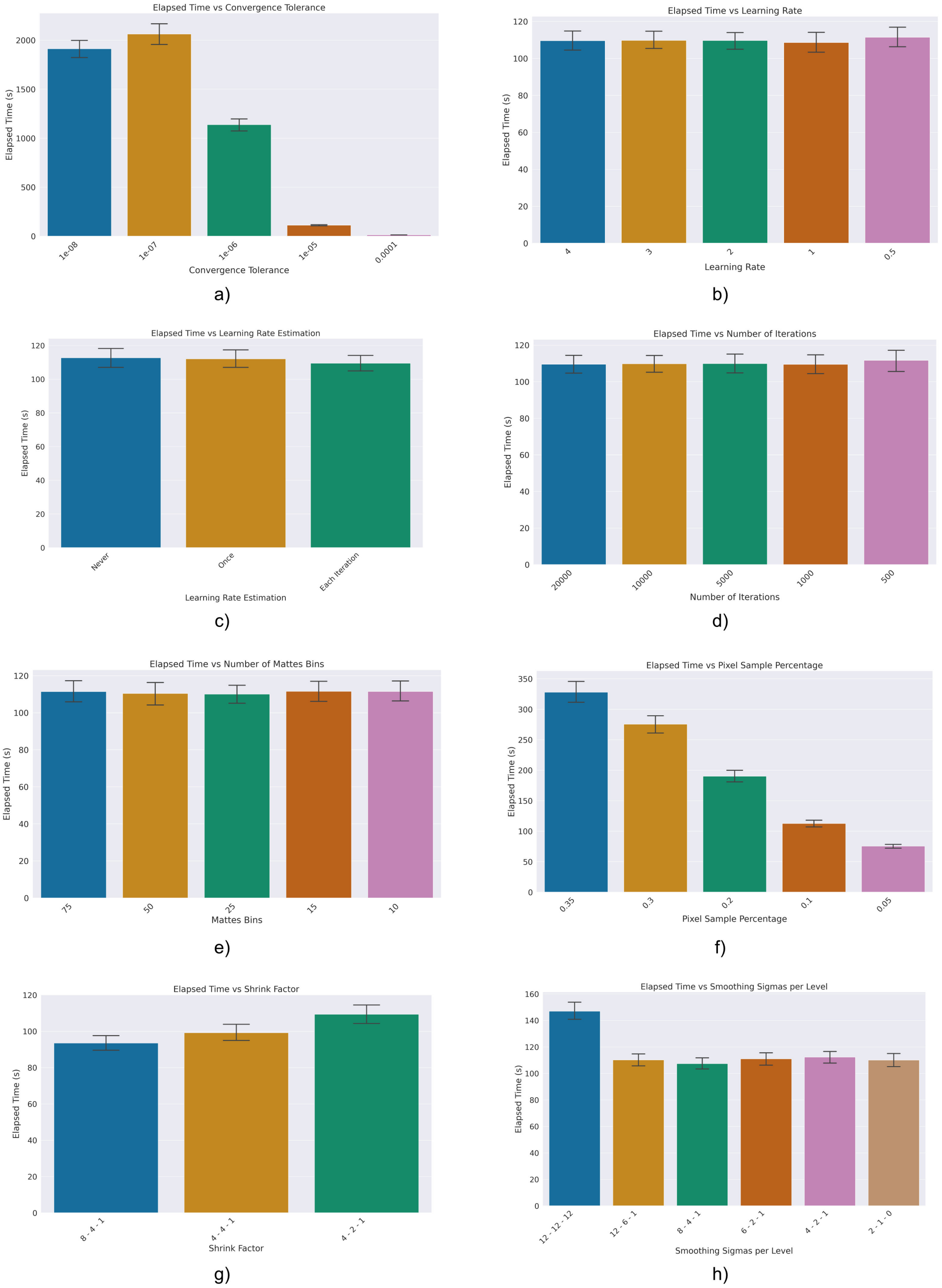
i)

Figure A.2: The results of the parameter sweep for the rigid-only alignment. The y-axis of all graphs displays the average time taken to align a tile in seconds. The x-axis displays the specified parameters of the alignment algorithm. The following panels display the effects of their specified parameter: a) convergence tolerance, b) learning rate, c) learning rate estimation, d) minimum step, e) number of iterations, f) number of Mattes bins, g) number of pixel samples as a percentage of the image, h) the configured shrink factors, and i) the configured smoothing sigmas per level.



a) b) c) d) e) f) g) h)

Figure A.3: The results of the parameter sweep for the B-Spline alignment. The y-axis of all graphs displays the Mattes metric. The x-axis displays the specified parameters of the alignment algorithm. The following panels display the effects of their specified parameter: a) convergence tolerance, b) learning rate, c) learning rate estimation, d) number of iterations, e) number of Mattes bins, f) number of pixel samples as a percentage of the image, g) the configured shrink factors, and h) the configured smoothing sigmas per level.



a)

b)

c)

d)

