Automated analysis of colorectal cancer

Alexander Ian Wright

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

School of Medicine & School of Computing

October 2017

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Publications

The following papers have been published out of the work of this thesis:

Chapter 3

Wright, A. I., Magee, D., R., Quirke, P., Treanor, D. (2014). Towards automatic patient selection for chemotherapy in colorectal cancer trials. *Proc. SPIE 9041, Medical Imaging 2014: Digital Pathology, 90410A* (March 20, 2014).

Wright, A. I., Grabsch, H. I., Treanor, D. E. (2015). RandomSpot: A web-based tool for systematic random sampling of virtual slides. *Journal of Pathology Informatics*. 2015; 6:8.

Chapter 4

Wright, A. I., Magee, D. R., Quirke, P., Treanor, D. E. (2015). Prospector: A web-based tool for rapid acquisition of gold standard data for pathology research and image analysis. *Journal of Pathology Informatics*. 2015; 6:21

Chapter 5

Wright, A. I., Magee, D. R., Quirke, P., Treanor, D. E. (2016). Incorporating local and global context for better automated analysis of colorectal cancer on digital pathology slides. *20th Conference on Medical Image Understanding and Analysis (MIUA 2016) Proc Comp Science*. Vol 90, pp 125-131.

In all cases, the work has been undertaken by the candidate (as the primary author), with the co-authors providing a supervisory role, directing the research and providing feedback on the writing of the papers.

Acknowledgements

I would like to thank the following people for their advice and assistance during my PhD.

- My supervisors, Dr Darren Treanor, Dr Derek Magee and Prof Phil Quirke for providing support and guidance throughout this project. The multidisciplinary nature of this work requires in-depth knowledge of a wide variety of areas, and the sheer breadth of knowledge that my supervisory team has, both intimidated and enabled me to develop my research skills and profile in this area. They have also been instrumental in developing my understanding of wider strategies, in terms of academic and industry collaboration, University research output, public engagement and external funding bodies.
- Yorkshire Cancer Research, who have supported me financially and enabled me to conduct this research project.
- The pathologists in LICaP that helped me to further my understanding of pathology, histology and cancer morphology, specifically Dr Gordon Hutchins, Dr Emily Clarke, Dr Nick West, Mr Eu-Wing Toh, Dr Heike Grabsch and Dr Scarlet Brockmoeller.
- The Virtual Pathology team, that have kept the servers, websites and digital slides all online and reliable for the duration of this project. They have also constantly been my sounding board, providing advice and feedback on ideas, presentations and more, as well as keeping me focused on my PhD - Martin Waterhouse, Michael Hale, Dave Turner and Catriona Dunn.
- The researchers that generated the millions of hand-labelled images, and provided me with copies of their data - Emma Tinkler-Hundal, Mitt Dattani, Matt Hale and Victoria Montrose.
- Dr Duane Carey for providing advice, assistance and chai latte, whenever I needed to visit the School of Computing.

Finally, I would like to thank Daisy for her encouragement and support throughout my writeup and for humouring me when I needed to discuss ideas at whatever unearthly hour my inspiration struck.

Abstract

Colorectal cancer (CRC) is the second largest cause of cancer deaths in the UK, with approximately 16,000 per year. Over 41,000 people are diagnosed annually, and 43% of those will die within ten years of diagnosis. The treatment of CRC patients relies on pathological examination of the disease to identify visual features that predict growth and spread, and response to chemoradiotherapy. These prognostic features are identified manually, and are subject to inter and intra-scorer variability. This variability stems from the subjectivity in interpreting large images which can have very varied appearances, as well as the time consuming and laborious methodology of visually inspecting cancer cells.

The work in this thesis presents a systematic approach to developing a solution to address this problem for one such prognostic indicator, the Tumour:Stroma Ratio (TSR). The steps taken are presented sequentially through the chapters, in order of the work carried out. These specifically involve the acquisition and assessment of a dataset of 2.4 million expert-classified images of CRC, and multiple iterations of algorithm development, to automate the process of generating TSRs for patient cases. The algorithm improvements are made using conclusions from observer studies, conducted on a psychophysics experiment platform developed as part of this work, and further work is undertaken to identify issues of image quality that affect automated solutions. The developed algorithm is then applied to a clinical trial dataset with survival data, meaning that the algorithm is validated against two separate pathologist-scored, clinical trial datasets, as well as being able to test its suitability for generating independent prognostic markers.

Table of contents

Abbreviations used

ANN	Artifical Neural Network
AUC	Area Under the Curve (ROC curves)
CAD	Computer Aided Diagnosis
CART	Classification And Regression Tree analysis
CCA	Connected Components Analysis
CCD	Charge-Coupled Device
CNN	Convolutional Neural Network
CR07	The MRC CR07 ColoRectal cancer trial
CRC	Colorectal Cancer
CRT	Chemoradiotherapy
DL	Deep Learning
DP	Digital Pathology
FFPE	Formalin-Fixed, Paraffin-Embedded tissue
GLCM	Grey Level Co-occurrence Matrices
H&E	Haematoxylin and Eosin Stain,
HCI	Human-Computer Interaction
HDAB	Haematoxylin and DiAminoBenzidine stains
HPC	High Performance Cluster
HR	Hazard Ratio
HSV	Hue Saturation Value / Intensity colourspace
HTTP	HyperText Transfer Protocol
KNN	k Nearest Neighbours classifier
KM	Kaplan Meier survival curves
LBP	Local Binary Patterns
ML	Machine Learning

MRF	Markov Random Fields
NB	Naïve Bayes classifier
OD	Optical Density
QC	Quality Control
QUASAR	The QUick And Simple And Reliable randomised CRC study
RBM	Restricted Boltzmann Machine
ReLU	Rectified Linear Units
RF	Random Forests
RGB	Red Green and Blue colourspace
ROC	Receiver Operator Characteristic
ROI	Region of Interest
RVM	Relevance Vector Machine classifier
SCRT	Short-course radiotherapy
SLIC	Simple Linear Iterative Clustering of superpixels algorithm
SRS	Systematic Random Sampling
SVM	Support Vector Machine
SVS	Scanscope Virtual Slide (image format)
TB	TeraBytes
TCD	Tumour Cell Density
TMA	Tissue MicroArray
TNM	Tumour, lymph Nodes and Metastasis staging system
TSR	Tumour-Stroma Ratio
WSI	Whole Slide Imaging
XML	eXtensible Markup Language

Figures

Tables

List of equations

"Is technology the missing piece of the puzzle for cancer?"
Alex Wright (2011) University of Leeds Postgraduate Conference: Showcase [1]

Chapter 1 - Introduction

1.1 Problem statement

Colorectal cancer (CRC) is the second largest cause of cancer deaths in the UK, with approximately 16,000 per year. Over 41,000 people are diagnosed annually, and 43% of those will die within ten years of diagnosis [2]. Treatment of patients typically involves a combination of surgery, chemotherapy and radiotherapy, which is invasive, toxic, and in some cases, lethal. Not all patients respond to therapy, and therefore it is important to identify which will benefit from treatment before treatment decisions are made. These decisions are based on characteristics of the patient's cancer, which are diagnosed by a pathologist. The diagnosis of a cancer is performed on an ex-vivo biopsy that has been sectioned, placed on a glass slide and stained for visual inspection, traditionally under a standard microscope.

Traditional glass slides can be digitised using high resolution slide scanners to create large images of tissue, called virtual slides. The University of Leeds Department of Pathology and Tumour Biology has been routinely digitising slides for the past 14 years for research use [3] , using Leica-Aperio (formerly Aperio) digital slide scanners [4]. Glass slides are typically scanned at 0.5 microns per pixel, or 20x resolution, which creates an image approximately one gigapixel in size.

Virtual slides are an excellent basis for the practical implementation of computer vision and Machine Learning (ML) techniques, as they facilitate high throughput, calibrated imaging of pathology images. Manual diagnoses require laborious visual inspection of patient tissue that is subjective and prone to inter-observer variation, and as a result, robust methods for accurate and reliable quantification of cancer tissue are highly desirable. Despite many recent attempts, no methods for routine application to histological analysis have been widely adopted [5]. Clearly there is a need to quickly identify and accurately diagnose CRC to provide appropriate targeted treatment to patients that will respond to invasive therapies. Therefore, successfully

implementing strategies to more efficiently detect and characterise CRC will ultimately increase patient quality of life and survival rates.

The ratio of tumour to stroma within a cancer (TSR) uses the ratio of tumour epithelium to its surrounding connective tissue, to generate a quantifiable metric for determining patient survival. There are multiple methodologies for calculating TSR (discussed in 2.2.3), but the concept of TSR in general is a consistently validated measurement in pathological analysis for determining prognosis [6-8]. However, calculating this ratio manually requires either estimation of the quantities of tissue within a tumour, or sampling enough areas within the region of interest (ROI) to generate reliable statistics. The process of calculating TSR is therefore time consuming, subjective and prone to both inter and intra-observer variation. Automation of this task using computer vision is desirable, as it offers the possibility of objective, standardised quantification.

1.2 Research questions

The work presented in this thesis focuses on the automation of calculating TSRs on digital slide images. In order to attempt to achieve this objective, the following questions have been identified:

- Can the manual analysis of TSR be facilitated by using virtual slides?
- How can manual analysis results be obtained to develop ground truth datasets for both reliable ML training and effective evaluation of any automated solutions?
- Can an automated solution replicate manual scoring with an adequate level of agreement, using the ground truth data?
- What is the minimum amount of visual information required for maximising agreement between observers for manual scoring of TSR?
- Can algorithm performance be improved using conclusions drawn from observations of pathologist scoring?
- What are the issues that affect automated histopathological image analysis?
- How does the automated solution perform when these issues are not present or mitigated?

- Using patient survival data as the gold standard, does the algorithm outperform manual scoring?

These questions form the basic structure of the work conducted in this thesis, as an ordered set of projects which follow a sequential methodology. Each unit of work uses conclusions and questions drawn from previous work to either improve these or motivate new research.

1.3 Thesis overview

The work in this thesis is split into 8 chapters, and is outlined below.

Chapter 2 presents background research relevant to the automated analysis of TSR on CRC tissue, providing insight into the pathology of CRC and current methods for acquisition and inspection of patient tissue., A brief history of digital pathology is presented, and an exploration of appropriate computer vision and ML methods that can be applied to the automated analysis of histopathology images. Finally, current systems that use computer vision and ML to solve visual pathological tasks are evaluated.

Chapter 3 presents the RandomSpot system, a web-based systematic random sampling (SRS) system for use with digital slides, and shows how the system has helped researchers to calculate the TSR in their own work. The chapter then details how the system generates data for independent researchers, that is reusable for training and validating image analysis solutions, and presents a repository for the expert-classified ground truth data, called RandomSpotDB. Work is then undertaken to evaluate one of the datasets contained in the RandomSpotDB, the QUASAR clinical trial, in terms of the image data within it, and a thresholding-based algorithm is applied as a first step to ascertain whether TSR can be calculated using simple image processing techniques. Finally, a ML algorithm is developed to learn features of over 106,000 pre-classified histopathological images, retrieved from RandomSpotDB as image locations (spots). Experiments are performed to ascertain the effect of image size, classifier type and classifier parameters on algorithm performance.

Chapter 4 explores human interaction with digital images, to ascertain optimal scoring conditions for histopathological image analysis. This is done with the intent of mimicking these conditions in future work, to improve the automated solution. Out of the need to generate ground-truth data and agreement statistics, when manipulating scoring conditions, the Prospector system was developed. The web-based experiment platform is presented, and used to identify optimal image size for manual scoring. The results from this experiment identify an

appropriate minimum image patch size for the automated solution, and lead to the conclusion that pathologists use surrounding contextual information to correctly score tissue at the centre of the image patch. A pilot study is also presented, that looks at pathologist agreement correlated to the tissue staining levels. Results show that stain level affects agreement, but the dataset is too imbalanced to generate a solution for identifying poorly stained slides automatically. This motivates work in Chapter 6.

Chapter 5 uses the conclusions from Chapter 4, to sequentially modify the ML algorithm presented in Chapter 3, so that the impact of each algorithm modification can be individually assessed. Five enhancements to the algorithm are presented, based on conclusions drawn from the pathologist experiments. These use local contextual analysis and global slide staining properties to modify the feature vectors of each patch accordingly. Unsupervised segmentation algorithms are assessed for segmentation accuracy and computational speed, and a hybrid algorithm is created, applying normalised cuts to cluster superpixels using computed pairwise similarity metrics. The final product of the chapter is a combination of the modifications applied.

Chapter 6 focuses on the cases where algorithm-pathologist agreement is lowest, to identify visual artefacts that may affect image analysis. The prospector system is used to manually apply quality control (QC) checks on all slides in the QUASAR dataset, applying the pre-identified visual artefacts as categories. Using only the QC-approved slides, the final algorithm from Chapter 5 is re-evaluated, in order to ascertain whether accuracy improves when removing slides that have been flagged for issues. Finally, the dataset is used to extend the pilot study from Chapter 4, to train a ML classifier on the appearances of suboptimal slides, to create an automated QC algorithm.

Using a second clinical trial dataset, the algorithm accuracy is validated in Chapter 7, and the algorithm is assessed for suitability as a real-world prognostic predictor. TSRs generated are correlated to survival using a pre-published method to stratify the patients into two groups, TSR high and TSR low. Previous studies have shown that TSR high groups have better survival statistics, and so the algorithm generated TSRs are evaluated against this criterion, as well as the significance value comparing the dissimilarity between the two patient groups.

Chapter 8 concludes the thesis by summarising the work in each of the chapters, in terms of the achievements that the work has output. Algorithm statistics are summarised so that conclusions can be drawn about how appropriate the presented automated solution is, and future work regarding improvements and extensions to the project are outlined.

Chapter 2 - Background

"Automation of the acquisition and interpretation of data in microscopy has been a focus of biomedical research for almost a decade. In spite of many serious attempts, mechanical perception of microscopic fields with a reliability that would inspire routine application still eludes us." - Judith Prewitt, 1966

2.1 Introduction

Interpretation of the visual characteristics of disease (phenotype) is vital for understanding invasive behaviour and predicting response to therapy. The problem statement in section 1.1 outlines the need for consistent and reliable analysis of patient tissue, so that appropriate decisions can be made in terms of treatment options. Accurately predicting response to chemoradiotherapy (CRT) and similar treatments facilitates longer life expectancy where patients are predicted to respond well, and maximises patient quality of life where invasive and toxic treatments will not stop or slow disease progression. Furthermore, in a healthcare climate of efficiency savings, avoiding giving treatments to patients that will not benefit from them reduces costs to health services.

2.1.1 Chapter overview

Digital pathology has become a rapidly growing field of research, due to the advent of high resolution digital slide scanners, capable of scanning glass slides up to and beyond 400x magnification (a resolution of 0.25 microns per pixel). The acquisition of large gigapixel images allows patient tissue to be visually inspected on standard computer displays (which comes with risks, discussed in 2.3), and enables researchers to develop image analysis algorithms which

have the capacity to address some of the issues that affect manual analysis of disease, mentioned previously. The work in this chapter aims to provide a comprehensive background to the field of digital pathology and its use in analysing CRC both manually and automatically. The chapter is divided into six sections:

1) The introduction, which provides a brief overview of the chapter, and outlines potential benefits of digital pathology and computer assisted diagnosis.
2) An overview of colorectal cancer, a brief account on the anatomy of the disease, and a description of how it is analysed by pathologists to determine a patient's prognosis.
3) A background on the field of digital pathology, in terms of its history, development, and impact in clinical application.
4) A review of current image analysis techniques and methods that have the potential to facilitate the development of computer aided diagnosis algorithms and systems, with respect to CRC images.
5) Exploration of currently developed solutions to automating histopathology image analysis, and identification of the successes and limitations of those solutions.
6) A summary based on findings from the chapter, indicating methods with which to direct the research for this thesis.

2.1.2 Potential benefits of automating phenotype analysis

The digitisation of cancer patient tissue facilitates the development and integration of computer assisted diagnosis (CAD) for histopathological examination. Successful systems have the potential to benefit patients, researchers, clinicians and funding bodies in the following ways:

1. Provide consistent and reliable metrics that correlate with survival, with low rates of variability.
2. Increase throughput of case analyses, reducing time taken to return results to patients.
3. Identify statistical information which is impossible for manual inspection to accurately assess without considerable effort.
4. Reduce or eliminate the need to manually assess routine patient cases, so that more time can be spent on cases which require further inspection.
5. Reduce or replace costs spent on manual routine analysis.

However, the development and implementation of CAD algorithms and systems is non-trivial, due to the complex nature of histopathology images, and large quantities of data required for analysis. Some of these challenges, specific to this work, are discussed in the next section.

2.1.3 Challenges to Computer-Assisted Diagnosis (CAD)

The successful integration of CAD systems into routine pathological workflow faces several key issues which must be addressed before such systems are widely accepted. These issues include (but are not limited to) **accuracy** of analysis, to provide trusted output that is useful to the pathologist, as well as the **speed** of which those outcomes are generated. The systems must be well equipped for handling large volumes of throughput, commonly referred to as "**big data**", and the **cost** of the systems and infrastructure that supports them must be minimised. Due to the complex nature of such analyses, human-computer interaction, or **HCI design** must optimise the balance between simplicity for rapid, easy to use systems, and flexibility, so that all possible modifications can be made to customise and improve analysis. Finally, another consideration is that traditional light microscopy is still used for training pathologists, and therefore **uptake** is slower amongst those that have either no access or no desire to use these expensive, and currently experimental technologies.

2.1.3.1 Accuracy of algorithms

The accuracy of any given automated analysis not only needs to sufficiently reduce error to a degree that is acceptable to experimental design, but also requires extensive validation on a variety of datasets, to be trusted. Simply put - systems that are not trusted do not get used. Many CAD systems require extensive parameter optimisation (sometimes referred to as tuning) for tasks such as nuclear, membrane or microvessel detection. This often acts as a barrier to end users, who are not familiar with computer vision terminology or experimental design, and can lead to suboptimal and underperforming algorithms. Accuracy of CAD systems can be calculated using a variety of methodologies, typically comparing algorithm output to ground truth data, consisting of expert-labelled images, and assessing the rate at which true and false positive and negative classes are detected.

2.1.3.2 Speed of analysis

Automation or partial automation of the pathologist task has the potential to speed up workflow, and return results and treatment decisions back to patients faster than with traditional human scoring. Increased pressures on healthcare services from growing and aging populations require higher throughput of patient cases, and so optimising turnaround time is essential to maintaining these services. CAD has the potential to speed up analysis time by processing images prior to pathologist review, highlighting areas of tissue that are more likely to require in depth attention, and providing trusted output for routine images. The speed at which this data is generated

requires optimised code and systems, and will benefit from high-power processing clusters, which can analyse these large images using distributed and parallel processing.

2.1.3.3 Volume of data

Digital pathology slides can be gigapixel-sized images that contain hundreds (or thousands, depending on magnification and compression) of megabytes of visual information. Glass slides are produced routinely in pathology laboratories, and have large amounts of patient data associated with each sample. Handling this data efficiently is essential to developing efficient and useable CAD systems.

2.1.3.4 Cost of infrastructure

Implementing CAD systems requires expenditure in the supporting infrastructure to set up and maintain systems and workflow, such as digital slide scanners, servers and appropriate end user workstations. Adding requirements for extra workflow steps, in terms of equipment and staff incurs significant costs, and the financial benefits of fully digitising pathology labs is difficult to predict (see 2.3.4).

2.1.3.5 HCI and design

The design of CAD systems needs to consider the large amounts of visual, clinical and meta data associated with digital pathology slides. Displaying the minimum amount of data required for specific tasks may help speed up workflow, but at the expense of comprehensive data analysis. By using intuitive visualisation techniques, such as hierarchical displays and content drill-down, CAD systems have the potential to convey large amounts of information to end users, that allow pathologists to make informed decisions about patients and treatment.

2.1.3.6 Uptake of systems

There are many organisational, technical and financial issues that affect the uptake of CAD systems. Pathologists should be the focus when implementing digital pathology solutions, in terms of systems design, but also expectation management. Digital pathology is still a relatively new discipline that is not routinely used in practice. As such there are no standard practices for training pathologists to use digital slides and associated software, that are comparable to traditional microscope training. This may lead to preference of the older imaging modalities, or conversely, may lead to unrealistic expectations of what digital pathology and CAD systems can do.

The issues of implementing successful CAD systems in practice are by no means limited to these considerations. Research at Leeds is currently being conducted, studying the integration of digital pathology into routine laboratory workflow, and is discussed in section 2.3.4.

2.2 Colorectal cancer

This section provides an overview of colorectal cancer, a brief account on the anatomy of the disease, and a description of how it is analysed by pathologists to determine the best course of action for directing a patient's treatment.

Colorectal cancer (CRC) is the second highest cause of cancer related mortality in the UK, causing 15,903 deaths in 2014 [9]. The disease has a 10-year survival rate of 57%, due to the late stage at which most cases are identified. The diagnosis and treatment of CRC patients relies on the visual inspection of their biopsy samples to assess a range of phenotypic properties that pathologists use to classify the cancer in terms of the spread of invasion (stage) and the deformity of the cancer cells (grade).

2.2.1 Anatomy of CRC

The large bowel is the final part of the digestive tract, which consists of the colon and rectum (ergo, 'colorectal'). Both colon and rectum are tubular in structure, and made of a number of layers, labelled in Figure 1.

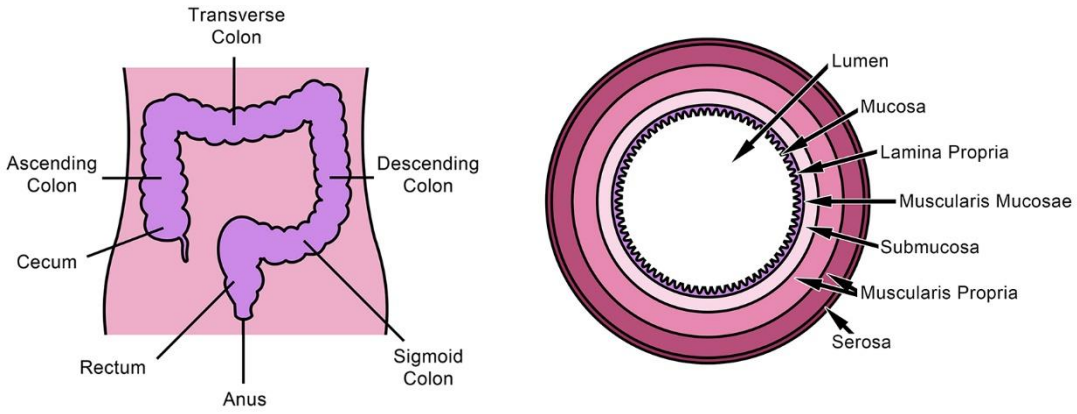


Figure 1 – Diagrams of colorectal anatomy and structure

Left: Labelled diagram of bowel anatomy
Right: Schematic diagram of the structure of a cross section of the bowel

The mucosa consists of cells that line the internal bowel wall, and are arranged in a formation of structures similar in appearance to connected test-tubes. These structures, called crypts, are glands which secrete mucin, and absorb water and bile salts. Crypts consist mainly of epithelial cells, that form the structure of the glands, and mucin is produced by specialised epithelial cells called goblet cells. The reproductive cycle of epithelial and goblet cells is approximately 4 days, through sustainable and asymmetric division of stem cells, located at the base of each crypt [10]. In CRC, the abnormal division of these cells leads to neoplasia and invasive tumours.

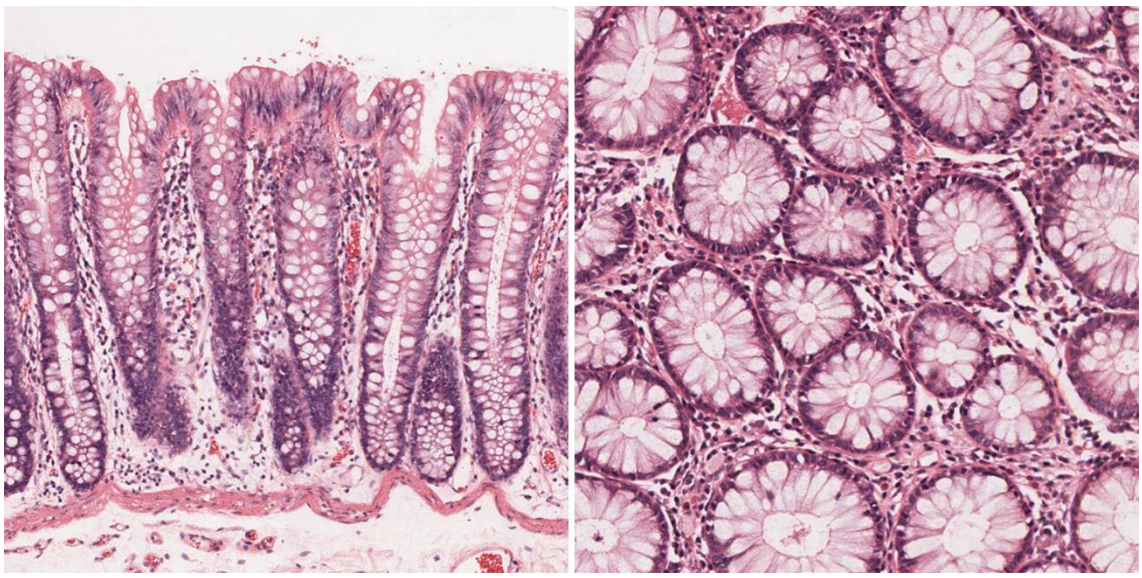


Figure 2 – Longitudinal-section and cross-section views of bowel crypts

Left: Longitudinal-section view of bowel crypts, with the luminal aspect at the top and the muscularis mucosa at the bottom.
Right: Cross-section view of bowel crypts

The lamina propria supports the mucosa, and contains loose connective tissue. This tissue consists of occasional fibroblasts and elastic tissue but also contains small vessels and lymphocytes, and is separated from the mucosa by a thin layer of smooth muscle called the muscularis mucosa, which can be seen at the bottom of the left image in Figure 2. This layer is significant in the diagnosis of CRC, as tumours that penetrate it change classification from benign to invasive (malignant), which changes predicted outcomes and treatments for patients. The submucosa supports the mucosal layer, and contains lymphatic vessels (lymphatics), blood vessels and clusters of immune cells (lymphoid aggregates). Cancers that spread to the submucosa have the capacity to metastasise via lymphatics to lymph nodes (lymph node metastasis) or via the venous system which drains through the portal system, although patterns of metastasis vary between colon and low rectal cancers [11]. The muscularis propria consists of two layers of smooth muscle, the inner circular layer that controls the contraction of the lumen,

and the outer longitudinal muscle that controls longitudinal movement along the bowel. The serosa is the outer lining of the bowel.

2.2.2 Tissue and glass slide preparation

Traditionally, CRC is visually inspected using the light microscope and glass slides. The process of obtaining samples for analysis is comprised of four main steps:

Collection: Cancer tissue samples are retrieved ex-vivo from the patient via a surgical procedure or a biopsy, from the primary or metastatic tumour site.

Embedding: Once removed from the body, the tissue sample is fixed, processed and then embedded in a mountant, such as paraffin wax, to create a rigid structure tissue block that preserves the state of the tumour as much as possible, and allows for sectioning.

Sectioning: The block is sliced using a microtome, to produce a section, typically with a thickness of 5 microns, and sections are floated onto the surface of a temperature-controlled bath of water that allows the tissue to be attached onto a glass slide. Once on the glass slide, the tissue appears colourless, and requires staining to enhance the visual features of the tissue.

Staining: In order to prepare the tissue for visual inspection, specific stains are applied that highlight structures or proteins that are to be analysed (depending on the stain used). Typically, a primary stain is applied to enhance contrast in cellular detail, and a second stain (counterstain) is applied to enhance the appearance of tissue morphology.

2.2.2.1 Haematoxylin and Eosin stain (H&E)

In CRC, the combination of Haematoxylin and Eosin stains (H&E) is routinely applied for morphological assessment, meaning that all tissue is stained the same way, highlighting nuclear and structural components. This makes H&E a relatively inexpensive stain that can be used for general inspection of cancer tissue, allowing trained pathologists to visually inspect tissue at a cellular level, as well as a structural level, for local and global contextual analysis of the patterns of normal and neoplastic cell growth. As a simplification, Haematoxylin stains nuclei blue, and appears purple when co-localised with the counterstain Eosin, which stains tissue structures pink. Examples of H&E stained tissue are shown previously in Figure 2.

2.2.2.2 Variability in glass slide preparation

The process of preparing glass slides for pathological analysis requires skilled laboratory staff that understand the need for consistency in the production of glass slides. However, making

these slides routinely is a repetitive and often time-critical task, and as such, human error in terms of inconsistencies in appearance are common. Minor inconsistencies are accepted as the standard in histopathology, and are generally easily accounted for by pathologists that have enough expert knowledge to understand such variation, but some variation affects pathological analysis beyond compensation, and so glass slides are rejected and not analysed. It is important to factor in this consideration when developing CAD systems, since computers do not have expert knowledge and subsequently may be adversely affected in terms of accurate and reliable analysis output when analysing suboptimal material.

Variability can be found through all stages of glass slide production, and can affect the appearance of the glass slide tissue in multiple ways. Some are discussed below.

Collection: The process of extracting the cancer may or may not retrieve the whole tumour, affecting how much visual information is available to analyse the invasive edge, the extent of invasion and spread.

Fixation: Any delay in placing the tissue in a preservative such as formalin can allow deterioration in the cells and tissue structures. The length of fixation can also affect the component biochemistry of the cells.

Embedding: The medium in which the tissue is embedded affects its appearance (e.g. paraffin wax compared to frozen sections), and how well the tissue is preserved. This affects structural appearance and in some cases the cell membranes may disintegrate or perforate, rendering visual analysis of cell morphometry much more difficult.

Sectioning: The microtomes used for sectioning are high precision instruments that use medical grade blades for slicing tissue precisely at a thickness set by the operator. However, if the machine is not calibrated or the blade is not kept sharp, the resulting tissue sections may be inconsistent thicknesses (making stain uptake inconsistent), or tissue may be scored or torn, instead of cleanly sliced. Calcium in tissues can lead to holes and imperfections, and during the process of transferring sections onto the glass slides, folds may occur, which creates areas of overlapping tissue.

Staining: The chemical compounds used for staining tissue samples are variable in terms of the colour yielded, depending on the batch produced or the vendor purchased from. The application of stains to the tissue is also subject to variation, depending on the length of time the tissue is submerged in the stains [12]. As previously mentioned, the thickness of the tissue affects the amount of stain uptake and the overall colour of the slide. Finally, the age, coverslip and storage conditions of the glass slides affects the stains in terms of fading.

2.2.3 Manual examination of CRC on glass slides

CRC is the end product of the neoplastic development of tissue within the colon or rectum, which begins in the epithelial layer. CRC is visually assessed, and diagnosed pathologically using phenotypic information derived from observable patterns of growth (or lack thereof) [13]. Uncontrolled, abnormal growth in the bowel is caused by excessive division of cells in the base of one or more crypts, and is referred to as neoplasia [14]. Neoplastic cells deviate from normal maturation (developing from a stem cell into a differentiated, functioning epithelial or goblet cell), in terms of shape and behaviour. Groups of neoplastic cells can be classified as benign or malignant tumours. A tumour in the colorectum becomes malignant (a cancer) when it has penetrated the muscularis mucosa, and if it detaches from the primary tumour, it is considered to have metastasised.

Traditionally, pathologists examine cancer biopsies stained with H&E on glass slides to assess how much the cancer has spread (staging), and how fast the cancer is likely to spread (grading). Once these characteristics are obtained, appropriate action can be taken regarding treatment of the patient. These two assessments can be made using standard methods of classification via visual observations of the tissue. Multiple methods for assessing stage and grade of CRC tissue exist globally, but this section focuses on methods used in the UK.

2.2.3.1 Cancer staging

Staging refers to the assessment of the current state of a patient's cancer, and was originally formally assessed using the Dukes staging system, proposed in 1932 [15]. This classified tumours into one of three groups, based on the spread of the disease: A) limited to the bowel wall; B) invasion through the bowel wall, but not spread to regional nodes; C) cases where lymph node metastases are present. The tumour, lymph nodes and metastasis (TNM) staging system [16] was developed into the international system (with variations, depending on the organisation developing it). This system proposes that cancer classification can be standardised using sub-classifications of each of the three categories, illustrated in Table 1 [17]. TNM is well established with eight major revisions, although version five is currently routinely used in the UK [18]. Table 1 lists the TNM (version five) scoring methodology with descriptions of the categories.

TNM Stage	Description
T1	Tumour has grown into the submucosa
T2	Tumour has grown into the muscularis propria
T3	Tumour has grown into the serosa
T4	Tumour has perforated the peritoneum and / or locally spread to other organs
N1	Tumour has spread to 3 or less regional lymph nodes
N2	Tumour has spread to more than 3 regional lymph nodes
M0	No metastasis
M1	Presence of metastasis

Table 1 – TNM staging system (version 5)

Staging is grouped into these categories to minimise subjectivity and inter-observer variation, so that consistency in assessing patients can be maximised.

2.2.3.2 Cancer grading

Cancer grading relates to the aggressiveness of the tumour, and how likely it is to spread. Grading is categorised based on the visual assessment of differentiation, which is defined as the extent to which the cell (and subsequently the parent tissue structures) differs from how it should look at maturation. Cancer grading is assessed by assigning tumours one of three categories, well differentiated, moderately differentiated and poorly differentiated [19]. Cancer cells that are well differentiated have reached maturity and resemble the original (epithelial or goblet) cells from which they have been derived. Poorly differentiated cancer cells have little or no resemblance to their original form, and structures of poorly differentiated cells deviate from the shape of regular glandular structures. Moderately differentiated are midway between the two. Assessing the extent to which cells are deformed is subjective, and subsequently grading is prone to inter-observer variability.

2.2.3.3 Phenotypic prognostic markers

In addition to cancer staging and grading, many other prognostic markers have been identified that correlate with patient survival [20], some of which require analysis of data that cannot be interpreted from standard glass slides, such as DNA or RNA content, cell proteins or other molecular markers. However, many studies have generated successful prognostic markers from phenotypic information, derived from routine glass slides of patient tissue. Phenotypic prognostic markers in CRC include, the proportion of tumour to connective tissue, known as the tumour-stroma ratio (TSR) [6-8], the growth pattern of the invasive edge, known as tumour

budding [21-28], the configuration of the infiltrating margin of the tumour [29], the level of immune response [30], and the presence of lymphovascular invasion [31,32]. These phenotypes are typically calculated manually, and are difficult to estimate as continuous numeric features, leading to large variation between observers. As such, these image properties have the capacity to be improved by automating the task using computer vision.

2.2.3.4 Stereology and Systematic Random Sampling (SRS)

Quantitative histology is based on observations of 2-dimensional representations (sections) of 3-dimensional tissue structures, making the appearance of tissue structures different, depending on how they are cut (see Figure 2). By quantifying the amount of tissue on a slide, an assumption is inherently made that area in 2D is directly related to volume in 3D, which can lead to falsely interpreting the content of the tissue and the growth pattern of tumours [33].

Stereology is the 3D interpretation of stacked 2D cross sections of materials or tissues. Systematic Random Sampling (SRS) is used in stereology to generate unbiased and quantitative data, with the intention of sampling serial sections of tissue, generating estimates of total volumes and total numbers, as opposed to ratios [34]. For object counts (such as nuclei), many methods exist for generating sampling areas [35]. Volume calculations are made using the Cavalieri estimator, which calculates the volume of an object by summing the standardised area a to sampling points (probes) p and multiplying by section thickness T.

$$V = T \cdot (a/p) \cdot \sum p$$

Equation 1 - The Cavalieri estimator for stereology volumetric calculations

SRS is a sampling method that uses point grids (as opposed to optical graticules) to generate an equidistant set of sampling locations, that have a random starting point for the generation of the grid. For a given section of tissue, points are placed equidistantly over the region of interest, in a uniform distribution. Figure 3 illustrates an example of a grid of sampling points placed over CRC tissue.

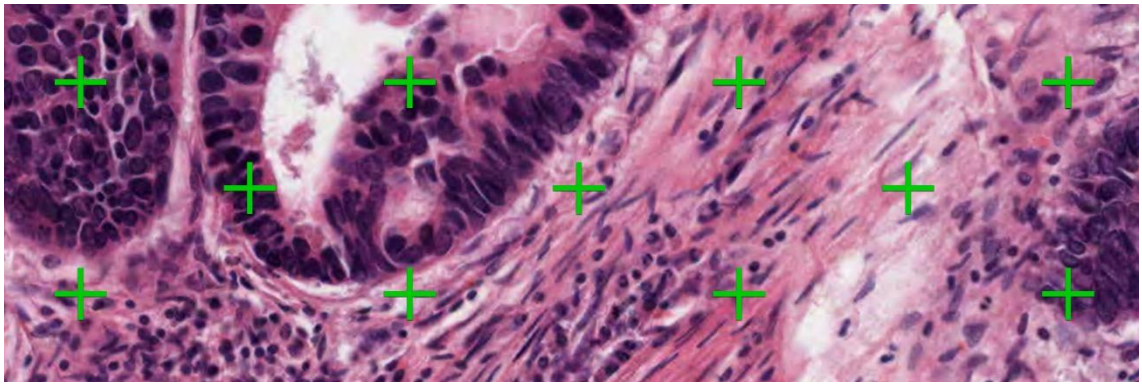


Figure 3 - Example of Systematic Random Sampling (SRS) points

Each sampling point is classified with its respective tissue type so that a quantifiable proportion of tissue types within the image can be calculated.

The number of sampling points can be optimised by first identifying an acceptable level of standard error SE for the mean count measure of the target tissue (e.g. five percent), and estimating the expected volume ratio of the target tissue V_e [36]. Equation 2 calculates the standard error relative to the mean value of the count which estimates V_e, where p is the required number of sampling points that fall on the target tissue (excluding other sampling points).

$$SE = \frac{\sqrt{1 - V_e}}{\sqrt{p}}$$

Equation 2 - Calculation for relative standard error in SRS

By setting the standard error to 5% and the volume estimation to 55% (i.e. 55% of the sampling area should contain the target tissue), we can estimate the number of points that are required to fall on the target tissue p, by solving the worked example in Equation 3.

$$\sqrt{p} = \frac{\sqrt{1 - 0.55}}{0.05}$$

Equation 3 - Worked example for calculating number of SRS points

The worked example shows that p is equal to 180 points, which are required to fall on the target tissue being estimated. The volume estimation of the worked example is 55% meaning that the total number of points required is 180 multiplied by 100 / 55, which equals 327 points. The same methodology of point estimation calculates the required number of points for a 50%

volume estimation at 400, with larger proportions requiring fewer sampling points, and smaller proportions requiring more sampling points, to be adequately represented.

SRS is a powerful tool for quantifying volumetric measurements from 2D image data. By removing inherent assumptions that are made by modelling volumetric calculations applied to 2D analysis (such as that areas of interest are homogeneously dispersed), the metrics derived from stereology reduce deviation from the true 3D representation (bias), and increase reproducibility (precision). However, appropriate calculations must be made to ensure the appropriate number of sampling points are taken, so that the minimum effect size (according to statistical power) can be satisfied [37].

A SRS tool based on stereological techniques is presented in Chapter 3.

2.2.3.5 Survival analysis of phenotypic markers

The method of assessing a phenotypic marker for prognostic significance involves stratifying sets of clinical trial patients into two or more groups. The groups are typically assessed as a reference group that is least at risk, or not exposed to risk, and one or more hazard groups that are predicted to be more at risk. These groups are based on either categorical phenotypic information, such as patients with poorly differentiated vs well differentiated cancers, or by identifying an appropriate threshold (referred to as cut-off) to apply to continuous data. Section 7.4.2.2 details a method using modified Receiver Operator Characteristic (ROC) curves to identify appropriate cut-offs. Once grouped, their correlation to survival is computed using Cox regression analysis, and Kaplan Meier survival curves are generated [38], so that any difference between groups can be evaluated. The most common way of assessing prognostic significance uses a log-rank test and hazard ratios confirm the extent to which the hazard group has an increased risk of death from the reference group [39].

2.2.4 The Tumour:Stroma Ratio (TSR)

The tumour:stroma ratio (TSR) is an observable metric quantifying the proportion of tumour epithelium to connective stroma within a patient's cancer [40]. Consistent generation and analysis of TSR is non-trivial, due to the complexity, variation and subjectivity of the task. TSR is believed to be an important factor in the development and progression of cancer, whereby stroma facilitates growth of cancerous epithelial tissue, such that cancers with a higher proportion of stroma have a poorer prognosis than cancers that have less connective tissue. Research relating to the "seed and soil" hypothesis describes this relationship between the two [41].

Recently, TSR has been found to be an independent prognostic marker in multiple cancers (see section 2.2.4) in breast [42-45], lung [46], oesophageal [47], gastrointestinal [47,48], colorectal [6,49-51], cervical [53], ovarian [54] and endometrial [55] cancers. It was observed from the publications detailing this research that there are two different reported methodologies for obtaining pathological assessment of TSR (with minor variations between studies), which were:

- Estimation based on visual observations
- Systematic random sampling (SRS) of a given region of interest

Visual estimations were typically made by two observers, and a large proportion of studies used a 50% cut-off to create a scoring system comprised of two bins: less than 50% and greater than or equal to 50%. Other variants used more bins, in either ten groups of 10% or twenty groups of 5%. The studies using SRS used specific software to generate a desired number of sampling points within a region of interest, classifying each of the generated sampling points to create a ratio between the sum of the tissue types counted (see the RandomSpot system in 3.2). Table 2 lists the number of publications using these methods and their variations.

Method	**Type of data generated**	**Studies**
Visual estimation	2 bins (<50%, >=50%)	8
Visual estimation	10 bins (of 10%)	3
Visual estimation	20 bins (of 5%)	1
Visual estimation	Continuous	1
SRS	Continuous	4

Table 2 - Methodologies used in TSR publications

The number of TSR studies observed are grouped based on their TSR generation methodology.

Both methods have advantages and disadvantages, most notably the trade-off between time and effort required, and accuracy of results. Analysis of reproducibility by Courrech-Staal et al [56] (not included in Table 2) showed that visual estimation of TSR in quartiles (four 25% bins) yielded a mean kappa agreement of 0.43 across three observers, whereas grouping the scores into two bins increased the agreement to 0.80. Eleven of the studies listed in Table 2 used multiple scorers (all of which used visual estimations) and reported agreement statistics. The mean kappa value of these reported statistics was 0.84 (S.D. = 0.04). SRS experiments did not use double scoring, however West et al reported agreement of 0.97 on a subset of 40 images from their dataset [6]. The difference in agreement statistics shows that using quantifiable metrics is more reproducible than visual estimations. However, the accuracy gains over estimation are only valid if appropriate sampling calculations (based on the calculation in Equation 3) have been made so that sampling error rates can be reduced. Also, the resulting

number of sampling points can be laborious and repetitive to inspect and therefore the methodology is prone to fatigue and inconsistencies. Note that in visual estimation-based studies, survival analysis was applied to cases which were split (stratified) into two groups with the cut-off between groups applied at 50%, whereas the SRS studies identified a method using a modified ROC curve to identify the cut off at 47% (see 7.4 for survival analysis detailing and using this method). Using SRS to quantitate multiple types of tissue within a cancer allows TSR to be calculated in multiple ways (see 3.3.2), and as such can be referred to using different terminology, such as Tumour Cell Density (TCD).

2.2.4.1 TSR in CRC

TSR in CRC is a strong prognostic indicator that has been consistently validated in numerous publications [6,50,51,52,57,58]. As with studies in other cancers, research has shown that the TSR in CRC predicts poorer response to therapy and clinical outcomes when cases exhibit higher proportions of stroma [59]. This provides a useful tool for directing targeted therapies at patients who will benefit from them, and avoiding giving toxic treatments to those who will not. TSR has the capacity to enhance the current TNM scoring system (see section 2.2.3) by including this information with routine patient assessment [60]. The majority of CRC TSR publications report that assessments are made on H&E stained slides, with most studies using visual estimations. Locating an appropriate sampling site for observing TSR is consistent amongst publications, with the consensus being the application of ROIs to the area of highest TCD along the luminal aspect of the tumour.

By suggesting a new prognostic metric for case reporting, extra work would have to be undertaken by the pathologist, per case. This extra work can be minimised by not requiring extra (non-routine) stains and making very general estimates, such as assigning a score of either above or below 50%. This would add very little to the pathologist workload, and could be easily implemented. However, research has not assessed the pathologist agreement levels on cases that fall close to either side of the cut-off (e.g. within ±10%), and it is predicted that agreement is much lower than the reported kappa values of 0.84 on cases falling within this range (see the assessment of research listed in Table 2). It is also not assessed what impact the agreement (or lack thereof) would have on patients with TSRs around this cut-off threshold.

To date, only research at Leeds reports using SRS to quantitate TSR, using the RandomSpot system (see section 3.2), which produces a precise continuous value. This method requires more time to analyse, with West et al reporting approximately 20 minutes per case, annotated with 300 sampling points [6]. The two-bin visual assessment of TSR can be simple to manually

integrate into routine pathological assessment of patient cases, at the expense of accuracy. Using more precise metrics when assessing patients should be encouraged where possible.

Finally, it should be noted that TSR does not account for tumour heterogeneity, i.e. describing whether the tumour is composed of dispersed islands of tumour, or whether the tumour epithelium is grouped together. Using a combination of TSR and heterogeneity would provide a more complete assessment of a tumour and perhaps prove to be a more strongly correlated prognostic marker.

2.2.5 Summary

Analysis of CRC is non-trivial, due to the variable appearance of neoplastic cells and glands. Current scoring systems account for variation in manual assessment by identifying finite categories with clear boundaries, such as the extent of invasion. However, these boundaries are still open to interpretation, and the amount of change compared to normal cells and crypts is made using subjective estimates. This makes manual inspection and analysis of the disease prone to inter-observer variability, which may ultimately affect decisions for patient treatment.

Traditional scoring methods use categorical bins to reduce scorer variation and subjectivity, which may be at the expense of better prognostic capabilities of more precise metrics.

The variability in the end product of the prepared tissue can be affected at every stage of the process, which can drastically affect the visual characteristics of the tissue in terms of colour, shape and content. The impact of these variations is accounted for in manual analysis, by using expert pathological knowledge. This variation must also be considered in automated solutions, which are more likely to have rigidly defined models of tissue.

The tumour:stroma ratio (TSR) is a validated prognostic indicator in CRC and other cancers, with multiple manual scoring methods that vary in precision and the amount of time and effort spent evaluating images by a pathologist. SRS provides a robust and accurate method of quantification, when appropriate calculations are made to minimise error rate in the sample size, but as the worked example shows, 400 points are required for an expected target frequency of 50%. Analysing this many points per case is time consuming, laborious and prone to inter-scorer variation, and so automation is highly desirable.

Despite these challenges, there is a clear benefit that consistent and reliable automated analysis of CRC will bring to the pathologist workflow, and ultimately patients.

2.3 Digital pathology

This section provides a background on the field of digital pathology, in terms of its history, development, as well as current and forecasted impact in clinical application.

2.3.1 History of digital pathology

Digital pathology (DP) is a rapidly growing discipline which encompasses a wide field of research in the histopathology domain, focusing on the acquisition of digital slide images, and the creation and management of data that can be derived from them. These fields include (and are not limited to) image analysis, artificial intelligence, high performance computing, human-computer interaction, psychophysics, informatics, data management, communication management and colour science. The current spike in interest in this field is due to the reduction in cost of high resolution digital slide scanners, which have been commercially available for approximately 15 years, combined with the increase in high powered multi-core processing available on single workstations and high-performance clusters (HPCs). However, the field of digital pathology extends back further to 1955, which saw the first attempt at digital microscopy, called the Cytoanalyzer [61]. An article from the Franklin Institute at the time writes,

"The device, called a Cytoanalyzer, scans the microscope images of the cells, automatically sorts them in its "mind" according to their characteristics, and classifies them as normal or suspicious." [62].

The characteristics that the article alludes to are based on the optical density of the cells. Figure 4 illustrates examples of cell images digitally captured by the Cytoanalyzer, printed as photomicrographs.

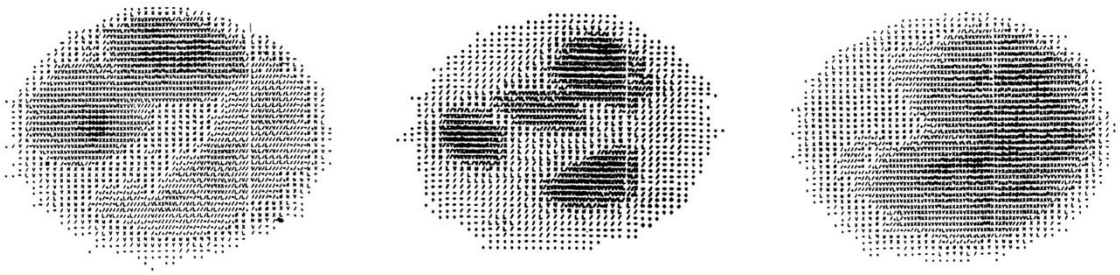


Figure 4 – Example photomicrograph images captured by the Cytoanalyzer [63]

Left: Eosinophil
Centre: Neutrophil
Right: Lymphocyte

The quote at the start of the chapter refers to the failure to utilise the Cytoanalyzer for routine clinical practice [63], however, one could be forgiven for thinking that the quote could have as easily been written in recent years, regarding image analysis solutions to modern digital slide images.

Until the development of high throughput digital slide scanners, microscope-mounted charge-coupled device (CCD) cameras were used to obtain low resolution digital images of microscopic fields of view for research into automated analysis of histopathology images [5]. Since the introduction of whole slide scanners in the early 2000s, the digital acquisition of full tissue samples at microscopic magnification, known as whole slide imaging (WSI) has become common in pathology research environments. Images can be scanned using a 20x, 40x or 80x objective, where images scanned with a 20x lens are scanned at a resolution of 0.5 microns per pixel. One standard digital slide image is approximately one gigapixel in size, and as such requires optimised compression (typically JPEG or JPEG2000 formats) to avoid generating images with prohibitive file sizes. As a result of the introduction of digital slide scanners, the number of publications presenting algorithms that attempt to automate histopathology image analysis exponentially increased [43,44]. However, research into routine clinical use of digital slides for primary diagnosis for pathologist interpretation is ongoing [45-47], with the US Food and Drug Administration (FDA) approving one single system for clinical use as of 2017 [69].

2.3.2 Digital slides vs glass slides

There are many advantages to using digital slides over glass slides [67]. These include immediate benefits to the pathologist, such as worldwide access to slides, when using web-facing digital slide servers, the ability to use computer monitors as inexpensive teaching and

collaboration tools instead of multiheaded microscopes, instantaneous sharing of images for obtaining second opinions, lack of stains fading in storage and eliminating the danger of slides being lost or damaged in transit. Other benefits rely on successful development and integration of CAD systems, which are not currently routinely accepted [70].

However, there are drawbacks to consider when assessing digital pathology as a replacement for traditional light microscopy. The process of generating digital slides still relies on the original laboratory workflow of glass slide production, in addition to adding a process to the pipeline, and therefore there are no efficiency gains in that respect. This extra step also requires skilled scanner operators to scan and apply quality control (QC) checks to every slide scanned. Extra hardware and digital slide storage must not only be purchased but also maintained by skilled technicians and network administrators. Digital slides are reliant on fast network or internet connections to view slides without progressive rendering of image tiles becoming distracting and frustrating. Also, end user hardware is not standardised, and the quality of the device used for viewing digital slides will greatly affect the experience.

2.3.3 Variability in digital slides

In addition to the issues of variability in glass slide preparation (section 2.2.2.2), the process of digital slide production, as well as the configuration of end user devices adds further scope for variation in slide appearance.

Scanner differences and variation

The make and model of scanner instrument can affect the appearance of the digital slides. Area scanners take traditional 2D photograph-style snapshots of a section of a whole slide image, and combine (stitch) them as image tiles. If the whole area is not uniformly illuminated, this can create issues with vignettes at the edges of each tile, and can create a gridline-like appearance over the whole slide image. Line scanners use linear-array detectors to try to mitigate this problem, but can instead create striping artefacts, if illumination is not optimal. As with traditional microscopes, digital slide scanners require a backlight to illuminate the tissue as the image is captured. Earlier digital slide scanners used halogen backlight bulbs which required warming up before reaching optimal brightness. Also, the maximum brightness of each bulb gradually reduces over time, meaning that slides scanned are not consistently lit. Brightness and colour calibration is intended to compensate for such variation, however, different scanner manufacturers use different calibration settings, which achieves varying images of the same tissue.

Colour calibration and variation

Histopathological analysis relies on accurate representation of colour so that structures in tissue and nuclei can be identified and diagnoses can be made. The colour of digital slide tissue can be affected by variation of physical factors, in both the glass slide production process (section 2.2.2.2), and the digital slide scanning process. Colour is also affected digitally by calibration of scanners, which transforms the red, green and blue (RGB) values in an attempt to change the scanned image colours to the true tissue slide colours [71]. This technique may also be used to digitally enhance colour or contrast to make slides easier to score, or look more striking for publications and media. The transformation parameters (colour profiles) are set during scanner calibration, typically using a colour calibration slide similar to a Macbeth colour chart [51,52]. This allows reference colours to be adjusted to their known values using a linear function that can be applied to subsequent scanned images [74]. Scanner calibration of this type may or may not be routinely enforced by scanner operators.

Variation of end user devices

Finally, the endpoint of the digital slide viewing pipeline affects both user experience, and the ease with which subtleties in differences between colour can be observed [54,55]. These issues are affected by monitor size and resolution, as well as contrast ratio, and colour calibration. Colour calibration of end user devices affects the visual appearance of digital slide images, as well as in-built calibration of slide viewing software, featuring enhancements or additional colour management profiles.

Using expensive, high resolution and high contrast medical grade monitors is likely to mitigate issues with colour calibration at the display side of the pipeline, and work at Leeds is currently being undertaken to establish to what extent such levels of contrast improve pathologist scoring [77]. By standardising the appearance of digital slides along all aspects of the digitisation pipeline, more consistent diagnoses can be made, and current barriers to using digital slides for primary diagnosis can be reduced.

However, the display resolution and colour management of display devices is less relevant to automated solutions, which obtain and process images independently of these device-level transformations. Therefore, automating image analysis tasks that require expensive visual display equipment may help to reduce future costs in routine diagnostic work.

2.3.4 Digital pathology slides for routine clinical use

The variation in both digital slide production and viewing, discussed in 2.3.3 means that end user experience can differ vastly between workstations and images produced by digital slide scanning centres. This variability has contributed to the slow uptake of digital pathology for routine clinical use [78], and at the time of writing, one digital slide viewing system (Phillips IntelliSite Pathology Solution) has been FDA approved for routine diagnostic evaluation of patient tissue [69]. Obstacles to widespread implementation of digital pathology include the additional investment and workflow steps required, in addition to existing laboratory practises [58,59]. Most importantly, is reassurance that the ability to maintain (if not improve) the current levels of speed, accuracy and consistency of pathologist scoring is preserved with the new technology [60,61]. Subsequently, microscope scoring is typically the gold standard for comparative studies using traditional and digital pathology imaging modalities [83]. Experiments into replicating fields of view similar to the microscope on high resolution displays often find that once familiarised, pathologists can navigate and score tissue as well as traditional methods [47,63,64], and time taken spent learning the interface is much quicker than that of a microscope [86].

2.3.5 Summary

Digital pathology currently has the potential to facilitate faster throughput of patient case analysis, by providing faster and more collaborative imaging modalities, and assistive workflow technologies. However, variability of the appearance of digital slides can be affected by every stage of the image generation pipeline – from obtaining patient specimens, through to viewing the scanned tissue on a display. Developing accepted and robust automation solutions requires standardisation and extensive validation to be accepted by national regulatory bodies, and very few systems currently are.

Automation of routine visual inspection tasks using computer vision and machine learning has the capacity to increase speed, accuracy and reproducibility of results, given that developed solutions are validated, and trusted by the pathology community.

2.4 Image analysis

This section presents image analysis techniques, methods and technologies that have the potential to facilitate the development of computer aided diagnosis algorithms and systems, with respect to CRC images.

2.4.1 Image retrieval

Digital slide images at the University of Leeds are scanned using Leica-Aperio digital slide scanners, which create JPEG2000 compressed BigTIFF image pyramid files with a proprietary extension, known as the ScanScope virtual slide format (SVS). At the time of writing, there are over 350,000 digital slides stored on the Leeds image server, which use 115 terabytes (TB) of hard disk storage. All images are made accessible over the web using the Leica-Aperio ImageServer software, which means images can be retrieved programmatically either over hypertext transfer protocol (HTTP), or via accessing the images directly on the server. By retrieving the images over HTTP, the ImageServer software retrieves the appropriate image tiles (set to 256x256 pixels in the scanner software) and combines them. The image is then converted into a web-friendly JPEG format and compressed using a quality input value set by the HTTP GET request. This method of transmission applied further compression to the images, creates a high demand on server resources, and is subject to transmission errors. Therefore it is preferable to access the slides locally, using vendor-neutral digital slide libraries such as OpenSlide [87], or the Open Microscopy Environment (OME) BioFormats package [88].

2.4.2 Analysis of colour and stain

Digital slide images represent colour using the RGB colourspace. This method uses additive colour mixing, whereby colours that are combined get brighter. This is representative of combining multiple wavelengths in the coloured light spectrum. Pixels can be analysed for their individual colour values using RGB values, however, these values are not particularly useful for separating colours based on hue. Transforming RGB values into the hue, saturation and value / intensity (HSV) colourspace creates values which are more intuitively represented by the human

visual system. Figure 5 illustrates that this transformation is useful for separating tissue counter-stained with Diaminobenzidine (HDAB).

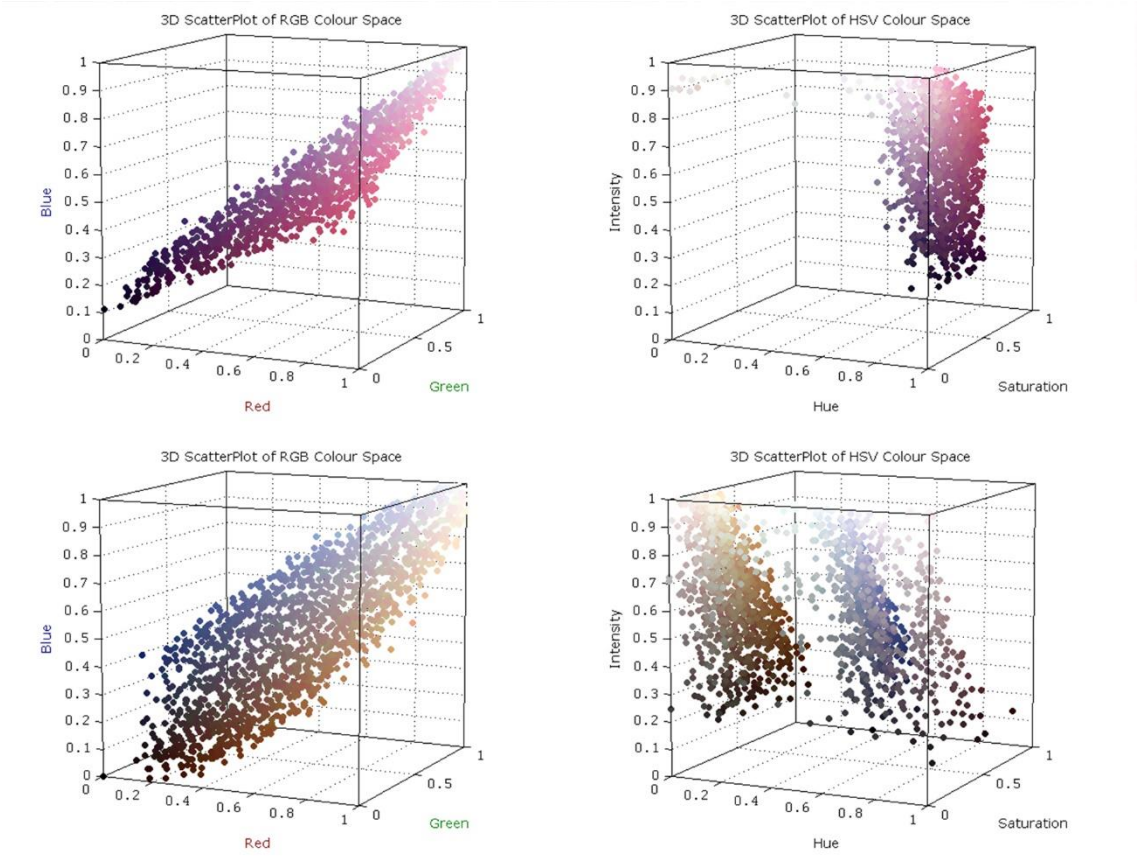


Figure 5 - 3D Scatterplots of representative RGB and HSV values

The values plotted are representative of pixel values found in H&E and HDAB stained slides.
Top Left: RGB values of H&E stained tissue
Top Right: HSV values of H&E stained tissue
Bottom Left: RGB values of HDAB stained tissue
Bottom Right: HSV values of HDAB stained tissue
The figure shows that HDAB staining is separable in the HSV colourspace, whereas the H&E staining is not.

The 3D scatterplot in the bottom right of Figure 5 shows that the HDAB stains can be separated and analysed individually with linear thresholding. However, this is not the case for H&E stains used in histopathological analysis.

In histopathology, stains that are combined get darker, which is known as subtractive colour mixing. By transforming the (additive) RGB colourspace, using values that represent pure Haematoxylin and Eosin, stain colours can be represented independently, digitally separating the stains [68,69]. The process of transforming the colourspace is called colour deconvolution, and one of the most commonly used methods in histopathology uses orthonormal transformation of the RGB colour space using predetermined vectors that represent the optical

density (OD) of individual staining components [91]. The resulting transformation produces images which have three pixel values representing three stain colours. Since typical histopathology staining uses only one counterstain, the third OD vector is usually calculated from the cross-product of the other two.

$$OD_C = -\log_{10}\frac{I_c}{I_{0,c}} = Ac_C$$

Equation 4 – Calculation of OD for pure stains

Equation 4 shows the calculation of OD for colour channel $_C$, where the ratio I_c / $I_{O,C}$ is the intensity of transmitted light, relative to the incident light (the transmission coefficient). *A* is the concentration of stain and c_C is the absorption factor of the pure stain. To transform the RGB image using OD values, the OD image is multiplied by the inverse of the OD matrix.

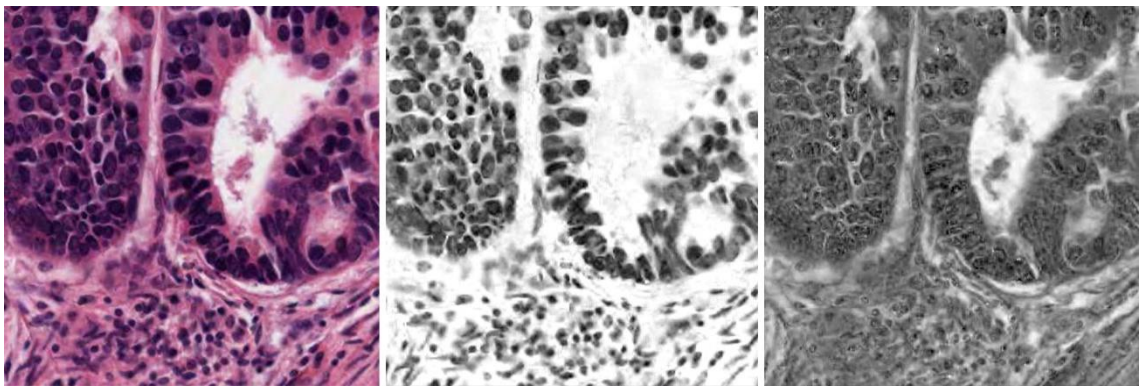


Figure 6 - H&E stains digitally separated by colour deconvolution

Left: Original Image
Centre: Haematoxylin colour channel
Right: Eosin colour channel
Note that the Haematoxylin channel image highlights nuclear components and the eosin channel highlights structural elements.

Note that the separated colour channels are single-value intensity images, and in many visualisations, often have a prototype stain colour applied to them for visual representation purposes only. The third colour channel (representing everything else) is not shown.

It should be noted that colour deconvolution follows Beer-Lambert's law, in that there is a linear dependency between stain concentration and OD. However, this is not true for all stains, most notably the brown DAB stain, which scatters light, breaking one of the prerequisites for colour deconvolution [71,72].

2.4.3 Spatial filtering

Spatial filtering is a general image processing technique that applies an operation to each pixel value in an image $I(x, y)$, using the intensity values of the surrounding pixels in neighbourhood N, with co-ordinates (u, v). The operation, based on convolution, applies a filter (or kernel) H, the same size as N, to each pixel in the image. The filter values are referred to as filter coefficients, rather than pixels. For linear spatial filtering, filter coefficients are multiplied with image pixels at the corresponding point in N.

$$g(u, v) = \sum_{(i,j) \in N} I(u + i, u + j)H(i, j)$$

Equation 5 – Linear spatial filtering

Equation 5 describes the process of linear spatial filtering, where $g(u, v)$ denotes the resulting value for pixel *(u, v)*. The sum of all values is taken from the pixels in neighbourhood *N*.

Two examples of linear spatial filters in image analysis are blurring and edge detection. Blurring functions use averaging filters, which take the average value of the pixel values in neighbourhood *N*, such that the filter coefficients are all set to 1, divided by the number of elements in the filter. This is so that the sum of the filter is equal to 1, and is referred to as normalisation. Edge detection filters are directional, in that they are applied in both horizontal and vertical orientations, and the mean value for all directional results is taken per pixel. Figure 7 shows examples of both filters applied to the same greyscale image of CRC tissue.

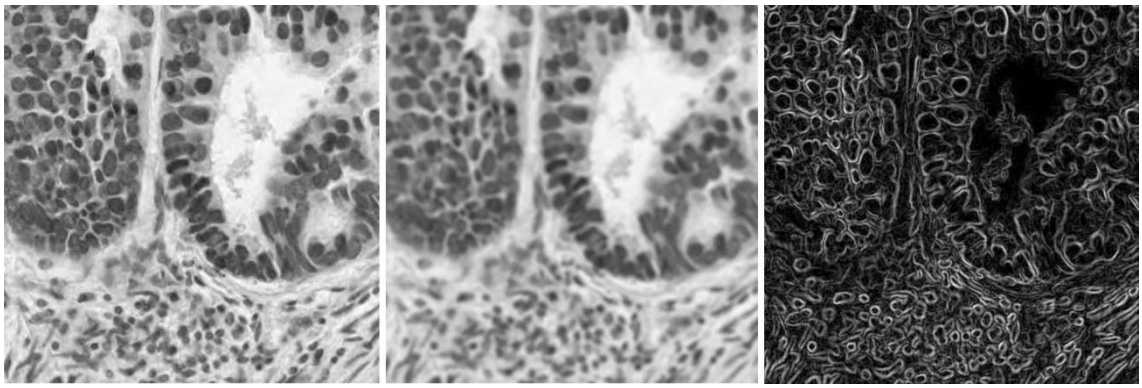


Figure 7 - Examples of two spatial filters on CRC tissue

Left: Original greyscale image taken from intensity channel of HSV image
Centre: Averaging filter applied, blurring the image
Right: Edge detection filter (Prewitt) applied, enhancing edges in the image
These filters form the basis of more advanced computer vision techniques.

Other filters provide variations on these two methods, such as Gaussian blurring, and Sobel edge detection. Median filtering, simply sets the pixel value (i, j) to the median value of all pixels in the neighbourhood N. This method is less affected by noise in the neighbourhood, when compared to the standard averaging (or box) filter, and so can be used for noise reduction in image analysis. This method may be particularly useful when removing sparse areas of nuclei (such as stroma when compared to tumour), when analysed at low power.

Convolutional Neural Networks are a form of Deep Learning, which independently learn spatial filters that appropriately model image data, instead of using human-generated filters or features, and are discussed in section 2.4.6.8.

2.4.4 Texture analysis

Texture is an important image feature which can be used to distinguish between areas of heterogeneous and homogeneous appearance [94]. Textures can be described as images (or areas of images) that contain repeated structures, often contain a degree of randomness, and can be modelled statistically. This section focuses on two common techniques that use statistical methods, although there are many others that fall under four categories, listed below [95].

- Statistical methods
  The spatial distribution of intensity levels within an image
- Geometric methods
  Derision of numeric values from texture elements, called primitives
- Model based methods
  Predefining image models that can be used to analyse or synthesise texture
- Signal processing methods
  Applying frequency analysis, computing features from filtered images

2.4.4.1 Grey Level Co-occurrence Matrices (GLCM)

Textural analysis using Grey Level Co-occurrence Matrices (GLCM) assesses differences between pixel pairs in a greyscale (intensity) image [96]. Initially a greyscale image is reduced from an 8-bit unsigned integer matrix with 256 grey values to N grey values, (typically 8), and an empty NxN matrix is constructed. For every pixel in the down-sampled image, a predetermined offset defines pixel pairs to compare to. Depending on the implementation, an offset of [1 0] would indicate a comparison to be made +1 along the x axis and 0 along the y axis. Some implementations specify offset in degrees. The down-sampled grey values of both

pixels represent co-ordinates in the NxN matrix, and the number of co-occurrences of the pixel pair values is stored at those co-ordinates. This 2D histogram is referred to as the GLCM.

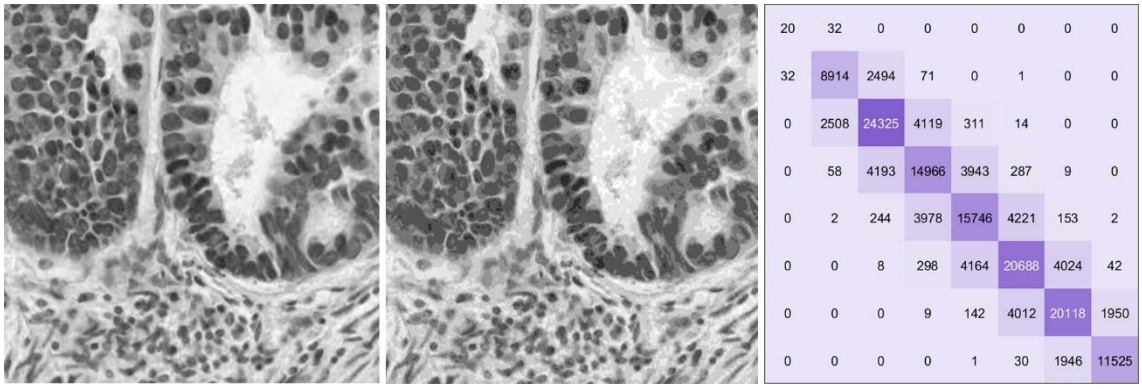


Figure 8 - Example of an 8x8 GLCM on CRC tissue

Left: Original greyscale image taken from intensity channel of HSV image
Centre: Down-sampled image to 8 grey values
Right: GLCM showing counts of co-occurrences between pixel pairs with offset [1 0]
Images with a higher proportion of pairs along the diagonal indicate a more homogeneous image.

Figure 8 illustrates the process of generating a GLCM for an image with an offset of [1 0]. However, as with the edge detection convolution kernels, GLCM calculation is a directional function, and therefore it is often more appropriate to repeat the process in four directions, and take the average value of each GLCM co-ordinate.

Once the GLCM has been generated, features can be calculated from the distribution of pixel pairs. Haralick features are calculations that quantify properties of texture, based on statistics of the computed GLCM [97]. These features include: contrast, which is a measure of how intensity differs between pixel pairs over the whole image; energy (also referred to as angular second moment), which is the sum of squared elements in the GLCM; homogeneity, which measures how closely the distribution in the GLCM fits the diagonal.

GLCM texture properties are commonly used and robust features for texture analysis, and may be useful for discriminating between different areas of CRC tissue. Methods using texture to segment and classify tissue are discussed in 2.5.3 and 2.5.4.

2.4.4.2 Local Binary Patterns (LBP)

The Local Binary Patterns (LBP) algorithm is a rotationally invariant texture descriptor, using neighbouring pixels to analyse differences in intensity [98]. Traditionally LBP is computed per-pixel using 8-neighbourhood connectivity, obtaining each value as a 1x8 vector. For every value in the vector, the current pixel *(i, j)* is applied as a threshold, where values lower than the threshold are set to zero, and values above the threshold are set to one. This yields a vector of 8

binary digits (bits), which is converted into an 8-bit number, and assigned as the value to pixel *(i, j)*.

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^P$$

Equation 6 – LBP value for a given pixel with neighbourhood size of 8

Equation 6 describes the process of generating an LBP value for a given pixel, such that *P* is the number of sampling points (neighbourhood size), and *R* is the radius. The value g_c represents the intensity value at the centre of the local neighbourhood, and g_p denotes the intensity values of the surrounding pixels. Figure 9 visualises this process.

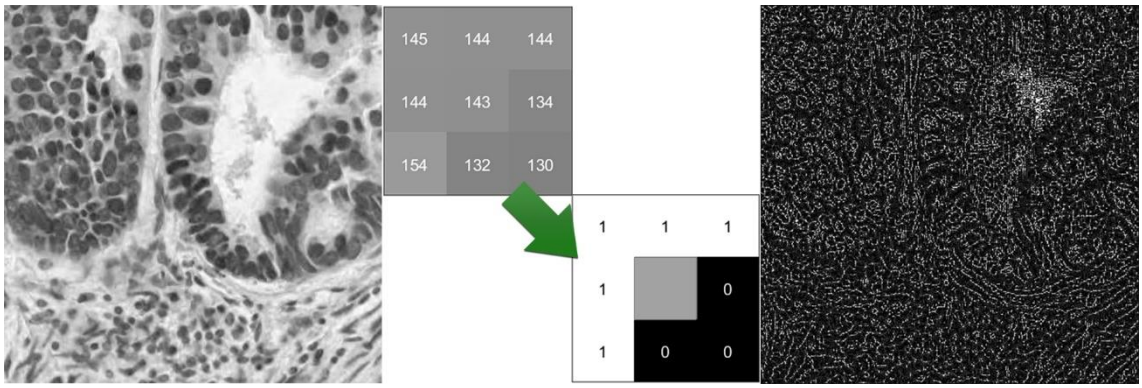


Figure 9 – Example of LBP operation performed on a single pixel neighbourhood

Left: Original greyscale image
Centre: Intensity values of pixel (i, j) with surrounding neighbourhood of size 8 and radius 1, and thresholded values, using pixel (i, j) as the threshold.
Right: Resulting LBP image

The LBP algorithm can be adapted to accommodate number of neighbouring values and radius size to act as a multi-resolution and rotationally invariant measure of texture [99]. Increasing the number of sampling points makes the resulting descriptor more accurate, at higher computational cost, and increasing the radius incorporates larger scales, but loses local information.

2.4.5 Segmentation

Image segmentation is the process of partitioning sections of a given image into multiple segments, typically into regions that represent the component objects of the overall image. There are many segmentation techniques that exist [100], and evaluation of such techniques is well documented [80-85]. A selection of segmentation algorithms is presented in this section, which may be useful for analysis of CRC images.

2.4.5.1 Thresholding

Thresholding is a simple and computationally inexpensive operation that segments objects based on intensity values falling above, below or equal to a threshold value. For a given intensity image $I(x, y)$, segmentation can be performed using thresholding described in Equation 7 to produce binary image $BW(x, y)$.

$$BW(x,y) = \begin{cases} 1 & I(x,y) \geq T \\ 0 & I(x,y) < T \end{cases}$$

Equation 7 - Binary image thresholding

Threshold T is used to determine which pixels in binary image BW are set to 1 and which are set to 0, so that foreground objects (pixels set to 1) can be analysed or further processed using connected components analysis (CCA) or binary image morphology.

Identifying thresholds that yield the best possible segmentation are critical for subsequent processing and analysis. However, due to the large amount of variation in CRC digital slide images (section 2.3.3), including non-uniform staining and background illumination levels across slides, identifying adaptive techniques that provide best case segmentations is non-trivial. One common method for solving this problem (known as Otsu's method) uses global clustering [107], which assumes an image is comprised of two classes with different intensity values, and the optimal threshold lies at the minima between the two peaks of the bimodal distribution represented by an intensity histogram. However, determining thresholds when minima between peaks are not well defined is difficult, and the concept of fuzzy sets is a proposed method to identify thresholds where boundaries may not be clear [108]. Furthermore, with respect to CRC images, multi-level thresholding may be required to partition images into foreground tissue and non-informative background, and then identify classes or objects within the foreground tissue. One method for achieving this is to remove background pixels that are a known value before applying a second segmentation. In this instance, the University of Leeds's Leica-Aperio scanners have a default background RGB colour setting of 240, 240, 240 [109], and so the appropriate threshold value for removing background pixels in an HSV intensity image would

be $T = 240$. The remaining values can be analysed to identify a second threshold between foreground and background tissue using an appropriate method.

$$DC_{Hf}(u,v) \in DC_H(x,y) \times BW_f(x,y) > 0$$

$$BW_{ft}(x,y) = DC_H(x,y) \leq \mu DC_{Hf}(u,v) - \frac{\sigma DC_{Hf}(u,v)}{2}$$

Equation 8 - Creating a foreground tissue mask

Equation 8 details this process, such that $BW_f(x,y)$ is the segmentation for foreground pixels, described in Equation 7, with $T = 240$. $DC_{Hf}(u,v)$ is the subset of foreground pixels in the Haematoxylin deconvolution channel $DC_H(x,y)$, using the foreground pixels in $BW_f(x,y)$ to remove background pixels. $BW_{ft}(x,y)$ is the resulting binary image segmentation of foreground tissue in $DC_H(x,y)$ when the thresholded using a calculation of the mean values of foreground Haematoxylin intensity, $DC_{Hf}(u,v)$ subtracted by half of one standard deviation of the same distribution $\sigma DC_{Hf}(u,v) \,/\, 2$.

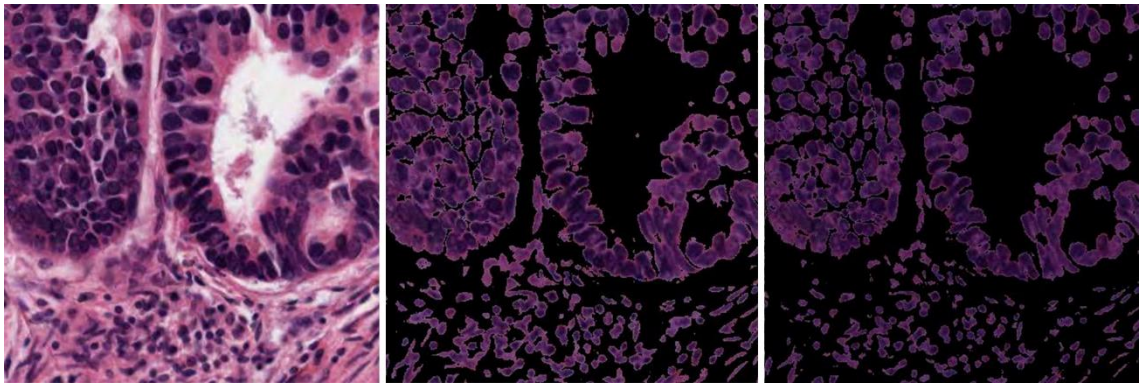


Figure 10 – Comparison of two thresholding methods, on CRC tissue

Both methods assume a bimodal distribution of foreground and background tissue, after background pixels have been removed.
Left: Original RGB CRC image
Centre: Thresholded image using Ostu's method on $DC_{Hf}(u,v)$
Right: Thresholded image using method from Equation 8 on $DC_{Hf}(u,v)$
The Otsu method shows that more pixels around the edge of the nuclei are retained using this method.

Figure 10 shows a comparison of methods using adaptive thresholding using the subset of pixels $DC_{Hf}(u,v)$. The segmentation is applied to all channels in the RGB image, and the original deconvolution result $DC_H(x,y)$ can be seen in Figure 6. It should be noted that thresholding and colour models may be applied in multiple colourspaces, or staining channels (using colour deconvolution).

2.4.5.2 Minima-seeded region-growing

Region-growing uses the concept of expanding seed points (pixels) into segmented areas by adding neighbouring pixels that fit a predefined criterion of similarity or homogeneity, such as difference in intensity, texture or (for deconvoluted stain channels) intensity of stain. This has the effect of grouping pixels based on their appearance. Seed points are calculated using salient parts of an image, depending on the desired segmentation result. In the case of CRC images, seed points may be chosen at pixels with the lowest intensity (areas of local minima), as they could represent the centres of nuclei (not including the nucleolus).

H-minima transform

The H-minima transform is a morphological method which supresses local minima in a given image [110]. The transform itself is performed using a non-linear filter NxN pixels in size, that sets each pixel to the lowest value of all pixels in the filter neighbourhood. The image result is modified with the value H added, which ensures that no value in the resulting image is lower than its original value. This creates larger areas of homogeneous minima, which are useful for seeding region growing segmentation algorithms.

Watershed algorithm

The watershed algorithm is a region-growing segmentation algorithm, which uses the concept of viewing intensity images as topographical maps [90,91], so that low intensity values are lower points (valleys), and high intensity values are higher points (peaks). By applying region-growing to areas of minima, the concept is likened to filling basins with water, until region boundaries meet. At this point a partition is created between the two, using the region's maximum intensity, representing a dam, or watershed, between two lakes, to prevent them merging. The resulting segmentation creates partitions between objects which are likely touching, and so is a useful tool for nuclear detection [113] The process is sensitive to noise, both over and under-segmentation, and is also dependent on regional minima representing the true number of objects in the image. Figure 11 shows an example of the H-minima transformed $DC_H(x, y)$ image, generating seeds for the watershed segmentation function, and the resulting segmentation partitioning the thresholded image $DC_{Hf}(u, v)$ from Figure 10.

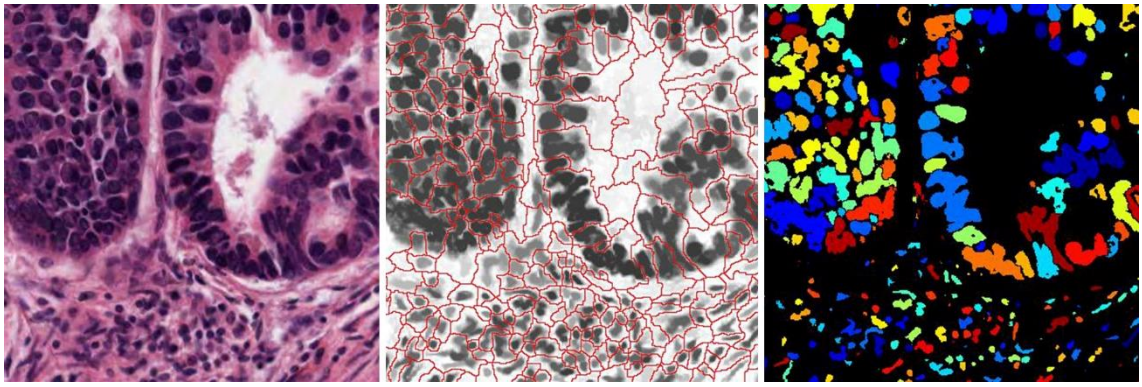


Figure 11 - Local minima and watershed segmentation of CRC image

Left: Original RGB CRC image
Centre: Image $DC_H(x, y)$, transformed by H-minima function, and watershed region boundaries
Right: Segmentation image $DC_{Hf}(u, v)$ from Figure 10 further partitioned by watershed boundaries
Note that the amount of clustered nuclei that have not been segmented using this method – this highlights that the number and distribution of seeds used for region growing are not sufficient.

It should be noted that the watershed algorithm, and other minima-seeded region-growing algorithms can be applied to pre-segmented nuclei (binary image) as a function specifically to separate touching objects. In this case, the seeds can be identified as minima in the complement (inverse) of the distance transform of the binary image, where the distance transform is a function that sets each pixel's value to the Euclidean distance in pixels to the nearest object.

2.4.5.3 Segmentation by clustering

Clustering is the process of grouping data points in an N-dimensional feature space, based on their similarity or distance. The feature space can apply to pixels, areas or objects for clustering pixels or segments into larger segments, and also to classify objects based on their feature similarity. This section focuses on clustering as a segmentation tool.

K-Means clustering

The K-Means clustering algorithm attempts to identify appropriate cluster groups based on the minimisation of the mean Euclidean distance between the prototype cluster centres and their associated data points [114]. The parameter k is set to determine the number of clusters to be used, and as such, the number of clusters to find needs to be known. For each cluster, a randomly selected start point is selected within the feature space, and data points are assigned to the nearest one. The mean of the distances for all associated data points to the cluster centre is calculated for all k clusters (hence K-Means). The position of the cluster centres is iteratively moved, so that the means can be recomputed to identify if the new positions reduce the mean

distance for all clusters. Because of this, K-Means clustering is thought of as a gradient descent algorithm. Equation 9 shows this formally, where S is the group of k data point sets, and μ_i is the mean of points in S_i.

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

Equation 9 – K-Means Clustering minimisation of sum of squares

K-Means clustering is a fast and efficient algorithm for segmenting images using a given feature space. Using this technique in various colourspaces is a common method in image processing, and may be useful for separating stain colours. However, the algorithm suffers from reliance on appropriate initialisation of the cluster centres, and as such, it can be considered good practice to run the algorithm several times to find the optimal solution. Also, since mean values are taken, the distance values are sensitive to noise.

Mean-shift segmentation

Mean-shift segmentation is another localised homogenisation technique using mean distance values to identify optimal clusters [115]. However, unlike K-Means clustering, Mean-shift does not have a random initialisation method, and does not require a predefined number of clusters. The algorithm analyses every data point in the feature space iteratively, and for each co-ordinate, all other points within a given radius are identified. The mean co-ordinate of all points within the radius is calculated, and this is set as the centroid for the next iteration, which repeats the calculation, until the convergence of all iterations in that cluster. The appropriateness of the radius size is dictated by the density of the clusters. For large datasets, a number of maximum iterations can be set (depending on the algorithm implementation).

Simple Linear Iterative Clustering (SLIC)

Simple Linear Iterative Clustering of superpixels (SLIC), is a localised clustering algorithm which, instead of identifying cluster centres in a given feature space, uses an regular grid of k seed points to create nearly uniform segmentations [116]. For each seed point, cluster centres are created in a five-dimensional space using LAB colour and x and y co-ordinates. The regular grid is deformed slightly by moving each centre to the co-ordinate with the lowest intensity value within an NxN neighbourhood. The surrounding pixels are then clustered to these new centres based on their values in the five-dimensional space, and connectivity is enforced. This method is efficient and computationally inexpensive [117], and may be useful for segmenting complex structures in CRC tissue.

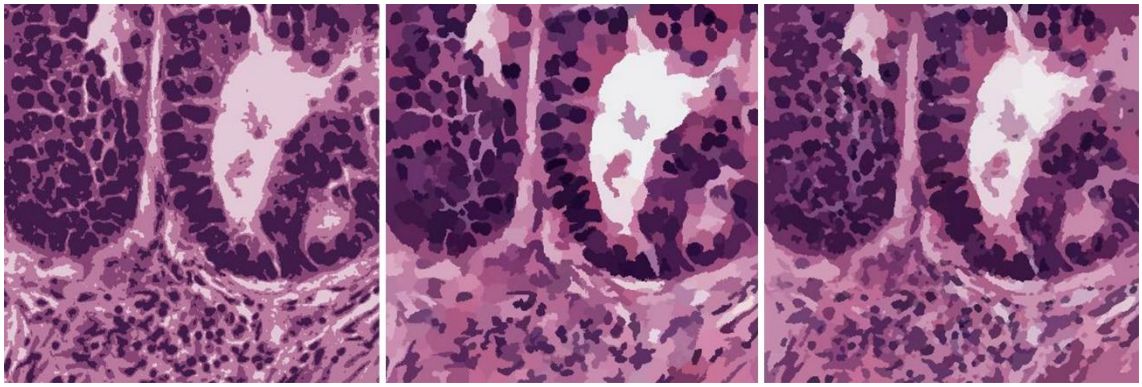


Figure 12 - Example of clustering methodologies on CRC tissue

Left: K-Means using 4 clusters
Centre: Mean shift with radius of 10 and minimum segment size of 100 pixels
Right: SLIC using a compactness value of 10 and merging radius of 1

Figure 12 shows examples of the three clustering algorithms applied to CRC tissue, using input parameters that yield visually similar results. Note that the clusters are coloured by their mean RGB values.

2.4.5.4 Graph-cut segmentation

Graph-cut segmentation models images as graphs, where graph G contains nodes or vertices V, that represent pixels in the image, and edges E, that are weighting functions called costs, using the similarity between the nodes. This is modelled as $G = \{V, E\}$. Removing enough edges in a graph to create two separate partitions is known as a cut, which has a calculated cost. Given two partitions A and B, the cost of the cut can be calculated as the sum of the edge costs $w(p, q)$, shown in Equation 10.

$$cut(A, B) = \sum_{p \in A, q \in B} w(p, q)$$

Equation 10 - The graph-cut cost function

Min-cut segmentation

The min-cut (or max-flow) method identifies cuts with the minimum cost value, as the most appropriate segmentation [118]. However, since the cost is a summation of all the edge weights between A and B, this method inherently penalises longer cuts, therefore is prone to making smaller partitions that are not necessarily optimal.

Normalised cut segmentation

Normalised cuts addresses the bias in the min-cut algorithm by normalising the sum of the costs, dividing them by their association [119]. Equation 11 shows that for a graph cut of two partitions (A, B), their normalised cut cost value is the sum of the cost value divided by the sum of all vertices touching A, and the cost value, divided by the sum of all vertices touching B.

$$Ncut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)}$$

Equation 11 – The Normalised cut cost function

Where $assoc(A,V)$ is the total connection from the vertices in A to all nodes in the graph, with $assoc(B,V)$ being respectively the same. Finding the minimum cost for cuts in this manner is computationally prohibitive, and therefore an approximate minimisation of the $Ncut$ value is made by solving a generalised eigenvalue problem.

Figure 13 - Comparison of graph cut methodologies on CRC tissue

Left: Original Image
Centre: Min-cut method partitioning into 16 (non-contiguous) regions
Right: Normalised cuts method partitioning into 16 (contiguous) regions
The min-cut example illustrates that partitions are made using cuts that have the lowest cost, and do not take into account segment size. The segments from the normalised cut method are more consistent in size.

Figure 13 shows a comparison between the min-cut and normalised cuts methods in CRC tissue, using a target number of 16 graph partitions. It is evident from these examples that the min-cut method prioritises cuts that are shorter in length. Note that the segments are coloured by their mean RGB values.

2.4.6 Supervised learning

It should be noted that the distinction between segmentation and classification is not always clear, due to some segmentation methods defining their own prototype classes based on similarity, homogeneity or distance, and grouping pixels by those metrics. In this respect, clustering may be considered a form of unsupervised learning, in that the clusters are made without any input on what they should be. This section focuses on classification of images using expert-labelled images and features to train machine learning (ML) algorithms.

2.4.6.1 Image features

Image features are descriptors that parameterise a specific characteristic over an entire image, or at a particular location within an image. Features are often continuous variables, but are not limited to single values, and so can be histograms, vectors, matrices, etc. Features are selected for their ability to discriminate between classes, and therefore the accuracy of the classifier is dependent on the features chosen. If x is a single feature, a set of n features is described as

$$f = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$$

and a feature set is represented as $S = [f_1, f_2, \ldots f_n]$ where each vector f has a classification label. The desired outcome is a trained classifier using f to predict its classification, independent of the label. Using regression and model fitting techniques, this can equate to a probability that the feature vector belongs to a class, for example, tumour: $P(tumour \mid f_n)$. In the case of binary predictions, the output value is thresholded (typically at 0.5), and in multi-class predictions, the highest value is selected. There are a number of prominent supervised learning algorithms that use feature vectors and training labels to generate statistical models representing object classes [120]. Some of these are explored in this section.

2.4.6.2 Naïve Bayes classifier

A Naïve Bayes classifier (NB) is based on Bayes Theorem, and uses conditional probability of the frequency of classifications in the training data (class labels), given some predictive and independent observations (features), to predict outcomes [121]. This makes the predictor unable to function when given unseen values, which results in a probability outcome of zero (known as the zero-frequency problem), and is solved using a smoothing function such as Laplace estimation [122]. For a given set of observations, their probability of resulting in a specific class

is modelled, based on the combined probability of the observed outcomes in the training data, normalised by the total frequency of the observed event occurrences (see Equation 12).

NB can be used for categorical or continuous observations, where, an assumption is made that continuous observations are normally distributed, which would not be the case in most CRC image features. If the classifier uses the assumption that a threshold of 0.5 separates the data appropriately, the probability an image segment being tumour, given two continuous values thresholded at 0.5, can be computed as

$$P(tumour \mid f) = \frac{P(\, x_1 \mid tumour)\, P(x_2 \mid tumour)}{P(x_1)\, P(x_2)}$$

Equation 12 – Naïve Bayes classifier computing probability of tumour, given features x_1 *and* x_2

where x_1 and x_2 are features in feature vector f. The features correspond to the mean intensity outputs from colour deconvolution in the haematoxylin channel $\mu DC_H(u, v)$ and the eosin channel $\mu DC_E(u, v)$ from $I(u, v)$, which is a subset of the whole image $I(x, y)$. The corresponding probability calculation is computed for the probability of those values being stroma, and the classification is selected based on the highest value.

The NB is known as a bad estimator, as it is limited by the severe assumptions that it builds the model on, making the algorithm inflexible, affecting accuracy [123]. However, due to this simplicity, NBs are computationally inexpensive. As such, it may be useful to use an NB as a benchmark to compare against other ML algorithms for CRC image data.

2.4.6.3 Logistic Regression as a classifier

Logistic regression uses linear regression in conjunction with a sigmoid (logistic) function to model class features, and fit a probability distribution between zero and one. For each feature in the feature vector, regression is used to identify a model that fits the distribution of data between one observed variable (feature) x_i and the dependent variable y. The generated regression models are straight lines with an intercept and a gradient, that have the capacity to predict values above and below zero. The model is normalised to fit a continuous scale of zero to one using the logistic function, shown in Figure 14.

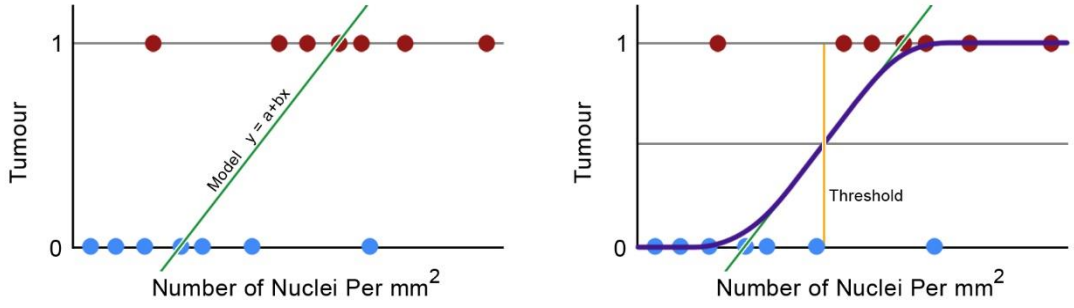


Figure 14 - Logistic Regression classification of tumour using one feature

Left: Linear regression model applied to the binary dataset (extending beyond 0 and 1 in the y axis) Right: The resulting model with the sigmoid function applied, and a threshold identified for number of nuclei per mm² where y = 0.5

Using the model, the probability of y (tumour) can be made from the feature x_i (number of nuclei per mm^2), and a classification threshold for splitting the data is typically determined at y = 0.5. Logistic regression is a computationally efficient method for modelling features, however, the linear model means that more complex non-linear data cannot be accurately represented. Also, the method assumes that there is no error in the output variable when training and is therefore sensitive to noise.

2.4.6.4 Support Vector Machines (SVM)

Support Vector Machines (SVM) conceptualise multi-dimensional space in order to attempt to create partitions between clusters within it [124]. For each element x in a given feature vector f, the value is treated as a co-ordinate in one dimension, such that a feature vector of size n can be plotted as a single point in n dimensional space. The classifier uses an optimisation function to identify a hyperplane for splitting the n dimensional data into two (binary) classes, by maximising the distance (margin) between the hyperplane and the nearest points (support vectors) in the feature space. Figure 15 visualises this concept for two features, $x_1$and x_2, showing examples of successful hyperplane partitions, each of which yields the same classification result, and the optimal hyperplane, which maximised the margin between the hyperplane and the support vectors.

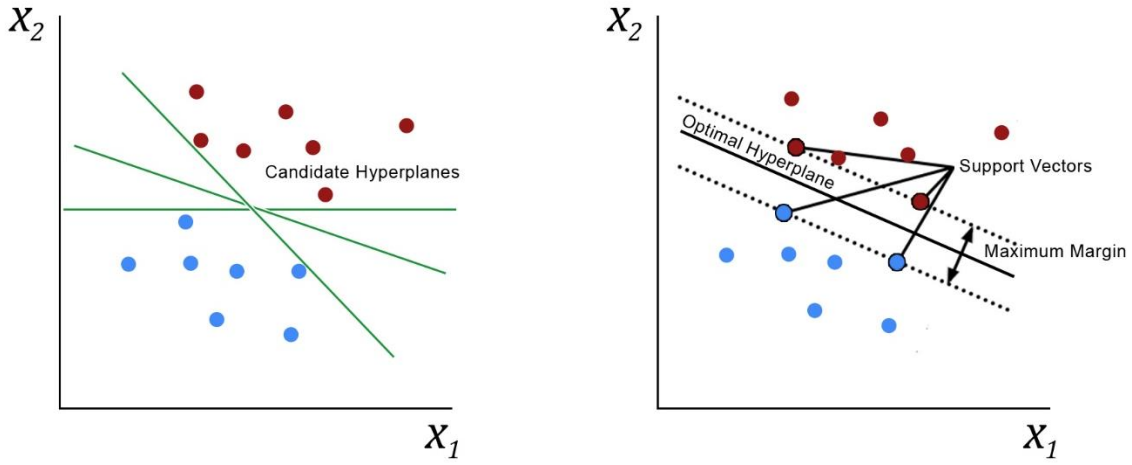


Figure 15 - Optimising the hyperplane margin for SVM classification

Left: Examples of multiple successful candidate hyperplanes
Right: The optimal hyperplane for the data points, by maximising the margin between the hyperplane and the nearest data points (support vectors)

Traditional SVMs provide binary classifications, which may be useful in CRC for separating tumour and stroma tissue, however, multi-class SVMs can be developed using different methodologies. One of the simpler methodologies uses multiple binary SVM classification models, and combines the predictions, making multiclass SVMs more computationally expensive [125]. SVMs also work on non-linearly separable data, by extending the optimisation problem to incorporate misclassification as a new variable that should be minimised. By extending the SVM implementation to classify non-linearly separable data, complex data such as CRC tissue features may be appropriately modelled [126].

2.4.6.5 Decision trees and Random Forests (RF)

The concept of decision trees is simple and widely used for both classification and regression, whereby classification trees generate categorical dependant variables, and regression trees output continuous dependent variables [127].

Decision trees

Classification And Regression Tree (CART) analysis is conceptually similar to a flow diagram, where a binary decision tree represents a Boolean function, using feature values as input to split the data into two groups [128]. Binary trees are made up of multiple nodes, whereby each node represents a test to perform on a given attribute. In random decision trees, the attribute used for splitting is chosen at random from the training dataset, and the threshold used to make the split is generated using regression to minimise the error in the resulting class predictions. The tree begins at a single (root) node, and the first split is made. Subsequent nodes split the data further,

and stop splitting the data when some stopping criteria has been met, such as the number or proportion of data points in the resulting class distribution. A node that reaches the stopping criteria is known as a terminal node, or leaf node, and each leaf node is assigned a classification or a number, depending on the type of CART analysis. For classification trees, the prediction class is set to the classification found most frequently in the resulting subset of data. Figure 16 shows the visual representation of a 2-class dataset with two splits on the left, and the equivalent decision tree process on the right.

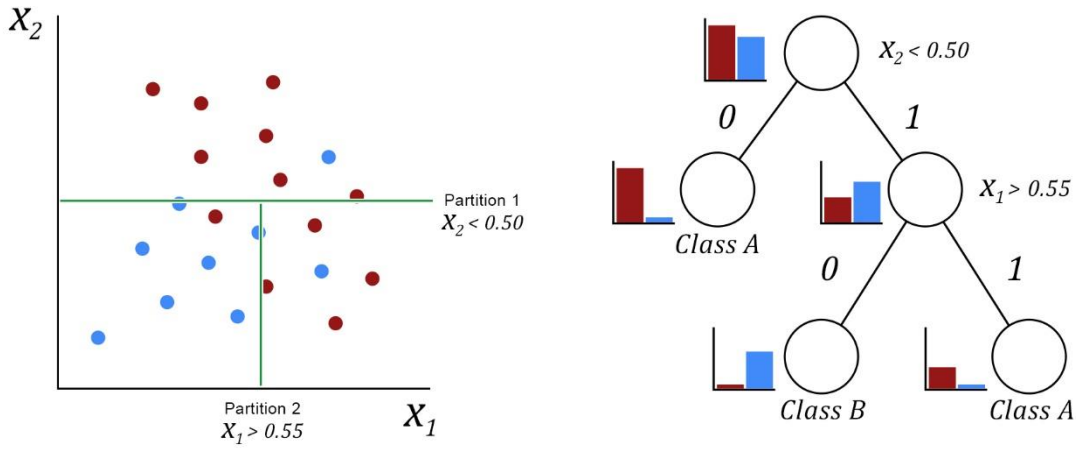


Figure 16 - Binary tree classification of 2-dimensional feature data

Left: Pre-classified two-dimensional data, with two partitions
Right: Binary tree representation with the resulting class distributions at each node

Note that in this example, optimum performance of the algorithm is dependent on splitting x_2 before x_1. The process of splitting the data is known as growing the trees, as the splits occur independently, and are not limited to a specific number of nodes. The process of fully growing a decision tree means that all leaf nodes in the tree have reached their stopping criteria. This typically generates large trees that are too specific for application to other datasets, or have over-fitted the training data. Pruning is a process that reduces the complexity of the fully-grown tree, in an attempt to create a more generalised model capable of being applied to other data. Pruning can be achieved by removing nodes, based on the effect the removal has on the overall error rate of the predictions, as well as more complex methods [129]. Decision trees are a useful and intuitive ML tool, that is unaffected by non-linear data. However, they are prone to overfitting training data, and can be sensitive to variance, to the extent that noise in the data will change the entire structure of the tree. Decision trees are also sensitive to biased data, since the stopping criteria will be met sooner, if the training data is imbalanced. The concept of using multiple trees as an ensemble classifier attempts to mitigate these issues, and Random Forests is one of the most commonly used methods.

Random Forests

Random Forests (RF) is a ML technique based on the concept of bootstrap aggregation (bagging) – that multiple weak predictors can combine to form a more robust one, by averaging noisy models to create a model with low variance [130]. In RF, a collection of n trees are grown, using randomly selected subsets (with replacement) from the training data. For each node in each tree, a predefined number of m features are selected at random, and the feature that provides the split of data with the lowest error is chosen. Each of the n trees are fully grown without pruning, and for each feature vector tested on the fully trained classifier, every tree is used to make a prediction. The predictions made by each of the classification models are aggregated as votes, and the modal average is used to select the final classification. For regression, the mean value of all predictions is taken.

Since each tree in the RF uses a subset of the feature set, the unused portion of the dataset, known as the out-of-bag samples, can be used to calculate the error rate of the model, and internally optimise the RF result by minimising the out-of-bag error [131].

Because of this methodology, RF is known as an ensemble classifier. Since the predictions are aggregated over n number of trees, the issue of overfitting is reduced, and so pruning is not necessary. This means that highly specific trees can be used and datasets with high dimensionality can be computed in the same manner. However, there is very little control over how models are generated [132], and the process of generating the trees is still sensitive to imbalanced data (data containing more examples of one class than another), although this can be mitigated using balanced sampling of the dataset, preproduction of feature vectors in the minority class, or applying a cost function to the dataset, based on the proportion of data in each class [133].

RFs are regarded as a computationally efficient and flexible solution for ML solutions, and have the capacity to learn classes with complex and high dimensional feature space, which makes them appropriate for application to CRC image data. However, independent analysis should be taken to assess whether RF are the most appropriate solution for learning CRC features [134].

2.4.6.6 Boosting

Boosting is an umbrella term for improving a given ML algorithm through combining multiple classifiers to make a more accurate prediction [135]. Like bagging in RF, boosting is an ensemble classification technique, used to average over noise and variance in the multiple models that have been generated. Boosting differs from bagging in that the output generated from each classifier is weighted, based on the out of bag error rate that each classifier generates.

For k classifiers, each ensemble classifier e is weighted by its error rate w. Equation 13 shows the difference between bagging and boosting calculations.

$$e = \frac{1}{k}\sum_{i=1}^{k} e_i \qquad e = \sum_{i=1}^{k} we_i$$

Bagging Boosting

Equation 13 - Bagging vs boosting

Boosting may also include an extra step, which is to discard entire classifiers if their error rate is too high, such as the AdaBoost method, which uses a cut-off of 50% [136]. Boosting is a useful technique for further avoiding bias in ensemble methods, by removing classifiers which contain higher levels of bias or noise. However, if the models are overfitting, then higher weighting is given to the models with the higher accuracy, and so bagging is often more effective.

2.4.6.7 Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) utilise the concept of biological neural networks to model a web of classifiers that resembles interconnected neurons in the brain [137]. An ANN consists of a graph structure, where neurons are represented by nodes, and connections are represented by edges. The network is structured in layers, with an input layer, multiple hidden layers and an output layer (see Figure 17).