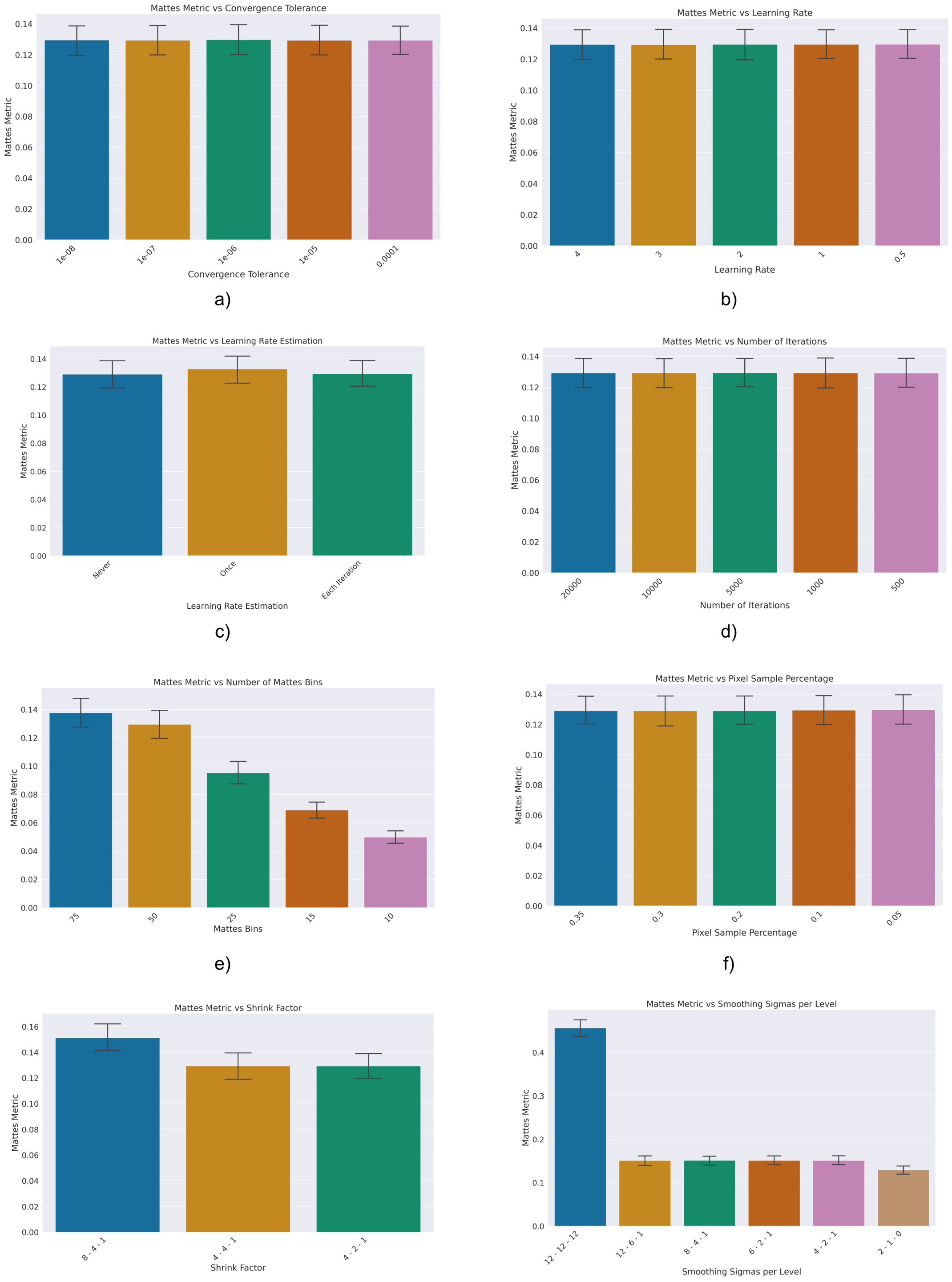
e)

f)

g)

h)

Figure A.4: The results of the parameter sweep for the B-Spline alignment. The y-axis of all graphs represents the average time taken to align a tile in seconds. The x-axis displays the specified parameters of the alignment algorithm. The following panels display the effects of their specified parameter: a) convergence tolerance, b) learning rate, c) learning rate estimation, d) number of iterations, e) number of Mattes bins, f) number of pixel samples as a percentage of the image, g) the configured shrink factors, and h) the configured smoothing sigmas per level.



a)

b)

c)

d)

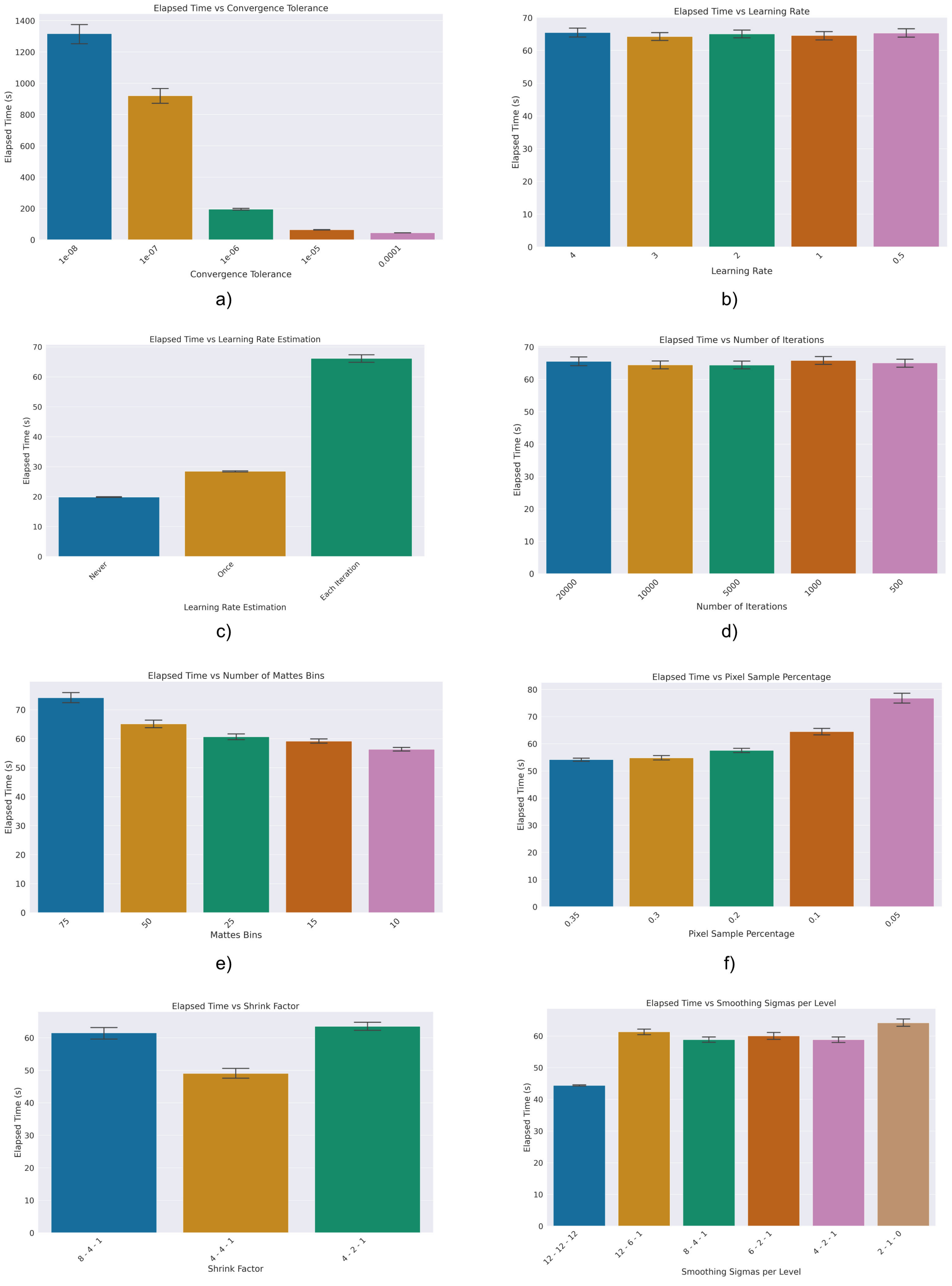
e)

f)

g)

h)

Figure A.5: The results of the parameter sweep for the Displacment-Field alignment. The y-axis of all graphs displays the Mattes metric. The x-axis displays the specified parameters of the alignment algorithm. The following panels display the effects of their specified parameter: a) convergence tolerance, b) learning rate, c) learning rate estimation, d) number of iterations, e) number of Mattes bins, f) number of pixel samples as a percentage of the image, g) the configured shrink factors, and h) the configured smoothing sigmas per level.



a)

b)

c)

d)

e)

f)

g)

h)

Figure A.6: The results of the parameter sweep for the Displacment-Field alignment. The y-axis of all graphs displays the average time taken to align a tile in seconds. The x-axis displays the specified parameters of the alignment algorithm. The following panels display the effects of their specified parameter: a) convergence tolerance, b) learning rate, c) learning rate estimation, d) number of iterations, e) number of Mattes bins, f) number of pixel samples as a percentage of the image, g) the configured shrink factors, and h) the configured smoothing sigmas per level.

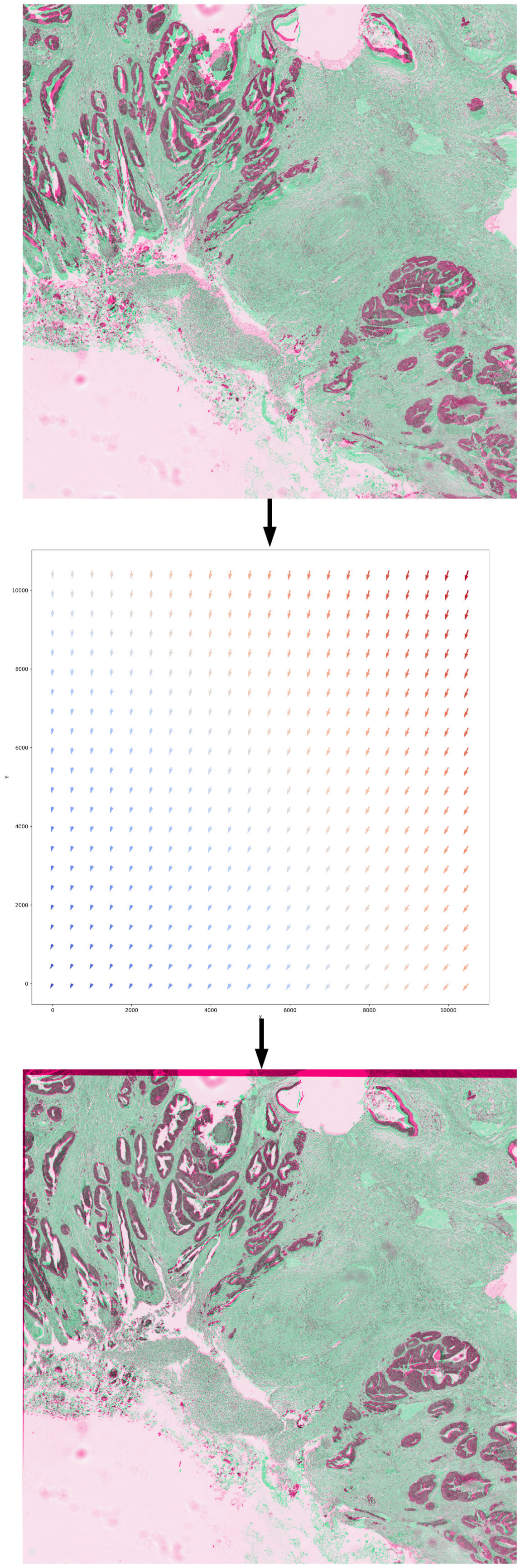


Figure A.7: An example of deformable high-level tile-based registration and alignment, using a displacement-field transform initialised by a rigid and affine transform with rotation, scaling, skewing and translation.