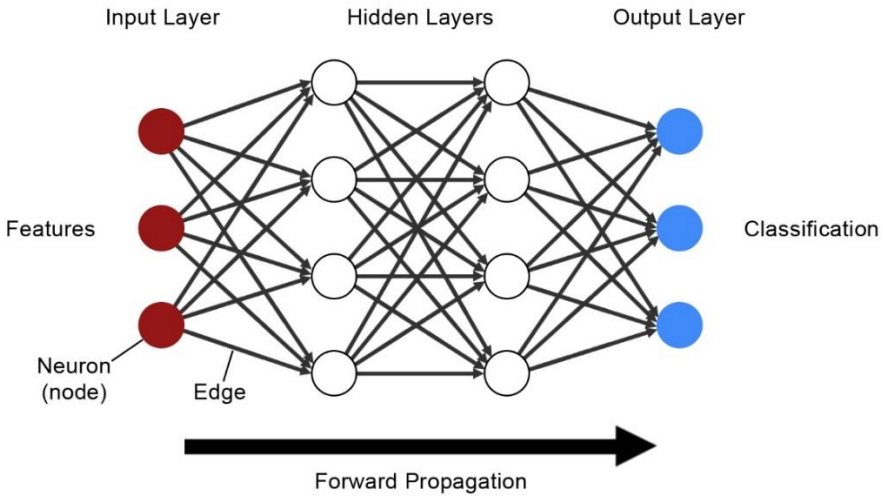


Figure 17 - Visual representation of an Artificial Neural Network (ANN)

Each neuron in the network represents a classifier, and edges model the relationships between the neurons in different layers. Classification results are passed forward (forward propagation), starting with features at the input layer, passing the output from each neuron to the next hidden

layer, until classification results are returned by the output layer. The classification is compared to the ground truth training data, and a cost is generated, representing the difference between the ground truth and the classification. The network is optimised using edge weights, and neuron biases, with the aim of minimising this cost [138].

The optimisation process incrementally alters weights and biases in the network over many training examples, and assesses the impact on the classification result. The rate at which changes to the weight or bias affects the cost is known as the gradient [139]. A higher (steeper) gradient indicates that changes to the weight or bias have a larger effect on the cost, and a low (shallow) gradient requires larger changes to the weight or bias to have the same effect. As the changes are performed iteratively, functions with higher gradients take less iterations to optimise, reducing training time.

The optimisation is performed backwards through the layers (back propagation) [140], so that a gradient at any point in the network is the product of all previous gradients up to that point. Therefore, having more layers in the network leads to exponentially smaller gradients in earlier layers. This ultimately means that changing weights and biases in these layers has very little effect on the cost, and takes much longer to optimise. This problem is known as the vanishing gradient [141], and has limited the use of large ANNs in practice until the advent of Deep Learning.

2.4.7 Deep Learning (DL)

Deep Learning (DL) is an umbrella term for a range of ML algorithms that aim to improve traditional specific classification algorithms, requiring predefined image features and large amounts of hand-labelled training data [142]. The architecture of DL algorithms is structurally similar to ANNs, but uses Restricted Boltzmann Machines (RBMs) to optimise layers in the network individually. By automatically defining features, independent of human intervention (and therefore error and biases), training data is partitioned into unlabelled sets, which can then be refined into labelled classes using relatively few hand-labelled examples [143]. Two common DL methodologies exist for image processing which are called Deep Belief Networks and Convolutional Neural Networks. These are discussed in the following sections.

2.4.7.1 Unsupervised learning using Restricted Boltzmann Machines

Restricted Boltzmann Machines (RBM) are shallow, two-layer restricted networks, that process data via both a forward and a backward pass in attempt to reconstruct the original input data [144]. The network is restricted in that no two nodes of a single layer are connected. Input

feature data is processed using weights and biases, in the same way as ANNs (section 2.4.6.7), but instead of generating output predictions and costs, the data is fed backwards, in attempt to reconstruct the input features (see Figure 18).

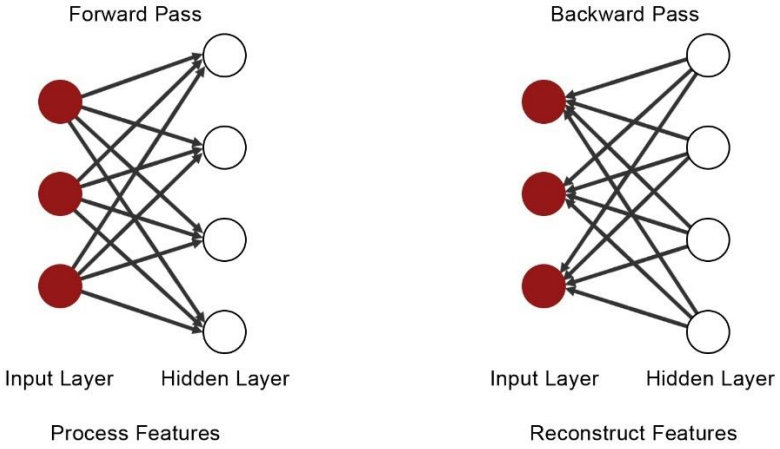


Figure 18 - Restricted Boltzmann Machine (RBM) structure and behaviour

The reconstructions are directly compared to their input features, using Kullback-Liebler (KL) divergence (relative entropy), so that the weights and biases can be optimised, as opposed to using the ANN method of calculating costs between output and ground truth. This means that the data does not need to be labelled, and therefore human classification error is eliminated. Also, features that do not reconstruct adequately after optimisation can be discarded, making the RBM a type of feature extractor, called an autoencoder [145].

2.4.7.2 Deep Belief Networks (DBN)

Deep Belief Networks (DBN) are structurally similar to ANNs, but used stacks of RBMs to internally pass self-trained RBM outputs from their respective hidden layers to the input layer of the next RBM [146]. This means that each layer is optimised before passing the output to the next layer, and so does not require back propagation, eliminating the vanishing gradient problem. By circumventing the vanishing gradient problem, DBNs can stack many layers in the (deep) network without requiring exponentially more resources to optimise weights and biases in earlier layers. The result is an optimised set of layers that improve in accuracy as data is passed along the network.

Since features are automatically selected, output classes from the network are initially unlabelled, and so require small examples of labelled input data to be classified. This small training set can also be used to make final adjustments to the weights and biases in the network, and increase accuracy further.

2.4.7.3 Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNN) are inspired by work by Hubel and Weisel (1968) describing the architecture of monkey striate cortex, which categorises cells attributed to visual sensory information in a hierarchical structure [147]. Cell functions are described as simple, complex or hypercomplex, referring to the processing of simple lines and edges, and combining them into shapes, ultimately combining those shapes into complex objects. Receptive fields are defined as small areas of the overall visual field, where processing of such an area will activate neurons based on its properties. The paper discusses that smaller receptive fields are associated with more highly developed brains (comparing monkeys to cats), and hierarchical processing of a visual field requires complex layered connections of many receptive fields.

Local connections with receptive fields

As opposed to traditional NN architecture, neurons in CNNs are not connected to every neuron in the next layer, as performing such operations on individual, or small groups of pixels would be computationally prohibitive when processing large images. Instead, images are processed using a two-dimensional window of a predefined size, with each window being referred to as a receptive field. Each receptive field is processed by one node in one layer of the network, using the three types of processing layers, convolution, transformation of the resulting images using a non-linear function such as Rectified Linear Units (ReLU), and pooling [148], illustrated in Figure 19. Note that the order of these layers is not limited to the order shown.

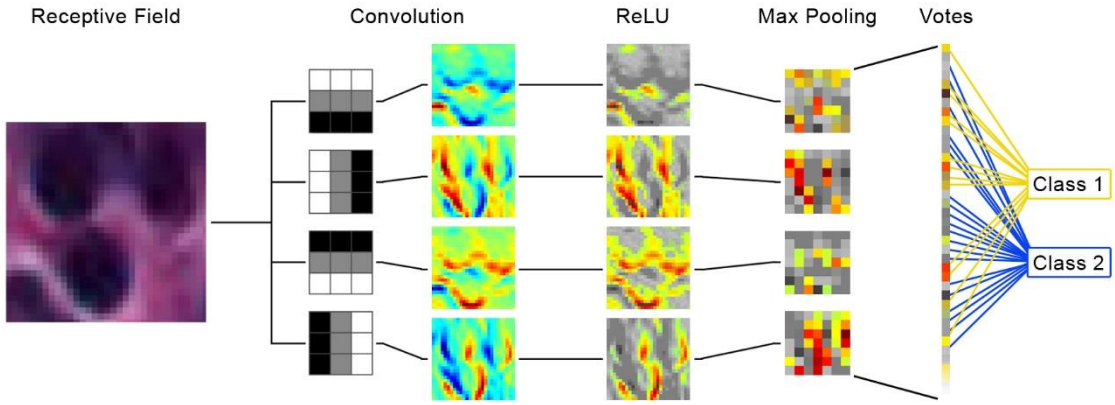


Figure 19 - Visualisation of the types of layers in a Convolutional Neural Network (CNN)

The illustration shows a receptive field containing CRC epithelial nuclei, which is processed by four rotated variants of the Prewitt edge detection convolution kernel (see 2.4.3), with the resulting four images being normalised by Rectified Linear Units (ReLU), and reducing the result using max pooling. The pixels in the output images after max pooling are reshaped into a list and their values are used for class voting.
Note that these images are for illustrative purposes only – pixel values are colourised using a heatmap, receptive fields are typically smaller, and filters are randomly generated.

Convolutional layer

A number of convolution kernels (also referred to as filters or weight matrices) are randomly generated and applied to the receptive field (see 2.4.3 for spatial filtering). This results in as many output images as there are filters. The filters are initiated randomly, and RBM autoencoding selects the most appropriate ones for identifying reproducible edges and basic shapes [149]. The result is a set of convolution filters that model generalisable low-level image features that can be used to identify basic properties of the receptive fields. Note that the filters in Figure 19 are rotated variations of the Prewitt filter, described in 2.4.3, and are for illustrative purposes only.

Rectified Linear Unit (ReLU) layer

Convolution filters in CNNs have a randomly generated set of weights with a numeric range of -1 to 1. Applying the filters to an image therefore may result in output images with negative values. Approximating the values to Rectified Linear Units (ReLU) is a form of normalisation, described by the formula $x = \max(0, x)$,which simply converts all negative numbers to zero [150]. The process is also referred to as activation, where the resulting image is called an activation map.

Pooling layer

Pooling reduces the size of the convolution result or activation maps by taking a single value of a sliding window of a predefined size, and iterates over an image with a given step size. Max-pooling is most common in CNNs due to computational efficiency, where for each window, the maximum value in that area is taken. Pooling is applied to reduce computational overheads and to prevent overfitting.

Fully connected layer and voting

The final layer in the CNN converts the resulting max-pooled images into a discretised list of numbers, and their values are used for voting for a class. The numbers are discretised using a soft-max function (similar to the sigmoid function in logistic regression, in section 2.4.6.3), and the resulting values are used as votes for classifications, typically using argmax to get the most likely class.

CNN structure and hierarchical layering

The structure of a given CNN can be altered by changing the order of, and by stacking multiple layers, depending on the design choices made by the programmer. By applying multiple sets of layers, activation images begin to contain models that represent higher level shapes and objects.

Regularisation

Also, referred to as dropout, regularisation randomly turns neurons off (sets weights to zero), to force the network to learn multiple representations of the same thing, increasing the generalisation of the trained network, and reducing overfitting.

2.5 Application of image analysis to digital pathology slides

This section provides an overview of currently developed solutions for automating histopathology image analysis, and identification of the successes and limitations of those solutions.

2.5.1 Overview

With the advent of digital slide scanners and high-performance computing in the last 15 years, research in the field of image analysis on histopathology images has substantially increased. [43,44,130-136]. This section highlights some of the current state-of-art applications and technologies in this area, with a focus on how these technologies might be used for CRC analysis.

2.5.2 Stain (colour) correction

The variable appearance of digital slides can be affected by biological, histological or digital issues, at all levels of the digital slide production workflow, discussed in section 2.3.3. Colour normalisation is the process of transforming the original image colourspace, for digital enhancement or standardisation, using a set of reference values to normalise to [158]. The normalisation process is traditionally based on the assumption that the colours can be corrected via a linear transform, or rotation of the colourspace, which may be derived using colour deconvolution [138-141], although non-linear methods also exist [142,143]

Magee et al present a method for normalising colour using context in addition to colour deconvolution [165]. The method takes a 'target image' as input, so that the colour values from it can be mapped to a 'transform image', that is to be normalised. Initially a classifier is trained on an unrelated dataset, using feature vectors calculated per-pixel. The vectors consist of the RGB values, combined with a 'context vector' based on a low-dimensional image colour histogram of the whole image. For training, vectors are calculated and assigned to one of three

classes: stain 1 (haematoxylin); stain 2 (eosin or DAB); background. The classes are assigned to pixels via manual segmentation and labelling. The labelled feature set is applied to ML, using an implementation of the Relevance Vector Machine method (RVM), which uses a methodology similar in principle to SVMs, but outputs probabilistic classification [166]. One RVM is trained on each of the three classes, so that predictions on unseen images output three probability values per-pixel, and each of the three values is normalised according to Equation 14.

$$P(Class_n \mid f) = \frac{P(R_n|f)}{\sum_{i=1}^{k} P(R_i|f)}$$

Equation 14 – Normalisation of RVM pixel class predictions

Where $Class_n \in \{Stain1, Stain2, Background\}$ and $P(R_n \mid f)$ is the probabilistic output from the RVM trained on pixels from $Class_n$ (with the other two classes being grouped as the negative examples), given the combined feature vector f, and k is the number of elements in $Class$.

RGB values from pixels classified with a probability value above a predefined threshold (reported in the paper as 0.99) are used to generate the OD vectors for colour deconvolution (see section 2.4.2). If less than 200 pixels satisfying this criterion are present in the image, then pixels with the top 200 probability values are chosen – this should be noted when reviewing the example images in Figure 20 and Figure 21.

Using the colour deconvolution methodology described in 2.4.2, both the target and the transform images derive OD values for independent colour deconvolution, using the classified pixels in each image. It is noted that OD values need to be derived independently, due to the variation of stain that may exist (discussed in 2.3.3). For both target and transform image, the three staining probability maps are generated, and pixels are classified, based on Algorithm 1.

```
for each stain channel n
  for each pixel p
    if all RGB OR DC channels in n(p) < t(black)
      Class = Black
    elseif n(p) > t(white)
      Class = White
    elseif RVM P(background) > t(background)
      Class = Background
    elseif RVM P(stain(n)) > t(foreground)
      Class = Stained
    else
      Class = Other
    end
  end
end
```

Algorithm 1 - Classification of pixels using RVM probability and DC channels

By classifying pixels into these groups, the black and white classified pixels can be ignored so that they remain unchanged when mapping. Statistics are generated from the remaining pixel classes (mean, and upper and lower 5th percentiles), which are used to map the transform image values to the target image values. Finally, the mapped values are transformed back into RGB using the colour deconvolution transform method. Figure 20 shows examples of original transform images and their normalised outputs.

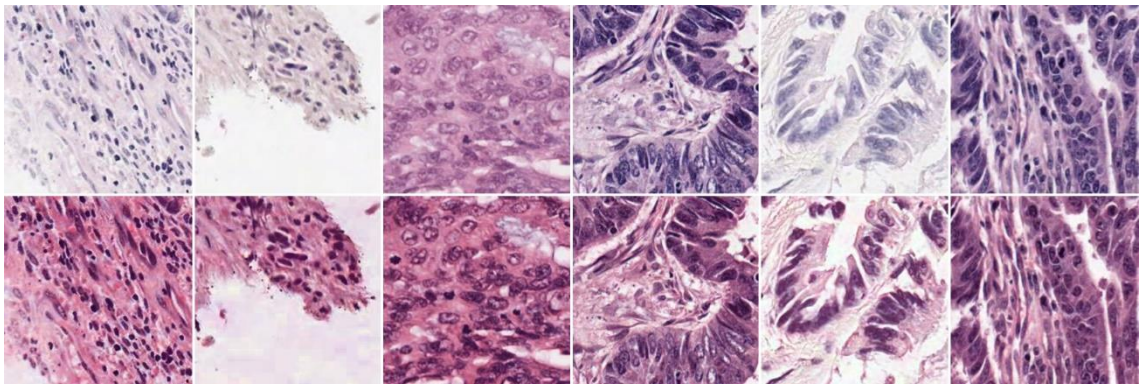


Figure 20 – Example of colour normalisation on six CRC images

Top: Transform images
Bottom: Normalised images
The normalisation process transforms the colour of each image to the colours of the target image (not pictured).

It is important to note that (as mentioned previously) if less than 200 pixels of a given class can be found in each image using the combined RVM classification method, colour deconvolution OD vectors will be generated from 200 pixels with the highest probability of that class.

Therefore, digital slide images exhibiting characteristics outside of expected colour ranges due to issues discussed in section 2.3.3, will not adequately be compensated for using normalisation techniques. Figure 21 shows examples of images too extreme in terms of weak staining, causing colour artefacts, or simply causing the algorithm to fail. Also, note that if large areas of lumen dominate the image, then the algorithm identifies background as foreground, and normalises it.

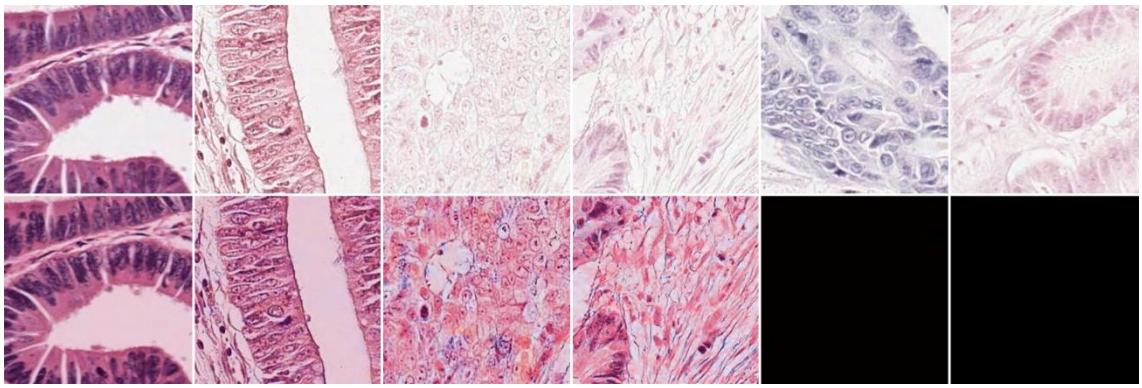


Figure 21 - Example of colour normalisation underperforming on six CRC images

Top: Transform images
Bottom: Normalised images
In these examples, there is not enough colour information in the original images to transform to a meaningful result.

Five of the six images in Figure 21 are extreme examples of weak staining that are part of the dataset presented in Chapter 3, and identified as not scorable in pathologist interaction studies in Chapter 4. It is important to consider that correction of stain does not account for other artefacts such as tissue folds, or tears, and identification of such artefacts may be useful for application to digital slides before normalisation and subsequent analyses [167].

The Magee et al colour normalisation algorithm is used in the image processing algorithms presented in Chapter 3.

2.5.3 Nuclear and cellular detection in CRC images

In CRC analysis, the ability to delineate between boundaries of tissue types and cells is a prerequisite to classification, staging, grading and diagnosis of disease. In automated analysis of CRC, the segmentation task is typically split into two disciplines – segmentation of cells and nuclei, and segmentation of tissue structures such as glands, vessels and nodules.

Cell and nuclear segmentation is a discipline over 50 years old [5]. Methods of segmentation have changed with imaging modalities and technological advances, yet the problem is not trivial

to solve, due to the variable appearance and characteristics of histopathological tissue (see section 2.2 and 2.3). CRC tissue contains various types of cells and nuclei which have different visual characteristics, and the detection of these cells typically involves segmentation or separation, and classification based on appearance. This separation is non-trivial due to the deformable size and shape of the nuclei, as well as touching and overlapping objects. Typical solutions developed for the problem of nuclear detection include: thresholding, morphological filtering and connected components analysis (see section 2.4.5.1) [147-149]; texture-based segmentation (see section 2.4.4) [150,151]; watershed transform (see section 2.4.5.2) [89,92,152-155]; clustering (see section 2.4.5.3) [176]; graph cuts (see section 2.4.5.4) [157,158]; active contours and deformable model fitting [159-161]. Recently, state-of-the-art technology known as Deep Learning has also been used to identify nuclei [181]. One issue with most reported methods is a lack of standardised benchmarks for evaluation, as solutions are typically only developed and tested on datasets that researchers have access to [182].

2.5.3.1 The Bennett et al algorithm

Bennett et al [183] present a nuclear detection algorithm based on the Hough transform [184], extending the parameter (and subsequently accumulator) space to detect more complex shapes [185], so that nuclei edge segments can be joined by the highest probable lines in the accumulator space [186]. Initially, colour deconvolution (see section 2.4.2) is applied to digitally separate the H&E stains, and the haematoxylin (nuclear) staining channel is used for analysis. The Canny edge detection filter (see section 2.4.3) is applied to generate a binary image of disjointed nuclear edge segments. Edges that are within distance d of another edge are joined recursively until the merged edge is greater than length l, and then any edges left that are shorter than length s are discarded.

To account for the varying elongation and rotation of nuclei, the Hough transform is applied multiple times with different hypothesis orientations. Hypothesis orientation is quantised into 4 angles, $\alpha = \{0, 45, 90, 135\}$, to reduce the number of iterations and increase computational efficiency. For each edge segment, only the centre pixel is plotted in a 5-dimensional parameter space, using x location, y location, nuclear aspect ratio, edge directions, and the hypothesis orientation, and the voting direction is specified towards the negative gradient direction, such that:

$$-\frac{\nabla I(x,y)}{\|\nabla I(x,y)\|} = -(\cos(\theta(x,y)), \sin(\theta(x,y)))$$

Equation 15 – Hough transform for edge segment voting in the negative gradient direction

where θ is the angle of the image gradients $\nabla I(x, y)$, with respect to the x axis. The voting space for each pixel is created using a pyramidal kernel with voting directions V_θ and scale r_{min} to r_{max}. The votes within the kernel area are weighted by a 2D gaussian kernel centred over the centre point of the pyramid kernel. Figure 22 visualises this voting space.

The process is repeated for each of the hypothesis orientations, and votes are aggregated from all four accumulator spaces. Centroids are then identified using the mean shift algorithm (see section 2.4.5.3). To reduce over detection, centroids that do not overlap binary objects in the haematoxylin channel image, thresholded by Otsu's method (see section 2.4.5.1), are discarded.

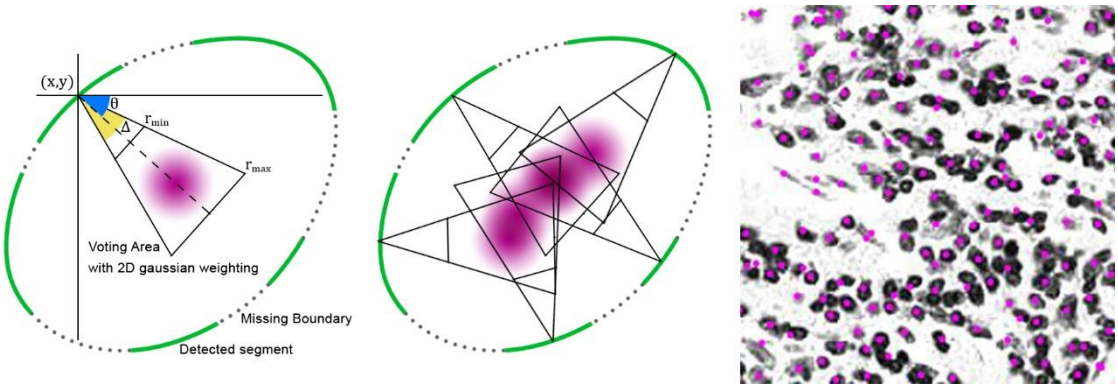


Figure 22 – Pyramid voting kernels for nuclear segment centroid voting

Left: Pyramid kernel with (magenta) 2D gaussian filter for one edge segment
Centre: Combined pyramid kernels for all segments showing centroid probability
Right: Detected nuclear centroids on deconvolution haematoxylin image after mean shift clustering

Edges are grouped by the centroids that they voted for, and the segments are then combined using ellipse fitting. The process joins the edges, minimising the distance measure of the set of points, using least squares error minimisation. Finally, some of the false positive detections are removed by identifying overlapping segmentations, and identifying the levels of haematoxylin inside the segmentation boundaries, compared to outside the boundaries. Figure 23 shows results from the algorithm detections.

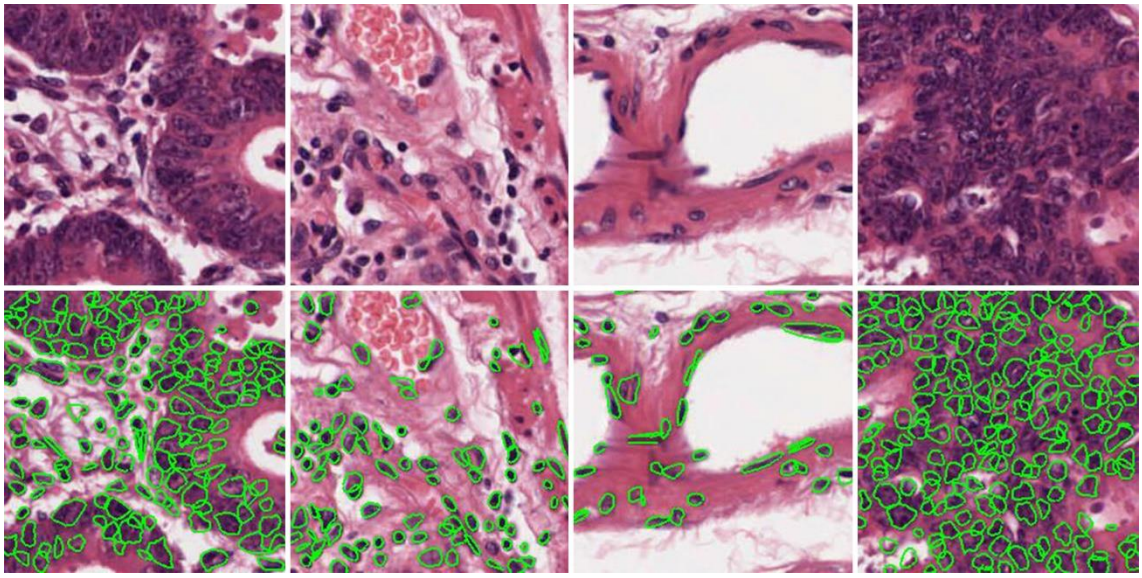


Figure 23 - Nuclear detection results for the Bennett et al algorithm [183]

Top: Original H&E stained images
Bottom: Nuclear detection results

The algorithm performs well on clearly separable nuclei, and copes adequately with densely clustered nuclei, with no input parameters (tuning).

The Bennett et al nuclear detection algorithm is used in Algorithm A, presented in Chapter 3.

2.5.4 Glandular detection in CRC images

Due to the complex and deformable nature of cancerous tissue, detecting glands and structures in CRC is a non-trivial task. Segmenting tissue structures such as tumour epithelium can be attempted using traditional feature based methods, such as texture [168,169], or traditional unsupervised segmentation such as graph cuts (see section 2.4.5.4) [170-173]. More recently, published methodology applies multi-scale analysis, using either a "top-down" approach, whereby characteristics of the tissue types are ascertained using general features of the tissue (typically at a scaled view, rather than native magnification) [174,175] or a "bottom-up" approach, that uses segmentation at native resolution (superpixels, clustering, nuclei detection) and combines visually similar regions [195]. Methods that use the bottom up approach have modelled glands as polygons using nuclei centres as co-ordinates [177-179], and are typically represented as graphical models such as Markov Random Fields (MRF) [199].

Table 3 lists publications that focus on detecting tumour in various cancers, recording the tissue type, stain, image size, number of images used in the development and testing of the algorithms, and their reported accuracy rating.

Primary author	Year	Tissue type	Stain	Image type	Image size	Image count	Accuracy
Reis [200]	2017	Breast	H&E	ROI	12x12	224	0.84
Bejnordi [201]	2017	Breast	H&E	WSI	-	646	0.96
Kainz [202]	2017	CRC	H&E	ROI	775x522	165	0.95
Kather [203]	2016	CRC	H&E	ROI	150x150	5000	0.99
Wang [198]	2016	Lung	H&E	TMA	300x300	9	0.80
Wang [198]	2016	Lung	IHC	TMA	500x500	9	0.78
Chen [204]	2015	Breast	H&E	ROI	1360x1024	1150	-
Rogojanu [205]	2015	CRC	H&E	ROI	-	50	0.97
Bianconi [206]	2015	CRC	IHC	ROI	161x161	1412	0.97
Hamilton [207]	2015	Lung	H&E	WSI	31x31	136	0.97
Paakkonen [208]	2014	Prostate	H&E	TMA	1280x960	-	0.79
van Engelen [209]	2013	Brain	H&E	ROI	800x800	13	0.76
McKenna [210]	2013	Breast	IHC	TMA	76x76	64	-
Angell [211]	2013	CRC	IHC	ROI	-	65	-
Mattfeldt [212]	2013	Prostate	H&E	ROI	512x512	103	0.82
Linder [213]	2012	CRC	IHC	TMA	80x80	1376	0.97
Doyle [214]	2012	Prostate	H&E	ROI	256x256	2000	0.86
Beck [215]	2011	Breast	H&E	TMA	-	1444	0.89
Law [216]	2011	Endometrial	IHC	ROI	200x200	60	0.85
Eramian [217]	2011	Oral	H&E	ROI	1300x1300	73	0.85
Signolle [218]	2010	Ovarian	IHC	ROI	512x512	61	0.72
Yang [219]	2009	Breast	IHC	TMA	1200x1200	3789	0.89
Huang [220]	2009	Prostate	H&E	ROI	512x384	205	0.95
Naik [221]	2008	Breast	H&E	ROI	-	54	0.82
Datar [222]	2008	Prostate	H&E	ROI	1024x768	109	-

Table 3 – List of published algorithms for segmenting tumour epithelium in histopathology images

List is sorted by year of publication. Some values are not reported in the publications, these are marked with a hyphen. Image type ROI = region of interest, TMA = Tissue MicroArray core image, WSI = Whole Slide Image, typically analysed in small image tiles / patches.

The data in this table provides a clear and concise set of data for drawing benchmark levels and noticing trends between types of analyses. For example, Figure 24 plots accuracy results of all algorithms, grouped by stain type. This shows that a reasonable state of the art benchmark level of accuracy is 86.5%, and that results on IHC images are slightly, but not significantly higher than results on H&E images ($p = 0.79$), but the spread of distribution is higher, and therefore reported accuracy results are not as consistent.

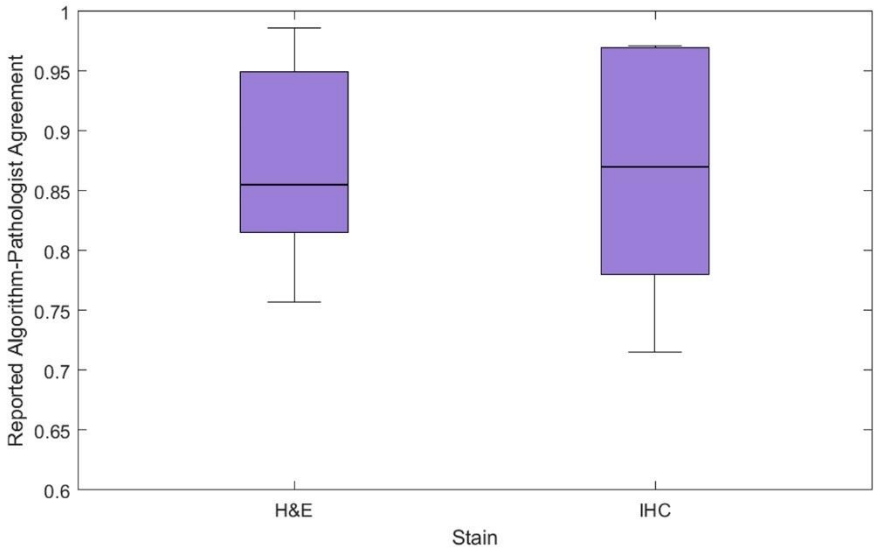


Figure 24 – Boxplots of published algorithm accuracies, grouped by stain type

Left: Image analysis applied to routine H&E images (mean = 0.87, median = 0.86, S.D. = 0.08)
Right: Image analysis applied to IHC images (mean = 0.86, median = 0.87, S.D. = 0.10)
The plot shows reported accuracy levels with state of the art algorithms. Image analysis results using IHC are slightly higher than on H&E, but not as consistent.

The slight increase in average accuracy on IHC image may be due to the tumour component of the tissue being highlighted by histochemical staining. In the case of antibodies such as cytokeratin (CK), this makes the tumour epithelium appear brown, and stroma blue. In terms of the segmentation task, this is simpler problem than on H&E. However, the spread of the distribution suggests that complicating factors may exist, such as quality of stain, or other issues relating to image variability, discussed in 2.3.3.

Figure 25 shows a scatterplot of image size used and accuracy reported in attempt to identify if there is a trend between the two. The x axis is plotted on a log scale due to the variability and spread of the number of images used.

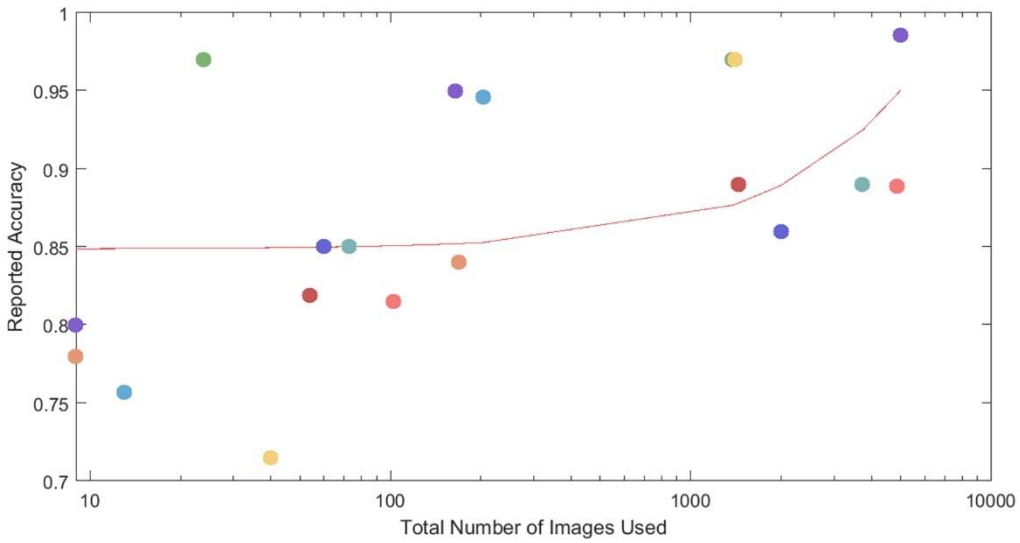


Figure 25 – Scatterplot accuracy vs total number of images use in the studies listed in Table 3

The overall trend shows a weak positive correlation ($R^2 = 0.18$). Note that the x axis is plotted on a log scale. The correlation suggests that algorithms using larger datasets lead to higher accuracy levels.

The scatterplot shows a weak positive correlation ($R^2 = 0.18$), which is exaggerated by the log scale in the higher numbers of the x axis. This implies that accuracy is likely to increase when using more images, and that the ideal number should be in the thousands, rather than hundreds.

2.5.4.1 The Kather et al algorithm

A large proportion of published tissue classification algorithms that are not based on segmentation use texture features to discriminate tissue classes [151,182-188]. Kather et al (2016) present a methodology for multi-class texture analysis in CRC [203], performed on greyscale images derived from H&E slides. The texture statistics generated were from six different methods: lower and higher order histogram features [227]; GLCM (see section 2.4.4.1); LBP (see section 2.4.4.2); Gabor filters [228]; Perceptron-like filters [206]; Combined feature sets. The features were tested against four different classifiers, 1 Nearest Neighbour (kNN) [229], two variants of SVMs and a decision tree classifier. Each classifier was trained using the texture features and one of eight hand labelled classes: Tumour epithelium; Simple stroma; Complex stroma; Immune cells; Debris (including necrosis, haemorrhage and mucus); Normal mucosal glands; Adipose tissue; Background. Ten-fold cross-validation is performed on a dataset of 10 hand-labelled slides, split into 5,000 150x150px images, and report 98.6% accuracy for the combined texture features using SVM with a radial basis function.

The Kather et al algorithm uses patch-based analysis to train classifiers using eight classes of CRC tissue. Of all the published algorithms found in the literature, this algorithm is the most

similar to Algorithm A presented in this research (Chapter 3). A comparison between the algorithms developed in this work and the Kather et al algorithm is made in the section 8.2.4.

2.5.4.2 The Gland Segmentation (GlaS) challenge

CRC gland segmentation was the focus of the Medical Image Computing and Computer Assisted Intervention society (MICCAI) annual conference in 2015, launching the colorectal gland segmentation challenge (GlaS) [230], which invited researchers to submit their algorithms for benchmark testing against other state of the art solutions, by validating developed solutions on the same dataset. One dataset was released prior to the challenge which allowed participants to train and test their algorithms (60 images), whereas an unseen dataset (20 images) was released on the day of the challenge. Participants had a 45-minute window to process them.

All but one of the top ten submissions used Convolutional Neural Network (CNN) technology (see 2.4.7) as a classifier [148], with the other method using K-Means clustering (see section 2.4.5.3) and a naïve Bayes classifier (see section 2.4.6.2). The nine entries using CNNs differed in the pre-processing and post-processing methodologies, the depth of the network, and the tissue class number and definitions. Figure 26 shows the distribution of accuracy results for all algorithms on the pre-released dataset and the unseen dataset.

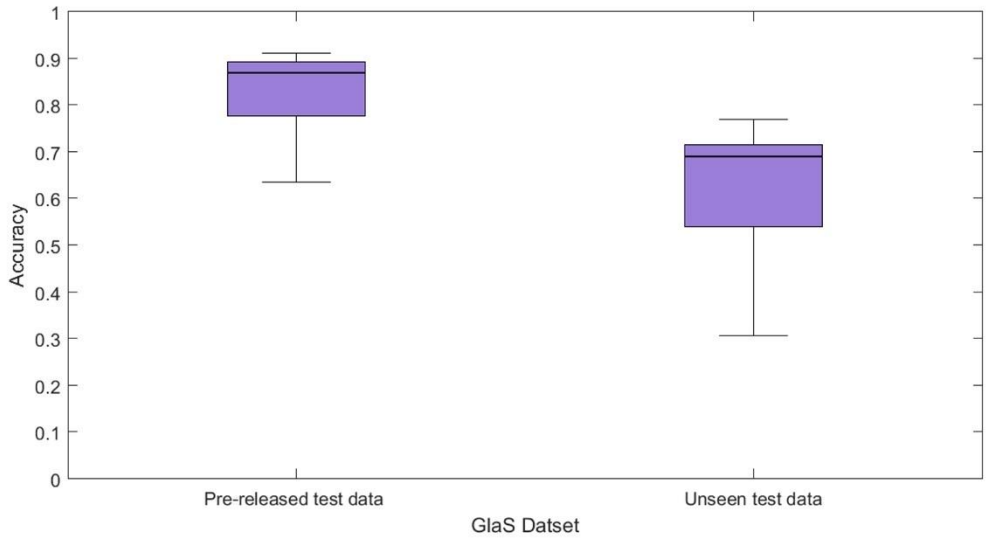


Figure 26 -Boxplots of GlaS challenge accuracy results, reported by dataset

Left: Pre-released test data results (mean = 0.82, median = 0.87, S.D. = 0.10)
Right: Unseen test data results (mean = 0.63, median = 0.69, S.D. = 0.14)
The graph shows that the algorithms developed on the pre-released test data were not generalised enough to repeat the same level of accuracy on the unseen dataset.

Accuracy in this case is reported as an F1 score, which is calculated as the harmonic average of precision and recall. Automatically segmented gland objects that intersect with over 50% of a ground truth gland are considered true positive, otherwise the object is considered a false positive. Algorithms performed significantly better on the pre-released test set compared to the unseen dataset, with means of 0.82 and 0.63 respectively. The distributions of the results were statistically significantly different to each other (t-test $p < 0.01$). Most of the methodologies used data augmentation to increase the number of training examples, in attempt to improve the generalisation of the algorithms. The results presented for the unseen dataset imply that the variation of image data was not adequately compensated for by this method. It is therefore an important message that adequate numbers of training examples and validation images are required when evaluating algorithm performance.

A comparison of segmentation algorithms for CRC images is presented in Chapter 5.

The GlaS challenge represents the general trend of research into CRC classification, shifting towards using CNNs as a black box to automatically identify visual features without bias in model or parameter creation [193]. Alternatively, Levenson et al hypothesise that understanding the cognitive and visual systems of animals may be useful as a gateway to understanding the processes involved in pathologically assessing images [233]. The paper explores the concept that pigeons have the capacity to distinguish between malignant and benign tumour samples, and using a touch sensitive interface, trained Columba Livia pigeons by automatically dispensing pellets to reward pecks on the screen that corresponded with a correct diagnosis. After 20 days training, pathologist-pigeon agreement rose to 88% on images presented at 20x magnification. The levity of this concept detracts from the message of the paper – that understanding their performance on visual tasks may assist in guiding future vision research. However, coining the term “flock-sourcing” (a take on crowd-sourcing) and suggesting pigeons have the potential to be cost effective medical image observers does not help.

CRC classification algorithms are developed in Chapter 3, 5 and 7 in this work to identify CRC tissue on digital slides, and also in Chapter 4 and 6 for the application of automated QC. The work presented in these chapters predates the popularity of the deep learning methodology.

2.5.5 Automated TSR analysis

Due to the emerging research into TSR as a prognostic marker, no articles have been published focusing on automating the task. TSR can be generated as a by-product of gland segmentation (assuming TSR is the ratio of gland area to other tissue detected in the image), and as such is

reported in three of the previously referenced publications, although it is not the focus of the studies [203, 206, 207]. Hamilton et al [207] report TSR analysis on 10 slides, yielding a high correlation to ground truth ($R^2 = 0.97$), using a pathologist to count over 500,000 cells as the benchmark. This is compared to their observations of human inter-observer variability, which is less well correlated ($R^2 = 0.32$). Kather et al [203] present a similar methodology to the work presented in Chapter 3 and Chapter 5, using multi-class image patches to learn sub-classes of tumour and stroma tissue. TSR values and ground truth correlations are not reported but mention that their work could be used to generate TSR. It should be noted that this work was published after the development of the algorithms in this thesis. Bianconi et al [206] report a 0.97 accuracy rating using a two class system (tumour and stroma), generating confusion matrices on a per-patch basis, but show no correlations to pathologist-generated ratios. It should be noted that the study uses IHC images which have the tumour component stained with DAB, and are more visually distinguishable than H&E images. In all three publications mentioned, SVM classifiers are used to learn image features and predict the classifications of image patches retrieved, in small image tile areas.

2.5.6 Quality assessment

The scope for variability of digital slide images has been outlined in section 2.2.2 from a histological and slide preparation perspective, and in section 2.3.3 in relation to digital slide scanning, and calibration of devices. These factors can influence the appearance of digital slides, and subsequently, how image analysis solutions perform on them. Research into colour standardisation (discussed in 2.3.3) and stain normalisation (discussed in 2.5.2) allows algorithms to process images that have a more consistent appearance, to improve accuracy. However, as seen from the adverse normalisation performance shown in Figure 21, extreme deviations in image quality lead these techniques to exacerbate the issues rather than improve them.

Quality assessment is a visual inspection task assigned to scanner operators, and the Quality Control (QC) procedure typically involves identifying slides that have digital scanning artefacts, such as out of focus areas or incorrect stitching of image tiles (striping artefacts on line scanners). However, histological issues on the slides themselves cannot be accounted for, such as weak staining, tissue folds, tears, or even permanent marker on the slide, used by researchers to delineate regions of interest. These issues have the potential to negatively affect image analysis solutions, whilst being considered a QC pass, and therefore should be considered when designing automated solutions.

There are few documented examples of research into this area in the literature, and typically QC algorithms focus on replicating the work of the digital slide scanner operator, evaluating scanning artefacts such as irregular illumination, overall brightness, contrast, colour separation and blur, as opposed to the content of the image [234] [235]. Ameisen et al [236] present a method of whole slide image analysis for automation of the QC task, that analyses a given slide using each layer of the tiled pyramid. Initially saturation in the HSV colourspace is used to identify background and weakly stained tissue, and image tiles considered to contain foreground objects (including artefacts) are further processed for blur, contrast, brightness and colour, and a threshold is set for each, in order to pass or fail the slide on that metric. Avanaki et al [237] present work on developing an algorithm to QC both scanning and histological artefacts using two separate estimators.

The issue of image quality is addressed in Chapter 6.

2.5.7 Validation against clinical data

The pathological image analysis methods presented in this section (2.5) are based on existing computer vision methods, and model human methods of scoring, which are evaluated against images scored by a human. As mentioned in the problem statement in 1.1, these methods are prone to inter and intra-scorer variation, meaning that most algorithm validation is performed on imperfect ground truth. Aeffner et al refer to this problem as the 'Gold Standard Paradox' [238], which points out that computer vision systems require validation against the pathologist as the gold standard to achieve validity – however, such systems are developed out of a need to address the flaws in human analytical methods, and so to compare them to humans as ground truth is not ideal.

By regressing the problem back to the original reason for pathological assessment – *prediction of patient survival and response to therapy* – it should be considered that the more appropriate evaluation strategy for automated pathological assessment is whether the algorithms perform this task better or worse than a human.

Currently, computer vision researchers have limited access to existing gold standard data, and generating such data is laborious and time consuming for pathologists. This has led to calls for shared repositories of data between researchers [239]. Survival data on the other hand, is generated by expensive clinical trials, that require years of patient follow-up data, and therefore clinical trial data centres typically do not disclose survival data, without certain controls. Researchers are blinded to the data, so that independent statisticians can perform hypothesis

testing on their behalf, so as not to over-engineer solutions to the datasets. Validation of image analysis algorithms against clinical datasets should be performed with this in mind.

Beck et al [215] remark that human pathological scoring relies on minimal observations, and that vast amounts of computer-generated image feature data have the capacity to identify phenotypic prognostic markers that are currently unknown, due to their complexity and infeasibility for the visual inspection task. The Computational Pathologist (C-Path) system is presented as a computer vision system that identifies 6,642 features from two breast cancer cohorts using H&E stained Tissue MicroArray (TMA) digital slide images (totalling 576 patient cases, with a mean value of 2 TMA cores per patient). The C-Path system was developed as a processing pipeline using the Definiens Developer XD image analysis environment [240], and as such, the exact methodology was not reported in the paper. The process of tissue analysis has been broken down into the steps outlined in Algorithm 2.

```
for each H&E TMA core image i (total n = 158)
  Partition i into superpixels
  for each superpixel p
    Segment nuclei in p
    Generate feature vector v (size = 31)
    Append v to feature set f
  end
end
Train Logistic Regression classifier on f
```

Algorithm 2 – Beck et al tumour-stroma classifier using Definiens Developer XD

The 31 features generated in Algorithm 2 are based on intensity, texture, size, shape, and relational properties of neighbouring superpixels. The subset of 158 images was hand labelled, and combined with the feature set to train an L_1-regularised logistic regression classifier (see section 2.4.6.3 for logistic regression) [241], and reported 89% accuracy using 8-fold cross validation.

Using the trained classifier, the entire dataset of 576 cases was automatically partitioned into tumour, stroma and unclassified regions. Within those regions, nuclear detection was applied so that their shape metrics could be used to identify regular and atypical epithelial nuclei, stromal round and spindled nuclei, and unclassified round and spindled nuclei. The set of 6,642 features was generated from these statistics, in combination with the features used in the previous classifier, and their spatial characteristics (such as distances of one nuclear type to a different type, or distance between tumour or stromal regions). For all metrics, the mean, standard deviation, minimum and maximum was calculated.

To compare against survival data, a Boolean set of training labels was made per case, splitting patients into groups "alive at five years" and "not alive at five years". The labels were combined with the final feature set and were trained and tested using 8-fold cross validation on a second L_1-regularised logistic regression classifier. The trained and tested algorithm was validated on an unseen dataset, where the algorithm stratified the patients into groups predicted to live beyond five years or not. Kaplan-Meier curves [242] and log-rank tests confirmed that the model had statistically significant prognostic capabilities on both of the datasets used ($p < 0.01$ for both). Hazard Ratios (HR) were not presented.

The visual features in the second classifier used were reported individually for their prognostic significance, using their coefficient estimates. Of the 6,642 features used, 11 (0.002%) yielded estimates with 95% confidence intervals that did not overlap zero (zero or negative probability). Seven of the top 11 were relational features assessing contextual relationships, such as distance between nuclei or regions. In the supplementary materials provided with the paper, coefficients are provided for 41 of the 6,642 features in descending (absolute) numerical order, and Figure 27 shows the absolute values of these features.

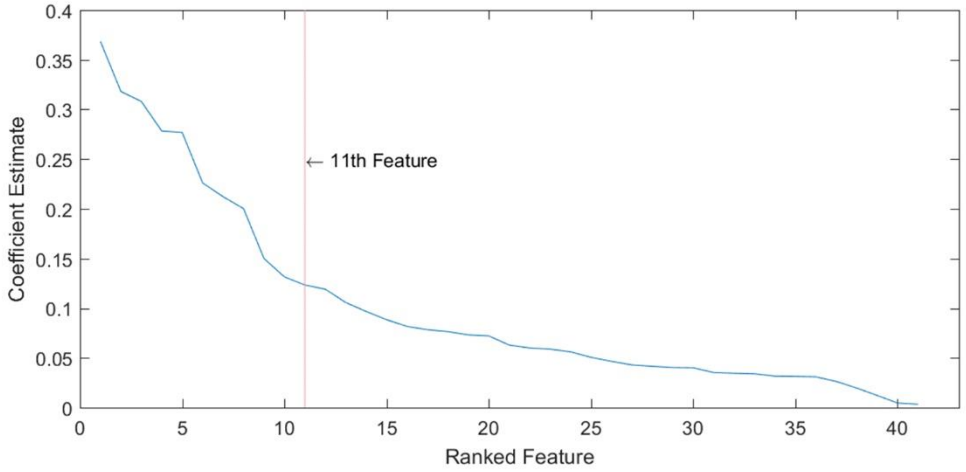


Figure 27 - Beck et al coefficient estimates of 43 algorithm features

Of the top 11 features, 3 were related to stroma, and an independent model built on these three features alone was as strong a predictor as one built on all 6,642. The top three features were also stronger than the remaining eight epithelial features combined.

One thing to note is the number of TMA core images is reported in the study - 576 cases with a total of 1286 core image. Tissue MicroArrays can contain upwards of 400 cores per slide and so the entire dataset may have been analysed using 3 or 4 slides, which would mean the algorithm does not account for histological or digital variability discussed in previous sections.

Chen et al present a similar approach, using 730 features generated from a combination of image features over 230 patient digital slide images, SVM classifiers and watershed nuclei detection, and 12 features were significantly associated with survival [204].

The algorithms developed in this thesis are evaluated against clinical data in Chapter 7.

2.5.8 Summary

Computer vision and ML within digital pathology research is an exciting and rapidly developing field, tackling a variety of subjects including stain normalisation, nuclear and structural segmentation, tissue classification, quality assessment, and automatic prediction of survival.

Image analysis provides a logical method for automating quantitation of subjective and laborious tasks, with consistent and predictable accuracy.

Many automated solutions exist for specific problems in histopathological analysis, and as such, are not generalisable to other tasks. Classifiers used for creating models of tissue are traditionally built using hand-selected features, of varying types and quantities.

Methodology for validation of computer vision and ML algorithms is typically limited to using cross validation, and the number of image used varies dramatically. Of the papers analysed in this chapter, numbers of digital slides ranged from 10 or potentially fewer (see discussion on Beck et al's work), to 230.

There are many algorithms that exhibit promising accuracy statistics, yet the uptake of an accepted solution in this area is ongoing. Without appropriate levels of validity in terms of the number of datasets used for validation, and perhaps even survival statistics, these solutions cannot be fully trusted.

2.6 Summary

The work presented in this chapter is already summarised in the appropriate sections, and is briefly listed below.

Colorectal cancer

- CRC has a high economic cost worldwide.
- CRC is non-trivial to analyse.
- Pathologists provide recommendations for CRC treatment based on prognostic indicators
- The current methodology for identifying prognostic indicators relies on manual assessment.
- Subjectivity and fatigue cause inter and intra scorer variability.

However, despite these challenges, there is a clear benefit that consistent and reliable automated analysis of CRC will bring to the pathologist workflow, and ultimately patients.

Digital pathology

- Digital pathology is an exciting and modern innovation that encompasses a broad spectrum of research fields.
- The digital slide creation pipeline crosses between laboratory practice and computer science, and there is scope to create variation in both areas.
- Digitisation of glass slides offers the capacity to develop automated solutions.

Automation of routine visual inspection tasks using computer vision and machine learning has the capacity to increase speed, accuracy and reproducibility of results, given that developed solutions are validated, and trusted by the pathological community.

Computer vision

- Computer vision is applied to digital slide images for enhancement, detection, classification, quantification and prediction.
- Many solutions exist, but not enough widely available datasets exist for robust algorithms to be extensively validated.

- Survival analysis can be replicated by machine learning, and has the capacity to learn new prognostic markers.

Thesis aims

Moving forward, the work in this thesis aims to:

- Generate a large dataset of pathologist scored CRC images, representative of routine work, to use as a robust bank of gold standard data
- Automate the pathologist scoring task for prognostic features
- Address issues with image quality that may affect image analysis
- Validate the automated solution on large amounts of data from the dataset developed
- Automate survival prediction

These five tasks are split across the next five chapters (3 to 7), involving algorithm development, refinement, web systems development, human computer interaction experimental design, algorithm refinement, automated image quality assessment, and survival analysis using the automatically generated prognostic markers.

Chapter 3 - Automation of systematic random sampling

3.1 Introduction

3.1.1 Chapter overview

The work in this chapter presents a systematic random sampling (SRS) system based on stereological methods (see section 2.2.3.4), for the manual analysis of histopathology images. The system is used to collect ground truth data to make a large repository of labelled image data, which is then used to develop an automated system for application of SRS to digital slides. The chapter is divided into five sections:

1) The introduction, which builds on the description of SRS and stereology from 2.2.3.4, and discusses how SRS can be used effectively for the analysis of CRC.
2) Presentation of the RandomSpot system, a SRS tool for manual quantitation of digital slide tissue, and how it has been used in clinical trial research to determine prognostic factors in CRC, and a description of RandomSpotDB, a repository for ground truth data relating to manually scored images using the RandomSpot system.
3) An overview of one CRC trial which used the RandomSpot system, the data that it generated, and how this data can be used as an input to train image analysis algorithms, with exploratory analysis of the dataset, and assessment of basic image processing techniques for their suitability of application to the dataset.
4) Development of a machine learning (ML) algorithm, which uses the original clinical trial image label data as a training set, for learning the appearance of tumour and stroma tissue.
5) Discussion of the work presented in this chapter and conclusions. The discussion focuses on the how well the algorithms perform compared to a human, the main challenges of processing histopathological image data, the quality control issues

involved when processing a longitudinal dataset and how validation can be flawed when comparing against subjective assessments.

3.1.2 Systematic random sampling for CRC tissue

See section 2.2.3.4 for a background on stereology and systematic random sampling.

Recap of 2.2.3: The manual quantification of tissue containing highly complex cellular structures within regions of interest (ROI) on pathological slides is laborious and prone to interobserver and intra-observer variation. Current quantification methods (scoring) rely on experienced pathologists to make informed estimations of the proportions of tissue, either on a Whole Slide Image (WSI), or within a given region of interest (ROI). These estimations are subjective, and require broad category bins to maintain inter and intra scorer consistency (typically three to four bins).

Virtual slide viewing software solutions often provide inbuilt tools for drawing boundaries around tissue types, or counting cells with mouse clicks. Although applying these techniques to gigapixel resolution images is possible, due to the number of cells per image, the application is not practical. SRS provides a feasible alternative to manual whole slide quantification, but can be prone to biases when choosing areas for sampling.

Systematic Random Sampling (SRS) allows accurate unbiased estimation of the proportion of classes while minimising the number of measurements needed. Traditionally, systematic random sampling involves placing a fixed grid (usually on an optical graticule) at a random seed point on a slide and counting objects under the points on the grid [35]. The efficacy of the SRS technique fundamentally relies on selecting an appropriate density of the sampling points within the grid. The density of the grid should be based on the estimation of the relative proportions of tissue within the area being analysed (low vs high frequency), as well as the distribution of those proportions (sparsely vs densely distributed) [243].

With an appropriate number of samples, the true frequency of objects in the whole tissue can be estimated effectively from the sampled frequency using the grid [33,208], which ensures higher reproducibility and consistency in scoring (see 2.2.3.4 for a worked example calculating optimal number of samples). However systematic random sampling is still laborious, requiring the user to make hundreds of measurements using a conventional microscope and an optical graticule. Errors are easily made, and it is difficult to pause and recommence work.

3.2 Systematic random sampling and ground truth acquisition

3.2.1 Aim

To develop a tool that applies SRS to digital slide images, within a given ROI, allowing researchers to quantitate the area and ratios of tissue types within that area.

3.2.2 Methods

The RandomSpot system [245] is a web-based HTML5 SRS tool designed to generate a uniform distribution of discrete, quantifiable targets within a given region of interest (ROI). These targets are used for systematic sampling of cancer tissue, to generate an estimation of the relative area of features within a given region.

3.2.2.1 The RandomSpot algorithm

Spots are created using a MATLAB compiled executable program, which initially generates a grid with an arbitrary number of spots, with a random seed to initiate the first spot [246]. Spots are spaced equidistantly using a hexagonal mesh, which is iteratively increased or decreased in size until the number of spots within the ROI matches the number chosen by the user, within the percentage tolerance specified.

Initially the minimum and maximum x and y values of the ROI are padded with borders 10% of the width and height respectively (see Figure 29). A single co-ordinate, randomly generated within the bottom left corner of the border (the seed point area) is used to instantiate the seed point for the grid. This random seed point ensures that sampling points are randomly (as well as systematically) placed. Horizontal grid spacing (distance) is defined by the constant d, and vertical grid spacing h is defined as $h = d \times \sin(r)$ where r is 60 degrees, expressed in radians, to generate an equidistant set of points, based on a hexagonal mesh structure. A hexagonal mesh structure is chosen over square or triangle in order to reduce sampling bias from edge effects of the grid shape, caused by having higher ratios of perimeter to area. Point distance d is initialised using the following formula:

$$d = \frac{P_w \times P_h}{T^2}$$

Equation 16 - Initialisation of RandomSpot mesh grid point distance

Where P_w is the width of the polygon and T is the target number of spots. Points are generated along each row and column, with the stopping criteria set as the maximum ROI x or y value plus the 10% border.

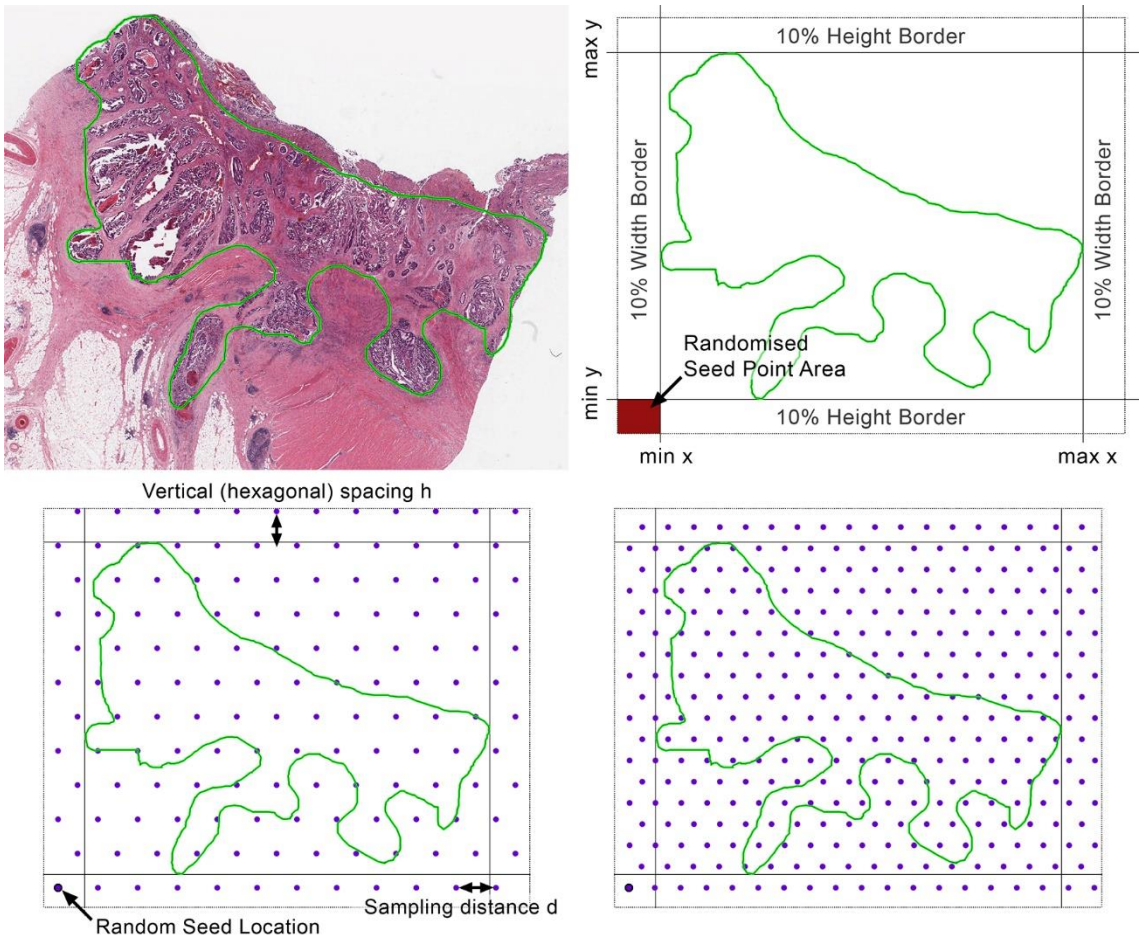


Figure 28 – RandomSpot mesh grid generation

Top Left: Original ROI drawn by a pathologist
Top Right: The isolated ROI showing the maximum and minimum x and y values, with the 10 height and width borders, and the randomised seed point area
Bottom Left: Example sampling grid with random seed point and sampling distance d
Bottom Right: Example sampling grid with the same random seed point and a smaller value for d

Points are created along the x axis with a spacing of d pixels, and every row is created using a spacing of h pixels. To create the hexagonal mesh, alternating rows are spaced with $d + d/2$ pixels.

After the mesh grid is generated, and stopping criteria are met, the number of sampling points within the ROI boundary are calculated. If the number of points falls within the target number, ± the tolerance value (set as a percentage of the target number), then the algorithm converts the sampling points to XML (eXtensible Markup Language) and saves them in the Leica-Aperio annotations file format. If the number of points within the ROI is too low, the grid is regenerated with d set to $d/(1.2 + k)$, and if the number of the points is too high, d is set to $d \times (1.1 + k)$, where k is a randomly generated number between 0 and 0.01. This process is repeated until the target number of points is fitted in the ROI, within the tolerance specified.

3.2.2.2 The RandomSpot system

RandomSpot is a web-based SRS system, designed by Wright et al at Leeds, using HTML5, jQuery and PHP, with a MySQL database and MATLAB compiled executable programs used for background data processing. As a result, the system is both platform and browser independent. RandomSpot is primarily modelled on two simple use cases:

1) Creating SRS points for a given region / several regions of interest
2) Collecting hand-labelled SRS points as expert classified co-ordinates

These use cases are explained in the following sections.

3.2.2.3 Use-case 1: Generating spots

Use case one utilises HTML5 and jQuery to create a simple, user-friendly interface for uploading XML regions of interest (ROIs). Regions of interest can be rectangular, elliptical or polygonal. Once uploaded, the RandomSpot algorithm places equidistant, systematic, randomly distributed spots within the ROIs. The number of spots required will rely on the expected frequency of the tissue type being measured. By default, the system sets 300 as the target number of spots, as this is optimal for minimising the coefficient of variation for a target frequency of 50% i.e. a normal two class system (see section 2.2.3.4)

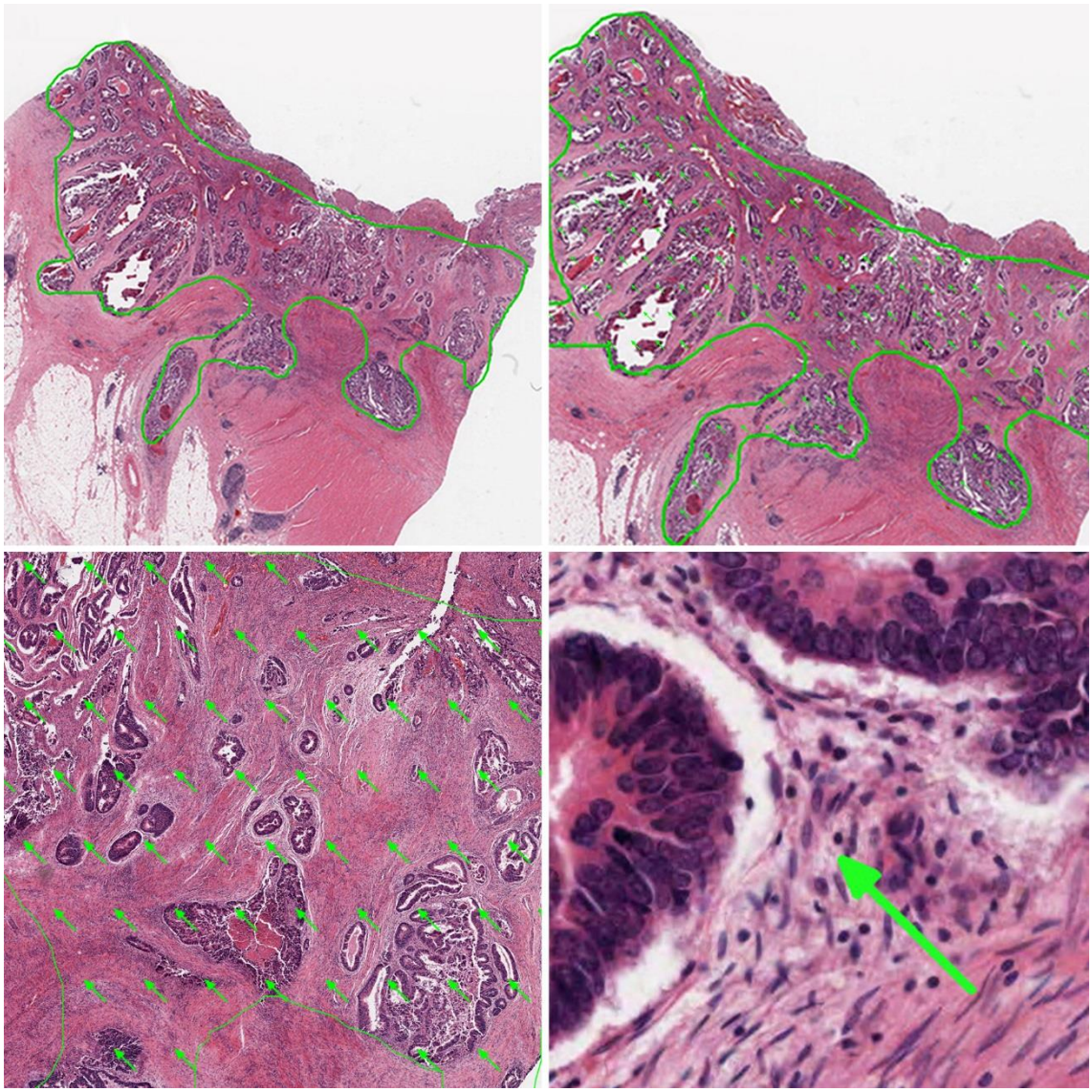


Figure 29 - Example of a virtual slide annotated using the RandomSpot software

Top Left: Hand-annotated polygonal ROI delineating a tumour boundary
Top Right: The resultant spots added to the ROI by RandomSpot
Bottom Left: Zoomed in view of sampling points
Bottom Right: Single Sampling point at 20x magnification

Figure 28 shows the ROI illustrated in Figure 29, overlaying the hexagonal mesh grid. The number of hexagons within the ROI is 373, which falls within the tolerance limits of the target number of spots set (400 ± 15%). Decreasing the tolerance level typically requires more iterations to fit the mesh grid to the ROI, increasing time taken to generate to sampling points.

Figure 29 illustrates how the software systematically adds these targets, known as 'spots', equidistantly within the ROI, and depicts the resolution at which a single spot is manually viewed and scored. The process of scoring involves manual visual inspection of each given spot, and recording the type of tissue at that location. The types of tissue (classes) that are

recorded are subject to the organ from which the tissue has been taken. Typical scoring systems involve giving a spot a number that corresponds to a pre-set selection of tissue classes, designed by the investigator. An example of the tissue classes used for TSR scoring in CRC can be found in Figure 33.

Each spot within the ROI is visually inspected and a classification is given for the type of tissue at that spot. The number of spots typically ranges from 50 to 300 per region on the slide. This number is dependent on the frequency of the observed objects in order to achieve a low coefficient of error (see 2.2.3.4).

Once all spots have been visually assessed and scored, statistics can be produced regarding the proportion and distribution of tissue types. One such metric is the proportion of epithelial cells to stromal cells within a cancer, known as the tumour-stroma ratio (TSR). Since the introduction of this sampling system, research at Leeds has shown that the proportion of tumour to stroma is a prognostic marker in colorectal cancer [6], and that the level of tumour cell density is a prognostic indicator of response to preoperative therapy [247].

3.2.2.4 Use-case 2: Collecting ground truth data

Use case two focuses on retaining the value of the hand annotated data, by encouraging users to submit their completed XML files after they have used the data for their own research. As with creating spots, XML files can be uploaded individually or as a zip archive. These files are processed, and each spot is added to the RandomSpot database – RandomSpotDB. These hand labelled locations are processed and stored as text classifications, paired with the URL strings containing x and y co-ordinates of the spot that has been scored, as well as the location of the virtual slide which the spot relates to. In addition to uploading XML files, users are encouraged to submit a scoring key, which matches the shortcut keys they used whilst scoring, to the actual semantic text classifications that are used when analysing the data. Also, to add further value to RandomSpotDB, each time the user submits their XML files, they are asked which type of tissue it is that they have scored. Having an extensive set of manually scored images, which is searchable by tissue type is extremely useful for training computer vision algorithms.

3.2.2.5 Collation of ground truth data into RandomSpotDB

The aim of the RandomSpotDB is to create a repository for the spot counting data that has been generated by experts for clinical trial studies and other research, so that the data can be reused for purposes such as image analysis and ML.

Using the annotation data collected from the RandomSpot system (section 3.2) a database of spot co-ordinates and manually classified labels was created. As per the usage guidelines of the

RandomSpot system, each XML annotation file was saved with the corresponding slide's unique file name (the Aperio SVS number). These XML files were grouped by the project that they were generated for, if available, the pathological data (tissue type, disease type), the stain applied to the slide, and the sampling method. The XML files were parsed using a custom script written in MATLAB, which extracted the XY co-ordinates of each spot, along with a numeric score that had been used to denote different types of tissue. Different projects used different numbering systems, so a key was generated for each project. Table 4 shows an example of a key used to convert numeric labels to text descriptions to store in the database.

Numeric Value	Tissue Type
0	Non-informative
1	Tumour
2	Stroma or Fibrosis
3	Necrosis
4	Vessels
5	Inflammation
6	Tumour Lumen
7	Mucin
8	Muscle

Table 4 - Example scoring key for one of the RandomSpotDB projects

Once the data was fully parsed, a MySQL relational database was populated with each spot saved as one record, using the schema described in Figure 30.

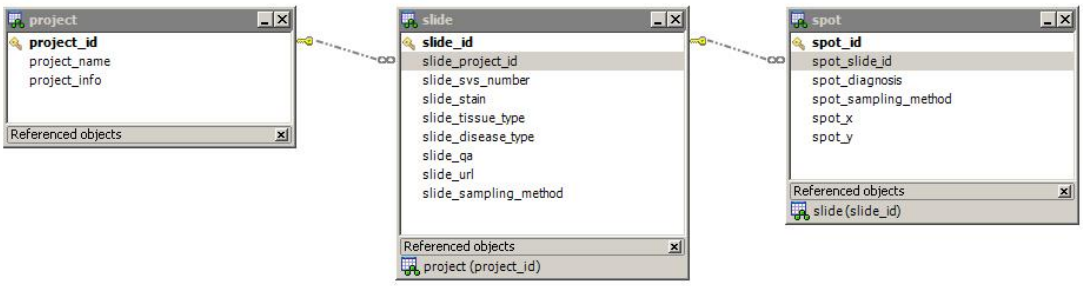


Figure 30 - ER Schema for RandomSpotDB

Note that the sampling methods field was stored in the slide table, meaning that if one slide was scored using multiple sampling methods, the spot data would be kept separate.

Once inserted into the database, the integrity of the data was checked by extracting each record and overlaying a coloured dot representing the classification label on a scaled thumbnail image of its parent slide at its correspondingly scaled XY co-ordinate. The labels were grouped into three groups, tumour (red), stroma (green) and other classifications (blue). Figure 31 shows examples of data that has been collated and overlaid on their parent slide thumbnail image.

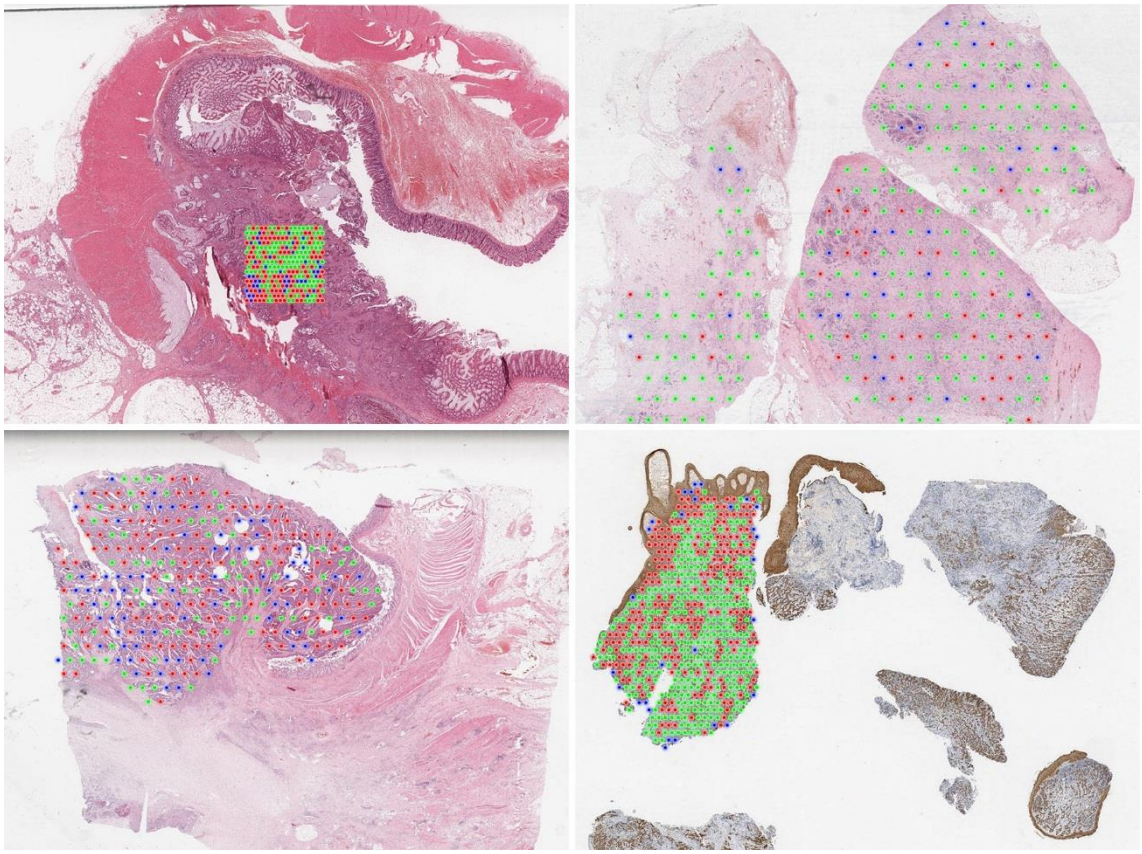


Figure 31 – Examples of RandomSpotDB ground truth data

Images visualising manual scores added to the Spot Counting training database from multiple clinical trials (red dot = tumour, green dot = stroma, blue dot = other)

The images are from four separate studies and have been sampled in different ways. Once the output images from the database had been generated, they were visually inspected (per database project) and verified.

3.2.3 Results

RandomSpot has been used extensively at Leeds [245] and is continually being improved iteratively, in response to user feedback. Since the introduction of this sampling system, research at Leeds has shown that the level of tumour cell density is a prognostic indicator of

response to preoperative therapy [247]. It has also been used in studies which have successfully identified the prognostic significance of the ratio of tumour to stroma in breast [42] and upper gastro-intestinal [48] cancers. Currently the system has been adopted by over 40 active users, which have generated over 21,000 sets of spots, with an estimated total of 6.3 million classifications (assuming 300 spots per ROI, and only one ROI per set). These sets of data are being used extensively by researchers, research students and in clinical trials [7,213-217]. After publication, the clinically valuable datasets have been reused and collected in RandomSpotDB.

With all the data parsed and inserted into the database, the total amount of data stored in the repository consisted of over 2.4 million expert-classified XY co-ordinates (spots). Table 5 shows the number of records in each table.

Spot Counting Database	**Contents**
Projects / Trials	32
Slides	11,989
Spots	2,470,628

Table 5 - The number of trials, slides and spots in the Spot Counting Database

3.2.4 Summary

The RandomSpot system provides a reproducible, quantitative method of estimating tissue proportions within CRC. However, manually inspecting each spot is a time-consuming task, with each spot taking an experienced pathologist an estimated five seconds (see section 4.2.4). Also, manually inspecting and classifying each spot still suffers from subjectivity, with mean pathologist agreement of 89% (see section 4.3). Therefore, automation of the spot counting task is highly desirable.

RandomSpotDB is a valuable resource for extensive training and validation of image analysis algorithms on histopathology images, containing over 2.4 million expert-classified images over approximately 12,000 digital slides, used in 32 independent clinical trials and research projects.

3.3 Exploratory analysis of the QUASAR dataset

3.3.1 Aim

To gain a better understanding of the QUASAR dataset, by analysing descriptive statistics of the SRS data from RandomSpotDB.

To identify issues which may affect automated solutions, using visual features of the digital slides, using statistical analysis and basic image processing techniques, by automating TSR generation and identifying images where performance is poorest.

3.3.2 Methods

3.3.2.1 The QUASAR CRC clinical trial

The QUick And Simple And Reliable (QUASAR) trial [253] was a multi-centre clinical trial that aimed to identify any survival benefits from applying chemotherapy to patients after curative resections of colon or rectal cancer. The trial consisted of 3239 patients, 2291 of which had stage II colon cancer, who were recruited between 1994 and 2003. Patients were randomised so that half received chemotherapy after surgery, and half did not. The trial was a longitudinal study, where five-year mortality and recurrence was recorded per patient. Therefore, when combined with patient demographics and pathological data, the trial is a valuable resource for applying statistical tests in order to identify prognostic indicators for CRC [254].

3.3.2.2 Image data

All patients in the QUASAR trial had pathological data recorded and Formalin-Fixed, Paraffin-Embedded (FFPE) biopsies, or blocks, were created for each patient. The blocks were sectioned into 5-micron thick samples, and stained with Haematoxylin and Eosin (H&E) to identify tumour regions. Glass slides were scanned using an Aperio[1] AT scanner at 20x objective zoom (0.5 microns per pixel).

[1] The QUASAR glass slides were scanned before Leica's acquisition of Aperio

3.3.2.3 Staining variation

Presented in this section are some example cases from the dataset. These have been grouped by similarity in terms of staining levels. This is to illustrate the effect of staining on the clarity of tissue structures within digital slides.

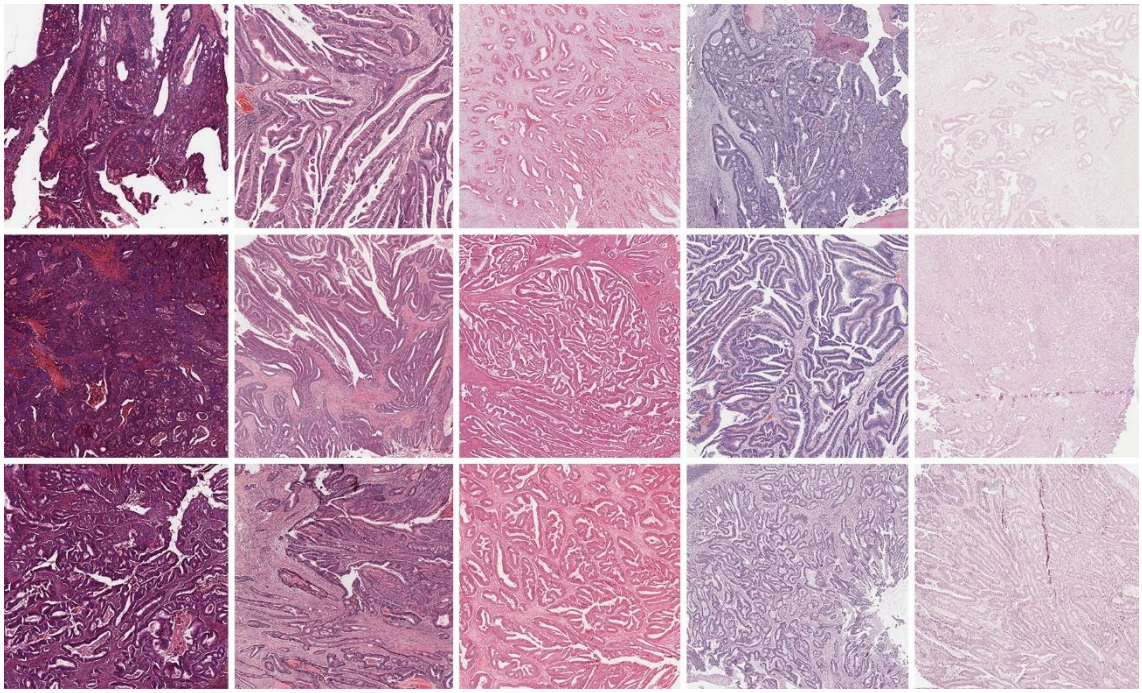


Figure 32 - Example cases, grouped by perceived levels of staining

Staining levels were subjectively assessed by a pathologist and are displayed in columns (left to right) as over stained, well stained, low levels of haematoxylin, low levels of eosin and weakly stained overall.

Based on the performance of the colour normalisation algorithm in 2.5.2, it is likely that the variation exhibited in these images is too extreme for the trained RVM to identify enough relevant foreground pixels in the weakly stained slides to apply colour deconvolution and stain mapping.

3.3.2.4 Scoring data

The RandomSpot SRS system (section 3.2) was applied to digital slides from the QUASAR dataset, to assess the prognostic capabilities of the ratio of tumour to stroma [52]. For each of the colorectal cancer cases scanned (n = 2,211), tumours were identified by a trained pathologist, and annotations were drawn around the whole tumour. In addition, a 3x3mm square was placed over the area with the highest perceived tumour cell density. A subset of 145 cases were sampled using 300 spots per region of interest (ROI), and all cases were sampled using a target of 50 spots per ROI, with a tolerance of 15%. This was to minimise the overall workload of the pathologist, whilst maintaining results that were statistically similar to the methodology

using 300 spots. Table 6 shows the full amount of scoring data generated by this study, held in the RandomSpotDB.

Sampling Location	ROI Boundary Type	# Sampling Points	# Cases
Whole tumour	Freehand polygon	300	145
Whole tumour	Freehand polygon	50	2210
Highest area of TCD	3x3mm square highest TCD	300	145
Highest area of TCD	3x3mm square highest TCD	50	2211

Table 6 - Pathologist-scored data generated by the QUASAR TSR study, available for image analysis training

Once the sampling points were generated, each one was individually assessed by either a trained pathologist or technician, according to the available scoring categories detailed in the key in Table 4. The scoring categories used were all sub-classes of the tumour or stroma categories, and Figure 33 illustrates examples of the different sub-classes for both tumour and stroma.

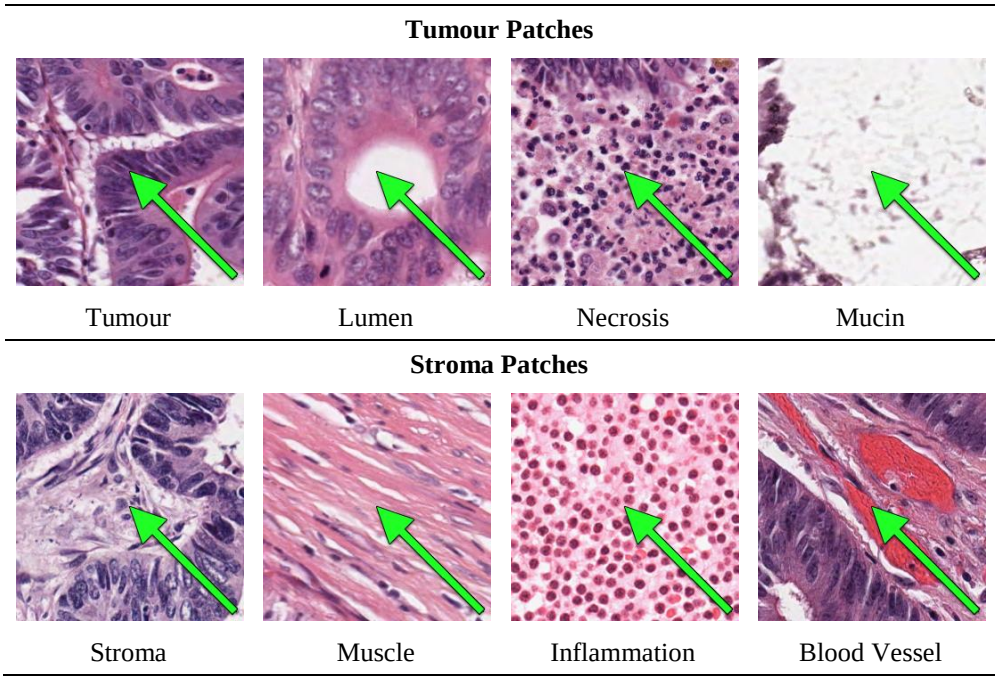


Figure 33 - Subtypes of tumour and stroma classes,

Green arrows illustrate the exact centre that determines the classification of each patch
Top: Left to right; tumour, lumen, necrosis, mucin
Bottom: Left to right; stroma, muscle, inflammation, blood vessels
Images shown are 256x256 pixels in size and extracted at 20x objective zoom.

Note that each of the sampling points were generated using the RandomSpot SRS system, and therefore classifications only apply to the centre of the images shown (at the tip of the

arrowhead). Examples of the four different sampling methods are shown in Figure 34, applied to the same digital slide. The TSR is slightly different in each of the examples (approximately 0.46 tumour to 0.32 stroma).

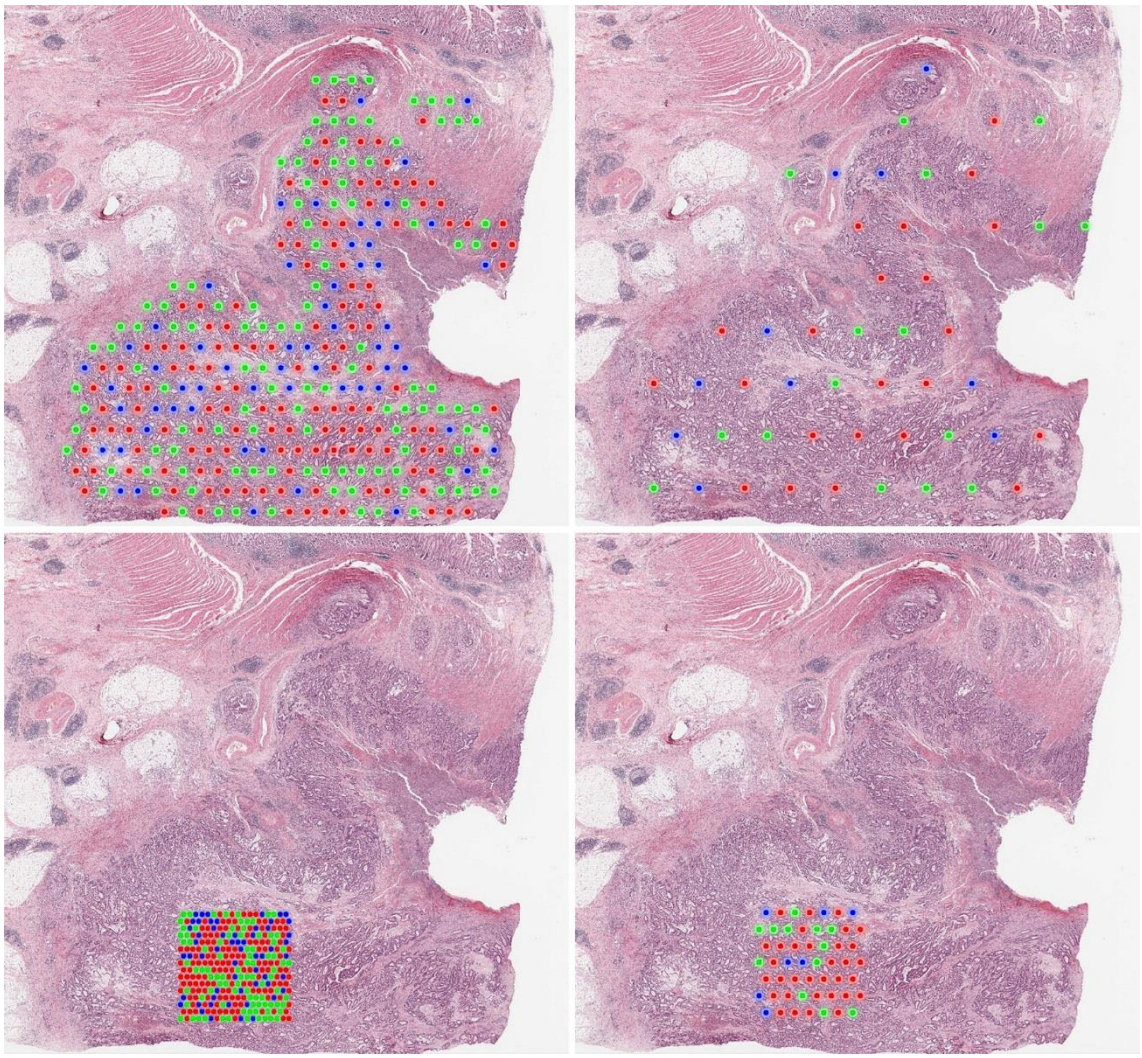


Figure 34 – Sampling methods used on the QUASAR trial data

The same tumour sampled by a pathologist, using the four methods described in Table 6
Top Left: 300 spots on the whole tumour
Top Right: 50 points on the whole tumour
Bottom Left: 300 points within a 3x3mm square placed over the area with the highest TCD
Bottom Right: 50 points within a 3x3mm square placed over the area with the highest TCD.
Red dots indicate tumour labels, green dots indicate stroma, and blue indicate other types of tissue / non-informative areas

The method of placing a 3x3mm box over the highest perceived area of tumour cell density was applied to the full dataset in the original study. This was because the statistics derived from this technique were more strongly correlated with survival, and therefore of higher clinical value. Note that because the 50-point sampled ROIs were applied to the full dataset and that the

original manually-scored study used the 3x3mm box annotations to confirm the TSR as a prognostic marker, the same dataset was used with this work.

Tumour cell density can be calculated in multiple ways, and for the original study (described at the start of this section), the manual analysis of the QUASAR dataset was calculated using Method 1 from Table 7 to identify the TSR as a prognostic indicator in CRC.

Method	Formula
1	$\frac{\sum T}{\sum T, S}$
2	$\frac{\sum T_{tumour}}{\sum T, S}$
3	$\frac{\sum T_{tumour}}{\sum T_{tumour},\ S_{stroma}}$

Table 7 - Methods for calculating TSR

Where $T \ni \{tumour, lumen, necrosis, mucin\}$ *and* $S \ni \{stroma, muscle, vessels, inflammation\}$.

The tumour and stroma parent classes each have four subclasses, which are described in Figure 33.

Method 1 is calculated as the ratio of the count of all tumour subclass observations to all tumour and stroma subclasses within a sampling area, and is referred to as the Tumou:Stroma ratio (TSR). Method 2 is calculated as the ratio of the total number of tumour subclass observations to all tumour and stroma subclasses, and is referred to as Tumour Cell Density (TCD). Method 3 is the ratio of the tumour subclass to the tumour and stroma subclasses only, ignoring all other data points. This method discards 8% of the data.

3.3.2.5 Feature analysis for image processing

Using the ground truth data (see section 3.3.2.1), the TSRs were plotted against six image intensity features, so that correlations between features and TSRs could be assessed as models for predicting TSR. The intensity features used were:

- Mean image HSV intensity
- Standard deviation of image HSV intensity
- Mean intensity of the stain deconvolution H channel
- Standard deviation of H channel intensity

- Mean intensity of the stain deconvolution E channel
- Standard deviation of E channel intensity

Features were calculated describing the overall sampling area, at a fixed zoom of 6x (2000x2000 pixel images from 20x slides at 0.5 microns per pixel). Figure 35 shows hex scatter plots of TSRs against mean and standard deviation of intensities for the cases (HSV intensity and stain intensity for deconvoluted H&E staining).

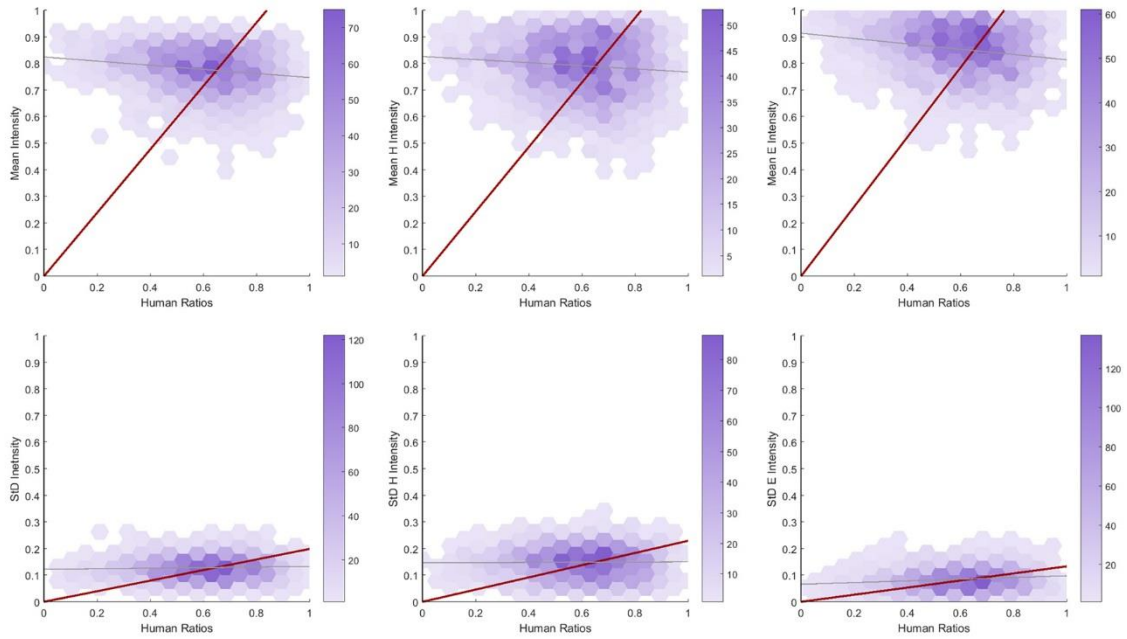


Figure 35 - Hex scatter plots of slide features compared to TSRs

Each figure plots TSRs along the x axis and the image feature values along the y axis
Heatmaps indicate the density of distribution with darker colours representing higher numbers, referred to by their scales

The resulting plots show no correlations between TSR and image features, and therefore it is concluded that these features alone are not sufficient for predicting TSR.

3.3.2.6 Thresholding

It was hypothesised that, due to the higher cellular density of epithelial tissue, tumour and stroma would be separable using thresholding on the Haematoxylin intensity image, generated by colour deconvolution (see section 2.4.2). This is an over-simplified observation that appears true in a large proportion of the observed images.

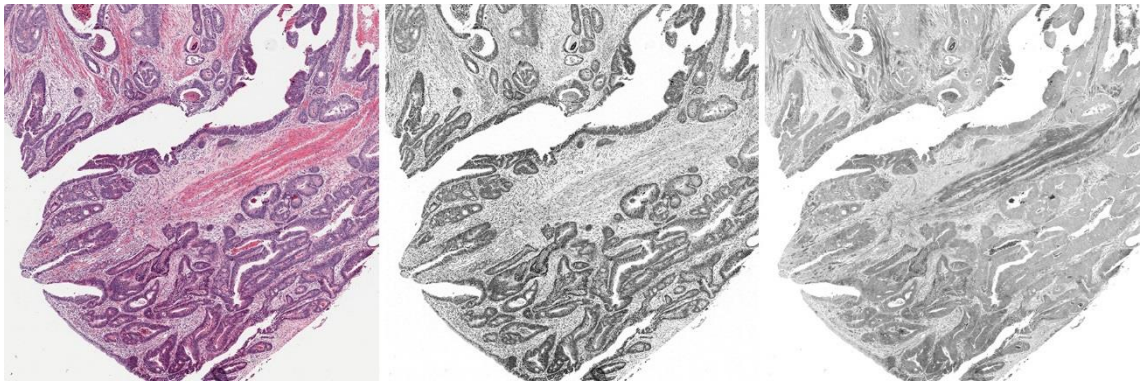


Figure 36 - Colour deconvolution applied to an example image from the dataset

Left: original image
Centre: The Haematoxylin channel
Right: The Eosin channel

The image size in all cases was a 3x3mm square, at 20x (0.5mpp) resolution, meaning that fibroblasts (stromal nuclei) would be included if the H channel image were simply thresholded. By using a simple median filter (size, 20x20 with symmetric border replication), these nuclei were filtered out as noise. Figure 37 illustrates that the median filtering process has the capacity to transform the pixel intensity values into an easily thresholdable bimodal distribution.

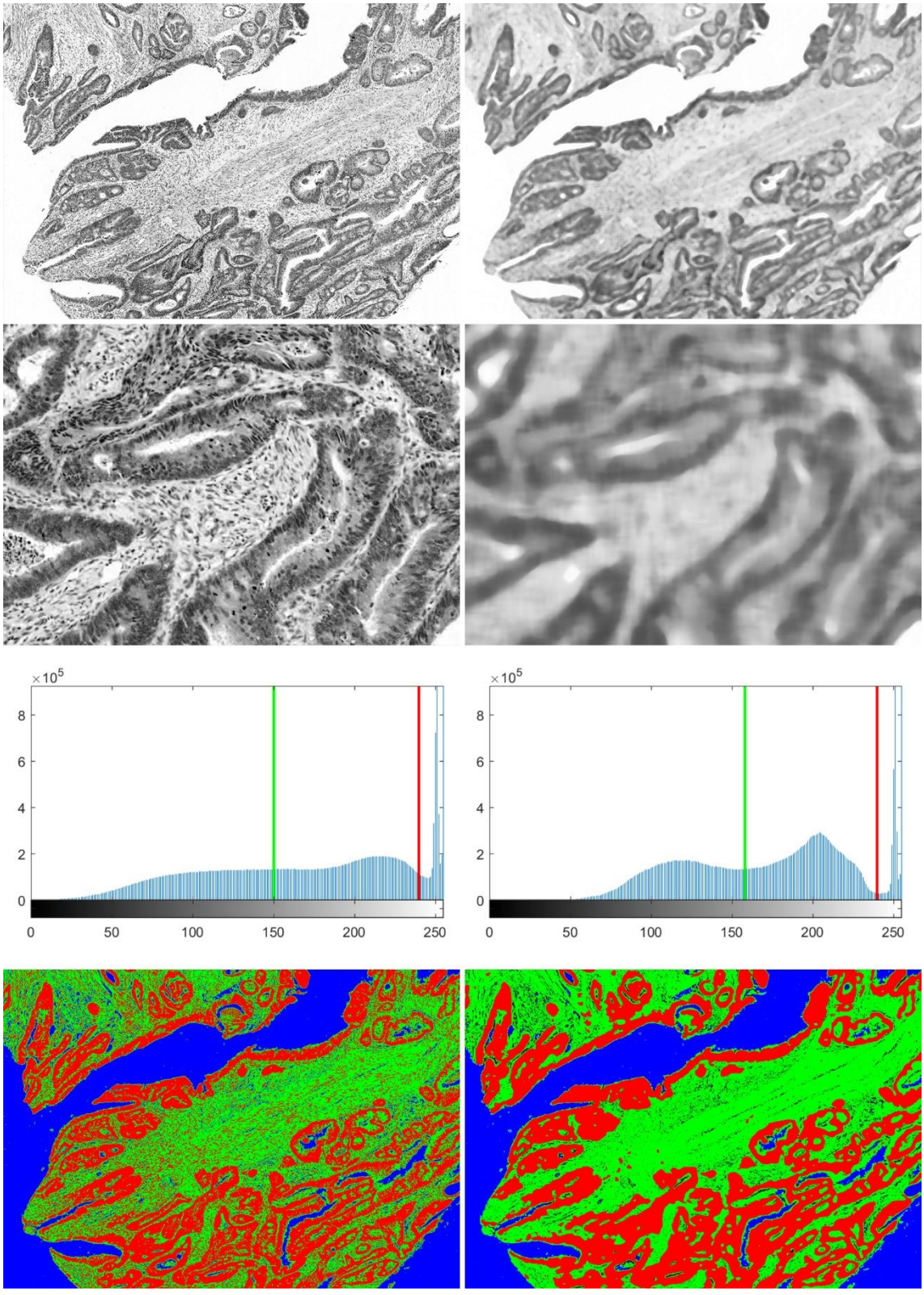


Figure 37 – Threshold comparison with and without median blur for filtering out stromal nuclei

Left: Haematoxylin image channel from Figure 36, thresholded using the method detailed in Equation 8 to separate tumour (red) from stroma (blue), and a static threshold to separate background (blue) of 240
Right: The original image is median filtered with a size of 50 (see 2.4.3).
Blurring the image affects the distribution of intensity values in the image histogram.

After using median filtering as a pre-processing step, the image was thresholded using the mean intensity of the non-background pixels (threshold = 240 intensity) – this assumed that the image had a bimodal distribution of pixel intensities, and that the bimodal distribution was approximately equal for both peaks. This means that images that are not well stained, or have unequal amounts of tumour and stroma will most likely not be analysed accurately using this technique.

After thresholding, the resulting image mask, which provided an area of candidate tumour pixels. The ratio of tumour to stroma could then be counted as the number of tumour pixels to the number of non-background pixels. For visualisation purposes, the mask was applied to the haematoxylin image, and a colour heatmap applied to it. The haematoxylin channel was used again as an alpha (transparency) map to create an overlay for the original RGB image (see Figure 38). Note that the threshold markup image combines the foreground tissue mask in the blue channel and the tumour mask in the red channel to create an RGB image, where co-localised foreground and tumour is represented as magenta.

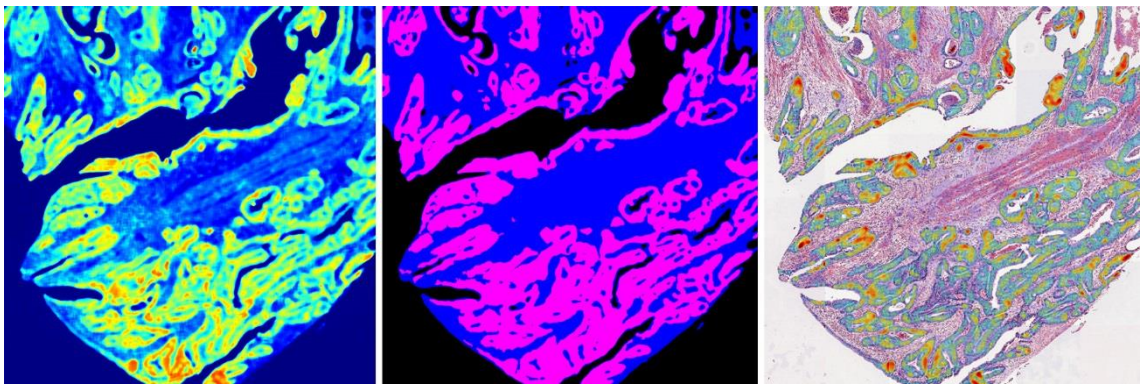


Figure 38 – Heatmap of haematoxylin stain intensity after median filter and threshold

Left: Blurred image heatmap
Centre: Thresholded image – magenta is tumour, blue is stroma
Right: The markup image applying the tumour mask to the heatmap, and overlaying the original image with an alpha (transparency) map equivalent to the Haematoxylin stain intensity

3.3.3 Results

3.3.3.1 Dataset observations

Using the full dataset from RandomSpotDB (section 3.2.2.5), the data were analysed to assess the distribution of classifications given. Figure 39 shows that out of all 360,698 expert-classified locations for the QUASAR dataset, 203,616 (56%) belonged to the Tumour class (Tumour,

Lumen, Necrosis and Mucin subclasses), 129,119 (36%) belonged to the Stroma class (Stroma, Muscle, Vessels and Inflammation), and 27,963 (8%) of them were non-informative.