A.2 Virtual Staining

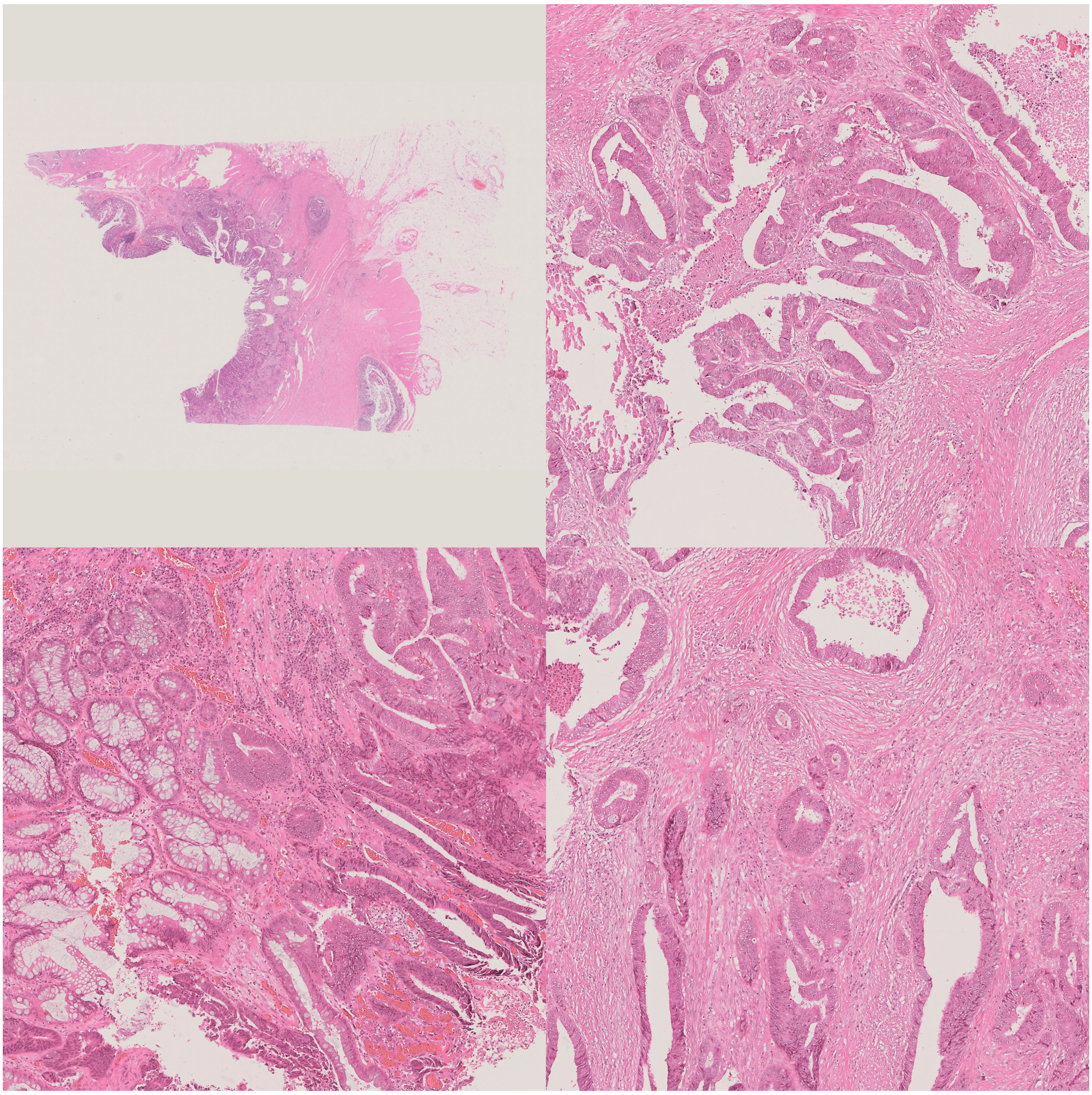


Figure A.8: A real haematoxylin and eosin whole slide image from the test dataset used as input for virtual stain translation in the UNet, Pix2Pix, CycleGAN and our proposed method. The WSI is displayed in the top left quadrant, and three magnified areas from the test dataset occupy the remaining three.

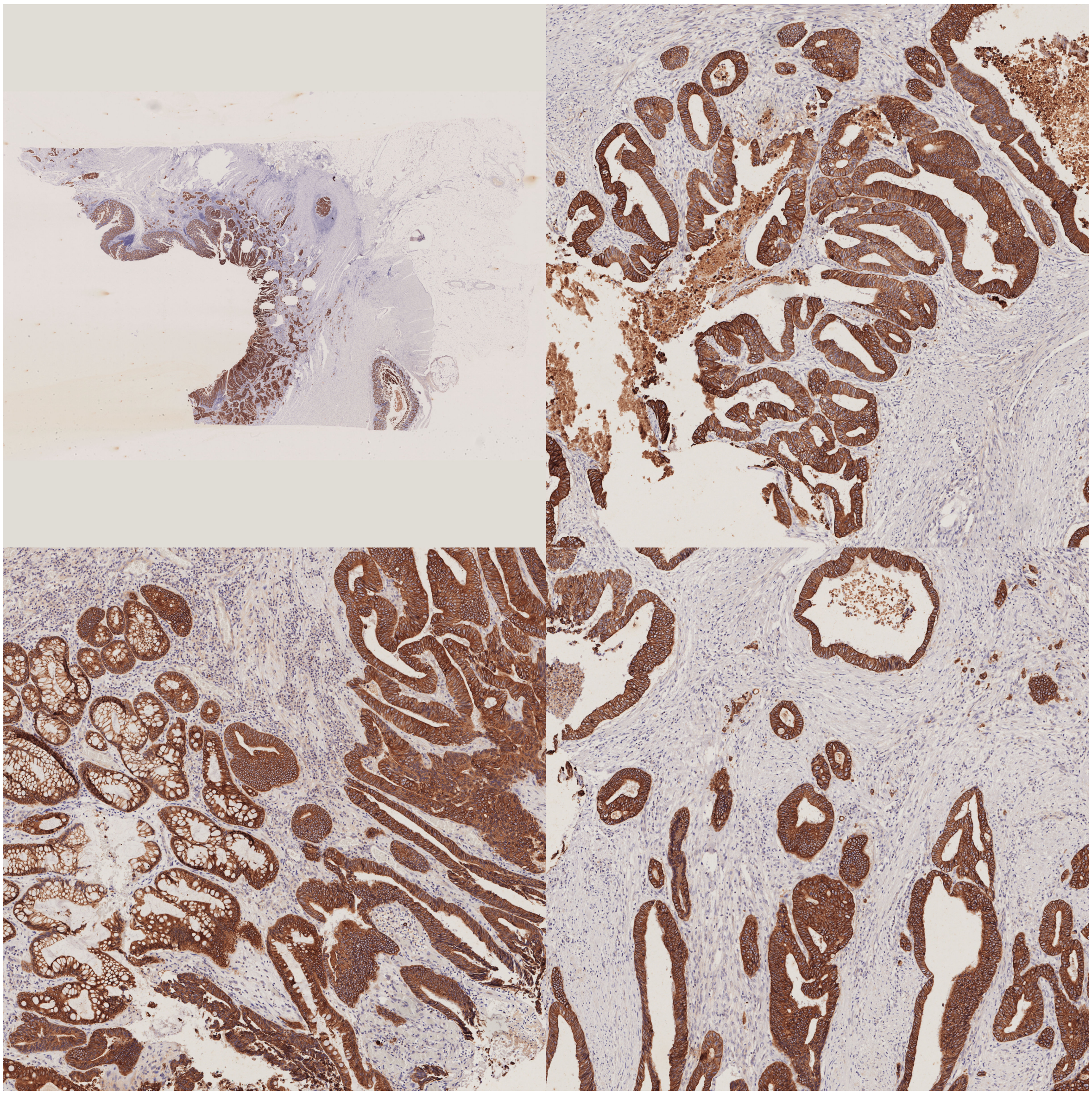


Figure A.9: A real AE1/AE3 IHC whole slide image from the test dataset used as reference for the UNet, Pix2Pix, CycleGAN and our proposed method. The WSI is displayed in the top left quadrant, and three magnified areas from the test dataset occupy the remaining three.