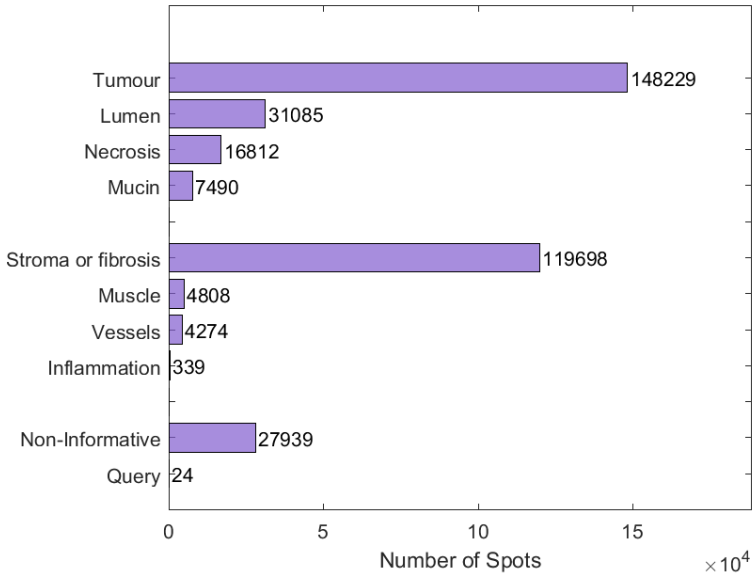


Figure 39 - Bar plot of all available data for the QUASAR set, grouped by classification

The bars are grouped by their parent class (top group being tumour, and middle group being stroma). The dataset has more examples of tumour than stroma, which may affect the training of ML algorithms.

Each of the 2,211 cases in the dataset had been scored using two of the four sampling methodologies described in Table 6: 50 spots within a 3x3mm box annotation over the area with the highest tumour cell density, and 50 spots within the whole tumour boundary. These data were separated and analysed independantly for their TSRs per case (Figure 40).

Method 1 was implemented in the original manual spot counting analysis study, applied to the QUASAR dataset (see 3.3.2.1), in order to generate the most statistically significant prognostic markers [52]. Therefore, to maintain comparability between the results of the original study and the algorithm output, the same method was applied to all subsequent analyses of this dataset.

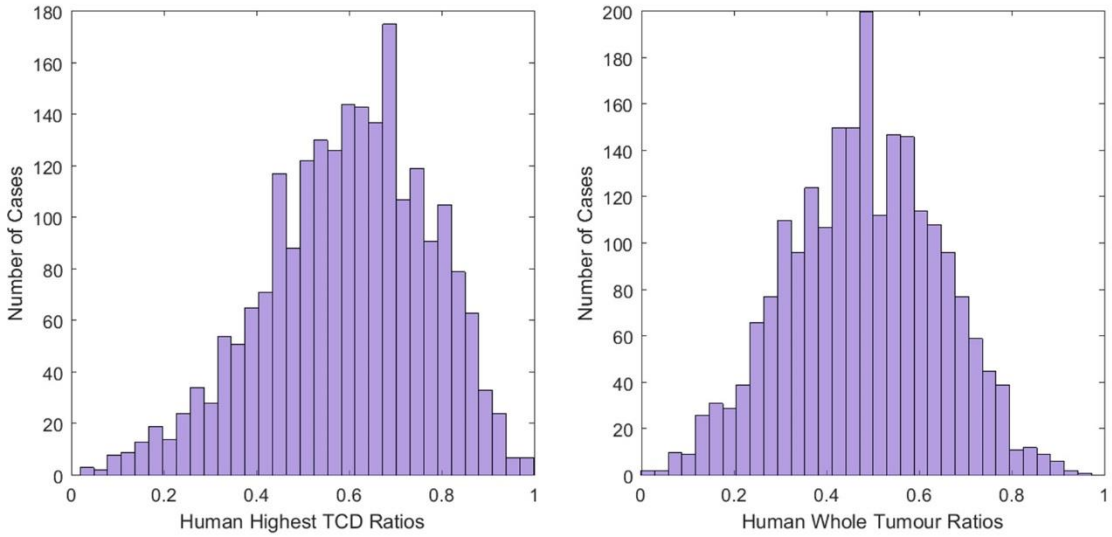


Figure 40 - Histograms of TSR distribution per case for the two pathologist sampling methods

Left: 3x3 mm box over the highest area of TCD (mean = 0.6, standard deviation = 0.18)
Right: Whole tumour region (mean = 0.48, standard deviation = 0.16)
The distribution shows that sampling ROIs placed over the perceived area of highest TCD yield higher TSR values.

The distributions illustrate that ROIs sampled in the areas of highest perceived TCD yield higher TSR values. For the highest TCD and whole tumour datasets, the means are 0.6 and 0.48 resepectively. The highest TCD dataset has a standard deviation of 0.18, whereas the whole tumour dataset has a standard deviation of 0.16. This indicates that the whole tumour dataset is more normalised, wherease the highest TCD dataset predicts higher TSRs, but with a stronger skew to the left of the mean.

3.3.3.2 Correlation of TSRs using different manual sampling methods

Using the highest TCD dataset as the ground truth, the remaining dataset (whole tumour) was used to compare the TSRs for both methodologies. The graphs in Figure 41 compare both sets of ratios generated for each case.

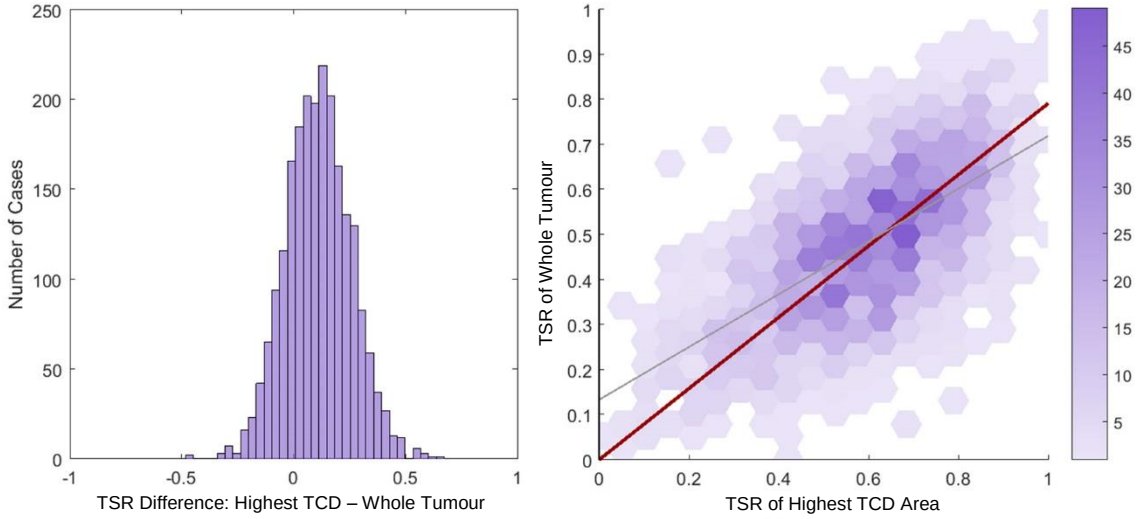


Figure 41 - Comparison charts of ground truth for the two pathologist sampling methods

Comparisons of 50 spot sampling method, using 3x3mm box over highest TCD area, to the whole tumour area (left) histogram of ratio differences per case (mean = 0.11, standard deviation 0.15) and (right) hex plot scatter heatmap (R^2 = 0.41).
The figures show that the relationship between the two sampling methods is consistently skewed due to the area of highest TCD having higher TSRs than whole tumour areas.

Using the differences per case (between sampling methods, highest TCD area minus whole annotations area), the histogram shows that the whole annotations underestimate the proportions of tumour. The comparison dataset has a mean bias of 0.11, and a standard deviation of 0.15. These results indicate that the whole tumour sampling method identifies a lower proportion of tumour by comparison, which is as expected, due to the location of the box annotations being at the area of highest TCD.

A paired samples T-Test rejected the null hypothesis that the pairwise mean comparison is equal to zero ($p < 0.01$). A Spearman's nonparametric rank correlation test confirms that a positive relationship extists (rho = 0.61) and that the correlation is significantly different from zero ($p <$

0.01). The fitted linear regression model (grey line on the plot) to the dataset has a coefficient of determination (R^2) of 0.41.

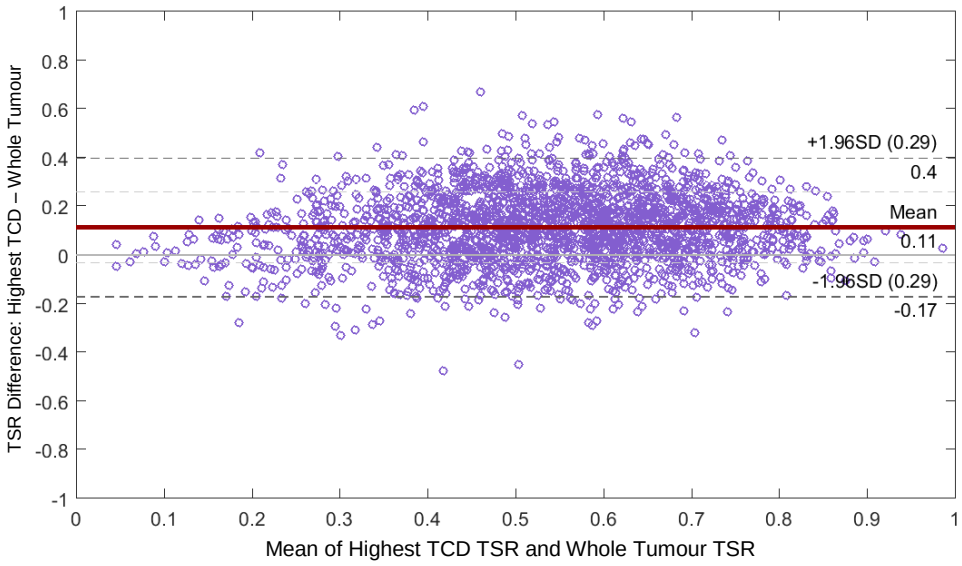


Figure 42 - Bland-Altman plot of TSRs generated by the two sampling methods

The BA plot shows that the differences between the two sampling methods have a large spread (C.I. = 0.29), but there is no obvious point in the distribution where this spread is likely to be larger or smaller. This means that the two methods have consistently high levels of variance, indicating that either human scoring methods are not consistent, or that cancers with high TSR in the highest area of TCD may not necessarily have high TSR overall.

The Bland-Altman plot in Figure 42 shows that the data has a bias of 0.11, and that the 95% confidence intervals (1.96 standard deviation, indicated by the outer dashed lines) are ±0.29 [255]. The mean bias of 0.11 indicates that the highest area of TCD is has 11% more tumour than the overall whole cancer ROI. The data shows that cancers with high TSRs in the area of highest TCD do not necessarily have high TSR rates over the whole tumour area. The spread of data gives an indication of how much this figure may vary from case to case, but may also be subject to human inconsistencies with scoring.

3.3.3.3 Image analysis results observations

Using the methods described in 3.3.2.6, each of the cases from the dataset was processed to identify tumour and stroma tissue, whereby their TSR was recorded, and a markup image denoting the presence of tumour was saved for visual inspection. Examples of some of the markup images are shown in Figure 43.

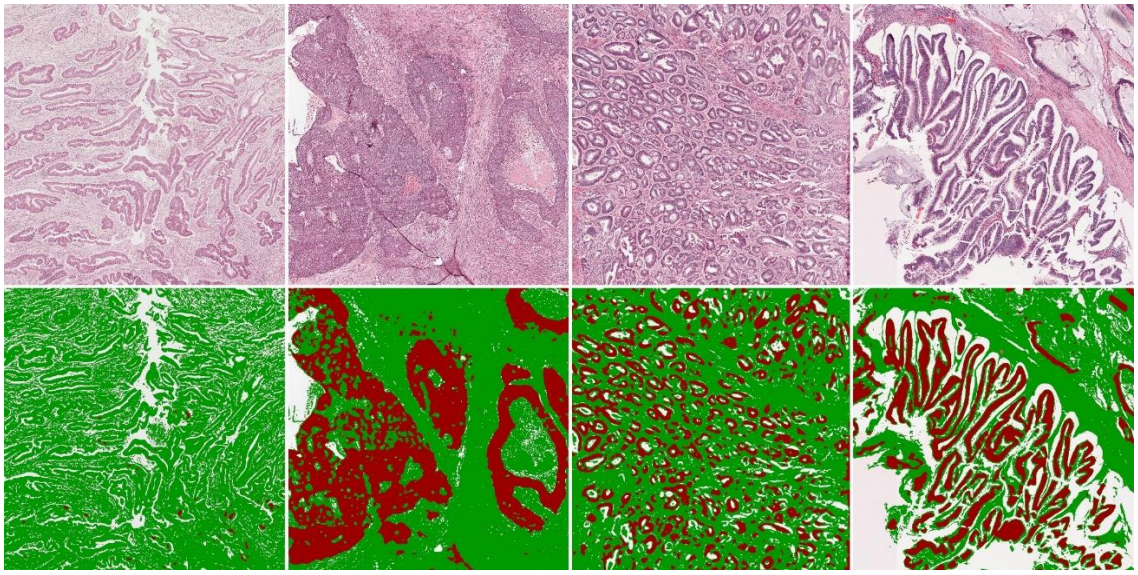


Figure 43 - Examples of image analysis markup from the initial thresholding-based algorithm

Top: Original images
Bottom: Markup images.
Green indicates tissue detected as stroma, red indicates tissue detected as tumour and white indicates detected background.
The examples show that using the thresholding method is not appropriate for images with lower contrast between tumour epithelium and stroma (poorer staining).

Overall the algorithm appeared to robustly identify tumour areas, but due to the median filtering of the haematoxylin channel, smaller areas of tumour, or transversely sliced glandular structures lost edge fidelity and were detected at smaller sizes. Areas of tumour lumen that contained mucin or debris were often detected as stroma due to their lighter appearance, instead of their position in relation to the tumour gland structures. Also, images that were poorly stained in terms of haematoxylin did not yield accurate detection results (example shown in top left of Figure 43).

3.3.3.4 Correlations of TSRs between manual and automated methods

Each of the cases from the full dataset were analysed using the techniques described to generate a percentage of tumour and stroma ratio (TSR) per case. This ratio was compared to ratios generated by pathologist scoring from the original study, and comparisons were made by subtracting the algorithm-generated ratio from the manually generated ratio per case.

Using the highest TCD spots as the ground truth, the algorithm-generated ratios were compared in order to assess how the algorithm results correlate with the manual data. The graphs in Figure 44 compare both sets of ratios generated for each case.

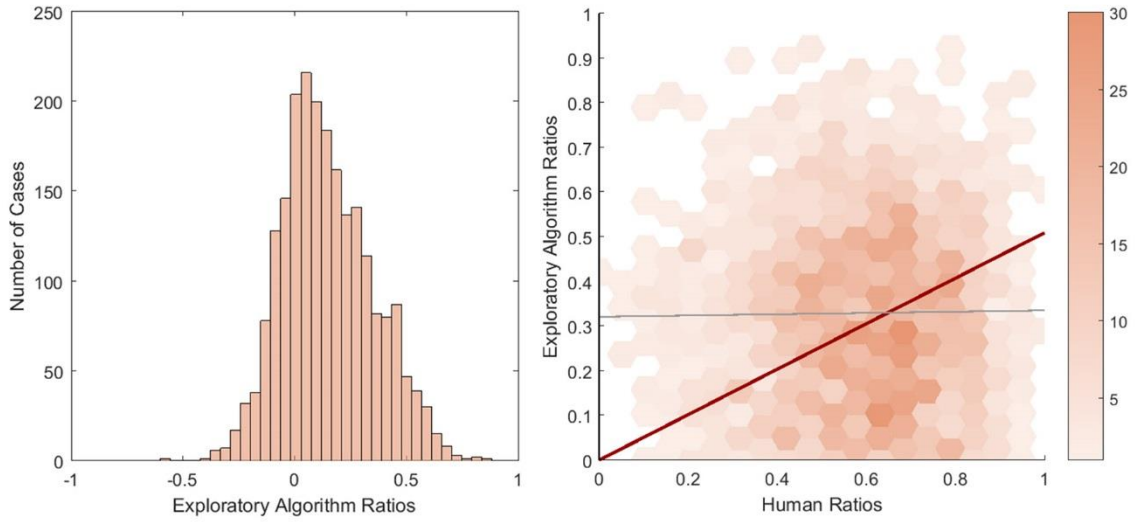


Figure 44 – Comparison plots of pathologist scores to thresholding-based algorithm

Left: Histogram of TSR differences per case
Right: Hex scatter plot showing distribution of the ratios
The histogram shows that the algorithm has a positive bias, meaning that it underrepresents tumour in the analysis. This is due to the median filter blurring small areas of tumour into larger areas of stroma.

Using the differences per case (ground truth highest TCD method, minus exploratory algorithm method), the histogram shows that the automated method underestimates the proportions of tumour. The comparison dataset has a mean bias of 0.27, and a standard deviation of 0.20. A paired samples T-Test rejected the null hypothesis that the pairwise mean comparison is equal to zero ($p < 0.01$).

The distribution of ratio differences indicates that the algorithm underestimates the proportions of tumour. This is likely due to the naïve assumption of the algorithm, that intensity is an appropriate feature for independently segmenting tumour from stroma. Issues with this assumption are explored in 3.4.3.3, however, the blurring method may also be too aggressive in cases where stains were weaker – such that the contrast between the resulting areas was too low to identify a useful threshold. Also, areas of lumen in the tissue are considered part of tumour tissue in the ground truth, but applying larger average filters over them raises the intensity over the overall area, which would cause the algorithm to incorrectly predict stroma instead.

A Spearman's nonparametric rank correlation test confirms that a positive relationship extists ($rho = 0.69$) and that the correlation is significantly different from zero ($p < 0.01$). The fitted linear regression model (grey line on the plot) to the dataset has a coefficient of determination (R^2) of < 0.01.

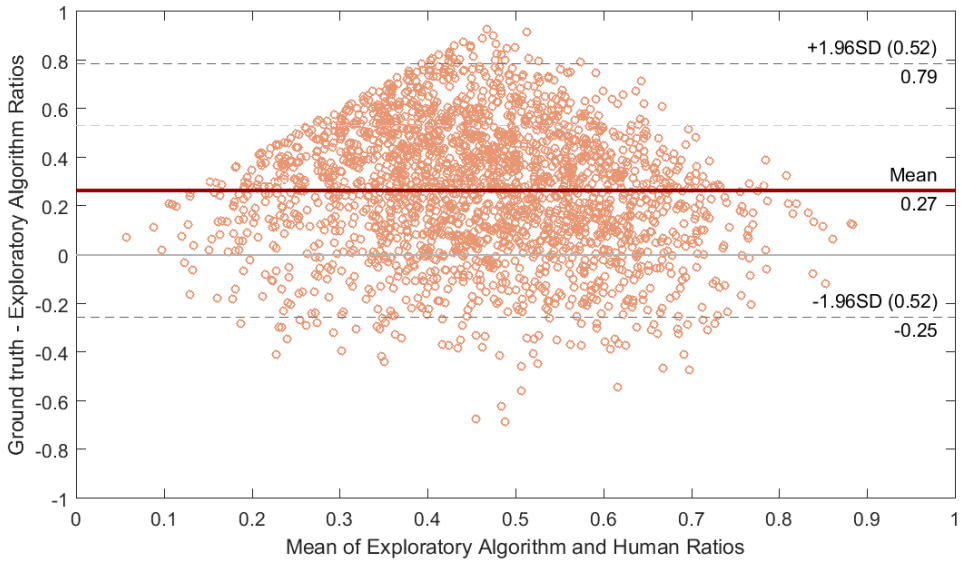


Figure 45 - Bland-Altman plot comparing TSRs generated by the algorithm and the manual analysis methods

The distribution of the BA plot shows the data has a high variance (C.I. = 0.52), with a positive bias, indicating the algorithm underrepresents stroma. The distribution also shows that the algorithm is likely to perform more poorly on cancers with a TSR of approximately 0.45, with better agreement on cases with more polarised values.

The Bland-Altman plot in Figure 45 shows that the data has a bias of 0.27, and that the 95% confidence intervals (1.96 standard deviation, indicated by the outer dashed lines) are ±0.52. The combined bias and spread of the distribution indicates that the algorithm is not an adequate substitute for human scoring, but provides a baseline with which to identify issues with performance and mitigate them.

3.3.3.5 Poor algorithm performance

The dataset was processed using the described methodology from 3.3.2.6 and observations were made where the algorithm performed poorly. Poor performance in this instance is defined as results from images where the difference between pathologist-generated TSR and algorithm-generated TSR falls outside of the 95% confidence intervals, shown in the resulting Bland-Altman plot of TSR difference distributions (see Figure 45). Observations were made by manually inspecting the markup of the processed images. Four recurring issues were identified:

1) Slides that are stained inconsistently compared to the majority of the dataset
2) Tissue that has a strong lymphocyte immune response
3) Images that had technical artefacts upon retrieval from the server
4) Images containing tissue that does not have the characteristics of a stage II cancer

These issues are addressed independently in the following sections.

Issue 1: Staining inconsistencies

Inconsistent staining affects the threshold levels for object detection and the output images from colour deconvolution. This means that either the thresholding method should compensate for such variation, or the images should be adjusted before analysis. Colour normalisation [165] provides a robust solution to staining variation (see section 2.5.2), and was applied as a pre-processing step. However, despite colour normalisation's robust method for correcting stain variation, there are small numbers of cases where the images have very little colour information (staining is faded or too weak), or the colour information is too far outside the expected range (extreme overstaining in one or two staining channels).

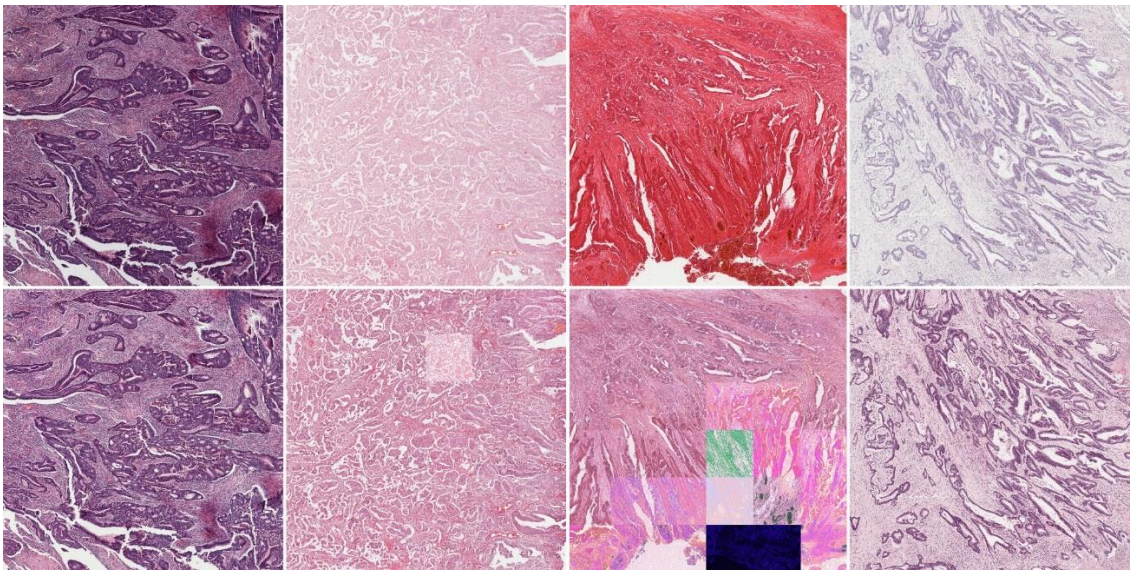


Figure 46 – Examples of colour normalisation on sub optimal images

Top: Examples of slides with sub-optimal staining
Bottom: The slides with colour normalisation applied in 1000x100 pixel blocks

Figure 46 shows examples of colour normalisation being applied to suboptimal images, in order to assess its suitability. Note that the block artefacts are created due to the normalisation being applied to 1000x1000 pixel image patches before concatenation of the final image. The examples show that colour normalisation is appropriate for use with images that are both too weak and too strongly stained, however some extreme cases may not be correctable. Applying colour normalisation to the whole slide prior to analysis may provide better colour correction than block-based normalisation, however this was not possible with the software available.

It is concluded that the application of colour normalisation should, for the most part, improve the overall consistency of slide images to allow more consistent analysis.

Issue 2: Lymphocytic infiltrate

Lymphocytes are immune cells that provide an inflammatory response to cells that are being attacked by viruses and diseases. These cells are dark, round and consistent in appearance, and appear in dense clusters. Inflammation in CRC is a sub class of the stroma category, and breaks the naïve assumption of the algorithm, that areas of tumour are darker than stroma. The density of lymphocytes in inflammation means that unlike stromal fibroblasts, lymphocytes are removed by median filtering alone.

Lymphocytes are typically consistent in size, shape and colour, with dark appearance and well-defined cell boundaries. By creating an edge (strength) probability map, the edge information can be used to identify areas with high levels of edge strength and subtract them from the tumour probability (median filtered haematoxylin image) map.

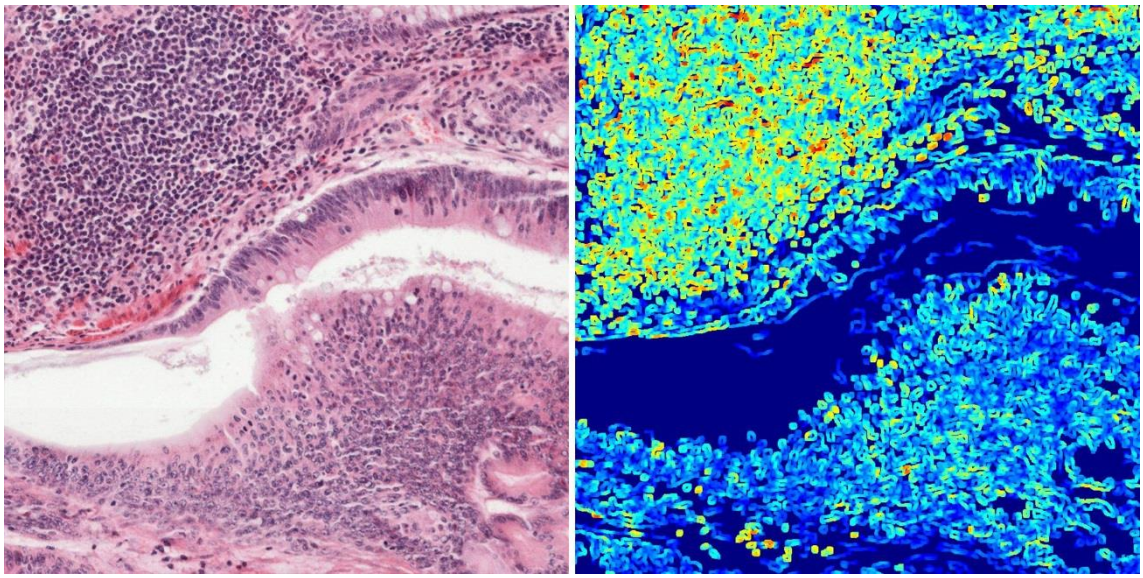


Figure 47 - Example of lymphocytic immune response

Left: Illustration of lymphocytes (top) having lower intensity than tumour (bottom), making simple intensity thresholding inaccurate
Right: Maximally filtered edge strength map using canny edge detector, showing lymphocytes have higher edge strength

Firstly, the image is filtered using a canny edge detector, and then to amplify the areas with strong edges, a maximum filter (size [5 5]) is applied. This image is subtracted from the tumour probability map.

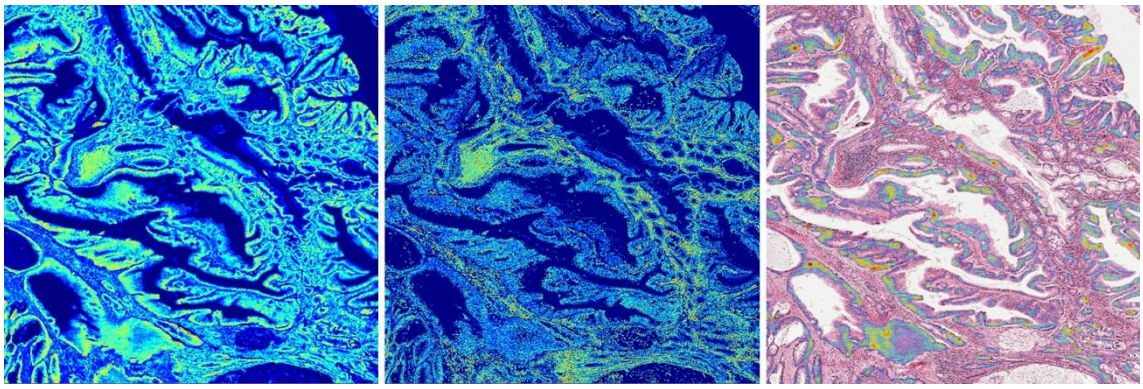


Figure 48 – Removing lymphocyte cells from analysis using edge intensity

Left: Stain intensity heatmap
Centre: Edge intensity ieatmap
Right: Original image, overlaid with subtracted heatmaps, removing areas of lymphocytes

The example images in Figure 48 illustrate that when subtracted from the haematoxylin stain image, the edge probability map has the capacity to remove lymphocytic infiltrate from the tumour probability map. Accounting for edge strength in future algorithm versions should be useful for mitigating this issue.

Issue 3: Image retrieval artefacts

Images retrieved over http via the Leeds digital slide server have an inbuilt size cap of 2000x2000 pixels. The larger the images retrieved, the more load placed on the server, which also causes more issues with latency and digital artefacts. In order to read in and process the 3x3mm ROIs (20x zoom or 0.5 microns per pixel), it is sensible to process the image in blocks.

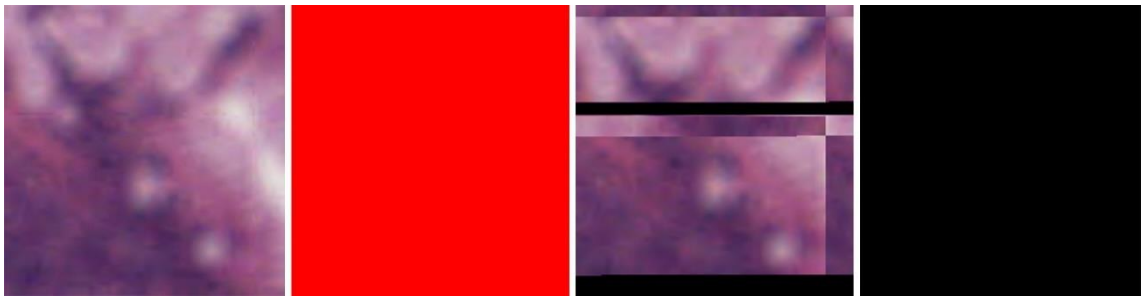


Figure 49 - Examples of image retrieval failure

Left to Right: Original (out of focus) 256x256 pixel image – out of focus but as intended, image retrieved over http using Aperio ImageServer (caused by memory leak), image retrieved over local filesystem using OME BioFormats package, image atypical

To reduce the overall memory requirements of the algorithm, the images were split into 1000x1000 tiles using a sliding window methodology with an additional 100-pixel border and concatenated. This methodology created two issues:

1) Sometimes tiles would be retrieved from the server without any visual information (all 3 of the RGB channels set to 0)
2) Colour normalisation is applied per image tile and so areas with little colour information are overcompensated (see 2.4.2).

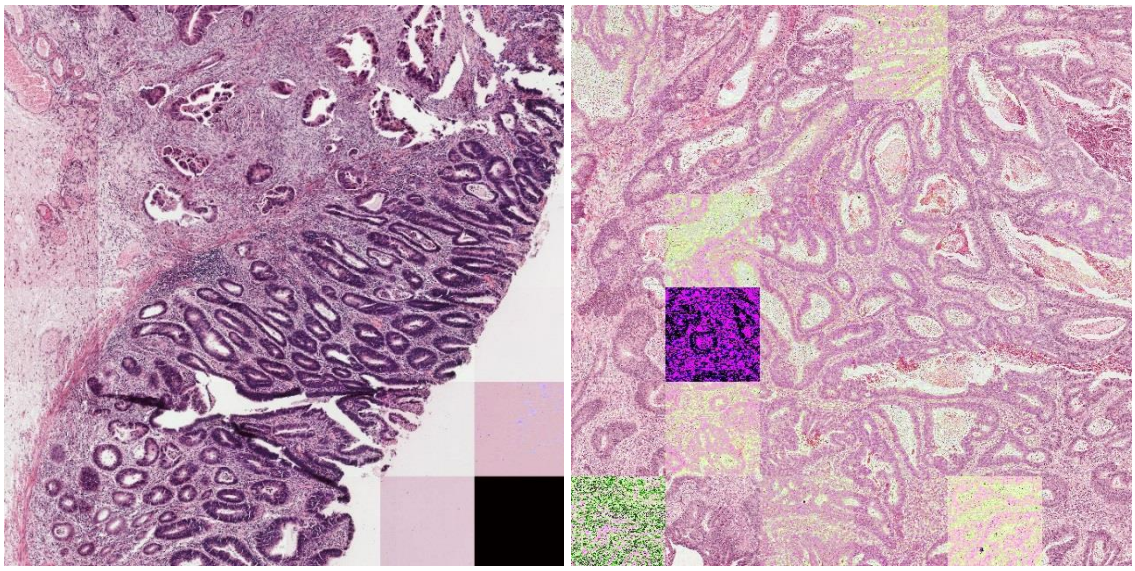


Figure 50 - Examples of artefacts caused by processing images in tiles

Applying colour normalisation to sliding window areas means that image tiles are corrected using local colour information (as opposed to global), which has the capacity to alter the processing results per block.

Issue 4: Pathological factors affecting analysis

Upon visual inspection, some of the cases in the dataset were negatively affecting TSR generation due to pathological factors. Examples of these include mucinous cancers, where large amounts of mucin are present in the images, poorly differentiated tissue, which has little to no structure, and appears homogeneous, and necrotic tissue, where cells have died and are scattered. In all of these cases, the assumption that tumour and stroma are visually separable using intensity is broken, meaning that the thresholding algorithm under performs on these images. Figure 51 illustrates examples of mucinous, undifferentiated and necrotic cancers, which do not contain any discernible glandular structures.

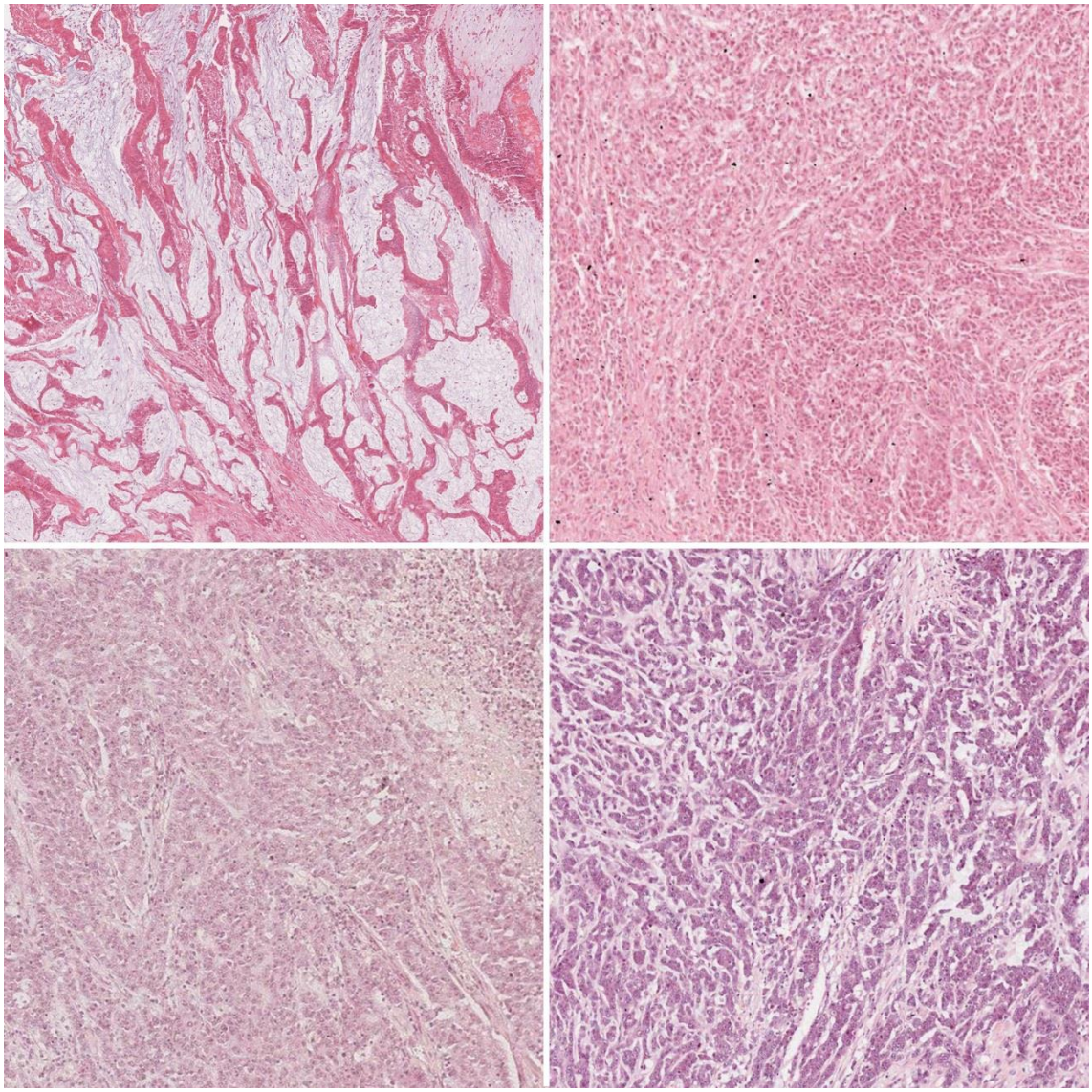


Figure 51 - Examples of tissue in the dataset that negatively affected algorithm performance

Top Left: Mucinous adenocarcinoma
Top Right: Undifferentiated tissue
Bottom Left: Undifferentiated tissue
Bottom Right: Necrotic tissue
The illustrated examples show that the pathological reasons for affecting TSR are likely due to their lighter and more homogeneous appearance. The lack of contrast between glands and stroma mean that thresholding assumptions based on structural differences within a given image are broken. Note that in all four cases, these images are considered to contain mostly tumour.

These images show examples of atypical cancers in stage II cancer patients, but to maintain completeness of the dataset, were not removed from the analysis. The appearance of these cancers has the capacity to negatively affect the results of image analysis algorithms due to their deviation in appearance to typically CRC tissue, and their homogeneous appearance. The lack of structural information in the images makes the task of gland segmentation very difficult, and there is very little contextual information, even given the large field of view (2000x2000

pixels). An in depth look at images that negatively affect image analysis performance id undertaken in Chapter 6.

3.3.4 Conclusions

The results presented are consistent with the methodologies that generated them. Sampling the highest area of TCD generates higher TSRs than whole tumour areas, creating a mean difference bias skewed towards 1.

The spread of data indicated by the confidence intervals in Figure 42 shows that the distribution of differences has a high amount of variance, which supports the motivation for this work, that manual scoring methodologies are subjective and largely inconsistent.

The area of highest TCD within a 3x3mm box annotation is used in previous publications, due to this method having the strongest correlation with survival this method will be the gold standard for further work.

The algorithm performs basic processing that accommodates for some but not all of the complex histopathological features.

The performance of the algorithm has a mean difference of +0.15, indicating that the algorithm underestimates the presence of tumour. This is confirmed by the visual inspections made of the markup images in section 3.3.3.3, and is due to smaller or thinner epithelial areas being filtered out as noise, as well as mucin and luminal debris being counted as stroma.

The spread of the distribution of ratio differences would appear to be too wide to make the algorithm of clinical use without further improvements. These improvements could be made by using the existing ground truth data to train a ML algorithm, in order to learn the appearances of different tissue types. This approach could reduce the false negative rate by identifying tumour lumen and excluding it from analysis. Section 3.4 adopts this approach.

3.4 Algorithm A: Automated SRS using a fixed patch size

3.4.1 Aim

To develop a ML algorithm for automatically generating TSRs, using the QUASAR dataset as the ground truth data for training.

3.4.2 Methods

The image data for the QUASAR clinical trial dataset, obtained using the RandomSpot system (detailed in section 3.2), was used to develop a ML based image analysis solution. Instead of using the whole 3x3mm annotation (as per the previous algorithm), the original image spot classifications were taken individually in order to extract an image patch using the spot as the centre coordinate. Using this methodology, each of the 2,211 cases had approximately 50 pre-scored spots, which were classified as one of the eight tissue types (illustrated in Figure 33). The ninth classification that was used for non-informative spots, which represented areas that could not be scored, such as non-luminal background pixels, or histological and digital artefacts, was discarded.

3.4.2.1 Processing pipeline

To ensure consistency in processing methodologies, and maximise the reliability of results, each algorithm that processed the RandomSpotDB datasets using ML used the following processing pipeline as a template for analysing tissue images. Any deviations from or extensions to the pipeline are explicitly mentioned in the corresponding methodology sections.

1. Image retrieval

Image co-ordinates were retrieved from the RandomSpotDB, and extracted at a given size, with the ground truth label lying at the centre of the image. Image patches were retrieved directly from the Leeds digital slide server hard drives using the OpenSlide library as opposed to over HTTP using the Leica-Aperio ImageServer software (see section 2.4.1). Using the QUASAR dataset, 2,211 patient cases were available for processing, consisting of 106,242 image patches,

scored with one of the eight subclasses illustrated in Figure 33. All images in this dataset were sampled using the 3x3mm box annotation described in section 3.3.2.

2. Stain variation correction

Stain variation was corrected using the Magee et al colour normalisation algorithm, detailed in section 2.5.2. To compensate for errors illustrated in Figure 21, an assessment was made using the intensity and saturation values of the normalised image patch, and images that did not meet these criteria were rejected in favour of the original image.

3. Image feature calculation

Image features were calculated for each patch, generating a feature vector consisting of 18 features, listed below:

- Median HSV hue
- Median HSV saturation
- Median HSV intensity
- Percentage of foreground area[2]
- Percentage of haematoxylin deconvolution channel foreground area
- Percentage of eosin deconvolution channel foreground area
- Median intensity of haematoxylin deconvolution channel foreground
- Median intensity of eosin deconvolution channel foreground
- Threshold value of foreground haematoxylin pixels as described by Equation 8
- Threshold value of foreground eosin pixels as described by Equation 8
- Percentage of haematoxylin deconvolution channel foreground area after threshold
- Percentage of eosin deconvolution channel foreground area after threshold
- Standard deviation of HSV intensity
- GLCM (section 2.4.4.1) energy value
- GLCM homogeneity value
- Nuclear detection count, using Bennett et al algorithm (section 2.5.3.1)
- Mean nuclei size
- Ratio of total nuclei area to total image area

[2] Foreground is defined as pixels with saturation > 0.04 and intensity < 240 for HSV colour images, and pixels < 240, or 0.94 for greyscale images.

4. Data splitting

After feature vectors for every patch in the dataset were generated, the data was randomly grouped (by case, rather than by patch) into ten folds for cross validation.

5. ML classifier training

Using the ten groups of feature vectors, a ML classifier was trained using nine of the groups. Unless otherwise specified, a RF algorithm (section 2.4.6.5) was implemented.

6. ML classifier testing

The trained ML algorithm was applied to the left-out group, using the generated model on the unseen data, so that predictions could be made for each of the feature vectors.

The training and testing processes were repeated for each of the ten groups, so that predictions were made for each image patch.

7. TSR generation

TSRs were generated for each case, using all three methods described in Table 7. However, for results analysis, Method 1 was used so that TSRs were comparable to the results of the original study.

8. Results analysis

Results analysis was performed on classification agreement, and TSR differences and correlation. For agreement, statistics were calculated from an 8x8 confusion matrix comparing predictions with ground truth for all sub classes. The matrix was condensed to a 2x2 matrix, grouping the four tumour subclasses and the four stroma sub classes, and agreement statistics were calculated again. Receiver Operator Characteristic (ROC) curves were generated for each tissue subclass, so that Area Under the Curve (AUC) could be calculated and assessments could be made as to which tissue types are modelled appropriately by the algorithm, and which are not.

TSRs were compared using ratio difference, subtracting pathologist-generated TSRs from algorithm generated TSRs, so that negative values indicated that the algorithm overestimated the amount of stroma in the image, and positive values indicated that tumour was underestimated, and 0 indicated perfect agreement. Correlation between algorithm and pathologist generated TSRs were also calculated.

3.4.2.2 Optimising algorithm parameters

With a computer vision algorithm established, three key questions were addressed, in order to optimise algorithm performance:

1. What is the optimum image patch size for processing these images?
2. What classifier provides the highest pathologist agreement?
3. What settings are most appropriate for use with the selected classifier?

For the initial parameter optimisation, the dataset was reduced to tumour and stroma classified image patches only. This reduced the number of images from 106,242 to 80,802.

Also, for the purposes of assessing impact input parameters on per-patch accuracy, TSRs were not generated.

1. Image patch size

Image patches were extracted using the x-y coordinate of each spot as the centre point. To assess the impact of patch size on the accuracy of the algorithm, images were extracted at 16, 32, 48, 64, 128, 256, 512 and 1024 pixels in width and height at native (20x) magnification, and processed as separate datasets. Figure 52 illustrates a scaled example of the image sizes, showing the same spot (with the same x-y coordinate), but with increasing amounts of visual information, going from left to right. Note that the black lines are to help identify the centre of the patch, and are for illustration purposes only.

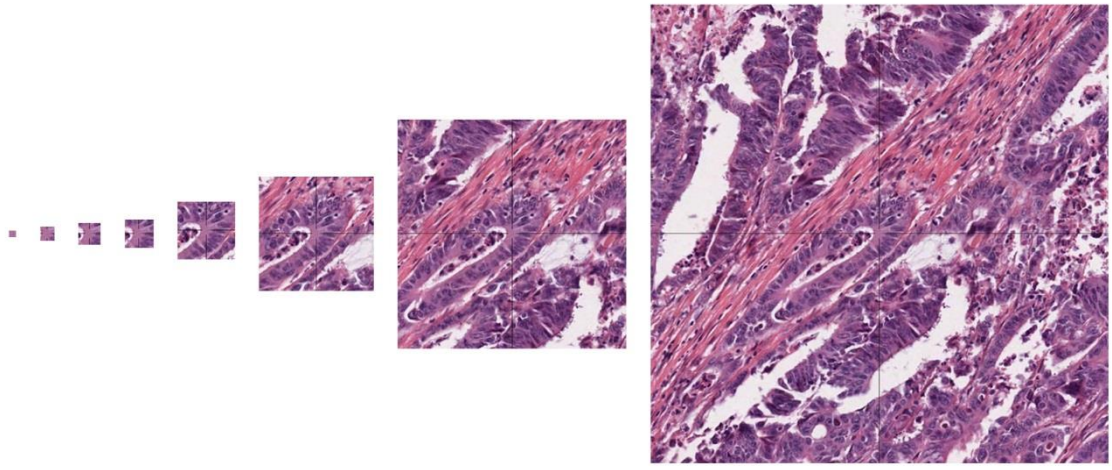


Figure 52 - Image patch sizes, ordered by size

Left to Right: 16, 32, 48, 64, 128, 256, 512 and 1024 pixels in both width and height.
Note that the images are proportionally accurate, but scaled down to fit the document.

Initially, an implementation of the RF algorithm was trained using 300 trees, with three predictors (features) sampled for splitting at each node. The ten-fold cross validation repeated training and testing for each of the seven patch sizes so that agreement statistics could be used to identify the optimal image patch size.

2. Classifier selection

Once the image size with the highest percentage agreement had been identified, several ML algorithms were used for training and testing, to assess their accuracy. In addition to the RF algorithm (section 2.4.6.5), the following four classifiers were used:

- A Naïve Bayes classifier (section 2.4.6.2)
- Logistic Regression (section 2.4.6.3)
- Support Vector Machine (section 2.4.6.4)
- AdaBoost applied to RF (section 2.4.6.6)

All algorithms processed the reduced dataset containing tumour and stroma patches only, extracting images at the most appropriate size, identified by the previous experiment (64x64 pixels), using ten-fold cross validation.

3. Parameter optimisation

With the optimal patch size (64x64 pixels) and ML algorithm identified (RF), an experiment to identify the optimal number of trees for the RF was created, again analysing the reduced dataset of 80,802 patches, and using ten-fold cross validation. The experiment was repeated for 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400 and 500 trees.

3.4.2.3 Analysis using optimised algorithm

The findings from the previous sections are presented in sections 3.4.3, which identify optimal patch size, ML classifier and parameters. These settings are used to create a ML algorithm on the full dataset of 2,211 cases, using a patch size of 64 pixels, and a RF classifier with 100 trees.

Analysis was performed using the pipeline described in 3.4.2.1, on the full set of 106,242 images. Spots classified as non-informative were discarded from the feature set for training.

3.4.2.4 Image processing hardware

Image processing, feature generation and classifier training and testing was implemented in MATLAB, and performed on a VMWare powered Windows Server 2008 R2 64-bit virtual machine with Intel Xeon processors and 32GB RAM. The MATLAB implementation utilised the parallel computing toolbox, assigning 8 processors to parallelise the task.

3.4.3 Results

3.4.3.1 Optimal algorithm parameters

4. Patch size

The algorithm was trained and tested on different patch sizes of 16, 32, 48, 64, 128, 256 and 1024 square pixels, to identify the optimal size for accuracy. Figure 53 contains boxplots for the mean percentages of correctly classified spots per cross validation fold, for each of the patch sizes.

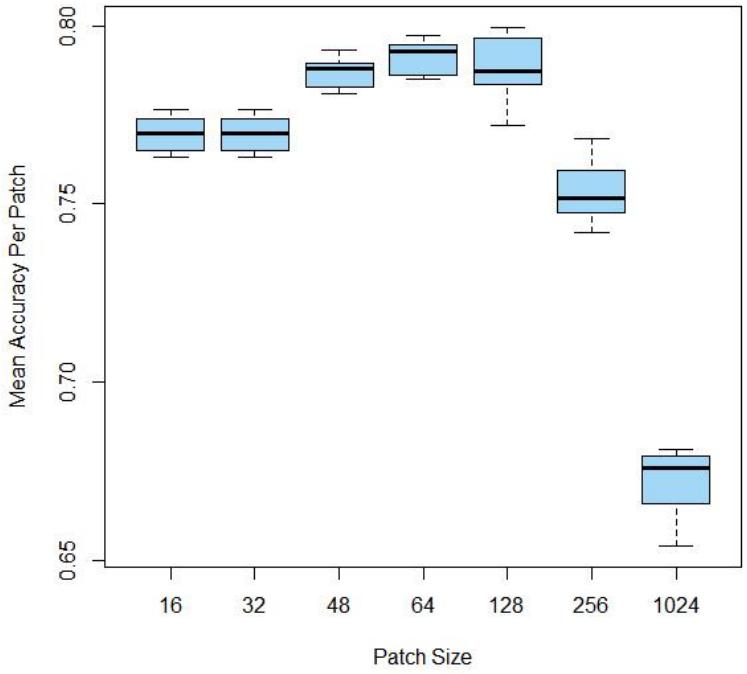


Figure 53 – Results of algorithm accuracy at multiple patch sizes

Note the vertical axis is scaled for readability. Highest mean algorithm agreement observed on 64x64 pixel patches (0.79) and decreases either side. Lowest agreement is recorded on 1024x1024 pixel images (0.67). This indicates that algorithm performance decreases with larger amounts of visual information, which may be due to containing averaged feature values form multiple classes (see discussion).

The accuracy of the algorithm is highest at the patch size of 64 square pixels in size and decreases with larger as well as smaller sizes. Table 8 contains the agreement statistics for the algorithm across all image patch sizes tested. When analysing the 64 square pixel image patches, the mean accuracy of the algorithm was 0.79 (SD < 0.01), has a sensitivity of 0.84 and a specificity of 0.81.

	Patch Size (x)						
	16	**32**	**48**	**64**	**128**	**256**	**1024**
Mean	0.77	0.77	0.78	0.79	0.78	0.75	0.67
STD	<0.01	<0.01	<0.01	<0.01	0.01	0.01	0.01
IQR	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table 8 – Agreement statistics for baseline algorithm on multiple image patch sizes

The table shows the statistical data represented in the boxplots in Figure 53. The algorithm accuracy peaks at 64x64 pixels and sizes 128x128 and above have larger variance.

2. Classifier selection

Once the optimal image patch size was identified at 64 square pixels, the algorithm was trained and tested using multiple classifiers to ascertain which had the best performance.

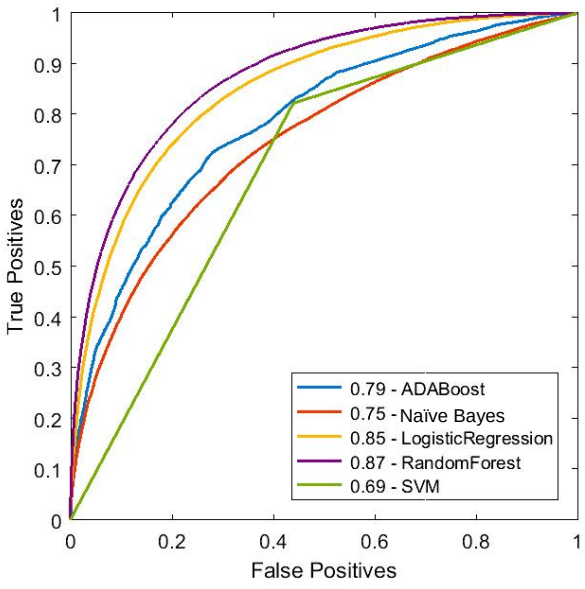


Figure 54 – ROC curves for multiple ML algorithms on the baseline algorithm feature set.

The curves show that the RF algorithm has the highest levels of pathologist agreement. The SVM curve contains one point due to outputting binary class predictions.

The ROC curves in Figure 54 show the performance of the five different algorithms used: Support Vector Machine (AUC = 0.69); Bayesian Networks (AUC = 0.75); AdaBoost applied to RF (AUC = 0.79); Logistic Regression (AUC = 0.85); RF (AUC = 0.87). The RF algorithm had the highest AUC, and so was selected for further analysis of pathologist agreement, and TSR correlation.

5. Parameter optimisation

The selected RF algorithm was optimised by identifying an appropriate number of decision trees. The box plots in Figure 55 show the algorithm-pathologist agreement for each of the ten folds of cross validation, per number of trees tested.

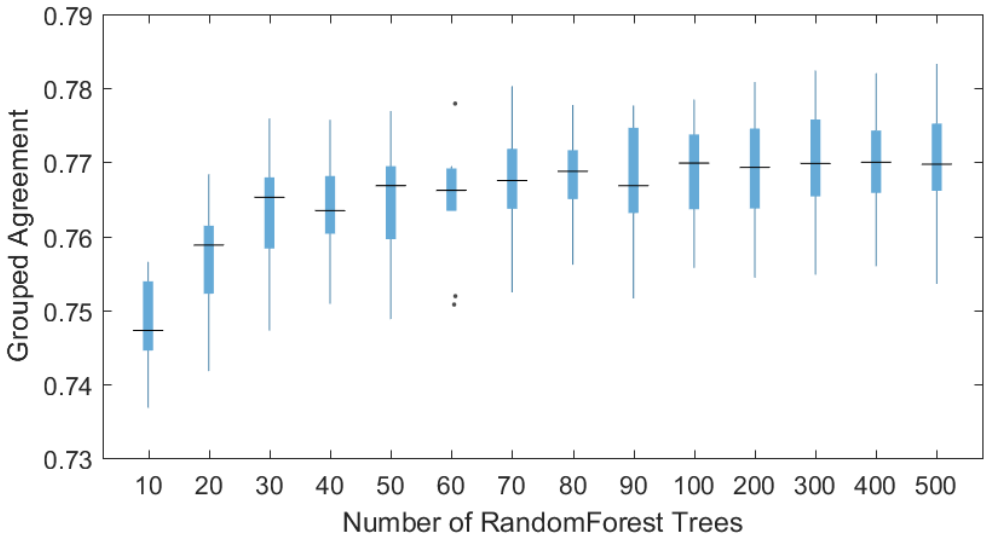


Figure 55 – Boxplots showing agreement between RF classifier and pathologists when using a range of trees to identify grouped agreement
The distributions indicate that using 100 trees provides the most appropriate trade-off between maximising accuracy and minimising computational cost.

In ascending order, the means of each distribution were 0.75, 0.76, 0.76, 0.76, 0.77, 0.77, 0.77, 0.77, 0.77, 0.77, 0.77, 0.77, 0.77, 0.77, 0.77. The minimal differences between values led to the decision to use 100 trees in further experiments.

3.4.3.2 Processing time

The Algorithm A feature set took 6,314 seconds to generate for 64x64 pixel images, using the methodology and hardware previously described. This equates to approximately 0.06 seconds per patch.

3.4.3.3 Agreement

With the optimal parameters identified, a 10-fold cross validation was applied to the dataset of 2,211 cases, and each iteration used 106,242 patches in total, with approximately 95,400 for training, and 10,600 patches for testing.

Agreement was calculated in two ways: calculating per-patch agreement for all eight tissue classes, and grouping the eight classes into their parent class: tumour and stoma. Figure 56 shows the confusion matrices for both evaluation methods of Algorithm A. The left matrix

shows the agreement between algorithm and pathologist for the eight tissue classes. The eight by eight matrix is subdivided into four quadrants, which group the eight classes into the two parent classes, tumour and stroma. The summations of these results into the parent classes are depicted in the two by two confusion matrix to the right.

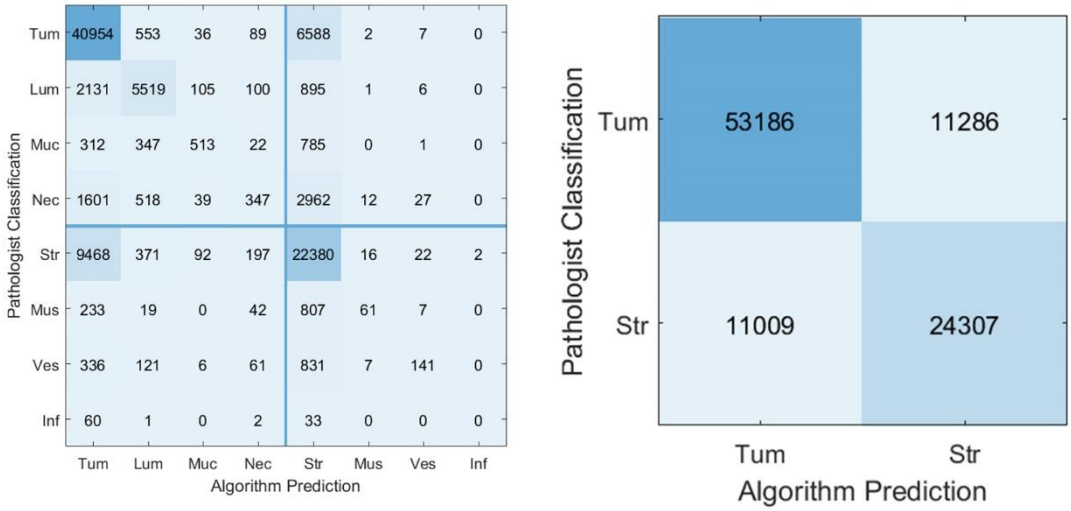


Figure 56 – Confusion matrices showing pathologist – Algorithm A agreement

Left: Confusion matrix of all 8 tissue subtypes from dataset. Accuracy = 70.06%, sensitivity (true positive rate / recall) = 0.70, kappa = 0.51 (moderate agreement)
Right: Confusion matrix of subtypes grouped into Tumour and Stroma parent classes. Accuracy = 77.66%, sensitivity (true positive rate / recall) = 0.82, specificity (true negative rate) = 0.69, kappa = 0.51 (moderate agreement)
The matrices show that the prediction error is not skewed towards tumour or stroma, and that the majority of false predictions are classifying the subclasses tumour and stroma.

The accuracy of the algorithm for the individual tissue classes was 70.06%, with a sensitivity of 0.70 and a kappa [256] of 0.51, indicating moderate agreement. Grouping the individual classes into the parent tumour and stroma classes yielded 77.66% accuracy, a sensitivity of 0.82, a specificity of 0.69 and a kappa value of 0.51, indicating moderate agreement. The algorithm has a higher number of false negatives than false positives (11,009 and 11,286 respectively).

The 2x2 confusion matrix shows that the distribution of false positives and negatives in the grouped results is relatively balanced. This means that the algorithm performance is less likely to be overfitting to a specific class. Looking at the 8x8 matrix shows that he majority of false positives are from incorrect predictions of stroma that should be tumour, and the overall false negative count is caused by tumour being classified as stroma, but also necrosis being classified as stroma. The confusion between tumour and stroma shows that there is a large overlap in the representation of these two classes in the feature space, and further work needs to be done to

separate the two classes. Necrosis can appear lighter and more sparsely textured than tumour due to the break-up of the cells, which is believed to be causing the ambiguity in the classifier between necrosis and stroma. Lumen and mucin are also more likely to be incorrectly classified as stroma than vice-versa, however, the algorithm has a high level of accuracy for detecting lumen. This is due to the simple and distinctive appearance of homogeneous areas of white background.

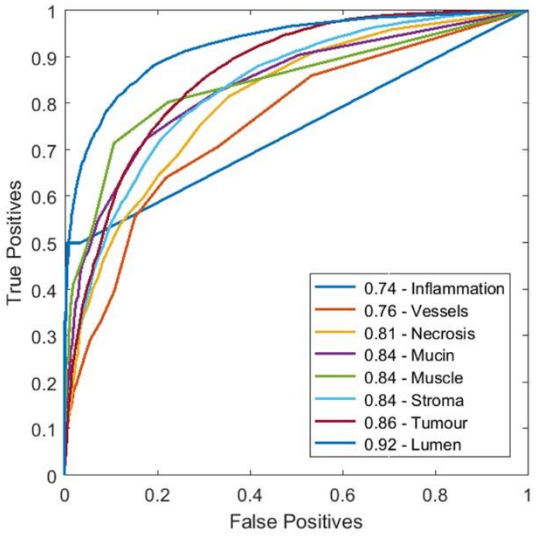


Figure 57 – ROC Curves for all 8 tissue subtypes, classified by Algorithm A

The graph shows Area Under the Curve for each tissue subtype:
Tumour parent class: Tumour (0.86), Lumen (0.92), Mucin (0.84), Necrosis (0.81)
Stroma parent class: Stroma (0.84), Vessels (0.76), Muscle (0.84), Inflammation (0.74)
The ROC curves show that the algorithm performs best on patches of lumen, and poorest on inflammation.

Figure 57 shows the ROC curves for each individual tissue class when classified by the votes-based algorithm. The tumour subtypes have an AUC of 0.86 for tumour, 0.92 for lumen, 0.84 for mucin and 0.81 for necrosis. The stroma subtypes have an AUC of 0.84 for stroma, 0.76 for vessels, 0.84 for muscle and 0.74 for inflammation. The mean AUC for all tissue types is 0.83.

3.4.3.4 Correlation of TSRs

Using the highest TCD spots as the ground truth, the algorithm-generated ratios were compared in order to assess how the algorithm results correlate with the pathologist data. The graphs in Figure 58 compare both sets of ratios generated for each case.

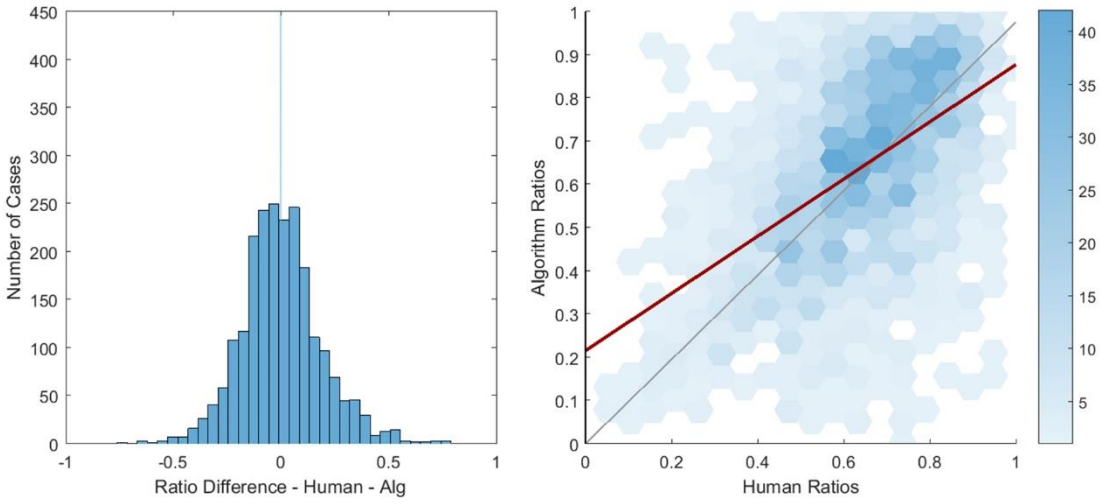


Figure 58 – Comparison plots of pathologist scores to Algorithm A

Left: Histogram of ratio differences generated by pathologist and regular segment algorithm. Distribution has a mean bias of 0 (median 0), and standard deviation of 0.19.
Right: Heatmap Correlation plot of the distribution of the ratios. Correlation has R^2 coefficient of 0.27. The mean bias of zero indicates that the algorithm is generating TSRs comparable to human scoring, but the standard deviation and correlation statistics

Using the differences per case (gound truth highest TCD method, minus baseline algorithm method), the histogram shows that the whole annotations overestimate the proportions of tumour. The comparison dataset has a mean bias of 0, and a standard deviation of 0.19. The distribution of the histogram indicates that the algorithm is not biased towards over estimating tumour or stroma, but the standard deviation indicates that the method is less reproducible than comparing the two pathologist-scoring methodologies observed in section 3.3.3. This may be due to the variability if the data in terms of staining and composition of structural information.

For assessing correlation between the pathologist and algorithm-generated TSRs, a paired samples T-Test did not reject the null hypothesis that the pairwise mean comparison is equal to zero ($p = 0.62$). A Spearman's nonparametric rank correlation test confirms that a positive relationship extists ($rho = 0.50$) and that the correlation is significantly different from zero ($p < 0.01$). The fitted linear regression model (red line on the plot) to the dataset has a coefficient of determination (R^2) of 0.27. The slope of the least squares line in Figure 58 indicates that algorithm performance is likely to overestimate the presence of tumour where it is low, and underestimate it, where it is high.

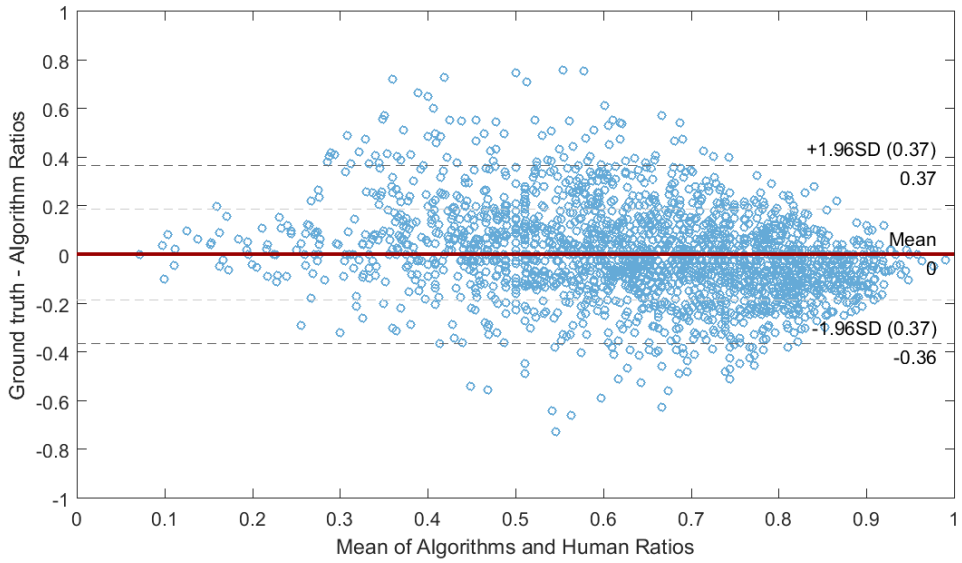


Figure 59 – Bland-Altman plot of pathologist and Algorithm A generated TSRs per case

Distribution has a mean bias of 0, with upper and lower limits of agreement of 0.37 and -0.36 respectively (±0.37).
The distribution shows that mean TSR values are higher than the thresholding-based algorithm and variance is likely to be higher in cases with a TSR of approximately 0.6.

The Bland-Altman plot in Figure 59 shows that the data has a bias of 0, and that the 95% confidence intervals (1.96 standard deviation, indicated by the outer dashed lines) are ±0.37. The distribution shows a smaller spread compared to the thresholding-based algorithm presented in 3.3.2, but still has larger variance than human scoring. Variance is also likely to be lower at more polarised TSR values.

Results figures for Algorithm A are also provided in Appendix D.1.

3.4.4 Conclusions

The baseline ML algorithm out performs the basic thresholding algorithm (described in 3.4) in terms of correlation. The mean bias of ratio differences observed is 0, which indicates that the algorithm TSRs are aligned with pathologist-generated TSRs. However, the spread of the data (standard deviation 0.19, C.I. ±0.37) indicate that although the distribution is centred around zero, the algorithm is not adequately consistent, when compared to pathologist agreement in section (standard deviation 0.15, C.I. ±0.29).

The thresholding algorithm is more likely to underestimate the presence of tumour, with a mean bias of 0.27. Since the thresholding algorithm does not report predictions per spot (rather a single ratio of tumour to stroma), accuracy per spot cannot be compared.

The highest baseline algorithm accuracy is 0.79, but this is when compared to a single human scorer, and therefore could be considered to be a suboptimal comparative measure. The highest accuracy reported uses the RF classifier and an image patch size of 64x64 pixels. The image accuracy reported (per patch size) suggests that the feature vectors become subject to a trade-off between containing insufficient amounts of visual data (patches are too small), and containing multiple tissue classes (patches are too big). Establishing how limiting fields of view affects pathologist scoring may provide insights into how the algorithm can be improved.

3.5 Discussion

Systematic random sampling is a valuable tool for quantitation of prognostic markers in CRC. The use of the RandomSpot system for applying SRS to virtual slides has facilitated the research into identification of the tumour stroma ratio as a prognostic marker. As a result, the pathologist-labelled coordinates have provided a database of valuable ground truth data, which have been used in this chapter to develop a ML algorithm in order to automate the task of spot counting.

The RandomSpot system has been used routinely at Leeds since its initial development in 2008, and has been used to identify prognostic markers in multiple types of tissue. The simplicity of the spot generation interface allows users to generate a simple sampling method for quantitation of tissue types within their chosen regions. This leads to higher reproducibility of results, and has facilitated numerous research projects in their analyses. The system itself generates XML annotation data which is then imported into Aperio ImageScope slide viewing software, which allows users to view image spots with unrestricted fields of view.

The raw scoring data for these projects has been collected in the RandomSpotDB, which currently contains 2.4 million spot locations with manual scores. This data is spread across multiple tissue types and stains, and has the potential for training image analysis algorithms to automate the task for different diseases. The spot classifications have not been counter scored or checked before entering into the database, and so using these data is subject to typical levels of human error and biases, that are encountered in day to day pathological analysis. This provides an interesting dilemma for using the data in both the training and evaluation of any supervised learning algorithms.

The QUASAR clinical trial dataset has been visually inspected by a pathologist using the RandomSpot system, using multiple sampling techniques. Ideally, the full dataset would have been analysed using the method placing 300 spots within the whole tumour ROI in order to provide a higher frequency of samples. Using higher frequency data should lead to a higher precision result with reduced error rates. However, due to time constraints this was not feasible in the original (pathological) study, and therefore the full dataset has only been analysed with sampling techniques using 50 spots. The method of placing a 3x3mm box over the highest

perceived area of tumour cell density was used for training the algorithm. This was because the statistics from this technique were more strongly correlated with survival in the original study, and therefore of higher clinical value. However, once the algorithm has been proved to work satisfactorily on this dataset, it would be useful to apply training and testing to all the available data to assess how that affects performance.

After examining the dataset, it became apparent that it the images are very varied in terms of phenotypic information, as well as staining quality, and artefacts. Some issues could be mitigated using colour normalisation, however, staining on some slides was so weak that normalisation could not rectify it. Processing was affected by tissue that contained little to no structural information (i.e. glands were not detectable), such as poorly differentiated tissue or large areas of necrotic tissue. Mucinous adenocarcinomas also affected analysis in that they appear as large light areas of the image and are classified as tumour. Due to the low frequency of these types of cancers in the dataset, it is proposed that these should be removed from training and testing. The presence of suboptimal and abnormal slides in the dataset motivated the work in Chapter 6 to automate a quality control (QC) procedure for flagging such images.

Digital artefacts arose from colour normalisation, if images contained almost no colour information, as discussed in 2.5.2. This was typically due to large areas of background, lumen or mucin, or tissue that was very weakly stained / faded. As a result, simple checks were built into the program in order to identify images that were entirely black (a result of colour normalisation failing), and these images were processed without normalisation applied.

Initial comparisons of the two sets of pathologist data compared TSRs from the two sampling methods, 50 spots on whole tumour regions, and 50 spots on 3x3mm boxes placed on the areas with the highest perceived tumour cell density. The data from both sets was generated using unrestricted fields of view on Aperio ImageScope software, and the results indicate that, as expected, sampling a smaller area with higher perceived tumour cell density, yields slightly higher TSRs.

It was initially hypothesised that TSRs could be simply calculated using thresholding in order to separate the lighter stroma from the darker tumour. The QUASAR dataset was used to develop an algorithm that used the nuclear staining channel from colour deconvolution to identify dense areas of nuclei as tumour (excluding lymphocytes). The purpose of developing such a simple algorithm was to identify where the methodology worked well and where it did not - and in the cases that it did not, ascertain the reasons why. Aside from the image variation and technical issues mentioned previously, areas of immune response (lymphocytes) broke the assumption of darker areas represent tumour cells. A simple attempt at mitigation used edge strength to

identify areas where probability of lymphocytes was higher in order to subtract them from the tumour probability map. In comparison to the ground truth, the algorithm is likely to underestimate the presence of tumour, which may be due to miscounting areas of tumour that appear lighter such as mucin, lumen and necrosis.

In order to fully utilise the RandomSpotDB data, a machine vision algorithm using artificial intelligence for supervised learning was implemented on the QUASAR dataset. This algorithm used basic features in order to train the classifier and was tested on multiple image sizes to find the most appropriate for maximising agreement with the pathologist data. The 64x64 pixel patch size was identified as the most accurate, and a variety of classifiers were used to assess which was most accurate. The RF algorithm provided the highest AUC and so was used for the purposes of assessing overall performance.

It was concluded that the 64x64 pixel patch size yielded the highest accuracy due to optimising the following two conditions:

1. Being large enough to contain useful visual data
2. Being small enough to only contain one tissue class within the image

Issues with larger patches are illustrated in Figure 60, where the classification of one patch taken at multiple resolutions can be affected by the surrounding visual information.

The three different image sizes contain varying percentages of the ground truth classification (stroma). From smallest to largest, the percentage of stroma in each image is 82%, 25% and 20%. Therefore, in the larger two images, the feature vectors generated are more likely to represent tumour in this example.

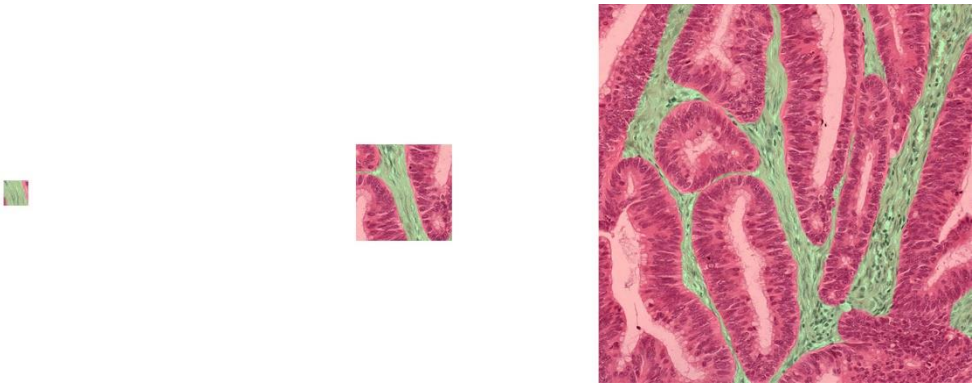


Figure 60 – Surrounding contextual information influencing feature generation in Algorithm A

Left: 64x64 pixel image patch, containing 18% tumour (red) and 82% stroma (green).
Centre: 256x256 pixel image patch, containing 75% tumour and 25% stroma.
Right: 1024x1024 pixel image patch, containing 80% tumour and 20% stroma.
Note in all cases, the classification is applied to the centre of the patch, which is stroma.
The figure shows that by increasing the amount of visual information in the image (increasing the size of the image patch), there is more scope for including more than one class (tissue type). The feature vectors used for ML mostly contain use average values to represent the appearance of the image, which in turn means that the visual characteristics derived from it are more likely to represent the class that is more prominent within the images. Therefore, by having larger image patches, without accounting for multiple classes within them, accuracy will decrease.

The results from Algorithm A suggest that larger images will negatively affect algorithm performance, but the ground truth data was scored by a pathologist on an unrestricted field of view (they were able to zoom in and out of the whole slide image). Utilising the unrestricted fields of view would allow them the use of surrounding contextual information to make a decision on the classification. Therefore, contextual information is important for human scoring, but this leaves an unanswered question of how much contextual information is required for a pathologist to confidently score a single spot. The answer to this question will facilitate the development of an algorithm that more closely mirrors manual scoring.

Chapter 4 - Optimisation of manual analysis conditions

4.1 Introduction

4.1.1 Chapter overview

The work in this chapter focuses on exploring optimal conditions for manual analysis of digitally scanned CRC tissue, using the conclusions drawn from work in the previous chapter. The primary objective of the work is the development of a user-centric data collection system for rapid acquisition of manual scoring data. The system (called Prospector) is initially used to create a simple experiment, which assesses the levels of agreement between the ground truth (pathologist) scores presented in Chapter 3, and pathologist scores when using a restricted field of view. The purpose of the experiment is to establish the amount of surrounding contextual information required for pathologists to score an image, which can then be used to optimise the image analysis solution. Also, a preliminary experiment is carried out with the objective of understanding the relationship between pathologist agreement and tissue staining levels, as a precursor to introducing a basic quality control factor into subsequent image analysis. The chapter is divided into five sections:

1) The introduction, which reiterates the discussion of the previous chapter, and details the questions raised from it.
2) An explanation of the web-based experiment platform, Prospector, which was specifically designed to capture expert ground truth data for assessing agreement and qualitative assessment under varying conditions.
3) An experiment using the Prospector system, created to identify levels of pathologist agreement when constrained by image size, and to determine appropriate amounts of surrounding context to include in the automated solution.

4) A second experiment using the Prospector system, created to identify levels of pathologist agreement when compared to tissue staining levels, in order to assess whether features related to tissue staining can be used as an appropriate quality control factor for accepting/rejecting images for image analysis.
5) Discussion of the work presented in this chapter and conclusions. The discussion focuses on the performance of the Prospector system, and potential uses of the system. The experiments are discussed, and the how the conclusions of each will affect the development of the algorithm in the next chapter.

4.1.2 Insights from Chapter 3

The results presented in Chapter 3 identified that the automated solution had the highest levels of agreement when analysing 64x64 pixel image patches. The main conclusion drawn from the data was that 64x64 pixel sized patches contained the optimal amount of visual information, which a) minimised distorting the feature vectors with image information from more than one tissue class, and b) avoided containing too little visual information to be of practical use. The original dataset was scored by a pathologist on an unrestricted field of view (a standard computer screen, with the ability to pan freely and to zoom in to native resolution), and so allowed surrounding contextual information to be factored into the scoring process. It was hypothesised that mimicking manual scoring more closely in terms of contextual analysis would improve the algorithm accuracy. However, the available data did not indicate how much surrounding contextual information is required in order to manually score an image. Establishing the amount of visual information required for a pathologist to score an image will inform how much visual information is appropriate for the automated solution.

4.1.3 Manual scoring on digital slides

The main application used for scoring virtual slides at the University of Leeds is Leica-Aperio's desktop application ImageScope, due to vendor-specific digital slide formats and viewing software. Scoring is typically achieved using hand-drawn annotations (stored in the specific Leica-Aperio XML format), using a mouse and keyboard. The pathologist's field of view is limited to the size of the workstation monitor, but the software allows pan and zoom actions in order to navigate the entire slide. Monitors are rarely colour-calibrated, and users have the option to enable or disable the in-built colour calibration functions within ImageScope. As

presented in Chapter 3, the ground truth data for the algorithm is simple x-y coordinates stored with a numerical key, denoting the tissue type at that coordinate.

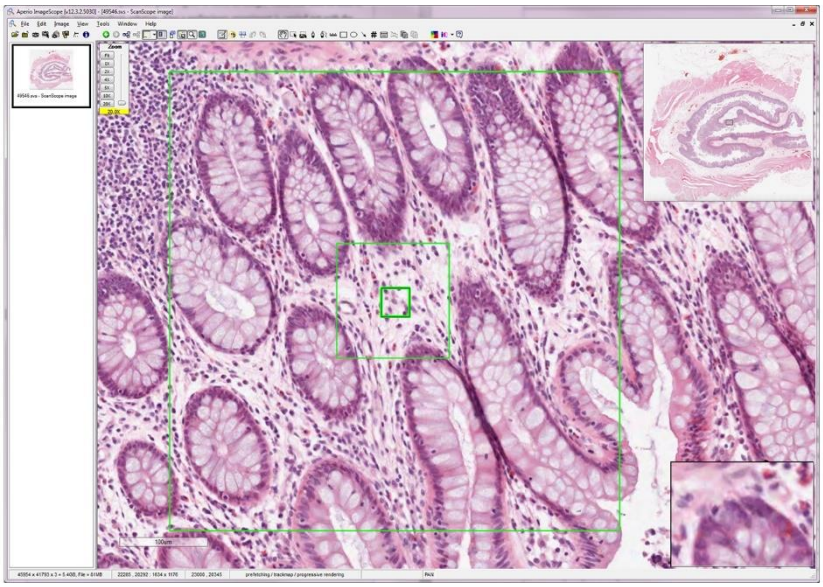


Figure 61 - The Leica-Aperio ImageScope software interface

The image view has been annotated with three different sized box annotations – 64, 256 and 1024 pixels in height and width. These represent some of the image patch sizes that were used in the evaluation of algorithm performance in Chapter 3.

In order to compare the ground truth data to pathologist scoring on restricted fields of view, it was concluded that the ImageScope software was not appropriate, due to the lack of functionality for masking areas of the screen that were not annotated. This resulted in a need for a system which is designed for the rapid acquisition of ground truth data, with the ability to present images with restricted fields of view.

4.2 The Prospector system

4.2.1 Aim

To develop a system that facilitates rapid acquisition of manual scoring data in order to create experiments for evaluating pathologist agreement when manipulating various conditions.

4.2.2 Introduction

Image analysis algorithms have the potential to either fully or partially automate visual inspection tasks to assist pathologists with their workload. However, due to the amount of variation in appearance between both tissue and disease types, complex algorithms need to be developed specifically for one purpose, rather than general image analysis pathologist tasks [64].

In order to be properly validated, image analysis algorithms must be trained on an extensive and varied set of pre-classified images [156,167,187,222-224]. For the trained algorithms to be trusted (let alone useful), classification of these images must be done by clinically trained pathologists with working experience of the tissue and disease being analysed [132,225,226]. However, labelling large quantities of data for training and testing is time consuming for pathologists and therefore expensive to generate. This often results in pathologists providing insufficient amounts of data for training algorithms and validation experiments [262]. Insufficient training and validation of pathological image analysis algorithms leads to overfitting to the ground truth, meaning that algorithms fail when exposed to the wide variety of real world image data.

Obtaining reliable ground truth data for experiments is time consuming, and requires effort from pathologists in addition to their daily workload [263]. Therefore, the process with which the pathologist generates this data should be as simple and effective as possible. Maximising the amount of ground truth data obtained, compared to the effort spent by the pathologist generating the data, will provide computer vision researchers with larger expert-classified data sets for training, testing and validation of their algorithms. Prospector is a lightweight web-based system specifically designed to capture pathologist scores and opinions rapidly.

4.2.3 Methods

Prospector is a web-based interactive pathology scoring system, using an HTML5 and jQuery powered interface, PHP server-side script, with a MySQL database back end. As a result, the system is both platform and browser independent. Figure 62 illustrates a simplified view of the network architecture used by participants of a given experiment set up on the system.

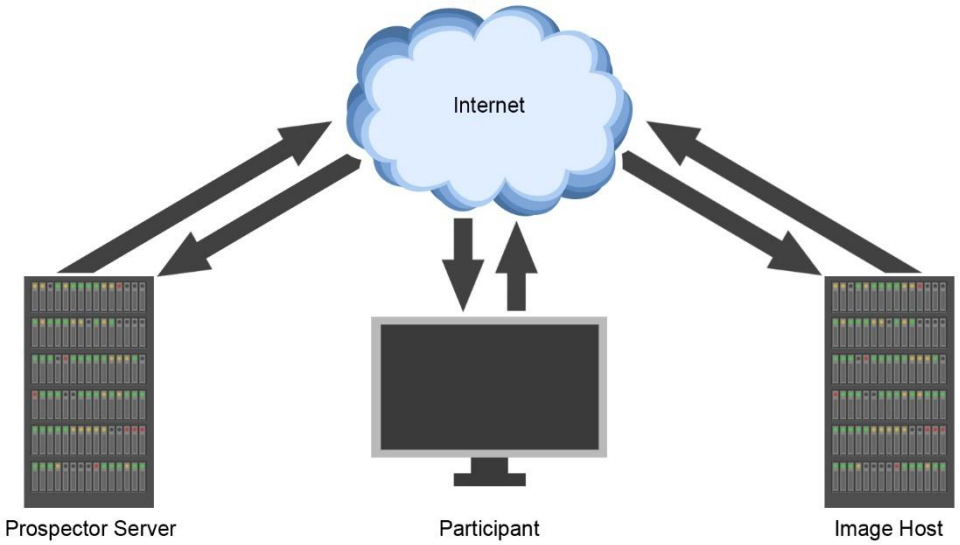


Figure 62 - Network architecture diagram of the Prospector system

Participant requests to load an experiment over http. Requests are processed by the Prospector server, and images are retrieved from the user-defined image host for any given experiment (anywhere on the web).

Prospector is primarily modelled on two simple use cases:

1) Administering an experiment
2) Participating in an experiment

These use cases are explained in detail in this section.

4.2.3.1 Use-case 1: Administering an experiment

Part one allows administrator users to create an experiment, whereby administrators are required to provide a comma separated value (CSV) list of URLs pointing to static images. These images should be hosted on a fast, reliable server, and should not exceed the expected size of the image viewing area (relative to participant monitor size), in order to maximise the efficiency of the experiment. Images can be any web browser enabled format (jpeg, png etc) and subject matter

can include, but is not limited to, micrographs, extracts from virtual slides, or macroscopic images. Images can also be generated from within predefined regions of interest on a virtual slide, creating randomly sampled, equidistant and systematically placed images, using the RandomSpot system. With the CSV image list created as a preparatory measure, administrators are required to step through a simple setup wizard, which consists of four screens.

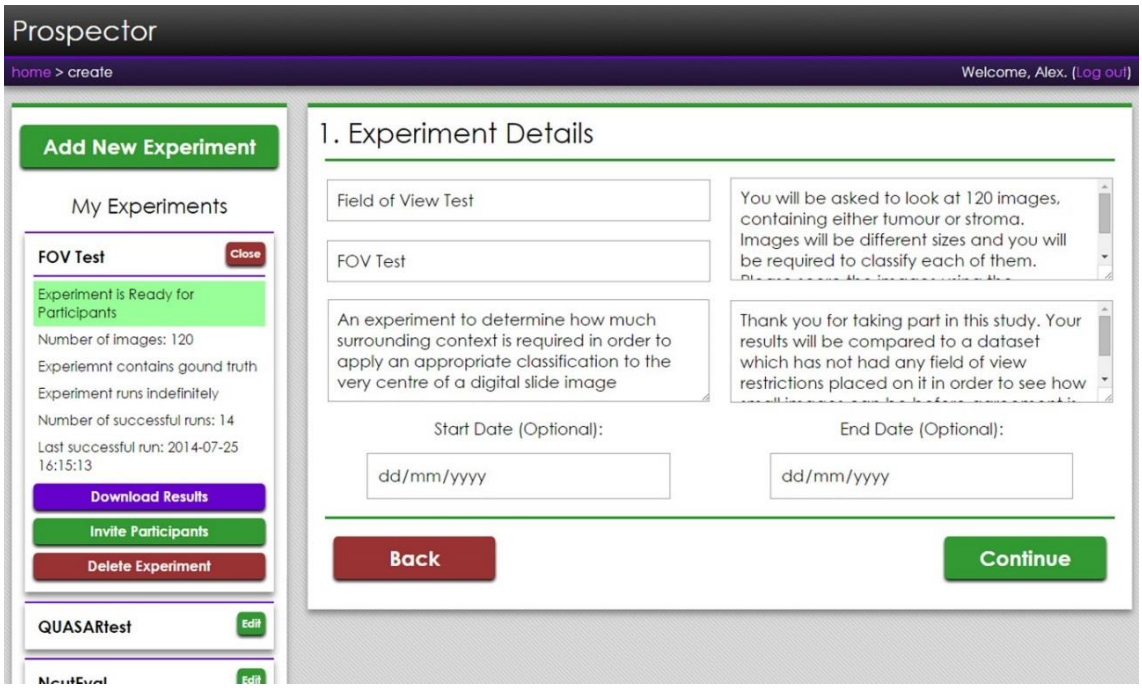


Figure 63 - The Prospector experiment setup screen

The left-hand pane shows a collapsible list of all the experiments that the administrator has made. From the expanded view, administrators can edit the experiment, download existing results, invite participants using a generic email template or delete the experiment. The right hand pane shows step one of the experiment setup wizard, asking for the experiment title, shortname, description, brief, debrief and optional start and end dates.

Initially the administrator is presented with a list of available experiments to view and edit, or to create a new experiment. This functionality is contained within the left-hand pane of the screen, shown in Figure 63, and existing experiments may be selected and edited from here. Built into the system is an automatic mailto email link that, which automatically opens a template email in the user's default email client containing a message that invites recipients to participate in the experiment, and provides them with a hyperlink to follow. The user simply needs to add recipient email addresses. Once participants have completed the experiment, their data can be downloaded (by the administrator) as a zip archive of CSV files, where one CSV file contains all the responses from one participant.

The first screen of the setup wizard asks the administrator to provide a name and description of the experiment, a brief and debrief, and optional start and end dates which can limit the period

of the experiment's availability to participants. If left blank, the experiment will run indefinitely. The brief and debrief are used to present to the participants before and after the experiment has taken place.

The second screen prompts users to upload their list of image hyperlinks with optional ground truth data. Ground truth should be a short text classification applied to each image in the list as the second column, denoting the correct or ideal classification that should be given by the participant. Providing existing ground truth data changes the experiment type from 'collection' to 'comparison', and is useful for studies compare levels of agreement, or the effects of controlled manipulation of pathologist scoring conditions (see the subsequent sections for examples of the 'comparison' experiment). These classifications will be used to compare to participant scores, so it is important that the text classifications provided can be matched to the available scoring categories (specified in screen three). Collection experiments are simply for obtaining ground truth from pathologists. Images can be presented to the user sequentially or in a random order, and can be rotated and translated randomly in order to prevent repetition biases.

The third screen is for setting the available scores with which participants may respond. Each possible score requires a name, a description and a shortcut key. Names of scores should be directly comparable to ground truth data, if provided. It should be noted that the system has specifically been designed to give single keystroke responses in order to maximise throughput of data.

The fourth and final screen of the wizard concerns privacy. By default, the experiment is open to anyone in the world, and only requires a name and valid email address to participate. Administrators may however provide a csv white-list of email addresses that are allowed to participate in their experiments. For ease of use, using an asterisk and email domain will whitelist all addresses from a particular organisation (e.g. *@leeds.ac.uk). The type of anonymity given to the participant can also be set with one of three options: 'forced anonymity', where all participation is anonymous; 'optional anonymity', where participants may choose to participate anonymously; 'no anonymity', where participant identity is required to be linked to their results. The default setting is 'optional anonymity'.

4.2.3.2 Use-case 2: Participating in an experiment

Part two allows users to participate in an existing experiment that has been created previously. The participant is asked to log in, providing their name and an email address, and then is presented with advice on how to setup their environment before continuing the experiment. This relates to room conditions, browser settings, screen size, brightness and contrast (using a calibration scale). Once the participant has optimised their conditions, they are presented with

the experiment brief (set by the administrator), and instructions on how to use the system before proceeding to the scoring screen (see Figure 64).

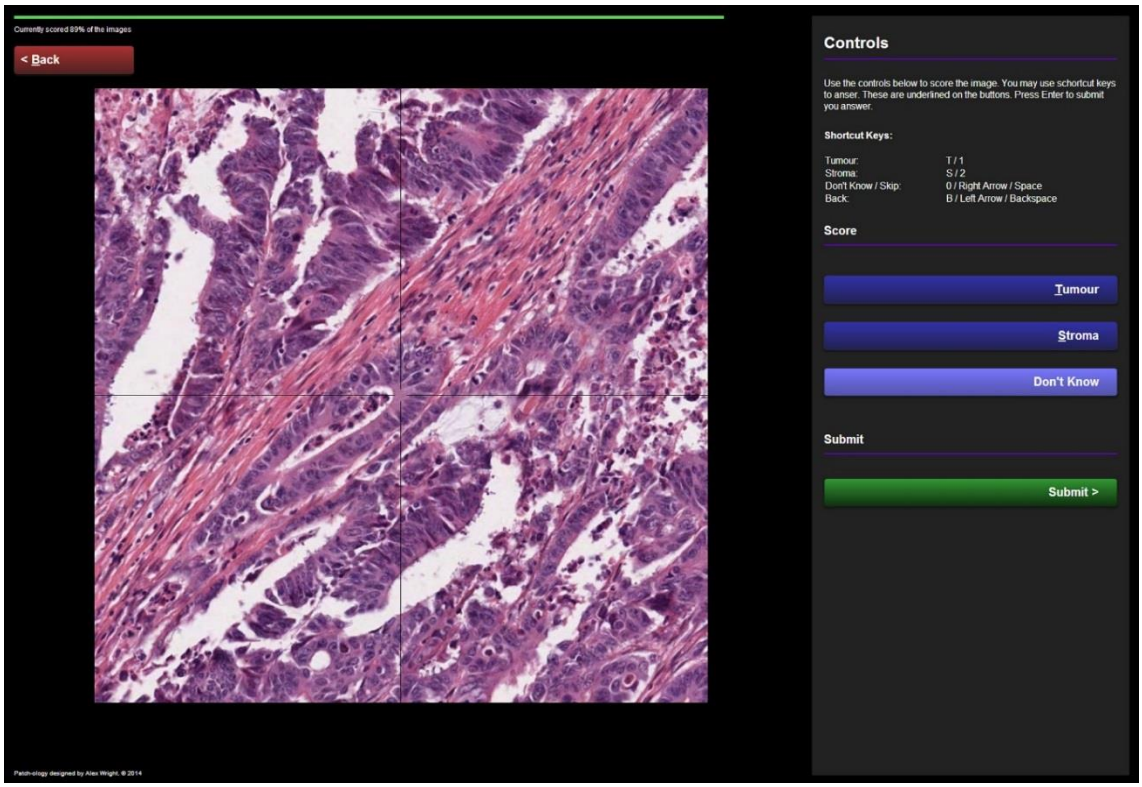


Figure 64 - The Prospector participant scoring screen

The figure shows an example experiment for scoring colorectal cancer tissue in order to identify the ratio of tumour to stroma. The screen consists of a large viewing area for the images, a control panel containing available scores and associated keyboard shortcut keys, a slim progress bar at the top, and a 'back' navigation button to correct scores made in error. Also, note that as an optional feature, crosshairs have been placed over the image to help participants identify the centre of the image, where the classification is to be made.

Figure 64 illustrates the scoring screen for the participant. The main image is a non-navigable, static snapshot, which has been embedded in the page using the http links in the image CSV file previously uploaded by the administrator. This is primarily in order to reduce time spent navigating and loading the image, but also allows the system to be able to apply random rotation and translation for prevention of repetition biases. The image in Figure 64 also has automatically placed guides over it, because in this example, the participant is being asked to identify the tissue type at the very centre of the image. The control panel on the right-hand side has been optimised for tablet users, with large, simple buttons. Desktop users are encouraged to use the keyboard shortcuts described at the top of the panel. These methods of scoring have been used to reduce as many clicks, taps or keystrokes as possible to increase the speed of data acquisition.

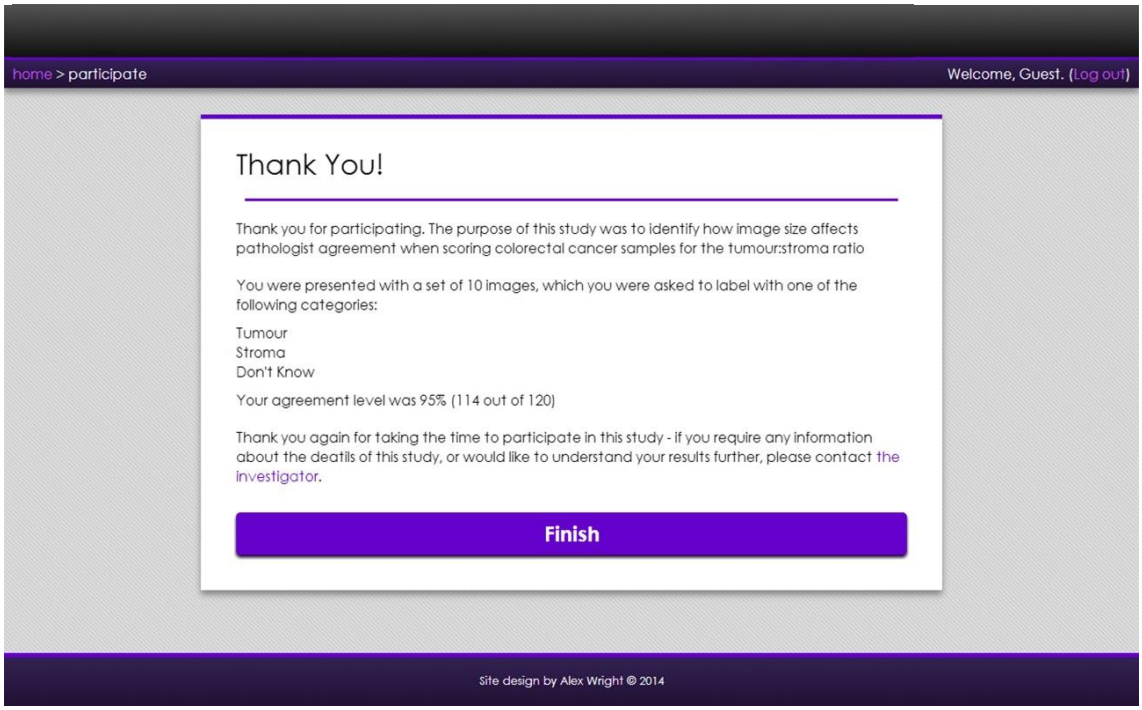


Figure 65- The Prospector experiment debrief screen

The debrief screen provides feedback on the participant's performance and informs them of the purpose of the study

Once all the images have been scored, the data is saved and instantly available for the administrator to download. If the experiment is comparing scores to existing ground truth, then the data is matched and added to the dataset. The participant is presented with a debrief screen (Figure 65) and if applicable, the level agreement with the original ground truth data.

Note that for end user simplicity, level of agreement is calculated as a percentage of the numbers of participant responses that matched the ground truth, as opposed to calculating more complex statistics such as kappa values. More complex statistical analysis can be performed by the administrator when downloading the raw results data.

4.2.3.3 Comparison of speed between Prospector and ImageScope

Prospector is designed to be a high throughput, rapid data acquaistion tool, and as such a simple pilot study was created in order to ascertain whether it outperformed data capture using existing methods. As a basic test to ensure the efficacy of the system, an input capture experiment comparing the scoring of data points generated by RandomSpot (section 3.2) was set up.

The experiment consisted of six cases from the QUASAR dataset, and already had XML annotations and RandomSpot sampling applied. The six cases were sampled using a target number of 50 spots per case, and a tolerance of 15%, creating a dataset of 313 spots in total.

Each spot was visually inspected by participants using the existing methodology, which involved loading the RandomSpot-generated XML annotations into Leica-Aperio ImageScope, and scoring them using the annotations window (Figure 66).

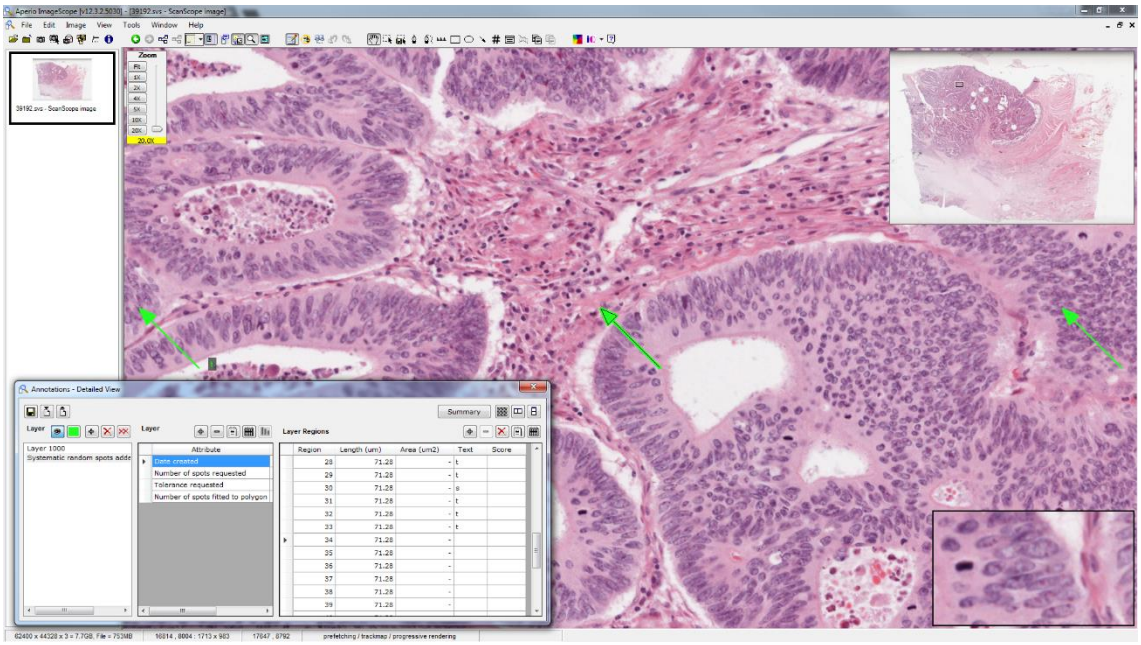


Figure 66 - ImageScope scoring interface using the annotations window

Each sampling point (spot) is analysed using the annotations window (bottom left). On typing in one of four responses (t = tumour, s = stroma, o = other, blank = unknown), participants can then press enter on the keyboard to jump to the next spot.

Participants were required to load each of the six digital slides using a pre-generated XML file containing a hyperlink to a slide (Leica-Aperio SIS file). After loading an image, participants were then required to load the corresponding pre-generated XML file, which contained the annotation and sampling points. These steps were taken prior to scoring. Users scored each spot individually by recording a letter T, S, O, or Space in the "Text" Field of the Annotations window. Once each slide was scored, Microsoft excel files were saved with the results (one per slide). The time taken to complete the whole task (of scoring the six slides and saving the data) was recorded.

The task using Prospector was set up as a single experiment for data capture, containing all 313 images from the six slides. The sample points were extracted as images 1024x1024 pixels in size prior to the experiment, using the co-ordinates from the RandomSpot-generated XML files. Four answers were available to choose from – "Tumour", "Stroma", "Other" and "Don't Know". Hotkeys were set to T, S, O and space respectively. Users were required to score the images, which were presented in a randomised order, and a CSV of results was automatically saved at the end. Time taken to score all 313 images was recorded.

One technician participated in the study, trained in spot counting techniques, and routinely used the methodology (RandomSpot and ImageScope) in day to day use.

4.2.4 Results

Prospector is a customisable web-based system for generating simple experiments which rapidly acquire data. Two such experiments are detailed in the subsequent sections (4.3 and 4.4), which in total capture 960 opinions on 240 images across two case studies, with a mean scoring time of 1.83 seconds per image. Also, the speed comparison study showed that the Prospector system was faster, taking 499 seconds to complete the task, as opposed to ImageScope taking 644 seconds (Table 9).

Method	**Total Time (s)**	**Average Time Per Spot (s)**
ImageScope	643.87	2.06
Prospector	499.23	1.59

Table 9 – Comparison of time taken to complete the experiment using both platforms

Performing the task on ImageScope took 644 seconds compared to Prospector which took 499 seconds.

The time taken using Prospector was 23% lower than the ImageScope method.

4.2.5 Conclusions

Prospector is a simple and powerful tool for assisting computer vision and clinical researchers in obtaining pathologist-classified image data. The pilot study showed the increase of efficiency of using Prospector over the current methodology for scoring using the RandomSpot-generated sampling points, by decreasing the time taken by 23%. However, the system was not designed to replace routine scoring, and the study does not consider time taken to set up the experiment in Prospector, and generating the images to score. Prospector is not intended as a replacement workflow for standard pathological tasks such as SRS.

Prospector is a suitable platform for obtaining ground truth data for identifying levels of agreement between pathologists when scoring under constrained fields of view.

4.3 Image size experiment

4.3.1 Aim

To identify an optimal minimum image size for manual analysis, by restricting fields of view to varying degrees, and assessing pathologist agreement compared to the unrestricted field of view.