Appendix B

Automated Tumour Bud Scoring by Deep Learning

B.1 AP Cohort

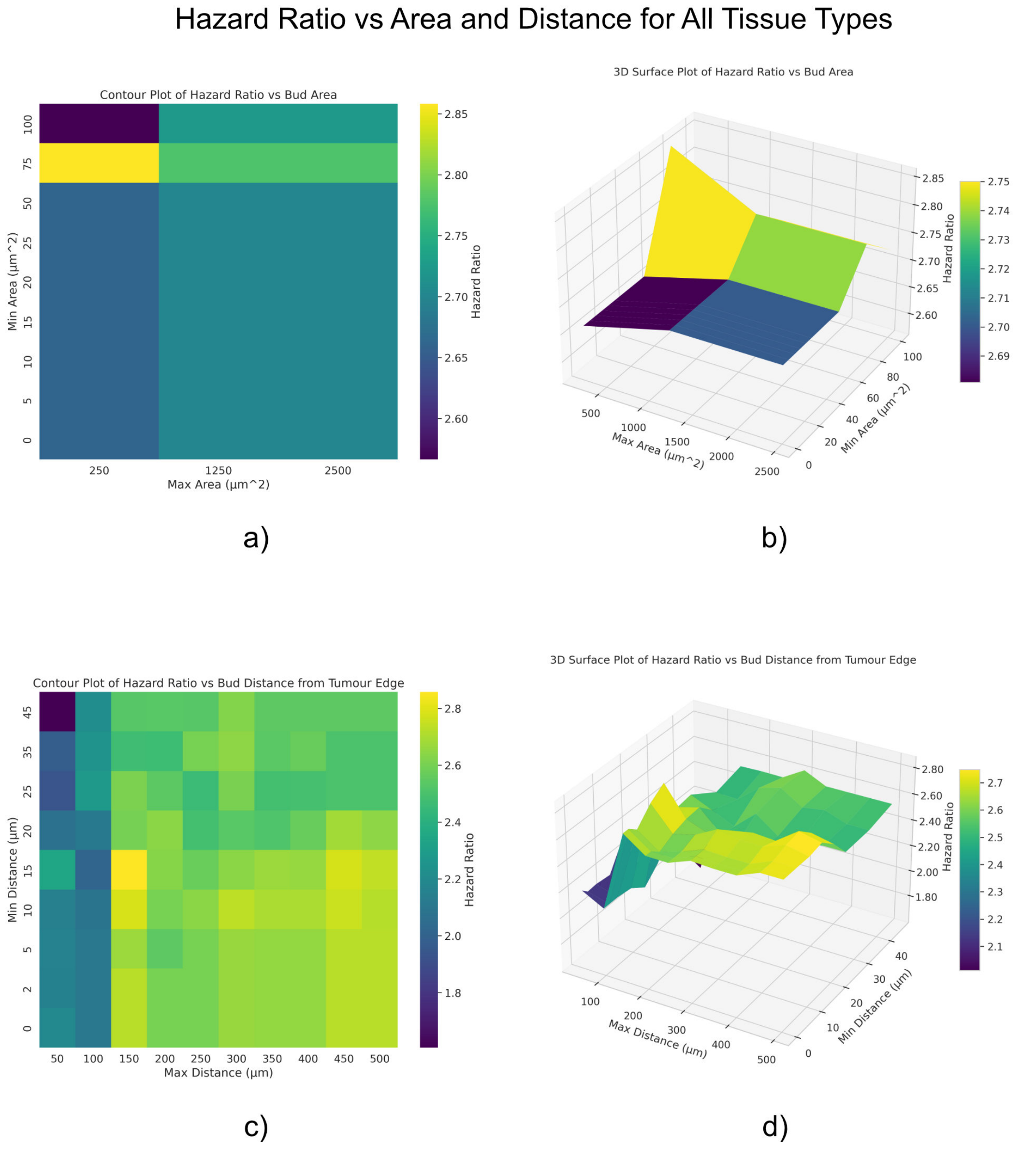


Figure B.1: Shown is the hazard ratio of the area and distance threshold parameters over all tissue classes in the AP cohort. Panel a) displays a 2D heatmap of the hazard ratios resulting from the minimum and maximum area thresholds evaluated. Panel b) displays a 3D contour plot of the values to better highlight change over the parameter space. Panel c) shows a 2D heatmap of the hazard ratios resulting from the minimum and maximum distance thresholds evaluated. Panel d displays a 3D contour plot of the values.

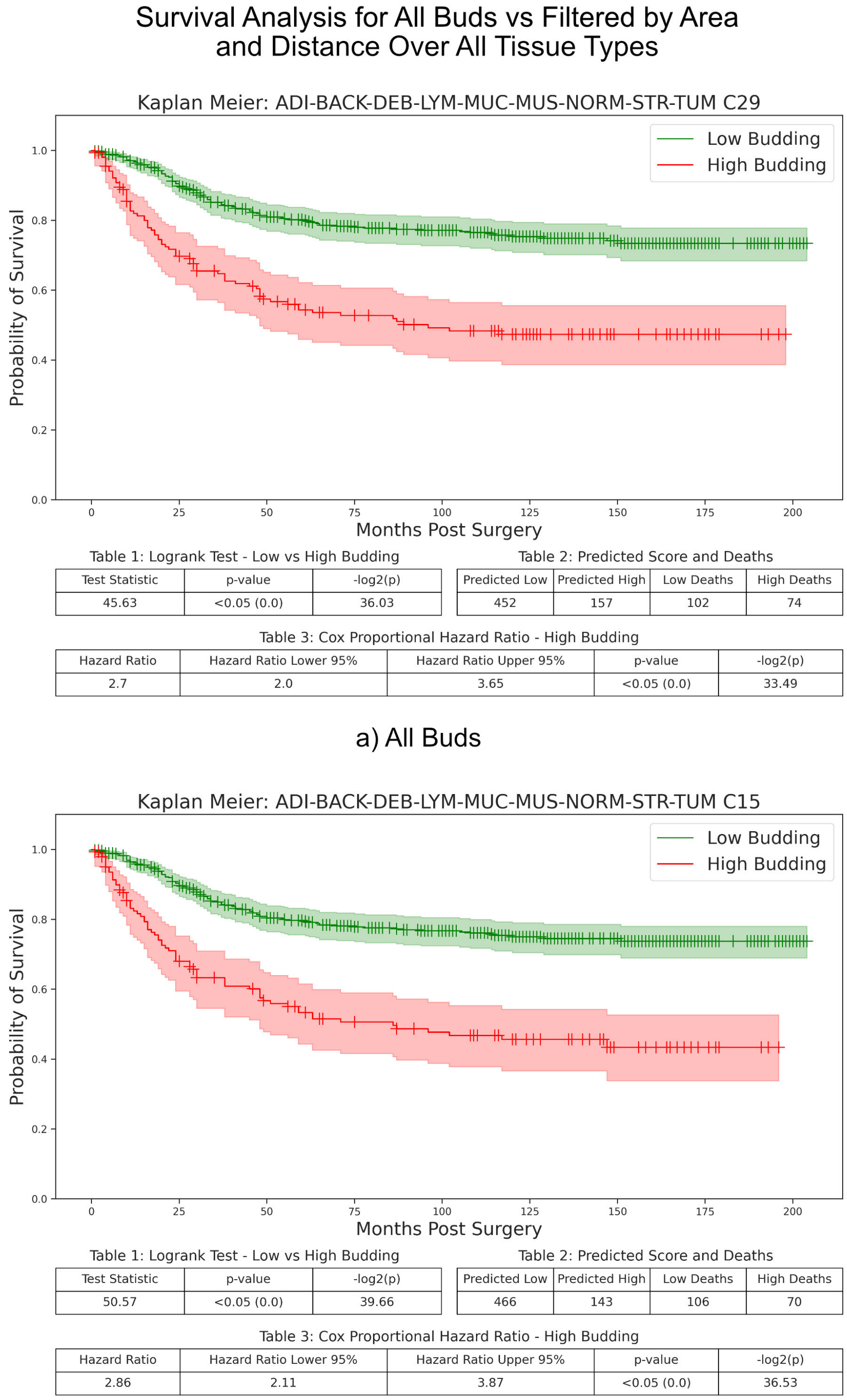


Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
45.63	<0.05 (0.0)	36.03

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
452	157	102	74

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
2.7	2.0	3.65	<0.05 (0.0)	33.49

a) All Buds

Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
50.57	<0.05 (0.0)	39.66

Table 2: Predicted Score and Deaths

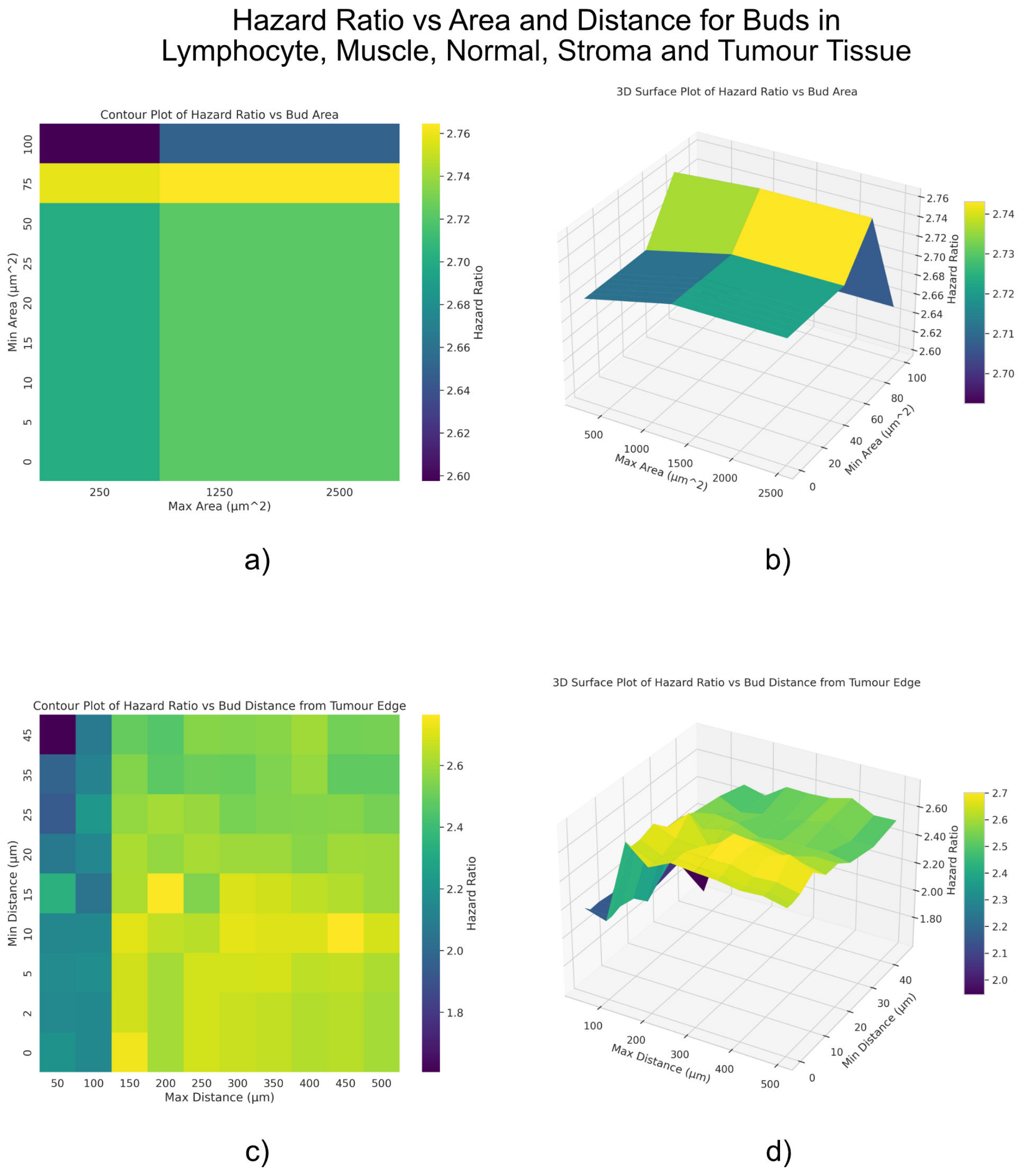
Predicted Low	Predicted High	Low Deaths	High Deaths
466	143	106	70