4.3.2 Introduction

The results from the baseline algorithm study presented in Chapter 3 show that the algorithm accuracy peaks when run using the RF algorithm, trained on features from image patches 64x64 pixels in size. The ground truth data was obtained using whole slide images (WSIs), with unrestricted fields of view, thus allowing pathologists to gain contextual information from the surrounding tissue. It is hypothesised that in order to improve algorithm accuracy, this surrounding contextual information should be taken into account, but the amount of context required is unknown. By using the RandomSpot co-ordinates from the QUASAR trial, the Prospector system can be used to assess to what extent limiting the fields of view for pathologists has on their agreement with the original ground truth.

4.3.3 Methods

The Prospector system was configured to create an interactive web-based experiment, with the aim of finding the minimum patch size that pathologists can accurately and consistently score at. Participants were chosen based on availability, and a mixture of pathologists and histology research technicians (n_p = 6) took part. Technicians were trained in using the RandomSpot system, and had already been using the system routinely for at least three months. Participants were presented with 120 images of pre-diagnosed haematoxylin and eosin (H&E) stained CRC, extracted from the spot counting data at three different sizes: 64x64, 256x256 and 1024x1024 pixels. Each image patch was randomly rotated or flipped to avoid biases from participants recognising repeated images of different sizes. Scores were collected based on whether the participant classified the centre of the images as tumour or stroma, or didn't know. These

classifications were subsequently compared with the ground truth spot counting data, where the original observer was not restricted to any specific level of context for analysis.

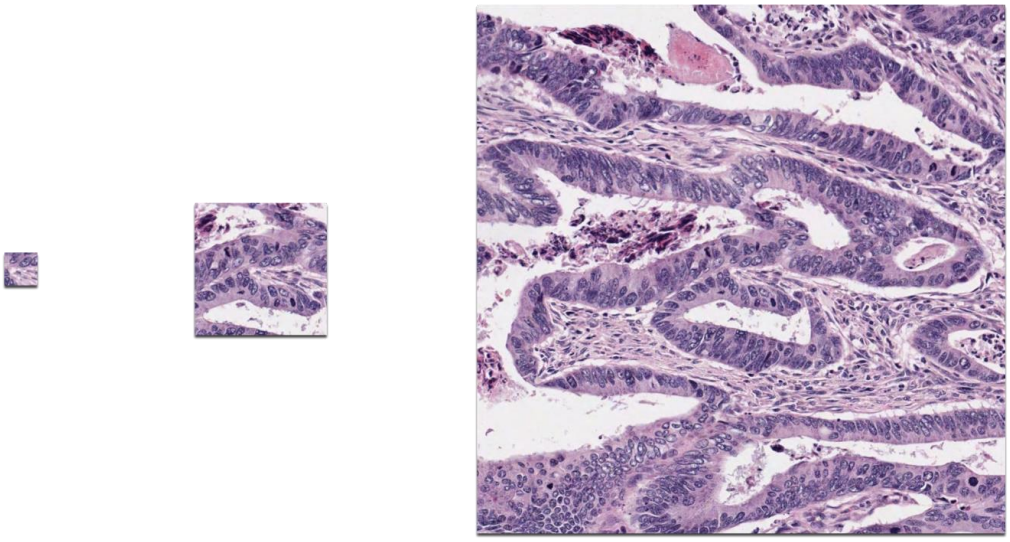


Figure 67 - Comparison of contextual information between patch sizes

The three images are proportionally displayed, showing 64x64 (left), 256x256 (centre) and 1024x1024 pixel (right) patches, illustrating that larger patches contain more tissue of varying types

All participants viewed the images on a 30-inch Dell 3008WPF monitor, running in an HTML5 compliant web browser (Google Chrome version 26.0.1410.64 m) in full screen mode at 2560 x 1600 resolution. This monitor was specifically chosen for the high dynamic contrast ratio (3000:1). After being briefed about the experiment, participants were sequentially presented with the images, and were given one of three options to score the image, “Tumour”, “Stroma” or “Don’t Know”. Each participant was presented with the same images in the same order. Upon completion, the participant was then presented with a debrief screen and a graph of how their scores correlated with the ground truth.

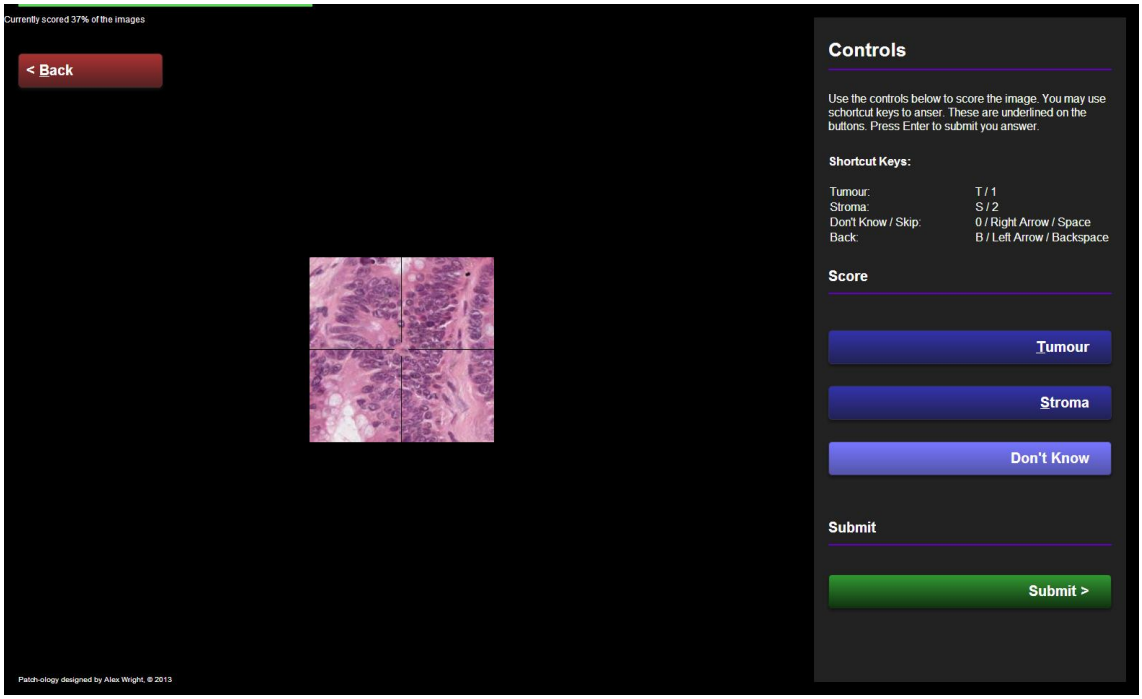


Figure 68 – The Prospector interface displaying an image at 256x256 pixels.

The screenshot of the shows the image patch display area, with progress bar (left) control panel with scoring buttons and shortcut keys (right)

The aim of the experiment was to identify how much contextual information is required in order for a pathologist to classify a given point on a piece of tissue. Forty images of colorectal cancer (CRC) tissue were used in the experiment, which had already been classified by a pathologist as part of a clinical trial study. The clinical trial used anonymised virtual slides scanned at 20x objective zoom (0.5 microns per pixel). The classifications were made by pathologists evaluating single point locations, within each virtual slide containing CRC tissue, in order to identify the type of tissue found at each of the locations. For the purposes of this study, classifications were simplified into one of two classes: tumour or stroma. The forty images for this experiment were randomly selected from the full set of 2,211 clinical CRC cases, containing over 100,000 pathologist-scored point locations. Each of the forty images were extracted at the virtual slide native resolution (0.5 microns per pixel), using three different sizes, in order to present to the participant's images with different amounts of visual contextual information surrounding the classified point. The image sizes were 64x64, 256x256 and 1024x1024 pixels. The intention of the study was to establish whether there were significant effects on the level of agreement between participants, when scoring images of different sizes.

	Image size		
	64x64	**256x256**	**1024x1024**
Tumour	20	20	20
Stroma	20	20	20

Table 10 - Number of patch size variations used for manual scoring

For the image patch size experiment, three sizes were used (64x64, 256x256 and 1024x1024 pixels), and for each size used, 20 images of tumour and 20 images of stroma were presented to the participant.

Six participants (3 trained pathologists and three technicians experienced in TSR scoring) were presented with the 120 images, and asked to classify each of them. As described previously, images were rotated and translated randomly to avoid repetition biases, and guides were placed over the images to explicitly illustrate the exact point that should be classified. Participant agreement was calculated by the system and presented in the experiment debrief. Results were collated by the Prospector system, and made available in CSV format for analysis.

4.3.4 Results

Participant responses were grouped by image size, and a percentage of agreement (with the ground truth) was generated per participant. This included the scores where a participant didn't know the answer. Figure 69 contains box plots showing the percentage agreement, for all participants, per patch size, compared to the baseline algorithm agreement using tenfold cross validation.

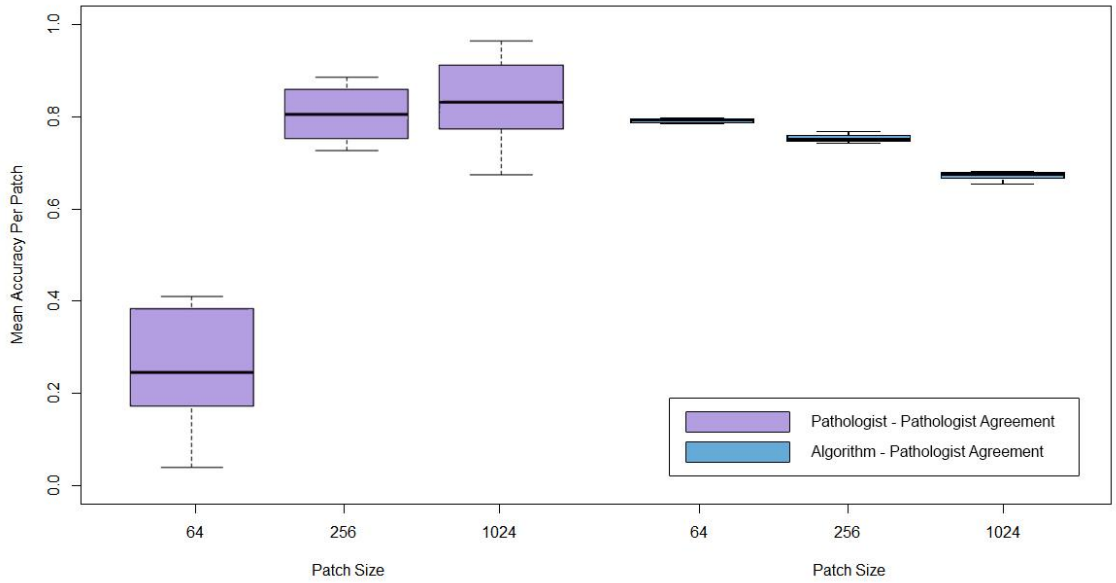


Figure 69 - Box plots of pathologist-pathologist agreement, and algorithm-pathologist agreement on three restricted fields of view

Left: Human agreement with ground truth, with means of 0.28, 0.80 and 0.85.
Right: Algorithm A agreement with ground truth, with means of 0.79, 0.75 and 0.67.
The boxplots show the levels of agreement between the pathologists and algorithm change when limited to three fields of view – 64x64, 256x256 and 1024x1024 pixels. The median line is represented in black, the coloured boxes indicate the inter quartile range, and the whiskers show the 95% confidence intervals of the distribution.
The plots show a positive relationship between patch size and accuracy for human scoring, and an inverse relationship between patch size and accuracy for algorithm scoring, indicating that the algorithm methodology does not adequately recreate the human scoring method.

For the manual scoring experiment results, a Friedman's ANOVA of the mean percentage of correct responses was conducted with image sizes (64x64, 256x256 and 1024x1024) as the repeated measures independent variable. Results revealed a significant effect of image size ($\chi 2(2) = 10.17$, $p < 0.01$). Pairwise comparisons revealed that participants performed significantly worse on the 64x64 pixel patches compared with patch sizes 256 ($Z = -1.25$, $p = 0.30$) and 1024 ($Z = -1.75$, $p < 0.01$). There were no significant differences in performance between patch sizes 256 and 1024 ($Z = -0.50$, $p = 0.39$).

Image Size			
64x64	**256x256**	**1024x1024**	***Total***
28 (13)	80 (6)	83 (10)	64 (10)

Table 11 - Mean accuracy and standard error for all participants of the image size experiment

The table displays statistics that are visualised in the boxplots presented in Figure 69, showing that participants perform significantly worse on 64x64 pixel images.

Further analysis of the results was conducted by distinguishing discordant answers from the answers where participants could not confidently score the images (the "Don't Know" category). Figure 70 highlights that the low levels of agreement exhibited by the 64x64 pixel category in Figure 69 are mostly due to the image rejection, rather than incongruous assessments.

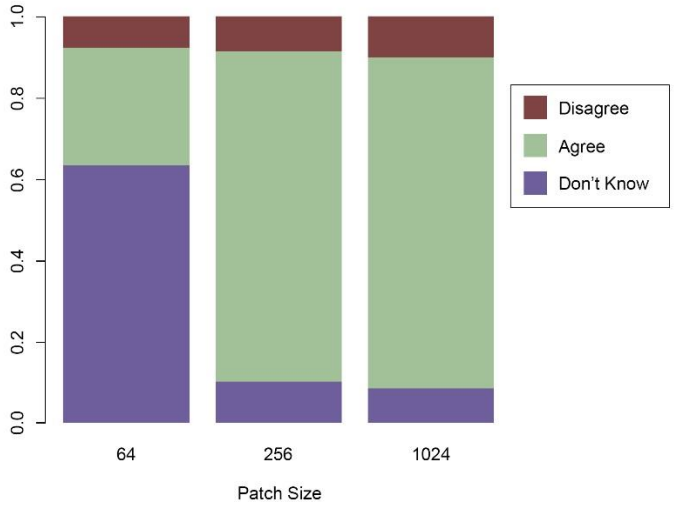


Figure 70 - Distribution of all participant responses per patch size

64x64 pixel patch size showed a 63.3% rejection rate by participants, with 256x256 and 1024x1024 pixel patches having 10% and 8.3% respectively. Incorrect responses stayed consistent across patch sizes at 7.6% (64), 8.6% (256) and 10% (1024).
The figure illustrates that the agreement levels presented in Figure 69 may misrepresent the pathologist scoring methods, in that a pathologist is more likely to reject an image at the smaller patch sizes, rather than disagree on a particular classification. The classification disagreement proportions stay consistent regardless of size.

For the 64x64 pixel images, 63.3% of the patches were rejected, where pathologists rejected these patches 79.16% of the time, compared with 47.5% for technicians. Of the patches that were not rejected, pathologists had higher levels of agreement for 256x256 square pixel patches, attaining 95.04% agreement, compared with 93.68% agreement on 1024x1024. Technicians

also showed this trend, scoring 85.8%and 84.4% respectively. Rejection rates of the two larger patches decreased, with 10% for 256x256 and 8.3% for 1024x1024.

4.3.5 Conclusions

The work in this section was conducted in order to establish a minimum size for image analysis algorithms to classify tissue, using appropriate levels of context (neighbouring tissue). The discussion in section 3.5 details the need for context, and Figure 60 illustrates how different levels of surrounding contextual information affect the classification results of algorithms that do not incorporate it (such as Algorithm A).

The experiment was successful in identifying that 64x64 pixel images were not appropriate for human inspection, whereas both 256x256 and 1024x1024 pixel images were. Since agreement levels were slightly higher on 256x256 pixel images (excluding image rejection rates), it is concluded that 256x256 pixels is an appropriate image size for scoring CRC tissue for TSR analysis.

4.4 Assessing agreement against tissue stain features

4.4.1 Aim

To establish the effect of tissue staining on pathologist agreement, using image features relating to staining levels as a basic quality control measure.

4.4.2 Introduction

The issues with tissue staining variation discussed in section 3.3.2.3 highlight that colour normalisation is an appropriate method for correcting for stain variation, within reason. As previously discussed, the images in the QUASAR dataset are from a longitudinal study where slides may not have been digitally scanned for a number of years after initially being stained. It is reasonable to exclude these images from the analysis, as they would typically be rejected by the pathologist for microscopic analysis. However, rejection of images should be treated as a last resort in order to minimise loss of data, and ultimately, the costs of re-sectioning and staining another slide from donor tissue blocks. In order to assess whether an image is suitable for analysis, the Prospector system was again used to present images to a trained pathologist, in order to assess the impact of staining levels on pathologist agreement. It was hypothesised that a lack of nuclear staining would impair the pathologist's ability to score the images, and agreement levels would be lower on images that had lower levels of staining.

4.4.3 Methods

4.4.3.1 Observer agreement

A similar methodology to the experiment in 4.3 was employed to identify how staining intensity, or lack thereof, affects a pathologist's ability to score images of tissue. A set of 240

spot co-ordinates (120 tumour and 120 stroma) was taken from the RandomSpot-sampled QUASAR dataset, and images were extracted at native resolution (scanned at 0.5 microns per pixel), at 256x256 pixels in size. This size was chosen based on the conclusions of the previous experiment (4.3.5).

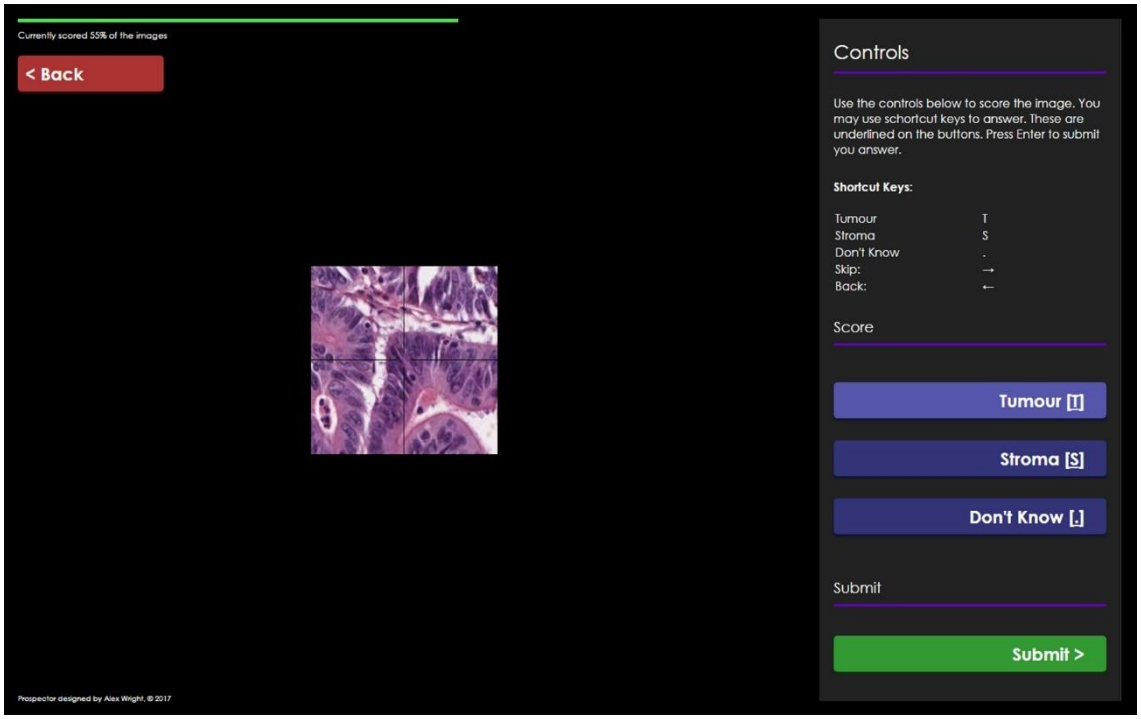


Figure 71 – The Prospector interface showing the image quality experiment

As with the previous experiment, participants were given the choice of three responses, tumour, stroma or don't know. The images were to be classified at the centre of the patch (where the original score was applied), and small black lines acted as guides to illustrate where the centre was.

Staining intensity values for each image were calculated prior to being presented to the pathologist. The values were generated from the means of three different channels of intensity; the HSV intensity channel, and the intensities of the deconvoluted staining channels for Haematoxylin and Eosin. Mean intensity was calculated for both the whole image and foreground pixels only. Foreground pixels were determined by statically thresholding HSV intensity less than 240 and HSV saturation greater than 0.04. Finally, for each image, adaptive threshold values were calculated for each of the three channels using Otsu's thresholding method (see 2.4.5.1).

The images were presented sequentially using the Prospector system. Since the images were not repeated, images were not randomly rotated or translated to prevent repetition biases. As with the previous study, scores were collected based on whether the participant classified the centre of the images as tumour or stroma, or didn't know. These classifications were again compared

with the ground truth spot counting data, where the original observer was not restricted to any specific level of context for analysis.

As with the previous study, participants viewed the images on a 30-inch Dell 3008WPF monitor, running in an HTML5 compliant web browser (Google Chrome version 26.0.1410.64 m) in full screen mode at 2560 x 1600 resolution. This monitor was specifically chosen for the high dynamic contrast ratio (3000:1). After being briefed about the experiment, participants were sequentially presented with the images, and were given one of three options to score the image, "Tumour", "Stroma" or "Don't Know". Each participant was presented with the same images in the same order. Upon completion, the participant was then presented with a debrief screen and a graph of how their scores correlated with the ground truth. This created three categories "agreement", "disagreement" and "rejection", where rejection represented the instances a participant could not use the visual information to make a confident classification of the tissue type ("Don't Know" response). One pathologist participated in the study, and responses were correlated against the previously generated staining intensity statistics.

4.4.3.2 Machine learning

A simple ML experiment was set up, using the precomputed image intensity values as the features, and the pathologist response as the ground truth. Using tenfold cross validation, a RF algorithm was trained to identify images that the pathologist would not answer. The experiment was performed twice, once for attempting to predict images with responses, and images that remain unanswered, and once for predicting whether the images are answered with agreement or disagreement, as well as whether they were unanswered.

4.4.4 Results

The average time taken to complete this experiment was approximately 454 seconds, equating to a mean scoring time of 1.89 seconds per image. Each of the scores given was compared to the ground truth data, and given one of three categories, answered with agreement, answered with disagreement and unanswered. The distribution of these responses was 206, 13 and 21 respectively, as illustrated in Figure 72.

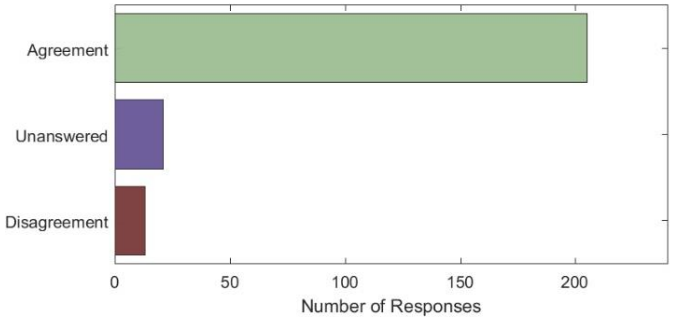


Figure 72 - Distribution of response types for the image quality experiment

Experiment used 240 images of 256x256 pixel size. Number of responses per category: agreement (206), disagreement (13) and unanswered (21). The distribution indicates that the dataset is too imbalanced to create an accurate model of the classes using a ML classifier.

4.4.4.1 Image rejection

The distribution of responses was analysed as three groups, and the mean feature values (types of image channel intensity) for each group was calculated. Images that were unanswered (participant response "Don't Know") were considered to be unsuitable for analysis (rejected). Figure 73 shows boxplots of the distribution of multiple types of intensities for the three responses.

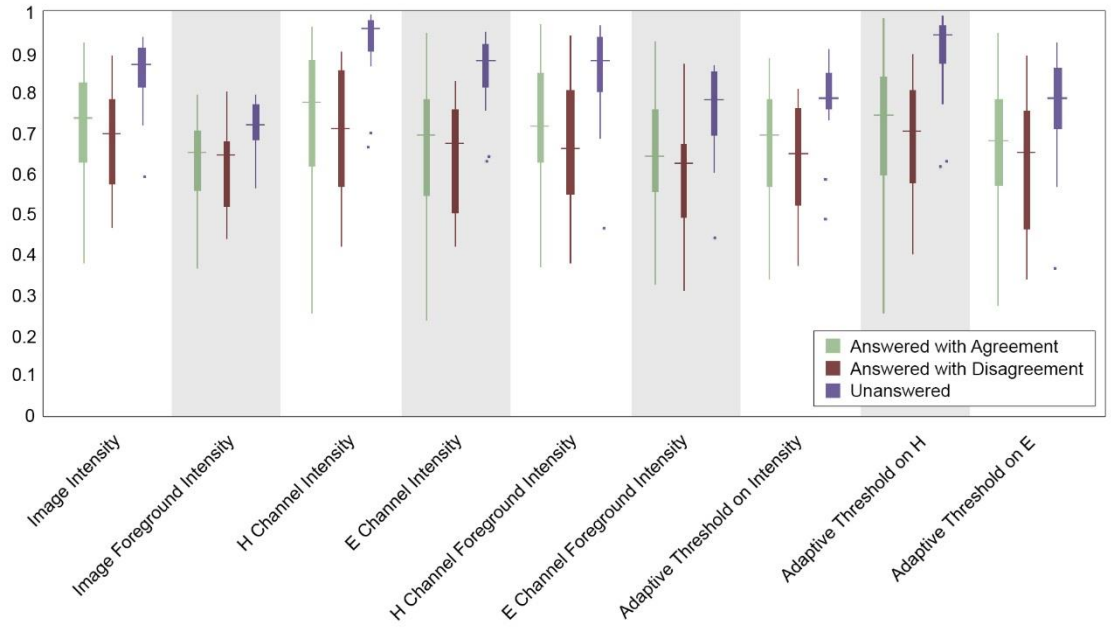


Figure 73 - Boxplots of mean feature values, grouped by response type

For each feature of every image scored by the pathologist, values were grouped by pathologist response type (answered with agreement to ground truth, answered with disagreement to ground truth, and unanswered), in order to ascertain whether these features were appropriate for predicting whether a pathologist would reject images, or whether they would affect agreement with other pathologists.

A one-way ANOVA of the feature values was conducted (per feature) with the participant response as the independent variable. For all features tested, results revealed a significant effect of image intensity on rejection rates ($p < 0.01$ for all features).

Pair-wise comparisons revealed that participants consistently rejected images with higher intensities, yielding significant differences between all intensity features for rejected and accepted images (Table 12). There were no significant differences between feature values for all images that were answered with both agreement and disagreement.

	P-value for Bonferroni-corrected pairwise comparisons between groups		
Image Feature	**Unsure vs Agreement**	**Unsure vs Disagreement**	**Agreement vs Disagreement**
Image Intensity	< 0.01	< 0.01	0.57
Image Foreground Intensity	< 0.01	< 0.01	0.78
H Channel Intensity	< 0.01	< 0.01	0.74
H Channel Foreground Intensity	< 0.01	< 0.01	0.86
E Channel Intensity	< 0.01	< 0.01	0.33
E Channel Foreground Intensity	< 0.01	< 0.01	0.38
Adaptive Threshold on Intensity	< 0.01	< 0.01	0.48
Adaptive Threshold on H	< 0.01	< 0.01	0.58
Adaptive Threshold on E	0.01	0.01	0.42

Table 12 - Bonferroni-corrected pairwise comparisons for post-hoc analysis of image features

For all image features, there were significant differences between images that had been rejected, and images that had been scored (either with agreement or disagreement). There were no significant differences between features of images that had been scored with agreement and disagreement.

4.4.4.2 Machine learning

The ROC curves in Figure 74 show the prediction accuracy of agreement (or disagreement and rejection) between pathologists, when analysing image features related to staining levels. Results show a higher AUC for predicting the binary response (answered and rejected = 0.79) versus further specifying the answered responses with pathologist agreement (agreement = 0.62, disagreement = 0.36 rejection = 0.77).

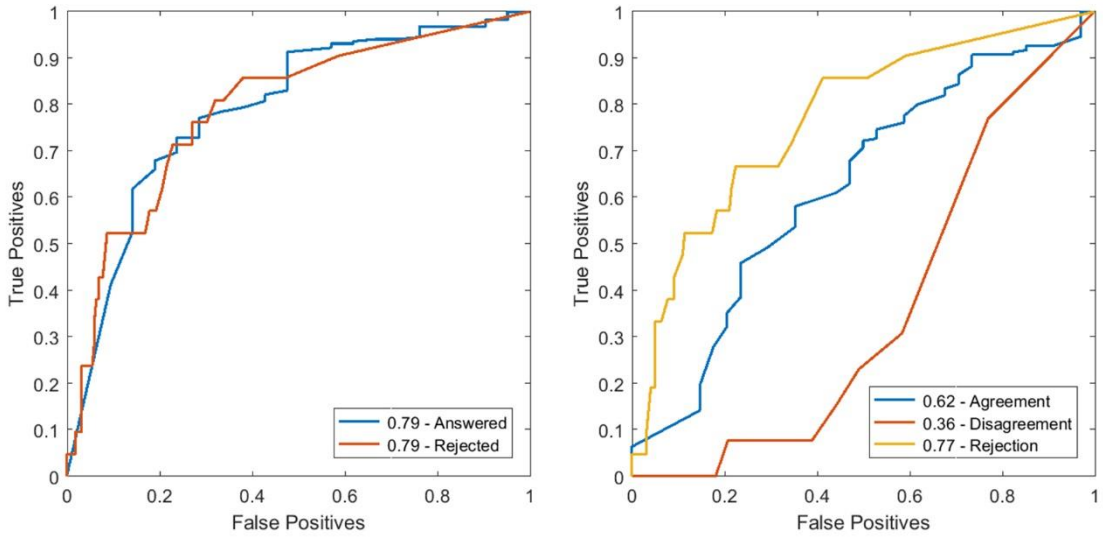


Figure 74 - ROC curves of trained RF algorithms for basic image QC

Two separate classifiers were trained independently to assess difference between agreement, disagreement and rejection. Both algorithms used the intensity values as features and the QC agreement as ground truth. One classifier was trained using rejected images and images that were scored, regardless of agreement (left) answered category includes answers that both agree and disagree with the ground truth (right) answers that agree with the ground truth are separated.

Using the algorithm for the binary predictions, the confusion matrix in Table 13 illustrates that the algorithm correctly identifies one image for rejection, and incorrectly accepts 20 images for analysis. The imbalance in the dataset groups means that the accuracy, sensitivity and specificity remain high, at 0.90, 0.98 and 0.95 respectively, whereas kappa agreement is 0.04.

		Algorithm	
		Accept	**Reject**
Pathologist	**Accept**	214	4
	Reject	20	1

Table 13 - Confusion matrix showing algorithm-pathologist agreement for QC

The table shows mean pathologist – algorithm agreement over 10 folds of cross validation for the binary QC dataset, where images that showed both agreement and disagreement counted as a QC pass. Accuracy = 0.90, sensitivity = 0.98, specificity = 0.95, kappa = 0.04 (poor agreement).

4.4.5 Conclusions

Image and stain intensity features are appropriate for identifying images that would be rejected for analysis by a pathologist. Analysis of the data showed no significant differences between

agreement and intensity, but showed a trend towards higher intensities being rejected by the pathologist (unsure category).

The prototype algorithm for automatically identifying images that should be rejected was created after the original experiment had been implemented, and as a result, the dataset was too imbalanced to create any firm conclusions. Results from the RF predictions indicate that training with more examples of rejected images should yield a more accurate classifier. This pilot experiment uses the data obtained from one pathologist and therefore the findings are simply meant to indicate trends in patterns of manual analysis, and for more concrete conclusions to be drawn, a larger study would be required.

It would be naïve to assume that tissue staining intensity is the only reason for pathologists to reject images for analysis, however, the results clearly indicate that obtaining appropriately stained images is fundamental to the visual inspection process. Therefore, assessing staining levels should be accounted for prior to analysis in any automated computer vision solutions.

4.5 Discussion

Obtaining acceptable quantities of ground truth from pathologists is a barrier to improving computer vision algorithms, and the generation of such data should be as efficient as possible. Out of this need, Prospector has been developed as a simple and powerful web-based system for rapid acquisition of ground truth data, whereby a set of pathologist-labelled images is generated. The system can also be used for comparing pathologist scores to existing ground truth data. This is beneficial for use in experiments either examining the efficacy of the characteristics of a given set of images, or manipulating conditions in order to understand their effects on pathologist agreement. The system has been designed to be as minimalistic as possible to expedite experiment setup time and minimise participation time for pathologists, providing experimenters with a platform for rapidly capturing pathologist scores. A simple pilot study showed a decrease in scoring time of 23% over current scoring methodology.

Prospector is used in two experiments presented in this chapter (4.3 and 4.4), which capture a total of 960 opinions on 240 images across the two studies, with a mean scoring time of 1.83 seconds per image. As such, Prospector demonstrates that it is an effective tool for experimenters wishing to analyse images using a simple, rapid interface. The two experiments illustrate that the types of analyses are not limited to gathering training data for computer vision algorithms or pathologists wishing to score their own clinical images. The "comparison" experiment methodology has the capacity to be incorporated into experiments where pathologist counter scoring is required for validation, or for validating image analysis algorithm results, by recruiting pathologists to score markup images. Its use could also be extended to studies gathering opinions on images from multiple pathologists.

Once familiarised with the shortcut keys, the mean time taken for a pathologist to score an image was less than 2 seconds for a simple three-class scoring system. The time taken to analyse images will be subject to the type of analysis, and the bandwidth of the client machines. As Prospector is a web-based system, it can also be used for worldwide collaborations, such as clinical trials or inter-observer studies. The images used need not be photomicrographs or virtual slides, as the system could be used for macroscopic images, clinical images, or snapshots of radiological images. Currently, Prospector only allows static images of virtual slides, as embedding navigable slides will slow down the user experience and is beyond the scope of the

project. Further extensions of the system might provide an opportunity for crowd-sourcing online image assessment (citizen science) experiments that do not require expert training to classify images.

The image size experiment concluded that 64x64 pixel images are not appropriate for manual scoring, and 256x256 pixel images have sufficient contextual information for basic visual inspection. This is the key finding in the work presented in this chapter.

The initial purpose of the experiment was to assess the minimum appropriate image size for maximising pathologist agreement on CRC tissue, however, it emerged that image rejection (being unable to score the images) was more salient than the level of pathologist agreement. The levels of pathologist-pathologist agreement (including rejection) were compared to levels of pathologist-algorithm agreement, which showed an inverse relationship. The work presented in Chapter 3 concluded that algorithm analysis required smaller image patches to ensure feature vectors contained information relating to one tissue class (see Figure 60 in section 3.5). However, the work presented in this chapter demonstrates that manual scoring requires more visual information in order to make appropriate classifications. It is concluded that human visual assessment requires the presence of structural information (context), including both tumour and stroma tissue classes, so that distinctions can be made between the two. Smaller images that contain single tissue classes are more likely to appear homogenous, and therefore do not contain the necessary visual information to make comparisons, and subsequently differentiate between types of tissue. Therefore, using a larger image patch size in an automated solution would require some form of contextual analysis to utilise this information and increase the representational acuity of the feature vectors. Image segmentation within larger patches may be necessary to minimise the probability of containing multiple tissue types within a given segment, whilst including the surrounding contextual information. Results from the image size experiment in 4.3 indicated that an appropriate amount of visual information is 256x256 pixels in size, which equates to a 128x128 micron field of view. This conclusion drives the contextual analysis research in Chapter 5.

Using the list of published algorithms (Table 3) in 2.5.4, image size was recorded for each study, so that comparisons could be made. Figure 75 shows two boxplot distributions for the image sizes used, separated by stain type. Note that some papers reported non-square image sizes, and so to maintain comparability, the square root of total area size is plotted on the graph.

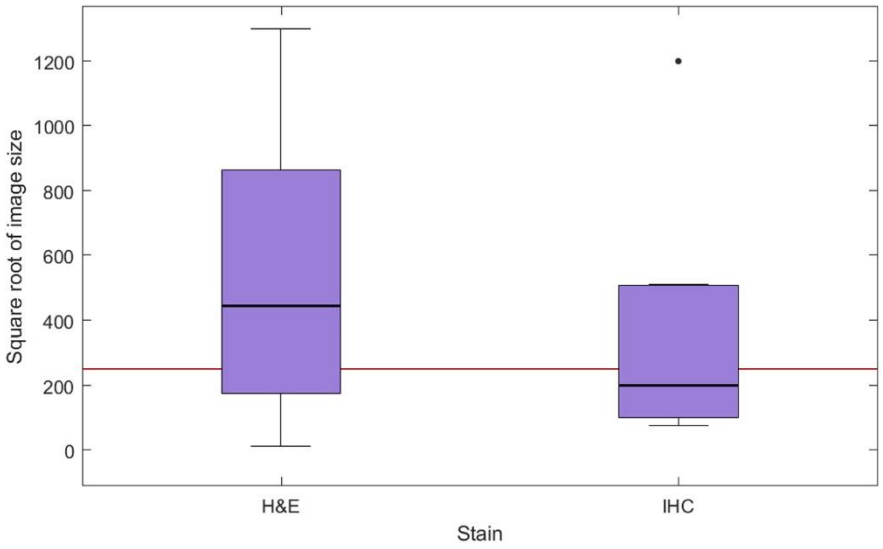


Figure 75 – Boxplots showing square root of image size in other studies, for H&E and IHC images

Left: Image size for 18 studies using H&E stained slides (mean = 527, median = 443, S.D. = 438)
Right: Image size for 8 studies using IHC stained slides (mean = 390, median = 200, S.D. = 401)
The horizontal line is placed at 256, which represents the optimised size found by the work in this chapter.
Image size is taken from the results presented in Table 3, and split by stain type. Note that some studies reported using non-square images for analysis and so the square root of total area size is taken to improve the comparability of results.

The boxplots show that studies analysing H&E stained tissue use larger image patch sizes compared to IHC (mean 527 compared to 390). The difference in distribution may be exaggerated due to the imbalanced numbers of studies for each stain type (H&E = 18, IHC = 8), but may also be due to the comparative complexity of the task - such that IHC images contain histochemical staining to highlight specific tissue components for visual inspection. This increased contrast facilitates both manual and automated analysis, which may result in less of a need for surrounding contextual information. It was also observed that studies using hand drawn annotations of whole epithelial glands (therefore requiring larger images) were performed on H&E images. Using the optimised patch size of 256x256 pixels is not appropriate for identifying large epithelial glands, but is more consistent with algorithms using a patch-based approach to segmentation and classification. It can be concluded from this that appropriateness of patch size is based on the appearance (size, contrast etc) of the objects being analysed.

Note that aside from the studies using freehand ROIs drawn by a pathologist (where the annotation dictates the image size), none of these papers presented rationale for the image size used. Reis et al [200] comment on image size optimisation as possible research area that could provide insight into the minimum amount of visual information required to distinguish between tissue types. The work in this chapter attempted to address this issue, and found that 256x256

was appropriate for manually distinguishing between tumour and stroma on H&E stained CRC images.

There are many issues in digital slide analysis pertaining to image quality, and as previously discussed, these issues can relate to many factors. However, one of the main issues is in the level of staining. The level of stain has the potential to affect image analysis algorithm performance, and the image quality experiment set out to identify if there was a way of automatically identifying whether an image was appropriate for analysis. The study concluded that image intensity values were a useful predictor in ascertaining whether a pathologist would reject an image, when presented using an appropriate field of view (determined by the previous Prospector experiment). However, the results generated from the data were imbalanced, which meant that training an automated solution had high sensitivity and specificity, but very low kappa agreement, and therefore is not usable in automated solutions. As this was a pilot test, further research into this area would need to include more participants, as well as a more balanced dataset, since the binary class distribution was 91% of accepted images and 9% of rejected. This could have been artificially balanced using repetition of data, but balancing the dataset using data augmentation would not have been appropriate in this case due to the extremity of the skew. For the purposes of the pilot study, the results were sufficient for directing further analysis, which is conducted in Chapter 6.

Using the conclusions drawn from this chapter, the following work has been identified for improving the automated solution presented in Chapter 3:

- Use an image patch size of 256x256 pixels in order to utilise surrounding contextual information.
- Apply unsupervised segmentation to the image patch in order to minimise the probability of containing multiple tissue types per segment.
- Analyse each segment individually and apply this contextual to analysing the segment that has the ground truth applied.
- Use global slide intensity features as a benchmark for assessing image segments

It is hypothesised that by applying these changes to the algorithm, the analysis techniques will be more aligned with that of a pathologist, and in turn should increase algorithm-pathologist agreement.

Chapter 5 – Algorithm improvement using contextual analysis

5.1 Introduction

5.1.1 Chapter overview

The work in this chapter uses the conclusions drawn from Chapter 4, in order to develop the ML based algorithm from Chapter 3, and improve per-patch agreement and TSR correlation with the ground truth data. These conclusions lead to work using unsupervised segmentation to generate features from image segments containing only one tissue class, and so are more representative of singular tissue class features. The conclusions also lead to work creating contextual analysis, at both local (native zoom) and global (whole slide) levels, in order to more closely mirror pathologist scoring methods. The chapter is divided into four sections:

1) The introduction, which summarises the conclusions from Chapter 3 and Chapter 4, and describes the direction of work that the conclusions suggest.
2) An experiment into unsupervised segmentation methods and the performance of multiple algorithms, in terms of their ability to create segments that only contain one tissue class, for application to CRC images.
3) Iterative development of the ML algorithm presented in 3.4, which sequentially applies multiple modifications to the algorithm, utilising the contextual information that was concluded as important for human visual assessment in 4.3. The work details six separate modifications, which are individually evaluated on the full dataset of 2,211 cases and 106,242 images.
4) Discussion of the work presented in the chapter and conclusions.

5.1.2 Conclusions from Algorithm A and human scoring

The ML algorithm presented in Chapter 3 showed that algorithm accuracy peaked on 64x64 pixel image patches, leading to two conclusions:

- Larger image patches were likely to contain visual information from multiple tissue classes, which would negatively affect the descriptive power of the feature vectors.
- Smaller image patches were not big enough to contain enough meaningful visual information

The behaviour exhibited by pathologists when visually classifying CRC images was assessed in the experiments presented in Chapter 4, which concluded that manual scoring required larger images in order to maintain optimal levels of agreement. It was concluded that the larger images provide more contextual information surrounding the centre of the image patch, where the ground truth classification is applied. The pilot image quality experiment identified a relationship between image brightness (a surrogate for level of staining) and pathologist rejection of images. This led to the conclusion that the overall level of staining, and therefore global appearance of the slide may affect the value of the visual data at a local level.

These findings suggest that by incorporating contextual information into image analysis, the methodology will more closely resemble expert manual scoring, and therefore improve agreement with the ground truth.

Based on these conclusions, changes to the algorithm were to be made in the following ways:

1) Use the smallest patch size appropriate for manual scoring (256x256 pixels).
2) Apply segmentation to divide the patch into non-square segments that are less likely to contain multiple tissue classes.
3) Use the segmentation results to analyse the local contextual information at the native resolution.
4) Assess the impact of incorporating global contextual analysis using whole slide features.
5) Combine both levels of contextual analysis in order to more closely mirror pathologist scoring.

These modifications were implemented sequentially to assess the impact of each one on the agreement with the ground truth data, before combining them.

5.1.3 Algorithm evaluation methodology

The work in this chapter presents several iterations of improvements to Algorithm A (section 3.4), which uses the same base methodology detailed in section 3.4.2.1, and expands on it in various ways.

Conclusions are drawn from the results of the evaluation, and these are used to improve the algorithm further. As such the development and evaluation of the algorithm is similar for each iteration. To avoid repetition in this chapter, the basic algorithm methodology and evaluation is recapped in this section only, and omitted in subsequent sections. Results are summarised in section 5.3.3, and all results figures (with exception to the final algorithm) are removed from the main text, and can be found in Appendix D.

5.1.3.1 Consistencies in algorithm development

Each algorithm uses the full available QUASAR dataset of 2,211 slides, which contains 106,242 usable image spot locations. Images are extracted using the OpenSlide library to process images directly, rather than reading them over http, and subjecting them to image compression. Based on the conclusions from 4.3, image patches are extracted at 256x256 pixels in size, in order to maintain surrounding contextual information, and processed for the same image features, unless stated otherwise. Each algorithm generated a feature set for each image patch, and the base features generated are detailed in 3.4.2.

5.1.3.2 Consistencies in algorithm evaluation

For each feature set, a RF classifier was used for training and testing, and evaluated using ten-fold cross validation for all available image patches in the full dataset. Cross validation grouping was randomised by slide rather than per image patch, to maintain comparability to manual scoring, and generation of TSR per slide. Using conclusions from the experiment in 3.4, the classifier used 100 trees and three predictors sampled for splitting at each node.

For each algorithm, agreement was calculated in two ways – using the eight individual tissue classes (per-class agreement), and by grouping the eight classes into one of the two parent classes, tumour or stroma. For each method, a confusion matrix was generated, and agreement statistics were calculated from them. ROC curves were generated for each class to ascertain algorithm performance on individual tissue types.

Finally, TSRs were generated for each case, using the algorithm predictions for every spot in the case sampling set. TSRs were calculated using all three methods described in Table 7, but were compared against the pathologist method used in the original study that the QUASAR annotations were generated for. In the case of the algorithms presented in this chapter, that corresponds to Method 1. The algorithm-generated TSRs were subtracted from the pathologist-generated TSRs (using the same method), in order to create a histogram of TSR differences, and statistics regarding the distribution were recorded. In this case, a TSR difference value of zero indicates perfect agreement, a value below zero indicates that the algorithm overestimates the presence of tumour, and above zero indicates that stroma is overrepresented in the results. A Bland-Altman plot is generated to visually inspect the distribution further, and a heatmap scatterplot is visualised so that correlation statistics can be calculated.

5.1.3.3 Consistencies in hardware and software used

Image processing, feature generation and classifier training and testing was implemented in MATLAB v9.0.0.341360, and performed on a VMWare powered Windows Server 2008 R2 64-bit virtual machine with Intel Xeon processors and 32GB RAM. The MATLAB implementation utilised the parallel computing toolbox, assigning 8 logical processors to parallelise the task.

5.1.4 Algorithm reference table

The work in this thesis details the iterative development of a computer vision algorithm, for the generation of TSR on CRC images. Incremental improvements to the algorithm are made, and evaluated independently so that the effects of each modification can be observed. There are 11

Name	Description
Algorithm A	Presented in 3.4, RF-based classifier using a fixed image patch size of 64x64 pixels.
Algorithm B	The methodology from Algorithm A, using a patch size of 256x256 pixels.
Algorithm C	The methodology from Algorithm B, using regular partitioning of the image patch to split into five equal-sized segments in order to analyse context, generating tissue class predictions for each of the segments, with a confidence metric.
Algorithm D	The methodology from Algorithm C, using RF classifier votes for each tissue class (8 sets of votes per segment), for each segment, per image (as opposed to one prediction and one confidence metric per segment, per image).
Algorithm E	Incorporation of the unsupervised segmentation algorithm from 5.2 in order to extract image segments more likely to contain one single tissue type in them.
Algorithm F	Combination of the unsupervised segmentation of Algorithm E with the contextual information processing of Algorithm D, creating an algorithm that uses local contextual analysis to improve pathologist agreement.
Algorithm G	Application of global (whole slide) contextual analysis to Algorithm E to mitigate issues with overall staining variation.
Algorithm H	Combination of Algorithm F and Algorithm G, creating an improved algorithm that uses both local and global contextual analysis to improve pathologist-algorithm agreement.
Algorithm I	The application of Algorithm H to a subset of the QUASAR dataset that has been manually QC approved
Algorithm J	The application of Algorithm H to the CR07 dataset
Algorithm K	The application of Algorithm H to a subset of the QUASAR dataset, that has been automatically assessed and substandard cases removed by the QC algorithm presented in 6.5.

Table 14 - Names and descriptions of all algorithms developed up to and including Chapter 5

5.2 Unsupervised segmentation

5.2.1 Aim

To assess multiple unsupervised segmentation algorithms for their suitability for calculating TSR in CRC digital slide images, and to identify the most appropriate algorithm for maximising accurate segmentations whilst minimising computation cost.

5.2.2 Methods

Note that Algorithm C and D (presented in 5.3) were created and evaluated prior to the image segmentation work in this section. The conclusions from the results of these two algorithms identified the need for unsupervised segmentation rather than regular partitioning. The work is presented in this order so that the ML algorithm iteration work is reported sequentially and concisely.

The results from the development and evaluation of Algorithms C and D indicate that using fixed partitions as a form of image segmentation is too rigid to create meaningful segments of tissue components within CRC images, and that the issue of containing multiple class types per image segment is not resolved. In order to attempt to limit the number of classes to one per image segment, unsupervised segmentation is explored in this section (see section 2.4.5). The chosen unsupervised segmentation method should create non-uniform segments that follow the tissue boundaries. It is hypothesised that this methodology will improve the quality of the feature vectors for individual classes and improve ML training sets, yielding a more accurate classifier. For work presented in this section, evaluation criteria are based on whether segmentation contains segments with more than one class.

5.2.2.1 Segmentation ground truth

To assess the suitability of multiple unsupervised segmentation algorithms on digital CRC images, a subset of data points was randomly selected (without replacement) from the QUASAR dataset. Using the conclusions from the previous experiments presented in Chapter 4, a 256x256 pixel patch size was chosen, with the original pathologist classification label at the centre of the patch. This was in order to minimise the size of the patch whilst maintaining

enough visual information for pathologists to classify the tissue at the centre. A trained technician with six years' experience of spot counting was recruited to apply classification labels to all tissue within each patch. Fifty-three patches were hand labelled using Adobe Photoshop CC 2016, using the brush tool to colour tissue class areas. Tumour was coloured red (RGB value [255 0 0]), stroma was coloured green (RGB value [0 255 0]), and other objects, including tissue, debris and background were coloured blue (RGB [0 0 255]). Figure 76 shows three examples of the original patch images and the hand-labelled classes within them.

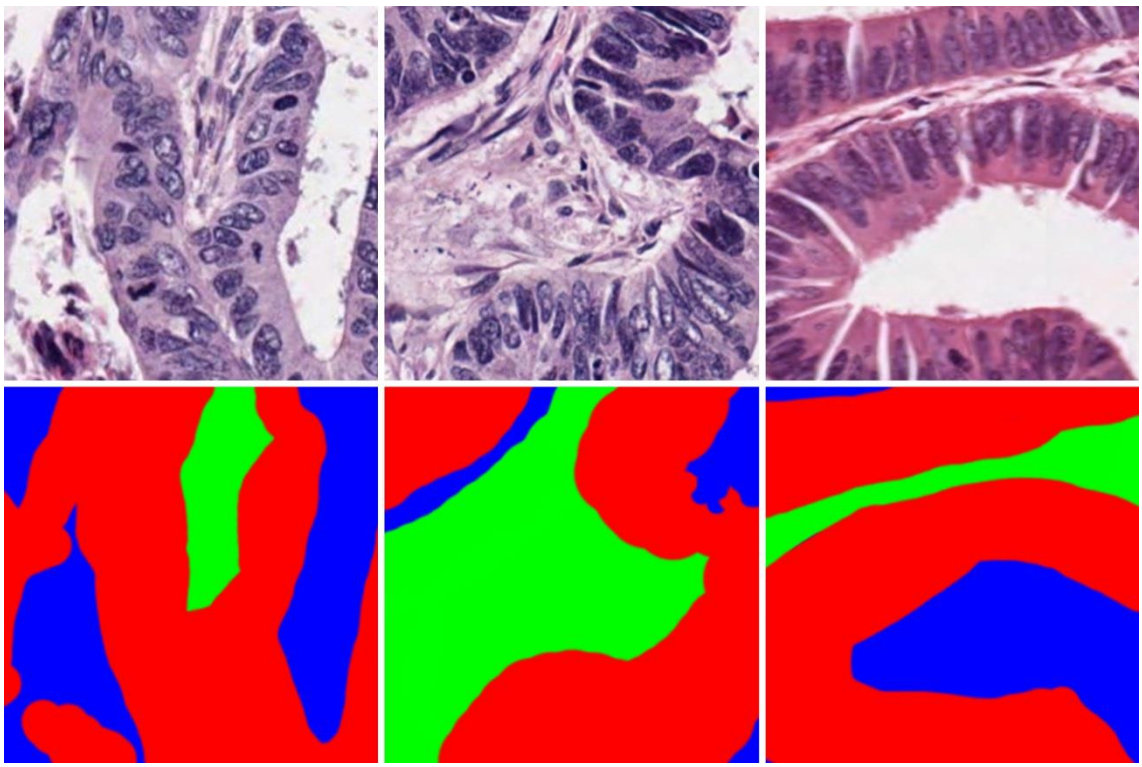


Figure 76 - Examples of images to be segmented and their hand drawn ground truth labels

Top: Original input images to be segmented.
Bottom: Hand drawn ground truth images, where red = tumour, green = stroma and blue = other.

The ground truth obtained in this manner can then be used to identify the presence of multiple tissue classes within a given automatically generated image segment.

5.2.2.2 Number of segments

For each segmentation methodology tested, the target number of partitions is addressed in this section. The number of segments made by each method was affected by the type of segmentation algorithm.

Class-based segmentation

For algorithms that do not account for spatial characteristics such as n-by-n neighbourhoods or Euclidean distance, segments may be grouped by features such as colour and intensity. As a result, pixels are clustered into non-contiguous groups. In this case, the number of clusters was set to the equivalent of the three ground truth classes: Tumour, Stroma and Other. This methodology allows the possibility of making single pixel segments, and so steps to avoid over-segmentation, such as median filtering, were taken as a pre-processing step (see the segmentation methods section).

Region-based segmentation

For algorithms that use localised clustering or other forms of spatial analysis, the number of clusters may be manually set. In this instance, two methodologies were compared in order to assess the most appropriate number to divide into. The first was to define a number of partitions that would create segments of a similar size to 64x64 pixels, which was the best-case image patch size identified by Algorithm A presented in Chapter 3. In this case, an image of 256x256 pixels will yield 16 segments exactly 64x64 pixels in size, illustrated by Figure 77.

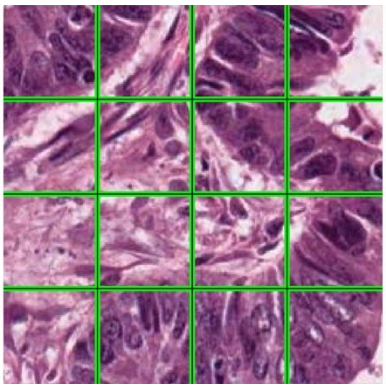


Figure 77 - Illustration of the 256x256 size patch being divided into 64x64 pixel patches

Note that 64x64 pixel patches yielded the highest accuracy when processed by Algorithm A, and so dividing the larger images into similar sized segments should maintain this level of performance.

The second methodology used an optimisation technique in order to automatically select the best number of segments. The technique applied multiple segmentations with different numbers of partitions (from 1 to 20 segments) to the same image patch, and attempted to identify an appropriate trade-off between minimising the number of segments and maximising similarity metrics. The mean standard deviation of intensity per segment was selected as the metric to identify the most appropriate number of partitions, and was selected as the feature to minimise (see Figure 78).

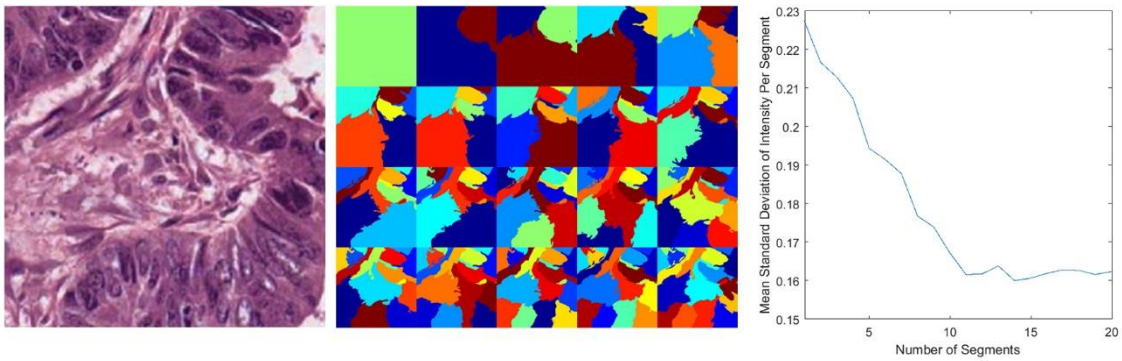


Figure 78 - Investigation into number of partitions vs segment standard deviation of intensity

Left: Original image to be segmented
Centre: 20 separate segmentations, ranging from one segment (top left) to 20 segments (bottom right) using the normalised cuts method
Right: Plot of number of segments against mean standard deviation of intensity per segment

The trade-off between minimising standard deviation (used to derive segmentation accuracy), and minimising the number of segments (indicating lower computational requirements) was calculated by identifying the smallest Euclidean distance between the pair of metrics in 2D space, and the zero origin (0,0) the line graph in Figure 78 indicates that 11 segmentations yield the most appropriate result for that image, using this methodology.

However, applying each segmentation method 20 times to every image in order to select the number with a minimised number of segments and mean standard deviation proved to be computationally expensive compared to a static number of 16 segments (596 seconds compared to 28 using a MATLAB implementation of normalised cuts for both methods). Given that the optimal image patch size for maximising pathologist agreement with Algorithm A was 64px, and that the automatic optimisation yielded minimal increases in segmentation accuracy, the methodology using a static number of 16 cuts was implemented.

5.2.2.3 Segmentation methods

Nine different unsupervised segmentation methods were used for evaluation, which either used class-based segmentation or region-based segmentation, depending on whether the algorithms used spatial information to cluster pixels. These methods are described below.

Method 1: Pixel intensity thresholding

Pixels were clustered into three different classes, depending on their intensity values (see section 2.4.5.1). In order to avoid over-segmentation, a median filter of size 75x75 pixels was applied to the intensity channel of the HSV colourspace (see explanation of the method in section 3.3.2.6). Background pixels were thresholded at a static value of 240, and the remaining

pixel values were binned into two groups, using a threshold of the mean intensity value, minus one half of the standard deviation of the intensity values. This created a label image of three non-contiguous classes, which meant that the number of segments produced was variable, depending on the image data.

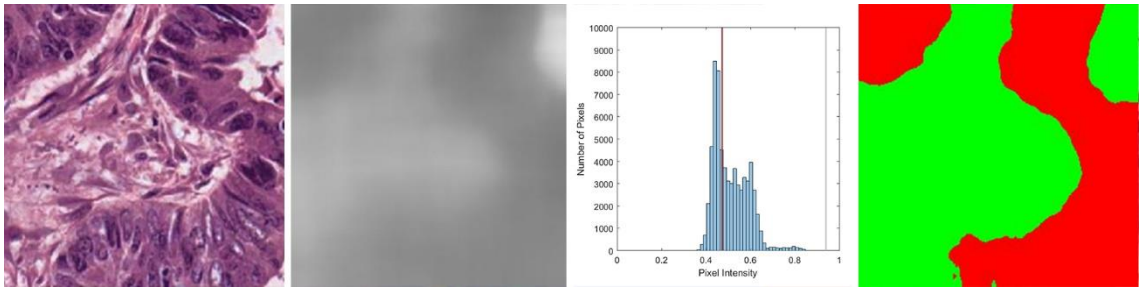


Figure 79 - Pixel intensity thresholding segmentation method

Left: Original Image
Centre Left: The median filtered greyscale image, using a 75x75 pixel kernel
Centre Right: The histogram of intensities after pre-processing. The grey line shows the background threshold of 240 (as a percentage of 255), and the red line shows the threshold of the mean intensity, minus half of one standard deviation
Right: The resulting segmentation (note that there is no background (pixels > 240 intensity) accounted for in this image

Method 2: Watershed segmentation

Segments were created using local minima as seed points, and regions were grown using the watershed segmentation technique (see section 2.4.5.2). As a pre-processing step, images were median filtered with a 75x75 pixel kernel on the intensity channel of the HSV colourspace, and intensity values were adjusted using automatic histogram stretching. An H-minima transform was applied to supress minima of a value less than 20 before applying the watershed segmentation algorithm.

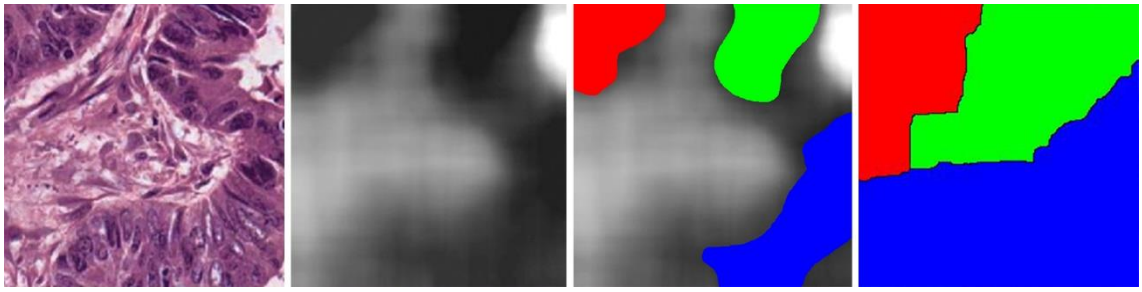


Figure 80 - Watershed thresholding segmentation method

Left: Original Image
Centre Left: Greyscale image patch, with histogram stretching to create areas of minima
Centre Right: Identified areas of minima
Right: Watershed segmentation using minima-seeded regions

Method 3: Texture-based segmentation

A local entropy filter was applied, using a 9x9 kernel to the intensity channel from the HSV colourspace. The resulting image was thresholded using a static threshold of 0.8 in order to identify areas of high levels of textural appearance (higher local intensity variance). Once binarized, morphological clearing was used to remove objects smaller than 2000 pixels in size, using a closing neighbourhood of 9x9 pixels.

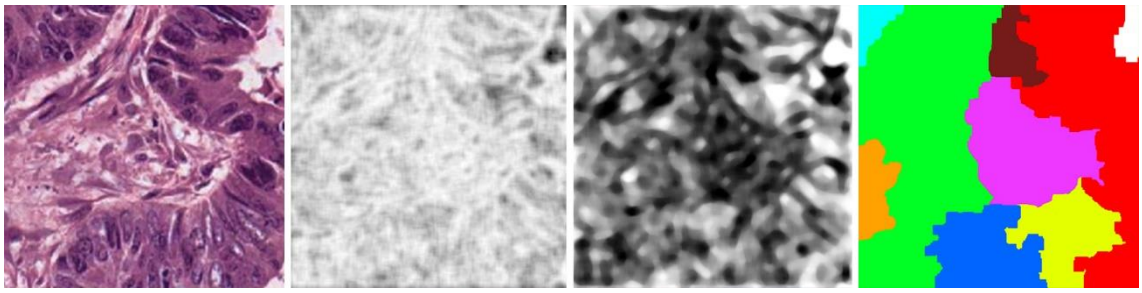


Figure 81 - Texture based thresholding segmentation method

Left: Original Image
Centre Left: Entropy-based texture filter on greyscale image patch
Centre Right: Histogram equalisation applied to the entropy image
Right: Segmented image, using the adjusted texture image to identify regions of similar values.

Method 4: K-Means clustering

Images were median filtered by applying a 75x75 kernel across each of the RGB channels of the three-tensor matrix in order to reduce noise before applying K-Means clustering (see section 2.4.5.3). The images were clustered into three prototype classes based on the RGB colour of each pixel, in an attempt to generate non-contiguous segmentations for tumour, stroma and background.

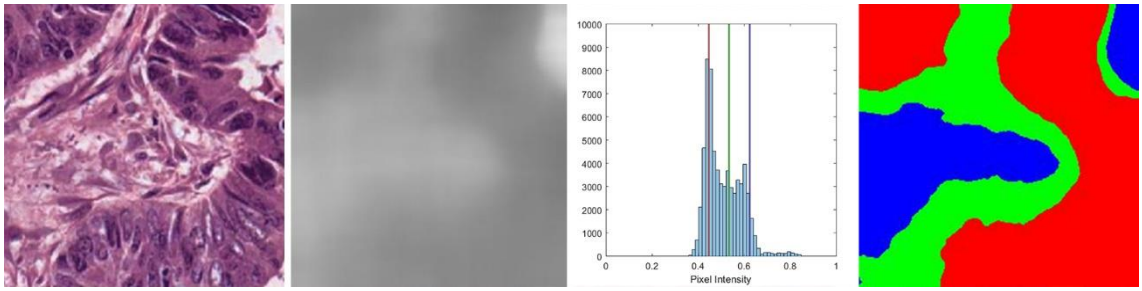


Figure 82 - K-Means Clustering image segmentation method

Left: Original Image
Centre Left: Greyscale image patch with median filter of size 75x75
Centre Right: Intensity histogram with 3 cluster centres overlaid in red, green and blue
Right: Segmented image into 3 non-contiguous classes

Method 5: Mean-shift segmentation

Images were median filtered by applying a 10x10 kernel across each of the RGB channels of the three-tensor matrix in order to reduce noise before applying mean-shift segmentation (see section 2.4.5.3). The parameters used for mean shift segmentation were a spatial bandwidth of 20, range bandwidth of 20 and a minimum segment size of 5000 pixels.

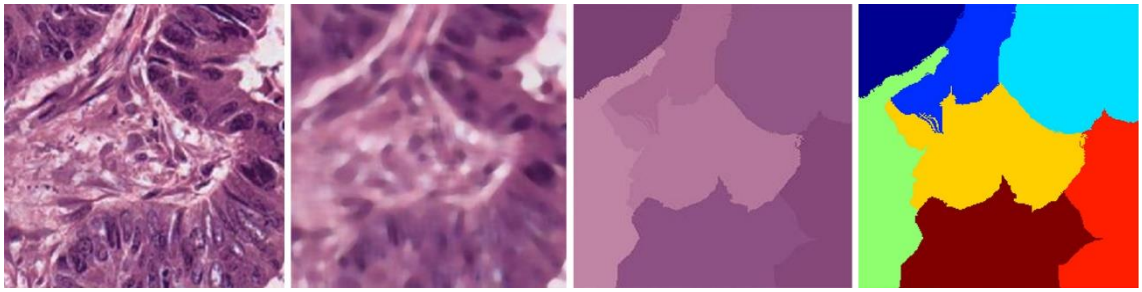


Figure 83 - Mean-shift segmentation method

Left: Original Image
Centre Left: Median-filtered image using 10x10 kernel over the individual RGB channels
Centre Right: Mean-shift segmentation using prototype colours to label the segments
Right: Segmented image

Method 6: SLIC – Simple Linear Iterative Clustering of superpixels

Using the SLIC algorithm (see section 2.4.5.3), segments were created by clustering around a randomly sampled equidistant grid of seed points (similar to the RandomSpot system in 3.2.2). Images were median filtered on individual RGB channels using a kernel size of 10x10. The target number of clusters (and so grid seed points) was 16, but due to the random nature of the grid application, that number was variable. Weighting between colour and spatial information

was set to 20 and a merging radius (for merging small segments) was set to 7 pixels. The cluster centres were calculated using mean colour values.

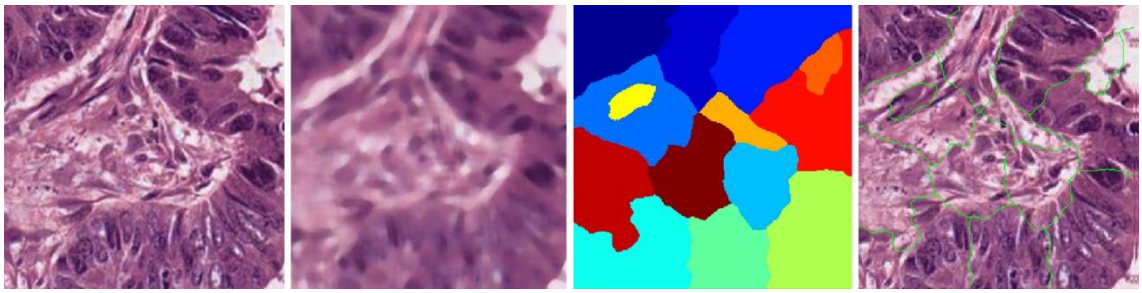


Figure 84 – SLIC superpixel segmentation method

Left: Original Image
Centre Left: Median-filtered image using 10x10 kernel over the individual RGB channels
Centre Right: Segment labels generated by SLIC algorithm
Right: Segment boundaries overlaid on original image

Method 7: Graph cuts

Graph cuts (see section 2.4.5.4) was applied to the median-filtered RGB patch image, using a filter kernel size of 10x10 on each of the three channels. K-Means clustering was applied to identify 3 class prototypes. For each of the classes, the Euclidean distance image was calculated for each pixel, to each of the three class prototypes. A smoothing cost constant of 1.5 was applied to enhance neighbourhood constraints and create larger segments. Graph cuts optimisation was applied to the resulting distance images, in order to create non-contiguous labels for each of the three prototypes.

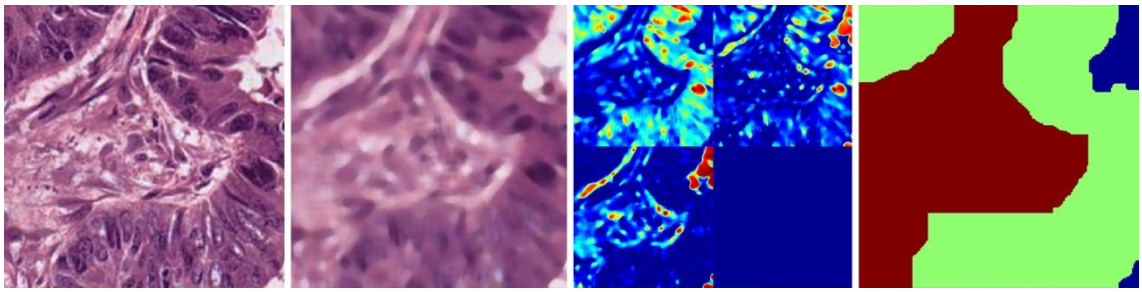


Figure 85 – Graph-cut based segmentation method

Left: Original Image
Centre Left: Median-filtered image using 10x10 kernel over the individual RGB channels
Centre Right: Heatmap of pixel distances to class prototypes (3 used, bottom right is blank)
Right: Segmented image, partitioning prototypes into 3 clusters

Method 8: Normalised cuts

The normalised cuts algorithm (see section 2.4.5.4) was applied to the intensity channel of the HSV colourspace using the pixel values as a 2D graph problem. The intensity channel was chosen as opposed to the deconvoluted H or E channel because of the loss of structural tissue information that would occur. Per-pixel comparisons were made using several features to compute the affinity matrix. These features were differences in intensity and texture values, Euclidean distance and the maximum edge strength between the two pixels (known as intervening contours, illustrated in Figure 88). The graph was partitioned into 16 clusters per image using the dissimilarity metrics to apply the graph cuts.

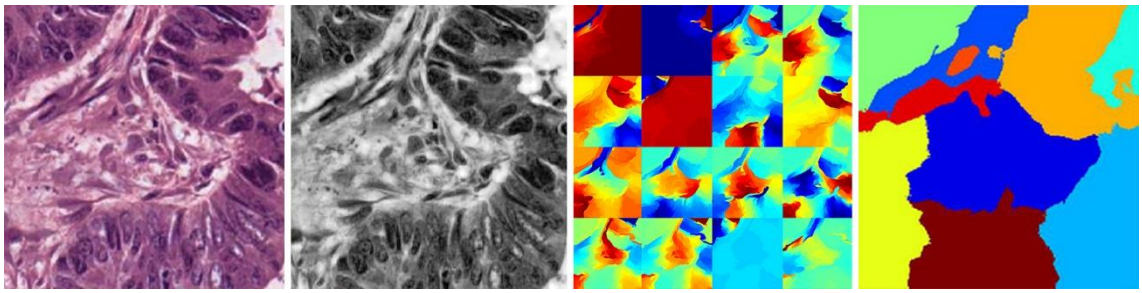


Figure 86 - Normalised Cuts algorithm applied to a greyscale image patch

Left: Original Image
Centre Left: Greyscale image showing the Haematoxylin channel of a 256x256 pixel H&E image patch.
Centre Right: Eigenvalues of the 16 clusters of pixel features for segmentation.
Right: Normalised cut segmentation into 16 segments.

5.2.2.4 Extension to segmentation methods

Through the development and application of the 8 described methods, it became apparent that the normalised cuts method provided an accurate solution, but at a high computing cost. This was due to the algorithm being applied to a symmetric affinity matrix, which calculated a similarity metric for every pixel pair, meaning that the 256x256 pixel image required 65,536 feature vectors to be computed (one for each pixel), and 4,294,967,296 pairwise comparisons between sets of features to be computed.

Method 9: Normalised cuts to cluster superpixels

To reduce the size of the affinity matrix, and therefore the number of pairwise comparisons to be computed, the normalised cuts algorithm was applied to a set of superpixels (instead of per-pixel). Each image was over-segmented into superpixels using the SLIC algorithm, using a target number of 168 superpixels, which divided the 256x256 pixel image into areas

approximately the same size as one nucleus (10µm^2, given that the slides were scanned at 0.5µm per pixel).

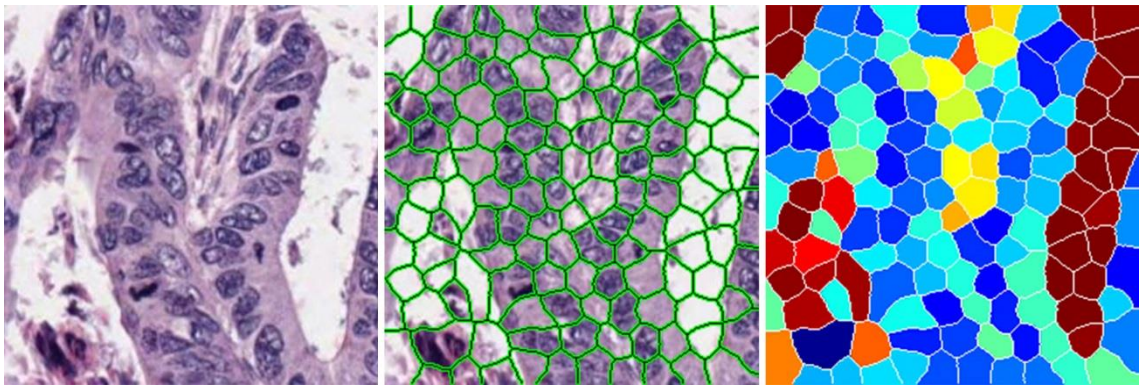


Figure 87 - Example of SLIC superpixel clustering

Left: 256x256 pixel image patch of H&E stained CRC tissue.
Centre: Superpixel clustering to perform over-segmentation.
Right: Colour heatmap of median superpixel intensity values.

For every superpixel, statistics were calculated to be used in the feature vector for generating the similarity metric between superpixel pairs. For every superpixel pair, a similarity metric was computed, which used the same features as the Normalised Cuts algorithm, except that median intensity was used instead of pixel intensity and Euclidean distance was calculated between superpixel centroids. The similarity metric consisted of Euclidean distance, median intensity difference, mean absolute difference of intensity, texture difference (comparing localised texture features called textons [264]) and maximum intervening contours. Figure 88 illustrates these features.

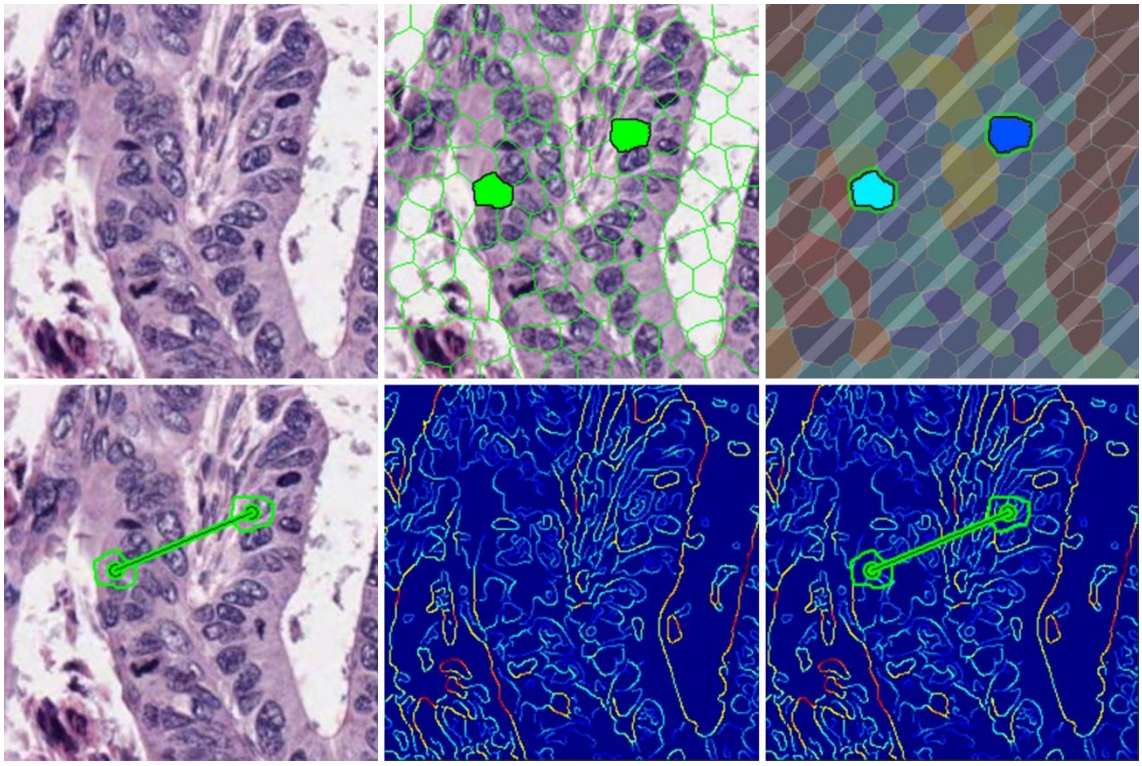


Figure 88 - Visualisations of features used for calculating the superpixel similarity metric for Normalised cuts

Top Left: Original 256x256 H&E stained CRC image.
Top Centre: Over-segmentation boundaries and highlighted superpixel pairs to compare.
Top Right: Highlighted superpixel pairs on colour heatmap, comparing median intensity values.
Bottom Left: Euclidean distance between centroids of superpixel pairs.
Bottom Centre: Colour heatmap of edge intensities.
Bottom Right: Euclidean distance between superpixel centroid pairs, denoting the line to calculate maximum edge intensity value along (intervening contours).

Similarity metrics were used to create a symmetric affinity matrix, and the normalised graph cuts algorithm was applied so that superpixels were clustered into image segments. As previously discussed, a fixed number of sixteen segments was chosen in order to create images approximately 64x64px in size, which yielded the highest accuracy result using the original algorithm. Figure 89 illustrates the superpixel clustering process combined with the affinity matrix to make an image segmentation.

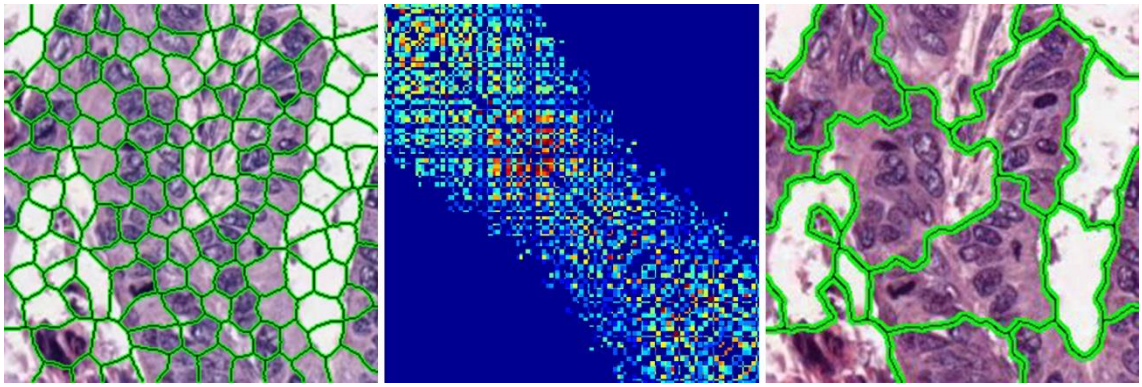


Figure 89 - Superpixel clustering combined into segments using normalised cuts on custom affinity matrix

Left: Over-segmentation using SLIC superpixel clustering
Centre: 168x168 affinity matrix comparing all superpixel pairs
Right: Image segmentation applying normalised cuts to superpixel affinity matrix

With all nine segmentation methods established, they were each run on the 53 images to compare to the hand-labelled ground truth data.

5.2.2.5 Evaluation methods

Each of the segmentation methods generates a set of segment labels and no classification data. The success of the method depends upon successfully delineating boundaries without over-segmenting the image. The aim of the evaluation for this work was to identify the algorithm that optimised the following criteria:

- Maximise the accuracy of the segmentation
- Minimise the time taken to compute

Accuracy, in this instance was the agreement with the ground truth. Since the boundaries of the tissue in the image were of no standardised appearance, and objects in the image had no consistent shape, assessing segmentation by comparing boundary tracing would be less meaningful than identifying whether the segments generated had one single class of tissue within them. The methodology for evaluation used the ground truth as an image mask to apply to each generated segment, and the percentage area of the three ground truth classes was calculated per segment. The tissue class with the highest percentage coverage per segment was considered to be the most appropriate classification for that segment, which meant that the percentage coverage of that particular class would be the percentage accuracy of the segmentation. This was repeated for all segments in the image, and repeated for all images, for every method. Figure 90 illustrates the combination of the segmentation result and the ground truth in order to calculate the percentage accuracy for that segmentation. The bottom right image in the figure shows a segmentation with a coverage of 5% tumour, 95% stroma and 0%

other. Since stroma has the largest coverage, the accuracy of that particular segmentation is 95%.

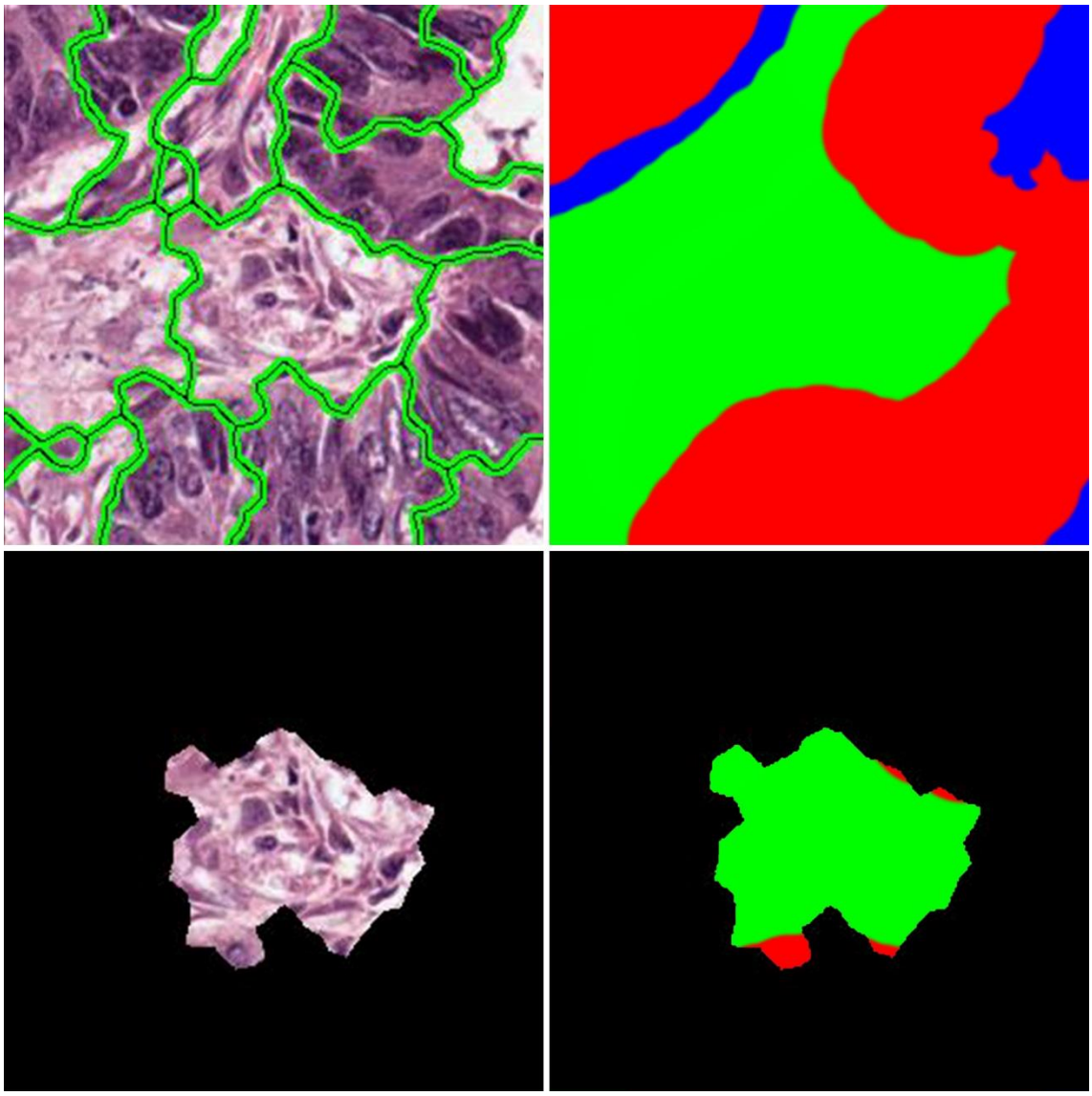


Figure 90 - Evaluation of segmentation algorithms comparing to ground truth

Top Left: Algorithm segmentation result (SLIC + Normalised cuts in this example).
Top Right: Hand-drawn ground truth label image.
Bottom Left: Single segment for evaluation.
Bottom Right: Segment mask applied to ground truth image, for establishing percentage of tissue types within segmentation result (5% Tumour, 95% Stroma, 0% Other).

The percentage accuracy was calculated for every segment, for every image, for every segmentation method.

5.2.3 Results

Results for each method were calculated as mean processing time per image, the mean number of segments each method generated per image, the success rate (mean percentage area covered of one class per segment, per image), and the standard deviation of the success rate. Table 15 displays these data for all nine segmentation methods.

Method	Processing time	Mean # segments	Success rate	SD
Intensity thresholding	4.12s	6.90	0.89	0.09
Watershed segmentation	4.16s	4.52	0.78	0.16
Texture thresholding	0.23s	8.23	0.92	0.09
K-Means clustering	4.25s	9.79	0.90	0.07
Mean-shift segmentation	2.79s	5.10	0.91	0.07
Superpixel clustering	1.18s	15.65	0.92	0.05
Graph cuts	0.82s	3.65	0.90	0.10
Normalised cuts	20.18s	16.00	0.97	0.02
Hybrid clustering	10.71s	16.00	0.93	0.05

Table 15 - Results for all nine segmentation algorithms

Processing time is recorded per image, mean number of segments and success rates are calculated across all segments for all images in the test set. The normalised cuts algorithm yields the highest accuracy with a success rate of 0.97, but has the highest computational cost of 20.18 seconds.

Algorithms were assessed on the evaluation criteria described in 5.2.2.5

5.2.3.1 Maximise accuracy of segmentation

Segmentation accuracy (success rate) of the algorithm is defined as the mean percentage area coverage of a single tissue class per image segment, per image (see Figure 90). The highest success rate is generated by the Normalised Cuts algorithm, which has a rate of 0.97, and a standard deviation of 0.02. The lowest success rate is generated by the watershed segmentation algorithm, which has a rate of 0.78 and a standard deviation of 0.16. Figure 91 displays the success rates of all nine algorithms in boxplots for visual comparison.

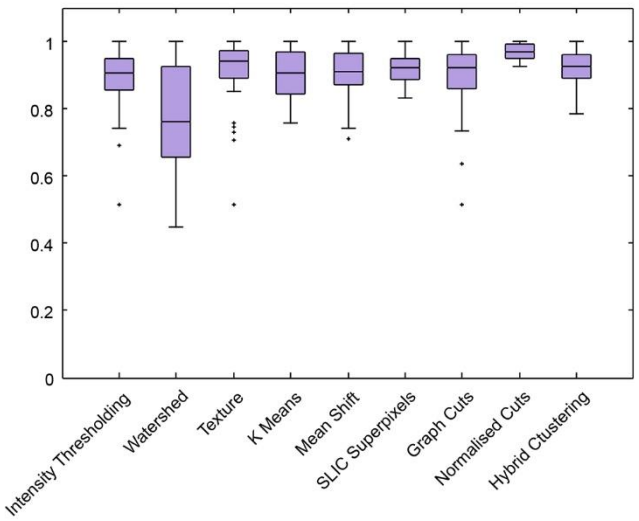


Figure 91 - Boxplots showing mean success rate of each segmentation method

Success rate is the highest percentage of area covered by one of the three classes provided by the ground truth markup images, per image segment, per evaluation image. The normalised cuts method yields the highest segmentation accuracy, and watershed segmentation yields the lowest.

For the unsupervised segmentation results, a Friedman's (repeated measures) ANOVA of the mean percentage of maximum segment area was conducted with segmentation method (total = 9) as the repeated measures independent variable. Results revealed a significant effect of segmentation method ($\chi 2 = 139.82$, $p < 0.01$). Post-hoc analysis of pairwise comparisons using the Wilcoxon signed rank test are presented in Figure 92.

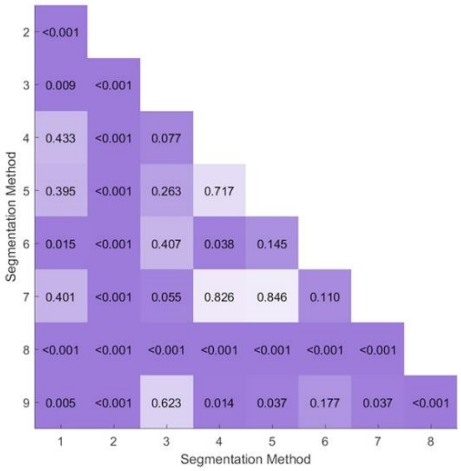


Figure 92 – Wilcoxon signed rank test significance values for all segmentation method success rate pairwise comparisons

Note that Bonferroni correction changes the significance cut-off value from 0.05 to 0.006, and subsequently, significance is reported to three decimal places. Only the results of the normalised cuts and watershed algorithms report statistically significant differences, meaning that all other algorithms perform similarly.

Note that Bonferroni correction changes the significance cut-off value from 0.05 to 0.006, due to the nine groups being compared.

5.2.3.2 Optimise computational time

Each algorithm was individually run on the same set of 53 images, and the time taken to process each was recorded. The texture-based segmentation algorithm took the least amount of time to process the images, averaging 0.23 seconds per image, and the normalised cuts algorithm required the most processing time, taking 20.18 seconds per image. The distribution of times per algorithm is presented in Figure 93.

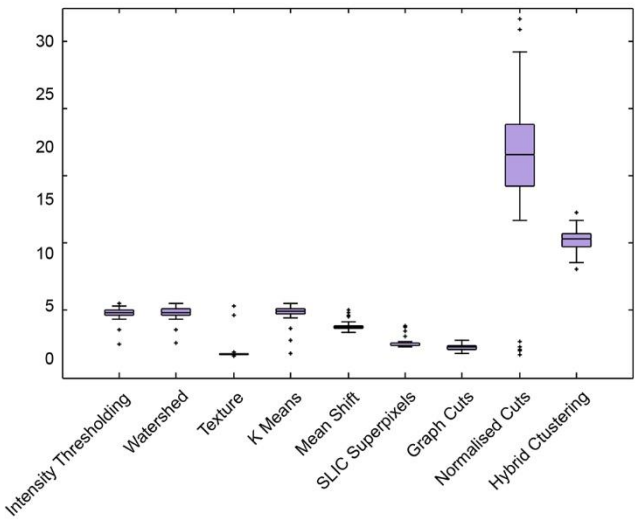


Figure 93 - Boxplots of segmentation method processing times

Normalised cuts algorithm took the longest time to compute with a mean time of 20.18 seconds per image. The texture-based segmentation algorithm required the least processing time, taking 0.23 seconds per image.

The purpose of the hybrid clustering algorithm was to improve on the normalised cuts processing time, and Figure 94 illustrates the reduction in complexity of the affinity matrix between the two algorithms. However, the reduction in computing time per algorithm appeared disproportionate to the decrease in computations required.

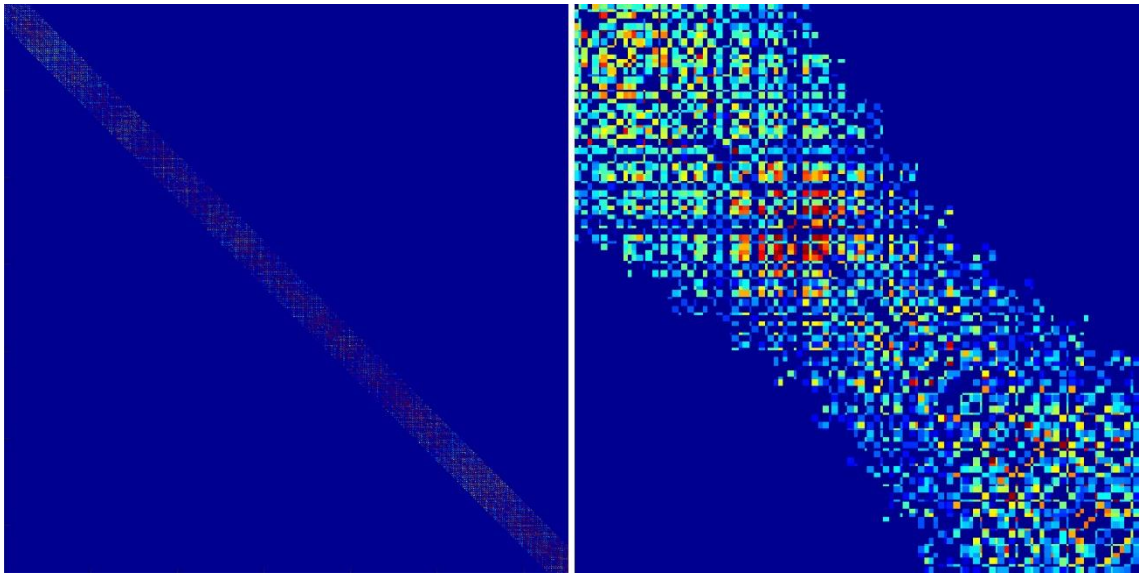


Figure 94 - Affinity matrix comparison between two segmentation methods

Left: Affinity matrix for the Normalised Cuts image segmentation algorithm, $65{,}536^2$ in size.
Right: Affinity matrix for the Hybrid superpixel clustering algorithm, 168^2 in size.

The features that were calculated for the superpixels were to be used in the ML algorithm further down the processing pipeline. Therefore, for ease of comparison, the computational gains that would occur elsewhere in the algorithm were subtracted from the results. Table 16 displays the modified mean processing time per image, which reduces the hybrid clustering algorithm from 10.71 seconds to 2.53 seconds per image, reducing time taken over the original normalised cuts algorithm by 87%.

Method	**Processing Time**
Intensity Thresholding	4.12s
Watershed Segmentation	4.16s
Texture Thresholding	0.23s
K-Means Clustering	4.25s
Mean Shift Segmentation	2.79s
Superpixel Clustering	1.18s
Graph Cuts	0.82s
Normalised Cuts	20.18s
Hybrid Clustering	2.53s

Table 16 - Time taken to perform segmentations when removing feature calculation from the hybrid clustering method

The table shows that the hybrid clustering algorithm is more comparable to the previous methods in terms of computation time.

The modified timing results are displayed in the box plots in Figure 95.

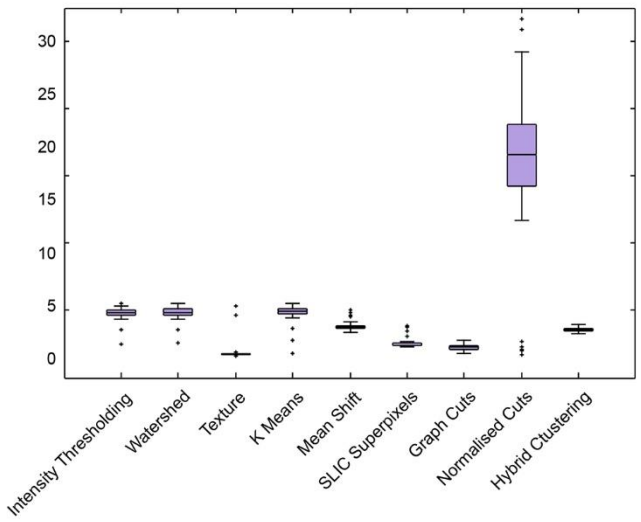


Figure 95 - Boxplots of segmentation method processing times removing feature calculation from the Hybrid Clustering method

The graph illustrates that hybrid the clustering method requires less computation time when accounting for feature vector reuse. The normalised cuts method requires significantly more processing time.

Finally, to assist with selection of an appropriate algorithm, Figure 96 plots the two unsupervised segmentation algorithm optimisation criteria (mean segmentation accuracy and processing time per segment). In this plot, segmentation error is simply the inverse of the previously presented success rate, in order to visualise that the smallest distance to [0,0], which identifies the most suitable algorithm for the task.

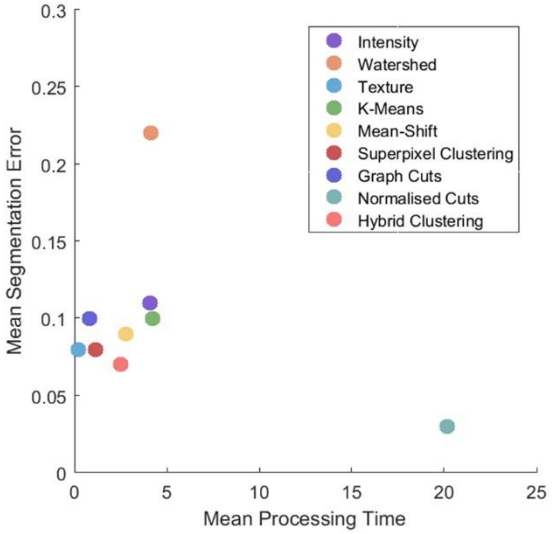


Figure 96 - Scatter plot illustrating optimisation criteria for unsupervised segmentation methods

The two criteria are processing time and algorithm accuracy, and in this plot, segmentation error (the inverse of segmentation accuracy) is used so that plot points closer to [0,0] are more suited to the task. The figure shows that the hybrid clustering method is the most successful at minimising error and time.

5.2.4 Conclusions

The normalised cuts algorithm has the highest mean accuracy rate of 97%, at the expense of a 20.18 second computing time per image patch. Pairwise comparisons revealed that the normalised cuts algorithm is statistically higher than all other algorithms ($p < 0.001$ for all comparisons). However, the cost in computing time is currently considered too high for processing large clinical datasets.

Optimisation of the affinity matrix using nucleus-sized superpixels decreases computing time by half. However, this time is disproportionately high compared to the reduction in size of the affinity matrix, due to inefficient feature calculation. The resulting features are to be reused in the ML of the image analysis algorithm, which reduces processing time elsewhere in the algorithm, and therefore was discounted from the results analysis. This reduced the processing time of the hybrid clustering algorithm to 2.53 seconds.

The trade-off between segmentation accuracy and computational performance illustrated in Figure 96 meant that the hybrid clustering algorithm was chosen for implementation in the development of the artificial intelligence algorithm (Algorithm E and onwards).

5.3 Iterative algorithm development using unsupervised segmentation and contextual analysis

5.3.1 Aim

To improve Algorithm A, by iteratively applying modifications based on the conclusions from the previous studies.

5.3.2 Methods

The incremental improvements to Algorithm A were made by applying one modification at a time in order to assess their effect on agreement individually. All algorithms and their modifications are described in Table 14. For the purposes of clarity, each algorithm has been assigned a letter, so that they may be referred to concisely in figures.

Using the methodology from Algorithm A presented in 3.4, and the conclusions drawn from the observer experiments in section 4.3 and 4.4, each algorithm was developed, focusing on one single conclusion drawn from the previous work.

As with Algorithm A, the image data from the QUASAR clinical trial dataset was used to train and test the algorithm using 10-fold cross validation over 2,211 cases, containing 106,242 expert-classified image coordinates (see section 3.3).

All algorithms presented in this section extracted image patches at a size of 256x256 pixels. The patch size of 256x256 pixels was chosen so that the surrounding visual information (context) could be included in the analysis, based on the conclusions from the image size experiment presented in 4.3. This contextual information was averaged in the feature vectors of Algorithm A and B (discussed in section 3.5), which led to the conclusion that some form of segmentation

was required to create image segments that contained only one tissue class, and subsequently feature data which more accurately represented each class.

All processing was performed on the hardware detailed in section 5.1.3.

5.3.2.1 Algorithm C and Algorithm D: Contextual analysis using regular partitioning

Note that Algorithm C and D were created and evaluated prior to the image segmentation work in section 5.2. The conclusions from the results of these two algorithms identified the need for unsupervised segmentation rather than regular partitioning. The work is presented in this order so that the ML algorithm iteration work is reported sequentially and concisely.

Algorithm C and D attempt the segmentation task using fixed size and shape partitions of a given image patch. Several methods of dividing the patch into regular segments were considered, and an appropriate partitioning method was selected, based on the following specifications:

1. Each partition should have the same area size in pixels (but not necessarily shape).
2. Minimise the number of partitions, and subsequently the number of feature vectors generated per image.
3. Avoid segment boundaries crossing the centre of the image patch (the ground truth classification point).
4. Utilise all of the information from the extracted image.

A visual example of the partition design chosen is displayed in Figure 97, applied to a CRC patch. Note that the resulting partitions may or may not contain multiple tissue classes per patch.

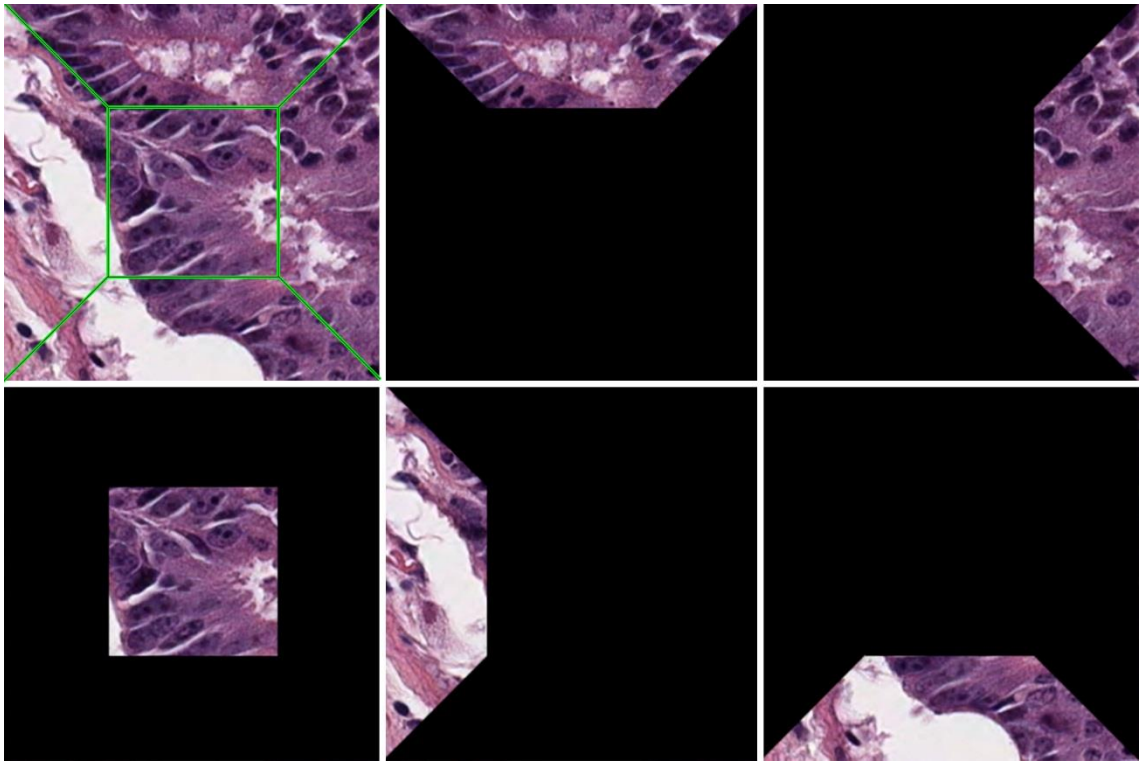


Figure 97 - Illustrated examples of image segments using regular segmentation method D

Regular partitioning is applied in order to reduce the probability of containing more than one tissue class per partition, as well as obtaining contextual information within the image patch. Each partition is a consistent size in terms of pixel area (13107px^2), but not shape. The centre partition is a square of 114x114 pixels.

To combine the visual information from each of the partitions into a single feature vector per patch, two methodologies were implemented and assessed as separate algorithms (Algorithm C and D). Both methods generated feature vectors for all five segments of the patch, using the 18 features detailed in Algorithm A in section 3.4. Once generated, the features were used to train the RF classifier (optimised in section 3.4.2.2) using features calculated from the centre partition only (where the ground truth classification was applied).

The trained classifier was applied to the surrounding partitions in order to generate features for the final feature vector of the image patch. The new feature vector for the centre partition was created using the original features of the centre partition, and either the classifier **predictions** (Algorithm C), or the **votes** (Algorithm D) for each class. The algorithms differed in the following way:

- The prediction-based algorithm had one prediction per partition, with a confidence metric (largest percentage of the 8 sets of votes from the RF classifier)
- The votes-based algorithm had 8 sets of votes for each partition

Algorithm C: Predictions-based features

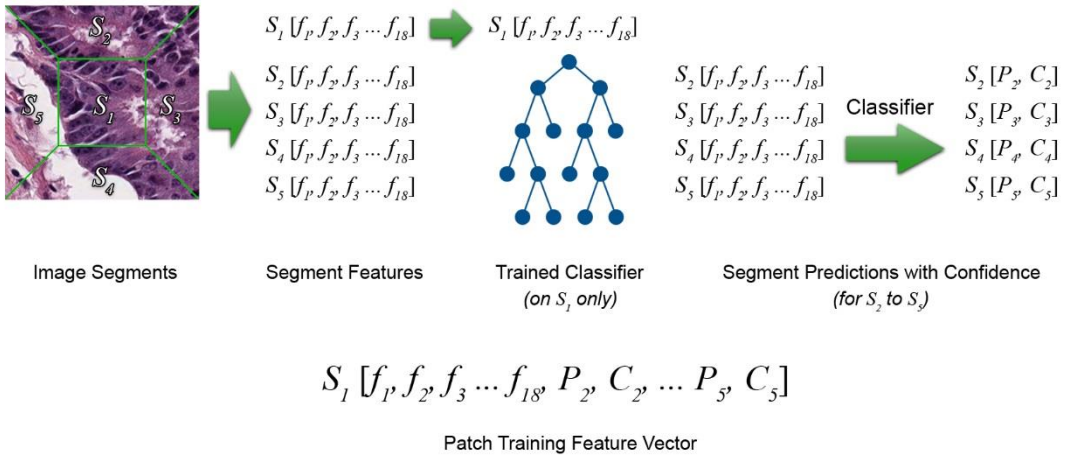


Figure 98 - Diagram illustrating the feature vector generation process for Algorithm D - with classifier predictions and confidence values

The diagram shows that for every partition, two values were recorded – the numeric class prediction and the probability of that class, as a percentage of the total votes given by the RF classifier.

The centre partition features were concatenated with a numeric class prediction for each of the remaining four partitions, as well as a confidence metric, which was calculated as the percentage of the votes that the predicted class obtained (see Figure 98). This created a feature vector of 26 columns.

Algorithm D: Votes-based features

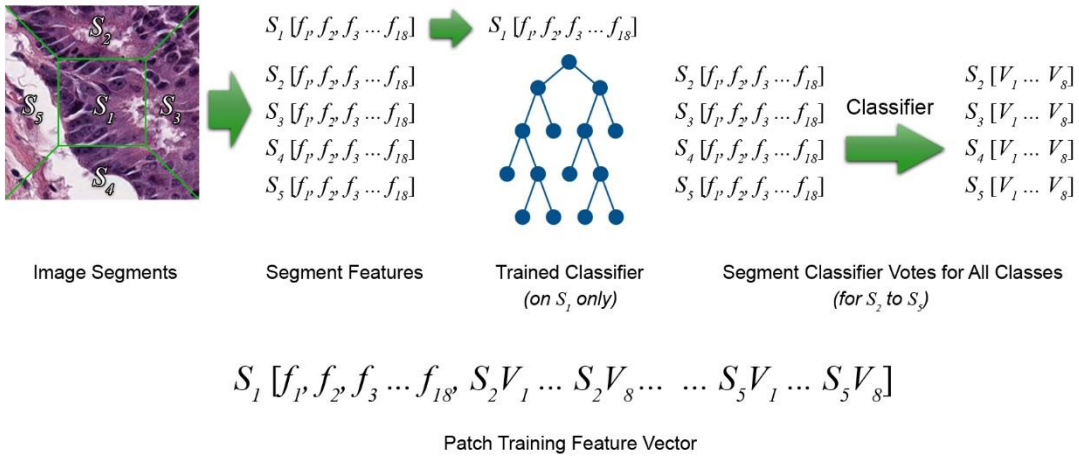


Figure 99 - Diagram illustrating the feature vector generation process for Algorithm D - with classifier votes for each class, per segment

The diagram shows that for every partition, eight values are recorded – the number of votes obtained for each of the tissue classes, as distributed by the RF classifier.

The centre partition features were concatenated with the number of votes for each of the eight candidate classifications, for all four of the neighbouring partitions (see Figure 99). This created a 50-column feature vector.

5.3.2.2 Algorithm E: Unsupervised segmentation of image patches

Algorithm E incorporates the unsupervised segmentation from 5.2 in order to extract image segments more likely to contain one single tissue type in them, mitigating the issue of larger image patches containing too much surrounding contextual information, discussed in 3.5 and illustrated in Figure 60. The algorithm was developed focusing on identifying a single segment at the centre of the patch, with the aim of that segment containing only one tissue class.

Using the unsupervised segmentation method developed in section 5.2, Algorithm A was extended to extract images at 256x256 pixels, and identify the image segment at the centre of the patch (where the ground truth classification was applied). This created a non-regular image segment, in terms of size and shape. This was performed with the intent of containing only one tissue class, for generating more precise feature vectors.

5.3.2.3 Algorithm F: Contextual processing of segmented of images

Combining the contextual analysis methodology from Algorithm D in section 5.3.2.1, with the unsupervised segmentation methodology detailed in section 5.2, Algorithm F focuses on utilising the contextual data surrounding the centre of the patch, with non-uniform segment boundaries.

Algorithm F used a similar methodology to Algorithm D, in that it applied a trained classifier to the neighbouring segments of the patch, and used the classifier votes for each segment in the feature vector for that patch. The trained classifier developed in Algorithm E was used to provide votes for candidate classes, for the neighbouring patches. The contextual information had two key differences to Algorithm D's contextual analysis:

1. The neighbouring segments were of a non-uniform shape
2. The number of neighbouring segments was not constant

The variable number of neighbours, and therefore number of votes to add to the patch feature vector, was accounted for by combining the number of votes for each class, for all segments (eight features, one for each tissue class). The non-uniform segmentation of the neighbours meant that cumulative counting of votes per neighbour would favour multiple smaller segments. This issue was mitigated by using the percentage length of the segment boundary that overlapped the centre segment perimeter boundary, as a weighting function of the votes for that neighbour (see Figure 100).

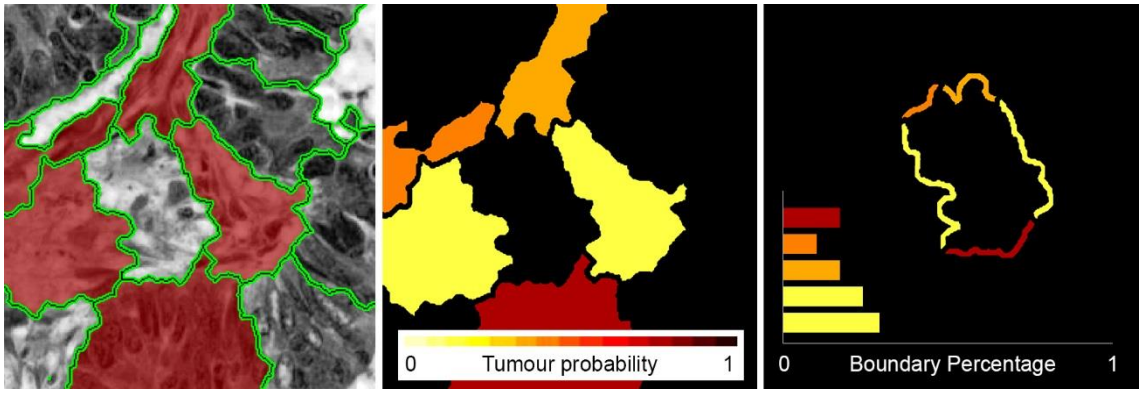


Figure 100 – Perimeter-based tumour probability weighting

Left: Neighbouring image segments comprised of clustered superpixels, highlighted in red
Centre: Colour heatmap of tumour segment probability on neighbouring segments
Right: Neighbouring segment probabilities as a percentage of centre segment boundary length

By using the summation of the surrounding tissue class probabilities, weighted by the length of the adjacent perimeter length as a percentage, artefacts (in the form of disproportionate counts of class votes) from over-segmented regions are avoided.

The resulting feature vector contained the original 18 features from section 3.4.2, with eight additional features: the boundary-weighted votes for each of the eight tissue classes.

5.3.2.4 Algorithm G: Global contextual analysis of stain

There are many factors involved when considering the global appearance of histological tissue samples. Most notable is the inconsistency in colour, which can be attributed to issues such as section thickness, variations in staining chemical compounds, application of the stains, quality of the glass slides and cover slips, length of time between slide generation and digital slide scanning, slide scanning hardware, and colour profiles of the digital scanners to name a few. The clinical dataset used for the development of the algorithms in this chapter was designed to provide five-year survival data, and as such, has large variation in appearance (see Figure 101).

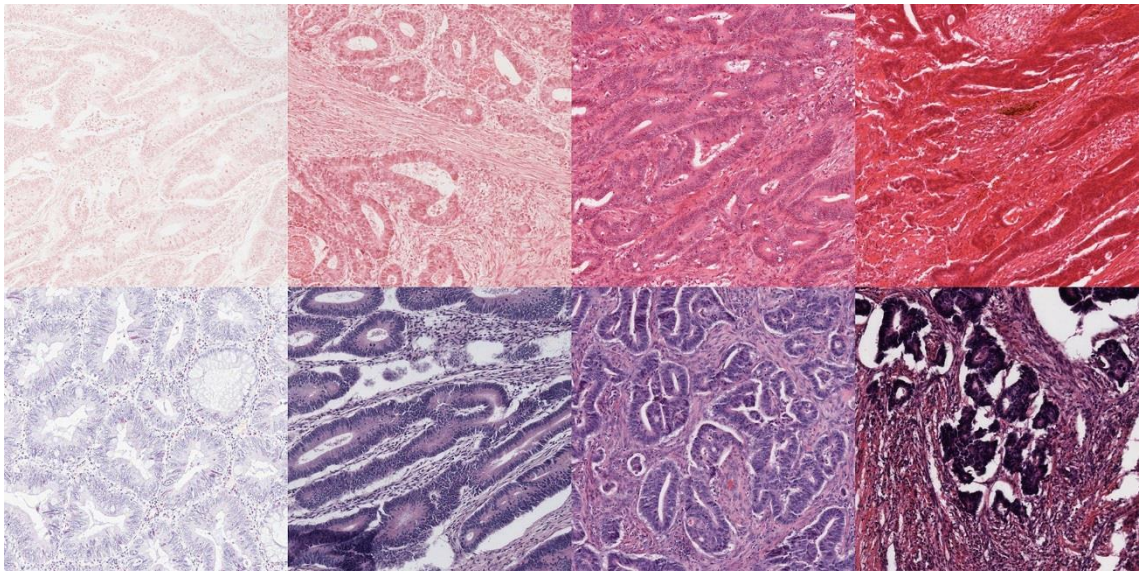


Figure 101 - Examples of different levels of staining in the clinical dataset

The top row illustrates different levels of eosin staining from low to high when haematoxylin staining is weak. The bottom row illustrates examples of weak to strong haematoxylin staining.

This variation becomes problematic when using image and stain intensities as training features, and therefore must be compensated for in the algorithm. Colour normalisation provides a robust method for correcting for inconsistent staining in tissue, however overcompensates in areas that do not require normalising, such as larger areas of mucin, lumen, tissue tearing or retraction artefact. Section 3.3.3.5 briefly explores the effect of colour normalisation on correcting staining issues, and highlights that normalisation is unable to correct for stains that are extremely weak, as there is not enough visual information to be corrected. Without appropriate correction of the stains, features based on colour and intensity will become less powerful image features for training classifiers.

By examining the intensity values of the two classes; tumour and stroma, there is too much overlap in the dataset to create a simple linear classifier. This is detailed in the exploratory analysis, in section 3.3. However, when plotting these intensity values in conjunction with the overall intensity of the slide (the global property), there is a more distinguishable distribution of the two classes (see Figure 102).

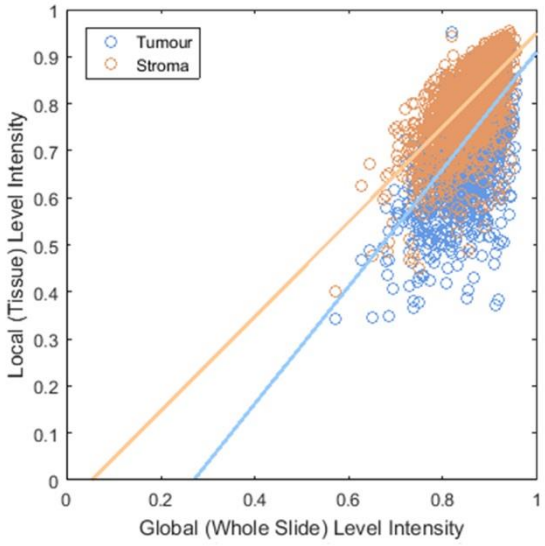


Figure 102 - Scatterplot of correlation between local and global intensity for tumour and stroma patches

Global intensity values are median HSV intensity of an 8x zoomed out view of the whole slide (excluding background pixels) - compared to the mean value of median intensity of tumour and stroma intensities for RandomSpot sampled locations within that slide. The least-squares line of best fit is calculated for both groups.

Figure 102 plots the average (mean) median intensity of tumour and stroma image patches (y axis) against the median intensity of their parent whole slide (x axis) at low power (scaled to 3x objective zoom). This indicates that the visual information derived from local features can be enhanced when combined with global features. It was proposed that using visual features of global contextual information in conjunction with the local image features will give a reliable comparative measure, with respect to colour variation.

Five global features were calculated to make a small context vector, similar to Magee et al's work discussed in 2.5.2. However, the features calculated were all median values (as the distribution of intensities was not expected to be normally distributed), so that they could be used to modify existing feature values from the vector described in 3.4.2.1. The five values were:

- Median hue
- Median saturation
- Median intensity
- Median haematoxylin staining channel intensity
- Median eosin staining channel intensity

Median haematoxylin and eosin values were calculated after applying a simple threshold to

remove background pixels (mean intensity minus half of one standard deviation intensity). These global features were calculated once per slide, prior to image patch analysis. Global values were subtracted from their respective feature vector values for each local image patch in order to create the feature set for training and testing. This resulted in feature values that ranged from -1 to 1.

As with Algorithm A, the image data from the QUASAR clinical trial dataset was used to train and test the algorithm (see section 3.4.2).

5.3.2.5 Algorithm H: Combined contextual analysis using all algorithm modifications

Algorithm H uses the methodology from Algorithm G to calculate global slide characteristics prior to local patch feature calculation. Again, these global characteristics (whole slide) were based on image intensities, so that overall staining levels could be ascertained, and used to explain how these levels affected the appearance of local tissue at native magnification. The local contextual analysis (per-patch) used the hybrid segmentation method of superpixel clustering and normalised cuts to group superpixels based on a distance-weighted similarity metric (see section 5.2). Once image patches were segmented, the pre-trained RF classifier was applied to each of the centre segment's neighbours, so that the class votes generated could be combined using a weighting based on the length of the segment boundary that shared the perimeter with the centre segment (see section 5.3.2.3). As with Algorithm F, these weighted votes were concatenated with the 18 features (see section 3.4.2) for the centre segment.

5.3.3 Results

All results presented in this chapter are summarised in this section. Individual figures for all algorithm results can be found in Appendix D.

5.3.3.1 Processing time

Processing time was calculated for generating the image feature set for all patches, for all algorithms. The processing time is indicative of the algorithms complexity, but it is important to note that each algorithm was run using the same hardware and software configuration (detailed in section 3.4.2.4), and the task was parallelised across eight logical processors. Training and testing was computed independently of the feature generation processing, and the training times are recorded as a mean value of the time (in seconds) taken to train each of the ten folds of data for cross validation.

Algorithm		Feature generation (s)	Training per fold (s)
A	Fixed patch size (64x64)	6,314	159
B	Fixed patch size (256x256)	13,765	200
C	Regular segmentation predictions	214,189	157
D	Regular segmentation votes	214,306	185
E	Unsupervised segmentation	242,090	191
F	Local context	295,829	184
G	Global context	250,499	185
H	Combined context	303,764	190

Table 17 - Processing times for all algorithms

The table shows the algorithms processing times for feature generation, and classifier training (per cross validation fold). The feature generation times show that processing time required increases with the complexity of the algorithms. Training times remain similar due to the size of each feature vector falling between 18 and 50 features. Algorithm B shows the highest training time, which may be attributed to the large patch size without context - containing more overlap and ambiguity in the feature space, requiring the RF to make more splits before convergence.

It should be noted that for all algorithms except Algorithm A (fixed patch size of 64x64 pixels), the image patch size processed was 256x256 pixels, following the conclusions of the observer experiments in Chapter 4.

Table 17 shows that there is increase in feature generation processing time, in relation to the complexity of the algorithm. The training times show less variance as the feature vectors for each algorithm are similar sizes. Time taken to train the RF algorithm is higher in Algorithm B, despite the feature vectors being smaller than Algorithms C and D. This may be due to the fact that the feature vectors are harder to separate, given that they contain statistics derived from visual information of larger images, likely to contain multiple tissue classes.

5.3.3.2 Algorithm agreement

Algorithm agreement is calculated on a per-image patch basis, comparing pathologist ground truth labels to algorithm predictions, and is explained in 3.4.3. Table 18 lists the agreement (also referred to as accuracy) for two the methods of evaluation. The first is the mean accuracy result of the ten-fold cross validation applied to the image patches from the full 2,211 cases, using all eight tissue classes (per-class agreement). The second uses the same accuracy results, and groups the tissue classes into their parent class, either tumour or stroma. The tumour and stroma parent classes are used to generate the prognostic marker, the TSR.

Algorithm	Per-patch accuracy	Kappa	Mean AUC	Grouped accuracy	Sensitivity TPR	Specificity TNR	Kappa
A	0.7006	0.51	0.83	0.7766	0.82	0.69	0.51
B	0.6399	0.39	0.81	0.7605	0.85	0.60	0.46
C	0.7037	0.51	0.86	0.7972	0.87	0.66	0.55
D	0.7032	0.51	0.86	0.7988	0.88	0.66	0.55
E	0.6916	0.50	0.85	0.7674	0.81	0.68	0.49
F	0.7172	0.54	0.88	0.7936	0.85	0.70	0.55
G	0.7115	0.53	0.87	0.7773	0.82	0.70	0.52
H	0.7285	0.56	0.89	0.8048	0.85	0.72	0.57

Table 18 – Agreement statistics for all algorithms

The table displays mean accuracies for the ten-fold cross validation performed on all the reported algorithms. Per-class agreement represents algorithm-pathologist agreement for all eight of the tissue classes, and grouped agreement combines the eight classes into the tumour or stroma parent class. The statistics show that incremental improvements translate to small increases in accuracy in the algorithm predictions, with the combined context algorithm yielding the highest accuracy levels.

The accuracy of the algorithms is reported to four decimal places due to the results falling within a narrow percentage bracket. Table 18 shows that Algorithm H exhibits the highest agreement for both per-patch comparisons and grouped patch comparisons, as well as having the highest mean AUC and kappa value. This supports the initial hypothesis drawn from previous conclusions that contextual information is required in order to improve algorithm accuracy. It is unsurprising that the progression of algorithm development improves algorithm accuracy, given that the modifications made were based on conclusions from previous iterations. Comparing effects of the individual improvements on accuracy shows that local contextual analysis (Algorithm F) has a higher impact then global contextual analysis (Algorithm G), which is expected, given that accuracy is assessed on local patch-level classifications. Combining the two levels of contextual analysis improves the algorithm further over either of the methods being used independently.

All algorithms have higher sensitivity than specificity, which means that tumour is overrepresented by the classifier. This may be due to the distribution of the training data classes (more tumour than stroma – see 3.3.3), but may also be due to the more variable appearance of stroma compared to tumour. In the context of medical image analysis, false positives are preferred over false negatives.

Table 19 lists all receiver-operator characteristic (ROC) AUC statistics for each tissue class using each algorithm.

Algorithm	Tum.	Lum.	Muc.	Nec.	Str.	Ves.	Mus.	Inf.	Mean
A	0.86	0.92	0.84	0.81	0.84	0.76	0.84	0.74	**0.83**
B	0.80	0.83	0.83	0.74	0.82	0.72	0.88	0.88	**0.81**
C	0.86	0.91	0.90	0.81	0.85	0.77	0.89	0.91	**0.86**
D	0.86	0.91	0.90	0.81	0.85	0.77	0.89	0.91	**0.86**
E	0.86	0.94	0.89	0.77	0.83	0.76	0.82	0.91	**0.85**
F	0.88	0.95	0.93	0.82	0.86	0.80	0.89	0.94	**0.88**
G	0.88	0.93	0.90	0.81	0.84	0.79	0.89	0.89	**0.88**
H	0.89	0.94	0.92	0.83	0.87	0.81	0.90	0.95	**0.89**

Table 19 – AUC for pathologist-algorithm agreement for all tissue types

AUC is displayed for all tissue subclasses, for each algorithm presented. The table shows a steady progression of increasing AUC for all subclasses as algorithm development iterations are made.

Algorithm H has the highest AUC for all tissue classes, with the exception of lumen and mucin in Algorithm F. This is likely due to the compensation for stain variation in Algorithm H, meaning Algorithm F will have more feature values relating to weakly stained tissue, and subsequently may be more sensitive to such values. Algorithm H shows highest AUC on inflammatory cells, which is consistently high across all algorithms since the introduction of local context. This may be due to contextual inferences made by the algorithm that inflammatory cells are likely to be surrounded by stroma, in spite of the comparatively low number of examples in the dataset for training and testing.

All context-based algorithms perform consistently well on lumen patches, which is due to the relatively simple appearance in terms of texture and colour. Also, AUC for inflammation is consistently high for all context-based algorithms, which may be due to the contextual information (i.e. inflammation is found in stroma) providing more accurate description of the class in the feature vector. This theory is supported by the non-contextual algorithms performing more poorly on inflammation. Given that blood vessels have a distinct bright red appearance (when filled with blood cells) it is perhaps surprising that performance on vessel images is consistently poorest. This may be explained by both the comparatively low frequency of examples, but also that blood vessels may or may not have red blood cells within them, and so empty vessels may be classified as lumen.

Finally, the algorithm performs slightly better on tumour patches than stroma ones. This may be due to the less consistent appearance of stroma, which may be sparse or densely packed with nuclei such as fibroblasts, depending on the proximity to the epithelial cells.

5.3.3.3 Per-class agreement

Each algorithm was processed using ten-fold cross validation. The ten subsequent accuracy results from each fold (per algorithm) are plotted in the boxplots in Figure 103. The mean accuracies of each algorithm are presented previously in Table 18.

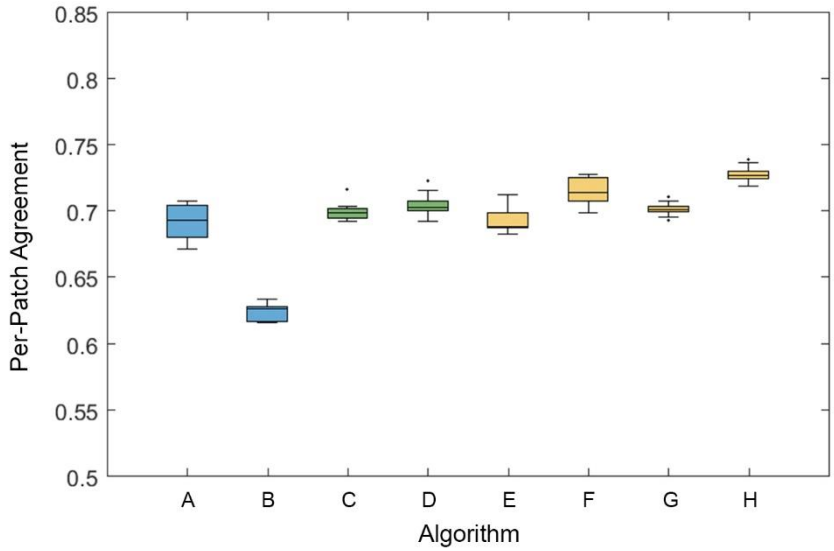


Figure 103 – Boxplots showing all algorithm accuracies per tissue class

The box plots show the ten accuracy results obtained by using 10-fold cross validation on all eight algorithms, analysing the eight tissue classes individually (per-class agreement). The mean accuracies of each algorithm are presented in Table 18. The figure shows that algorithm accuracy improves with the addition of contextual analysis, and peaks when combining both local and global context.

Algorithm H exhibits the highest mean accuracy (0.73), and the global context algorithm (Algorithm G) has the lowest standard deviation (< 0.01). The progression of algorithm development shows that the implementations of contextual analysis improve per-class agreement individually, with the local contextual analysis exhibiting higher levels of improvement compared to global contextual analysis. Combining both local and global contextual information improves the results further. The graph also highlights that without contextual analysis, using an image patch size of 256x256 pixels negatively affects performance to a significant degree (Algorithm B - $p < 0.001$ for all comparisons), which supports the previous conclusion that image patch size affects the amount of contextual information in the image, which has the capacity to negatively affect the feature vector's representation of a single class.

For algorithm accuracy results, a Friedman's (repeated measures) ANOVA was conducted with algorithm type (total = 8) as the repeated measures independent variable. Results revealed a significant effect of algorithm methodology ($\chi 2 = 51.50$, $p < 0.01$). Post-hoc analysis of pairwise comparisons using the Mann-Whitney test are presented in a cropped confusion matrix in Figure 104.

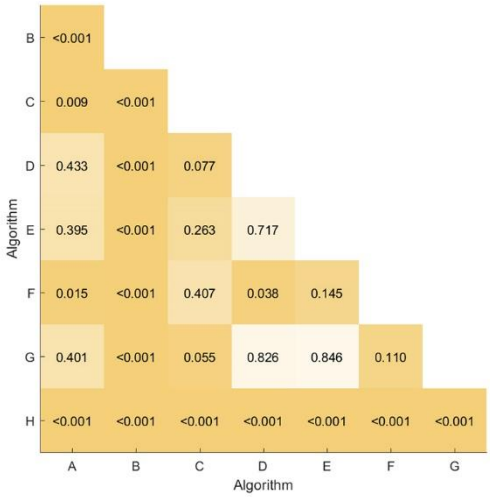


Figure 104 – Mann-Whitney test significance values for all algorithm accuracy pairwise comparisons

The matrix shows pairwise comparisons between algorithm accuracy results. Comparisons with a significance value less than the cut-off are considered significantly different to each other. Algorithm B and H are significantly different from all other algorithms. Note that Bonferroni correction changes the significance cut-off value from 0.05 to 0.063, and subsequently, significance is reported to three decimal places.

Note that Bonferroni correction changes the significance cut-off value from 0.05 to 0.063, due to the eight groups being compared. By accepting the null hypothesis (significance values above 0.063), the test indicates that the difference between the two sets of results has a mean difference of 0. Rejecting the null hypothesis indicates that the distributions of results are significantly different from one another. The matrix shows that the results of both algorithms B and H are significantly different to all other algorithms. Algorithm B is significantly worse than all others, due to processing of the larger patch size without contextual analysis. Algorithm H combines local and global contextual analysis, and exhibits significantly higher levels of agreement than all other algorithms.

5.3.3.4 Grouped agreement

The ten accuracy results from each cross-validation fold (per algorithm) are plotted in the boxplots in Figure 105. The mean accuracies of each algorithm are presented in Table 18.

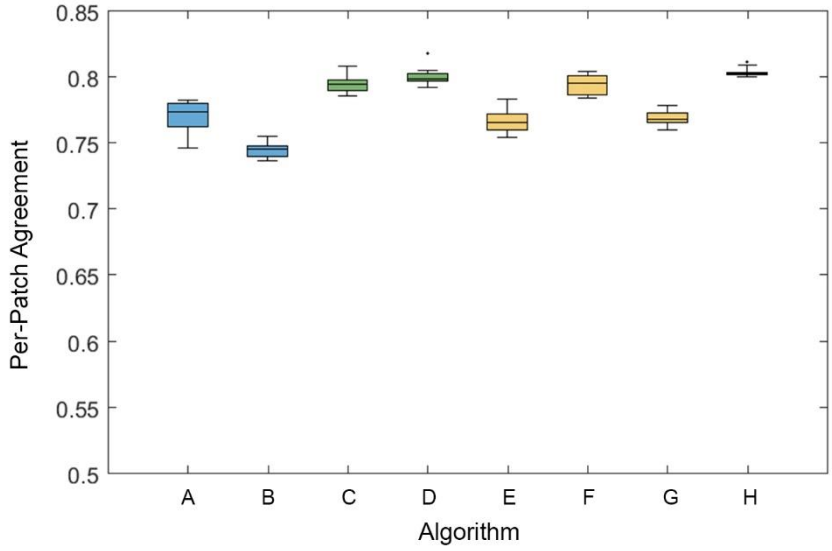


Figure 105 – Boxplots showing all algorithm accuracies grouped by tumour or stroma

The box plots show the ten accuracy results obtained by using 10-fold cross validation on all eight algorithms, and grouping the eight tissue classes into the two parent tissue classes, tumour and stroma (grouped agreement). The figure shows that algorithm accuracy increases as algorithm development iterations are made. However, the static segmentation Algorithms C and D have higher commparative values than when comparing individual class agreement. The mean accuracies of each algorithm are presented in Table 18.

The overall relationship of accuracy to algorithm development is similar to the per-class agreement results shown in Figure 103, with the exception of Algorithms C and D being visibly higher in relation to the other algorithms. This may be due to both algorithms having larger (albeit rigid) image segments, which have the potential to contain multiple tissue classes, and so their feature vectors have the capacity to model image segments containing two or more classes. The rigidity of the segments could mean that these averaged feature vector values are useful for determining contextual information when grouped by their parent classes. However, it is more likely that this is an artefact of grouping incorrect classifications into a correct parent class. Algorithm H exhibits the highest mean accuracy (0.80) as well as the lowest standard deviation (< 0.01), indicating that unsupervised segmentation is the better method.

For algorithm accuracy results, a Friedman's (repeated measures) ANOVA was conducted with algorithm type (total = 8) as the repeated measures independent variable. Results revealed a significant effect of algorithm methodology ($\chi 2 = 61.07$, $p < 0.01$). As with the per-class agreement results, post-hoc analysis of pairwise comparisons using the Mann-Whitney test are presented in a cropped confusion matrix, in Figure 106.

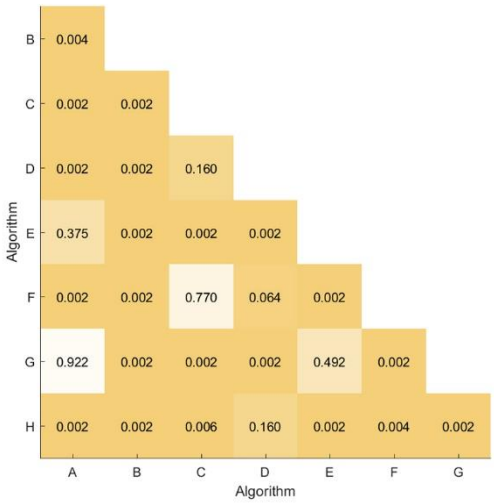


Figure 106 – Mann-Whitney test significance values for all algorithm grouped accuracy pairwise comparisons

As with Figure 104, the matrix shows pairwise comparisons between all algorithms using their grouped accuracy statistics. In this case, only Algorithm B is significantly different to all other algorithms, and the combined context algorithm shows no significant difference over Algorithm D. It is believed that this is due to an artefact of the grouping process. Note that Bonferroni correction changes the significance cut-off value from 0.05 to 0.063, and subsequently, significance is reported to three decimal places.

Note that Bonferroni correction changes the significance cut-off value from 0.05 to 0.063, due to the eight groups being compared. The matrix shows that only the results of Algorithm B are significantly different to all other algorithms. Algorithm H (combined context) is not significantly different to Algorithm D (regular segmentation using votes). However, given that the algorithms are significantly different when analysing the individual tissue classes, the accuracy of Algorithm C and D may be an artefact of grouping classes which were originally misclassified.

5.3.3.5 Correlation of TSRs

The TSRs are calculated per case (total 2,211), using the ratio of the number of all patches that are classified with one of the four tumour subtypes, to the total number of patches for the case. The generated ratios are subtracted from the pathologist-generated ratios, so that the mean difference and standard deviation of the distribution can be calculated. Table 20 displays these values for all eight algorithms. Note that Algorithms B and C were not reported previously, as they were minor variations of Algorithms A and D, which were more accurate per-patch.

Algorithm	Mean bias	Median	SD	R^2	C.I.
A	0	0	0.19	0.27	±0.37
B	-0.03	-0.06	0.20	0.28	±0.39
C	-0.04	-0.05	0.16	0.43	±0.31
D	-0.04	-0.06	0.15	0.45	±0.30
E	0.01	0	0.17	0.29	±0.33
F	-0.01	-0.02	0.16	0.37	±0.31
G	0.01	0	0.17	0.27	±0.33
H	0	0.02	0.15	0.40	±0.29

Table 20 – All algorithm TSR difference results

The mean difference and standard deviation relate to the distribution of the ratio generated by the pathologist manual scoring, minus the algorithm generated ratio. This means that a mean difference lower than 0 indicates the algorithm over-estimates tumour. The table shows that mean bias shows very little change over the algorithms, but the spread of data decreases, meaning that the TSR distribution is more closely aligned with human scoring methods.

Algorithm H exhibits the most favourable results with a mean difference of 0 and the lowest standard deviation (0.15). As with the reported accuracy and AUC, the algorithm development iterations improve the results when contextual analysis is added, and peak with the combined local and global context algorithm. Interestingly, Algorithm D shows the highest correlation to ground truth ratios, with an R-squared of 0.45. However, TSR is calculated on grouped accuracy, so algorithm D's better correlation may be an artefact of the grouping process - especially given that the algorithm produces results with the furthest mean bias from zero. The R-squared correlation value for Algorithm H is 0.40, which when compared to Hamilton et al's previously published correlation of 0.97 (see section 2.5.5), does not appear to be state of the art. However, the number of slides used in the publication totals 10, compared to the 2,211 slides of the QUASR trial, and so it would be interesting to compare the two algorithms on the same dataset to see if that difference changes.

Most of the algorithms have a mean bias below 0, meaning that they over-estimate the presence of tumour in a given slide. This is consistent with the specificity and sensitivity observed in the previous tables and figures. The mean bias of all algorithms appears appropriate for replicating human scoring, but the standard deviation of the distribution should be minimised also. Algorithm H has the lowest standard deviation of 0.15, indicating that the algorithm is the most capable of producing consistently well correlated TSRs, but further work should be done in attempt to reduce this spread, and increase the algorithm performance further. Examining cases

where the algorithm performs poorly is the focus of section 6.2, with the intent of minimising the standard deviation and confidence intervals reported in this section.

Figure 107 displays boxplot distributions of the ratio differences per algorithm, per case. This is a visual representation of the data displayed in Table 20. The declining spread of the data on the context-based algorithms is visualised by the narrower boxplots.

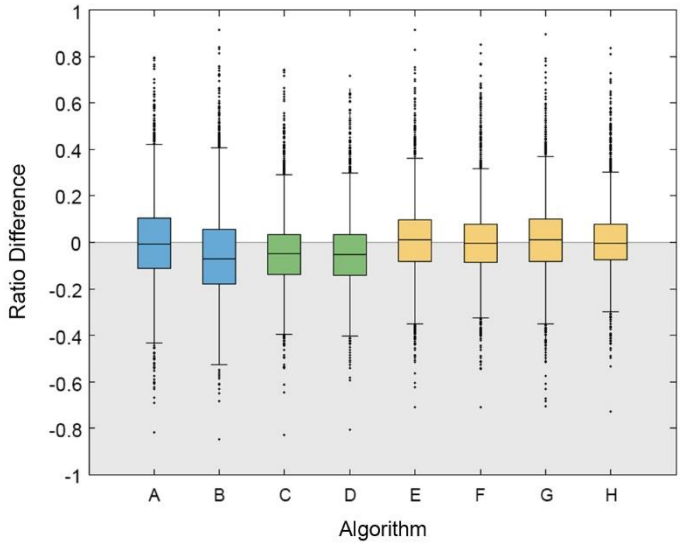


Figure 107 – Boxplots showing distribution of mean difference between pathologist and algorithm-generated ratios for all algorithms

The box plots visualise the distributions described in Table 20. Mean bias of all algorithms is close to zero, and the spread of data gets smaller with each algorithm iteration, indicating that the algorithm-generated TSRs become more similar to pathologist scoring with each algorithm improvement.

Note that the ratio difference is calculated by subtracting the algorithm generated ratio from the pathologist generated ratio, and so a mean bias below zero (the grey area on the plot) indicates that the algorithm overestimates the presence of tumour, and a mean bias above zero indicates that the algorithm underestimates the presence of tumour. Both Figure 103 and Figure 105 illustrate that the incremental improvements in accuracy affect the distribution of TSR differences, making the spread (standard deviation and confidence intervals) smaller, and the bias (mean and medians) closer to zero. The distributions of these ratio differences are made clearer in the Band-Altman plots for each individual algorithm, in Appendix D.1 to D.8.

5.3.4 Conclusions

The algorithm development presented in this chapter shows gradual and incremental improvements in accuracy, when compared to Algorithm A. These incremental improvements

support the conclusions from the work in Chapters 3 and 4 that were used to develop the algorithm further.

Regular segmentation as presented in Algorithms C and D does not account for the variable appearance of CRC, specifically in relation to the shape of class boundaries. Unsupervised segmentation yields higher agreement levels and therefore is more appropriate to the task of CRC analysis.

Local contextual analysis improves algorithm-pathologist agreement, therefore is an important factor when aligning the algorithm methodology to the pathologist visual scoring task.

Accounting for stain variation by assessing global contextual features improves accuracy, but not to the same degree as local contextual analysis. Combining the two improves accuracy over all other algorithms.

The final algorithm combines methodologies developed from the conclusions of previous observer experiments in Chapter 4, and shows an overall improvement in accuracy, agreement and correlation with the ground truth data.

The image dataset (QUASAR) used in this is work highly valuable for its associated survival data. However, the dataset itself contains suboptimal images (see 3.3), which adversely affect image analysis in terms of generating accurate feature vectors, and the subsequent training of classifiers and making class predictions. Analysing the quality of this dataset in detail has the potential to improve the image analysis performance. This drives the research in Chapter 6.

5.4 Discussion

The work in this chapter is based on the conclusions drawn from Chapter 3 and Chapter 4, which identified a need for contextual analysis surrounding point co-ordinates from SRS sampling, where the gold standard classification is applied.

The results from processing Algorithm A at multiple image sizes led to conclude that larger image patches contained multiple tissue classes and therefore distorted the features calculated for training the classifier. The human interaction experiments highlighted that Algorithm A's lack of contextual information processing was one of the key differences when comparing to pathologist visual assessment. In order to incorporate the surrounding contextual information into the algorithm, a larger image patch size was selected, based on the minimum size that maximised pathologist agreement when scoring with a restricted field of view. A minimum patch size of 256x256 pixels was identified as appropriate for pathologist scoring, due to there being no significant differences between pathologist agreement on images 256x256 pixels in size or larger. To maximise algorithm efficiency, this size was implemented in all iterations of development presented in this chapter (as opposed to using larger patch sizes).

Initially, a contextual analysis algorithm was developed using regular segments (Algorithms C and D), dividing the patch into five equally sized (but not shaped) segments. This preserved a single segment that overlapped the centre of the patch, where the original ground truth classification was applied. A classifier was trained on the features generated from the centre segment of each patch in the dataset, so that the classifier could be applied to the surrounding segments, and generate predictions and votes for candidate classes. The surrounding information was incorporated into the centre segment's feature vector in two separate ways – using the classifier predictions and a confidence metric, comprised of a percentage of the votes given to that tissue class (Algorithm C), and using the number of votes for each class for each segment (Algorithm D). Both methods took a similar amount of time to generate the feature vector (see section 5.3.3), and the votes-based method took slightly longer to train using the RF classifier, this was due to the larger feature vector (containing the number of votes of each of the eight tissue classes, for each of the four neighbouring image segments). Algorithm D yielded a slightly higher level of agreement, which was not statistically significant (see 5.3.3), this is arguably due to Algorithm C disregarding some of the contextual information. By disregarding

all of the votes and using the number of the predicted class votes as a percentage, there was no information regarding the other candidate classes. To clarify, the votes-based method (Algorithm D) contained all the votes for each of the eight subclasses, which meant that information was preserved. For example, if a segment is predicted to be tumour with less than half of the votes, the distribution of the other votes may be important to the final classification. Another reason could simply be that the votes-based algorithm had a larger feature vector, and so less likely to underfit the data (the predictions-based algorithm had 26 features as opposed to 50).The TSRs generated by the algorithm slightly overestimated tumour compared to the pathologist (see 5.3.3), which is unsurprising, given the algorithm's relatively low specificity (true negative rate) of 0.66, compared to the sensitivity of 0.88. The results of the algorithm led to the conclusion that the rigid segmentation techniques did not adequately solve the issue of image segments / patches containing multiple tissue classes, and therefore creating ambiguity in the feature set. This led to an experiment to identify an appropriate unsupervised segmentation method.

Unsupervised segmentation of colorectal cancer tissue is non-trivial due to the many biological, histological and technical considerations that affect the variation in appearance of tissue on digital slides. This variation affects the appearance of slides in terms of colour, texture and shape, and so model based segmentation was not considered appropriate for this task. Several unsupervised segmentation methods were evaluated against expert-labelled ground truth data, and a hybrid method of superpixel generation and normalised cuts clustering of superpixels was chosen. This was due to the prohibitive amount of processing time required by the normalised cuts algorithm, when generating the per-pixel pairwise feature values to construct the affinity matrix. Unsupervised segmentation was applied by initially discretising the pixel data of each image patch into a much smaller set of superpixels, which approximately represented the size of one nucleus. This resulted in a graph cut problem which was less computationally expensive than calculating similarities per-pixel, whilst maintaining segment boundaries at a cell level. The inclusion of unsupervised segmentation meant that training images were less likely to include more than one tissue class per segment, thus reducing ambiguity in the feature set.

Initially the unsupervised segmentation was applied to the dataset, and the centre segment was used to calculate the image feature set. This was trained and tested without any contextual analysis, and in this respect, the methodology was identical to Algorithm A. Algorithm E was extended to incorporate local contextual information, in a similar method to Algorithm D. However, Algorithm E did not generate a fixed number of neighbours to the centre segment of the image patch, and therefore neighbour perimeter length-based weighting was applied to the votes per class, and a summation of these weighted votes was appended to the feature vector of

the centre segment. This improved the algorithm accuracy compared to previous algorithms to a statistically significant degree. This improvement (specifically over Algorithm D) is due to the unsupervised segmentation, allowing each segment to more accurately represent one single class, and therefore create more precise feature vectors.

By applying segmentation to the image patches, and a trained classifier to the surrounding segments, the predicted classes and votes provide contextual information which was used to weight the contextual features for the centre segment. In previous work (Chapter 4), this local contextual information has been shown to be important to pathologists for scoring images, and therefore should be a consideration when developing image analysis solutions to automate the pathologist task. Currently the algorithm design uses the assumption that higher amounts of surrounding tissue increase the likelihood of that same classification, and is a naïve assumption that could be improved in future work. This could be extended to include important biological structural rules, for example, lumen must be surrounded by tumour, or that a candidate lumen segment lying between stroma and tumour is more likely to be a retraction artefact. By incorporating these rules, the algorithm may model human behaviour more closely, which is likely to improve agreement levels further.

A separate algorithm was also developed so that contextual information was also applied at the macro level, in order to provide information about the slide as a whole. This global information is important when considering sets of data that have been stained inconsistently, or in this case, as part of a five-year longitudinal study, where the fading of stains before digital slide scanning was inevitable. By combining the features of the slide with the feature set for each image patch, the algorithm adapts the feature set to maintain the visual relationship between tumour and stroma, based upon such global features. The algorithm results showed a statistically significant increase in accuracy over Algorithm E without context, but did not improve on Algorithm F (local contextual analysis). This suggests that the global context is not as important to the patch-level classification as local context. This work could be extended to incorporate a prediction of the type of cancer being analysed, which could help improve classification accuracy. For example, mucinous adenocarcinomas are considered mostly tumour, but the light appearance of the mucin is likely to be classified as weakly stained stroma at the microscopic level.

Finally, both levels of context (global and local) were combined into one algorithm (Algorithm H), so that the global and local contextual information could be used to improve the classification accuracy. Algorithm H yielded the highest per-patch agreement with the ground truth, both on individual tissue classes and when grouping into the two parent classes, tumour and stroma. The incorporation of context was identified as important in Chapters 3 and 4, and by incrementally adding in these modifications to the algorithm, it is evident that the results

support these observations. Analysis of significance values shows that Algorithm H accuracy is statistically different to all algorithms, except Algorithm D, but only when analysing grouped accuracy, rather than per-patch accuracy. It is assumed that the increased accuracy of Algorithm D may be an artefact of grouping the predictions into the parent classes, based on the observations of the confusion matrices in Appendix D.4, where a high number of errors are reported as subclasses of tumour, being classified as other tumour subclasses. Algorithm H also yielded the most reliable TSRs, with a mean difference (to the pathologist generated ratios) of 0, and the smallest standard deviation (0.15). The increase in accuracy per-patch combined with the improved TSR correlation shows that the patch-level accuracy gains are more likely to be consistent across the whole dataset, as opposed to correcting for issues in relatively few slides. The processing time of each algorithm was recorded, and the time taken for feature set generation varied significantly, depending on the algorithm. All context-based algorithms took over 200,000 seconds, with Algorithm H taking over 300,000 seconds, due to the combination of both local and contextual analysis. Algorithm A took 6,314 seconds, processing 64x64 pixel patches. All algorithms used the same hardware and software configuration, utilising parallel processing across eight logical processors. With the task already parallelised, the time taken to compute is less significant, given that the number of processors could be increased, depending on the hardware used. The feature generation was for over 100,000 images, and arguably, in a practical setting, this would only need to be done for training – then application of the trained classifier would be done on a per slide basis.

The incremental changes to the algorithm presented in this chapter show gradual, and relatively small improvements. The quality of the dataset is referred to in Chapter 3, and is a major contributing factor to issues caused by ambiguity in the feature set. Given that the dataset is so large, the variability accounted for in Algorithm G may not be sufficient, and it would be beneficial to inspect where algorithm performance is poorest. This could be achieved by looking at the cases that had the largest mean differences in TSRs, and quantifying the issues present (if any). Accuracy and agreement may also be improved if an image quality check was enforced, so that poor images could be rejected prior to analysis. This has the potential to improve the classifier, and subsequently the predictions made by it. This could be achieved by extending the pilot quality control study from section 4.4, and the work in Chapter 6 focuses on the issue of image quality, and how it affects analysis.

Finally, in order to assess the algorithms for clinical use, they must be applied to a clinical dataset where the correlation to survival data can be calculated, to establish whether the statistics generated correlate to survival, as a prognostic phenotypic marker.

Chapter 6 – Effects of quality control on algorithm accuracy

6.1 Introduction

6.1.1 Chapter Overview

The work in this chapter extends the previous work in Chapter 4, using the Prospector system to visually assess images for their suitability to the task of visual histopathological analysis. A Quality Controlled (QC) dataset is generated for all cases, which labels each case with its QC metric. The dataset is then used to assess the impact of image quality on algorithm performance, in terms of classification agreement, as well as the differences between pathologist and machine-generated TSR results. Finally, the dataset is used to try to develop an automated solution for identifying issues with virtual slides before they are processed, extending the work in 4.4. The chapter is divided into six sections:

1) The introduction, which reiterates the findings from the image quality pilot study in Chapter 4, and addresses the limitations of that study, proposing the need for higher levels of detail and more results.
2) Analysis of previous algorithm performance on the QUASAR dataset, for the identification of key issues that affect image quality, that commonly affect image analysis algorithms.
3) Utilisation of the identified image quality issues as QC categories, to apply quality control labelling to the whole image dataset, using the Prospector system.
4) The effect of image quality on Algorithm H (section 5.3.2.5), and how the algorithm performs on a dataset of slides that have been through QC screening for image analysis.
5) The development of an algorithm that has been trained to identify slides which may be likely to fail image quality control checks, using the dataset generated from 6.3.
6) Discussion of the work presented in this chapter and conclusions.

6.1.2 Image quality

The issue of image quality has been repeatedly addressed throughout this thesis. The importance of obtaining good quality images that are consistently sectioned, stained and scanned is highlighted in these sections, and methods such as colour normalisation and global contextual analysis have been employed to mitigate these issues. However, there are cases where image variability is too extreme for this level of adaptive analysis, and in most cases, images would not be confidently analysed by a pathologist. The issue of quality control in histopathological analysis extends along the entire pipeline of image generation, from tissue retrieval, to staining, to colour calibration of scanning. Histopathological analysis is a demanding and high throughput discipline, and it is difficult to maintain consistent standards within single laboratories, let alone cross-site collaborations that may have different reagents, staining compounds, protocols and practises. It is beyond to scope of this work to address laboratory processes, but the quality of images being passed to image analysis algorithms is an issue that must be addressed prior to processing.

6.1.3 Tissue stain features affecting manual scoring

Previous work detailed in section 4.4 briefly addresses the level of staining as a quality control factor in a simple pilot study. The study used pathologist agreement as an indicator of adequate quality, and whether certain levels of staining affect it. The study used 250 256x256 pixel sized patch images, and the conclusions drawn were that there was too little data to assess whether image intensity is a useful surrogate for image quality. This chapter details the work carried out to extend this pilot study.

6.2 Algorithm worst case performance

6.2.1 Aim

To evaluate the cases where algorithm agreement was lowest, in order to establish any recurring issues that affect image analysis results.

6.2.2 Methods

Using the algorithm performance from Algorithm H in section 5.3.2.5, the TSRs were generated per-case, and subtracted from the ratios generated by the pathologists, from the ground truth dataset. The resulting distribution of differences ranged from -1 to 1, with a difference of zero meaning that the algorithm and pathologist generated the same ratio for that case. Each of the ratio differences generated for the 2,211 cases were ordered by absolute difference, and the 100 cases with the highest absolute difference were manually inspected to identify any issues with the cases that may have adversely affected image analysis. Figure 108 highlights the data points for the 100 highest absolute differences in the dataset (when analysed by Algorithm H) in red on the Bland-Altman plot.

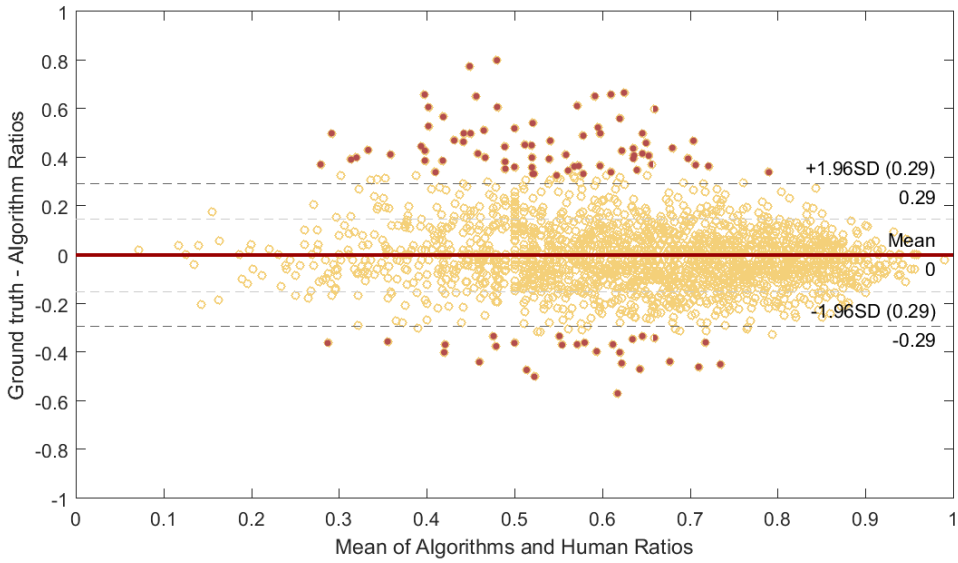


Figure 108 – Bland-Altman plot of Pathologist - Combined Context algorithm generated TSRs, highlighting the 100 cases with the highest absolute difference in red.

Cases were visually inspected by the author on a 30-inch Dell 3008WPF monitor using Leica-Aperio ImageScope virtual slide viewing software, in full screen mode at 2560 x 1600 resolution. This monitor was specifically chosen for the high dynamic contrast ratio (3000:1), so that differences in staining intensities were more easily observed. Answers were recorded manually in a spreadsheet, where the perceived level of both Haematoxylin and Eosin stains were given a score from zero to three (no visible stain, weak staining, acceptable staining, strong staining). An additional column for other observations was used for noting issues that would also affect analysis. This meant that multiple issues could be observed per case, and these observations were analysed for key themes and issues.

6.2.3 Results

Out of the observations made from the 100 cases, 12 recurring issues were identified (total 245 issues). Table 21 lists the issues with the frequency of their appearance in the identified subset of data.

Issue	Count
Weak haematoxylin	61
Weak eosin	47
Poorly differentiated tumour	38
Tissue tears	25
Weak overall staining	20
Debris	19
Tissue folds	14
Coverslip edges	11
Necrotic tissue	4
Sectioning	2
Bubbles	2
Mucinous tissue	2

Table 21 – Recurring quality issues identified in cases with poorest agreement

Issues manually identified on 100 cases exhibiting the highest absolute difference between pathologist and algorithm generated TSRs. Note that multiple issues could be reported per case (total = 245).

Of the issues identified, the top five relate to either some form of weak staining, or artefact that makes the average appearance of the slide lighter in appearance. This accounts for 78% of the issues observed.

Figure 109 shows cases displaying each of the issues identified in Table 21. Note that these are singular examples, and the visual characteristics of these issues can still vary from the images shown. Note that an explanation of these categories can be found in the description column of Table 22.

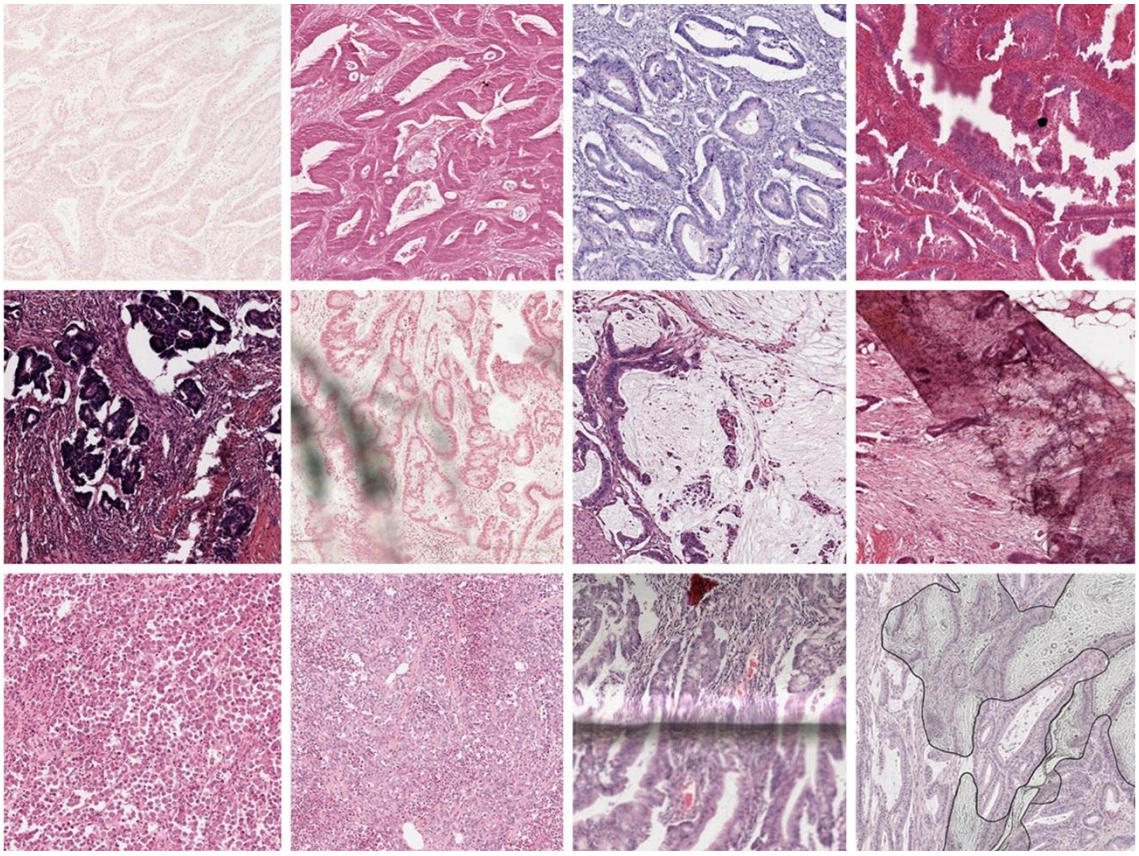


Figure 109 - Visual examples of the 12 QC categories identified from the top 100 worst correlated cases

In reading order from top left to bottom right: weak overall staining, weak haematoxylin staining, weak eosin staining, tissue tearing, tissue sectioning artefact (tissue is too thick), debris on slide, large areas of mucin, tissue folded over, poorly differentiated tumour, necrotic tissue, coverslip edge and bubbles on the slide.

By visually inspecting these images, it is clear that their appearances break some of the assumptions of the algorithm. For example, in using an unsupervised segmentation algorithm, it is assumed that there are existing structures and shapes which can be visually distinguished using their boundaries. This is not the case for poorly differentiated, necrotic and mucinous tissue. An early assumption made in 3.3 was that tumour is darker than stroma, and so quality issues that affect the brightness of images would also break this assumption, especially staining that is too strong or weak, and tissue folds. Tissue tears would likely be represented as lumen in the image, and debris, bubbles and coverslip edges may be interpreted as large areas of dark tumour.

6.2.4 Conclusions

There are several issues that appear to negatively affect the results generated by the automated analysis of CRC. Some of these issues are characteristic of the type of disease being analysed, but are not part of the scope of this work, and other issues are caused by slide preparation.

The 12 categories established from analysis of the cases exhibiting poorest algorithm performance are variable in appearance, and appear to affect the image analysis results in multiple ways (creating ratios that fall outside of the standard deviation either side of the mean). The majority of observations (78%) related to issues that increase the average intensity of the overall slide, such as staining levels. This relates to the issues discussed in 2.3.3, and observed in 3.4.3. Algorithm A initially attempted to mitigate similar staining issues using colour normalisation (2.5.2), but was replaced by global analysis of image features in Algorithm G, due to artefacts from small images with little colour information. The issues presented in these top 100 worst cases seem to be too extreme for the global analysis to compensate for.

Application of image analysis to a dataset free from these artefacts should improve analysis results, and assessing the full dataset for these issues would help create an optimal training set.

6.3 Quality control experiment

6.3.1 Aim

To use the QC issues observed in section 6.2 as scoring categories in a study utilising the Prospector system (presented in section 4.2), to fully assess the QUASAR dataset (presented in 3.3.2) for suitability for analysis in Algorithm H (presented in section 5.3.2.5).

6.3.2 Methods

The Prospector system (see section 4.2) was used to develop an experiment that required the user to visually assess each of the digital slide images from the 2,211 cases, and apply a categorical label, denoting whether the image was suitable for image processing, or had any visual defects which had the potential to affect algorithm training and/or testing. The categorical labels used in the experiment were derived from the QC issues identified in 6.2. A list of these categories and a description of the issues that they represent is presented in Table 22.

Category	Description
Accept	Image is suitable for analysis
Weak staining	Overall stain is visually too weak
Weak haematoxylin	Haematoxylin stain is visually too weak
Weak eosin	Eosin stain is visually too weak
Tissue tears	Tissue has tearing, affecting appearance
Sectioning	Sectioning artefacts present, such as score marks or the tissue is sliced too thickly
Debris	Debris either on tissue or on the coverslip is affecting appearance
Tissue folds	Tissue is folded
Mucinous tissue	Tissue contains large areas of mucin
Necrotic tissue	Tissue contains large areas of necrosis
Poorly differentiated tumour	Tissue has no discernible structure
Bubbles	Bubbles under the coverslip have formed
Coverslip edges	The coverslip edge overlaps the tissue

Table 22 – Quality Control categories used in the Prospector QC experiment

Categories were chosen based on visual inspection of the 100 worst cases in terms of algorithm agreement levels (section 6.2).

The experiment was run using one participant (the author), a trained technician who had six years' experience working with closely with pathologists on CRC digital slides. Initially, the participant was trained to identify these issues by a pathologist who completed the study on a test set of 250 cases. Where images had more than one of the QC issues, the issue that was perceived to have the highest detrimental effect to visual analysis was selected. Since the cases had no QC metrics to serve as ground truth, the experiment was set to 'collection' mode (as opposed to 'comparison').

The 2,211 case images were presented sequentially as a static (non-zoomable) jpeg image (compression quality 100), embedded into the system web page. Cases were visually inspected on a 30-inch Dell 3008WPF monitor running in an HTML5 compliant web browser (Google Chrome version 26.0.1410.64 m) in full screen mode at 2560 x 1600 resolution. This monitor was specifically chosen for the high dynamic contrast ratio (3000:1), so that differences in staining intensities were more easily observed. Answers were recorded automatically by the system in CSV format. Figure 110 shows the Prospector scoring window, with an example case image to be analysed.

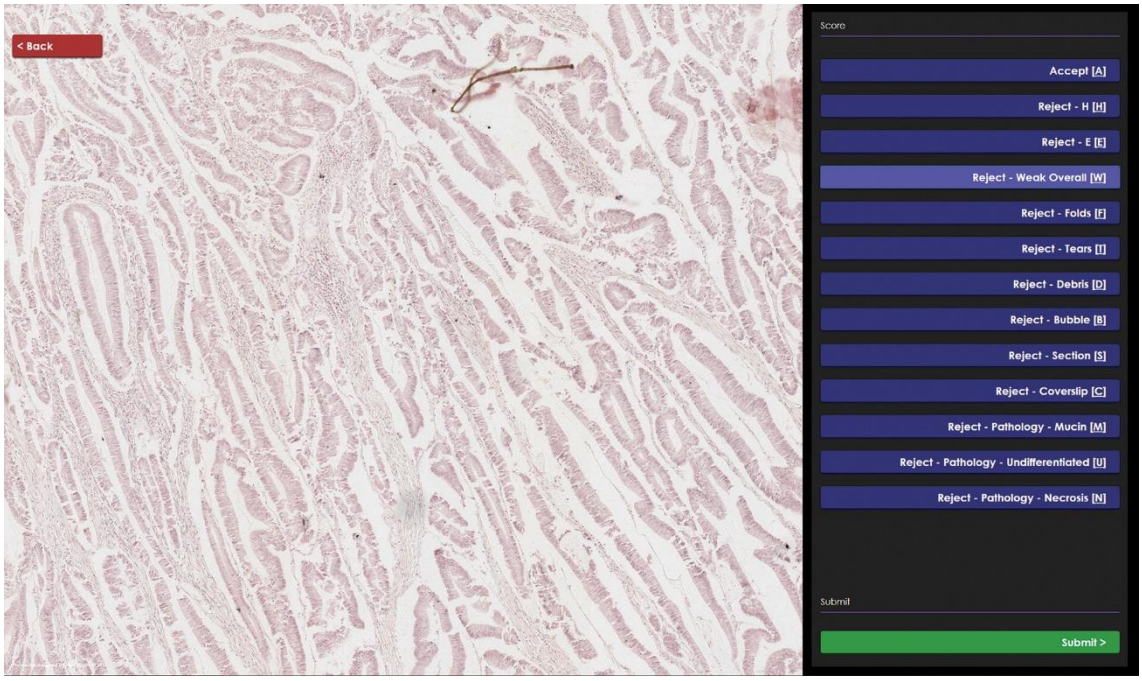


Figure 110 – Prospector scoring interface for QC experiment

Left: Viewing window showing a low-power static image of the digital slide to be assessed.
Right: The available qualitative scoring categories in the participant control panel.

The participant was presented with a briefing screen, explaining the purpose of the study, and a contrast calibration scale to ensure the monitor could adequately display the images. The participant was then required to visually inspect all images in the dataset and apply the single most appropriate classification to each case. Once all images had been scored, the participant was presented with a debrief screen, again, detailing the purpose of the study. Since this experiment required such a large volume of images to be visually inspected, ten-minute breaks were incorporated into the experiment for every 50 minutes of participation.

6.3.3 Results

The experiment took 10,856 seconds (181 minutes) to complete the inspection of all 2,211 cases, averaging a time of 4.91 seconds per image. This excluded the 10-minute breaks taken every 50 minutes. It was observed that the number of available classifications affected the complexity of the UI, which slowed the scoring process.

The results of the experiment were saved by the Prospector system and stored in CSV format. Each row in the list of results contained the case number, the participant name, a hyperlink to

the image and the scoring classification. The distribution of responses is displayed in Figure 111. Note the explanations of each category are listed in Table 22.

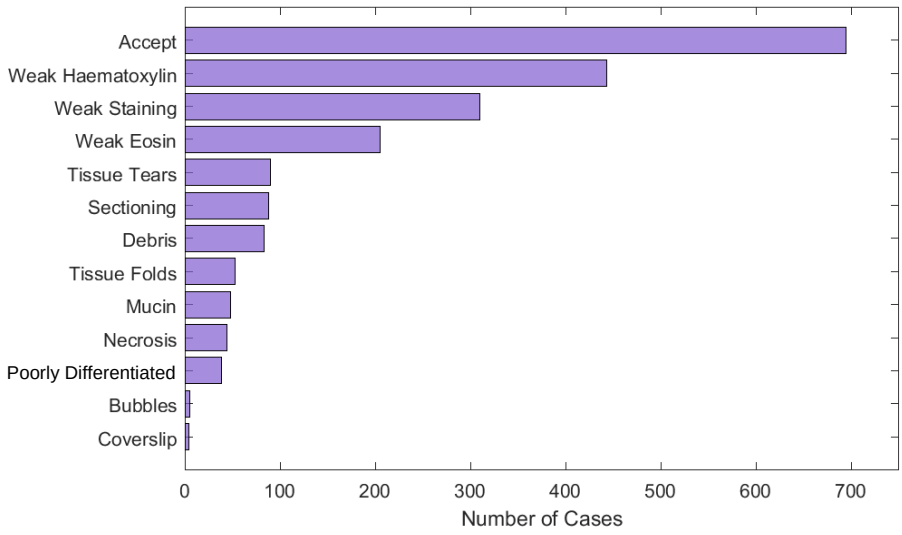


Figure 111 – Distribution of responses from QC experiment

The distribution of results, in order of frequency is as follows: Accept (701), weak haematoxylin (443), weak overall staining (310), weak eosin staining (205), tissue tears (90), sectioning artefacts (88), debris (83), tissue folds (52), large areas of mucin (48), large areas of necrosis (44), poorly differentiated tumour (38), bubbles over the tissue (5) and coverslip edges (4).
The distribution of QC classes shows that over half of the dataset is considered suboptimal in terms of visual representation of the tissue. This shows that the dataset itself may be causing the algorithm results to be lower than if they were applied to a set of virtual slides of higher visual quality.

The QC classification labels for each slide meant that further analysis of previously generated results could be undertaken. The TSR data generated by the previously presented algorithms was grouped by the QC label applied to each case, in order to visualise the effect of each issue on the automatic analysis. Figure 112 shows the TSR differences, between the pathologist scoring and Algorithm H (presented in 5.3.2.5). Note that a skew to the left of zero indicates the algorithm over estimates tumour compared to the pathologist, and to the right, the algorithm over estimates stroma.

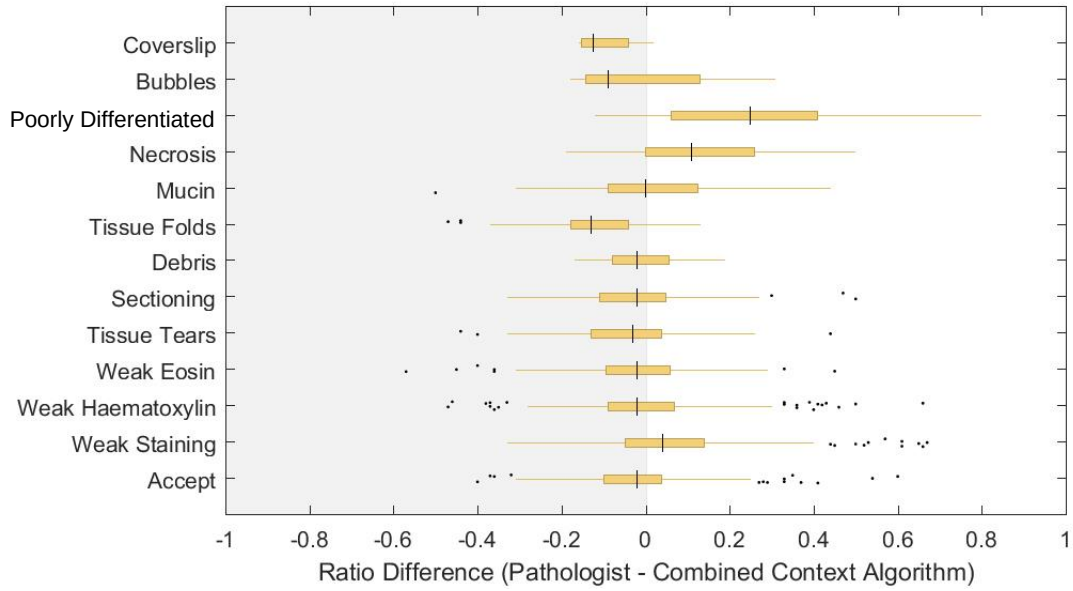


Figure 112 – Boxplot of TSR differences grouped by QC classification

The TSR differences are between pathologist and combined contextual algorithm H. The black line on the box plots shows the median difference for the distribution. The figure shows that different QC classifications affect the agreement of TSR between pathologist and algorithm to different degrees. Poorly differentiated tissue affects the TSR most of all, which may be due to the homogenous appearance being classified as stroma.

The TSRs are grouped by the QC label that was applied to the slide that they are generated from. The mean ratio difference for Algorithm H was 0, with a median difference of -0.02. The boxplots indicate that some of the QC metrics affect the TSR more than others (weak staining, tissue folds, necrosis, undifferentiated tissue, bubbles and coverslip artefacts). Table 23 displays the statistics for the distributions of TSR differences for each QC category.

QC Category	# Cases	Mean	Median	SD
Accept	701	-0.02	-0.02	0.11
Weak staining	310	0.06	0.04	0.16
Weak haematoxylin	443	-0.01	-0.02	0.14
Weak eosin	205	-0.02	-0.02	0.14
Tissue tears	90	-0.04	-0.03	0.14
Sectioning	88	-0.01	-0.02	0.14
Debris	83	-0.01	-0.02	0.10
Tissue folds	52	-0.13	-0.13	0.13
Mucin	48	0.01	0	0.17
Necrosis	44	0.14	0.11	0.19
Poorly differentiated tumour	38	0.25	0.25	0.24
Bubbles	5	0	-0.9	0.20
Coverslip edges	4	-0.10	-0.12	0.08

Table 23 – TSR difference distribution statistics for individual QC groups

The table presents the statistics from Figure 112, confirming that poorly differentiated tissue affects pathologist-algorithm TSR difference the most. Necrosis, tissue folds and coverslip edges all affect the TSR difference by 10% or more. Weak overall staining levels affect analysis more than weak haematoxylin or weak eosin individually. This may be because structural information is preserved if one of the two stains is adequate, allowing distinctions to be made between the tissue types.

6.3.4 Conclusions

Over half of the number of cases in the dataset used for developing the algorithms in previous chapters have at least one issue that is perceived to have a negative impact on image analysis.

Analysis of the algorithm performance on individual QC categories shows that the categories affect pathologist-algorithm agreement to varying degrees.

The weak haematoxylin and weak eosin categories have mean differences of -0.01 and -0.02 respectively, indicating that the global contextual analysis used in Algorithm H adequately compensates for stain variation.

Poorly differentiated tumour had the largest effect on TSR difference, which may be due to the lack of structural information, and similar visual characteristics of uniform distribution of homogeneous tissue.

Identification of QC issues prior to image analysis may be useful for improving accuracy results. Application of image analysis to the accepted dataset only should yield higher accuracy results.

6.4 Algorithm I: Effect of quality control on algorithm performance

6.4.1 Aim

Using only the accepted cases from the QC data (section 6.3), assess the impact of image quality on the performance of Algorithm H.

6.4.2 Methods

The 2,211 cases of the QUASAR dataset were reduced to the 701 cases accepted by the QC process in 6.3. The full dataset of cases was originally generated by Algorithm H presented in section 5.3.2.5. This dataset was filtered based on the case number associated with the QC metric. This left 33,736 image feature vectors for training and testing.

The RF classifier was trained and tested using the ten-fold cross validation method on the full set of remaining feature vectors, previously described in 3.4.2.

6.4.3 Results

6.4.3.1 Processing time

The image feature set was previously generated by the Combined Context alforithm in 5.3.2.5, and filtered based on case number. The RF model building (training) took an average of 44.46 seconds per cross validation fold on the smaller dataset.

6.4.3.2 Agreement

Agreement was calculated in two ways: calculating the agreement for all eight tissue classes, and grouping the eight classes into their parent class: tumour and stoma. Figure 113 shows the results for both of these methods in the confusion matrices.

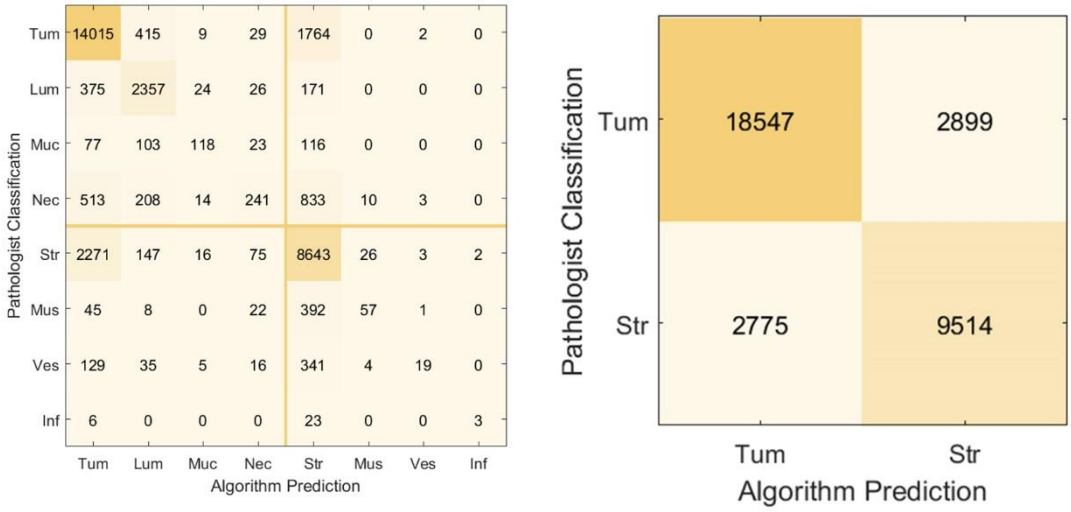


Figure 113 - Confusion matrices showing pathologist – Combined Context algorithm agreement on QC approved cases only

Left: Confusion matrix of all 8 tissue subtypes from dataset. Accuracy = 75.45%, sensitivity (true positive rate / recall) = 0.75, kappa = 0.60 (substantial agreement).
Right: Confusion matrix of subtypes grouped into Tumour and Stroma parent classes. Accuracy = 83.18%, sensitivity (true positive rate / recall) = 0.86, specificity (true negative rate) = 0.77, kappa = 0.64 (substantial agreement).
The distribution of classification errors is similar to those presented for Algorithm A in 3.4.3. The algorithm appears to have a balanced distribution of false positive and negatives, with a slight bias to overrepresenting stroma. Again, the highest proportion of errors is due to tumour being classified as stroma and vice versa, and necrosis being classified as stroma.

The accuracy of the algorithm for the individual tissue classes was 75.45%, with a sensitivity of 0.75 and a kappa of 0.6, indicating substantial agreement. Grouping the individual classes into the parent tumour and stroma classes yielded an 83.18% accuracy, a sensitivity of 0.86, a specificity of 0.77 and a kappa value of 0.64, indicating substantial agreement. The algorithm has a higher number of false negatives than false positives (2,899 and 2,775 respectively).

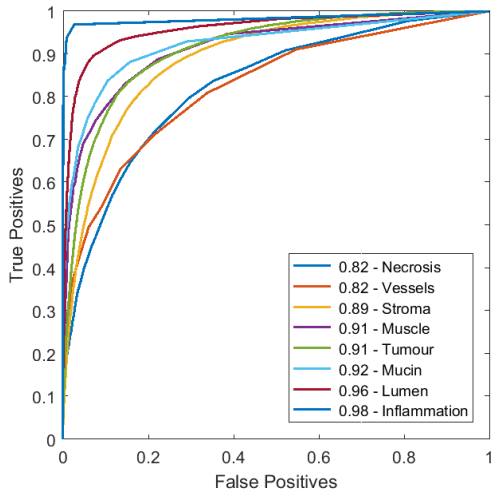


Figure 114 - ROC Curves for all 8 tissue subtypes, classified by Algorithm H on QC approved cases only

The graph shows Area Under the Curve for each tissue subtype:
Tumour parent class: Tumour (0.91), Lumen (0.96), Mucin (0.92), Necrosis (0.82)
Stroma parent class: Stroma (0.89), Vessels (0.82), Muscle (0.91), Inflammation (0.98)
The algorithm performs best on patches containing inflammation and least well on necrosis. However, there are only 32 instances of necrosis presented in the confusion matrix in Figure 113, most of which are classified as stroma. Necrosis is also more likely to be classified as stroma, due to the lighter, less dense appearance.

Figure 114 shows the ROC curves for each individual tissue class when classified by Algorithm H. The tumour subtypes have an AUC of 0.91 for tumour, 0.96 for lumen, 0.92 for mucin and 0.82 for necrosis. The stroma subtypes have an AUC of 0.89 for stroma, 0.82 for vessels, 0.91 for muscle and 0.98 for inflammation. The mean AUC for all categories is 0.90.

6.4.3.3 Comparison to manual scoring

For comparison, the algorithm grouped accuracy results are plotted against the pathologist agreement statistics generated on the best-case data from the image patch size experiment in Chapter 4.

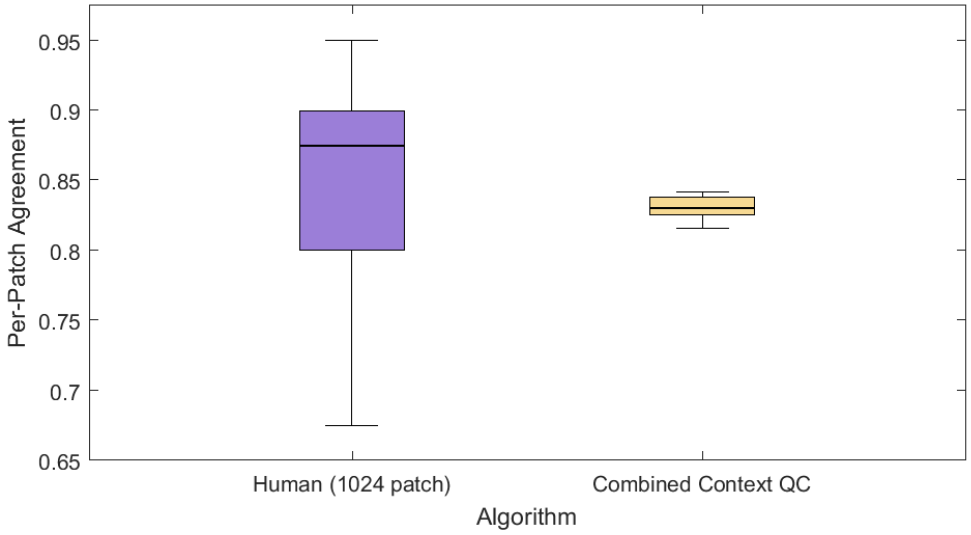


Figure 115 – Comparison boxplots for pathologist - pathologist agreement and pathologist -algorithm agreement

Left: Pathologist - pathologist agreement (mean = 0.85, median = 0.88., SD = 0.10, IQR = 0.10)
Right: Pathologist - algorithm agreement (mean = 0.83, median = 0.83., SD = 0.01, IQR = 0.01)
Mann-Whitney test P = 0.30
Human agreement is generated from six participants scoring 40 images 1024x1024 pixels in size, using the Prospector system (section 4.3). The boxplot shows a difference in means of 0.02, and that the distributions are not significantly different. The IQR of the algorithm agreement is smaller than pathologist agreement by a factor of 10, indicating that the algorithm is suitable for replicating human scoring with more predictable levels of error.

The Mann-Whitney test accepts the null hypothesis that the two distributions are from independent samples with equal means and unknown variance ($p = 0.30$). This indicates that the two methods are statistically similar.

6.4.3.4 Correlation of TSRs

As with Algorithm H, TSR was calculated for all the QC-approved cases and compared to the ratios calculated from manual scoring.

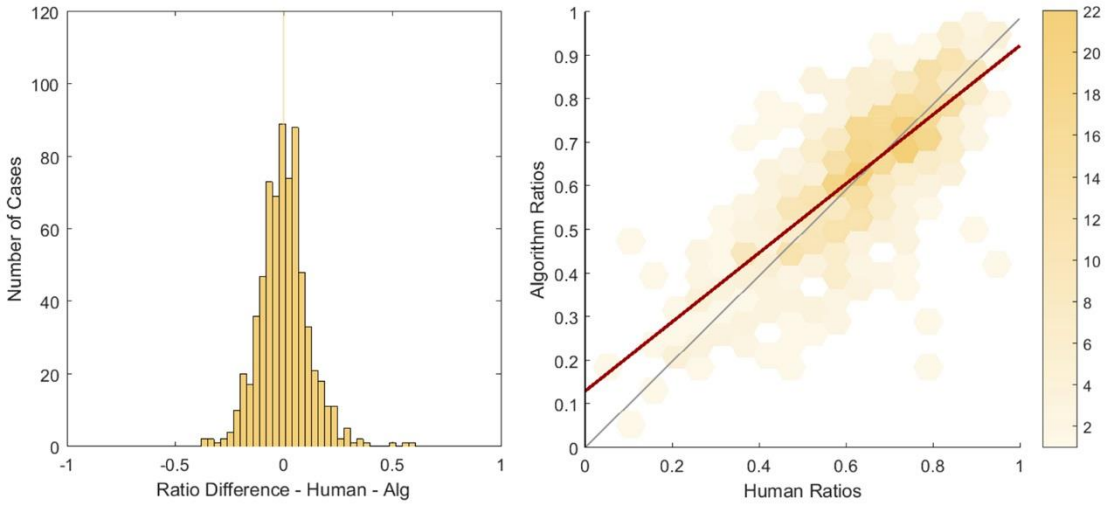


Figure 116 - Histogram and Heatmap Correlation Plots for Algorithm H on QC approved cases only

Left: Histogram of ratio differences generated by pathologist and regular segment algorithm. Distribution has a mean bias of 0 (median 0), and standard deviation of 0.11.
Right: Heatmap Correlation plot of the distribution of the ratios. Correlation has R^2 coefficient of 0.58. The distribution and correlation show higher levels of agreement compared to the analysis applied to the dataset with no QC applied.

Using the QC approved differences per case (ground truth highest TCD method, minus Combined Context algorithm method), the comparison dataset has a mean bias of 0, and a standard deviation of 0.11. A paired samples T-Test rejected the null hypothesis that the pairwise mean comparison of pathologist and algorithm TSRs is equal to zero ($p < 0.01$). A Spearman's nonparametric rank correlation test confirms that a positive relationship exists (rho = 0.74) and that the correlation is significantly different from zero ($p < 0.01$). The fitted linear regression model (red line on the plot) to the dataset has a coefficient of determination (R^2) of 0.58.

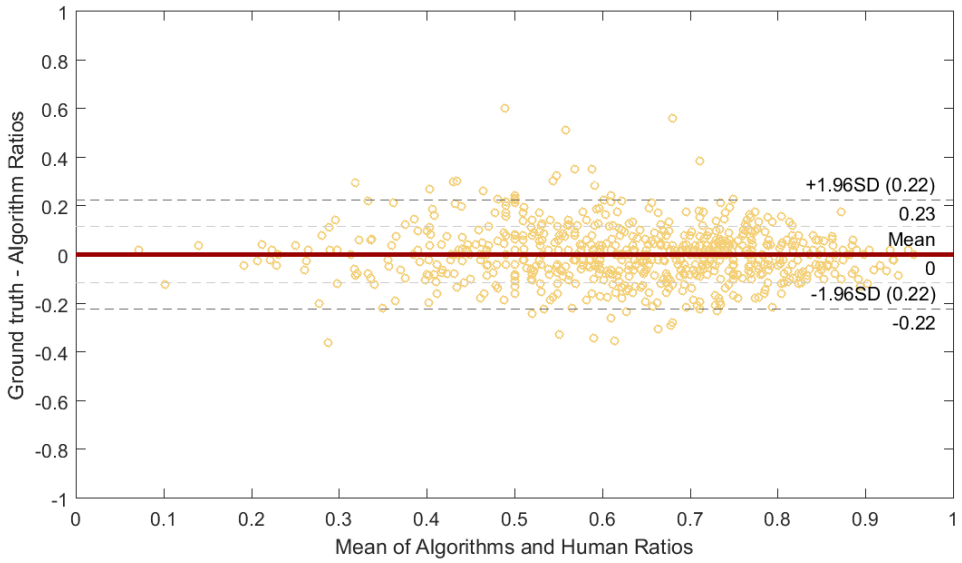


Figure 117 - Bland-Altman plot of Pathologist and Regular Segmentation algorithm-generated TSRs per case

Distribution has a mean bias of 0, with upper and lower limits of agreement of 0.23 and -0.22 respectively (±0.22). The plot shows a smaller level of variance then previous algorithms, with a consistent distribution that does not have higher variance at certain TSR levels.

The Bland-Altman plot in Figure 117 shows that the data has a bias of 0, and that the 95% confidence intervals (1.96 standard deviation, indicated by the outer dashed lines) are ±0.22. This improves over Algorithm A, which yielded upper and lower limits of agreement at ±0.37.

6.4.3.5 Results comparison

For clarity in assisting evaluation of the algorithm performance, Figure 118 displays boxplots of agreement for the human interaction experiment, performed on the Prospector system (in section 4.3), Algorithm H cross-validation results from section 5.3.2.5, and Algorithm H on the QC approved dataset (Algorithm I).

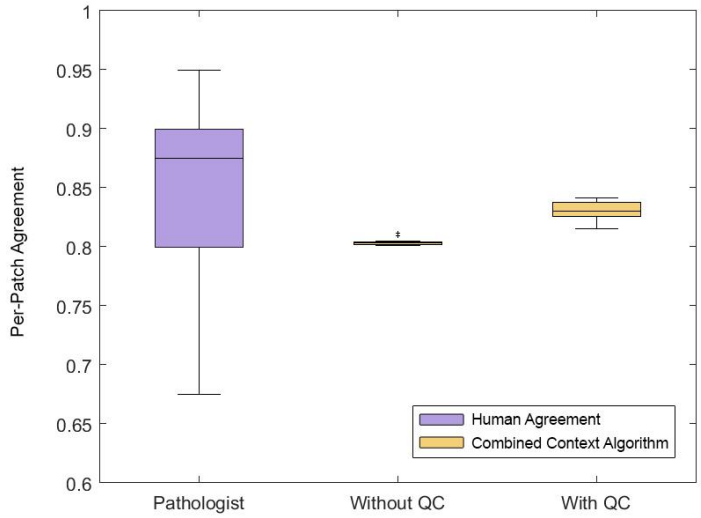


Figure 118 – Boxplot showing between-pathologist agreement and Algorithm H, with and without QC

Mean agreement for pathologist = 0.88, Combined Context without QC = 0.81, Combined Context with QC = 0.83
Note that pathologist agreement was calculated from the image size experiment, in section 4.3. The results from Algorithm H were when trained and tested on either the full dataset or the QC'd subset of slides. The figure shows that applying a QC filter on the dataset increases accuracy by 3%, and is more aligned with pathologist scoring.

To reiterate the previous findings, the pathologist agreement statistics were generated from six participants, looking at 40 images, compared to 106,262 images for Algorithm H, and 33,736 images for Algorithm I. The pathologist results were taken from the dataset using 1024x1024 pixel image patches, which yielded the highest level of agreement (but not statistically significantly so). Both algorithm results were generated from 256x256 pixel images, with Algorithm I, showing an increase median accuracy from 0.81 to 0.83, and a Mann-Whitney test reveals a statistically significant difference in distribution ($p < 0.01$). The pathologist agreement has a higher median accuracy, but has a larger spread of results. The pathologist boxplot data was generated from six participants instead of the algorithm methodology using ten-fold cross validation.

Note that another dataset was generated, which trained Algorithm H on the QC approved dataset, and tested on the full dataset. However, this reduced accuracy from the original combined context algorithm (H) and so was not reported. It is believed that this was because the RF classifier was trained on best-case images, which accounted for 32% of the total number of cases. It was therefore considered that reducing the number and type of training examples in this manner had a detrimental effect on the overall performance.

6.4.4 Conclusions

The results of this experiment have established that QC is important for increasing agreement to ground truth data, and improving correlation between pathologist and algorithm-generated TSRs.

The per-patch accuracy of Algorithm H increases from 80.50% to 83.18%. This is the first algorithm presented in this work which generates a kappa value of substantial agreement (0.64). The comparison of pathologist and Algorithm I accuracy distributions shows that they are not statistically different ($p = 0.30$). The mean difference of TSRs is 0, and the standard deviation of the distribution of differences per case is 0.11.

The increase in accuracy comes at a cost – since over half of the dataset was flagged as having at least one QC issue that would negatively affect image analysis. Removing this much data from a clinical trial is infeasible in practical applications, however, the dataset used for this work is abnormally faded, and the QC approved dataset is more representative of live clinical datasets.

Since the number of feature vectors in the training and testing dataset is reduced by 68%, algorithm performance is improved, but the results become less consistent.

Automation of the QC process based on overall features would be a useful tool as a pre-analysis step for automatic analysis of virtual slides.

6.5 Automation of quality control

6.5.1 Aim

To use the QC labelled dataset, generated in 6.3, for generating features and training a RF classifier to detect slides that are likely to fail a pathologist quality control check.

6.5.2 Methods

Quality control was superficially addressed as a pilot study in section 4.4, using levels of staining intensities. The main conclusion from that experiment was that images with less stain present were more likely to be rejected by a pathologist. This work in this section extends that study by looking at the QC issues identified from 6.3, and developing an image analysis solution for automatically flagging virtual slides with potential QC issues, if any.

Using the dataset of quality control labels generated from the work in section 6.3, cases were analysed for image features, using their respective 3x3mm box annotation, at the area of highest tumour cell density (described in 3.3.2.4). Images were retrieved at native zoom using block processing of image tiles, 2000x2000 pixels in size. The visual features generated were either using area-based quantification (pixel classification and area coverage using masks), or value-based quantification (pixel intensity values and statistics on those values). Image masks were created using both static and dynamic thresholds. In order to identify viable image foreground pixels, background pixels were isolated using a threshold of HSV intensity greater than 240, and black objects (representing debris, coverslip edges and bubbles) were identified as pixels with HSV intensity less than 30. Dynamic thresholds were applied to the foreground pixels to identify areas where staining was darker (to identify folds in tissue), using a threshold of one standard deviation below the mean foreground pixel intensity, and lighter (to identify areas of relative weak staining within the image), using a threshold of one standard deviation above the mean intensity of the foreground pixel values. The percentage area of these pixel classes present in the image was recorded in the feature vector, and Figure 119 illustrates these thresholds on the histogram of intensities for the image displayed. Using an edge strength heatmap (described in 5.2.2.4), basic textural properties were derived, using statistics from the intensities of the

edges. Finally, the median and standard deviation for individual HSV colour channels and Haematoxylin and Eosin channels were used after the (white) background pixel threshold was applied.

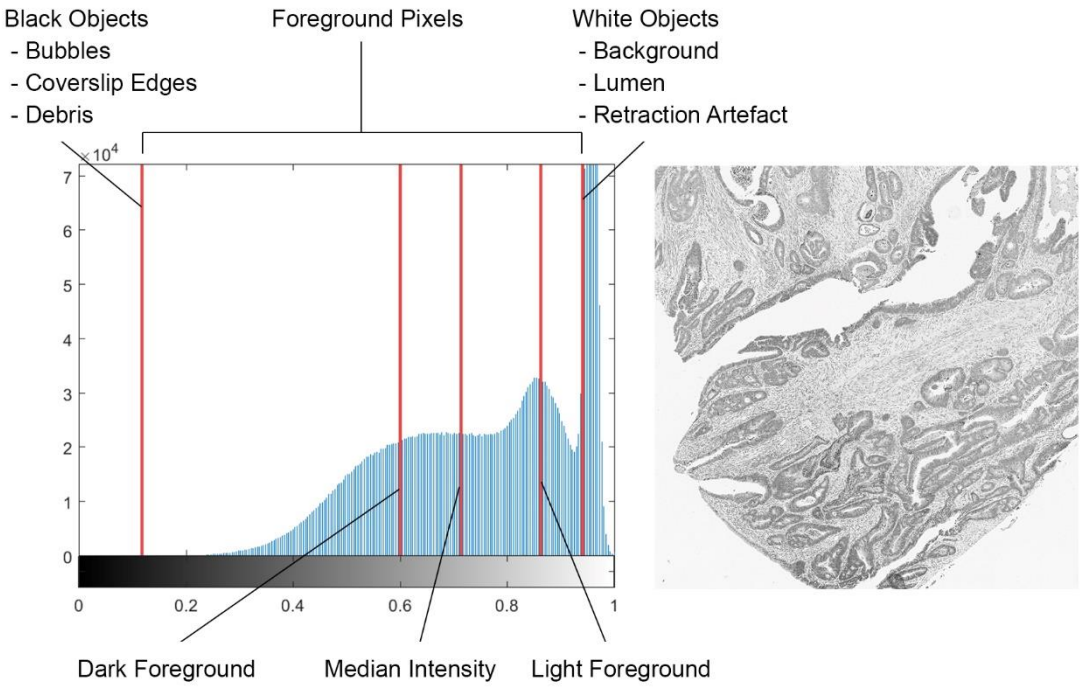


Figure 119 - Example of intensity cut-off values for pixel groupings

Left: Annotated histogram of all pixel intensities for the greyscale image on the right, with red lines indicating cut-off points. Black object threshold is set to 30 (of 255). White object threshold is set to 240 (of 255). Median intensity, and light and dark foreground values are calculated on the remaining pixel values (called foreground pixels).
Right: HSV intensity image of 3x3 mm area of tissue from the QUASAR dataset

A RF classifier was trained and tested using 10-fold cross validation of the 16-column feature vector, using the parameters detailed in 3.4.2.3. The classifier was trained and tested using three separate methods: with all 13 individual classes, by grouping the 12 reject classes into four parent classes, and binary accept/reject classes. The parent classes and groupings are displayed in Table 24.

QC category	Short code	QC group	QC binary
Accept	Acc	Accept	Accept
Weak overall staining	wkO	Staining	Reject
Weak haematoxylin	wkH	Staining	Reject
Weak eosin	wkE	Staining	Reject
Tissue tears	Trs	Preparation	Reject
Sectioning	Sec	Preparation	Reject
Tissue folds	Fld	Preparation	Reject
Mucin	Muc	Pathology	Reject
Necrosis	Nec	Pathology	Reject
Poorly differentiated tumour	PrD	Pathology	Reject
Debris	Dbr	Objects	Reject
Bubbles	Bub	Objects	Reject
Coverslip edges	Cov	Objects	Reject

Table 24 – QC categories grouped by type of QC issue

The table distils the 12 reject categories into four groups: staining issues, preparation issues, pathological issues and foreign objects. The right-hand column simply groups the 12 reject categories into one.

The processing was performed on the hardware detailed in section 3.4.2.4.

6.5.3 Results

6.5.3.1 Processing time

The QC feature set took 35,834 seconds to generate, using the methodology and hardware previously described. This equates to approximately 16.21 seconds per case. Training and testing for the feature set using individual QC categories took an average of 2.11 seconds to build each of the models of the 10-folds. Grouping the reject categories into the four parent classes yielded a quicker mean model-building time of 1.36 seconds per fold, and the binary classification took 1 second exactly.

6.5.3.2 Agreement

Agreement was calculated in three ways: calculating per-category agreement for all 13 QC categories, grouping the 13 classes into five parent classes (see Table 24), and grouping the 13 categories into a binary class dataset of accept and reject.

Figure 120 shows the confusion matrices for both evaluation methods of Algorithm E. The left matrix shows the agreement between algorithm and pathologist for the 13 QC categories. The 13 by 13 matrix is subdivided into five quadrants, which group the 12 reject classes into the four parent classes described. The summations of these results into the binary classes are depicted in the two by two confusion matrix to the right. Note for this evaluation, true positives were the correct prediction of the accepted QC class, and true negatives were the correct prediction of the rejected QC class.

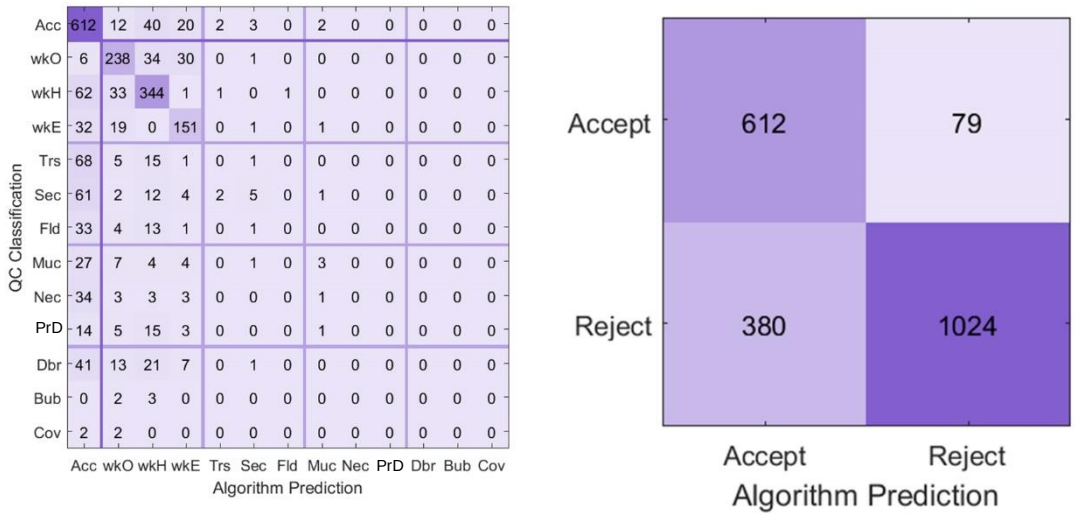


Figure 120 - Confusion matrices showing algorithm agreement for QC classification on all 13 categories

Left: Confusion matrix of all 13 QC subtypes from dataset. Accuracy = 64.58%, sensitivity (true positive rate / recall) = 0.65, Kappa = 0.53 (moderate agreement). The darker lines indicate the distinction between accepted images and rejected images, and the lighter lines delineate between the grouped QC categories.
Right: Confusion matrix of subtypes grouped into Accept and Reject parent classes. Accuracy = 78.09%, sensitivity (true positive rate / recall) = 0.89, specificity (true negative rate) = 0.73, kappa = 0.55 (moderate agreement).
The distribution of predictions shows that the algorithm is more likely to accept QC failed slides, which may be due to the distribution of classes in the training set. The algorithm shows highest levels of agreement on classes related to staining levels.

Figure 120 shows that the accuracy of the algorithm when applied to the individual QC categories was 64.85%, with a sensitivity of 0.65 and a kappa of 0.53, indicating moderate agreement. Grouping the individual categories into the parent accept/reject classes yielded a 78.09% accuracy, a sensitivity of 0.89, a specificity of 0.73 and a kappa value of 0.55, indicating moderate agreement. The algorithm has a higher number of false positives than false negatives (380 and 79 respectively).

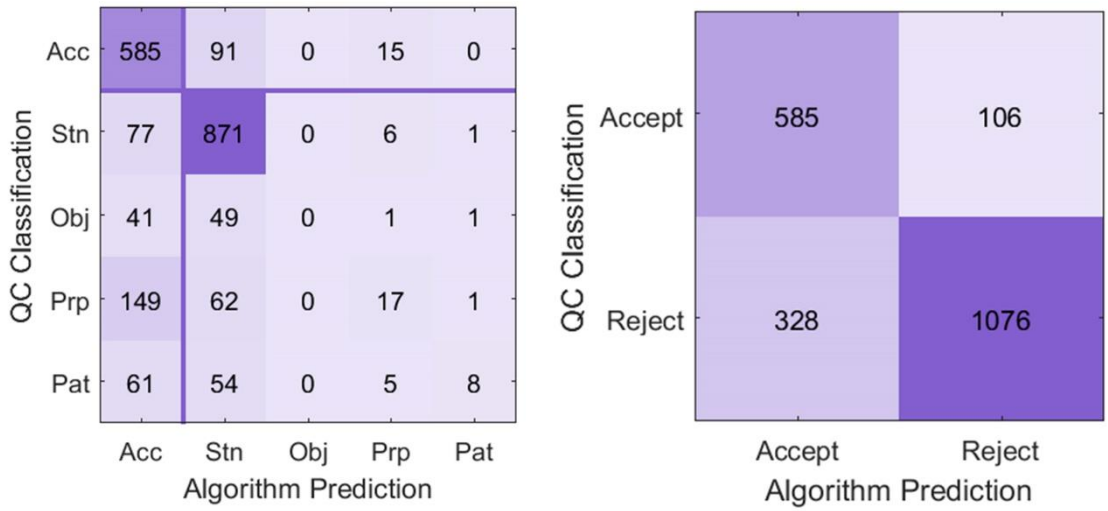


Figure 121 - Confusion matrices showing algorithm agreement for QC classification when trained and tested on using the five grouped category labels

Left: Confusion matrix of all five QC subtype groups from dataset. Accuracy = 70.69%, sensitivity (true positive rate / recall) = 0.83, kappa = 0.52 (moderate agreement). The darker lines indicate the distinction between accepted images and rejected images.
Right: Confusion matrix of subtypes grouped into Accept and Reject parent classes. Accuracy = 79.28%, sensitivity (true positive rate / recall) = 0.85, specificity (true negative rate) = 0.77, kappa = 0.57 (moderate agreement).
Grouping the QC issues into parent classes increases overall prediction accuracy by 1%. The confusion matrices again show that prediction error is lower in cases with staining issues.

Figure 121 shows that the accuracy of the algorithm when applied to the grouped QC categories was 70.69%, with a sensitivity of 0.83 and a kappa of 0.52, indicating moderate agreement. Grouping the individual categories into the parent accept/reject classes yielded a 79.28% accuracy, a sensitivity of 0.85, a specificity of 0.77 and a kappa value of 0.57, indicating moderate agreement. The algorithm has a higher number of false positives than false negatives (328 and 106 respectively).

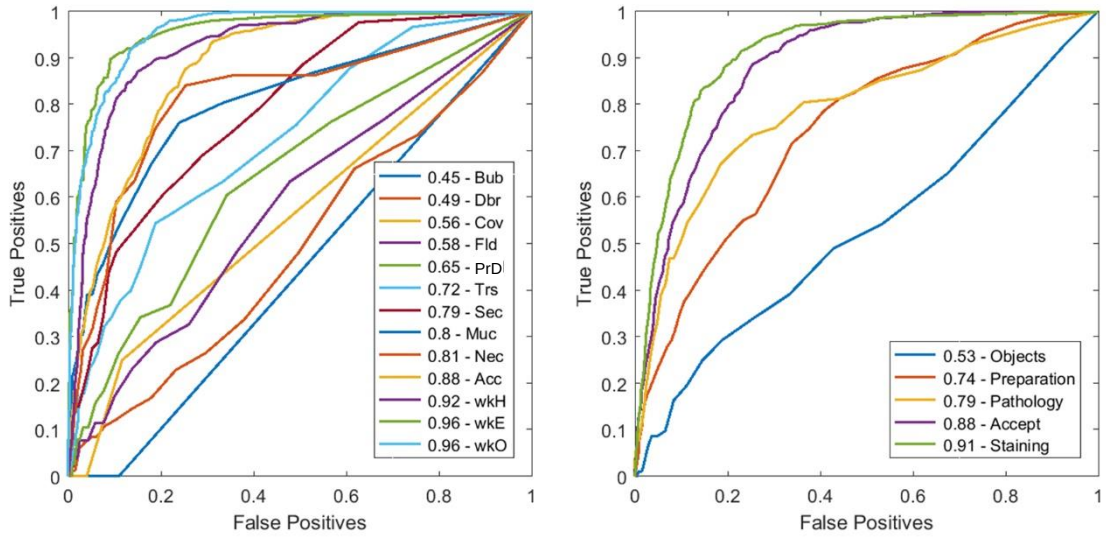


Figure 122 - ROC Curves for algorithm predictions of QC classifications

The graph shows Area Under the Curve for each QC classification:
Left: ROC curves for the individual QC categories – weak overall staining (0.96), weak eosin staining (0.96), weak haematoxylin staining (0.92), accepted images(0.88), large areas of necrotic tissue (0.81), large areas of mucin lakes (0.80), sectioning artefacts (0.79), tissue tears (0.71), poorly differentiated tumour (0.65), tissue folds (0.58), coverslip edges obscuring tissue (0.56), debris on the slide (0.49), bubbles under the coverslip (0.45)
Right: ROC curves for the grouped QC categories - staining issues (0.91), accepted images (0.88), pathological issues (0.79), slide preparation issues (0.74), foreign objects (0.53).
The ROC curves reiterate the findings that QC prediction is too variable to be of practical use, with the curve for bubbles showing a prediction rate less than 0.5 (equivalent to random). QC issues relating to stain have the highest AUC, and the lowest AUC is generated by predictions of issues relating to foreign objects on the slide.

Figure 122 shows both sets of ROC curves for the algorithm results presented in Figure 120 and Figure 121. The left figure shows the algorithm results for training and testing on individual QC categories, compared to the grouped categories on the right.

The individual QC categories have an AUC of 0.96 for weak overall staining, 0.96 for weak eosin staining, 0.92 for weak haematoxylin staining, 0.88 for accepted images, 0.81 for large areas of necrotic tissue, 0.80 for large areas of mucin lakes, 0.79 for sectioning artefacts, 0.72 for tissue tears, 0.65 for poorly differentiated tumour, 0.58 for tissue folds, 0.56 for coverslip edges obscuring the tissue, 0.49 for debris on the slide and 0.45 for bubbles under the coverslip. The prediction of individual QC categories had a mean AUC of 0.74. The grouped QC categories have an AUC of 0.91 for staining issues, 0.88 for accepted images, 0.79 for pathological issues, 0.47 for preparation issues and 0.53 for foreign objects. The mean AUC for the grouped categories was 0.77.