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
2.86	2.11	3.87	<0.05 (0.0)	36.53

b) Filtered by distance minimum 15μm, maximum 150μm.
Area minimum 75 μm^2 , maximum 250 μm^2.

Figure B.2: Shown is the Kaplan Meier plot of the survival of patients with no filtering and the best cutoff of 29 buds per 0.785mm^2 objective field. Panel b) displays the Kaplan Meier plot of bud scores over the GRI cohort determined after the filtering of detected buds based on the most performant values of a distance minimum of 15μm, area minimum of 75μm^2, a distance maximum of 150μm and an area maximum of 250μm^2, with the best cutoff of 15 buds per 0.785mm^2 objective field.



a)

b)

c)

d)

Figure B.3: Shown is the hazard ratio of the area and distance threshold parameters over the lymphocyte, muscle, normal, stroma, and tumour tissue classes in the AP cohort. Panel a) displays a 2D heatmap of the hazard ratios resulting from the minimum and maximum area thresholds evaluated. Panel b) displays a 3D contour plot of the values to better highlight change over the parameter space. Panel c) shows a 2D heatmap of the hazard ratios resulting from the minimum and maximum distance thresholds evaluated. Panel d displays a 3D contour plot of the values.

Survival Analysis for All Buds vs Filtered by Area and Distance Over Lymphocyte, Muscle, Normal, Stroma and Tumour Tissue

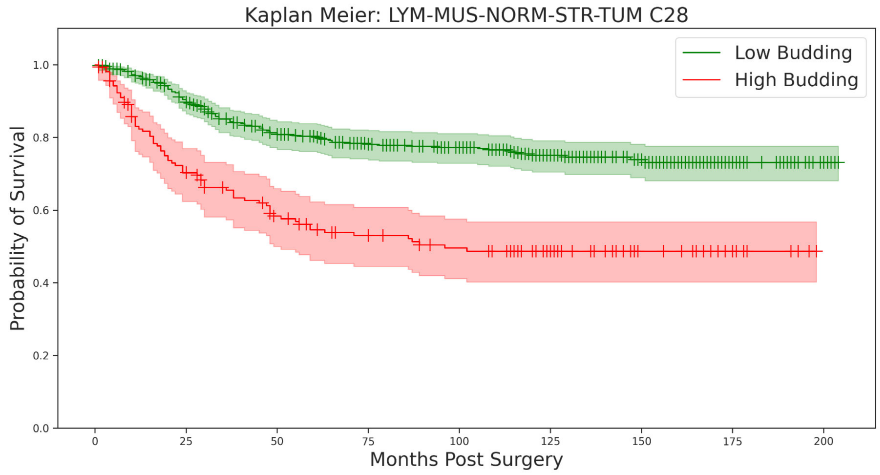


Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
42.21	<0.05 (0.0)	33.51

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
449	160	102	74

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
2.61	1.93	3.52	<0.05 (0.0)	31.32

a) All Buds

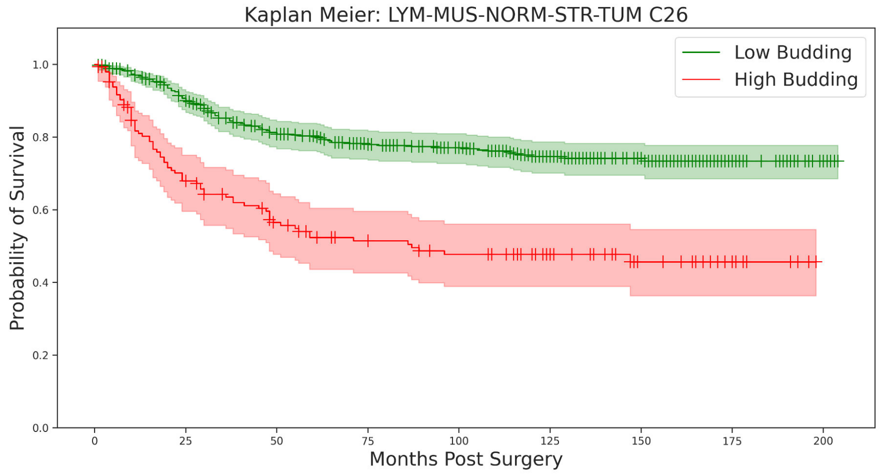


Table 1: Logrank Test - Low vs High Budding

Test Statistic	p-value	-log2(p)
47.36	<0.05 (0.0)	37.3

Table 2: Predicted Score and Deaths

Predicted Low	Predicted High	Low Deaths	High Deaths
460	149	105	71

Table 3: Cox Proportional Hazard Ratio - High Budding

Hazard Ratio	Hazard Ratio Lower 95%	Hazard Ratio Upper 95%	p-value	-log2(p)
2.76	2.04	3.74	<0.05 (0.0)	34.56

b) Filtered by distance minimum 10µm, maximum 450µm.
Area minimum 75 µm^2, maximum unlimited.