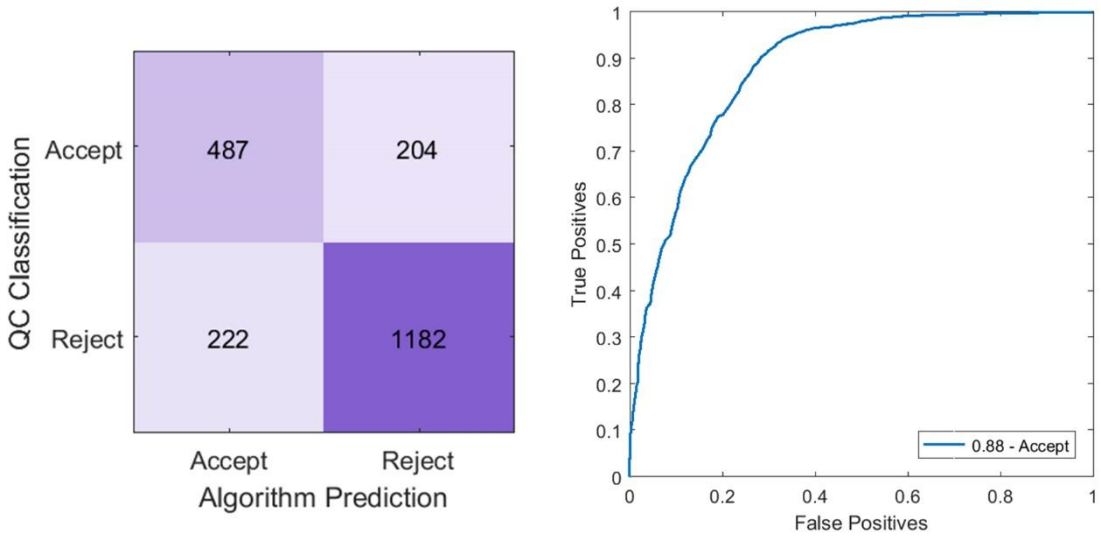


Figure 123- Confusion matrix and ROC curve showing algorithm agreement for QC classification when trained and tested on using the binary (accept-reject) labels

Left: Confusion matrix of all five QC subtype groups from dataset. Accuracy = 79.67%, sensitivity (true positive rate / recall) = 0.70, specificity (true negative rate) = 0.84, kappa = 0.54 (moderate agreement). Right: The graph shows Area Under the Curve for the Accept binary QC classification (0.88) By simply grouping the results into accept or reject, accuracy does not improve. The AUC for accept remains consistent (0.88) for all three methods tested. This means that the training examples are too varied to be of practical use when combined into one reject category.

Figure 123 shows both the confusion matrix and the ROC curve for the binary accept/reject QC categories. Accuracy of the algorithm is recorded at 79.67%, with a sensitivity of 0.70, a specificity of 0.84 and a kappa value of 0.54, indicating moderate agreement. The AUC for the accept rate is 0.88.

6.5.3.3 Post-hoc analysis: Algorithm as a QC tool for staining levels

Whilst analysing the results for this section, it was observed both agreement and AUC for stain-related categories was notably higher than the other rejection categories. As such, further analyses were conducted to identify the algorithm performance when only applied to weak staining as a QC metric. The feature vectors for all 2,211 cases were filtered by QC metric, leaving cases labelled with ‘Accept’, ‘Weak Overall’, ‘Weak Haematoxylin’ or ‘Weak Eosin’. The number of remaining rows was 1,647.

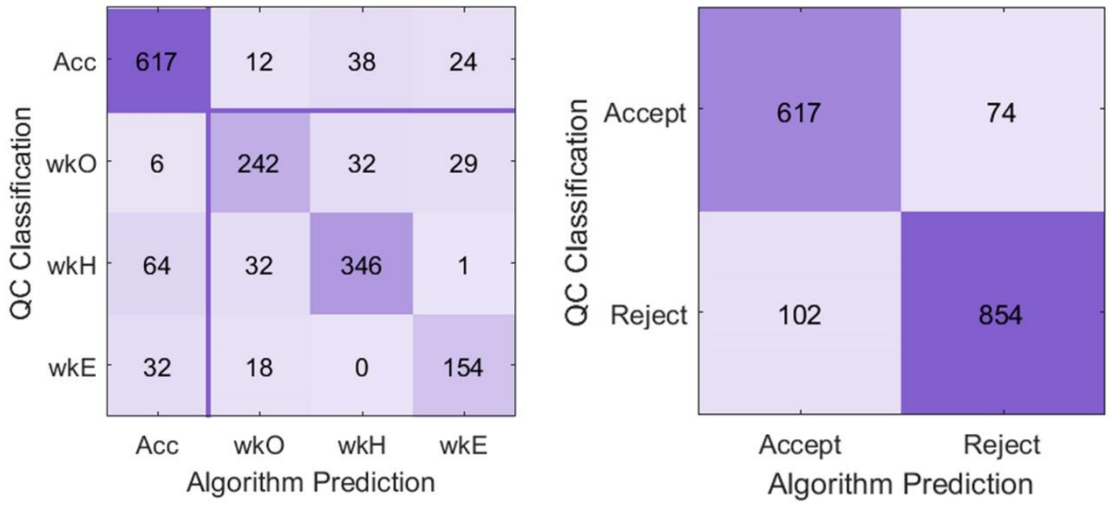


Figure 124 - Confusion matrices showing algorithm agreement for stain-related QC classification, trained and tested on the four classes only

Left: Confusion matrix of all 13 QC subtypes from dataset. Accuracy = 82.51%, sensitivity (true positive rate / recall) = 0.82, specificity (true negative rate) = 0.83, kappa = 0.75 (substantial agreement).
Right: Confusion matrix of subtypes grouped into Accept and Reject parent classes. Accuracy = 89.31%, sensitivity (true positive rate / recall) = 0.89, specificity (true negative rate) = 0.89, kappa = 0.78 (substantial agreement).
The confusion matrices show that when analysing staining as a QC metric isolated from other issues, the accuracy level is significantly higher. This means that the feature vectors used in the algorithm are suited to staining levels more so than objects. This is because the features used in this algorithm relate to intensity only.

Figure 124 shows that the accuracy of the algorithm when applied to the stain-based QC categories was 82.51%, with a sensitivity of 0.82 and a kappa of 0.75, indicating substantial agreement. Grouping the individual categories into the parent accept/reject classes yielded an 89.31% accuracy, a sensitivity of 0.89, a specificity of 0.89 and a kappa value of 0.78, indicating substantial agreement. The algorithm has a higher number of false positives than false negatives (102 and 74 respectively).

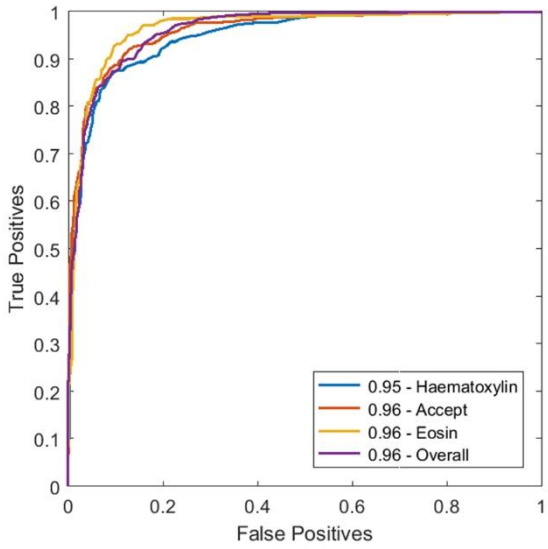


Figure 125 - ROC Curves for algorithm predictions of stain-based QC classifications

The graph shows Area Under the Curve for each stain-based QC classification:
Weak overall staining (0.96), weak eosin staining (0.96), accepted images (0.96), weak haematoxylin staining (0.95).
The curves show that the algorithm performs equally well on all staining related QC issues, with exception to the identification of weakly stained haematoxylin (AUC is 0.01 lower). This is due to the algorithm predicting more weakly stained haematoxylin slides as acceptable. This may be due to the ground truth data being biased towards assessing haematoxylin stain more stringently, since it is used to highlight nuclei and epithelial components. There are more issues related to weak H than any other QC issue.

The stain-based QC categories had an AUC of 0.96 for weak overall staining, 0.96 for weak eosin staining, 0.96 for accepted images and 0.95 for weak haematoxylin staining. The mean AUC was 0.96.

6.5.4 Conclusions

By using the image features described in 6.5.2, and the ground truth data generated in 6.3, QC can be applied to digital slide images with 79% accuracy, and a kappa value of 0.57. Grouping the categories before training and testing yields slightly higher levels of accuracy per-image, but not kappa agreement. Analysis of ROC curves shows that some individual QC categories have higher AUC values compared to when part of bigger groups. The loss of specific class definitions in the model is therefore detrimental to the individual class accuracy, but not the overall result.

As with the pilot data in 4.4, the dataset is imbalanced, and therefore model predictions favour the classes with the higher training examples.

Post-hoc analysis revealed that application of the algorithm to the staining-related QC issues yielded much higher accuracy, agreement and AUC statistics. The size of the dataset used for this analysis was reduced by 26% (to 1,647 cases), after removing cases not related to staining issues.

The algorithm presented has potential use for identification of staining issues as a QC metric in routine digital pathology applications. However, the algorithm is less successful at identifying other QC-related issues.

6.6 Discussion

The work in this chapter extends the pilot study from 4.4, which processed 256x256 pixel image patches for features, and trained a classifier based on labels generated from agreement statistics using the Prospector system.

The results from Algorithm H presented in 5.3.2.5 are analysed, and 100 cases with the highest levels of disagreement to the pathologist-generated TSRs are identified. These cases are qualitatively assessed for any visual issues that may affect image processing results. As a result, 12 QC issues were identified, and these were used as categories in the subsequent sections. The process of visually inspecting the cases was done on a standard computer screen, and results noted in excel. Most issues noted were extracted from comments made about the slides, but some (such as weak levels of stain) were pre-identified, and a simple stain presence metric was generated for each case. The number of cases was not scientifically selected, and there was no change or cut-off identified in the distribution of the TSR differences to suggest that the worst 100 cases would contain all the QC issues present in the dataset.

The manual QC experiment in 6.3 presented a single participant with all 2,211 cases, and the participant manually selected whether each case contained one of the QC issues (identified from the previous work), or was acceptable for analysis. This presented several issues, most notably the fact that the participant manually analysed such a large collection of images, that even with regular breaks, the dataset is likely to contain errors from fatigue, as well as the expected subjectivity. Another issue with the way that the experiment was conducted was that the participant could only select one of the QC issues. This meant that a slide presenting more than one of the conditions that would cause a case to be rejected, would be classified by the most visually prominent issue (adding another layer of subjectivity). Also, a number of QC issues identified were not whole-slide specific. Such issues didn't affect all of the tissue (which may or may not be adequate for analysis otherwise), meaning that even though these issues existed, by simply identifying them on the slide and removing that area from the analysis, the problem would be mitigated. This consideration relates more to object-based issues, such as debris, folds, tears, bubbles etc. The distribution of QC classes was unbalanced – 32% of the dataset was accepted for analysis without issue, 20% of the dataset had insufficient levels of haematoxylin staining present, 14% of the cases were weakly stained in both staining channels, and 9% had

weak eosin staining. This meant that approximately 75% of the data was accounted for by these four categories (out of 12). As briefly explored in 4.4, having an imbalanced dataset is not ideal for ML.

Having applied QC labels to the entire dataset, Algorithm H was further analysed to assess the pathologist-algorithm TSR differences, grouped by QC category. This information assisted in identifying which of the QC metrics had the biggest impact on skewing the data, which led to suggestions on how to improve the algorithm further. The distribution of ratio differences for weakly stained slides (H, E or both), was largely unaffected by the perceived inadequate levels of stain, which may be (at least in part) due to the global contextual analysis that factors the whole slide appearance into the local patch analysis.

With the dataset split by QC categories, Algorithm H was run on the accepted cases only. This was a subset of 701 from the 2,211 cases, and as such reduced the local feature set from 106,242 to 33,736 feature vectors. The reduction in data made the reproducibility of the algorithm slightly lower, but increased the mean accuracy by 3%. This increase equates to 1,012 images. The indications from this study are that reducing the number of suboptimal slides for automated analysis will improve accuracy, at the cost of losing data.

The final portion of work in this chapter focused on using the existing QC labels in order to apply them to global slide statistics, and use that data for training a classifier to learn the appearance of QC issues. The metrics used for analysis were similar to Algorithm G, using median pixel values as well as percentage of pixels in each image within a defined intensity range. The feature vectors were trained and tested using individual QC metrics as the labels, as well as grouping the QC labels into parent classes, and grouping into the simple accept-reject binary classes. All accuracy results were roughly the same (±0.5%) after grouping into accept/reject categories, but AUC of individual categories indicated that the accuracy of stain related QC fails was much more reliable.

Post-hoc analysis of the QC algorithm explored the accuracy of results when just applied to stain-based QC issues. The agreement results for the algorithm were higher (89% with 0.78 kappa agreement), and the ROC curves for all classes were 0.95 or above, indicating the algorithm is adequate for application as a stain quality tool. The fact that stain-based QC identification was much more reliable suggests that the features used to analyse the slide images were only suitable for that task, and identification of other types of fails that are localised or object-based need other image analysis methodologies to compensate for them.

The use of automated QC in a clinical dataset is a potentially important tool to remove images that will not yield accurate results for automated solutions. However, this methodology would

mean that data would be lost, and this is not an optimal solution. It may be more viable to use such a tool as a method for flagging cases where optimal QC conditions are not met, either before or after analysis, allowing the pathologist to review the cases which may have less accurate results.

Chapter 7 – Application to clinical data

7.1 Introduction

7.1.1 Chapter overview

The work in this chapter evaluates Algorithm H presented in Chapter 5, applied to a clinical dataset so that the algorithm's practical use for predicting survival can be evaluated. The chapter is divided into five sections:

1) The introduction, which briefly outlines the clinical datasets available via the RandomSpotDB (presented in Chapter 3), the availability of survival data within those datasets, and the algorithm selection for processing that data.
2) Exploration of the MRC CR07 dataset, describing the original project, the digital slides used for an experiment using SRS on the cases, and associated spot counting data stored in RandomSpotDB.
3) Application of Algorithm H to the CR07 dataset, in order to evaluate its accuracy using the same methodology as previous chapters.
4) Analysis of survival statistics of groups of patients, stratified by identifying appropriate cut-off points using cases ordered by algorithm-generated TSRs.
5) Discussion of the work presented in this chapter and conclusions. The discussion focuses on the how well the algorithms perform compared to the manual scoring and concludes by discussing whether the algorithm is appropriate for application to clinical data.

7.1.2 Clinical datasets

The RandomSpotDB (presented in 3.2.2.5) contained 32 separate projects that used the RandomSpot system (section 3.2.2.2) for various research purposes. Of the 32 projects, 13 were

from clinical trial studies, including the QUASAR dataset (section 3.3.2.1). This dataset was originally chosen for the work in previous chapters due to the large number of patient cases (2,211 available in the system) that had been scored by a pathologist, and the survival data associated with that trial (including survival and response to chemoradiotherapy). However, survival data was only accessible via the trial's national data centre, which required independent statisticians to run tests on the data. During the course of algorithm development, it emerged that it would not be feasible to retrieve survival analyses in a timescale appropriate to the project, and so other datasets were considered. One dataset in the RandomSpotDB, which had readily accessible survival data for independent analysis was selected instead of the QUASAR dataset. The trial, referred to as the CR07 dataset, is presented in section 7.2.

7.1.3 Algorithm selection for survival analysis

Of all the algorithms presented in this work, Algorithm H (section 5.3.2.5) exhibited the highest level of algorithm-pathologist agreement, and so was considered the most appropriate for application to survival analysis. The application of the algorithm to the CR07 dataset is presented in 7.3, and uses the same methodology from Algorithm H, presented in Chapter 5.

7.2 The CR07 dataset

7.2.1 The CR07 trial

The Medical Research Council (MRC) CR07 and National Cancer Institute of Canada (NCIC) clinical trial consisted of 1,350 patients with operable adenocarcinomas of the rectum, which were randomised in 80 centres across four countries [265]. Patients were randomised between two trial arms: arm 1 patients had preoperative short-course radiotherapy (SCRT, n=674) and arm 2 patients had selective postoperative chemoradiotherapy (CRT, n=676). Survival of patients was recorded in terms as the number of days that had elapsed between the patient entering the trial, and the final follow-up session for that patient. The status of the patient was recorded, in terms of whether they were alive, or had died of cancer or other causes at the time of the final follow-up session, in terms of whether the patient's cancer progressed to other organs (metastasised).

7.2.2 Image data

A subset of cases from the CR07 trial that were resected in Leeds were identified for a study using the RandomSpot system (section 3.2.2.2) for manual generation of TSRs [6]. These cases were fixed in paraffin wax blocks, sectioned into 5-micron thick samples, and stained with Haematoxylin and Eosin (H&E) to visually identify tumour regions. Glass slides were scanned using an Aperio AT scanner using a 20x objective (0.5 microns per pixel), using JPEG2000 compression quality 50.

It is important to note that the CR07 image data contained images from both arms of the randomised trial. This meant that half of the image data (n=674) was of tissue that had already undergone SCRT, and as such reduced the amount of tumour in the tissue. This reduction in tumour subsequently affects the TSR of the images, and does not correspond to the published prognostic significance of TSR in tumours. Therefore, for image analysis validation in terms of per-image patch pathologist agreement (section 7.3), both arms of the trial were combined to form the full dataset, and for survival analysis (section 7.4), the datasets for each of the trial arms were separately analysed.

7.2.3 Scoring data

The RandomSpot system was used on a subset of cases from the CR07 trial to manually quantify the TSR as a prognostic marker for CRC. The scoring data for these cases was generated using the same methodology described in section 3.2.2.2, and once generated and analysed for its original purpose, the scoring data was added to the RandomSpotDB (section 3.2.2.5). TSR was calculated using Method 2 in Table 7, which is referred to as Tumour Cell Density (TCD) in the original study.

Three different methods for sampling the tumours were used in the original study, outlined in Table 25. Each method used 300 sampling points, and the number of cases stored in the RandomSpotDB is recorded.

Sampling location	**ROI boundary type**	**# Sampling points**	**# Cases**
Whole tumour	Freehand polygon	300	344
Luminal TCD	3x3mm square highest TCD on luminal aspect	300	76
Greatest TCD	3x3mm square highest TCD away from luminal aspect	300	5

Table 25 – Sampling methods applied to the CR07 dataset

The three sampling methods used for previous studies calculating the TSR manually on the CR07 dataset and the number of cases sampled with each, available in RandomSpotDB [247]

The whole tumour method involved manually drawing a polygon around the whole tumour on the slide, using the freehand polygon tool in ImageScope. The Luminal TCD used a 3x3mm rectangle region of interest, manually placed over the area with highest perceived level of tumour cell density. The Greatest TCD method was only applied if there appeared to be an area on the slide which had higher perceived tumour cell density away from the luminal aspect. Figure 126 illustrates these methods with examples.

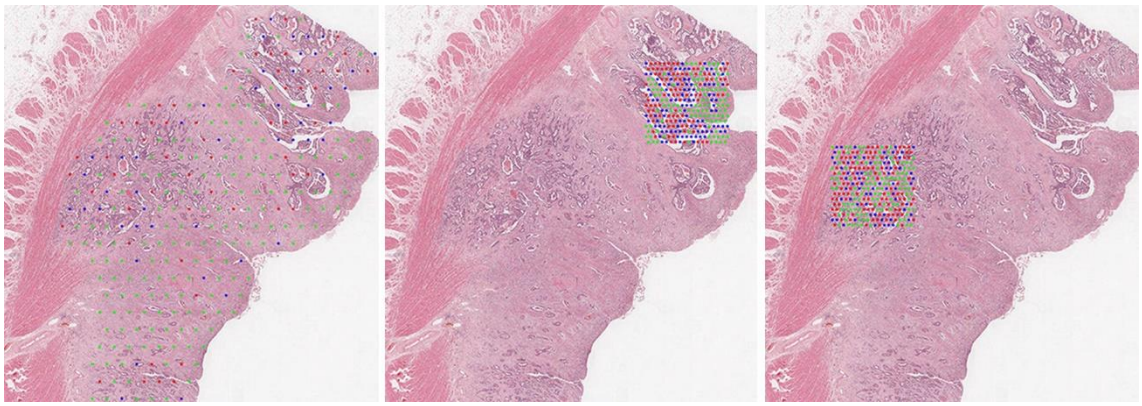


Figure 126 - Sampling methods applied to the MRC CR07 clinical trial dataset

Left: Whole tumour annotations using 300 sampling points (with tolerance of 15%)
Centre: Luminal TCD annotations using 300 sampling points (with tolerance of 15%)
Right: Greatest TCD annotations using 300 sampling points (with tolerance of 15%)
Note that sampling points in these examples are colour coded: red for tumour, green for stroma and blue for other classes.

In most cases, the Luminal TCD annotation contained the highest area of tumour cell density on the whole slide, and so Greatest TCD was not generated. The dataset from RandomSpotDB also contained one double scored whole tumour case, and 4 biopsy cases, totalling 430 sets of annotations for training and testing.

Note that the original TSR study using the CR07 dataset used the Luminal TCD annotations to show that the TSR is an independent prognostic marker. However, due to the smaller number of cases available in RandomSpotDB that used the Luminal TCD method, survival analysis in section 7.4 is performed on the whole tumour annotations dataset.

7.2.4 Data observations

Distribution

Using the full dataset from RandomSpotDB (section 3.2.2.5), the data were analysed to assess the distribution of classifications. Figure 127 shows that out of all 128,878 expert-classified locations for the CR07 dataset, 37,158 (42%) belonged to the Tumour class (Tumour, Lumen, Necrosis and Mucin subclasses), 62,937 (49%) belonged to the Stroma class (Stroma, Muscle, Vessels and Inflammation), and 11,697 (9%) of them were non-informative.

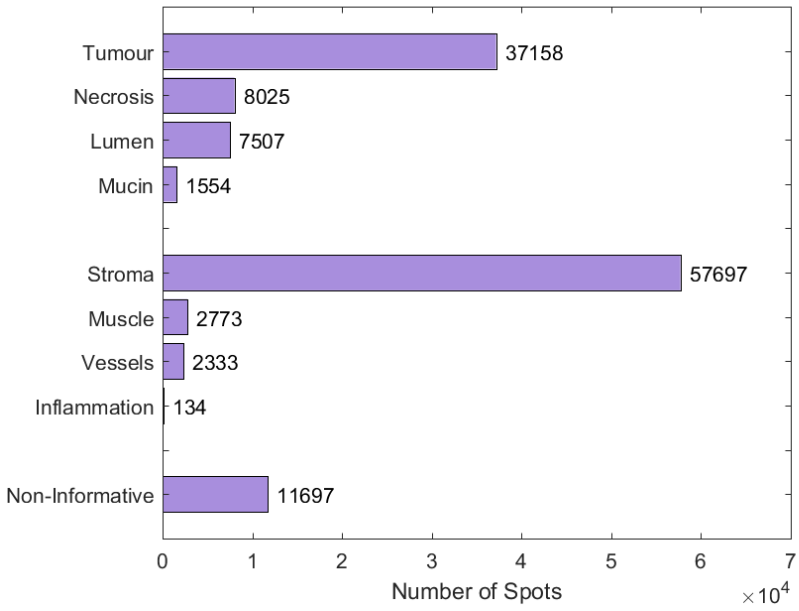


Figure 127 - Bar plot of all available data for the CR07 set, grouped by classification

Bars are grouped by parent classes, tumour (top), stroma (middle) and non-informative (bottom). The distribution has a higher proportion of stroma to tumour. This is the inverse of the QUASAR dataset, and likely due to the sampling methods chosen. The CR07 dataset uses sampling from the whole tumour area, and therefore is likely to have lower TSR ratees than when sampling from the areas of highest TCD.

TSR

For each of the 430 cases in the dataset, TSRs were generated, and their distributions plotted in Figure 128, spearating the two different sampling methods into separate histograms so that their distributions could be compared. Note that in the original CR07 TCD study, the TSR was calculated differently to the QUASAR dataset, using Method 2, instead of Method 1, described in section 3.3.2.4.

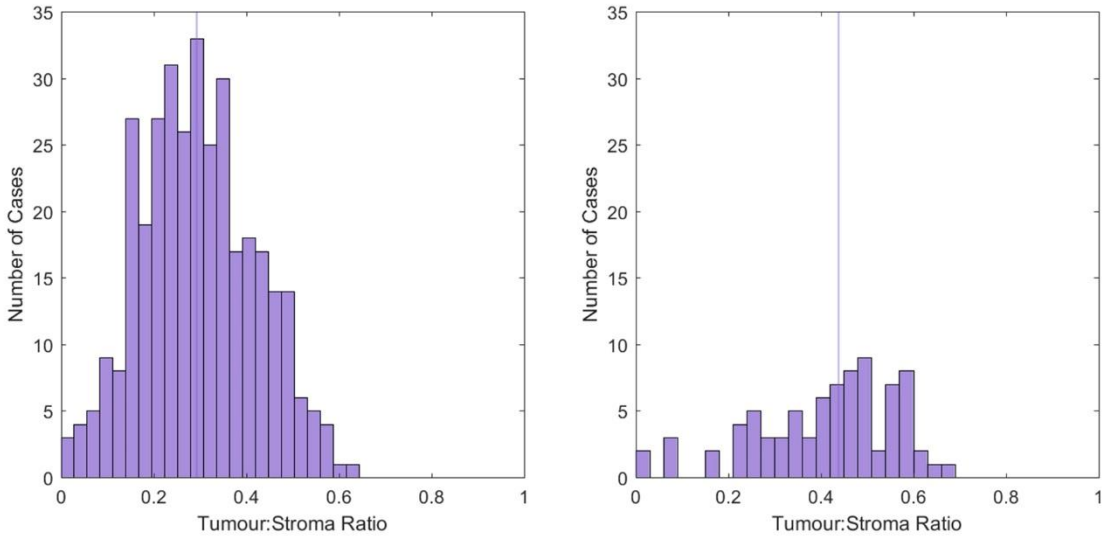


Figure 128 – TSR distribution for both sampling methods of the CR07 dataset, available in RandomSpotDB

Left: Whole tumour region (hand drawn polygon) with 300 sampling points (mean = 0.30, median = 0.29, SD = 0.12)
Right 3x3mm box annotation over the area with highest tumour cell density with 300 sampling points (mean = 0.41, median = 0.44, SD = 0.15)
The histograms confirm that the sampling methods affect the distributions of TSR. Despite the lack of examples for sampling the area of highest TCD, it is still clear that this method yields higher TSRs.

For the whole tumour and highest TCD datasets, the means are 0.30 and 0.41 resepectively. The whole tumour dataset had a standard deviation of 0.12 and the highest TCD dataset has a standard deviatin of 0.15. For both of the datasets, both Shapiro-Wilk and Kolmogorov-Smirnov tests reject the null hypothesis that the data is from a standard normal distribution (both tests $p < 0.01$).

Histograms of TSR by trial arm are presented in 7.4.2.

7.2.5 Summary

The data from the CR07 dataset contains proportionally more stroma sampling points than the QUASAR dataset. The original TCD study found that survival for the CR07 cases was lower than the other trials used to generate the results. This is consistent with the previous findings that lower proportions of tumour indicate a worse prognosis.

As with the QUASAR dataset, the differences in TSR distributions of the sampling methods are caused by either sampling at the area of highest TCD or the overall tumour. Previous studies found that sampling at the highest area of TCD is more strongly correlated to survival, however

the amount of data available for this work means that reproducing survival analysis on the areas of highest TCD would discard 81% of the data.

There are a smaller number of cases in the CR07 dataset than the QUASAR dataset (430 and 2,211 respectively), but there are more individual expert-classified labels in the CR07 dataset available for analysis, due to sampling 300 spots per region of interest rather than 50.

7.3 Algorithm J: Algorithm H applied to the CR07 dataset

7.3.1 Aim

To apply Algorithm H to the CR07 dataset, and establish algorithm accuracy, validating it against a different clinical trial.

7.3.2 Methods

The methodology for Algorithm H presented in section 5.3.2.5 was used to train a RF classifier using the globally-weighted local features, described in the global context algorithm (Algorithm G) section in 5.3.2.4. Once trained, the algorithm used the local contextual algorithm (Algorithm F) methodology to generate the classifier predictions at the centre of each image patch (section 5.3.2.3), making the final feature set containing 116,591 vectors. Once the feature dataset (with weighted predictions) was generated, another RF classifier was trained and tested using 10-fold cross validation. This algorithm is referred to as Algorithm J.

The study was also implemented using the trained QC stain classifier presented in section 6.5.3.3, to establish whether there were any suboptimal slides in the dataset that should be excluded for image analysis. The methodology of the original QC algorithm used the 3x3mm box annotations from the QUASAR dataset to train on, which had higher density of tumour and higher levels of stain intensity (weaker staining). Therefore, the algorithm was not appropriately trained for the whole annotations used in the CR07 dataset. It was presumed that the whole annotations would have a different overall appearance, thus rendering the original trained classifier ineffectual. The original QC algorithm was applied to the dataset to test the presumption, and rejected 55% of the cases. The algorithm was subsequently retrained using the QUASAR whole annotations, using a NaN mask over areas that were outside of the freehand annotation. This algorithm is referred to as Algorithm K.

The processing was performed on the hardware detailed in section 3.4.2.4.

7.3.3 Results

7.3.3.1 Processing time

Algorithm J took 339,566 seconds (94.32 hours) to generate the image patch feature set, using the methodology and hardware previously described. This equates to approximately 2.91 seconds per patch.

7.3.3.2 QC analysis

The QC algorithm presented in Chapter 6 was modified with training data from the QUASAR dataset, using whole tumour, freehand ROI annotations, instead of the 3x3mm boxes originally used. This was re-implemented in order to maintain comparability between datasets, so that the QC algorithm was trained on whole annotations from the QUASAR dataset and was tested on the CR07 whole annotations only. This reduced the dataset from 430 to 344 cases before QC was applied. Of the 344 remaining cases, 264 slides were approved for image analysis by the automated QC method (77%), 43 slides were identified as having suboptimal levels of haematoxylin (13%), 25 slides were classified as having weak eosin staining (7%), and 7 slides were flagged for weak overall staining (2%). The five slides unaccounted for were cases that had been double scored, and so to prevent duplicates, one of the annotations was excluded from analysis for each pair. By removing cases flagged for QC issues, the size of the dataset of spot features reduced by 31%, from 116,591 to 80,493. Due to the comparatively smaller number of cases compared to the QUASAR series, in depth analysis was performed on the full set of CR07 dataset to maximise the number of results, and QC results were applied as a filter afterwards.

Creating the new trained classifier using whole annotations took 5,470 seconds of processing time, and application of the algorithm to the CR07 dataset took 2,326 seconds, using the hardware previously described.

7.3.3.3 Accuracy (agreement)

Agreement was calculated in two ways: calculating per-patch agreement for all eight tissue classes, and grouping the eight classes into their parent class: tumour and stoma. Figure 129 shows the confusion matrices for both evaluation methods of Algorithm J. The left matrix shows the agreement between algorithm and pathologist for the eight tissue classes. The eight by eight matrix is subdivided into four quadrants, which group the eight classes into the two

parent classes, tumour and stroma. The summations of these results into the parent classes are depicted in the two by two confusion matrix to the right.

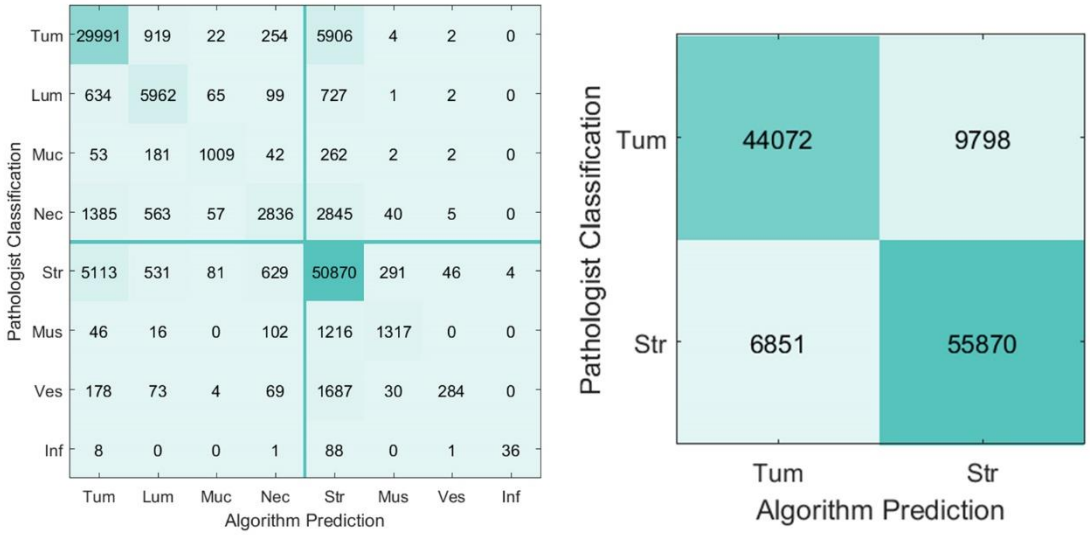


Figure 129 – Confusion matrices showing pathologist – Algorithm J agreement using the CR07 dataset for training and testing

Left: Confusion matrix of all 8 tissue subtypes from dataset. Accuracy = 79.17%, sensitivity (true positive rate / recall) = 0.79, kappa = 0.66 (substantial agreement)
Right: Confusion matrix of subtypes grouped into Tumour and Stroma parent classes. Accuracy = 85.72%, sensitivity (true positive rate / recall) = 0.82, specificity (true negative rate) = 0.89, kappa = 0.71 (substantial agreement)
The matrices show that the algorithm applied to the CR07 dataset is more likely to overestimate stroma compared to tumour. This shift (compared to other algorithms overestimating tumour) is likely due to there being more examples of stroma in the training set. As with all other algorithms, the majority of false positives and negatives are from incorrect predictions of tumour and stroma subclasses, with incorrect predictions of necrosis having the third highest number of false predictions. The false positives and false negatives exhibited in this figure are lower than that of any other algorithm presented.

The accuracy of the algorithm for the individual tissue classes was 79.17%, with a sensitivity of 0.79 and a kappa of 0.66, indicating substantial agreement. Grouping the individual classes into the parent Tumour / Stroma classes yielded an 85.72% accuracy, a sensitivity of 0.82, a specificity of 0.89 and a kappa value of 0.71, indicating substantial agreement. The algorithm has a higher number of false negatives than false positives (9,798 and 6,851 respectively).

The results show a significant improvement over Algorithm H (6% increase in grouped accuracy, Mann-Whitney test P = < 0.01), which uses the same methodology applied to the QUASAR dataset (as opposed to the CR07 dataset). The confusion matrices show that Algorithm J is more likely to overestimate the presence of stroma by 2.52% (9,798 false negatives (8%) compared to 6,851 false positives (6%)), where Algorithm H is more likely to overestimate tumour (10,568 false positives (10%) compared to 10,174 false positives (10%)).

This may be due to the faded staining in the QUASAR dataset used for algorithm H, leading to feature vectors for tumour containing higher intensity values, which would be more likely to represent stroma in a well stained dataset. It may also be due to the balance of examples in the dataset – 49% of the CR07 dataset is stroma, compared to 35% in the QUASAR dataset. The CR07 dataset has a mix of patients with preoperative SCRT and postoperative CRT, meaning that half of the dataset is comprised of images that have areas of tumour destroyed by radiation therapy, leading to higher proportions of stroma. Having more examples of stroma in the training set could lead to algorithm biases, which may be mitigated by data balancing, or augmentation of the dataset.

A comparison of results can be found in section 7.4.3.

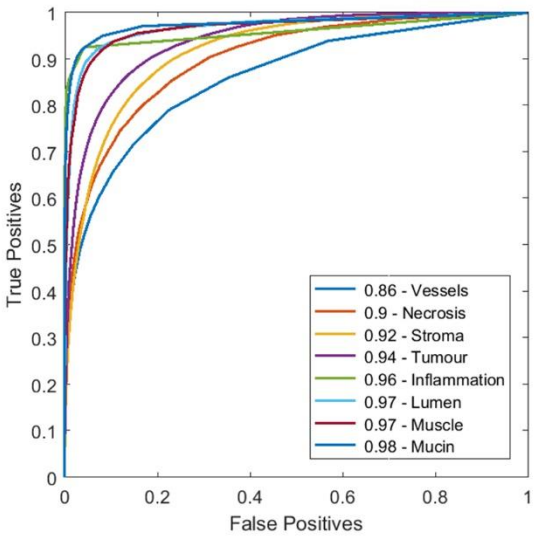


Figure 130 - ROC Curves for all 8 tissue subtypes of the CR07 dataset, classified by Algorithm J

The graph shows Area Under the Curve for each tissue subtype:
Tumour parent class: Tumour (0.94), Lumen (0.97), Mucin (0.98), Necrosis (0.90)
Stroma parent class: Stroma (0.92), Vessels (0.86), Muscle (0.97), Inflammation (0.96)
The curves show that the algorithm is least accurate at correctly predicting vessels, and has the highest AUC for mucin. Despite the larger number of training examples, the algorithm is still better at predicting tumour than stroma. This suggests that the variable appearance of stroma is more difficult to model accurately.

Figure 130 shows the ROC curves for each individual tissue class when classified by Algorithm J. The tumour subtypes have an AUC of 0.94 for tumour, 0.97 for lumen, 0.98 for mucin and 0.90 for necrosis. The stroma subtypes have an AUC of 0.92 for stroma, 0.86 for vessels, 0.97 for muscle and 0.96 for inflammation. The mean AUC for all subtypes is 0.94.

7.3.3.4 Comparison to manual scoring

For comparison, the algorithm grouped accuracy results are plotted against the pathologist agreement statistics generated on the best-case data from the image patch size experiment in Chapter 4.

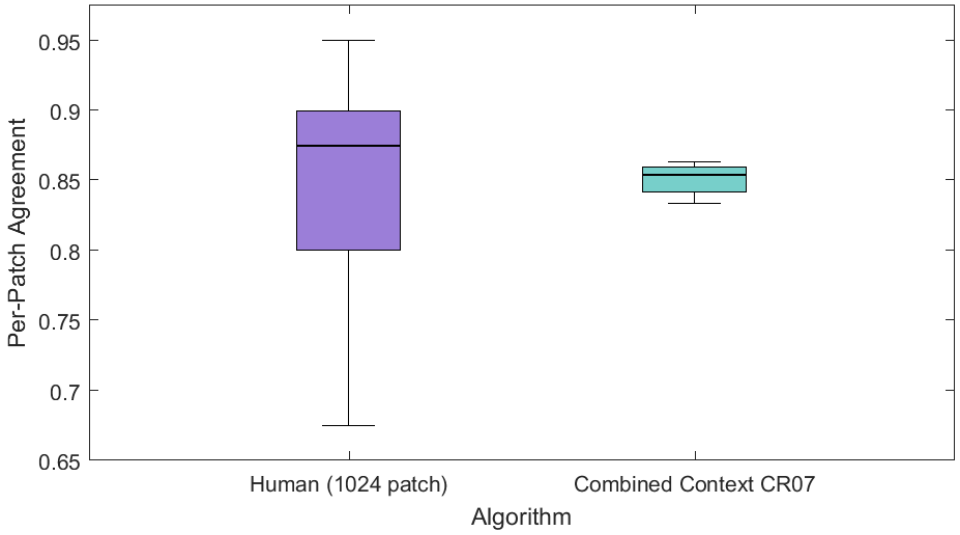


Figure 131 – Comparison boxplots for pathologist-pathologist agreement and pathologist-algorithm agreement

Left: Pathologist - pathologist agreement (mean = 0.85, median = 0.88., SD = 0.10, IQR = 0.10)
Right: Pathologist -algorithm agreement (mean = 0.85, median = 0.85., SD = 0.01, IQR = 0.01)
Mann-Whitney test P = 0.69
Human agreement is generated from six participants, each scoring 40 images 1024x1024 pixels in size, using the Prospector system (section 4.3). The boxplots show that the automated analysis of the CR07 dataset is comparable to the agreement levels of human scoring.

Note the slight difference between accuracy reported by the Figure 129 confusion matrix compared to the Figure 131 boxplot. This is due to the grouping methods per cross validation fold for boxplot results (mean of 10 accuracy levels), and grouping applied after all cross-validation predictions have been made for the calculation of the confusion matrix statistics (one single accuracy level).

A Mann-Whitney test accepts the null hypothesis that the two distributions are from independent samples with equal means and unknown variance ($p = 0.69$). This indicates that the two methods are statistically similar.

7.3.3.5 Correlation of pathologist and algorithm TSRs

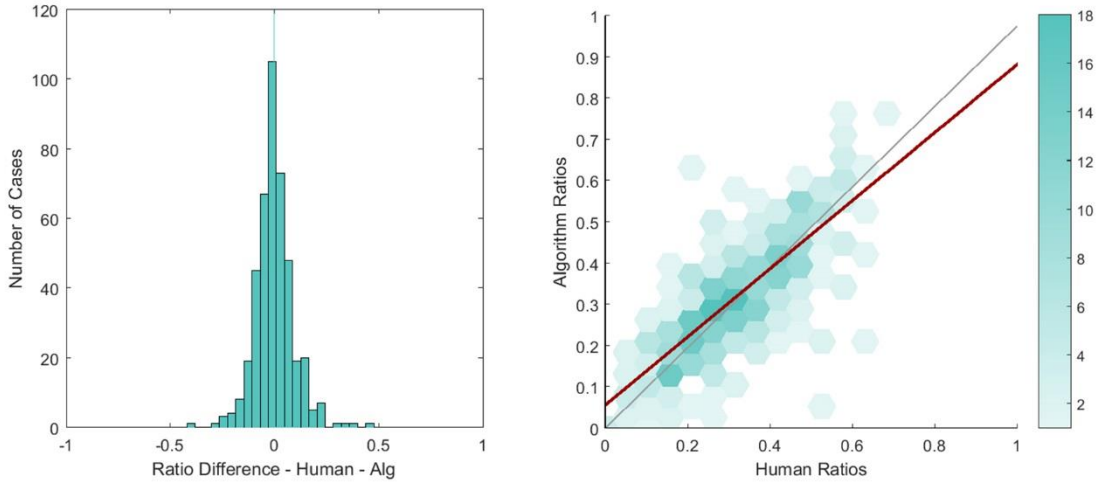


Figure 132 – Histogram and Heatmap Correlation Plots for Algorithm J on the CR07 dataset

Left: Histogram of ratio differences generated by pathologist and regular segment algorithm. Distribution has a mean bias of 0 (median 0), and standard deviation of 0.09.
Right: Heatmap Correlation plot of the distribution of the ratios. Correlation has R^2 coefficient of 0.61. The histogram and correlation plot show that the TSRs generated by pathologist and machine are more similar than previous algorithms.

Using the differences of TSR per case (ground truth TSRs minus algorithm TSRs), the histogram shows that Algorithm J shows no bias to overestimating tumour or stroma. The comparison dataset has a mean bias of 0, and a standard deviation of 0.09. The distribution of the TSR differences shows that the algorithm performs similarly to pathologist scoring, and has the lowest variance of any algorithm presented in this work. This confirms that conclusions from Chapter 6, that the algorithm performs better on better quality images.

A paired samples T-Test accepted the null hypothesis that the pairwise mean comparison of pathologist-algorithm TSRs is equal to zero ($p = 0.90$). A Spearman's nonparametric rank correlation test confirms that a positive relationship exists ($rho = 0.78$) and that the correlation is significantly different from zero ($p < 0.01$). The fitted linear regression model (red line on the plot) to the dataset has a coefficient of determination (R^2) of 0.67. The distribution of the correlation shows that the TSRs are closer to zero, which is caused by the sampling area being appliedto the whole tumour rather than the area of highest TCD.

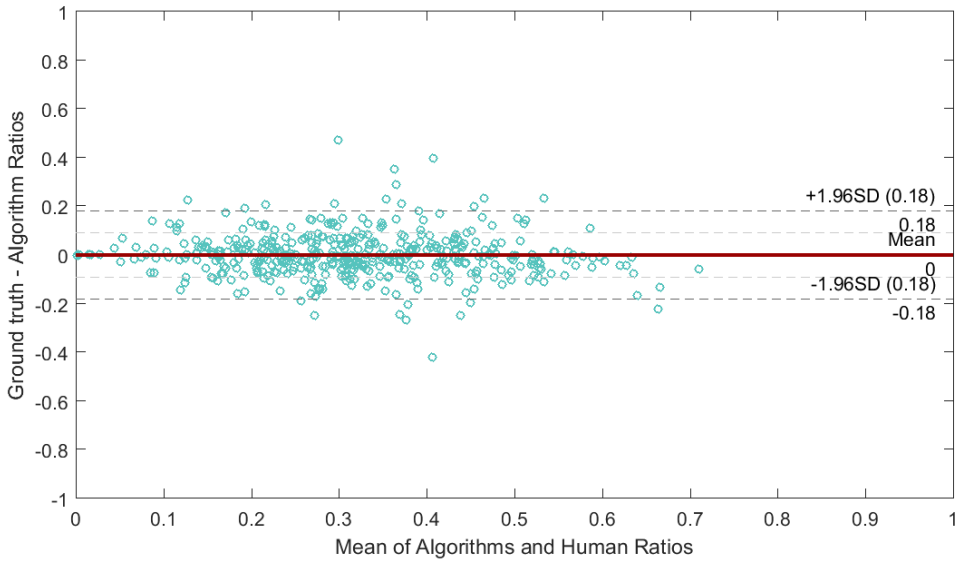


Figure 133 - Bland-Altman plot of Pathologist and Combined Contextual Analysis algorithm-generated TSRs per case, using the CR07 dataset

Distribution has a mean bias of 0, with upper and lower limits of agreement of 0.18 and -0.18 respectively (±0.18). This distribution shows that the algorithm consistency is slightly higher at lower levels of TSR, with more outliers in the distribution above 0.3 on the x axis.

The Bland-Altman plot in Figure 133 shows that the data has a bias of 0, and that the 95% confidence intervals (1.96 standard deviation, indicated by the outer dashed lines) are ±0.18.

7.3.3.6 Algorithm performance on QC dataset

Algorithm K ran on the reduced dataset that had been automatically approved for analysis by the automatic QC algorithm. As a result, the algorithm agreement statistics improved minimally, with per-patch agreement increasing from 79.17% to 79.56%, and grouped agreement increasing from 85.72% to 86.08%. A paired samples T-test confirmed that the distributions were not statistically different to the dataset where QC was not applied ($p = 0.15$ and $p= 0.29$ for per-patch and grouped agreement methods respectively), and Figure 134 visualises this in the respective boxplots. Because of these findings, further analyses are not presented in this section, but can be observed in Appendix D.11.

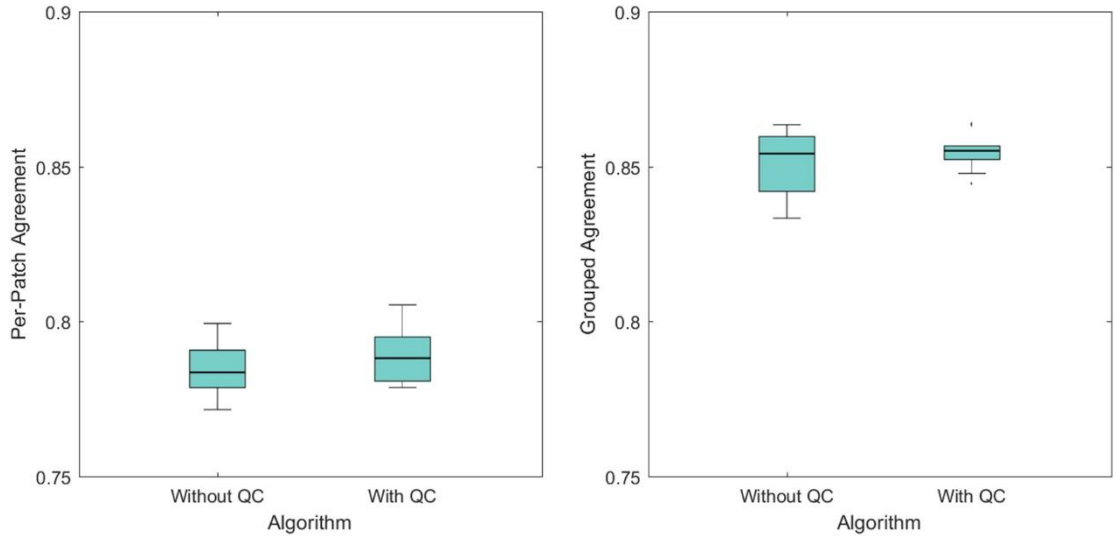


Figure 134 – Boxplots comparing Algorithm J and K agreement (with and without automatic QC)

Left: Per-patch agreement statistics for the algorithm without QC (mean = 0.7917) and with QC (mean = 0.7956). Mann-Whitney test indicates no statistically significant differences between groups (P = 0.13). Right: Per-patch agreement statistics for the algorithm without QC (mean = 0.8572) and with QC (mean = 0.8608). Mann-Whitney test indicates no statistically significant differences between groups (P = 0.38). The boxplots indicate that applying QC to the dataset does not improve the accuracy results enough to justify removing the slides and discarding their data.

The boxplots in Figure 134 illustrate that the statistical difference between the two algorithms in both per-patch and grouped agreement datasets is not significant. However, the distribution for Algorithm K (with QC) exhibits reduced variance when grouping the tumour and stroma sub classes, with an inter quartile range of 0.0044, compared to 0.0177 – a reduction of 75%.

7.3.3.7 Summary of algorithm development

To help summarise the progression of iterative algorithm development presented in this thesis, Figure 135 displays the pathologist agreement statistics from the experiment in 4.3, compared to the cross-validation accuracy results when grouping the eight tissue classes in to the parent tumour or stroma classes. There is a clear progression indicating that algorithm accuracy improves with the alterations applied.

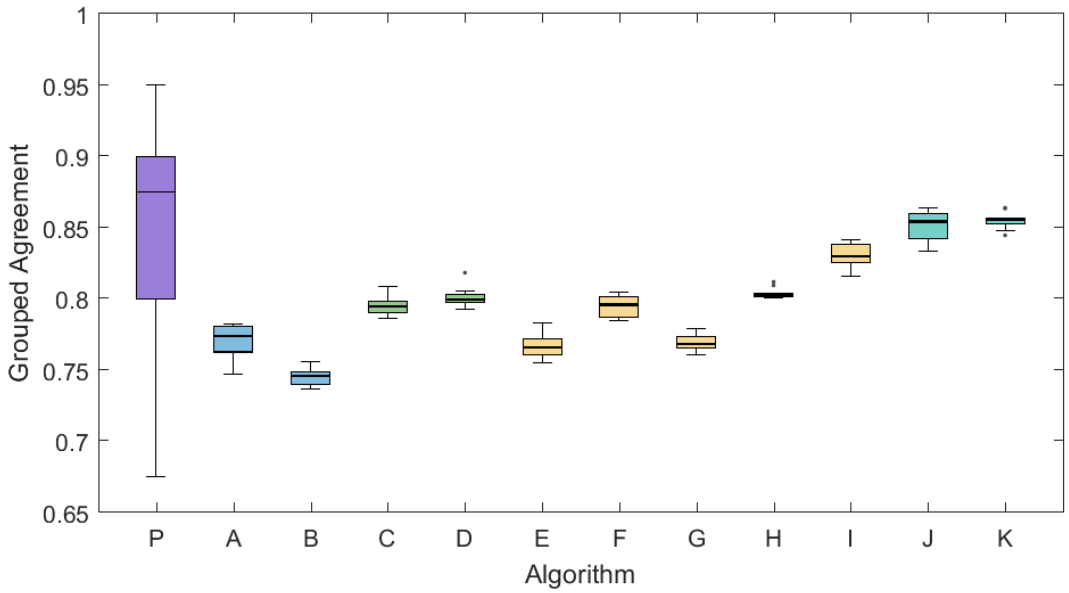


Figure 135 – Boxplots of pathologist-pathologist and pathologist-algorithm agreement results, grouped into tumour and stroma parent classes

Algorithms A-K can be identified using the name columns in Table 14. The boxplot labelled P is for pathologist agreement, which was generated from the best-case performance from the image size experiment in section 4.3. The boxplots show gradual improvements in accuracy over the development of the algorithms, with the application to the CR07 dataset yielding the highest results. Mean accuracy rates are identical between pathologist and Algorithm K (both 0.85), however human scoring is higher when using median accuracy (shown in the box plots), with agreement of 0.88 compared to 0.85. The spread of the distribution however makes the algorithm a more consistent evaluator.

For algorithm accuracy results, a Friedman's (repeated measures) ANOVA was conducted with algorithm type (total = 11) as the repeated measures independent variable. Results revealed a significant effect of algorithm methodology ($\chi 2 = 19.29$, $p < 0.04$). Post-hoc analysis of pairwise comparisons using the Mann-Whitney test can be found in the cropped confusion matrices in Appendix E. The significance testing reveals that only Algorithms B is statistically significantly different to all other algorithms. Algorithms J and K are significantly different compared to all other algorithms, but are not significantly different to each other. This means that using the CR07 dataset significantly improves algorithm performance, but applying the QC algorithm to the CR07 set does not. These findings confirm the observation that the CR07 dataset is of better visual quality, and that optimised datasets yield higher agreements statistics. The mean difference of TSR generated by pathologist and algorithm shows a similar trend of improvement. Figure 136 displays mean TSR difference with standard deviation and confidence intervals for all algorithms presented.

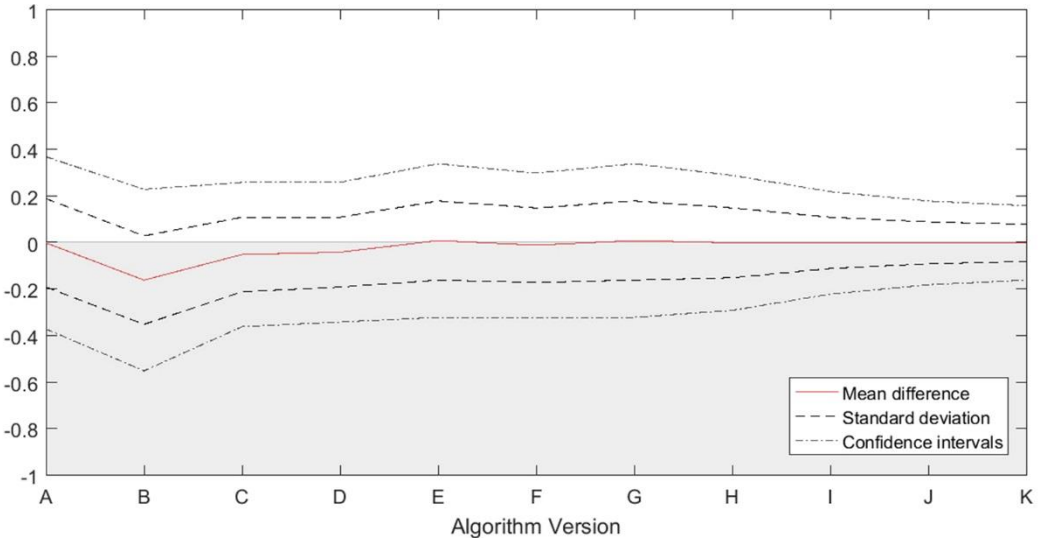


Figure 136 - Line plot of mean difference between pathologist and algorithm-generated TSRs

The plot shows the mean difference (red line) of pathologist and algorithm-generated TSRs, which progressively improves (gets closer to zero) over the algorithm iterations (x axis). The dashed line shows the width of the standard deviation of the difference distribution and the dotted line shows the confidence intervals that represent 95% of the data. The distributions become more tightly aligned in the later contextual algorithms, and algorithms on higher quality datasets.

Most notably, the standard deviation and confidence intervals of the TSR difference distributions become progressively narrower over iterations, meaning that the improvements made to the algorithm yield more tightly aligned distributions of agreement. The improvement of this distribution becomes more apparent starting with the combination of both local and global context, and improved further with QC and the higher quality CR07 dataset. It should be noted that all algorithms except Algorithm A use a patch size of 256x256 pixels, which without context (Algorithm B) performs the poorest. It is the contextual analysis that allows analysis of the 265x256 pixel patches to surpass the performance of the 64x64 pixel algorithm.

The progression of algorithm development and improved agreement means that the final algorithm performance is the most similar to human scoring, and may be appropriate for testing against survival data.

7.3.4 Conclusions

Algorithm J yields the highest accuracy rate of any results set presented so far (79% per patch and 86% grouped by tumour and stroma parent classes). The improvement in accuracy is reflected in the increase in AUC for all tissue subtypes (mean AUC = 0.94), the correlation (R^2

= 0.67) and TSR differences (mean = 0, SD = 0.09), and the Bland-Altman statistics (mean = 0, upper and lower limits of agreement = ±0.18).

The improvement over the results from the same algorithm applied to the QUASAR dataset (even with QC filtering applied) indicate that the CR07 dataset is more appropriate for image analysis. From undocumented visual inspection of the digital slides, it is apparent (but not conclusive) that ML is improved due to image quality rather than accuracy of pathologist data labelling. The QC algorithm has a higher pass rate on the CR07 dataset, with 77% of the CR07 dataset (344 cases) being accepted, compared to 32% of the QUASAR dataset (2,211 cases). A Chi Squared test shows that the proportions are significantly different ($p < 0.01$) indicating that this observation is valid.

The application of QC to the dataset required retraining of the algorithm on whole tumours rather than 3x3mm boxes over areas of highest TCD. Once retrained, the QC algorithm rejected 23% of the slides, but there was no manual scoring to validate these results. The remaining dataset yielded a slight, but not statistically significant increase in accuracy. The slight increase in accuracy was not a worthwhile trade-off, when considering the loss of data from removing the suboptimal slides. Therefore, it was decided that the smaller QC'd dataset would not be used when applying the algorithm to survival data.

7.4 Survival analysis on CR07 dataset

7.4.1 Aim

To use the TSRs generated by Algorithm J on the CR07 dataset, in order to stratify patients into groups for survival analysis. The work should conclude by assessing whether the automated solution is a feasible alternative to manual scoring.

7.4.2 Methods

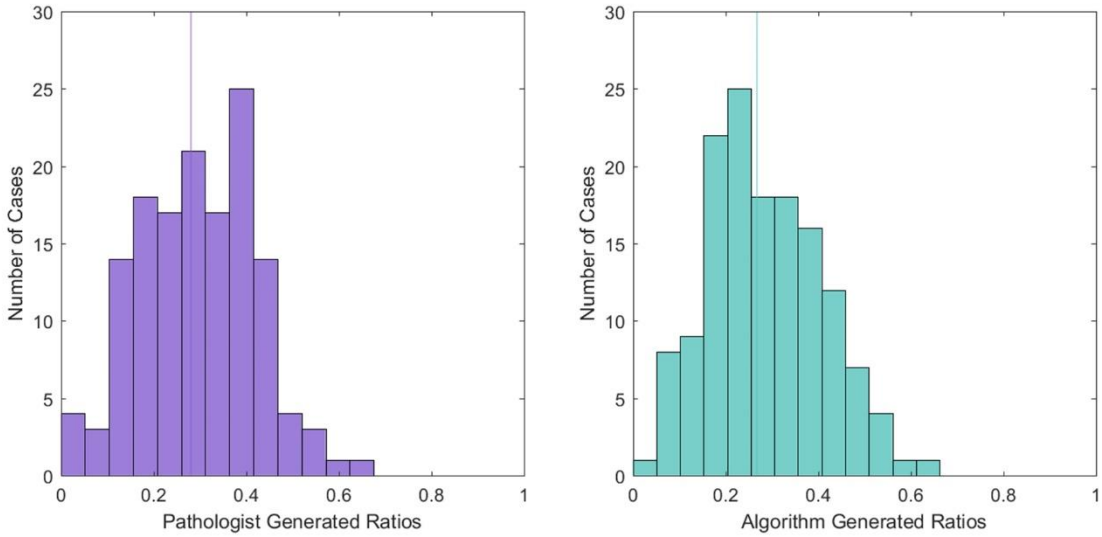
Using the TSRs generated by Algorithm J on the CR07 dataset (presented in 7.3.3), the data were prepared for analysis by identifying cases deemed appropriate for analysis. The preparation methodology is presented in the next section.

7.4.2.1 Case selection

Of the 430 TSRs generated, 314 had unique case numbers, which was due to the slides being sampled using different methods (see 7.2.3). One of the unique 314 cases was not scored using the whole tumour annotation method, and was discarded, leaving 313 cases with TSRs and survival data. Out of the remaining 313 cases, 142 were from the arm of the study that were given preoperative SCRT (arm 1), and 171 were from the arm that were given selective postoperative CRT (arm 2). As discussed in 7.2.2, survival analysis was applied separately to individual trial arm groups.

Distributions of TSRs for both trial arms, using both pathologist and algorithm methodologies are presented in Figure 137 for trial arm 1 and trial arm 2. Note that as with 7.2 and 7.3, TSRs were calculated using Method 2 detailed in section 3.3.2.4, as opposed to Method 1 which was used in the QUASAR trial. This method yielded lower ratios, and so the histograms show average values closer to 0, when compared to the other ratio calculation method.

Trial arm 1 (preoperative SCRT)



Trial arm 2 (selective postoperative CRT)

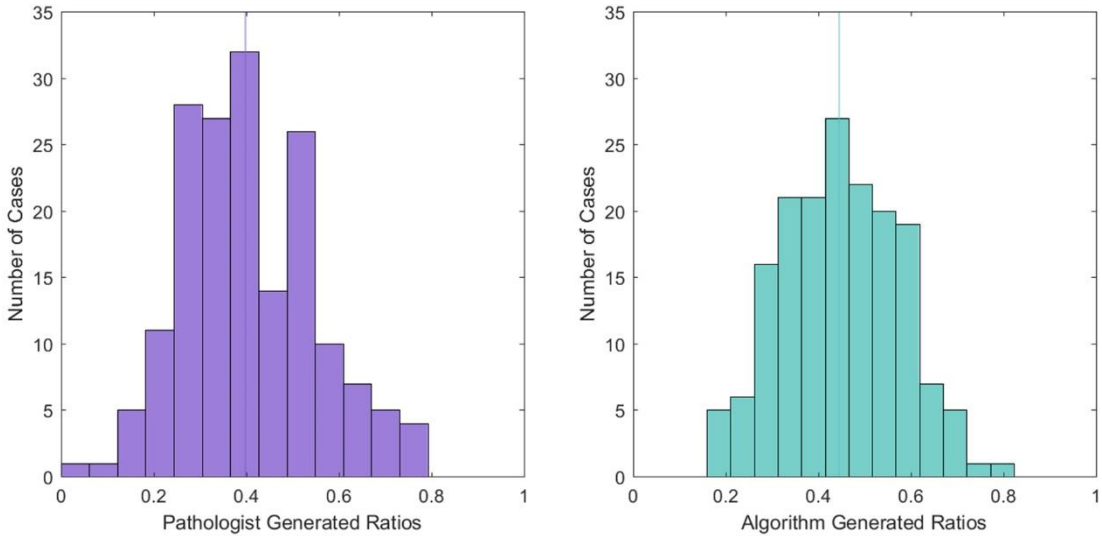


Figure 137 - Distribution of CR07 TSRs for generated by algorithm and pathologist, for trial arm 1 (top and trial arm 2 (bottom)

Left: Arm 1 pathologist ratios. Mean = 0.29, median = 0.28, SD = 0.13.
Right Arm 1 algorithm ratios. Mean = 0.29, median = 0.27, SD = 0.13.
Bottom Left: Arm 2 pathologist ratios. Mean = 0.41, median = 0.40, SD = 0.14.
Bottom Right Arm 2 algorithm ratios. Mean = 0.45, median = 0.45, SD = 0.13.
The distributions show that TSRs are higher when analysing cases with preoperative SCRT, and that the TSRs generated are comparatively lower than when analysis patients without preoperative SCRT. Both these observations are represented in the pathologist and algorithm-generated TSRs.

Both sets of distributions show that algorithm and pathologist-generated ratios are comparable when analysing trial arms individually, with the algorithm ratios more closely aligned to the ground truth on patients with preoperative SCRT. The distributions show that cases with

preoperative SCRT exhibit lower ratios, which is indicative of tumour cells being destroyed by radiation therapy.

The case survival data available consisted of the following fields: the number of days to the final follow-up date; the state of the patient at the final follow-up date (alive, died of cancer, or died of other causes); whether the patient's cancer had progressed (metastasised); and the number of days since starting the trial that the progression had been identified. Table 26 shows the number and proportion of patient statuses at the time of their final follow-up.

Trial arm	Alive	Cancer death	Other death	Total	Progression
1	86 (61%)	17 (12%)	39 (27%)	142	6 (4%)
2	96 (56%)	30 (18%)	45 (26%)	171	25 (15%)

Table 26 – Survival statistics for patients in both arms of CR07 trial

Cancer stage and grade was not included in the dataset, neither was recurrence (other research has shown that TCD is a good predictor of recurrence, but that data was not available for this analysis work [52]).

7.4.2.2 Patient stratification

Prior to survival analysis, cases were stratified into groups, based on a threshold of TSR, referred to here as a cut-off. Initially a test to assess the linearity of the TSR results was established, based on existing work using the RandomSpot system [48]. Cases were sorted into ascending TSR order, and split into equally sized groups (which may or may not have equal numeric intervals between TSR cut-offs), and their survival curves plotted. This method was repeated for stratifying into two, three and five equal groups. By splitting the data into this format, observations could be made to see if there is any clear proportionate threshold that differentiates survival rates, based on TSR.

The main goal of the analysis however was to use the same methodology as the previous CR07 SRS study, by splitting the cases into two groups: TSR High and TSR Low. The groups were divided based on a pre-identified cut-off value, which was calculated per dataset.

The method for finding appropriate cut-offs for stratification of patients was presented in the original work that used the CR07 dataset to identify TSR as a prognostic marker [6]. Using an iterative cut-off value of 0.01 to 1 in increments of 0.01, sensitivity and specificity was calculated based on the correct prediction of survival between groups less than the cut-off value (TSR low), and greater than or equal to the cut-off value (TSR high). For the analysis, the TSR low group was the 'at-risk' group. A modified ROC curve was created using the cut-off value as

the x-axis and the sensitivity plus the specificity as the y-axis. The cut-off value with the highest sensitivity plus specificity was deemed the most appropriate. Examples of these curves can be found in Appendix F.

7.4.3 Results

Results are summarised in Table 27.

7.4.3.1 Group linearity

The results in this section use the survival statistics, discussed in 7.4.2.1, split into either two, three or five equal sized groups, so that observations can be made based on the linearity of the distributions, in relation to patient survival. Kaplan Meier survival curves (KM) [242] for both cancer specific survival and overall survival, for two equal sized groups ordered by TSR, are presented in Figure 138 for trial arm 1 and Figure 139 for trial arm 2. The legends on each of the graphs show that the cut off points for stratifying the patients into groups.

Trial arm 1 (preoperative SCRT)

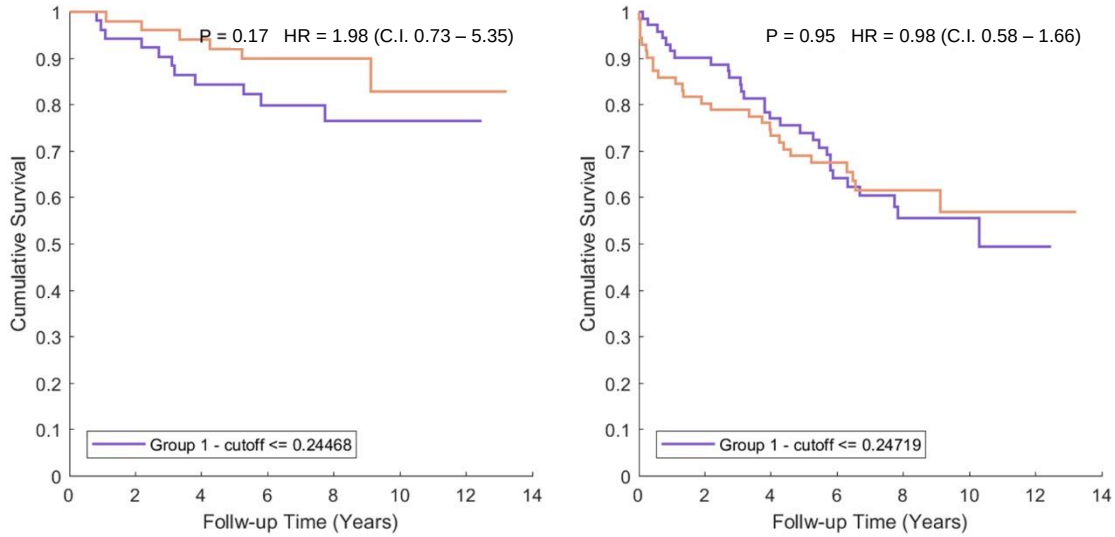


Figure 138 – KM curves for arm 1 cases stratified into equally sized groups using ordered TSR generated by Algorithm J

Cases were ordered by ascending TSR, and grouped into equal sized groups. The legends on the graphs show the upper cut-offs for TSR low group.
Left: Cancer specific survival
Right Overall survival
Note that Group 1 is the equivalent of the TSR low group when using 2 groups only. The graphs show that stratification of groups using the median value does not create significant differences in survival.

A log-rank test for comparing the distributions of the two-group method showed no significant differences between groups for both cancer and overall survival in arm 1 ($p = 0.17$ and $p= 0.95$ respectively) or arm 2 ($p = 0.44$ and $p = 0.77$ respectively).

Trial arm 2 (selective postoperative CRT)

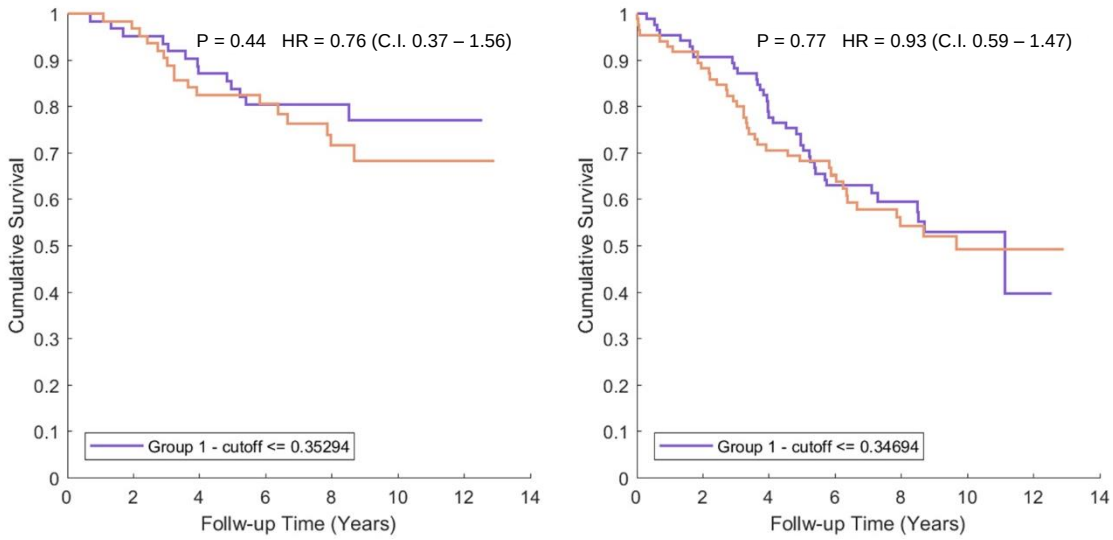


Figure 139 – KM curves for arm 2 cases stratified into equally sized groups using ordered TSR generated by Algorithm J

Cases were ordered by ascending TSR, and grouped into equal sized groups. The legends on the graphs show the upper cut-offs for TSR low group.
Left: Cancer specific survival
Right Overall survival
Note that Group 1 is the equivalent of the TSR low group when using 2 groups only. The graphs show that stratification of groups using the median value does not create significant differences in survival.

Analysis stratifying the patients into three and five equal group sizes, ordered by TSR was also attempted. These results are presented in Appendix F.1 for arm1 and Appendix F.2 for arm 2. Both the three and five-group methodology also showed no significant test values, when comparing all groups independently. These values are displayed for clarity in a cropped confusion matrix, with description, in the respective Appendix section.

The grouping of patients into equal sizes using ordered TSRs shows no significant differences between groups.

7.4.3.2 Survival using modified ROC curves for cut-off

The results in this section use the survival statistics, discussed in 7.4.2.1, split into two groups, TSR high and TSR low, where previously published studies showed that the TSR high group should correlate with higher survival rates. It should be noted that cut-offs closer to 0 will create smaller groups of TSR low cases, and larger cut off values will create larger groups of TSR low cases.

For both pathologist and machine generated ratios, TSR high and low groups were created using the modified ROC curves (see section 7.4.2.2) to identify the most appropriate cut-off point.

Subsequently a KM curve was generated for each group, using the number of days passed between the trial beginning and the final follow-up session occurred, for each patient. The status of the patient at the final follow-up was recorded as alive, cancer death or other cause of death. Figure 140 and Figure 141 show the survival curves for cancer specific survival (patient either alive or died of cancer at final follow-up) for arms 1 and 2 respectively. Figure 142 and Figure 143 uses cases from all three categories to show overall survival for trial arms 1 and 2 respectively.

Cancer specific survival on trial arm 1 (preoperative SCRT)

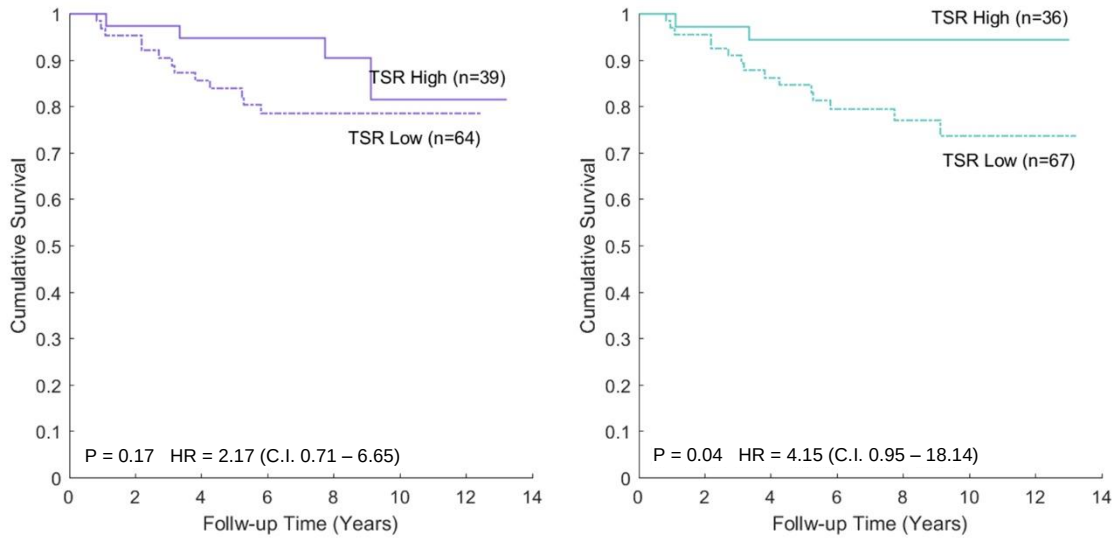


Figure 140 - KM survival curves for cancer specific survival on arm 1 using pathologist and machine generated TSRs

Left: KM survival curves for pathologist-generated ratios, thresholded by modified ROC cut-off (0.26). Log-rank test had a significance value of 0.17 comparing the TSR High group (n = 39) to TSR Low group (n = 64), with a Hazard Ratio of 2.17 (confidence intervals 0.71 – 6.65).

Right: KM survival curves for algorithm-generated ratios, thresholded by modified ROC cut-off (0.32). Log-rank test had a significance value of 0.04 comparing the TSR High group (n = 36) to TSR Low group (n = 67), with a Hazard Ratio of 4.15 (confidence intervals 0.95 – 18.14).
Note that modified ROC curves used for generating cut-offs can be found in Appendix F.
The curves show that algorithm-generated TSRs using a dynamic cut-off creates a significant stratification of patient groupings that predicts cancer-specific survival for patients with preoperative SCRT, whereas using the same methodology for pathologist-generated ratios does not.

The log-rank test of the groups, split using pathologist-generated ratios, accepts the null hypothesis that the two survival distributions are not significantly different (p = 0.17). TSR high and low groups numbered 39 and 64 respectively, with the TSR low group yielding a Hazard Ratio (HR) of 2.17 (C.I. = 0.71 – 6.65). The distribution of the groups using algorithm-generated ratios rejects the log-rank test null hypothesis, with a smaller significance value than

pathologist assessment (p = 0.04). HR for the TSR low group was 4.15 (C.I. = 0.95 – 18.14), with a TSR high group of 36 and TSR low group of 67 in size.

Cancer specific survival on trial arm 2 (selective postoperative CRT)

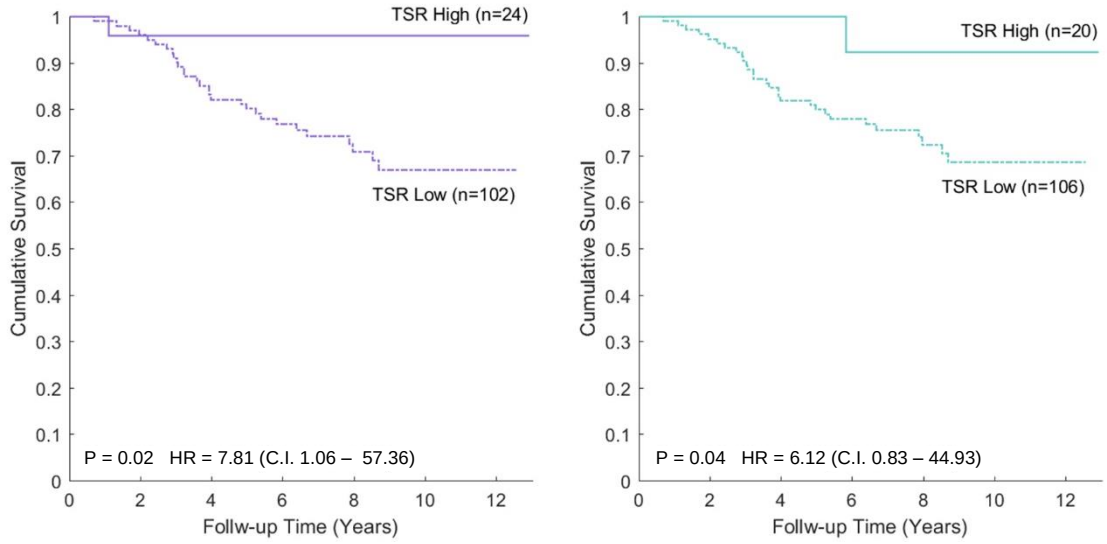


Figure 141 –KM survival curves for cancer specific survival on arm 2 using pathologist and machine generated TSRs

Left: KM survival curves for pathologist-generated ratios, thresholded by modified ROC cut-off (0.47). Log-rank test had a significance value of 0.02 comparing the TSR High group (n = 24) to TSR Low group (n = 102), with a Hazard Ratio of 7.81 (confidence intervals 1.06 – 57.36).

Right: KM survival curves for algorithm-generated ratios, thresholded by modified ROC cut-off (0.47). Log-rank test had a significance value of 0.04 comparing the TSR High group (n = 20) to TSR Low group (n =106), with a Hazard Ratio of 6.12 (confidence intervals 0.83 – 44.93).
Note that modified ROC curves used for generating cut-offs can be found in Appendix F.
The curves show that algorithm-generated TSRs using a dynamic cut-off creates a significant stratification of patient groupings that predicts cancer-specific survival for patients with selective postoperative CRT. However, using the same methodology for pathologist-generated ratios creates a more significant stratification (p = 0.02 compared to 0.04).

The log-rank test of the groups, split using pathologist-generated ratios, rejects the null hypothesis that the two survival distributions are not significantly different (p = 0.02). TSR high and low groups numbered 24 and 102 respectively, with the TSR low group yielding a Hazard Ratio (HR) of 7.81 (C.I. = 1.06 – 57.36). The distribution of the groups using algorithm-generated ratios accepts the log-rank test null hypothesis, with a larger significance value than pathologist assessment (p = 0.04). HR for the TSR low group was 6.12 (C.I. = 0.83 – 44.93), with a TSR high group of 20 and TSR low group of 106 in size.

Overall survival on trial arm 1 (preoperative SCRT)

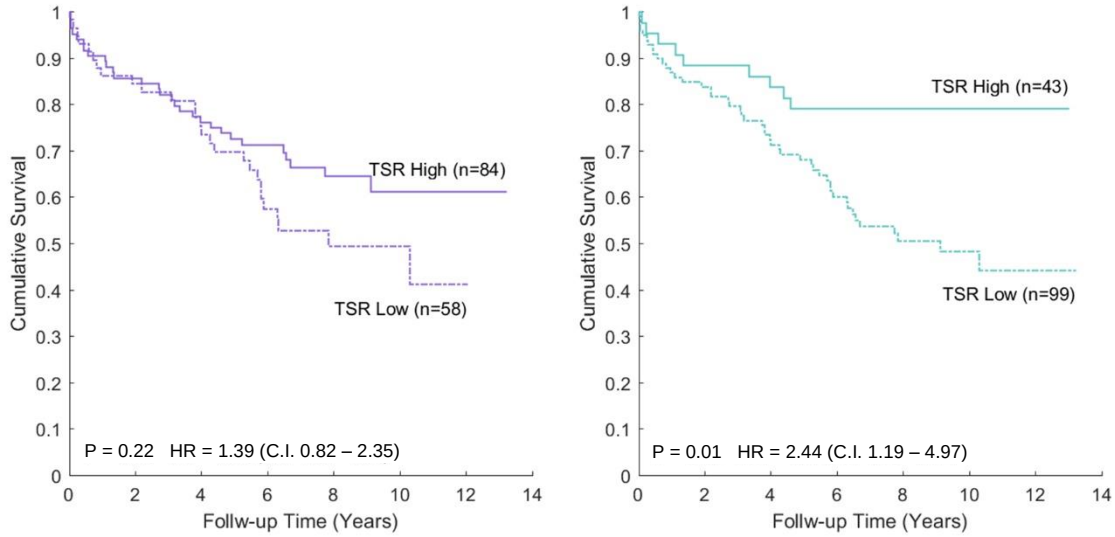


Figure 142 - KM survival curves for overall survival on arm 1 using pathologist and machine generated TSRs

Left: KM survival curves for pathologist-generated ratios, thresholded by modified ROC cut-off (0.20). Log-rank test had a significance value of 0.22 comparing the TSR High group (n = 84) to TSR Low group (n = 58), with a Hazard Ratio of 1.39 (confidence intervals 0.82 – 2.35).

Right: KM survival curves for algorithm-generated ratios, thresholded by modified ROC cut-off (0.32). Log-rank test had a significance value of 0.01 comparing the TSR High group (n = 43) to TSR Low group (n = 99), with a Hazard Ratio of 2.44 (confidence intervals 1.19 – 4.97).
Note that modified ROC curves used for generating cut-offs can be found in Appendix F.
The curves show that algorithm-generated TSRs using a dynamic cut-off creates a significant stratification of patient groupings that predicts overall survival for patients with preoperative SCRT, whereas using the same methodology for pathologist-generated ratios does not.

The log-rank test of the groups, split using pathologist-generated ratios, accepts the null hypothesis that the two survival distributions are not significantly different (p = 0.22). TSR high and low groups numbered 84 and 58 respectively, with the TSR low group yielding a HR of 1.39 (C.I. = 0.82 – 2.35). The distribution of the groups using algorithm-generated ratios rejects the log-rank test null hypothesis, with a smaller significance value (p = 0.01) than algorithm generated TSRs for cancer specific survival. The Hazard Ratio for the TSR low group was 2.44 (C.I. = 1.19 – 4.97), with a TSR high group of 43 and TSR low group of 99 in size.

Pathologist analysis is outperformed by the automated solution on trial arm 1 patients, which may be due to algorithm assessment being more consistent, especially across more difficult cases where tissue has been affected by radiation therapy. However, the group stratification for both methods yields different proportions of TSR high and low group sizes, due to the cut-off levels being identified at 0.20 for pathologist ratios and 0.32 for algorithm ratios.

Overall survival on trial arm 2 (selective postoperative CRT)

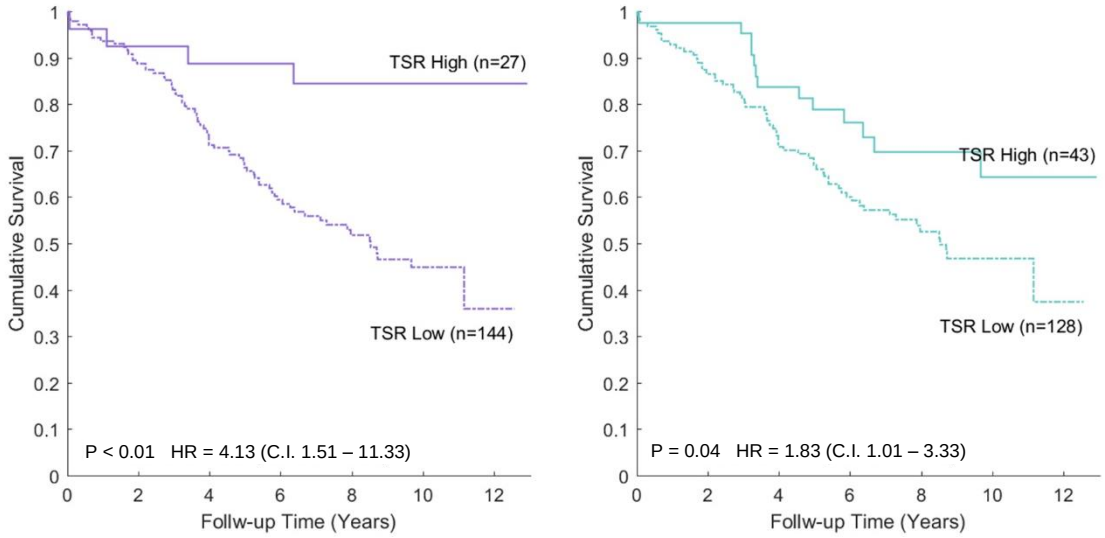


Figure 143 - KM survival curves for overall survival on 2 using pathologist and machine generated TSRs

Left: KM survival curves for pathologist-generated ratios, thresholded by modified ROC cut-off (0.47). Log-rank test had a significance value of < 0.01 comparing the TSR High group (n = 27) to TSR Low group (n = 144), with a Hazard Ratio of 4.13 (confidence intervals 1.51 – 11.33).

Right: KM survival curves for algorithm-generated ratios, thresholded by modified ROC cut-off (0.43). Log-rank test had a significance value of 0.04 comparing the TSR High group (n = 43) to TSR Low group (n = 128), with a Hazard Ratio of 1.83 (confidence intervals 1.01 – 3.33).
Note that modified ROC curves used for generating cut-offs can be found in Appendix F.
The curves show that algorithm-generated TSRs using a dynamic cut-off creates a significant stratification of patient groupings that predicts overall survival for patients with selective postoperative CRT. However, using the same methodology for pathologist-generated ratios creates a more significant stratification (p < 0.01 compared to 0.04).

The log-rank test of the groups, split using pathologist-generated ratios, rejects the null hypothesis that the two survival distributions are not significantly different ($p < 0.01$). TSR high and low groups numbered 27 and 144 respectively, with the TSR low group yielding a HR of 4.13 (C.I. = 1.51 – 11.33). The distribution of the groups using algorithm-generated ratios also rejects the log-rank test null hypothesis, with a significance value of 0.04. The HR for the TSR low group was 1.83 (C.I. = 1.01 – 3.33), with a TSR high group of 43 and TSR low group of 128 in size.

Algorithm-generated TSR shows a statistically significant stratification of patient groups on trial arm 2, yet underperforms compared to human analysis. This may be due the cut-off level identified, which has created a slightly more balanced set of TSR high and TSR low cases (but is still imbalanced). Using a static threshold of 0.74 for both human and machine results may show improved algorithm performance.

Using the same methodology as 7.4.3.2, progression was analysed instead of survival, creating modified ROC curves to identify appropriate group cut-offs and inverse KM plots (Figure 144 for arm 1 and Figure 145 for arm 2).

Cancer progression on trial arm 1 (preoperative SCRT)

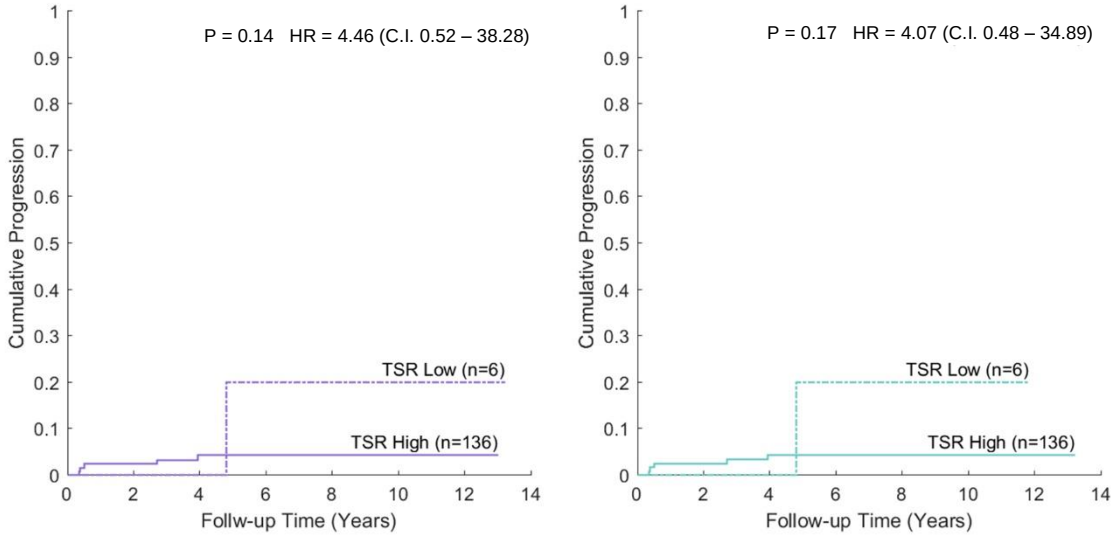


Figure 144- Progression curves for arm 1 using pathologist and machine generated TSRs

Left: Pathologist-generated ratios thresholded by the modified-ROC modified ROC cut-off (0.42). Los-rank test had a significance value of 0.14 comparing the TSR High group (n = 136) to TSR Low group (n = 6), with a Hazard Ratio of 4.46 (C.I. = 0.52 – 38.28).

Right: Algorithm-generated ratios thresholded by the modified-ROC modified ROC cut-off (0.44). Log-rank test had a significance value of 0.17 comparing the TSR High group (n = 136) to TSR Low group (n = 6), with a Hazard Ratio of 4.07 (C.I. = 0.48 – 34.89).
The graphs show that neither pathologist nor algorithm TSRs were able to generate preoperative SCRT patient groupings that predicted progression with significance. The lack of events in the dataset may contribute to this.

The log-rank test of the groups split using pathologist-generated ratios accepts the null hypothesis that the two survival distributions are not significantly different (p = 0.14). Group sizes for high and low TSR were 136 and 6 respectively, and TSR low group had a HR of 4.46 (C.I. = 0.52 – 38.28). The distribution of the groups using algorithm-generated ratios is similar, with a significance value that also accepts the null hypothesis that the two curves are not significantly different (p = 0.17), equally sized TSR high and low groups, and a HR of 4.07 (C.I. = 0.48 – 34.89).

Only six of the 142 cases in the dataset had cancers that progressed, and the algorithm and pathologist methods both identified different cut-offs which created a TSR low group of only six patients. Both methods contained the same single patient (out of the six in the dataset) that had progressed, hence the identical curve step size and time. The log-rank significance values

are different due to the differences in the other patients in the group (all of which were alive at their final follow-up, but were at different times), which can be seen by the length of the curves.

Cancer progression on trial arm 2 (selective postoperative CRT)

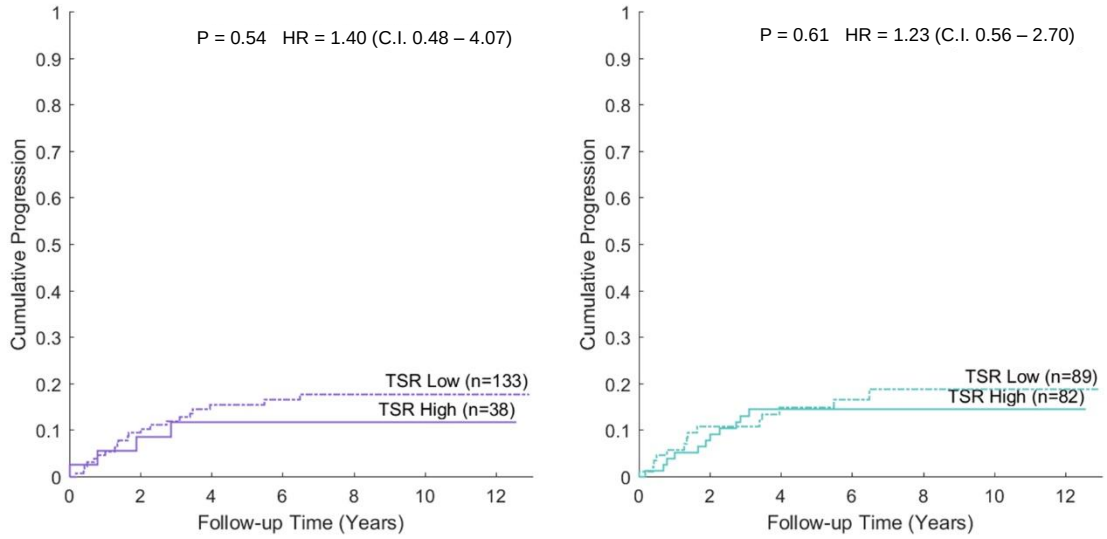


Figure 145 - Progression curves for arm 2 using pathologist and machine generated TSRs

Left: Pathologist-generated ratios thresholded by the modified-ROC modified ROC cut-off (0.27). Log-rank test had a significance value of 0.54 comparing the TSR High group (n = 38) to TSR Low group (n = 133), with a Hazard Ratio of 1.40 (C.I. = 0.48 – 4.07).

Right: Algorithm-generated ratios thresholded by the modified-ROC modified ROC cut-off (0.34). Log-rank test had a significance value of 0.61 comparing the TSR High group (n = 89) to TSR Low group (n = 82), with a Hazard Ratio of 1.23 (C.I. = 0.56 – 2.70).
The graphs show that neither pathologist nor algorithm TSRs were able to generate selective postoperative CRT patient groupings that predicted progression with significance.

The log-rank test of the groups split using pathologist-generated ratios accepts the null hypothesis that the two survival distributions are not significantly different (p = 0.54). Group sizes for high and low TSR were 38 and 133 respectively, and TSR low group had a HR of 1.40 (C.I. = 0.48 – 4.07). The distribution of the groups using algorithm-generated ratios also accepts the null hypothesis, with a significance value of 0.61. Algorithm-stratified group sizes were 82 for TSR high and 89 for TSR low, with a TSR low group HR of 1.23 (C.I. = 0.56 – 2.70).

Trial arm 2 contains 25 cases that progressed, out of 171 in total (15%). The progression curves for trial arm 2 show less extreme step sizes due to the proportional increase in number of cases that progressed over trial arm 1.

7.4.3.3 Summary of survival results

Table 27 presents the survival statistics for the different trial arms, methods and analysis types. The algorithm-generated TSRs outperform pathologist-generated TSRs on trial arm 1, but not trial arm 2.

Arm	Method	Analysis type	Cut-off	n TSR high	n TSR low	Log-rank P	HR	C.I.
1	Linear Split	Cancer specific	0.24	51	52	0.17	1.98	0.73 - 5.35
1	Algorithm	Cancer specific	0.32	36	67	0.04	4.15	0.95 – 18.14
1	Pathologist	Cancer specific	0.26	39	64	0.17	2.17	0.71 – 6.65
1	Linear Split	Overall	0.25	71	71	0.95	0.98	0.58 – 1.66
1	Algorithm	Overall	0.32	43	99	0.01	2.44	1.19 – 4.97
1	Pathologist	Overall	0.20	84	58	0.22	1.39	0.82 – 2.35
1	Algorithm	Progression	0.44	136	6	0.17	4.07	0.48 – 34.89
1	Pathologist	Progression	0.42	136	6	0.14	4.46	0.52 – 38.28
2	Linear Split	Cancer specific	0.35	63	63	0.44	0.76	0.37 – 1.56
2	Algorithm	Cancer specific	0.47	20	106	0.04	6.12	0.83 – 44.93
2	Pathologist	Cancer specific	0.47	24	102	0.02	7.81	1.06 – 57.36
2	Linear Split	Overall	0.35	135	136	0.77	0.93	0.59 – 1.47
2	Algorithm	Overall	0.43	27	144	0.04	1.83	1.01 – 3.33
2	Pathologist	Overall	0.47	43	128	< 0.01	4.13	1.51 – 11.33
2	Algorithm	Progression	0.34	82	89	0.61	1.23	0.56 – 2.70
2	Pathologist	Progression	0.27	38	133	0.54	1.40	0.48 – 4.07

Table 27 – All statistics from survival analysis

The table shows that algorithm-generated TSRs create statistically significant stratifications when predicting both cancer-specific and overall survival in both patients with preoperative SCRT, and selective postoperative CRT. The algorithm outperforms pathologist-generated TSRs on trial arm 1, but not trial arm 2. Neither methodology predicts progression with statistical significance, and applying a linear split to the data does not predict survival. Therefore, it can be concluded that algorithm-generated TSRs may be an appropriate substitute for human scoring.

The table shows the results for pathologist and machine methods of analysis for cancer specific survival, overall survival and progression, for both trial arms. The columns display statistics for the cut-off value (to stratify the patient groups), the numbers of cases in the high and low TSR

groups, the log-rank test significance value, and the hazard ratios (HR) for the TSR low groups. All of these results can be viewed in their respective figures throughout 7.4.3.

The modified ROC curve method for identifying appropriate TSR cut-off points shows that cut-offs generated from algorithm TSRs are higher than those generated from pathologist TSRs on patients with preoperative radiotherapy. These cut-offs are lower than those reported in the literature due to the original study presenting findings from the selective postoperative CRT patients only (arm 2). Cut-offs generated using TSRs generated from patients with selective postoperative CRT are identical to previously published results.

Algorithm-generated TSRs yield statistically significant differences between TSR high and low groups, when analysing patients with preoperative SCRT (log rank $p = 0.04$ for cancer-specific survival and $p = 0.01$ for overall survival). The HR confidence intervals for cancer-specific survival suggest that this significance is not replicated using a cox proportional hazards model (HR = 4.15, C.I. = 0.95 – 18.14), which may be due to the smaller number of cases compared to the overall survival dataset. In all cases for trial arm 1, the algorithm-generated-TSRs outperform pathologist-generated TSRs for survival prediction.

For patients with selective postoperative CRT, the algorithm-generated TSRs also generate statistically significant differences between TSR high and low groups for cancer-specific and overall survival ($p = 0.04$ for both). As with arm 1, the confidence intervals for cancer-specific survival suggest there is not enough data to show significance when analysing hazard ratios (HR = 6.12, C.I. = 0.83 – 44.93). For arm 2, pathologist-generated TSRs outperform algorithm-generated TSRs for survival prediction.

Progression was not significantly predicted by either pathologist or algorithm generated TSRs on either of the trial arms. However, both distributions have almost similar appearances, log-rank p-values and hazard ratios for patients with preoperative SCRT. For patients with selective postoperative CRT, there were also no significant differences between TSR high and low groups, showing log-rank significance values of 0.54 for pathologist-generated TSRs, and 0.61 for algorithm-generated TSRs.

7.4.4 Conclusions

The survival curves in this chapter show that algorithm-generated TSRs predict both cancer and overall survival with statistically significant differences between high and low TSR groups, on patients with preoperative SCRT ($p = 0.04$ for cancer specific survival and $p = 0.01$ for overall survival). Stratification of patients with selective postoperative CRT also showed significant

differences for both cancer specific survival and overall survival (both $p = 0.04$). However, HR confidence intervals suggest that algorithm predictions for cancer survival require more data to support the log-rank test results.

The algorithm outperforms manual analysis in terms of survival prediction for both cancer specific survival and overall survival on patients with preoperative SCRT, but not on patients with selective postoperative CRT.

For both methods of analysis, on both trial arms, there were no significant results for progression analysis.

Stratifying patients into two, three and five equal sized groups, when ordered by TSR did not predict survival for either trial arm.

The original study that generated the pathologist TSRs used 3x3mm annotation boxes for sampling at the area where TCD was perceived to be highest. This data was not available in the RandomSpotDB repository, but had more consistent correlation to survival, and so obtaining this dataset would be useful to confirm the algorithm's prognostic capabilities.

7.5 Discussion

The development and evaluation of Algorithms A to I were performed entirely on a single CRC dataset (QUASAR), and conclusions were made that validation could be improved by exposing the algorithm to more sets of real-world data. The MRC CR07 trial was chosen out of the datasets from RandomSpotDB to validate the algorithm across other CRC clinical trials. One of the reasons for this was that the anonymised follow-up data for the trial was held locally in Leeds and so survival analyses could be calculated. The original available data presented in this work (from the QUASAR trial) contained 1,898 more patient cases, and so had the potential for much more robust survival analysis. However, it became apparent that it was infeasible to retrieve survival analyses from the centralised (and protected) data centre, in a timescale appropriate to the project.

Due to the larger sample sizes used in the CR07 trial (300 spots compared to 50 per case), a similar number of labelled points existed in each data set (106,242 for QUASAR and 116,591 for CR07). This meant that despite the lower number of patients for survival data, algorithm validation using per-patch agreement was comparable. The distribution of tissue classes in the CR07 dataset had a higher proportion of stroma than the QUASAR dataset. The original paper analysing these tumours reports that the cases had an overall poorer prognosis, which is consistent with the literature, and suggests patients with low TSRs (proportionally higher levels of intra-tumoural stroma, or a lower TCD) have poorer response to therapy.

By analysing pathologist-algorithm agreement on a patch-by-patch basis, Algorithm J exhibits an 86% accuracy rate, which is 3% higher than on the QUASAR dataset, after manual optimisation using the QC process from 6.3, and 6% higher than the full dataset, without QC applied (Algorithm I). Application of the QC algorithm (modified for use on whole tumour annotations) reduced the size of the dataset of spot features reduced by 31%, from 116,591 to 80,493, and showed no significant differences in accuracy over the original dataset, so was not used for further analysis. This is likely due to the dataset having less stain variation and overall better-quality slides (where quality is based on the criteria described in Chapter 6). The algorithm has a higher number of false negatives (predicted stroma when tumour) than false positives (by 2.53%), which is likely due to the balance of classes within the dataset, meaning that stroma is more highly represented. The higher number of false negatives on the CR07 set is

lower than the number of both false positives and negatives when performed on the QUASAR set. The distribution of TSR differences is also improved on the CR07 dataset, with a mean and median difference of 0, and a standard deviation of 0.09. It should be noted that the method for calculating TSR on the CR07 dataset was different to the QUASAR dataset (Method 2 and Method 1 respectively from Table 7), which meant that TSRs were lower (closer to 0 before calculating difference) in the CR07 dataset. Without further analysis of the dataset, it is unclear whether image quality or the expert-labelling is the leading contributing factor to the overall improvement in performance.

With an improved algorithm generating results that correlated well with pathologist-generated TSRs, Algorithm J was predicted to generate equally well correlated survival statistics. The original West et al study used 3x3mm box annotations for sampling at the area of highest TCD, to show prognostic significance. Unfortunately, the largest set of sampling co-ordinates available for the CR07 dataset in the RandomSpotDB was on polygonal whole tumour regions of interest (313 unique cases compared to 76). This meant that the TSRs being analysed in this study were lower (closer to 0), which compressed the data (especially on patients with preoperative SCRT), meaning that the automatically generated cut-offs (for group stratification) were more sensitive to variation in the dataset than previous studies.

A point perhaps not important to this research, but worthy of consideration is that by using whole annotations, the ratio of tumour to stroma may be indicative of cancers that are more advanced, and therefore more infiltrative, since they will likely have a higher proportion of stroma on the invasive front. This could mean that the TSR is a surrogate for cancer staging (see 2.2), rather than being an independent prognostic marker. However, recent research has shown that there is no association between TSR and cancer stage [52].

The method for automatic grouping of patients used modified ROC curves proposed in the original study. This method created a clear peak for selecting the appropriate cut-off point, and examples of curves can be found in Appendix F. The cut-offs generated created TSR high and low group sizes that were more balanced in patients with preoperative SCRT than patients with selective postoperative CRT. This may be due to the more consistent appearance of tissue that has gone through radiotherapy (containing less tumour overall), affecting TSR. It is beyond the scope of this work (and the author's knowledge) to pathologically interpret these results.

The proportion of group sizes presented in the original study is also imbalanced, having a higher number of TSR high (n = 110) to TSR low (n = 35), with a cut-off of 0.47. This cut-off was replicated in the results for selective postoperative CRT, but not preoperative SCRT, which was lower for all methods. This shows consistency between studies (despite differences in sampling

method), where results presented in the original study are for postoperative CRT only. It is expected that cut-off values would be lower for patients in trial arm 1, where preoperative SCRT has reduced the amount of tumour in the tissue.

The method for generating cut-offs based on the modified ROC curve (plotting sensitivity and specificity of survival prediction) effectively means that the methodology engineers the most appropriate threshold for the stratification task. This is not appropriate when trying to develop a tool for generating a prognostic marker for use with any CRC dataset. Most of the literature reports using a cut-off of 0.5. This is due to the visual inspection task being estimated into two, four, five or ten bins, and so fits with the methodology. Using more precise SRS methods, the cut-off was consistently reported at 0.47. This value was also reported by the algorithm on trial arm 2 patients (patients with selective postoperative CRT), but the value was lower on patients with preoperative SCRT. This finding is supported by the distributions of TSR reported in the histograms of Figure 137, which show that TSR on patients that have had their cancer exposed to radiation therapy is lower than those who have not.

Significance testing on TSR high and low groups in arm 1 showed that pathologist-generated TSRs were not prognostic on patients with preoperative SCRT, with p values of 0.17 for cancer specific survival and 0.22 for overall survival. Algorithm-generated TSRs out-performed pathologist assessment, with significance values of 0.04 for cancer specific survival, and 0.01 for overall survival. HR confidence intervals contradicted the significance values for the cancer-specific survival analysis, which may be due to the smaller group sizes and small number of events in the TSR high group. The overall lower TSRs of the patients with preoperative SCRT means that the distribution of TSRs was more compressed, and therefore more sensitive to noise. It is therefore predicted that the superior algorithm performance for trial arm 1 is perhaps due (at least in part) to consistency and reproducibility of analysis. It is concluded that the algorithm performance is appropriate for survival prediction on patients that have undergone preoperative SCRT.

Applying significance testing to TSR high and low groups for arm 2 showed that pathologist-generated TSRs predict survival with log-rank significance values of 0.02 for cancer specific survival and <0.01 for overall survival. Algorithm-generated TSRs did not perform as well as the pathologist-generated TSRs, with significance values of 0.04 for both cancer specific survival and overall survival. It is concluded that algorithm performance may be appropriate for survival prediction of patients that have selective postoperative CRT, but investigations into causes of the confidence intervals for cancer specific survival would need to be undertaken.

Stratification of patients for cancer progression was not significant for either methodology, on either trial arm. All results exhibited non-significant log-rank test values, with pathologist and algorithm-generated ratios yielding 0.14 and 0.17 respectively on trial arm 1, and 0.54 and 0.61 respectively for arm 2. For arm 1, the distribution of groups was one cause of the poor performance (TSR low = 6, TSR high = 136), with only 4% of the available cases presenting cancers that had progressed (n = 6). Both pathologist and algorithm methodologies identified the same single patient in the TSR low group, which accounted for 17% of the cases in that group, and so the progression curves appear similar. A dataset containing a proportionally higher number of cases with progression may have yielded better significance results for stratification of the data. However, progression analysis in arm 2 had a proportionally higher number of patients with cancers that had progressed (25 out of 171, or 15%), and also did not yield statistically significant results.

In summary:

- Algorithm J yields statistically significant results when analysing differences between TSR high and low groups on all survival data, using the log-rank test.
- For patients with preoperative SCRT, in both presented methods of survival analysis, Algorithm J exhibits results that outperform manual analysis.
- For patients with postoperative CRT, in both presented methods of survival analysis, pathologist assessment exhibits results that outperform automated analysis.
- When analysing progression, both the algorithm and pathologist show equally poor stratification between groups, for both trial arms.

To conclude, the Algorithm J shows promising results for automatic prediction of survival that outperforms human analysis on patients with preoperative SCRT, but not postoperative CRT. The dataset that was analysed in this work was not optimal, for the reasons previously discussed, and so application of the algorithm to areas of highest TCD may improve results further.

Chapter 8 – Discussion

8.1 Thesis overview

The motivation for the work presented in this thesis is discussed in Chapter 1. The treatment of CRC patients relies on pathological examination of the disease to identify visual features that predict growth and spread, and response to chemoradiotherapy. These prognostic features are identified manually, and are subject to inter and intra-scorer variability. This variability stems from the subjectivity in qualitatively (or semi-qualitatively) interpreting the images, and the time consuming and laborious methodology of visually inspecting cancer cells.

The work in this thesis presents a systematic approach to developing a solution to address this problem for one such prognostic indicator, the tumour:stroma ratio (TSR). The steps taken are presented sequentially through the chapters, in order of the work carried out. These specifically involve the acquisition and assessment of a large dataset of expert-classified images of CRC and multiple iterations of algorithm development, in order to automate the process of generating TSRs for patient cases. The algorithm improvements are made using conclusions from observer studies, conducted on a platform developed as part of this work, and further work is undertaken to identify issues of image quality that affect automated solutions. The developed algorithm is then applied to a clinical trial dataset with survival data, meaning that the algorithm is validated against two separate pathologist-scored, clinical trial datasets, as well as being able to test its suitability for generating independent prognostic markers against patient survival.

8.2 Achievements

The work conducted in this thesis delivers the following key achievements:

1. A web-based quantitative Systematic Random Sampling (SRS) tool used for clinical research (the RandomSpot system), and a ground truth repository of pathological images containing over 2.4 million expert-classified labels (RandomSpotDB).
2. A web-based experiment platform used for rapid acquisition of ground truth data (Prospector), and psychophysics experiments that lead to the identification of criteria for maximising manual scoring agreement.
3. A novel hybrid unsupervised segmentation algorithm that maximises CRC tissue segmentation accuracy and minimises computational cost.
4. A novel, iteratively designed Machine Learning (ML) algorithm that automatically calculates TSR with 86% accuracy, compared to manual scoring, using contextual information from the local surrounding tissue to improve classifications and global stain characteristics to compensate for variation in the dataset,
5. An automatic stain-based Quality Control (QC) algorithm that labels slides with 89% accuracy, compared to manual scoring.
6. Validation of the TSR algorithm on two clinical trial datasets, and human survival data, showing that the algorithm is capable of predicting survival on suboptimal datasets.

These achievements are summarised in the subsequent sections.

8.2.1 RandomSpot and RandomSpotDB

RandomSpot uses SRS to generate semi-quantitative assessments of digital slide tissue, where traditionally qualitative or semi-qualitative methods are used (see section 2.2). The system generates digital slide annotations which are used to generate statistical analyses of the proportions of tissue type within their specified ROI. This data is also reusable for image analysis development.

RandomSpot has been used in multiple publications to generate precise numeric quantitation of tissue types within a cancer, and has been used to prove the prognostic capability of TSR in CRC, breast, endometrial and upper GI cancers. RandomSpot was developed out of a need to quantitate WSIs more quickly, and no other SRS systems were publicly available for use with digital slides at the time of development. This methodology improves upon the visual estimation used in other publications by providing a specific continuous variable with comparatively low inter scorer variability (see section 2.2.4 for a review on TSR).

The data from some of these studies was acquired and stored in RandomSpotDB, which currently contains over 2.4 million data points with expert-classified labels. As discussed in Chapter 7, some of the projects stored in RandomSpotDB are from clinical trials which have survival data associated with them, allowing algorithms to be validated against outcomes rather than the maximisation of agreement with subjective pathologist assessments.

The ground truth is stored as single x-y coordinates with an expert-classified label for each point. The incorporation of this data into a ML solution posed an interesting problem, in that the single point annotations contained no tissue boundaries to delineate between neighbouring tissue classes. This meant that without some form of further segmentation, surrounding visual information that was not considered relevant to the classification label would be included in analysis of the data points.

The RandomSpotDB dataset is a useful tool for developing and validating image analysis solutions, and has the potential to be a valuable collaborative research dataset to open to the digital pathology community. This will be explored in future work, and may be potentially released as a community exercise similar to the GlaS challenge [230], where researchers are invited to try their algorithms on the data.

8.2.2 Prospector and psychophysics experiments

Pathologist time is expensive, and difficult to obtain for research projects. The Prospector system was specifically designed to minimise the amount of time and cognitive effort required, to obtain adequate numbers of ground truth data, by facilitating capture of high throughput of ground truth information.

Prospector was used to present pathologists with pre-scored images of three different sizes, in order to identify an appropriate minimum size for maintaining levels of agreement. It was found that reducing the pathologists field of view when scoring images reduces the agreement between them, and 256x256 pixel images (at 20x objective or 0.5 microns per pixel) was identified as an appropriate minimum size. Figure 146 illustrates how this finding compares to image sizes used in previous research.

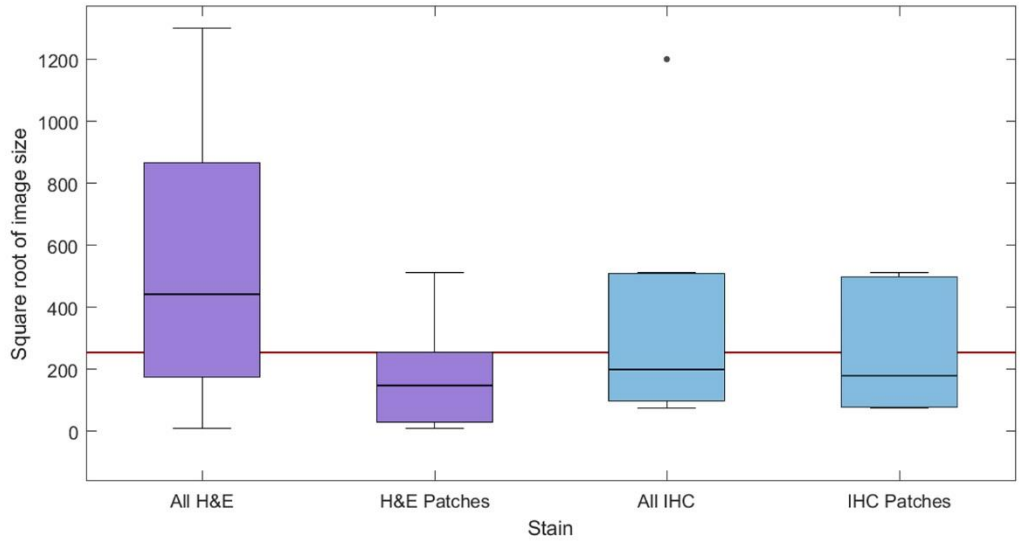


Figure 146 – Boxplot distributions of image size used in previous publications compared to psychophysics experiment findings

The plot breaks down the results previously presented in Figure 75, so that image patch size data is represented separately to ROI data. The horizontal line is placed at 256, which represents the optimised size found in research..
Image size is taken from the results presented in Table 3, and split by stain type. Note that some studies reported using non-square images for analysis and so the square root of total area size is taken to improve the comparability of results.

The boxplot data is taken from Table 3 and presents a broken down view of Figure 75 (from section 4.5) into all image sizes per stain, and image sizes for algorithms that use patch-based analysis. The patch size found by the research presented in Chapter 4 is higher than the size of published H&E patches (mean = 178x178 pixels, median = 150x150 pixels), which may be due to the contextual information not being taken into account in other studies. It should be noted that there were no publications that justified the reason for the patch size used, or that presented analyses on multiple patch sizes (unless using them for multi-scale analysis).

It was concluded that the image size impacted the pathologist's ability to score images by including or excluding contextual information that allows them to understand what the image shows. The contextual information surrounding the centre of the patch was likely to contain useful structural cues that enabled the observer to orientate their viewpoint and assess the centre of the image with higher levels of confidence (shown by the high rate of rejection in smaller image sizes). However, this extra information was averaged out in the feature vectors of the initial algorithms, instead of being used as separate contextual features, creating more overlap in the feature space. This drove the research in Chapter 5.

A pilot study also found that weak staining levels were consistently rejected for scoring by a pathologist, where one pathologist viewed 250 images, and accepted 86% of them. This resulted in an imbalanced dataset of pathologist classifications, and attempts at making a trained RF classifier to automatically QC images did not adequately detect slides that would be rejected based on stain. These initial findings and limitations drove the research in Chapter 6.

The prospector system itself was designed specifically to minimise choice and increase the speed at which answers are recorded. It was observed that digital pathology systems with complex user interfaces are rarely used - and if they are, they are rarely used effectively. By designing the system to be used with touch enabled devices, more possibilities are opened up in terms of collaborative or ad hoc data collection projects. One limitation with the system is that it uses images hosted on external URLs. This was a deliberate design choice, so that the image hosting and subsequent data transfer rates are solely the responsibility of the user (experiment designer). By restricting the participant to view static fields of view, the experiment can be effectively restricted to conditions set by the designer, but not being able to navigate embedded virtual slides is a limitation which affects the usability of the system, and limits the experimental design choices.

Currently prospector is used internationally by multiple research institutions for rapid acquisition of ground truth data. The system also has the capacity to be used in citizen science projects to collaboratively generate big data, such as Candido et al's work [266] enlisting over 100,000 participants to look for oestrogen receptors in TMA cores.

8.2.3 Unsupervised segmentation algorithm

The unsupervised segmentation algorithm was developed out of a need to separate areas of tissue within an image patch, that only contained one tissue class. The algorithm combined superpixel clustering with normalised cuts to generate a smaller affinity matrix and reduce processing time. The hybrid segmentation method reported a segmentation accuracy of 0.93, which was the second highest accuracy rating of the nine algorithms tested. The normalised cuts algorithm (using per-pixel pairwise comparisons) yielded 0.97 segmentation accuracy, but increased in computational time by 698%. The hybrid clustering algorithm was selected for incorporation into the ML algorithm based on the trade-off between speed and accuracy.

One thing that is not measured in this work is the importance of segmentation accuracy. One logical assumption of the work is that accuracy should be maximised, but a minimum acceptable level was not established. This is a common theme when discussing comparisons to

ground truth data, especially when inter-scorer variability is high. Pathologist agreement will be discussed in more detail in the next section, but for the segmentation task, the effect of the 4% accuracy difference between normalised cuts and the hybrid solution was not investigated. From working closely with pathologists, and inspecting ROI annotation data, assessment of digital pathology slides is rarely precise to the levels of segmentation accuracy exhibited by these algorithms. This may be a noteworthy factor when trying to attain perfect human-algorithm agreement.

The comparison plots between the hybrid clustering technique and other segmentation methods showed that normalised cuts performed significantly better than all other algorithms, and watershed segmentation performed significantly worse. Had computation time not been a factor in algorithm development, then the normalised cuts method may have improved the ML algorithm performance, however, with a 698% increase in computing time, the feature vectors would have taken approximately 24.5 days to complete on the QUASAR dataset.

When processing time is considered in digital pathology research, it is often placed below accuracy in terms of priority. This is logical given that the field is focused on increasing levels of agreement to ultimately maximise the efficacy patient treatments. Also, the research setting that such algorithms are typically developed in does not place time-critical demands upon processing. A common justification for accepting slower algorithm times is that processing can be computed in batches prior to pathological inspection, which is a reasonable approach to implementing such algorithms. However, with a view to implementing these algorithms into routine medical practice, consideration must be taken that the routine volumes of slides generated will greatly exceed the number of images that researchers report training and testing their algorithms on.

8.2.4 Automated TSR algorithm

The base methodology of the automated TSR algorithm used an optimised RF classifier to learn features of image patches that were generated by the RandomSpot system and stored in RandomSpotDB. This methodology was modified based on conclusions from processing results and psychophysics experiments in order to improve the agreement statistics between pathologist ground truth and algorithm-generated results. The algorithm modifications are described in Table 14, and were implemented sequentially to assess the effect of each on performance.

Figure 147 reiterates the accuracy levels reported by other publications, presented in Table 3. The red line on the plot indicates the 0.86 accuracy level reported by Algorithm K, which falls

slightly above the average accuracy level reported on H&E images (0.85) and below the average accuracy level reported on IHC images (0.87).

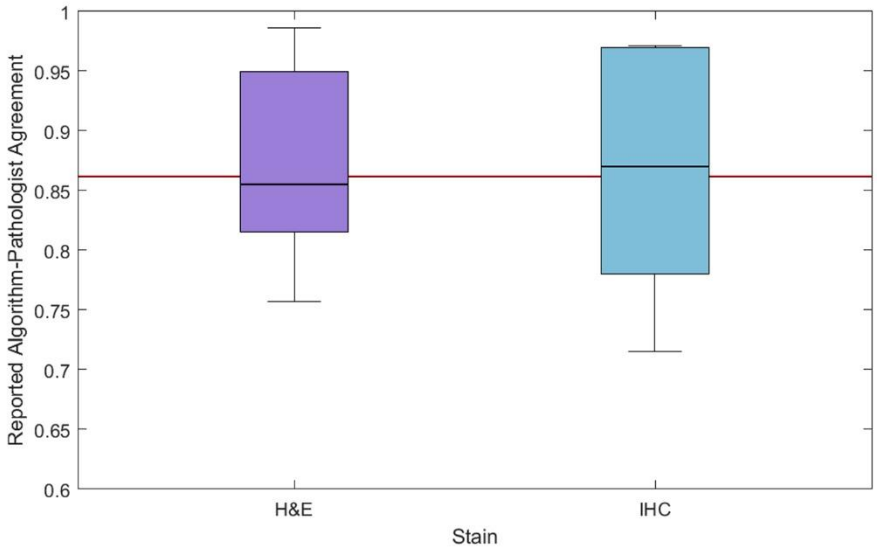


Figure 147 – Published algorithm-pathologist agreement results compared to Algorithm J

The boxplot distributions are taken from Table 3, and grouped by stain type. The red line represents the 86% accuracy level attained by the algorithm development in this thesis. The plot shows that the developed algorithm attains similar accuracy levels to published solutions.

The accuracy of the developed solution therefore is comparable to the current state of the art solutions. It is predicted that the increased application of deep learning in this area is likely to increase the average reported accuracy of digital pathology image analysis, and in the near future, the distribution of results presented in Figure 147 is likely to change. However, given the results in the GlaS challenge (mean accuracy = 0.82 for pre-released image dataset and 0.63 for the unseen dataset), it seems that deep learning solutions still require expert knowledge to appropriately design and tune the network architectures to optimise results. It should also be noted that the training, testing and validation of the algorithm developed in this work used over 200,000 expert-labelled patch images, and the highest number in the literature reported 5,000 (generated from 8 slides containing 625 patches). Arguably, this amount of data facilitated the accuracy of the algorithm, in that there were many more images available for algorithm training. However, this meant that the RF classifier was exposed to much more variability and suboptimal quality images, which equates to more overlap of classes in the feature space. It would be interesting to see how the published solutions perform on the same dataset.

Kather et al [203] presented a research methodology similar to that described in this work. The incorporation of multi-class CRC patches was reported as novel, despite being published two years after the work from Chapters 3 and 4 [259]. The paper reports using 5000 multi-class

patches, 150x150 pixels in size, using 8 hand-annotated WSIs to create 625 patches for each slide. The authors reported 98.6% accuracy for two-class (tumour and stroma patches only) agreement, and 87.4% accuracy for eight-class agreement. The two-class classification method discarded all other tissue types (as opposed to grouped them as sub-classes), so that the dataset was reduced to 1250 patch images for that part of the analysis. The methodology examined several classifiers using multiple textural features on greyscale image patches (converted from RGB). No segmentation was performed, and no contextual analysis was applied, making the algorithm more comparable to Algorithm A from this research. Even with such a small dataset (8 slides), the authors anecdotally noted a large amount of stain variation, which was neither quantified nor illustrated with examples.

The eight classes used differed from the ground truth classes presented in this thesis. They were: tumour epithelium; simple stroma; complex stroma; immune cells; debris or mucus; mucosal glands; adipose tissue; background. Through experience working with CRC slides, it is predicted that separating stroma into simple and complex may have increased algorithm accuracy, if it were included in the ground truth data.

The accuracy reported by this Kather et al are significantly higher than the algorithms presented in this thesis. Despite the lack of images, justification of patch size and contextual analysis, the methodology uses a rigorous approach to evaluating classifiers and texture features, and reports impressive agreement statistics. Investigations into alternate texture features and image patch classes will be undertaken in future work.

The incremental development of the automated SRS algorithm presented in this thesis follows a traditional computer vision approach to an image analysis problem, in that images are analysed for hand-crafted features that are believed to best represent the differences in the classes, before training a classifier and testing unseen images with ground truth labels. This approach is being overtaken by deep learning practices, which allow the algorithm to automatically generate and select features that minimise prediction error. The work in Chapters 4 and 5 specifically aim to better model the methods of human analysis by incorporating contextual information into the feature vectors, whereas deep learning solutions may model completely different behaviour that maximises agreement levels in an unrelated way. This may not be important, as it has been observed that deep learning is often seen as a black box, and that as long as the results are improved over traditional methods, the technicalities of the methodology does not matter. This attitude is more useful when considering ML methods as tools, and the application of the most appropriate tool for the task is more important than the methods themselves. This is especially important when considering the implementation of automated solutions into routine clinical practice.

Reiterating the point made regarding accuracy in 8.2.3, the development of automated pathological image analysis solutions typically focuses on 100% agreement as the ultimate goal. Given that the mean inter-observer agreement observed in Chapter 4 is 0.85 on optimised images (0.89 when excluding the rejected images), trying to attain 100% agreement should not be the target. Pathological assessment is applied to tissue images so that predictions can be made about a specific disease, and how best to treat it. By comparing algorithm performance to clinical outcomes, a better benchmark is set for testing the practical use of the algorithm. It is this conclusion that drives the research in Chapter 7.

The implementation of the developed automated SRS system into routine pathological practice requires that the analysis of the images is robust, but also that the misclassified images are easier to identify and correct than discarding the results and doing the analysis manually. This requires extensive research and development of user interface design, grounded in cognitive science and heuristics. This is beyond the scope of this research, but it is important to consider that automated solutions that do not work with 100% accuracy will need human review and intervention. From this perspective, making the results of an automated solution easy to review and adjust is much more important than researching into new methods to increase algorithm accuracy by a few percent.

8.2.5 Automated QC algorithm

The automated QC algorithm was developed using a dataset of hand-labelled images which were assessed for visual issues, which had been pre-identified as scoring categories after analysing cases where pathologist-algorithm agreement was lowest. The study was limited to one participant, due to the number of slides required to analyse (2,211), and despite built-in breaks during the experiment, results may well have been affected by fatigue. The participant was only required to assign one category to each slide, and so slides that exhibited more than one issue only had the most prominent issue recorded. In most cases that had more than one visual artefact, the most prominent issue was weak staining. This contributed to the imbalance of QC class examples in the training data (701 slides in the largest subset, and 4 in the smallest), and may have contributed to the subsequent algorithms poor performance on artefacts other than staining. Because of this, the QC algorithm was reduced in functionality to identify poor staining only. Further development of this algorithm may benefit from object detection, especially for bubbles and debris, where general pixel-level features overlap with other classes in the feature space.

Application of the (stain) QC algorithm to the QUASAR dataset flagged suboptimal slides with 89% accuracy, and application of the ML algorithm to the optimised subset of slides increased accuracy by 3%. However, the approved subset of slides amounted to 32% of the total number of slides, and so despite the overall low quality of the dataset, this QC process may not be feasible in a routine setting.

Image quality is clearly an important factor for image analysis algorithms, and results indicate that pre-screening slides based on some quality metric would improve analysis overall. However, automation of image quality in digital pathology is not widely researched, and publications are few. Ameisen et al [236] implement a quality assessment algorithm with a focus on blur detection (pun intended), but do not report any levels of accuracy or agreement to some form of quality-based ground truth. Avanaki et al [267] simulate blur on a set of H&E stained breast tissue images (scanned at Leeds where manual QC is applied by the scanner operator to flag and rescan slides with blur), and use Mittal et al's [268] natural image quality evaluator (NIQE) metric to compare perception of quality as assessed by a pathologist, using a scale of one to ten. Results showed that the NIQE score was successful at identifying images where pathologists identified quality as low. Robust blur detection (of out of focus areas) would be a very useful tool for slide scanner operators or to build into slide scanning software to minimise the effort involved in the manual QC of scanned slides. The work on image quality in this thesis was focused on levels of stain, since it accounted for 43% of the whole QUASAR dataset and 63% of all QC issues identified. However, assessing image quality using perspectives from other fields of research would benefit future work in this area.

The use of automated QC in a practical routine environment requires thorough consideration in terms of where to implement the point of use, such as on the scanner side, or prior to image viewing, or prior to image analysis. The implications of discarding slides based on a quality metric mean that slides may need to be re-scanned, re-stained or re-cut, and all of these options involve extra workload, which may have been suboptimal, yet acceptable to human visual inspection. This suggests that since image quality affects image analysis algorithms the most, the quality issues should be handled at that point of use. Rather than use the QC algorithm as a tool to simply discard slides and data based on visual appearance, application of such a QC metric may be more suited to discarding training images if they do not meet the quality criteria, or flagging up slides which are likely to affect image analysis due to their quality issues. This would allow users of automated solutions to review analyses of cases that do not meet quality criteria with more scrutiny, knowing that suboptimal images would likely result in suboptimal analyses.

8.2.6 Validation of the algorithm on survival data

Algorithm H was applied to the CR07 dataset and showed significant improvement over all previous algorithms, yielding 86% accuracy, and a kappa value of 0.71 on a set of cases that had not been through any QC process. The automatic (stain) QC algorithm was applied to the dataset and rejected 23% of the cases. Since the CR07 dataset was not manually analysed for QC issues, there was no ground truth to evaluate the accuracy of the QC process. Removing the QC-rejected slides increased the accuracy of the algorithm by less than 1%, (without statistical significance).

The improved performance of the algorithm on the CR07 dataset indicated that the dataset was of better quality than the QUASR one. In general, the cases appeared of higher and more consistent quality staining, with more visual contrast between tissue components. However, there may be other factors contributing to the accuracy gains, such as more accurate ground truth labels, or more appropriate sampling methods. The sampling areas in the new dataset used freehand ROIs of the whole tumour, compared to the 3x3mm box annotations of the QUASAR dataset. The box annotations were placed over the area with the highest TCD, which means tumour epithelium is denser, and more likely to be squashed together, distorting the appearance of the tissue. This is an observation that may not be pathologically accurate, however, it is important to note that there may be multiple contributing factors to the increase in accuracy on the CR07 dataset besides image quality.

Using the pre-published method for stratifying patient groups for survival analysis, the cut-off values generated by the human and machine TSRs were lower than other published studies for patients with preoperative SCRT, but were the same when analysing patients with selective postoperative CRT. This is because previously published studies did not report survival (or subsequent cut-offs) on patients with preoperative SCRT, because their findings were not prognostic. The algorithm-generated TSRs were prognostic on patients with preoperative SCRT. This may be due to the tissue being broken up by radiation therapy, and the subsequent mix of tumour epithelium and necrotic tissue is very difficult to manually analyse. The pathological interpretation of the performance difference is beyond the scope of this work.

The algorithm-generated survival curves showed that the algorithm outperformed manual analysis for both caner and overall survival when analysing patients with preoperative SCRT, generating statistically significant differences between patient groups (using the log-rank test), where the manual analysis did not. For patients with selective postoperative CRT, algorithm performance exhibited higher significance values than groups stratified by pathologist-generated TSRs, but were still statistically different using the log-rank test. However, for cancer survival,

the smaller and more imbalanced group sizes may have contributed to the larger ranging confidence intervals when analysing hazard ratios (stratification of groups was still statistically significant), which may require further investigation in future work.

The issue of generating cut-off values to stratify patient groups based on the modified ROC curve is contentious. Previous publications used two, four, five or ten categories to visually estimate TSR, and so creating a cut-off at 50% was convenient for the methodology used. The statistical method for generating the cut-off yielded a consistent cut-off of 47% across multiple publications. This number was also replicated by the algorithm-generated TSRs for patients that did not have preoperative SCRT. The contentious issue surrounds generating cut-offs optimised for a specific set of cases. The cut-offs for trial arm 1 patients were different for both human and machine-generated TSRs. This is unsurprising, given that patients that have had radiation therapy will have less tumour, and so their TSRs will be lower, but the characteristics of their cancer may not have changed. Again, it is beyond the scope of this work to pathologically interpret these results, but over-engineering the cut-off to the dataset should be avoided where possible. Using a fixed cut-off of 0.47 on the trail arm 1 cases would make the resulting stratification not significant, which is why analysis on these cases is unreported in the literature.

Analysis of survival data is very difficult. The data is of significant clinical value and is often protected by statisticians that work solely with the clinical trials units. This prevents researchers from over-engineering solutions to the data. Due to the difficulty of this task, very few publications use survival data to validate their algorithms, and the ones that do are more focused on the clinical methods, and use commercial algorithms as tools to generate the statistics to try to identify new prognostic markers, rather than focus on algorithm validation.

Using an automated solution to generate TSRs and predict survival carries very serious implications. In order to implement such a system into routine clinical practice, it must be considered that if results are accepted without some form of review, the algorithm may incorrectly score a patient's cancer, resulting in an incorrect prediction of response to therapy, and therefore advise an inappropriate treatment. Algorithms need to be approved by regulatory bodies such as the US Food and Drug Administration (FDA) or the European Commission's European Conformity (CE) mark to be used in clinical settings, which require proof of extensive validation to be accepted. Research into automated analysis of tissue typically focuses on very specific problems, such as the TSR in CRC, and often these algorithms are too specific to be of widespread use. Incorporating multiple image analysis tools into a software suite or an automated pathology workflow would be much more likely to be of widespread clinical value, and therefore would have more chance of being approved for primary diagnosis.

8.3 Conclusions and future work

The work presented in this thesis uses two clinical datasets to address the automatic generation of one prognostic marker in CRC. When considering the variation between observers, the results exhibited by the algorithm indicate that the produced solution is an adequate surrogate for manual scoring.

The manual scoring experiments presented in this thesis demonstrate pathologist inter-scorer variability, and highlight the subjective and laborious nature of scoring large histopathological images. With that consideration in mind, it is perhaps not the maximisation of (ultimately flawed) pathologist agreement that the algorithms should look to achieve, rather the consistent and accurate prediction of prognostic features. However, the algorithm development process is made more robust when validated against a reasonable benchmark, before validating it on patient survival data.

Survival analysis is performed on a dataset generated by a sampling method that is less well correlated with survival, and still shows significance when stratifying patient groups into high and low TSR groups. Both human and machine methods did not show significance between groups when analysing progression, and the dataset analysed did not include data on stage or recurrence. There is still a large amount of image data and clinical trials in the RandomSpotDB dataset that has not been analysed, and so applying the algorithm to optimally sampled survival data, with more clinical details will be part of future work to evaluate the capabilities of the algorithm further.

Deep Learning (specifically CNNs) is a promising technique for analysing images without introducing human biases in feature selection and learning the composition of objects in a hierarchical method - from simple lines and edges, up to shapes and complex structures. The timescale of this work meant that the popularity of the Deep Learning technologies emerged towards the end of the project, and will be explored in future work.

There are other prognostic markers that the developed solution may be able to generate (such as the density of tumour infiltrating lymphocytes), and the data generated by the algorithms may also be useful for identifying new prognostics markers (such as distribution of tissue within tumours). These are interesting research questions that the author intends to address after this project concludes.

Bibliography

[1] A. I. Wright, "Problem solved? Postgraduate research image of the year portrays high-tech solution to cancer puzzle," *Times Higher Education*, 2012. [Online]. Available: https://www.timeshighereducation.com/news/problem-solved-postgraduate-research-image-of-the-year-portrays-high-tech-solution-to-cancer-puzzle/418593.article. [Accessed: 27-Aug-2017].

[2] Cancer Research UK, "Bowel Cancer Statistics," 2017. [Online]. Available: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer. [Accessed: 26-Aug-2017].

[3] A. I. Wright, M. C. Waterhouse, and D. E. Treanor, "Virtual Pathology at the University of Leeds," 2017. [Online]. Available: http://www.virtualpathology.leeds.ac.uk/. [Accessed: 26-Aug-2017].

[4] Leica Biosystems, "Aperio Digital Pathology Slide Scanners," 2017. [Online]. Available: http://www.leicabiosystems.com/digital-pathology/aperio-digital-pathology-slide-scanners/. [Accessed: 26-Aug-2017].

[5] E. Meijering, "Cell Segmentation : 50 Years Down the Road," *Signal Process. Mag. IEEE*, vol. 29, no. 5, pp. 140–145, 2012.

[6] N. P. West, M. Dattani, P. McShane, G. Hutchins, J. Grabsch, W. Mueller, D. Treanor, P. Quirke, and H. Grabsch, "The proportion of tumour cells is an independent predictor for survival in colorectal cancer patients.," *Br. J. Cancer*, vol. 102, no. 10, pp. 1519–23, May 2010.

[7] E. Toh, P. Brown, I. Botterill, and P. Quirke, "(," 2014.

[8] M. Dattani, P. McShane, D. Treanor, G. Hutchins, J. Grabsch, J. M. Brown, H. Thorpe, D. G. Jayne, P. J. Guillou, W. Mueller, P. Quirke, and H. Grabsch, "Proportion of

Tumour Cells and Stroma – An Inexpensive but Reliable Predictor of Patient Survival in Colorectal Cancer," *J. Pathol.*, vol. 216, no. 1, p. S17, 2008.

[9] Cancer Research UK, "Bowel Cancer Incidence Statistics," *CRUK Cancer Statistics*, 2014. [Online]. Available: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/incidence. [Accessed: 06-Apr-2016].

[10] P. Quirke, "Colorectal pathology teaching package," Leeds, UK, 2011.

[11] M. Riihimäki, A. Hemminki, J. Sundquist, and K. Hemminki, "Patterns of metastasis in colon and rectal cancer.," *Sci. Rep.*, vol. 6, p. 29765, 2016.

[12] A. Gray, A. Wright, P. Jackson, M. Hale, and D. Treanor, "Quantification of histochemical stains using whole slide imaging: development of a method and demonstration of its usefulness in laboratory quality control.," *J. Clin. Pathol.*, vol. 68, no. 3, pp. 192–199, Dec. 2015.

[13] F. Marongiu, S. Doratiotto, M. Sini, M. P. Serra, and E. Laconi, "Cancer as a disease of tissue pattern formation.," *Prog. Histochem. Cytochem.*, vol. 47, no. 3, pp. 175–207, Oct. 2012.

[14] F. Bosman, F. Carneiro, R. Hruban, and N. Thiese, "Carcinoma of the Colon and Rectum," in *WHO Clasification of Tumours of the Digestive System*, 4th ed., 2010, pp. 134–146.

[15] C. E. Dukes, "The classification of cancer of the rectum," *J. Pathol. Bacteriol.*, vol. 35, no. 3, pp. 323–332, Nov. 1932.

[16] P. Denoix, "Nomenclature classification des cancers," *Bull Inst Nat Hyg*, vol. 7, pp. 743–748, 1952.

[17] L. H. Sobin, M. K. Gospodarowicz, and C. Wittekind, *Colon and Rectum*, 7th ed. Chichester, West Sussex, UK: Wiley-Blackwell, Oxford, 2012.

[18] P. Quirke and E. Morris, "Reporting colorectal cancer.," *Histopathology*, vol. 50, no. 1, pp. 103–12, Jan. 2007.

[19] R. H. Riddell, R. E. Petras, G. T. Williams, and L. H. Sobin, *Atlas of Tumor Pathology: Tumors of the Intestines*, Third. American Registry of Pathology, 2003.

[20] C. C. Compton, L. P. Fielding, L. J. Burgart, B. Conley, H. S. Cooper, S. R. Hamilton, M. E. Hammond, D. E. Henson, R. V Hutter, R. B. Nagle, M. L. Nielsen, D. J. Sargent,

C. R. Taylor, M. Welton, and C. Willett, “Prognostic factors in colorectal cancer. College of American Pathologists Consensus Statement 1999.,” *Arch. Pathol. Lab. Med.*, vol. 124, no. 7, pp. 979–94, Jul. 2000.

[21] a Lugli, E. Karamitopoulou, and I. Zlobec, “Tumour budding: a promising parameter in colorectal cancer.,” *Br. J. Cancer*, vol. 106, no. 11, pp. 1–5, Apr. 2012.

[22] H. Kanazawa, H. Mitomi, Y. Nishiyama, I. Kishimoto, N. Fukui, T. Nakamura, and M. Watanabe, “Tumour budding at invasive margins and outcome in colorectal cancer.,” *Colorectal Dis.*, vol. 10, no. 1, pp. 41–7, Jan. 2008.

[23] S. Kazama, T. Watanabe, Y. Ajioka, T. Kanazawa, and H. Nagawa, “Tumour budding at the deepest invasive margin correlates with lymph node metastasis in submucosal colorectal cancer detected by anticytokeratin antibody CAM5.2.,” *Br. J. Cancer*, vol. 94, no. 2, pp. 293–8, Jan. 2006.

[24] P. D. Caie, A. K. Turnbull, S. M. Farrington, A. Oniscu, and D. J. Harrison, “Quantification of tumour budding, lymphatic vessel density and invasion through image analysis in colorectal cancer.,” *J. Transl. Med.*, vol. 12, no. 1, p. 156, Jan. 2014.

[25] Y.-H. Lai, L.-C. Wu, P.-S. Li, W.-H. Wu, S.-B. Yang, P. Xia, X.-X. He, and L.-B. Xiao, “Tumour budding is a reproducible index for risk stratification of patients with Stage II colon cancer.,” *Colorectal Dis.*, vol. 16, no. 4, pp. 259–64, Apr. 2014.

[26] H. Nizze, M. Barten, and F. Prall, “Tumour budding as prognostic factor in stage I/II colorectal carcinoma,” *Histopathology*, vol. 47, no. 1, pp. 17–24, 2005.

[27] J. R. Jass, M. Barker, L. Fraser, M. D. Walsh, V. L. J. Whitehall, B. Gabrielli, J. Young, and B. A. Leggett, “APC mutation and tumour budding in colorectal cancer,” *J. Clin. Pathol.*, vol. 56, pp. 69–73, 2003.

[28] F. Prall, “Tumour budding in colorectal carcinoma.,” *Histopathology*, vol. 50, no. 1, pp. 151–62, Jan. 2007.

[29] E. Karamitopoulou, I. Zlobec, V. H. Koelzer, R. Langer, H. Dawson, and A. Lugli, “Tumour border configuration in colorectal cancer : proposal for an alternative scoring system based on the percentage of infiltrating margin,” pp. 464–473, 2015.

[30] K. M. Ropponen, M. J. Eskelinen, P. K. Lipponen, E. Alhava, and V. M. Kosma, “Prognostic value of tumour-infiltrating lymphocytes (TILs) in colorectal cancer,” *J. Pathol.*, vol. 182, no. 3, pp. 318–324, 1997.

[31] H. Chang, L. Loss, and B. Parvin, "Nuclear segmentation in H and E sections via multi-reference graph-cut (MRGC)," *Int. Symp. Biomed. Imaging*, 2012.

[32] J. W. Huh, J. H. Lee, H. R. Kim, and Y. J. Kim, "Prognostic significance of lymphovascular or perineural invasion in patients with locally advanced colorectal cancer," *Am. J. Surg.*, vol. 206, no. 5, pp. 758–763, 2013.

[33] T. M. Mayhew and H. J. Gundersen, "If you assume, you can make an ass out of u and me': a decade of the disector for stereological counting of particles in 3D space.," *J. Anat.*, vol. 188, pp. 1–15, Feb. 1996.

[34] S. a Tschanz, P. H. Burri, and E. R. Weibel, "A simple tool for stereological assessment of digital images: the STEPanizer.," *J. Microsc.*, vol. 243, no. 1, pp. 47–59, Jul. 2011.

[35] B. T. Hyman, T. Gomez-Isla, and M. C. Irizarry, "Stereology: a practical primer for neuropathology," *J. Neuropathol. Exp. Neurol.*, vol. 57, no. 4, pp. 305–310, 1998.

[36] W. A. Aherne and M. S. Dunnill, *Morphometry*, 1st ed. Bath, UK, UK: Edward Arnold (Publishers) Ltd, 1982.

[37] K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò, "Power failure: why small sample size undermines the reliability of neuroscience," *Nat. Rev. Neurosci.*, vol. 14, no. 5, pp. 365–376, 2013.

[38] D. R. Cox, S. Society, and S. B. Methodological, "Regression Models and Life-Tables," *J. R. Stat. Soc. Ser. B*, vol. 34, no. 2, pp. 187–220, 1972.

[39] J. T. RICH, J. G. NEELY, R. C. PANIELLO, C. C. J. VOELKER, B. NUSSENBAUM, and E. W. WANG, "A Practical Guide to Understanding Kaplan-Meier Curves," *Otolaryngol. Head. Neck Surg.*, vol. 143, no. 3, pp. 331–336, 2010.

[40] J. Wu, C. Liang, M. Chen, and W. Su, "Association between tumor-stroma ratio and prognosis in solid tumor patients: a systematic review and meta-analysis.," *Oncotarget*, vol. 7, no. 42, pp. 68954–68965, 2016.

[41] R. R. Langley and I. J. Fidler, "The seed and soil hypothesis revisited - the role of tumor-stroma interactions in metastasis to different organs," *Int J Cancer*, vol. 128, no. 11, pp. 2527–2535, 2012.

[42] C. L. Downey, S. A. Simpkins, J. White, D. L. Holliday, J. L. Jones, L. B. Jordan, J. Kulka, S. Pollock, S. S. Rajan, H. H. Thygesen, A. M. Hanby, and V. Speirs, "The prognostic significance of tumour-stroma ratio in oestrogen receptor-positive breast

cancer.," *Br. J. Cancer*, vol. 110, no. 7, pp. 1744–7, Apr. 2014.

[43] E. M. De Kruijf, J. G. H. Van Nes, C. J. H. Van De Velde, H. Putter, V. T. H. B. M. Smit, G. J. Liefers, P. J. K. Kuppen, R. A. E. M. Tollenaar, and W. E. Mesker, "Tumor-stroma ratio in the primary tumor is a prognostic factor in early breast cancer patients, especially in triple-negative carcinoma patients," *Breast Cancer Res. Treat.*, vol. 125, no. 3, pp. 687–696, 2011.

[44] T. J. A. Dekker, C. J. H. Van De Velde, G. W. Van Pelt, J. R. Kroep, J. P. Julien, V. T. H. B. M. Smit, R. A. E. M. Tollenaar, and W. E. Mesker, "Prognostic significance of the tumor-stroma ratio: Validation study in node-negative premenopausal breast cancer patients from the EORTC perioperative chemotherapy (POP) trial (10854)," *Breast Cancer Res. Treat.*, vol. 139, no. 2, pp. 371–379, 2013.

[45] S. Ahn, J. Cho, J. Sung, J. E. Lee, S. J. Nam, K. M. Kim, and E. Y. Cho, "The prognostic significance of tumor-associated stroma in invasive breast carcinoma," *Tumor Biol.*, vol. 33, no. 5, pp. 1573–1580, 2012.

[46] T. Zhang, J. Xu, H. Shen, W. Dong, Y. Ni, and J. Du, "Tumor-stroma ratio is an independent predictor for survival in NSCLC," *Int. J. Clin. Exp. Pathol.*, vol. 8, no. 9, pp. 11348–11355, 2015.

[47] K. Wang, W. Ma, J. Wang, L. Yu, X. Zhang, Z. Wang, B. Tan, N. Wang, B. Bai, S. Yang, H. Liu, S. Zhu, and Y. Cheng, "Tumor-stroma ratio is an independent predictor for survival in esophageal squamous cell carcinoma," *J. Thorac. Oncol.*, vol. 7, no. 9, pp. 1457–1461, 2012.

[48] Y. Wu, H. Grabsch, T. Ivanova, I. Tan, and J. Murray, "Comprehensive genomic meta-analysis identifies intra-tumoural stroma as a predictor of survival in patients with gastric cancer," *Gut*, vol. 62, no. 8, pp. 1100–11, 2013.

[49] P. Aurello, G. Berardi, D. Giulitti, A. Palumbo, S. M. Tierno, G. Nigri, F. D'Angelo, E. Pilozzi, and G. Ramacciato, "Tumor-Stroma Ratio is an independent predictor for overall survival and disease free survival in gastric cancer patients," *Surgeon*, vol. 15, no. 6, pp. 329–335, 2017.

[50] W. E. Mesker, J. M. C. Junggeburt, K. Szuhai, P. de Heer, H. Morreau, H. J. Tanke, and R. A. E. M. Tollenaar, "The carcinoma-stromal ratio of colon carcinoma is an independent factor for survival compared to lymph node status and tumor stage.," *Cell. Oncol.*, vol. 29, no. 5, pp. 387–98, 2007.

[51] A. Huijbers, R. A. E. M. Tollenaar, G. W. V Pelt, E. C. M. Zeestraten, S. Dutton, C. C. McConkey, E. Domingo, V. T. H. B. M. Smit, R. Midgley, B. F. Warren, E. C. Johnstone, D. J. Kerr, and W. E. Mesker, "The proportion of tumor-stroma as a strong prognosticator for stage II and III colon cancer patients: Validation in the victor trial," *Ann. Oncol.*, vol. 24, no. 1, pp. 179–185, 2013.

[52] G. G. A. Hutchins, D. Treanor, A. Wright, K. Handley, L. Magill, E. Tinkler-Hundal, K. Southward, M. Seymour, D. Kerr, R. Gray, and P. Quirke, "Intra-tumoural stromal morphometry predicts disease recurrence but not response to 5- fluorouracil - results from the QUASAR trial of colorectal cancer," *Histopathology*, vol. 72, no. 3, pp. 391–404, 2018.

[53] J. Liu, J. Liu, J. Li, Y. Chen, X. Guan, X. Wu, C. Hao, Y. Sun, Y. Wang, and X. Wang, "Tumor-stroma ratio is an independent predictor for survival in early cervical carcinoma," *Gynecol. Oncol.*, vol. 132, no. 1, pp. 81–86, 2014.

[54] Y. Chen, L. Zhang, W. Liu, and X. Liu, "Prognostic Significance of the Tumor-Stroma Ratio in Epithelial Ovarian Cancer," *Biomed Res. Int.*, vol. 2015, 2015.

[55] H. Panayiotou, N. M. Orsi, H. H. Thygesen, A. I. Wright, M. Winder, R. Hutson, and M. Cummings, "The prognostic significance of tumour-stroma ratio in endometrial carcinoma," *BMC Cancer*, vol. 15, no. 1, p. 955, 2015.

[56] E. F. W. Courrech Staal, V. T. H. B. M. Smit, M. L. F. Van Velthuysen, J. M. J. Spitzer-Naaykens, M. W. J. M. Wouters, W. E. Mesker, R. A. E. M. Tollenaar, and J. W. Van Sandick, "Reproducibility and validation of tumour stroma ratio scoring on oesophageal adenocarcinoma biopsies," *Eur. J. Cancer*, vol. 47, no. 3, pp. 375–382, 2011.

[57] J. H. Park, C. H. Richards, D. C. McMillan, P. G. Horgan, and C. S. D. Roxburgh, "The relationship between tumour stroma percentage, the tumour microenvironment and survival in patients with primary operable colorectal cancer," *Ann. Oncol.*, vol. 25, no. 3, pp. 644–651, 2014.

[58] S. O. Hynes, H. G. Coleman, P. J. Kelly, S. Irwin, R. F. O'Neill, R. T. Gray, C. McGready, P. D. Dunne, S. McQuaid, J. A. James, M. Salto-Tellez, and M. B. Loughrey, "Back to the future: routine morphological assessment of the tumour microenvironment is prognostic in stage II/III colon cancer in a large population-based study," *Histopathology*, vol. 71, no. 1, pp. 12–26, 2017.

[59] W. E. Mesker, G. J. Liefers, J. M. C. Junggeburt, G. W. Van Pelt, P. Alberici, P. J. K.

Kuppen, N. F. Miranda, K. A. M. Van Leeuwen, H. Morreau, K. Szuhai, R. A. E. M. Tollenaar, and H. J. Tanke, "Presence of a high amount of stroma and downregulation of SMAD4 predict for worse survival for stage I-II colon cancer patients," *Cell. Oncol.*, vol. 31, no. 3, pp. 169–178, 2009.

[60] G. W. van Pelt, T. P. Sandberg, H. Morreau, H. Gelderblom, J. H. J. M. van Krieken, R. A. E. M. Tollenaar, and W. E. Mesker, "The tumour-stroma ratio in colon cancer; the biological role and its prognostic impact," *Histopathology*, 2018.

[61] W. E. Tolles, "The Cytoanalyzer - an example of physics in medical research," *Trans. N. Y. Acad. Sci.*, vol. 17, no. 3, pp. 250–256, Nov. 1955.

[62] T. F. Institute, "Electronic computer helps detect cancer cells," *J. Franklin Inst.*, vol. 261, no. 5, pp. 587–588, 1956.

[63] J. M. Prewitt and M. L. Mendelsohn, "The analysis of cell images.," *Ann. N. Y. Acad. Sci.*, vol. 128, no. 3, pp. 1035–53, Jan. 1966.

[64] M. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener, "Histopathological Image Analysis: A Review," *IEEE Rev Biomed Eng*, vol. 2, pp. 147–171, 2009.

[65] S. Park, A. V Parwani, R. D. Aller, L. Banach, M. J. Becich, S. Borkenfeld, A. B. Carter, B. a Friedman, M. G. Rojo, A. Georgiou, G. Kayser, K. Kayser, M. Legg, C. Naugler, T. Sawai, H. Weiner, D. Winsten, and L. Pantanowitz, "The history of pathology informatics: A global perspective.," *J. Pathol. Inform.*, vol. 4, p. 7, 2013.

[66] E. Goacher, R. Randell, B. Williams, and D. Treanor, "The diagnostic concordance of whole slide imaging and light microscopy: A systematic review," *Arch. Pathol. Lab. Med.*, vol. 141, no. 1, pp. 151–161, 2017.

[67] B. Williams, D. Bottoms, and D. Treanor, "Future-proofing Pathology The case for clinical adoption of digital pathology," *J. Clin. Pathol.*, p. [Epub ahead of print], 2017.

[68] D. R. J. Snead, Y.-W. W. Tsang, A. Meskiri, P. K. Kimani, R. Crossman, N. M. Rajpoot, E. Blessing, K. Chen, K. Gopalakrishnan, P. Matthews, N. Momtahan, S. Read-Jones, S. Sah, E. Simmons, B. Sinha, S. Suortamo, Y. Yeo, H. El Daly, and I. A. Cree, "Validation of digital pathology imaging for primary histopathological diagnosis," *Histopathology*, vol. 68, no. 7, pp. 1063–1072, 2016.

[69] S. Caccomo, "FDA allows marketing of first whole slide imaging system for digital

pathology," *FDA News Release*, 2017. [Online]. Available: https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm552742.htm. [Accessed: 02-Sep-2017].

[70] D. Treanor, "Virtual slides: an introduction," *Diagnostic Histopathol.*, vol. 15, no. 2, pp. 99–103, Feb. 2009.

[71] W. C. Revie, M. Shires, P. Jackson, D. Brettle, R. Cochrane, and D. Treanor, "Color Management in Digital Pathology," *Anal. Cell. Pathol.*, vol. 2014, pp. 1–2, 2014.

[72] E. L. Clarke, C. Revie, D. Brettle, R. Wilson, C. Mello-Thoms, and D. Treanor, "Colour Calibration in Digital Pathology: the Clinical Impact of a Novel Test Object," *Diagn. Pathol.*, vol. 1, no. 8, p. 44, 2016.

[73] P. A. Bautista, N. Hashimoto, and Y. Yagi, "Color standardization in whole slide imaging using a color calibration slide.," *J. Pathol. Inform.*, vol. 5, no. i, p. 4, 2014.

[74] E. L. Clarke and D. Treanor, "Colour in digital pathology: a review," *Histopathology*, vol. 70, pp. 153–163, 2017.

[75] W. S. Campbell, G. A. Talmon, K. W. Foster, S. M. Lele, J. A. Kozel, and W. W. West, "Sixty-five thousand shades of gray: Importance of color in surgical pathology diagnoses," *Hum. Pathol.*, vol. 46, no. 12, pp. 1945–1950, 2015.

[76] E. L. Clarke, C. Mello-Thoms, D. Magee, and D. Treanor, "A response to Campbell WS, Talmon GA, Foster KW, Lele SM, Kozel JA, West WW. Sixty-five thousand shades of gray: importance of color in surgical pathology diagnoses. HUM PATHOL 2015;6:1945–50," *Hum. Pathol.*, vol. 56, pp. 204–205, 2016.

[77] E. L. Clarke, A. Sykes, D. Brettle, A. I. Wright, A. Boden, and D. E. Treanor, "Development of a novel quality assessment tool for digital microscopy," *J. Pathol. Inform.*, vol. (in press), p. ß, 2017.

[78] L. Pantanowitz and Y. Yagi, "Comment on 'Quality evaluation of microscopy and scanned histological images for diagnostic purposes': Are scanners better than microscopes?," *J. Pathol. Inform.*, vol. 3, p. 14, 2012.

[79] J. Griffin and D. Treanor, "Digital pathology in clinical use: Where are we now and what is holding us back?," *Histopathology*, vol. 70, no. 1, pp. 134–145, 2017.

[80] D. J. Hawkes, "From clinical imaging and computational models to personalised medicine and image guided interventions," *Med. Image Anal.*, vol. 33, pp. 50–55, 2016.

[81] D. Treanor, B. D. Gallas, M. A. Gavrielides, and S. M. Hewitt, "Evaluating whole slide imaging: A working group opportunity.," *J. Pathol. Inform.*, vol. 6, p. 4, 2015.

[82] T. J. Fuchs and J. M. Buhmann, "Computational pathology: Challenges and promises for tissue analysis," *Comput. Med. Imaging Graph.*, vol. 35, no. 7–8, pp. 515–530, 2011.

[83] B. J. Williams, P. DaCosta, E. Goacher, and D. Treanor, "A Systematic Analysis of Discordant Diagnoses in Digital Pathology Compared With Light Microscopy," *Arch. Pathol. Lab. Med.*, p. arpa.2016-0494-OA, 2017.

[84] D. Treanor, N. Jordan Owers, J. Hodrien, J. Wood, P. Quirke, and R. A. Ruddle, "Virtual reality Powerwall versus conventional microscope for viewing pathology slides: an experimental comparison," *Histopathology*, vol. 55, no. 1, pp. 294–300, 2009.

[85] R. A. Ruddle, R. G. Thomas, R. Randell, P. Quirke, and D. Treanor, "The Design and Evaluation of Interfaces for Navigating Gigapixel Images in Digital Pathology," *ACM Trans. Comput. Interact.*, vol. 23, no. 1, pp. 1–29, 2016.

[86] R. Randell, R. A. Ruddle, C. Mello-Thoms, R. G. Thomas, P. Quirke, and D. Treanor, "Virtual reality microscope versus conventional microscope regarding time to diagnosis: An experimental study," *Histopathology*, vol. 62, no. 2, pp. 351–358, 2013.

[87] A. Goode, B. Gilbert, J. Harkes, D. Jukic, and M. Satyanarayanan, "OpenSlide: A vendor-neutral software foundation for digital pathology.," *J. Pathol. Inform.*, vol. 4, p. 27, Jan. 2013.

[88] Open Microscopy Environment, "BioFormats | Open Microscopy Environment," *openmicroscpoy.org*, 2017. [Online]. Available: https://www.openmicroscopy.org/bio-formats/. [Accessed: 03-Sep-2017].

[89] P. J. Tadrous, "Digital stain separation for histological images," *J. Microsc.*, vol. 240, no. 2, pp. 164–172, Nov. 2010.

[90] R. Celis, D. Romo, and E. Romero, "Blind colour separation of H&E stained histological images by linearly transforming the colour space," *J. Microsc.*, vol. 0, no. 0, pp. 1–12, 2015.

[91] A. C. Ruifrok and D. A. Johnston, "Quantification of histochemical staining by color deconvolution," *Anal. Quant. Cytol. Histol.*, vol. 23, no. 4, pp. 291–299, 2001.

[92] D. L. Rimm, "What brown cannot do for you," *Nature*, vol. 200, p. 6, 2006.

[93] P. Haub and T. Meckel, "A Model based Survey of Colour Deconvolution in Diagnostic Brightfield Microscopy: Error Estimation and Spectral Consideration," *Sci. Rep.*, vol. 5, no. February, p. 12096, 2015.

[94] R. W. Conners and C. A. Harlow, "A theoretical comparison of texture algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-2, no. 3, pp. 204–222, 1980.

[95] M. Tuceryan, M. Tuceryan, A. K. Jain, and A. K. Jain, "The Handbook of Pattern Recognition and Computer Vision (2nd Edition), Texture Analysis," *Pattern Recognit.*, pp. 207–248, 1998.

[96] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man. Cybern.*, vol. 3, no. 6, pp. 610–621, 1973.

[97] C. C. Gotlieb and H. E. Kreyszig, "Texture descriptors based on co-occurrence matrices," *Comput. Vision, Graph. Image Process.*, vol. 51, no. 1, pp. 70–86, 1990.

[98] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," *IEEE Pattern Recognit.*, vol. 1, pp. 582–585, 1994.

[99] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[100] N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern Recognit.*, vol. 26, no. 9, pp. 1277–1294, 1993.

[101] Y. J. Zhang, "Evaluation and comparison of different segmentation algorithms," vol. 18, pp. 963–974, 1997.

[102] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation : A survey of unsupervised methods," *Comput. Vis. Image Underst.*, vol. 110, pp. 260–280, 2008.

[103] S. Umaa Mageswari, M. Sridevi, and C. Mala, "An experimental study and analysis of different image segmentation techniques," *Procedia Eng.*, vol. 64, pp. 36–45, 2013.

[104] Y. J. Zhang, "A Survey on Evaluation Methods for Image Segmentation," *Pattern Recognit.*, vol. 29, no. 8, pp. 1335–1346, 1996.

[105] F. Ge, S. Wang, and T. Liu, "Image-segmentation evaluation from the perspective of salient object extraction," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern*

Recognit., vol. 1, pp. 1146–1153, 2006.

[106] S. Bhattacharyya, "A brief survey of color image preprocessing and segmentation techniques," *J. Pattern Recognit. Res.*, vol. 1, pp. 120–129, 2011.

[107] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man. Cybern.*, vol. 9, no. 1, pp. 62–66, 1979.

[108] L. K. Huang and M. J. J. Wang, "Image thresholding by minimizing the measures of fuzziness," *Pattern Recognit.*, vol. 28, no. 1, pp. 41–51, 1995.

[109] Aperio Technologies Inc., "ScanScope Console User Guide MAN-0010, Revision C," San Diego, CA, USA, CA, USA, 2007.

[110] C. Jung and C. Kim, "Segmenting clustered nuclei using H-minima transform-based marker extraction and contour parameterization," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 10 PART 2, pp. 2600–2604, 2010.

[111] F. Meyer, "Topographic distance and watershed lines," *Signal Processing*, vol. 38, no. 1, pp. 113–125, Jul. 1994.

[112] F. Meyer, "The watershed concept and its use in segmentation : a brief history," pp. 1–11, 2012.

[113] N. Malpica, C. O. de Solórzano, J. J. Vaquero, a Santos, I. Vallcorba, J. M. García-Sagredo, and F. del Pozo, "Applying watershed algorithms to the segmentation of clustered nuclei.," *Cytometry*, vol. 28, no. 4, pp. 289–97, Aug. 1997.

[114] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *J. R. Stat. Soc.*, vol. 28, no. 1, pp. 100–108, 1979.

[115] K. Fukunaga and L. D. Hostetler, "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition," *IEEE Trans. Inf. Theory*, vol. 21, no. 1, pp. 32–40, 1975.

[116] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC Superpixels," *EPFL Tech. Rep. 149300*, no. June, p. 15, 2010.

[117] R. Achanta, A. Shaji, and K. Smith, "SLIC superpixels compared to state-of-the-art superpixel methods," *Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2281, 2012.

[118] Z. Wu and R. Leahy, "An optimal graph theoretic approach to data clustering: theory and\nits application to image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*,

vol. 15, no. 11, pp. 1101–1113, 1993.

[119] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[120] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *The 23rd international conference on Machine learning (ICML)*, 2006, pp. 161–168.

[121] I. Rish, "An empirical study of the naive Bayes classifier," *IJCAI 2001 Work. Empir. methods Artif. Intell.*, vol. 22230, pp. 41–46, 2001.

[122] P. Domingos and M. Pazzani, "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier," *Mach. Learn.*, vol. 768, pp. 105–112, 1996.

[123] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," *Proc. Twent. Int. Conf. Mach. Learn.*, vol. 20, no. 1973, pp. 616–623, 2003.

[124] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[125] C.-J. Hsu, Chih-WeiLin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 415–425, Jan. 2002.

[126] C. Hsu, C. Chang, and C. Lin, "A practical guide to support vector classification," vol. 1, no. 1, pp. 1–16, 2003.

[127] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

[128] S. B. Akers, "Binary Decision Diagrams," *IEEE Trans. Comput.*, vol. C-27, no. 6, pp. 509–516, 1978.

[129] Y. Mansour, "Pessimistic decision tree pruning based on tree size," *Proc. 14th Int. Conf. Mach. Learn.*, pp. 195–201, 1997.

[130] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[131] K. Xiong, "Roughened Random Forests for Binary Classification," Albany, State University of New York, 2014.

[132] G. Biau, "Analysis of a random forests model," *J. Mach. Learn. Res.*, vol. 13, pp. 1063–

1095, 2012.

[133] C. Chen, A. Liaw, and L. Breiman, “Using random forest to learn imbalanced data,” Berkelely, CA, 2004.

[134] M. Wainberg, B. Alipanahi, and B. J. Frey, “Are Random Forests Truly the Best Classifiers?,” *J. Mach. Learn. Res.*, vol. 17, no. 110, pp. 1–5, 2016.

[135] Y. Freund, R. E. Schapire, P. Avenue, and F. Park, “A Short Introduction to Boosting,” *J. Japanese Soc. Artif. Intell.*, vol. 14, no. 5, pp. 771–780, 1999.

[136] Y. Freund and R. E. Schapire, “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,” *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.

[137] D. J.E. and D. J.M., “Artificial neural networks,” *Cancer*, vol. 91, no. S8, pp. 1615–1635, 2001.

[138] S. A. Corne, “Artificial neural networks for pattern recognition,” *Concepts Magn. Reson.*, vol. 8, no. 5, pp. 303–324, 1996.

[139] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, “Learning to rank using gradient descent,” *Proceedings of the 22nd international conference on Machine learning*, vol. pages. pp. 89–96, 2005.

[140] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 9, pp. 533–536, 1986.

[141] Y. Bengio, P. Simard, and P. Frasconi, “Learning Long Term Dependencies with Gradient Descent is Difficult,” *IEEE Trans. Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[142] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[143] X. Du, Y. Cai, S. Wang, and L. Zhang, “Overview of deep learning,” *Proc. - 2016 31st Youth Acad. Annu. Conf. Chinese Assoc. Autom. YAC 2016*, pp. 159–164, 2017.

[144] G. E. Hinton, “A Practical Guide to Training Restricted Boltzmann Machines,” Tonronto, Canada, Canada, 2010.

[145] G. Hinton, “Where do features come from?,” *Cogn. Sci.*, vol. 38, no. 6, pp. 1078–1101, 2014.

[146] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets.," *Neural Comput.*, vol. 18, no. 7, pp. 1527–54, 2006.

[147] D. H. Hubel and T. N. Wiesel, "Receptive Fields and Functional Architecture of Monkey Striate Cortex," *J. Physiol.*, vol. 195, pp. 215–243, 1968.

[148] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2012.

[149] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, C. Hill, and A. Arbor, "Going Deeper with Convolutions," pp. 1–9, 2014.

[150] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," *Proc. 27th Int. Conf. Mach. Learn.*, no. 3, pp. 807–814, 2010.

[151] A. Belsare and M. Mushrif, "Histopathological Image Analysis Using Image Processing Techniques: An Overview," *Signal Image Process.*, vol. 3, no. 4, pp. 23–36, 2012.

[152] A. Kårsnäs, *Image Analysis Methods and Tools for Digital Histopathology Applications Relevant to Breast Cancer Diagnosis*. 2014.

[153] A. Madabhushi, "Digital pathology image analysis: opportunities and challenges," *Imaging Med.*, vol. 1, no. 1, pp. 7–10, Oct. 2009.

[154] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Med. Image Anal.*, vol. 33, pp. 170–175, 2016.

[155] M. G. Rojo, G. Bueno, and J. Slodkowska, "Review of imaging solutions for integrated quantitative immunohistochemistry in the Pathology daily practice.," *Folia Histochem. Cytobiol.*, vol. 47, no. 3, pp. 349–54, Jan. 2009.

[156] J. H. Sinard, "Review of ' digital pathology ' by Yves Sucaet and Wim Waelput," *J. Pathol.*, vol. 6, no. 12, pp. 11–12, 2015.

[157] T. Patel, "Review of 'Digital Pathology' by Liron Pantanowitz and Anil V Parwani," *J. Pathol. Inform.*, vol. 8, no. 37, pp. 9–10, 2017.

[158] L. Csink, D. Paulus, U. Ahlrichs, and B. Heigl, "Color Normalization and Object Localization Color Normalization and Object," *Rev. Lit. Arts Am.*, 1998.

[159] D. Magee, D. Treanor, P. Chomphuwiset, and P. Quirke, "Context Aware Colour Classification in Digital Microscopy," in *Medical Image Understanding and Analysis*,

2011, pp. 1–5.

[160] J. N. Kather, C.-A. Weis, A. Marx, A. K. Schuster, L. R. Schad, and F. G. Zöllner, "New Colors for Histology: Optimized Bivariate Color Maps Increase Perceptual Contrast in Histological Images.," *PLoS One*, vol. 10, no. 12, p. e0145572, 2015.

[161] J. Quintana, R. Garcia, and L. Neumann, "A novel method for color correction in epiluminescence microscopy," *Comput. Med. Imaging Graph.*, vol. 35, no. 7–8, pp. 646–652, 2011.

[162] X. Li and K. N. Plataniotis, "A Complete Color Normalization Approach to Histopathology Images Using Color Cues Computed From Saturation-Weighted Statistics," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 7, pp. 1862–1873, 2015.

[163] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 6, pp. 1729–1738, 2014.

[164] D. Bug, S. Schneider, A. Grote, E. Oswald, F. Feuerhake, D. Merhof, C. Vision, and O. Gmbh, "Context-based Normalization of Histological Stains using Deep Convolutional Features," no. Dlmia, pp. 1–6, 2017.

[165] D. Magee, D. Treanor, D. Crellin, M. Shires, K. Mohee, and P. Quirke, "Colour Normalisation in Digital Histopathology Images," in *Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*, 2009, pp. 100–111.

[166] M. Tipping, "Sparse Bayesian Learning and the Relevance Vector Mach," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[167] S. Wen, T. M. Kurc, Y. Goa, T. Zhao, J. Saltz, and W. Zhu, "A Methodology for Texture Feature-based Quality Assessment in Nucleus Segmentation of Histopathology Image," *J. Pathol. Inform.*, vol. 8, no. 38, pp. 1–12, 2017.

[168] C. Arteta, V. Lempitsky, J. A. Noble, and A. Zisserman, "Learning to detect cells using non-overlapping extremal regions.," in *MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2012, vol. 15, no. Pt 1, pp. 348–56.

[169] L. P. Coelho, A. Shariff, and R. F. Murphy, "Nuclear segmentation in microscope cell images: A hand-segmented dataset and comparison of algorithms," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2009. ISBI '09*,

2009, pp. 518–521.

[170] H. Kong, M. Gurcan, and K. Belkacem-Boussaid, "Partitioning histopathological images: an integrated framework for supervised color-texture segmentation and cell splitting.," *IEEE Trans. Med. Imaging*, vol. 30, no. 9, pp. 1661–77, Sep. 2011.

[171] J. Diamond, N. H. Anderson, P. H. Bartels, R. Montironi, and P. W. Hamilton, "The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia," *Hum. Pathol.*, vol. 35, no. 9, pp. 1121–1131, Sep. 2004.

[172] J. Shu, H. Fu, G. Qiu, P. Kaye, and M. Ilyas, "Segmenting overlapping cell nuclei in digital histopathology images," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 5445–5448, 2013.

[173] K. Ali, A. Jalil, M. U. Gull, and M. Fiaz, "Medical image segmentation using h-minima transform and region merging technique," *Proc. - 2011 9th Int. Conf. Front. Inf. Technol. FIT 2011*, pp. 127–132, 2011.

[174] S. Navlakha, P. Ahammad, and E. W. Myers, "Unsupervised segmentation of noisy electron microscopy images using salient watersheds and region merging.," *BMC Bioinformatics*, vol. 14, no. 1, p. 294, Jan. 2013.

[175] P. J. Schüffler, D. Schapiro, C. Giesen, H. A. O. Wang, B. Bodenmiller, and J. M. Buhmann, "Automatic single cell segmentation on highly multiplexed tissue images," *Cytom. Part A*, vol. 87, no. 10, pp. 936–942, 2015.

[176] Y. Zhou, D. Magee, D. Treanor, and A. Bulpitt, "Stain guided mean-shift filtering in automatic detection of human tissue nuclei," *J. Pathol. Inform.*, vol. 4, no. 6, p. 6, 2013.

[177] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images.," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 841–52, Apr. 2010.

[178] J. P. Vink, M. B. Van Leeuwen, C. H. M. Van Deurzen, and G. De Haan, "Efficient nucleus detector in histopathology images.," *J. Microsc.*, vol. 249, no. 2, pp. 124–35, Feb. 2013.

[179] P. Bamford and B. Lovell, "Unsupervised cell nucleus segmentation with active contours," *Signal Processing*, vol. 71, no. 2, pp. 203–213, Dec. 1998.

[180] D. Randell, A. Galton, S. Fouad, H. Mehanna, and G. Landini, *Model-Based Correction*

of Segmentation Errors in Digitised Histological Images. 2017.

[181] K. Sirinukunwattana, S. Raza, Y. Tsang, D. Snead, I. Cree, and N. Rajpoot, "Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images," *IEEE Trans. Med. Imaging*, vol. 62, no. February, pp. 1–1, 2016.

[182] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, "Methods for nuclei detection, segmentation, and classification in digital histopathology: A review-current status and future potential," *IEEE Rev. Biomed. Eng.*, vol. 7, pp. 97–114, 2014.

[183] A. Bennett, Y. Zhu, A. Wright, C. Verbeke, L. Hodgkin, D. Magee, V. Spiers, and D. Treanor, "A novel nuclear detection algorithm for the automatic analysis of immunohistochemistry staining," in *Joint Meeting of the Pathological Society of Great Britain and Ireland and the Dutch Pathological Society*, 2009.

[184] R. O. Duda and P. E. Hart, "Use of the Hough transformation to detect lines and curves in pictures," *Commun. ACM*, vol. 15, no. 1, pp. 11–15, 1972.

[185] D. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognit.*, vol. 13, no. 2, pp. 111–122, 1981.

[186] P. Chomphuwiset, D. Magee, and R. Boyle, "Nucleus classification and bile duct detection in liver histology," in *MICCAI Workshop on Machine Learning in Medical Imaging*, 2010, pp. 1–8.

[187] A. B. Tosun, M. Kandemir, C. Sokmensuer, and C. Gunduz-Demir, "Object-oriented texture analysis for the unsupervised segmentation of biopsy images for cancer detection," *Pattern Recognit.*, vol. 42, no. 6, pp. 1104–1112, Jun. 2009.

[188] K. Nguyen, A. K. Jain, and B. Sabata, "Prostate cancer detection: Fusion of cytological and textural features.," *J. Pathol. Inform.*, vol. 2, p. S3, Jan. 2011.

[189] A. B. Tosun and C. Gunduz-demir, "Graph Run-Length Matrices for Histopathological Image Segmentation," vol. 30, no. 3, pp. 721–732, 2011.

[190] S. A. Devaraj, T. Kumarasamy, V. B. G. Ganesh, C. C. C. Ranjan, A. G. Narayanan, A. Prof, B. E. Scholar, and F. Xavier, "Normalized cuts based segmentation of gastroentrology images using visual features," vol. 1, no. 11, pp. 1376–1382, 2014.

[191] J. Xu, A. Madabhushi, A. Janowczyk, and S. Chandran, "A weighted mean shift, normalized cuts initialized color gradient based geodesic active contour model:

applications to histopathology image segmentation," *Proc. SPIE*, vol. 7623, no. April 2016, p. 76230Y–76230Y–12, 2010.

[192] M. Veta, J. P. W. Pluim, P. J. Van Diest, and M. A. Viergever, "Breast cancer histopathology image analysis: A review," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 5, pp. 1400–1411, 2014.

[193] Y. Gao, W. Liu, S. Arjun, L. Zhu, V. Ratner, T. Kurc, J. Saltz, and A. Tannenbaum, "Multi-scale learning based segmentation of glands in digital colonrectal pathology images," *Proc. SPIE*, vol. 9791. p. 97910M–97910M–6, 2016.

[194] V. Roullier, O. Lézoray, V. T. Ta, and A. Elmoataz, "Multi-resolution graph-based analysis of histopathological whole slide images: Application to mitotic cell extraction and visualization," *Comput. Med. Imaging Graph.*, vol. 35, no. 7–8, pp. 603–615, 2011.

[195] H. Kalkan, M. Nap, R. P. W. Duin, and M. Loog, "Automated colorectal cancer diagnosis for whole-slice histopathology.," *Med. Image Comput. Comput. Assist. Interv.*, vol. 15, no. Pt 3, pp. 550–7, Jan. 2012.

[196] K. Sirinukunwattana, D. R. J. Snead, and N. M. Rajpoot, "A Stochastic Polygons Model for Glandular Structures in Colon Histology Images," *IEEE Trans. Med. Imaging*, vol. 34, no. 11, pp. 2366–2378, 2015.

[197] P. Chomphuwiset, D. Treanor, and D. Magee, "Liver Cell Type quantification Using Local appearance and Global Context," *Mach. Vis. Appl.*

[198] C. W. Wang, "Robust automated tumour segmentation on histological and immunohistochemical tissue images," *PLoS One*, vol. 6, no. 2, 2011.

[199] J. P. Monaco, J. E. Tomaszewski, M. D. Feldman, M. Moradi, P. Mousavi, A. Boag, C. Davidson, P. Abolmaesumi, and A. Madabhushi, "Detection of prostate cancer from whole-mount histology images using Markov random fields," *Microsc. Image Anal. Appl. Biol. {(MIAAB).} New York, {NY}, {USA}*, pp. 5–6, 2008.

[200] S. Reis, P. Gazinska, J. H. Hipwell, T. Mertzanidou, K. Naidoo, N. Williams, S. Pinder, and D. J. Hawkes, "Automated Classification of Breast Cancer Stroma Maturity from Histological Images," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 10, pp. 2344–2352, 2017.

[201] B. Ehteshami Bejnordi, J. Lin, B. Glass, M. Mullooly, G. L. Gierach, M. E. Sherman, N. Karssemeijer, J. Van Der Laak, and A. H. Beck, "Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images," *Proc. -*

Int. Symp. Biomed. Imaging, pp. 929–932, 2017.

[202] P. Kainz, M. Pfeiffer, and M. Urschler, "Segmentation and classification of colon glands with deep convolutional neural networks and total variation regularization," *PeerJ*, vol. 5, p. e3874, 2017.

[203] J. N. Kather, C.-A. Weis, F. Bianconi, S. M. Melchers, L. R. Schad, T. Gaiser, A. Marx, and F. G. Zöllner, "Multi-class texture analysis in colorectal cancer histology.," *Nat. Sci. reports*, vol. 6, no. 27988, pp. 1–11, 2016.

[204] J.-M. Chen, A.-P. Qu, L.-W. Wang, J.-P. Yuan, F. Yang, Q.-M. Xiang, N. Maskey, G.-F. Yang, J. Liu, and Y. Li, "New breast cancer prognostic factors identified by computer-aided image analysis of HE stained histopathology images," *Sci. Rep.*, vol. 5, no. 1, p. 10690, 2015.

[205] R. Rogojanu, T. Thalhammer, U. Thiem, A. Heindl, I. Mesteri, A. Seewald, W. Jäger, C. Smochina, I. Ellinger, and G. Bises, "Quantitative Image Analysis of Epithelial and Stromal Area in Histological Sections of Colorectal Cancer: An Emerging Diagnostic Tool," *Biomed Res. Int.*, vol. 2015, 2015.

[206] F. Bianconi, A. Álvarez-Larrán, and A. Fernández, "Discrimination between tumour epithelium and stroma via perception-based features," *Neurocomputing*, vol. 154, pp. 119–126, 2015.

[207] P. W. Hamilton, Y. Wang, C. Boyd, J. A. James, M. B. Loughrey, J. P. Hougton, D. P. Boyle, P. Kelly, P. Maxwell, D. McCleary, J. Diamond, D. G. McArt, J. Tunstall, P. Bankhead, and M. Salto-Tellez, "Automated tumor analysis for molecular profiling in lung cancer.," *Oncotarget*, vol. 6, no. 29, pp. 27938–52, 2015.

[208] J. Pääkkönen, N. Päivinen, M. Nykänen, and T. Paavonen, "An automated gland segmentation and classification method in prostate biopsies: an image source-independent approach," *Mach. Vis. Appl.*, vol. 26, no. 1, pp. 103–113, 2014.

[209] A. van Engelen, W. J. Niessen, S. Klein, H. C. Groen, K. van Gaalen, H. J. Verhagen, J. J. Wentzel, A. van der Lugt, and M. de Bruijne, "Automated segmentation of atherosclerotic histology based on pattern classification.," *J. Pathol. Inform.*, vol. 4, no. Suppl, p. S3, Jan. 2013.

[210] S. J. McKenna, T. Amaral, S. Akbar, L. Jordan, and A. Thompson, "Immunohistochemical analysis of breast tissue microarray images using contextual classifiers.," *J. Pathol. Inform.*, vol. 4, no. Suppl, p. S13, Jan. 2013.

[211] H. K. Angell, N. Gray, C. Womack, D. I. Pritchard, R. W. Wilkinson, and M. Cumberbatch, "Digital pattern recognition-based image analysis quantifies immune infiltrates in distinct tissue regions of colorectal cancer and identifies a metastatic phenotype," *Br. J. Cancer*, vol. 109, no. 6, pp. 1618–1624, 2013.

[212] T. Mattfeldt, P. Grahovac, and S. Lück, "Multiclass Pattern Recognition of the Gleason Score of Prostatic Carcinomas Using Methods of Spatial Statistics," *Image Anal. Stereol.*, vol. 32, no. 3, p. 155, 2013.

[213] N. Linder, J. Konsti, R. Turkki, E. Rahtu, M. Lundin, S. Nordling, C. Haglund, T. Ahonen, M. Pietikäinen, and J. Lundin, "Identification of tumor epithelium and stroma in tissue microarrays using texture analysis.," *Diagn. Pathol.*, vol. 7, no. 1, p. 22, Jan. 2012.

[214] S. Doyle, M. D. Feldman, N. Shih, J. Tomaszewski, and A. Madabhushi, "Cascaded discrimination of normal, abnormal, and confounder classes in histopathology: Gleason grading of prostate cancer," *BMC Bioinformatics*, vol. 13, no. 1, 2012.

[215] A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. van de Vijver, R. B. West, M. van de Rijn, and D. Koller, "Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival," *Sci. Transl. Med.*, vol. 3, no. 108, pp. 108–113, Nov. 2011.

[216] Y. N. Law, a M. Yip, and H. K. Lee, "Automatic measurement of volume percentage stroma in endometrial images using texture segmentation.," *J. Microsc.*, vol. 241, no. 2, pp. 171–8, Feb. 2011.

[217] M. Eramian, M. Daley, D. Neilson, and T. Daley, "Segmentation of epithelium in H&E stained odontogenic cysts," *J. Microsc.*, vol. 244, no. 3, pp. 273–292, 2011.

[218] N. Signolle, M. Revenu, B. Plancoulaine, and P. Herlin, "Wavelet-based multiscale texture segmentation: Application to stromal compartment characterization on virtual slides," *Signal Processing*, vol. 90, no. 8, pp. 2412–2422, 2010.

[219] L. Yang, W. Chen, P. Meer, G. Salaru, L. A. Goodell, V. Berstis, and D. J. Foran, "Virtual microscopy and grid-enabled decision support for large-scale analysis of imaged pathology specimens," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 4, pp. 636–644, 2009.

[220] P.-W. Huang and C.-H. Lee, "Automatic classification for pathological prostate images based on fractal analysis.," *IEEE Trans. Med. Imaging*, vol. 28, no. 7, pp. 1037–1050,

2009.

[221] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," in *Biomedical Imaging: From Nano to Macro, 2008. {ISBI} 2008. 5th {IEEE} International Symposium on*, 2008, pp. 284–287.

[222] M. Datar, D. Padfield, and H. Cline, "Color and texture based segmentation of molecular pathology images using HSOMS," *IEEE Int. Symp. Biomed. Imaging*, pp. 292–295, 2008.

[223] N. Rajpoot, "Texture classification using discriminant wavelet packet subbands," *Circuits Syst. 2002. MWSCAS-2002. …*, pp. 3–6, 2002.

[224] M. D. DiFranco, G. O'Hurley, E. W. Kay, R. W. G. Watson, and P. Cunningham, "Ensemble based system for whole-slide prostate cancer probability mapping using color texture features," *Comput. Med. Imaging Graph.*, vol. 35, no. 7–8, pp. 629–645, 2011.

[225] A. Adam, A. J. Bulpitt, and D. Treanor, "Texture Analysis of Virtual Slides for Grading Dysplasia in Barrett ' s Oesophagus .," in *Medical Image Understanding and Analysis*, 2011, pp. 1–5.

[226] K. Rajpoot and N. Rajpoot, "SVM optimization for hyperspectral colon tissue cell classification," in *MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2004, pp. 829–837.

[227] N. Aggarwal and R. K. Agrawal, "First and Second Order Statistics Features for Classification of Magnetic Resonance Brain Images," *J. Signal Inf. Process.*, vol. 3, no. 2, pp. 146–153, 2012.

[228] D. Gabor, "Theory of communication," *J. Inst. Electr. Eng.*, vol. 93, pp. 429–457, 1946.

[229] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[230] K. Sirinukunwattana, J. P. W. W. Pluim, H. Chen, X. Qi, P.-A. A. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, A. Böhm, O. Ronneberger, B. Ben Cheikh, D. Racoceanu, P. Kainz, M. Pfeiffer, M. Urschler, D. R. J. J. Snead, and N. M. Rajpoot, "Gland segmentation in colon histology images: The glas challenge contest," *Med. Image Anal.*, vol. 35, pp. 489–502, 2017.

[231] R. Marée, "Deep Learning for Classification of Colorectal Polyps on Whole-slide

Images," *J. Pathol. Informatics | Publ. by Wolters Kluwer -Medknow*, vol. 1, no. 19, pp. 1–4, 2017.

[232] J. Xu, X. Luo, G. Wang, H. Gilmore, and A. Madabhushi, "A Deep Convolutional Neural Network for segmenting and classifying epithelial and stromal regions in histopathological images," *Neurocomputing*, vol. 191, pp. 214–223, 2016.

[233] R. M. Levenson, E. A. Krupinski, V. M. Navarro, and E. A. Wasserman, "Pigeons (Columba livia) as Trainable Observers of Pathology and Radiology Breast Cancer Images.," *PLoS One*, vol. 10, no. 11, pp. 1–21, 2015.

[234] P. Shrestha, R. Kneepkens, J. Vrijnsen, D. Vossen, E. Abels, and B. Hulsken, "A quantitative approach to evaluate image quality of whole slide imaging scanners," *J. Pathol. Inform.*, vol. 7, no. 1, p. 56, 2016.

[235] N. Hashimoto, M. Yamaguchi, Y. Yagi, P. Bautista, and N. Ohyama, "Referenceless image quality evaluation for whole slide imaging," *J. Pathol. Inform.*, vol. 3, no. 1, p. 9, 2012.

[236] D. Ameisen, C. Deroulers, V. Perrier, and F. Bouhidel, "Automatic Image Quality Assessment in Digital Pathology: From Idea to Implementation," *Iwbbio*, pp. 148–157, 2014.

[237] A. R. N. Avanaki, K. S. Espig, A. Xthona, C. Lanciault, and T. R. L. Kimpe, "Automatic Image Quality Assessment for Digital Pathology," in *Breast Imaging: 13th International Workshop, IWDM 2016, Malm{ö}, Sweden, June 19-22, 2016, Proceedings*, A. Tingberg, K. Lång, and P. Timberg, Eds. Cham: Springer International Publishing, 2016, pp. 431–438.

[238] F. Aeffner, K. Wilson, N. T. Martin, J. C. Black, C. L. L. Hendriks, B. Bolon, D. G. Rudmann, R. Gianani, S. R. Koegler, J. Krueger, and G. D. Young, "The Gold Standard Paradox in Digital Image Analysis: Manual Versus Automated Scoring as Ground Truth," *Arch. Pathol. Lab. Med.*, vol. 141, no. September, p. arpa.2016-0386-RA, 2017.

[239] J. Weese and C. Lorenz, "Four challenges in medical image analysis from an industrial perspective," *Med. Image Anal.*, vol. 33, pp. 1339–1351, 2016.

[240] Definiens, "Research | Developer XD," *www.definiens.com*, 2017. [Online]. Available: http://www.definiens.com/solutions/research. [Accessed: 12-Sep-2017].

[241] S. S. Lee, H. Lee, P. Abbeel, and A. Y. A. Ng, "Efficient L 1 Regularized Logistic

Regression," *Compute*, vol. 21, no. 1, p. 401, 2004.

[242] E. L. Kaplan and P. Meier, "Nonparametric Estimation from Incomplete Observations," *J. Am. Stat. Assoc.*, vol. 5318910, no. 282, pp. 457–481, 1958.

[243] H. J. Gundersen, P. Bagger, T. F. Bendtsen, S. M. Evans, L. Korbo, N. Marcussen, A. Møller, K. Nielsen, J. R. Nyengaard, B. Pakkenberg, F. B. SØRensen, A. Vesterby, and M. J. West, "The new stereological tools: disector, fractionator, nucleator and point sampled Diagnosis., intercepts and their use in pathological research and diagnosis," *APMIS*, vol. 96, no. 7–12, pp. 857–881, 1988.

[244] L. M. Cruz-orive and R. Weibel, "Recent stereological a brief survey methods for cell biology :," *Am. J. Physiol.*, vol. 256, no. L, pp. 148–156, 1990.

[245] D. Treanor, M. Dattani, P. Quirke, and H. Grabsch, "Systematic Random Sampling with Virtual Slides: A New Software Tool For Tissue Research," in *Abstract J Pathol*, 2008, p. 216.

[246] D. Hill, "Practical Example: Placing a sampling grid in a polygon," *MATLAB Central*, 2008. [Online]. Available: http://blogs.mathworks.com/videos/2008/03/11/practical-example-placing-a-sampling-grid-in-a-polygon/.

[247] N. West, H. Grabsch, D. Treanor, D. Sebag-Montefiore, H. Thorpe, D. Jayne, H. Rutten, H. Swellengrebel, I. Nagtegaal, and P. Quirke, "Quantitative assessment of tumor cell density in rectal cancer following three different preoperative therapies compared to surgery alone," *Clin. Oncol.*, vol. 28, p. 15s, 2010.

[248] M. Elmoursi, D. Treanor, and N. A. B. Simpson, "Novel insights from using Stereology based volume estimation of Syncytial nuclear aggregates in Diabetic placenta," *Arch. Dis. Childhood. Fetal Neonatal Ed.*, vol. 99, no. 1, p. A158, 2014.

[249] M. Elmoursi, J. Stahlshmidt, D. Treanor, and N. A. B. Simpson, "Syncytial nuclear aggregates and villous capillary volume in IUGR placentas: A Stereology-based study on virtual slides," *Placenta*, vol. 35, no. 9, p. A11, 2014.

[250] M. D. Hale, M. G. Nankivell, W. Mueller, N. P. West, S. P. Stenning, A. I. Wright, D. E. Treanor, R. E. Langley, L. C. Ward, W. H. Allum, D. Cunningham, J. D. Hayden, and H. I. Grabsch, "The relationship between tumour cell density in the pre-treatment biopsy and survival after chemotherapy in OE02 trial oesophageal cancer patients," *J. Clin. Oncol.*, vol. 32, no. 3, p. 49, 2014.

[251] S. Earle, T. Aoyama, A. I. Wright, D. E. Treanor, Y. Miyagi, L. C. Hewitt, J. D. Hayden, T. Yoshikawa, H. I. Thygessen, and G. H. I, "Prognostic and predictive value of tumor-infiltrating immune cells in Japanese patients with stage II/III gastric cancer," *J. Clin. Oncol.*, vol. 32, no. 3, p. 46, 2014.

[252] K. J. Pettinger, L. Ward, A. Wright, D. Treanor, and H. I. Grabsch, "Immune Cell Density is Associated with Tumour Cell Density, Histological Subtype, Depth of Invasion and Outcome in Gastric Cancer," in *Journal of Pathology*, 2013, vol. 229, p. S21.

[253] QUASAR Collaborative Group, "Adjuvant chemotherapy versus observation in patients with colorectal cancer: a randomised study.," *Lancet*, vol. 370, no. 9604, pp. 2020–9, Dec. 2007.

[254] G. Hutchins, K. Southward, K. Handley, L. Magill, C. Beaumont, J. Stahlschmidt, S. Richman, P. Chambers, M. Seymour, D. Kerr, R. Gray, and P. Quirke, "Value of mismatch repair, KRAS, and BRAF mutations in predicting recurrence and benefits from chemotherapy in colorectal cancer," *J. Clin. Oncol.*, vol. 29, no. 10, pp. 1261–1270, 2011.

[255] D. Giavarina, "Understanding Bland Altman analysis.," *Biochem. medica*, vol. 25, no. 2, pp. 141–51, 2015.

[256] M. L. Mc Hugh, "Interrater reliability : the Kappa Statistic," *Biochem Med*, vol. 22, no. 3, pp. 276–282, 2012.

[257] P. Chomphuwiset, D. R. Magee, R. D. Boyle, and D. Treanor, "Context-Based Classification of Cell Nuclei and Tissue Regions in Liver Histopathology," in *Medical Image Understanding and Analysis*, 2011, pp. 1–5.

[258] C. H. Lim, D. E. Treanor, M. F. Dixon, and A. T. R. Axon, "Low-grade dysplasia in Barrett's esophagus has a high risk of progression," *Endoscopy*, vol. 39, no. 7, pp. 581–587, 2007.

[259] A. Wright, D. Magee, P. Quirke, and D. E. Treanor, "Towards automatic patient selection for chemotherapy in colorectal cancer trials," in *SPIE 9041, Medical Imaging 2014: Digital Pathology*, 2014, p. 90410A.

[260] J. D. Hipp, D. R. Lucas, M. R. Emmert-Buck, C. C. Compton, and U. J. Balis, "Digital slide repositories for publications: lessons learned from the microarray community.," *Am. J. Surg. Pathol.*, vol. 35, no. 6, pp. 783–6, Jun. 2011.

[261] J. D. Hipp, S. C. Smith, J. Sica, D. Lucas, J. A. Hipp, L. P. Kunju, and U. J. Balis, "Tryggo: Old norse for truth: The real truth about ground truth: New insights into the challenges of generating ground truth maps for WSI CAD algorithm evaluation.," *J. Pathol. Inform.*, vol. 3, p. 8, Jan. 2012.

[262] A. Laurinavicius, A. Laurinaviciene, D. Dasevicius, N. Elie, B. Plancoulaine, C. Bor, and P. Herlin, "Digital image analysis in pathology: benefits and obligation," *Anal. Cell. Pathol.*, vol. 35, no. 2, pp. 75–78, 2012.

[263] J. D. Hipp, J. Sica, B. McKenna, J. Monaco, A. Madabhushi, J. Cheng, and U. J. Balis, "The need for the pathology community to sponsor a whole slide imaging repository with technical guidance from the pathology informatics community.," *J. Pathol. Inform.*, vol. 2, p. 31, Jan. 2011.

[264] S.-C. Zhu, C. Guo, Y. Wang, and Z. Xu, "What are Textons?," *Int. J. Comput. Vis.*, vol. 62, no. 1/2, pp. 121–143, Apr. 2005.

[265] D. Sebag-Montefiore, R. J. Stephens, R. Steele, J. Monson, R. Grieve, S. Khanna, P. Quirke, J. Couture, C. de Metz, A. S. Myint, E. Bessell, G. Griffiths, L. C. Thompson, and M. Parmar, "Preoperative radiotherapy versus selective postoperative chemoradiotherapy in patients with rectal cancer (MRC CR07 and NCIC-CTG C016): a multicentre, randomised trial," *Lancet*, vol. 373, no. 9666, pp. 811–820, 2009.

[266] F. J. Candido dos Reis, S. Lynn, H. R. Ali, D. Eccles, A. Hanby, E. Provenzano, C. Caldas, W. J. Howat, L. A. McDuffus, B. Liu, F. Daley, P. Coulson, R. J. Vyas, L. M. Harris, J. M. Owens, A. F. M. Carton, J. P. McQuillan, A. M. Paterson, Z. Hirji, S. K. Christie, A. R. Holmes, M. K. Schmidt, M. Garcia-Closas, D. F. Easton, M. K. Bolla, Q. Wang, J. Benitez, R. L. Milne, A. Mannermaa, F. Couch, P. Devilee, R. A. E. M. Tollenaar, C. Seynaeve, A. Cox, S. S. Cross, F. M. Blows, J. Sanders, R. de Groot, J. Figueroa, M. Sherman, M. Hooning, H. Brenner, B. Holleczek, C. Stegmaier, C. Lintott, and P. D. P. Pharoah, "Crowdsourcing the General Public for Large Scale Molecular Pathology Studies in Cancer," *EBioMedicine*, vol. 2, no. 7, pp. 681–689, 2015.

[267] A. R. N. Avanaki, K. S. Espig, A. Xthona, C. Lanciault, and T. R. L. Kimpe, *Automatic image quality assessment for digital pathology*. 2016.

[268] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2013.

[269] A. I. Wright, *Using computer vision to improve cancer patient selection for*

chemotherapy. UK: YouTube, 2013.

Appendices

Appendix A - Publications arising from thesis work

1. Wright, A. I., Magee, D. R., Quirke, P., Treanor, D. E. (2016). Incorporating Local and Global Context for Better Automated Analysis of Colorectal Cancer on Digital Pathology Slides. *20th Conference on Medical Image Understanding and Analysis (MIUA 2016) Proc Comp Science*. Vol 90, pp 125-131.
2. Wright, A. I., Magee, D. R., Quirke, P., Treanor, D. E. (2015). Prospector: A web-based tool for rapid acquisition of gold standard data for pathology research and image analysis. *Journal of Pathology Informatics*. 2015; 6:21
3. Wright, A. I., Grabsch, H. I., Treanor, D. E. (2015). RandomSpot: A web-based tool for systematic random sampling of virtual slides. *Journal of Pathology Informatics*. 2015; 6:8.
4. Wright, A. I., Magee, D., R., Quirke, P., Treanor, D. (2014). Towards automatic patient selection for chemotherapy in colorectal cancer trials. Proc. SPIE 9041, Medical Imaging 2014: Digital Pathology, 90410A (March 20, 2014).

Appendix B - Presentations arising from thesis work

B.1 International oral presentations

Towards automatic patient selection for chemotherapy in colorectal cancer trials. (SPIE Medical Imaging, San Diego Conference Center, CA, United States. February 2014)

B.2 National oral presentations

Incorporating Local and Global Context for Better Automated Analysis of Colorectal Cancer on Digital Pathology Slides. (MIUA: Medical Image Understanding and Analysis, Loughborough University, UK. July 2016)

B.3 Local oral presentations

Automated analysis of virtual slides: are we nearly there yet? (Leeds Institutes of Molecular Medicine Postgraduate Symposium. Leeds, UK. April 2015)

B.4 Poster presentations

Wright A, Treanor D. (2013). Automatic analysis to calculate the tumour:stroma ratio in colorectal cancer. Presented at: Leeds Institute of Molecular Medicine postgraduate research symposium. University of Leeds, UK.

Wright A, Hutchins G, West N, Toh E, Magee D, Quirke P, Treanor D (2012). An automatic approach to improving patient selection for chemotherapy. Presented at: Showcase 2012 - The 3rd annual University of Leeds postgraduate research conference. University of Leeds, UK.

Wright A, Coe A, Dattani M, Toh E, Hutchins G, West N, Grabsch H, Magee D, Quirke P, Treanor D (2012). Automatic Image Analysis to Calculate the Cancer:Stroma Ratio in Colorectal Cancer. Presented at: Joint Meeting of the Pathological Society of Great Britain and Ireland and the Dutch Pathological Society. University of Sheffield, UK.

Appendix C - Awards arising from thesis work

2014

Robert F. Wagner Student Paper Award Digital Pathology Conference Finalist. *"Towards automated patient selection for chemotherapy in colorectal cancer trials"*. SPIE Medical Imaging Symposium, 17th February 2014, San Diego, US.

2013

Research Image of the Year (1st Place) *"Stitching it all together – my PhD Life as a Patchwork Quilt"*. University of Leeds 4th Annual Postgraduate Research Conference, Showcase, December 2013, University of Leeds, UK.

Research Video of the Year (1st Place) "*Using computer vision to improve cancer patient selection for chemotherapy".* [269] University of Leeds 4th Annual Postgraduate Research Conference, Showcase, December 2013, University of Leeds, UK.

2012

Public Engagement "Twitter Thesis" Competition (1st Place) *"Automatically Improving Patient Selection for Chemotherapy"*. University of Leeds 3rd Annual Postgraduate Research Conference, Showcase, December 2012, University of Leeds, UK.

- Interview printed in Medicine Maters Magazine, *"The future of research communication?"* August 2013, issue 37, p15

Research Image of the Year (2nd Place) *"Using Image Patches to Understand the Bigger Picture in Cancer."* University of Leeds 3rd Annual Postgraduate Research Conference, Showcase, December 2012, University of Leeds, UK.

- Printed in Times Higher Education Magazine, *"This mosaic of a pathologist diagnosing cancer uses more than 100,000 "patch" images of patient biopsies"*. January 2013, issue 2082, p10
- Also Printed in Leeds Reporter Magazine, January 2013, issue 571, p6

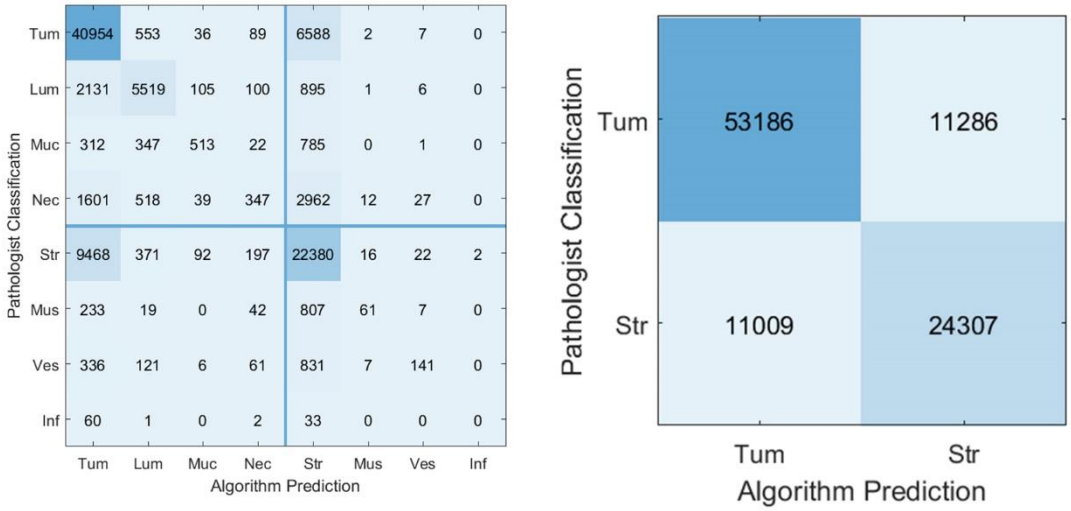
2011

University of Leeds 2nd Annual Postgraduate Research Conference, Showcase: Research Image of the Year (1st Place) – Is Technology the Missing Piece of the Puzzle for Cancer?

- Featured on Times Higher Education Website – "Problem solved? Postgraduate research image of the year portrays high tech solution to cancer puzzle". [1]
- Printed in Medicine Maters Magazine, *"Is technology the missing piece of the puzzle for cancer?"* August 2012, issue 35, p15
- Also Printed in Leeds Reporter Magazine, January 2012, issue 563, p4

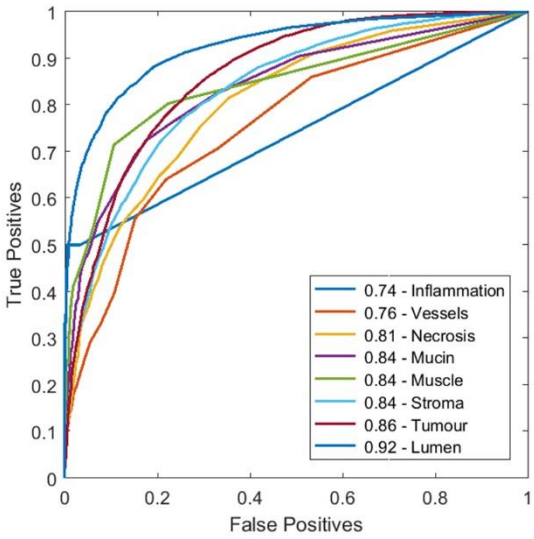
Appendix D - Algorithm results figures

D.1 - Algorithm A: Fixed patch size (64x64 pixels) algorithm



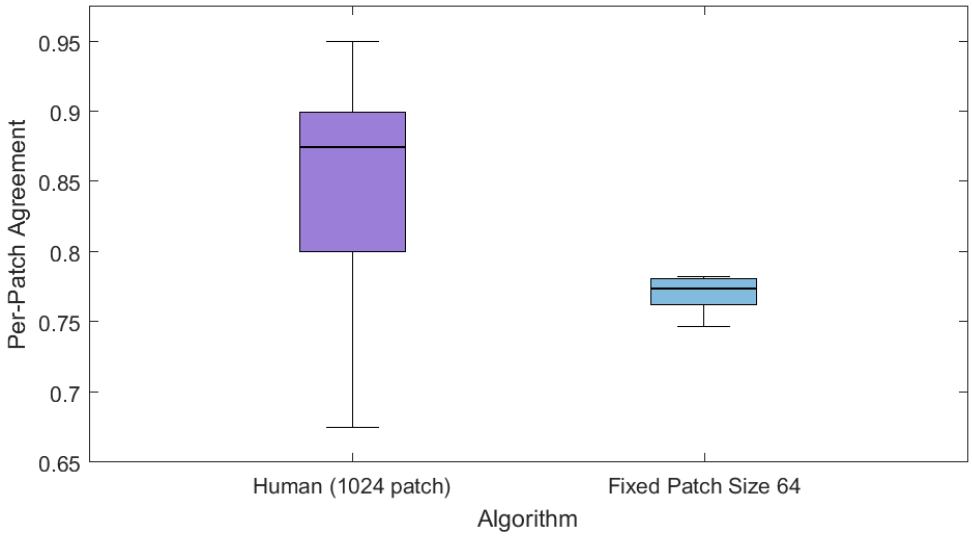
Confusion matrices showing pathologist – Fixed-Patch algorithm agreement

Left: Confusion matrix of all 8 tissue subtypes from dataset. Accuracy = 70.06%, sensitivity (true positive rate / recall) = 0.70, kappa = 0.51 (moderate agreement)
Right: Confusion matrix of subtypes grouped into Tumour and Stroma parent classes. Accuracy = 77.66%, sensitivity (true positive rate / recall) = 0.82, specificity (true negative rate) = 0.69, kappa = 0.51 (moderate agreement)



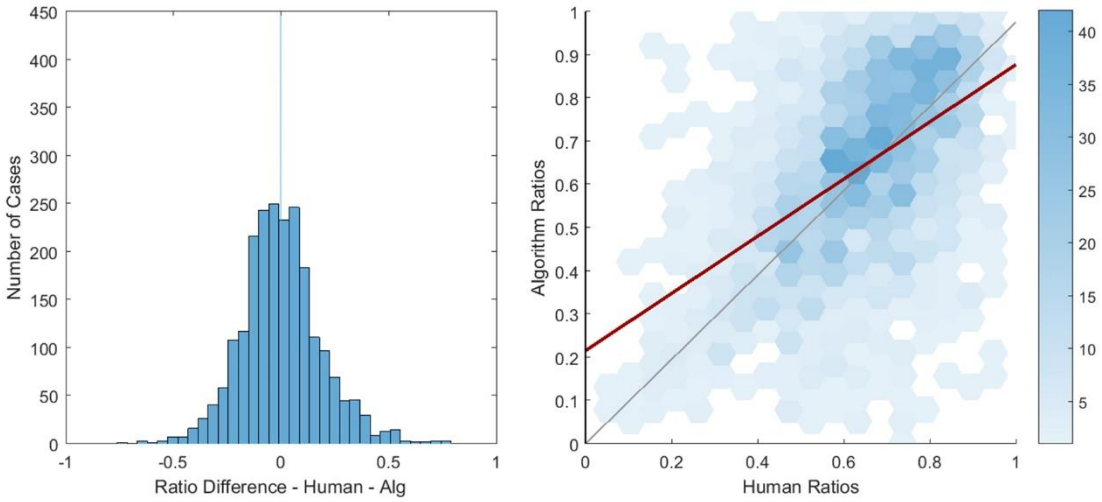
ROC Curves for all 8 tissue subtypes, classified by Algorithm A

The graph shows Area Under the Curve for each tissue subtype:
Tumour parent class: Tumour (0.86), Lumen (0.92), Mucin (0.84), Necrosis (0.81)
Stroma parent class: Stroma (0.84), Vessels (0.76), Muscle (0.84), Inflammation (0.74)



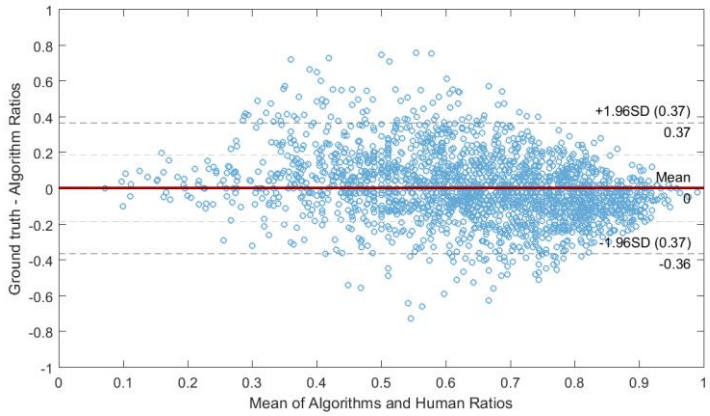
Comparison boxplots for pathologist-pathologist agreement and pathologist-algorithm agreement

Left: Pathologist - pathologist agreement (mean = 0.85, median = 0.88., SD = 0.10, IQR = 0.10)
Right: Pathologist -algorithm agreement (mean = 0.77, median = 0.77., SD = 0.01, IQR = 0.01)
Two sample T-Test P = 0.03
Human agreement is generated from six participants scoring 40 images 1024x1024 pixels in size, using the Prospector system (section 4.3).