Figure B.4: Shown is the Kaplan Meier plot of the survival of patients with no filtering and the best cutoff of 28 buds per 0.785mm^2 objective field. Panel b) displays the Kaplan Meier plot of bud scores over Lymphocyte, Muscle, Normal, Stroma and Tumour Tissue in the AP cohort determined after the filtering of detected buds based on the most performant values of a distance minimum of 10μm, area minimum of 75μm^2, a distance maximum of 450μm and an unlimited area maximum, with the best cutoff of 26 buds per 0.785mm^2 objective field.

Appendix C

Colorectal Tumour Aggression Prediction by Self-Supervised Deep Learning and Transformer Networks

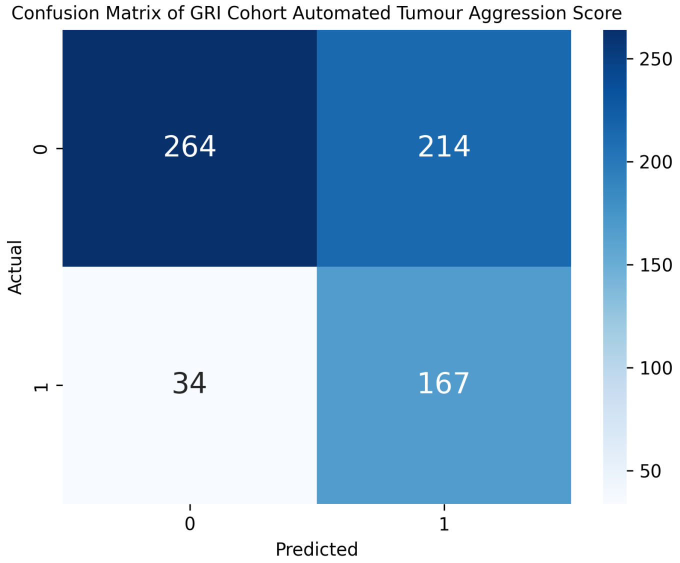


Figure C.1: The confusion matrix of the automated aggression score on the GRI cohort.

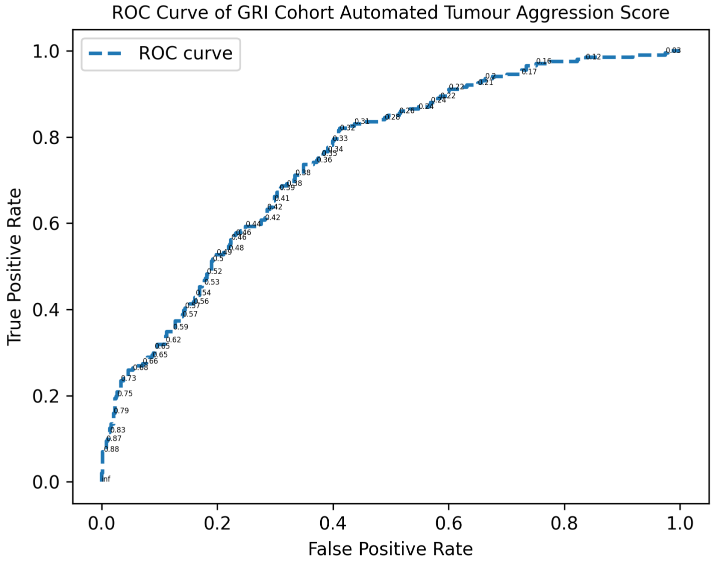


Figure C.2: The ROC curve of the automated aggression score on the GRI cohort.

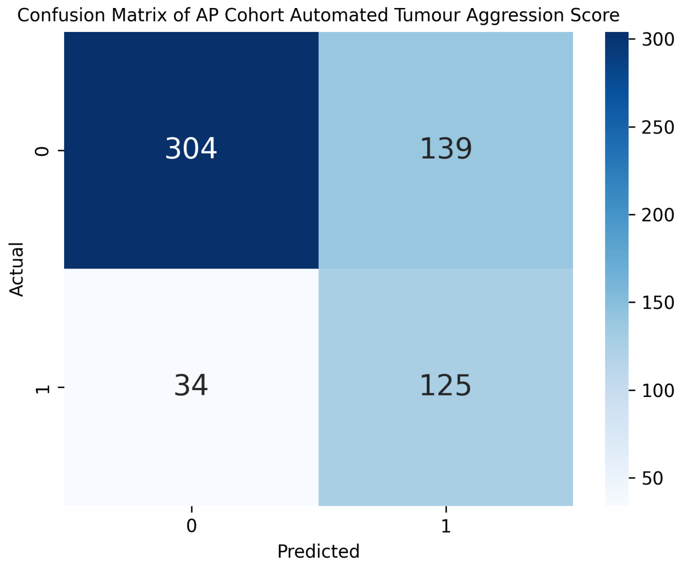


Figure C.3: The confusion matrix of the automated aggression score on the AP cohort.

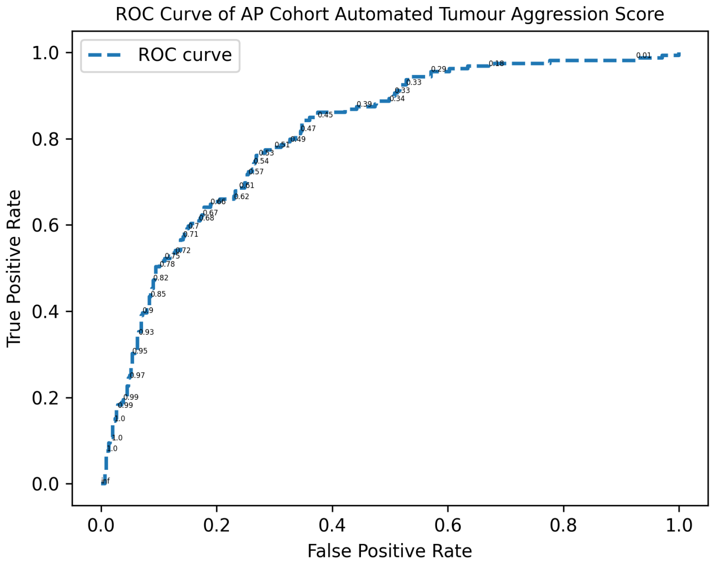


Figure C.4: The ROC curve of the automated aggression score on the AP cohort.