Comparison plots of pathologist scores to Algorithm A

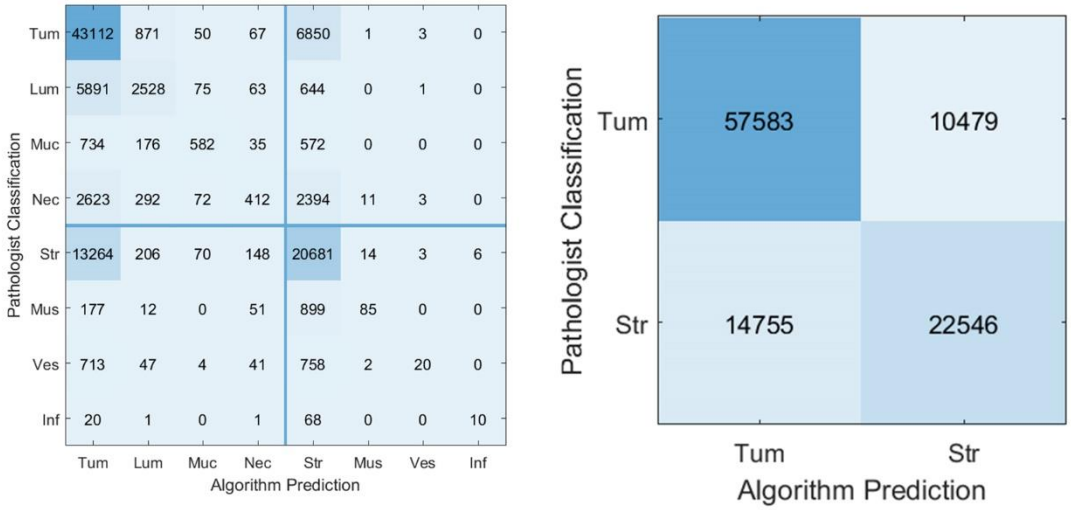
Left: Histogram of ratio differences generated by human and regular segment algorithm. Distribution has a mean bias of 0 (median 0), and standard deviation of 0.19.
Right: Heatmap Correlation plot of the distribution of the ratios. Correlation has R^2 coefficient of 0.27.



Bland-Altman plot of pathologist and Algorithm A-generated TSRs per case

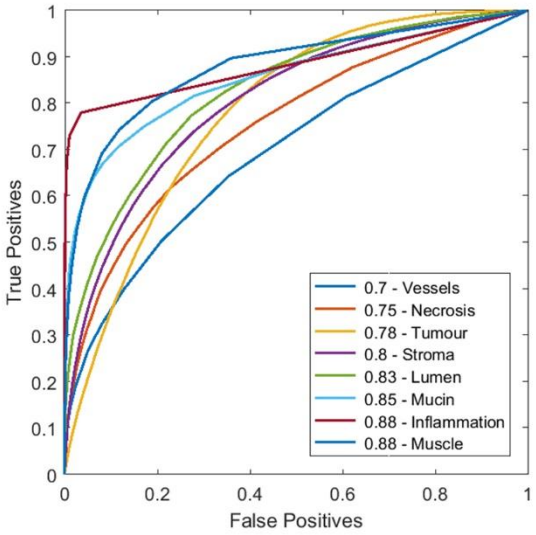
Distribution has a mean bias of 0, with upper and lower limits of agreement of 0.37 and -0.36 respectively (±0.37).

D.2 - Algorithm B: Fixed patch size (256x256 pixels) results



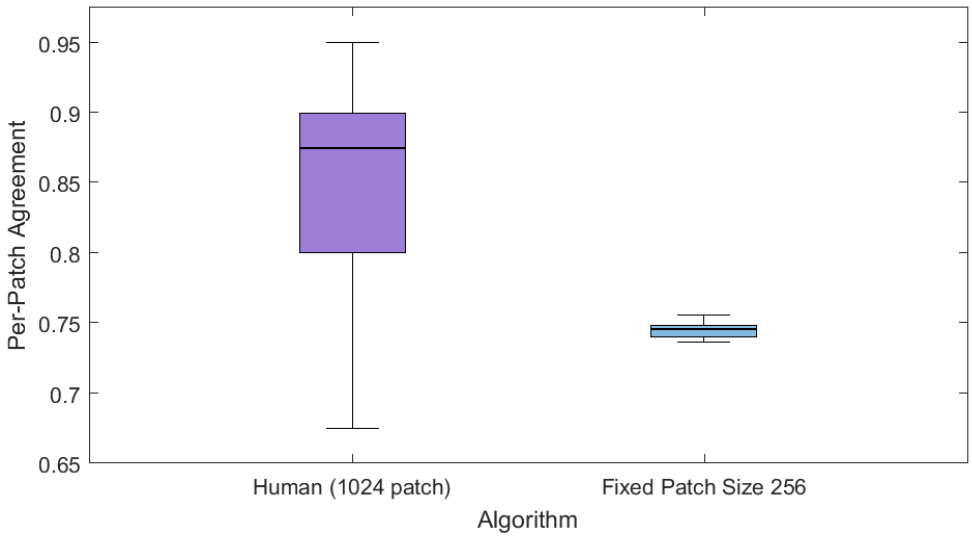
Confusion matrices showing pathologist – Algorithm B agreement

Left: Confusion matrix of all 8 tissue subtypes from dataset. Accuracy = 63.99%, sensitivity (true positive rate / recall) = 0.64, kappa = 0.39
Right: Confusion matrix of subtypes grouped into Tumour and Stroma parent classes. Accuracy = 76.05%, sensitivity (true positive rate / recall) = 0.84, specificity (true negative rate) = 0.60, kappa = 0.46



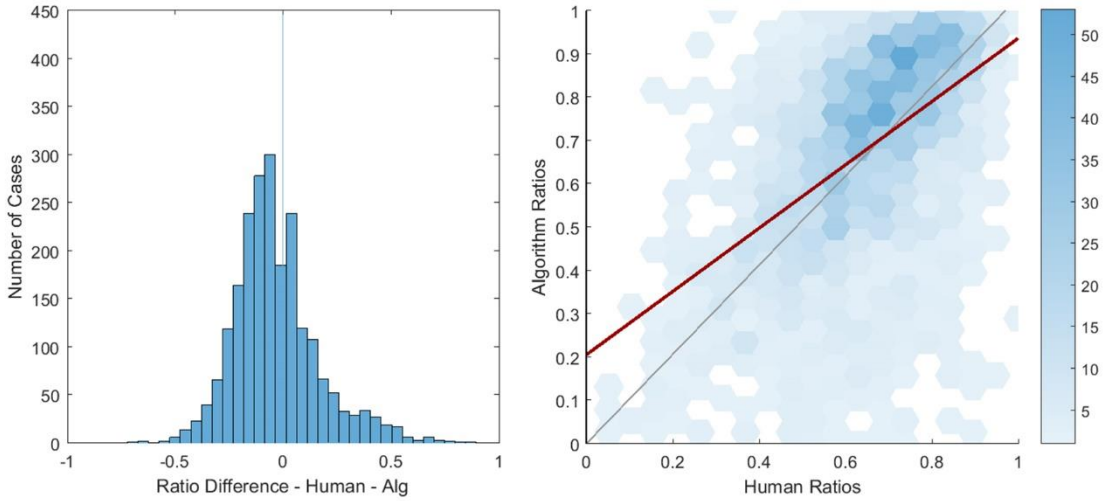
ROC Curves for all 8 tissue subtypes, classified by Algorithm B

The graph shows Area Under the Curve for each tissue subtype:
Tumour parent class: Tumour (0.78), Lumen (0.83), Mucin (0.85), Necrosis (0.75)
Stroma parent class: Stroma (0.80), Vessels (0.70), Muscle (0.88), Inflammation (0.88)



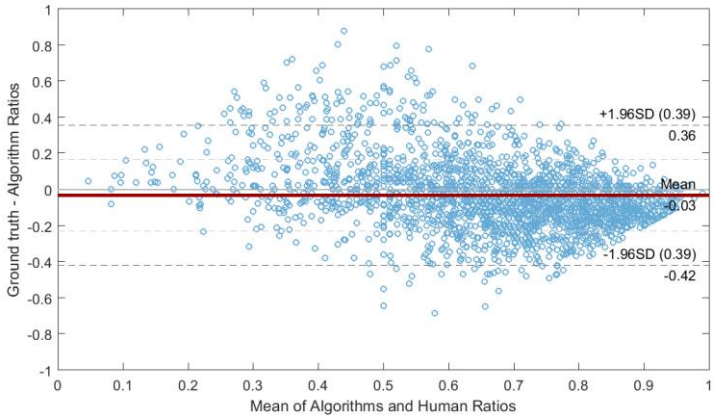
Comparison boxplots for pathologist - pathologist agreement and pathologist - algorithm agreement

Left: Pathologist - pathologist agreement (mean = 0.85, median = 0.88., SD = 0.10, IQR = 0.10)
Right: Pathologist -algorithm agreement (mean = 0.75, median = 0.75., SD = 0.01, IQR = 0.01)
Two sample T-Test P = 0.01
Human agreement is generated from six participants scoring 40 images 1024x1024 pixels in size, using the Prospector system (section 4.3).



Comparison plots of pathologist scores to Algorithm B

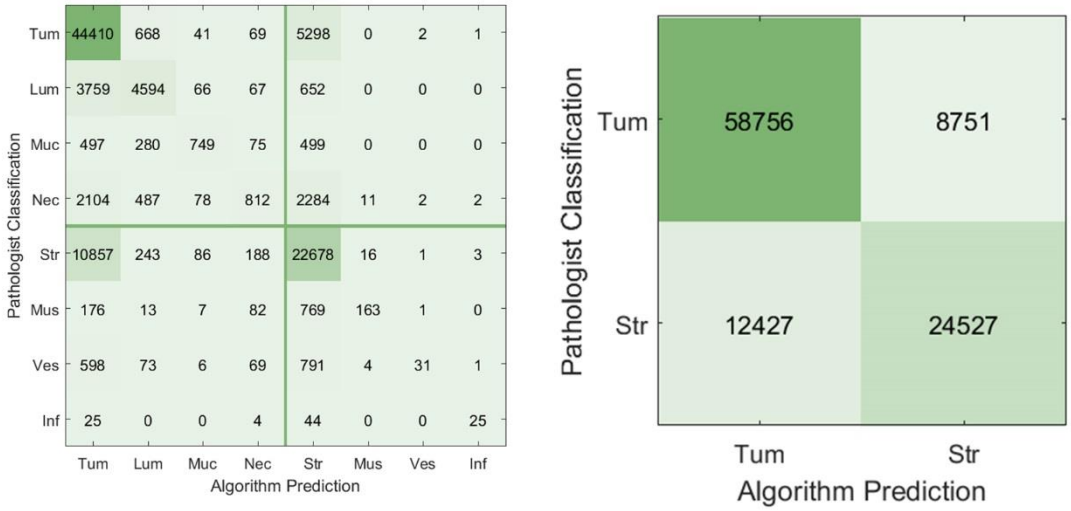
Left: Histogram of ratio differences generated by human and regular segment algorithm. Distribution has a mean bias of -0.03 (median -0.06), and standard deviation of 0.20.
Right: Heatmap Correlation plot of the distribution of the ratios. Correlation has R^2 coefficient of 0.28.



Bland-Altman plot of pathologist and Algorithm B-generated TSRs per case

Distribution has a mean bias of -0.03, with upper and lower limits of agreement of 0.36 and -0.42 respectively (±0.39).

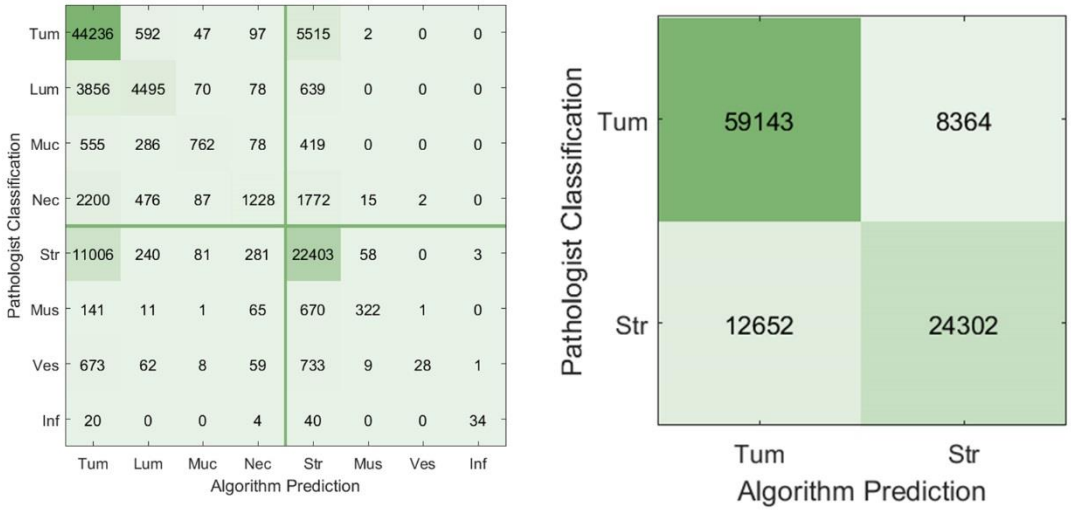
D.3 - Algorithm C: Fixed partition (with predictions) results



Confusion matrices showing pathologist – Algorithm C agreement

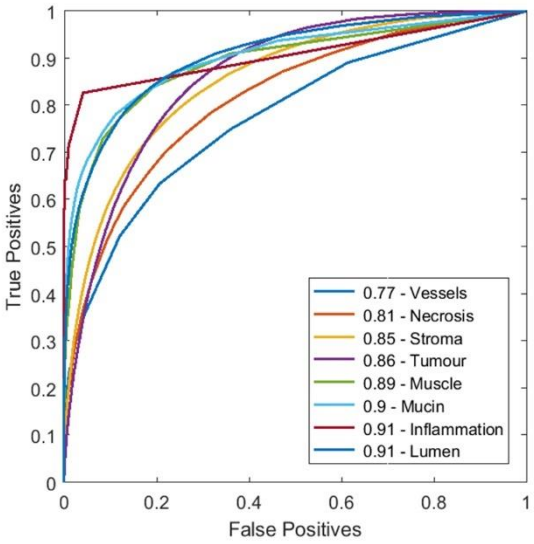
Left: Confusion matrix of all 8 tissue subtypes from dataset. Accuracy = 70.32%, sensitivity (true positive rate / recall) = 0.70, kappa = 0.51 (moderate agreement)
Right: Confusion matrix of subtypes grouped into Tumour and Stroma parent classes. Accuracy = 79.72%, sensitivity (true positive rate / recall) = 0.87, specificity (true negative rate) = 0.66, kappa = 0.55 (moderate agreement)

D.4 - Algorithm D: Fixed partition (with votes) results



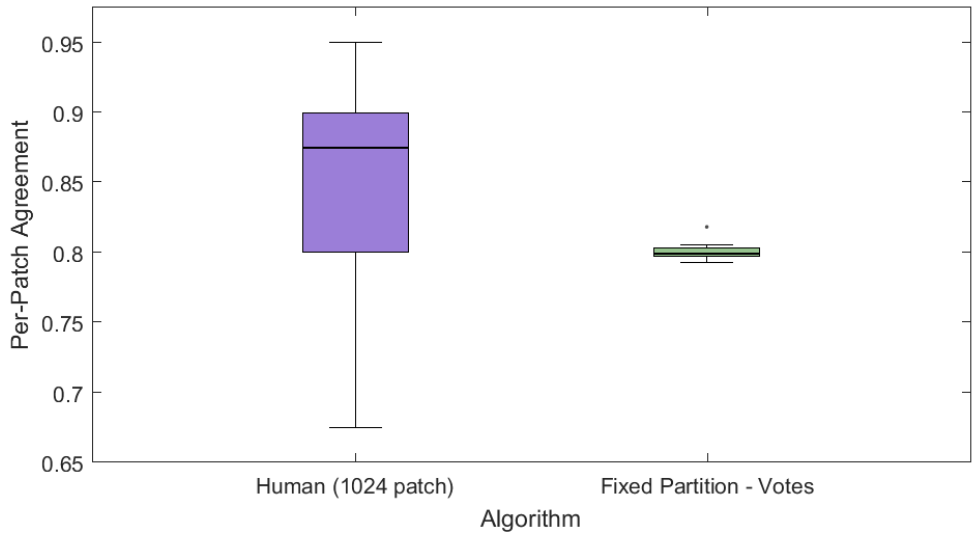
Confusion matrices showing pathologist – Algorithm D agreement

Left: Confusion matrix of all 8 tissue subtypes from dataset. Accuracy = 70.37%, sensitivity (true positive rate / recall) = 0.70, kappa = 0.51 (moderate agreement)
Right: Confusion matrix of subtypes grouped into Tumour and Stroma parent classes. Accuracy = 79.88%, sensitivity (true positive rate / recall) = 0.88, specificity (true negative rate) = 0.66, kappa = 0.55 (moderate agreement)



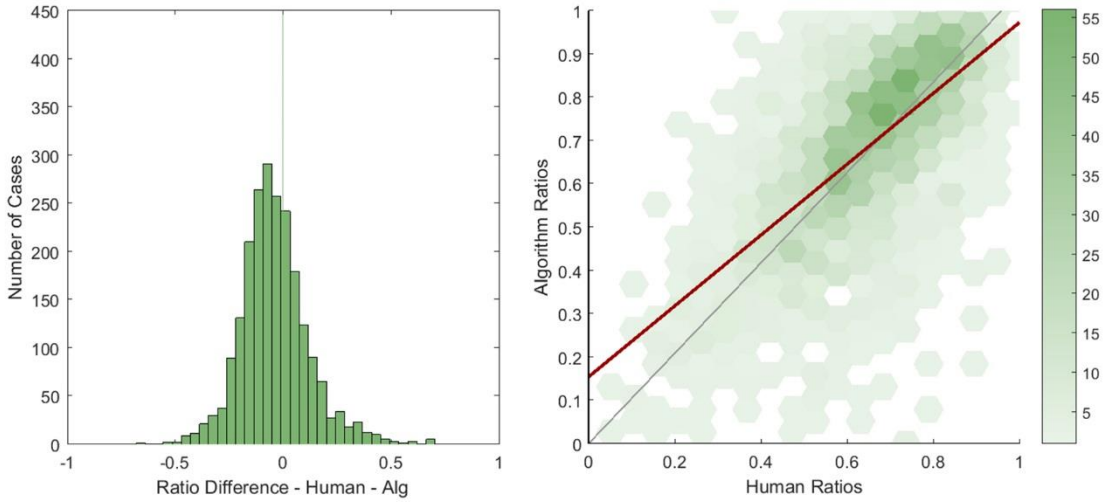
ROC Curves for all 8 tissue subtypes, classified by Algorithm D

The graph shows Area Under the Curve for each tissue subtype:
Tumour parent class: Tumour (0.86), Lumen (0.91), Mucin (0.90), Necrosis (0.81)
Stroma parent class: Stroma (0.85), Vessels (0.77), Muscle (0.89), Inflammation (0.91)



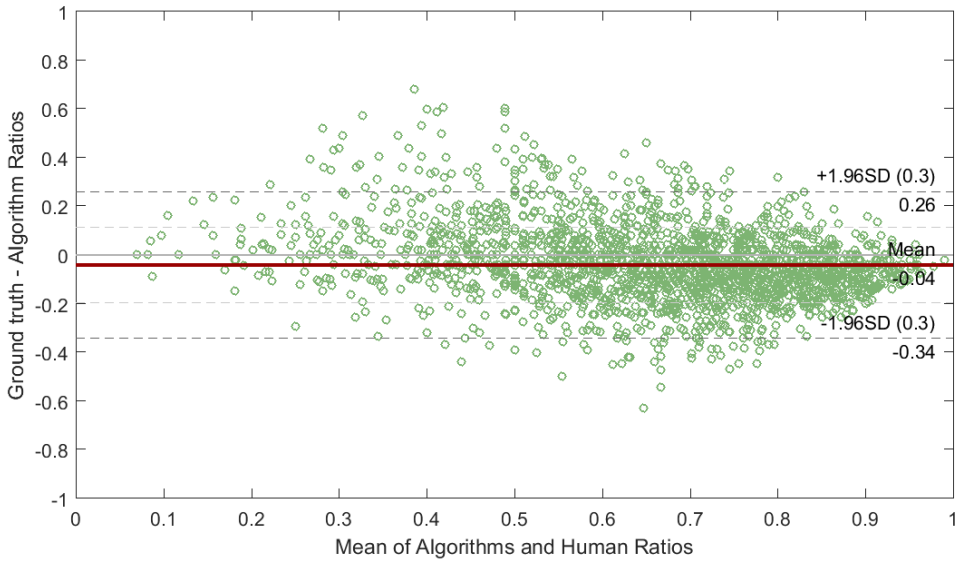
Comparison boxplots for pathologist - pathologist agreement and pathologist - algorithm agreement

Left: Pathologist - pathologist agreement (mean = 0.85, median = 0.88., SD = 0.10, IQR = 0.10)
Right: Pathologist - algorithm agreement (mean = 0.80, median = 0.80., SD = 0.01, IQR = 0.01)
Two sample T-Test P = 0.16
Human agreement is generated from six participants scoring 40 images 1024x1024 pixels in size, using the Prospector system (section 4.3).



Histogram and Heatmap Correlation Plots for Algorithm D

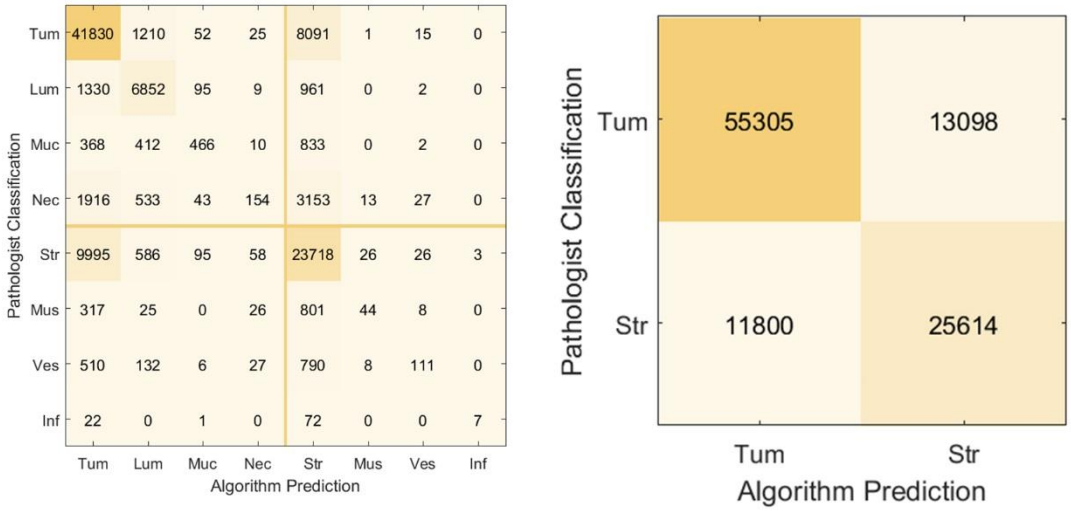
Left: Histogram of ratio differences generated by human and regular segment algorithm. Distribution has a mean bias of -0.04 (median 0.06), and standard deviation of 0.15.
Right: Heatmap Correlation plot of the distribution of the ratios. Correlation has R^2 coefficient of 0.45.



Bland-Altman plot of Pathologist and Regular Segmentation algorithm-generated TSRs per case

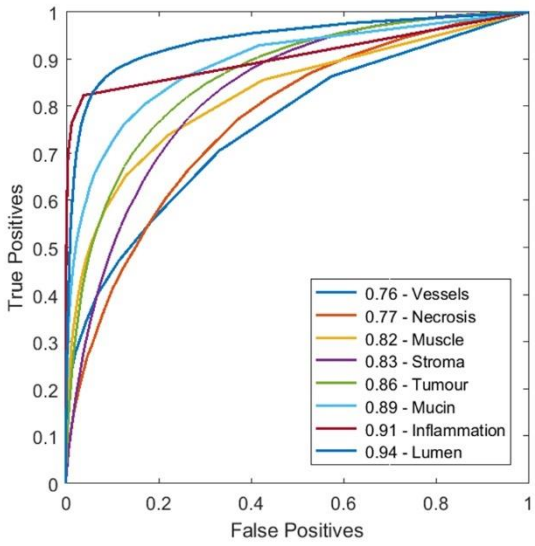
Distribution has a mean bias of – 0.04, with upper and lower limits of agreement of 0.26 and -0.34 respectively (±0.30).

D.5 - Algorithm E: Unsupervised segmentation results



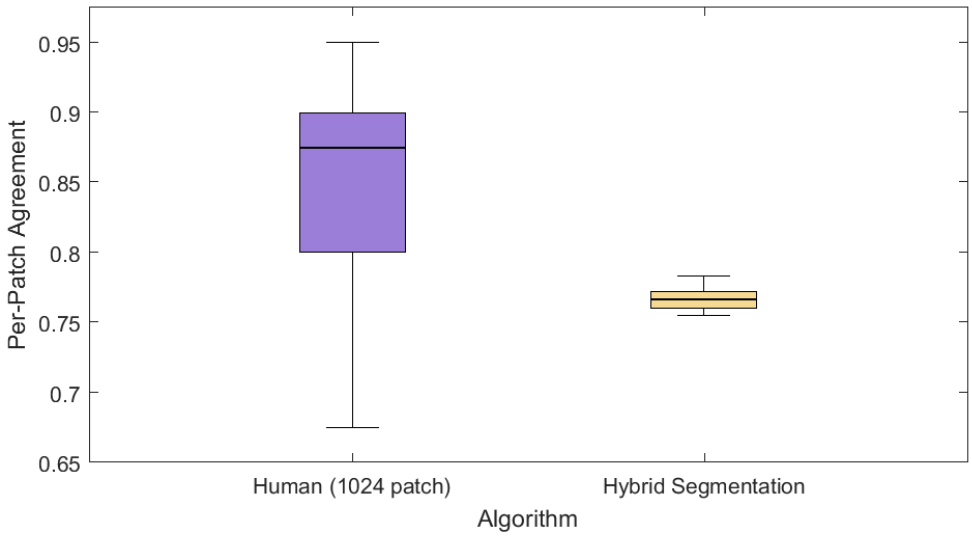
Confusion matrices showing pathologist – Algorithm E agreement

Left: Confusion matrix of all 8 tissue subtypes from dataset. Accuracy = 69.16%, sensitivity (true positive rate / recall) = 0.69, kappa = 0.50 (moderate agreement)
Right: Confusion matrix of subtypes grouped into Tumour and Stroma parent classes. Accuracy = 76.47%, sensitivity (true positive rate / recall) = 0.81, specificity (true negative rate) = 0.68, kappa = 0.49 (moderate agreement)



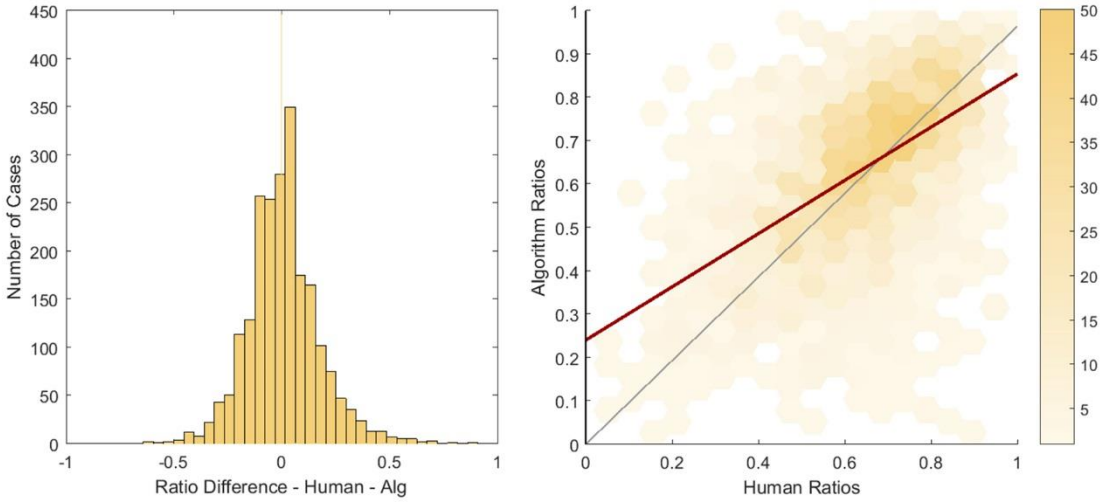
ROC Curves for all 8 tissue subtypes, classified by Algorithm E

The graph shows Area Under the Curve for each tissue subtype:
Tumour parent class: Tumour (0.86), Lumen (0.94), Mucin (0.89), Necrosis (0.77)
Stroma parent class: Stroma (0.83), Vessels (0.76), Muscle (0.82), Inflammation (0.91)



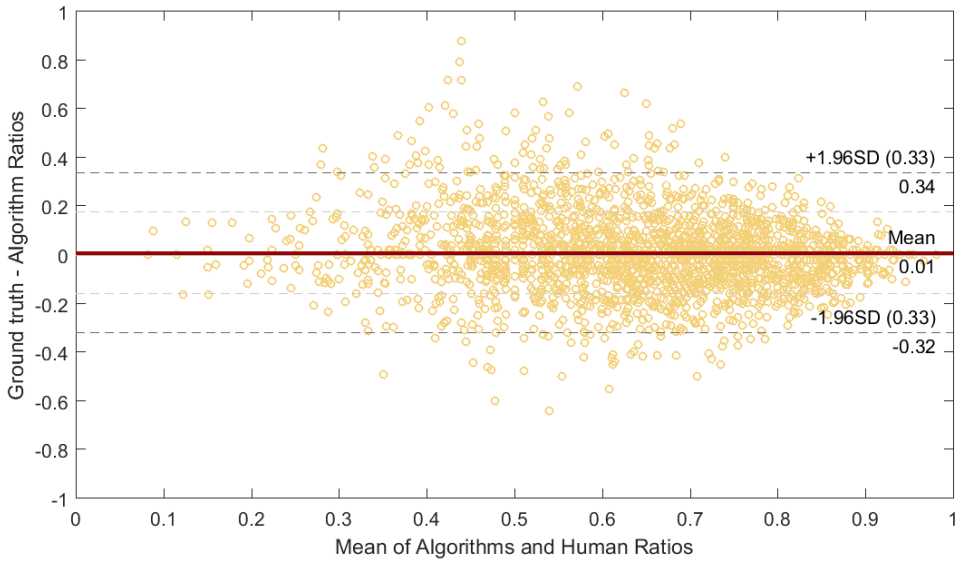
Comparison boxplots for pathologist - pathologist agreement and pathologist - algorithm agreement

Left: Pathologist - pathologist agreement (mean = 0.85, median = 0.88., SD = 0.10, IQR = 0.10)
Right: Pathologist - algorithm agreement (mean = 0.77, median = 0.77., SD = 0.01, IQR = 0.01)
Two sample T-Test P = 0.02
Human agreement is generated from six participants scoring 40 images 1024x1024 pixels in size, using the Prospector system (section 4.3).



Histogram and Heatmap Correlation Plots for Algorithm E

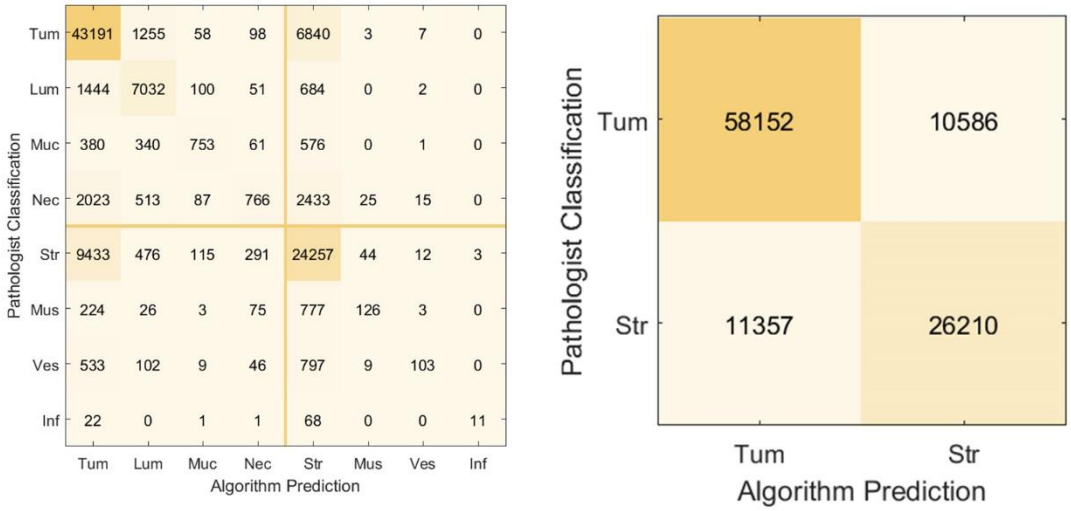
Left: Histogram of ratio differences generated by human and regular segment algorithm. Distribution has a mean bias of 0.01 (median 0.00), and standard deviation of 0.17.
Right: Heatmap Correlation plot of the distribution of the ratios. Correlation has R^2 coefficient of 0.29.



Bland-Altman plot of Pathologist and Algorithm E-generated TSRs per case

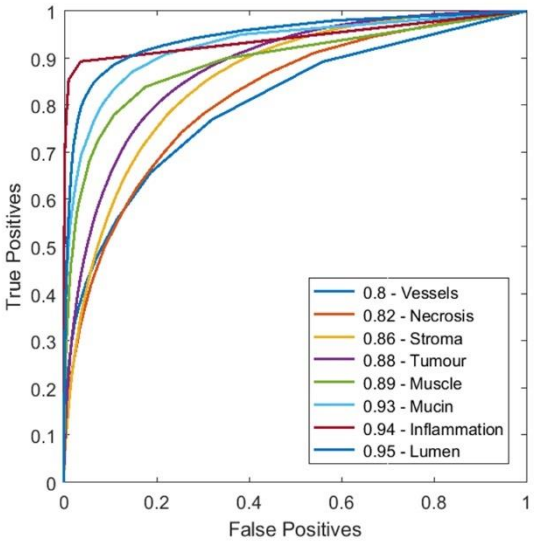
Distribution has a mean bias of 0.01, with upper and lower limits of agreement of 0.34 and -0.32 respectively (±0.33).

D.6 - Algorithm F: Local context results



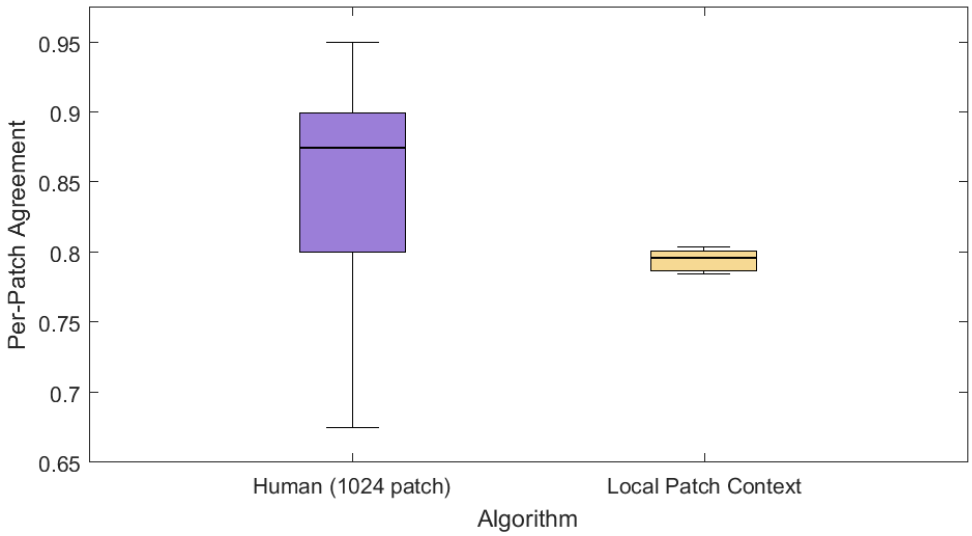
Confusion matrices showing pathologist - Algorithm F agreement

Left: Confusion matrix of all 8 tissue subtypes from dataset. Accuracy = 71.72%, sensitivity (true positive rate / recall) = 0.72, kappa = 0.54 (moderate agreement)
Right: Confusion matrix of subtypes grouped into Tumour and Stroma parent classes. Accuracy = 79.36%, sensitivity (true positive rate / recall) = 0.85, specificity (true negative rate) = 0.70, kappa = 0.55 (moderate agreement)



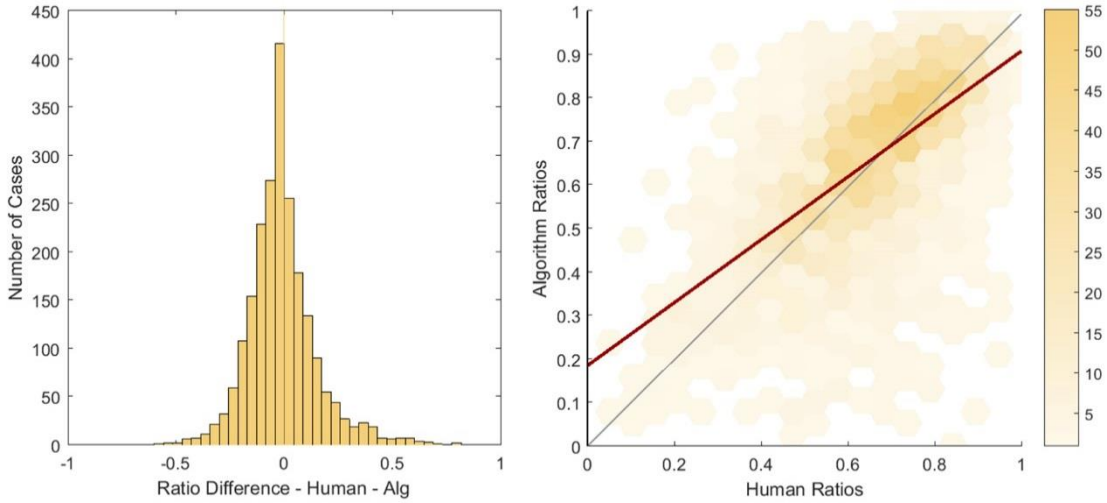
ROC Curves for all 8 tissue subtypes, classified by Algorithm F

The graph shows Area Under the Curve for each tissue subtype:
Tumour parent class: Tumour (0.88), Lumen (0.95), Mucin (0.93), Necrosis (0.82)
Stroma parent class: Stroma (0.86), Vessels (0.80), Muscle (0.89), Inflammation (0.94)



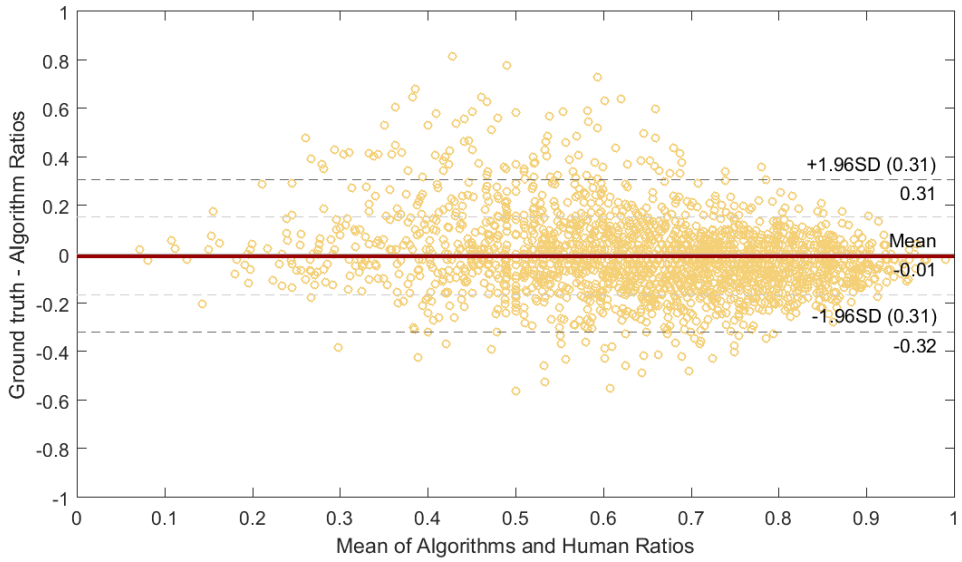
Comparison boxplots for pathologist - pathologist agreement and pathologist - algorithm agreement

Left: Pathologist - pathologist agreement (mean = 0.85, median = 0.88., SD = 0.10, IQR = 0.10)
Right: Pathologist - algorithm agreement (mean = 0.79, median = 0.80., SD = 0.01, IQR = 0.01)
Two sample T-Test P = 0.11
Human agreement is generated from six participants scoring 40 images 1024x1024 pixels in size, using the Prospector system (section 4.3).



Histogram and Heatmap Correlation Plots for Algorithm F

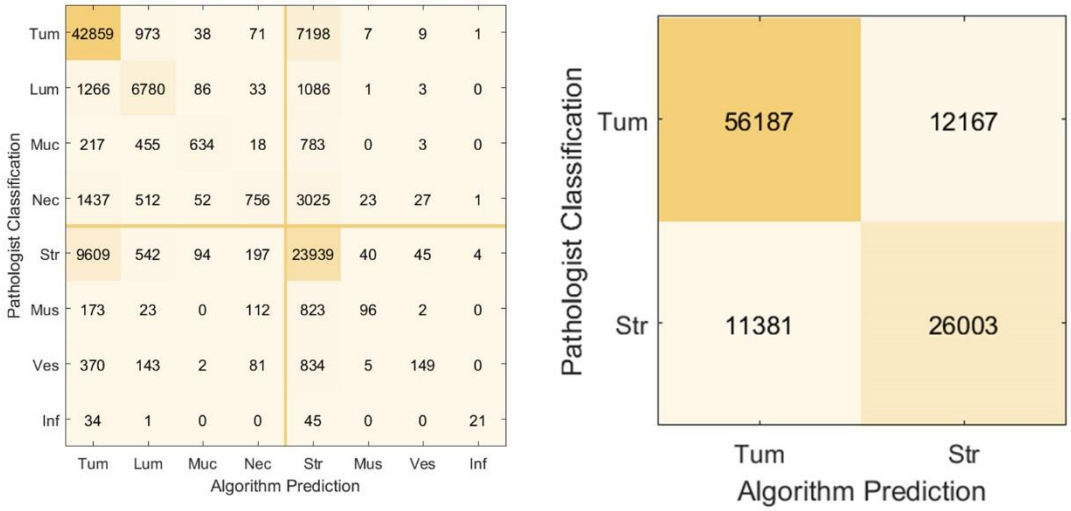
Left: Histogram of ratio differences generated by human and regular segment algorithm. Distribution has a mean bias of -0.01 (median -0.02), and standard deviation of 0.16.
Right: Heatmap Correlation plot of the distribution of the ratios. Correlation has R^2 coefficient of 0.37.



Bland-Altman plot of Pathologist and Algorithm F-generated TSRs per case

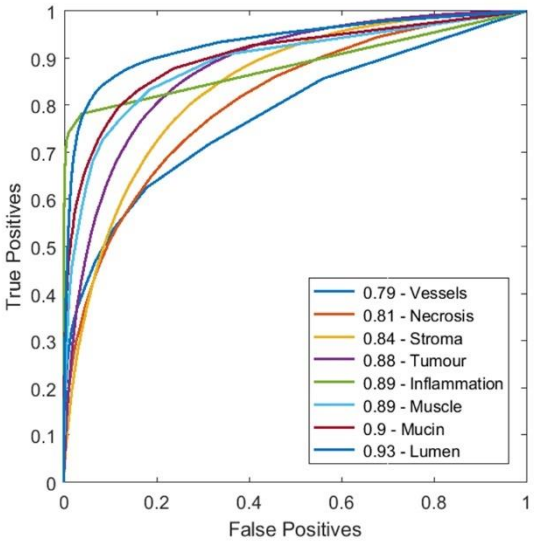
Distribution has a mean bias of – 0.01, with upper and lower limits of agreement of 0.31 and -0.32 respectively (±0.31).

D.7 - Algorithm G: Global context results



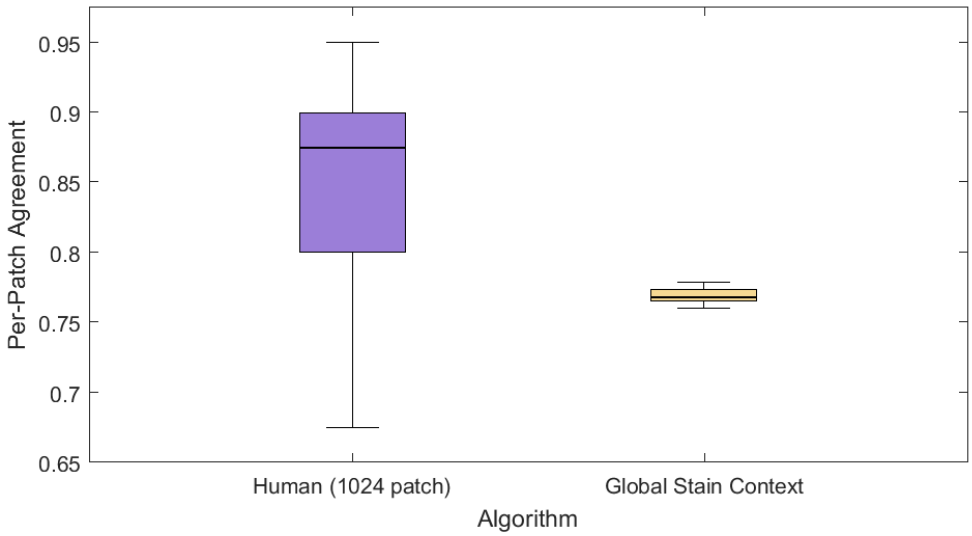
Confusion matrices showing pathologist – Algorithm G agreement

Left: Confusion matrix of all 8 tissue subtypes from dataset. Accuracy = 71.15%, sensitivity (true positive rate / recall) = 0.71, kappa = 0.53 (moderate agreement)
Right: Confusion matrix of subtypes grouped into Tumour and Stroma parent classes. Accuracy = 77.73%, sensitivity (true positive rate / recall) = 0.82, specificity (true negative rate) = 0.70, kappa = 0.52 (moderate agreement)



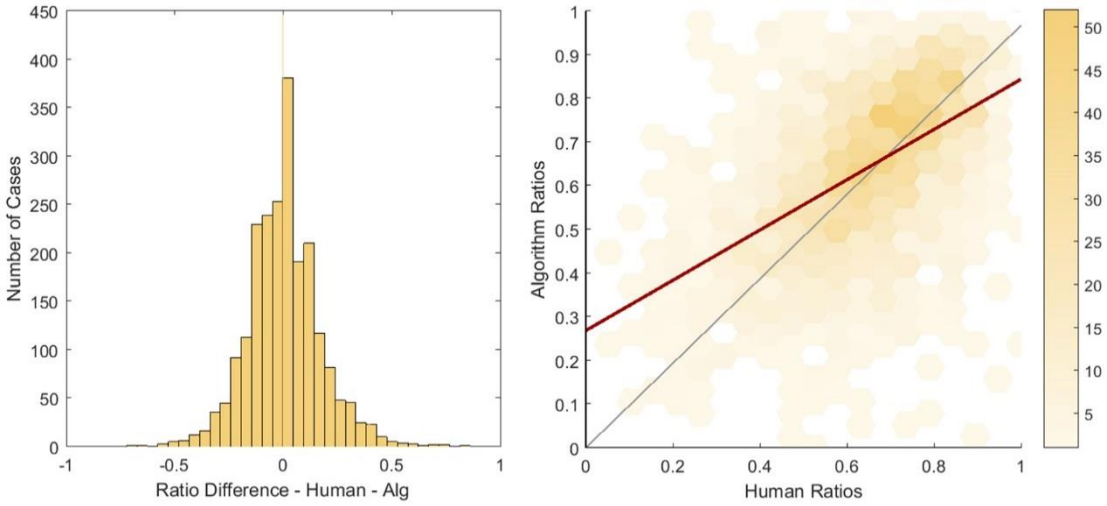
ROC Curves for all 8 tissue subtypes, classified by Algorithm G

The graph shows Area Under the Curve for each tissue subtype:
Tumour parent class: Tumour (0.88), Lumen (0.93), Mucin (0.90), Necrosis (0.81)
Stroma parent class: Stroma (0.84), Vessels (0.79), Muscle (0.89), Inflammation (0.89)



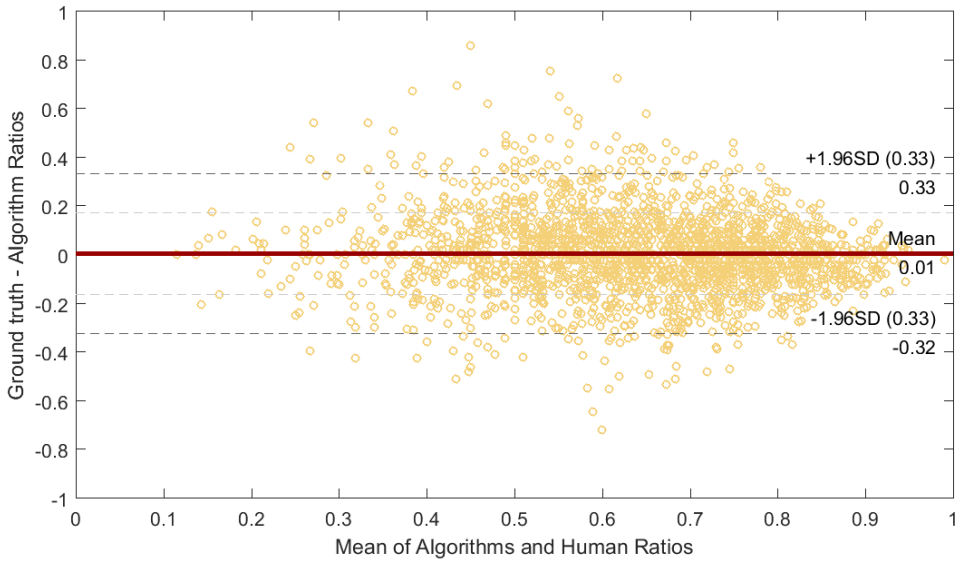
Comparison boxplots for pathologist - pathologist agreement and pathologist - algorithm agreement

Left: Pathologist - pathologist agreement (mean = 0.85, median = 0.88., SD = 0.10, IQR = 0.10)
Right: Pathologist - algorithm agreement (mean = 0.77, median = 0.77., SD = 0.01, IQR = 0.01)
Two sample T-Test P = 0.02
Human agreement is generated from six participants scoring 40 images 1024x1024 pixels in size, using the Prospector system (section 4.3).



Histogram and Heatmap Correlation Plots for Algorithm G

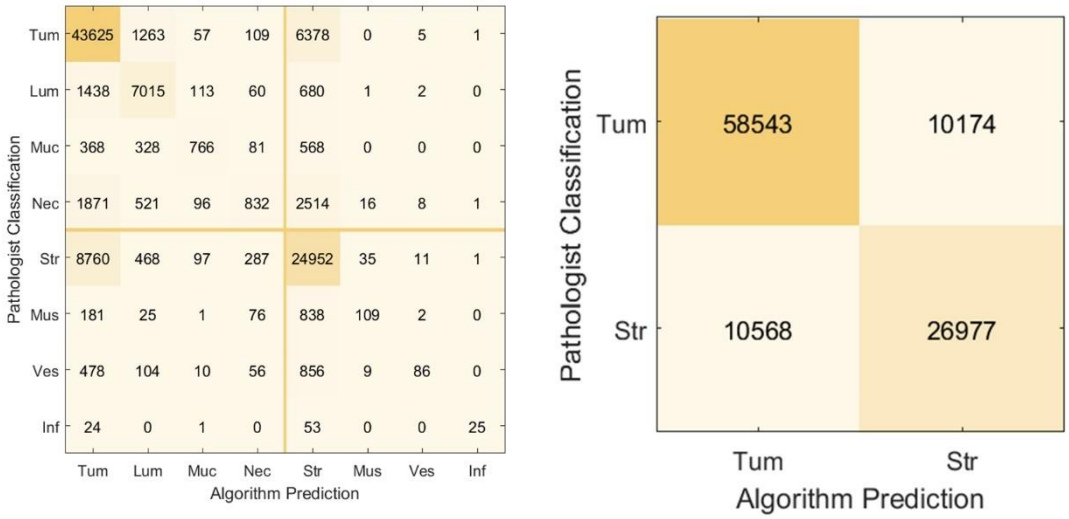
Left: Histogram of ratio differences generated by human and regular segment algorithm. Distribution has a mean bias of 0.01 (median 0), and standard deviation of 0.17.
Right: Heatmap Correlation plot of the distribution of the ratios. Correlation has R^2 coefficient of 0.27.



Bland-Altman plot of Pathologist and Algorithm G-generated TSRs per case

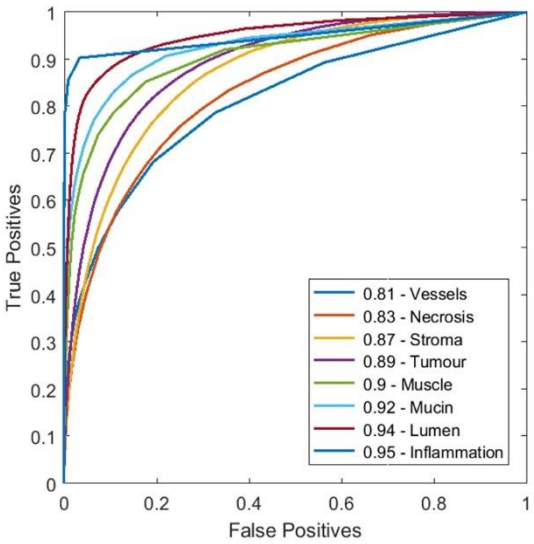
Distribution has a mean bias of 0.01, with upper and lower limits of agreement of 0.33 and -0.32 respectively (±0.33).

D.8 - Algorithm H: Combined context results



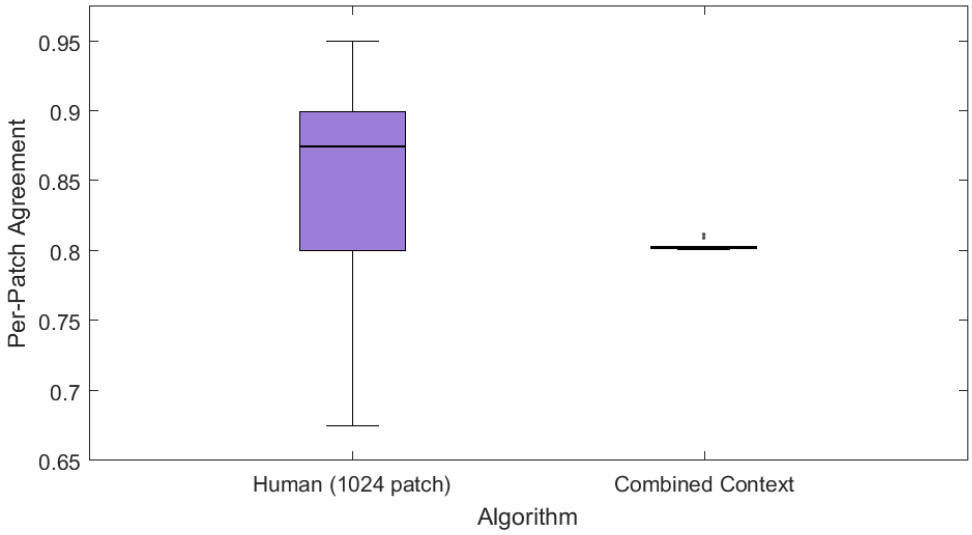
Confusion matrices showing pathologist – Algorithm H agreement

Left: Confusion matrix of all 8 tissue subtypes from dataset. Accuracy = 72.85%, sensitivity (true positive rate / recall) = 0.73, kappa = 0.56 (moderate agreement)
Right: Confusion matrix of subtypes grouped into Tumour and Stroma parent classes. Accuracy = 80.50%, sensitivity (true positive rate / recall) = 0.85, specificity (true negative rate) = 0.72, kappa = 0.57 (moderate agreement)



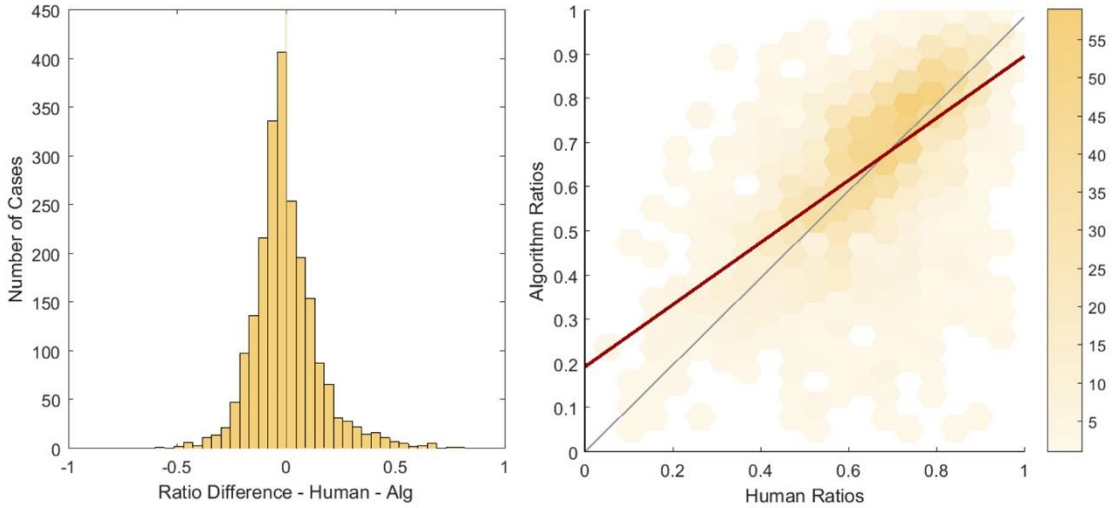
ROC Curves for all 8 tissue subtypes, classified by Algorithm H

The graph shows Area Under the Curve for each tissue subtype:
Tumour parent class: Tumour (0.89), Lumen (0.94), Mucin (0.92), Necrosis (0.83)
Stroma parent class: Stroma (0.87), Vessels (0.81), Muscle (0.90), Inflammation (0.95)



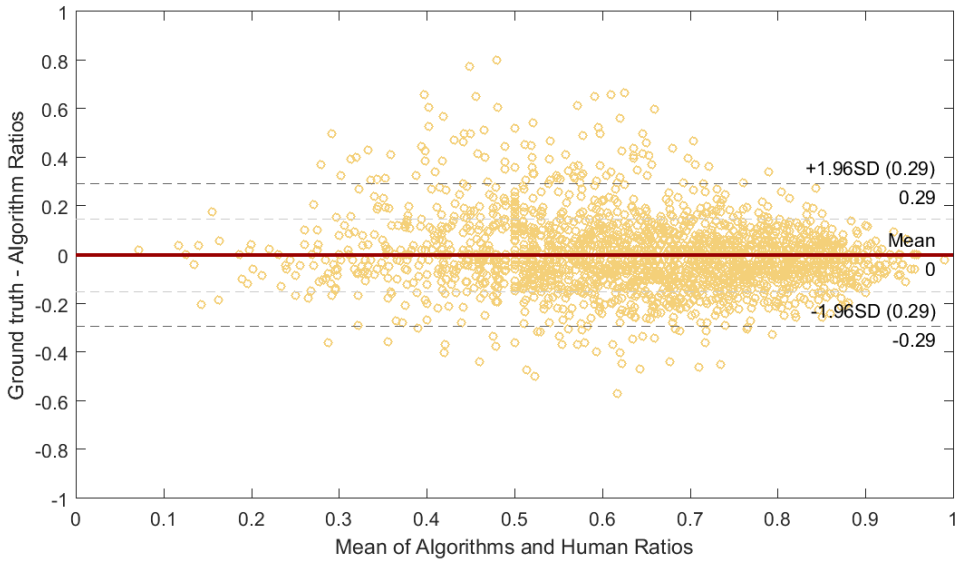
Comparison boxplots for pathologist - pathologist agreement and pathologist - algorithm agreement

Left: Pathologist - pathologist agreement (mean = 0.85, median = 0.88., SD = 0.10, IQR = 0.10)
Right: Pathologist - algorithm agreement (mean = 0.80, median = 0.80., SD <0.01, IQR <0.01)
Two sample T-Test P = 0.19
Human agreement is generated from six participants scoring 40 images 1024x1024 pixels in size, using the Prospector system (section 4.3).



Histogram and Heatmap Correlation Plots for Algorithm H

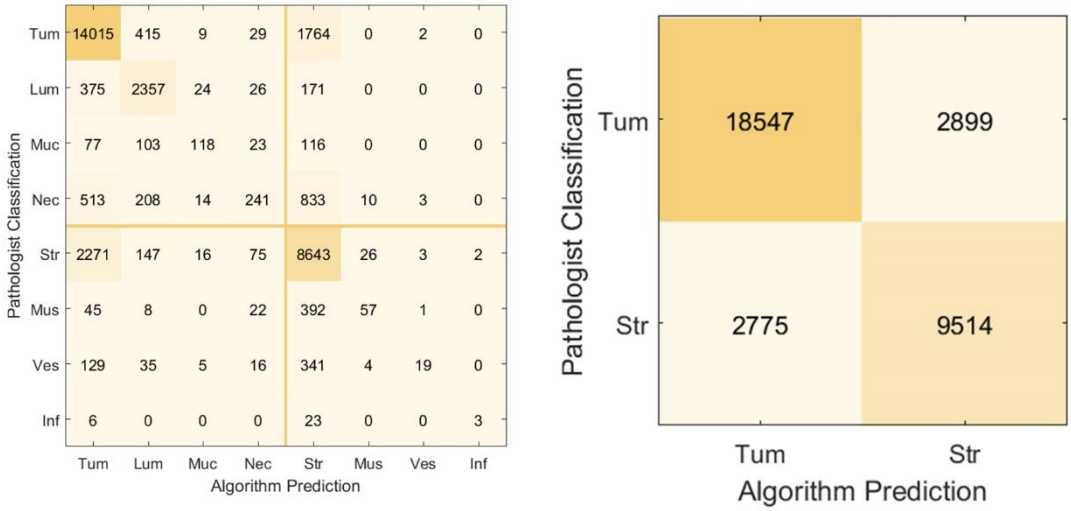
Left: Histogram of ratio differences generated by human and regular segment algorithm. Distribution has a mean bias of 0.00 (median -0.02), and standard deviation of 0.15.
Right: Heatmap Correlation plot of the distribution of the ratios. Correlation has R^2 coefficient of 0.40.



Bland-Altman plot of Pathologist and Algorithm H-generated TSRs per case

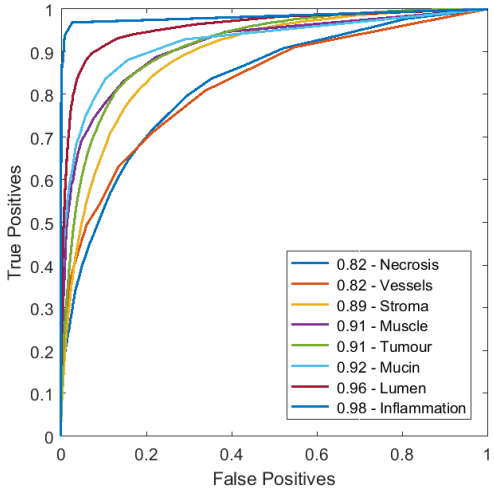
Distribution has a mean bias of 0, with upper and lower limits of agreement of 0.29 and -0.29 respectively (±0.29).

D.9 – Algorithm I: Combined context results – with QC



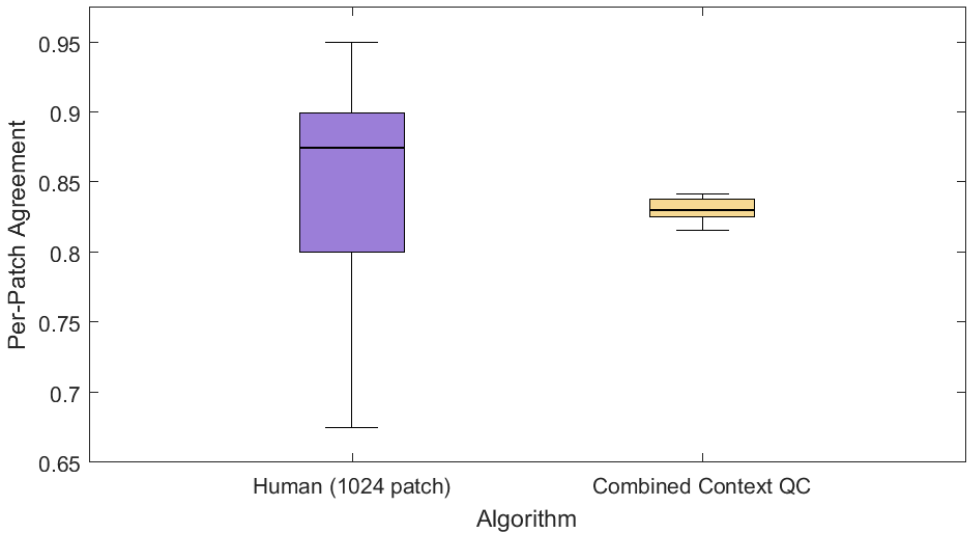
Confusion matrices showing pathologist – Algorithm I agreement

Left: Confusion matrix of all 8 tissue subtypes from dataset. Accuracy = 75.45%, sensitivity (true positive rate / recall) = 0.75, kappa = 0.60 (substantial agreement).
Right: Confusion matrix of subtypes grouped into Tumour and Stroma parent classes. Accuracy = 83.18%, sensitivity (true positive rate / recall) = 0.86, specificity (true negative rate) = 0.77, kappa = 0.64 (substantial agreement).



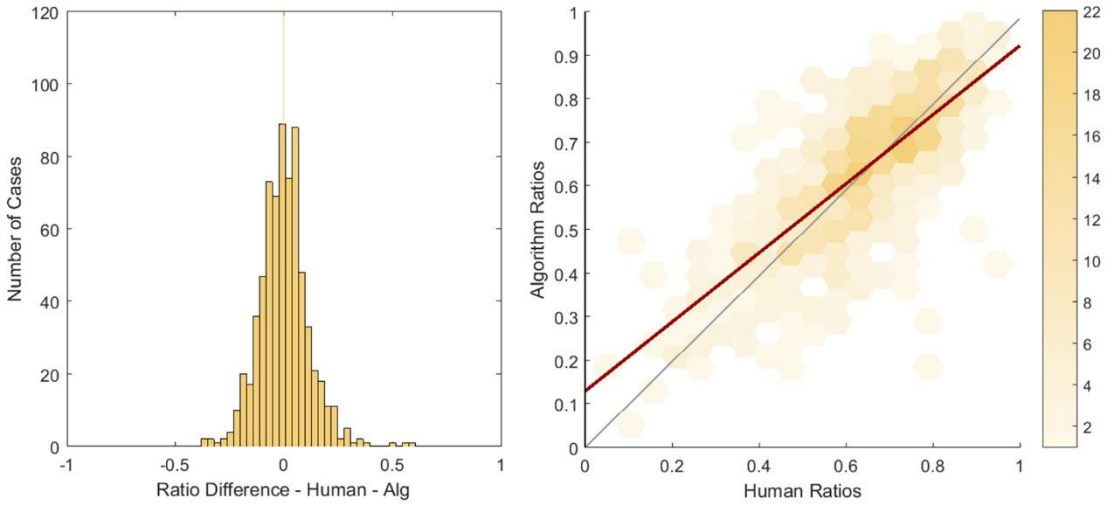
ROC Curves for all 8 tissue subtypes, classified by Algorithm I

The graph shows Area Under the Curve for each tissue subtype:
Tumour parent class: Tumour (0.91), Lumen (0.96), Mucin (0.92), Necrosis (0.82)
Stroma parent class: Stroma (0.89), Vessels (0.82), Muscle (0.91), Inflammation (0.98)



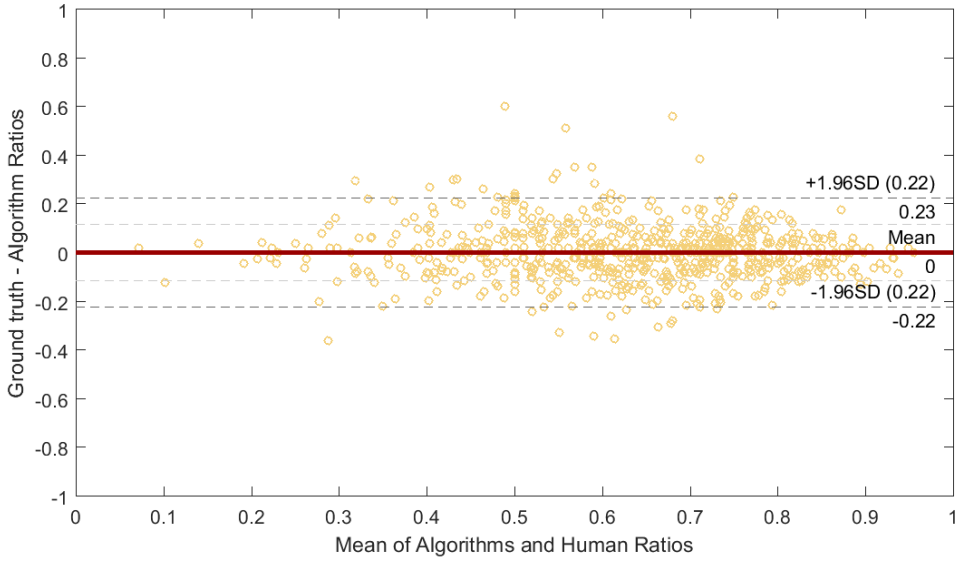
Comparison boxplots for pathologist - pathologist agreement and pathologist - algorithm agreement

Left: Pathologist - pathologist agreement (mean = 0.85, median = 0.88., SD = 0.10, IQR = 0.10)
Right: Pathologist - algorithm agreement (mean = 0.83, median = 0.83., SD = 0.01, IQR = 0.01)
Two sample T-Test P = 0.62
Human agreement is generated from six participants scoring 40 images 1024x1024 pixels in size, using the Prospector system (section 4.3).



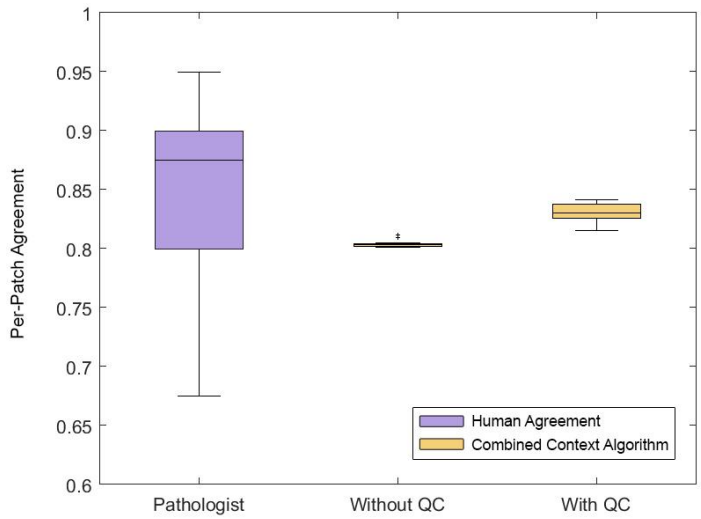
Histogram and Heatmap Correlation Plots for Algorithm I

Left: Histogram of ratio differences generated by human and regular segment algorithm. Distribution has a mean bias of 0 (median 0), and standard deviation of 0.11.
Right: Heatmap Correlation plot of the distribution of the ratios. Correlation has R^2 coefficient of 0.58.



Bland-Altman plot of Pathologist and Algorithm I-generated TSRs per case

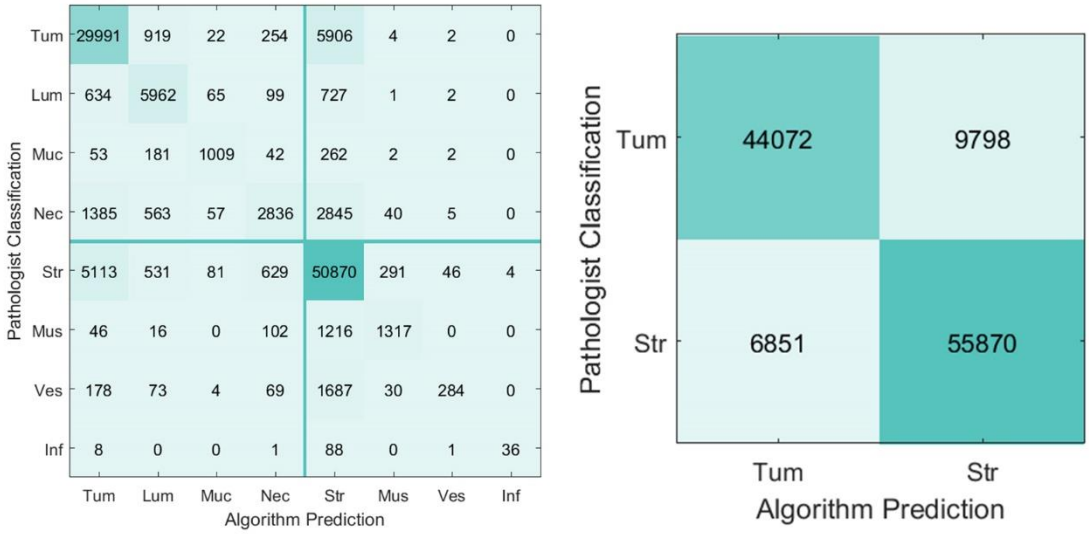
Distribution has a mean bias of 0, with upper and lower limits of agreement of 0.23 and -0.22 respectively (±0.22).



Boxplot showing between-pathologist agreement and the Combined Context algorithm, with and without QC

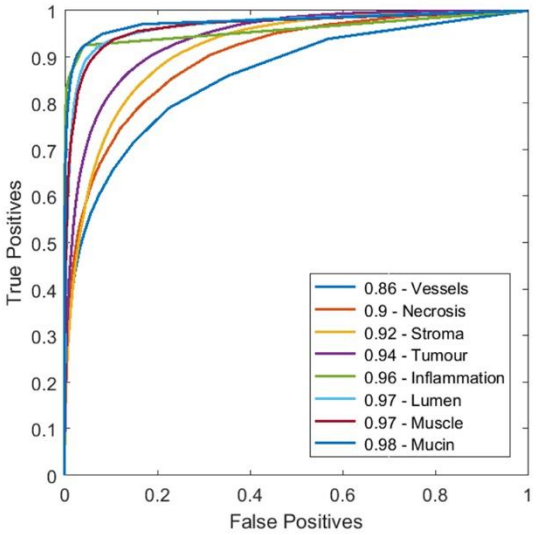
Mean agreement for pathologist = 0.88, Combined Context without QC = 0.81, Combined Context with QC = 0.83
Note that pathologist agreement was calculated from the image size experiment, in section 4.3. The results from the Combined Context algorithm were when trained and tested on either the full dataset or the QCed subset of slides

D.10 - Algorithm J: Combined context results CR07 dataset



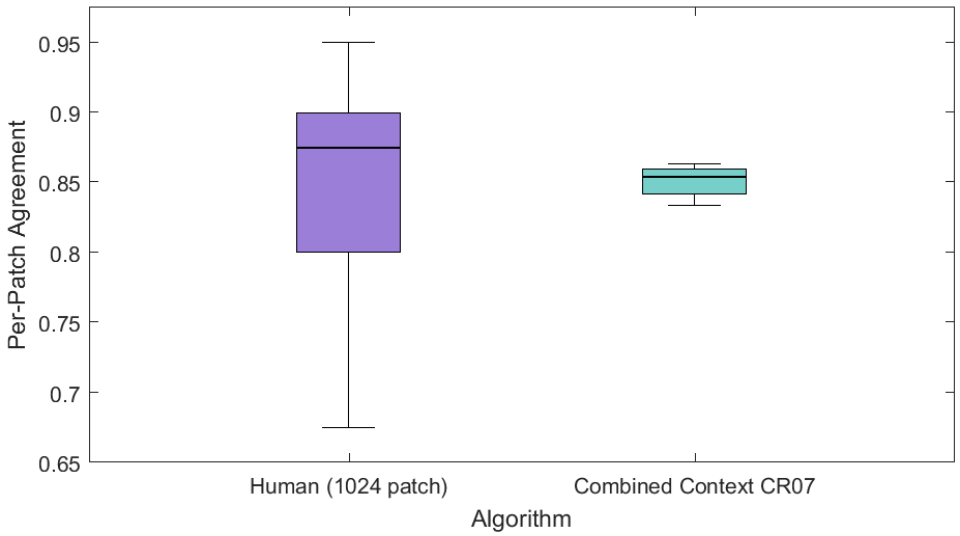
Confusion matrices showing pathologist – Algorithm J agreement

Left: Confusion matrix of all 8 tissue subtypes from dataset. Accuracy = 79.17%, sensitivity (true positive rate / recall) = 0.79, kappa = 0.66 (substantial agreement)
Right: Confusion matrix of subtypes grouped into Tumour and Stroma parent classes. Accuracy = 85.72%, sensitivity (true positive rate / recall) = 0.82, specificity (true negative rate) = 0.89, kappa = 0.71 (substantial agreement)



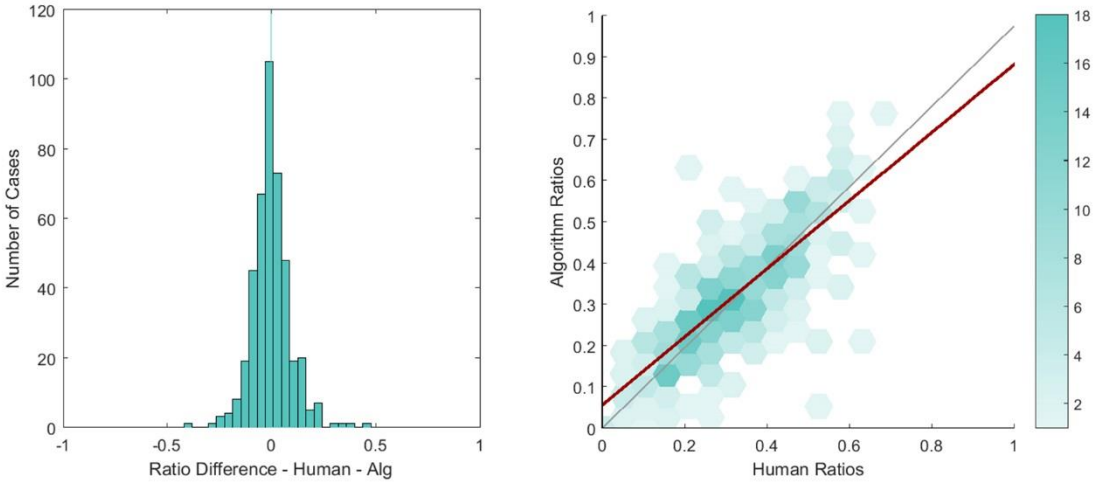
ROC Curves for all 8 tissue subtypes of the CR07 dataset, classified by Algorithm J

The graph shows Area Under the Curve for each tissue subtype:
Tumour parent class: Tumour (0.94), Lumen (0.97), Mucin (0.98), Necrosis (0.90)
Stroma parent class: Stroma (0.92), Vessels (0.86), Muscle (0.97), Inflammation (0.96)



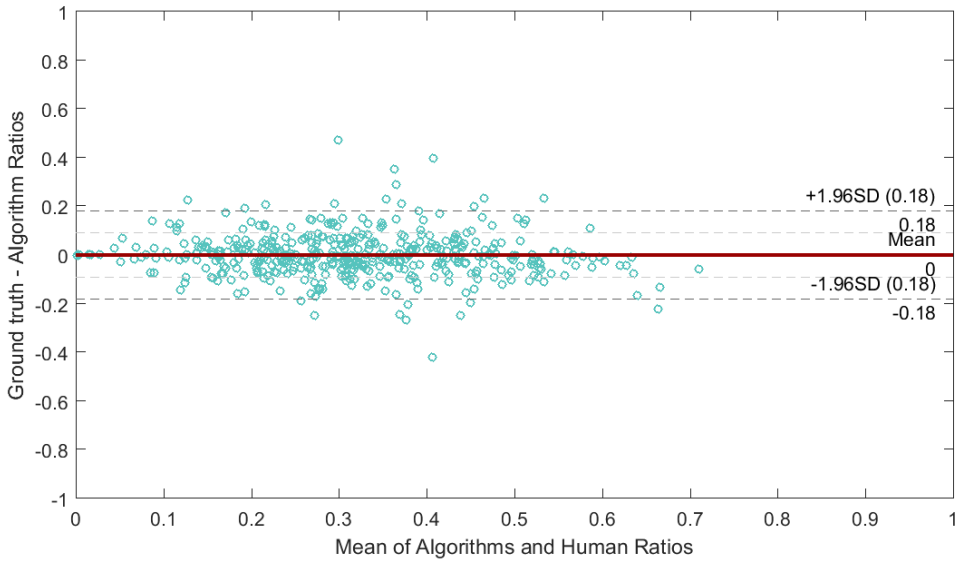
Comparison boxplots for pathologist - pathologist agreement and pathologist - algorithm agreement

Left: Pathologist - pathologist agreement (mean = 0.85, median = 0.88., SD = 0.10, IQR = 0.10)
Right: Pathologist - pathologist agreement (mean = 0.85, median = 0.85., SD = 0.01, IQR = 0.01)
Two sample T-Test P = 0.87
Human agreement is generated from six participants, each scoring 40 images 1024x1024 pixels in size, using the Prospector system (section 4.3).



Histogram and Heatmap Correlation Plots for Algorithm J

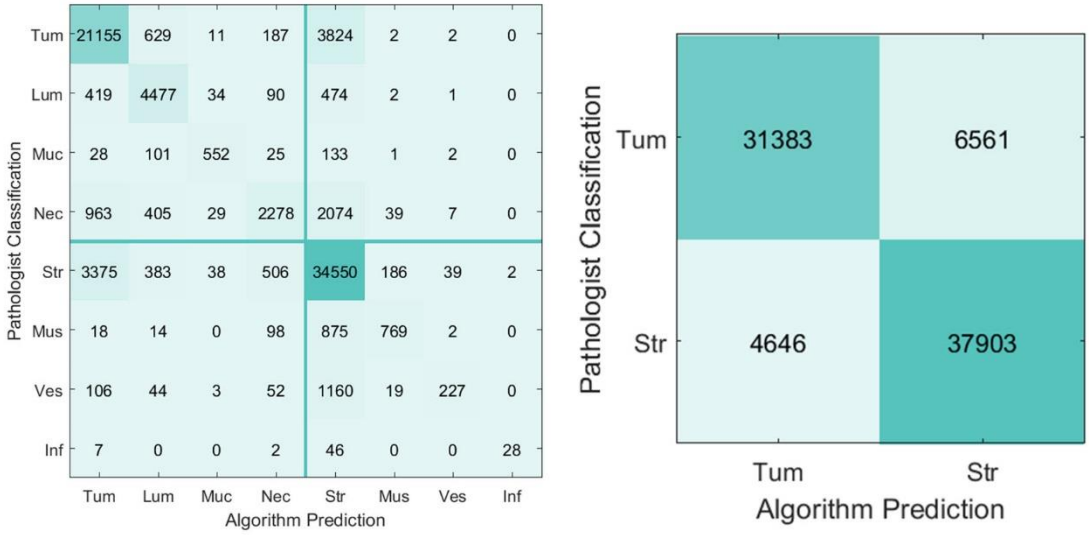
Left: Histogram of ratio differences generated by human and regular segment algorithm. Distribution has a mean bias of 0 (median 0), and standard deviation of 0.09.
Right: Heatmap Correlation plot of the distribution of the ratios. Correlation has R^2 coefficient of 0.61.



Bland-Altman plot of Pathologist and Algorithm J generated TSRs per case

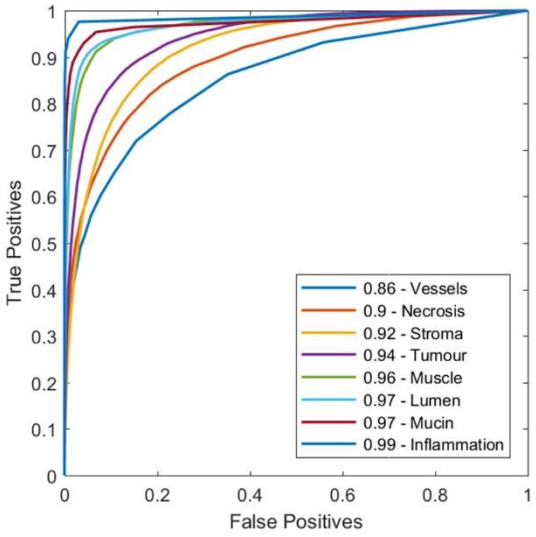
Distribution has a mean bias of 0, with upper and lower limits of agreement of 0.18 and -0.18 respectively (±0.18).

D.11 - Algorithm K: Combined context results CR07 dataset with QC



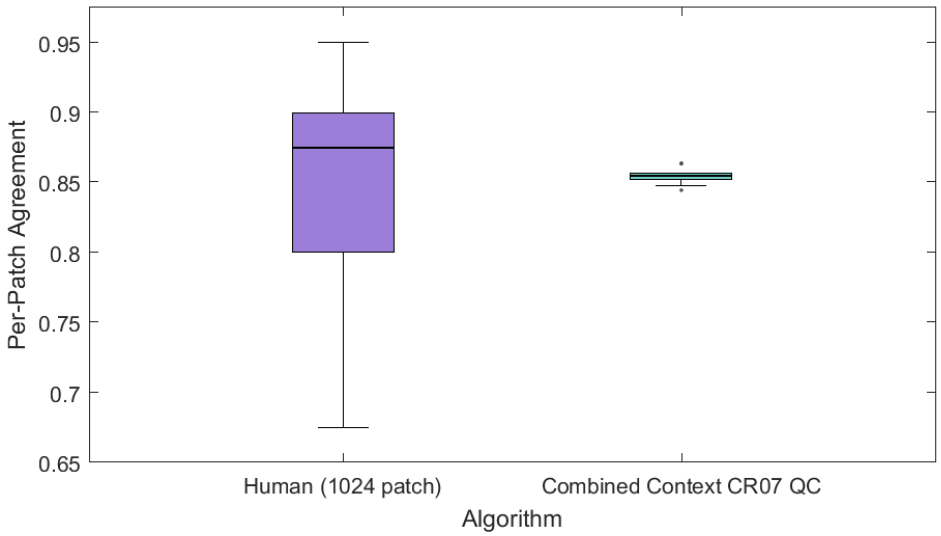
Confusion matrices showing pathologist – Algorithm K agreement

Left: Confusion matrix of all 8 tissue subtypes from dataset. Accuracy = 79.56%, sensitivity (true positive rate / recall) = 0.84, kappa = 0.67 (substantial agreement)
Right: Confusion matrix of subtypes grouped into Tumour and Stroma parent classes. Accuracy = 86.08%, sensitivity (true positive rate / recall) = 0.83, specificity (true negative rate) = 0.89, kappa = 0.72 (substantial agreement)



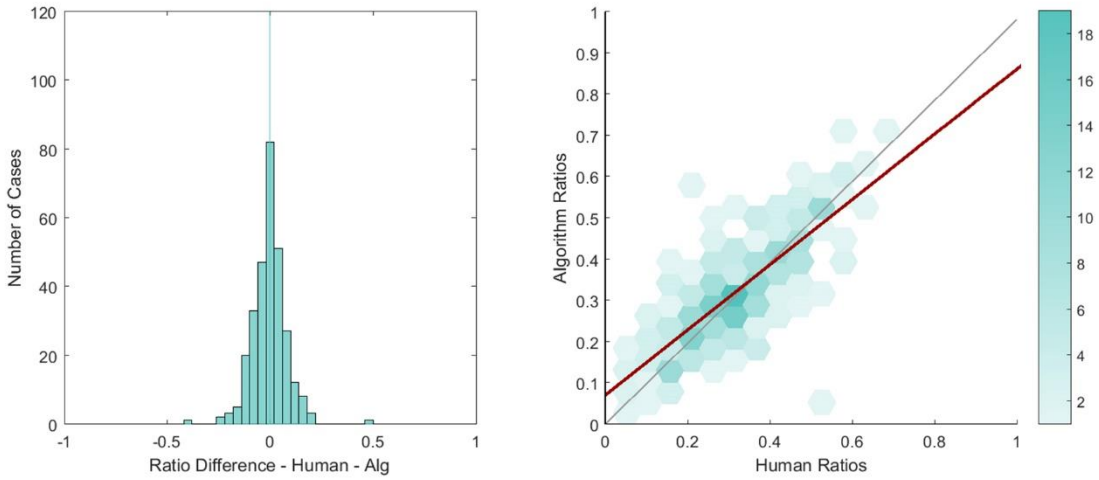
ROC Curves for all 8 tissue subtypes of the CR07 dataset, classified by Algorithm K

The graph shows Area Under the Curve for each tissue subtype:
Tumour parent class: Tumour (0.94), Lumen (0.97), Mucin (0.97), Necrosis (0.90)
Stroma parent class: Stroma (0.92), Vessels (0.86), Muscle (0.96), Inflammation (0.99)
Mean 0.94



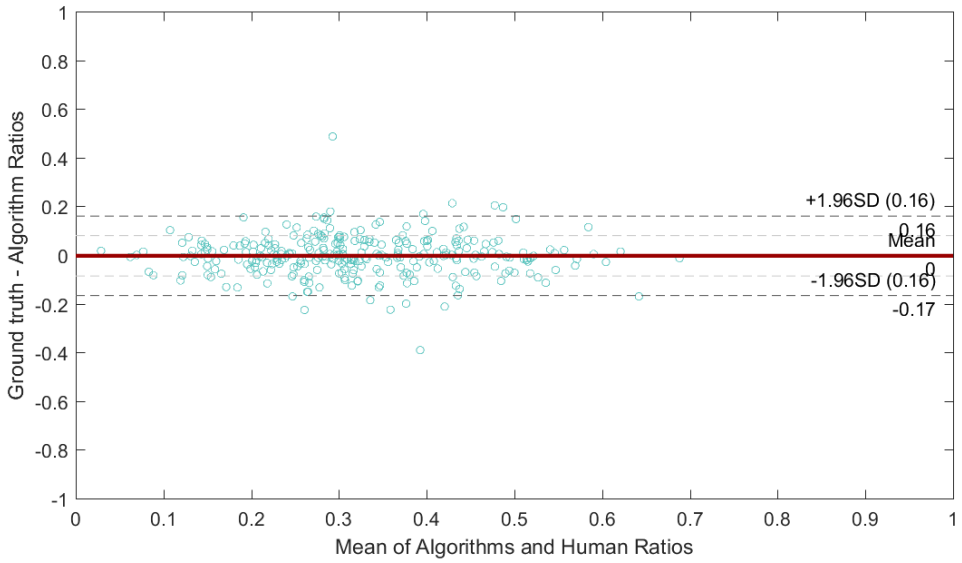
Comparison boxplots for pathologist - pathologist agreement and pathologist - algorithm agreement

Left: Pathologist - pathologist agreement (mean = 0.85, median = 0.88., SD = 0.10, IQR = 0.10)
Right: Pathologist - algorithm agreement (mean = 0.85, median = 0.86., SD = 0.01, IQR = 0.01)
Two sample T-Test P = 0.77
Human agreement is generated from six participants scoring 40 images 1024x1024 pixels in size, using the Prospector system (section 4.3).



Histogram and Heatmap Correlation Plots for Algorithm K

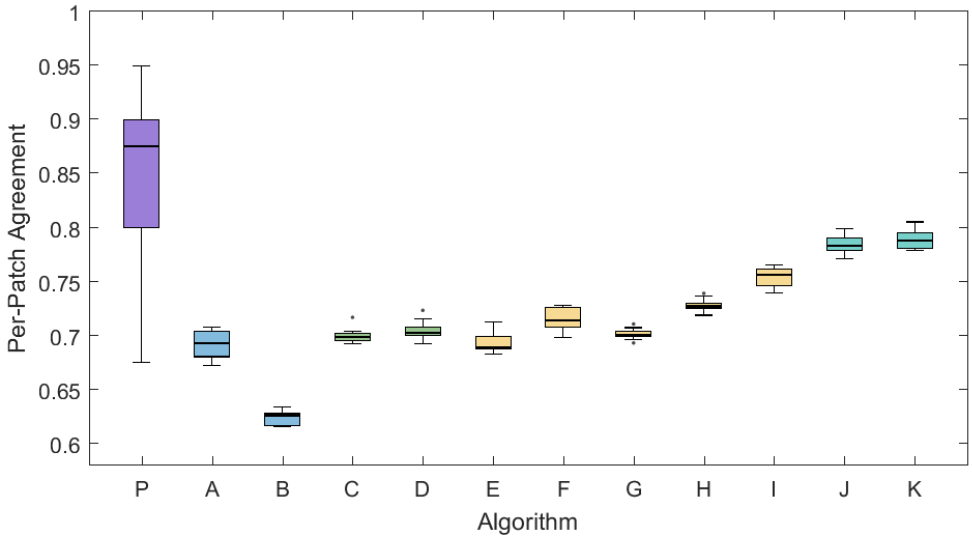
Left: Histogram of ratio differences generated by human and regular segment algorithm. Distribution has a mean bias of 0 (median 0), and standard deviation of 0.08.
Right: Heatmap Correlation plot of the distribution of the ratios. Correlation has R^2 coefficient of 0.61.



Bland-Altman plot of Pathologist and Algorithm K-generated TSRs per case, using the CR07 dataset

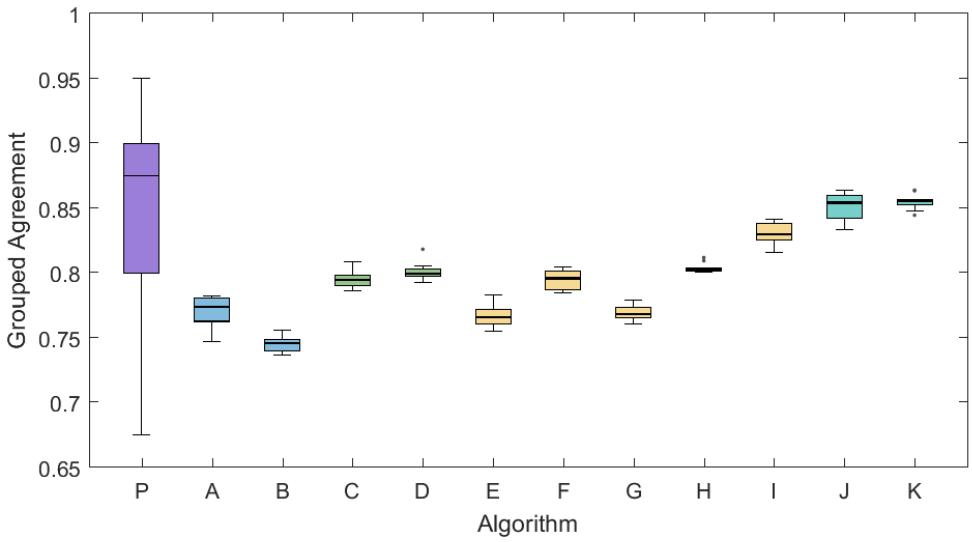
Distribution has a mean bias of 0, with upper and lower limits of agreement of 0.16 and -0.17 respectively (±0.16).

Appendix E – Comparison plots for all algorithms



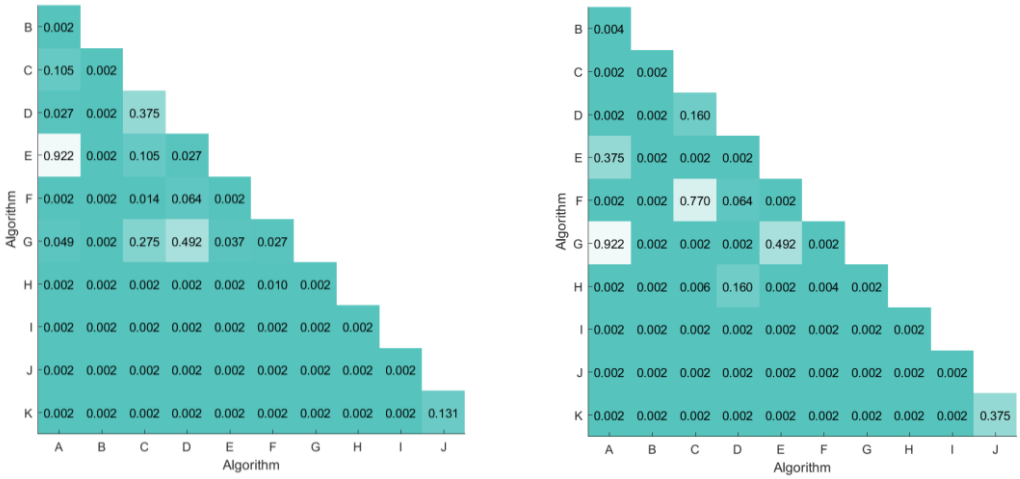
Boxplots of pathologist - pathologist (P) and human-algorithm agreement (A-K) for all algorithms

Note that algorithm agreement is calculated for all eight tissue subclasses, and human agreement is based on the binary tumour-stroma analysis presented in Chapter 4.



Boxplots of pathologist - pathologist (P) and human-algorithm agreement (A-K) for all algorithms

Note that both algorithm and human agreement is for binary tumour-stroma analysis

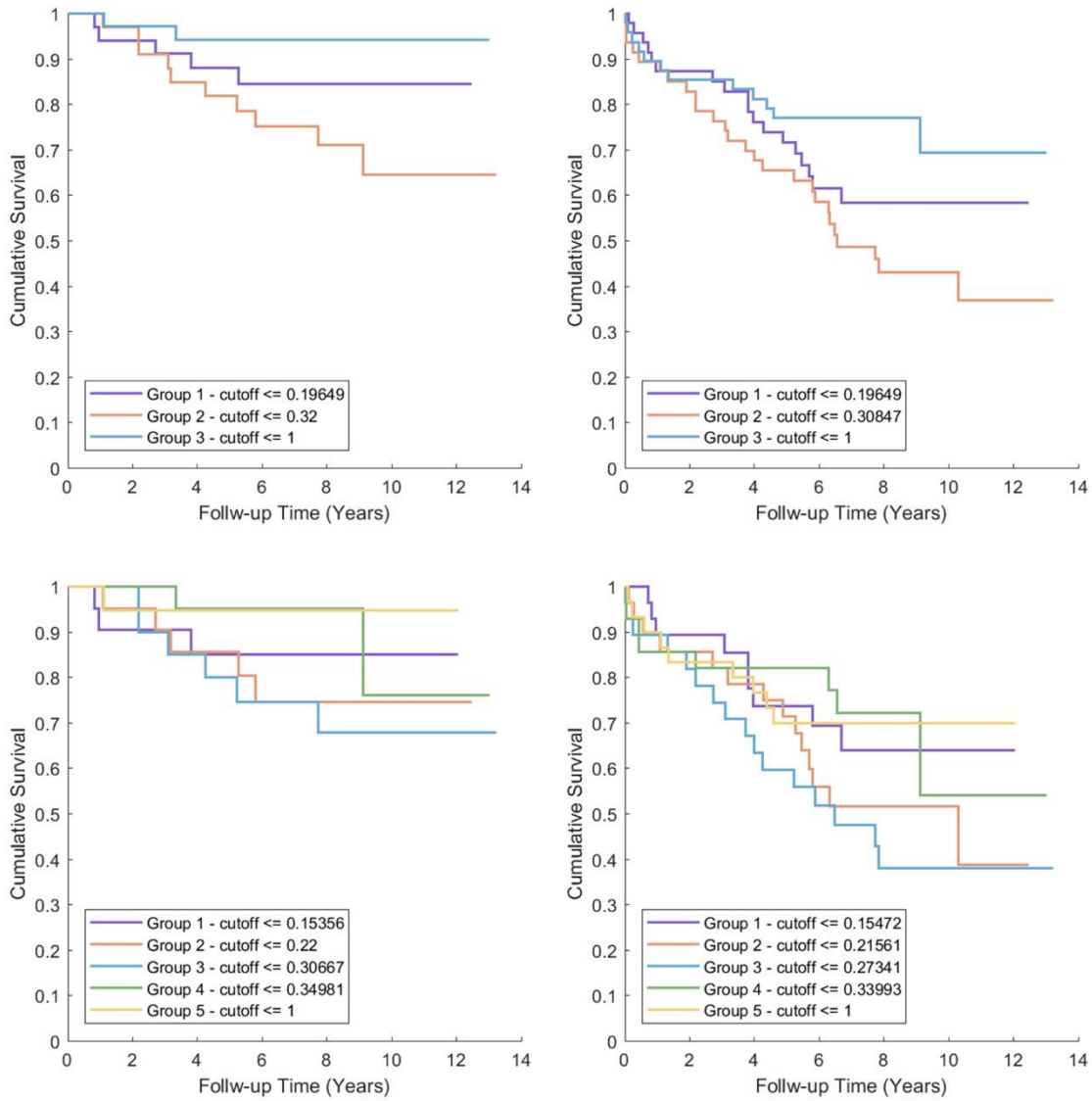


Reduced confusion matrices for Mann-Whitney tests between algorithm agreement statistics for per-patch (left) and grouped (right) agreement

Note that a value greater than 0.0045 indicates the distributions are similar (the compared mean differences are not significantly different from zero), due to Bonferroni correction

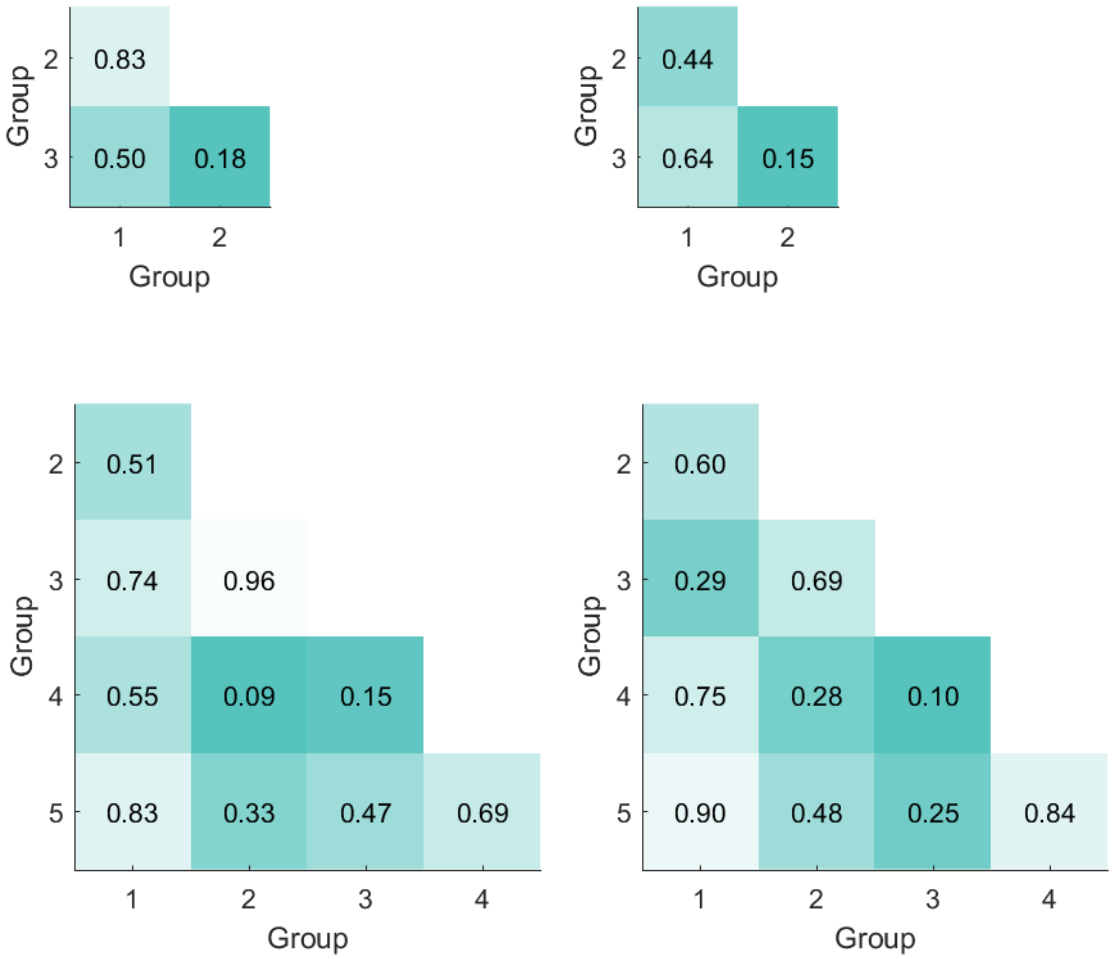
Appendix F – Survival analysis on CR07 dataset

F.1 – Arm 1: patients with preoperative radiotherapy



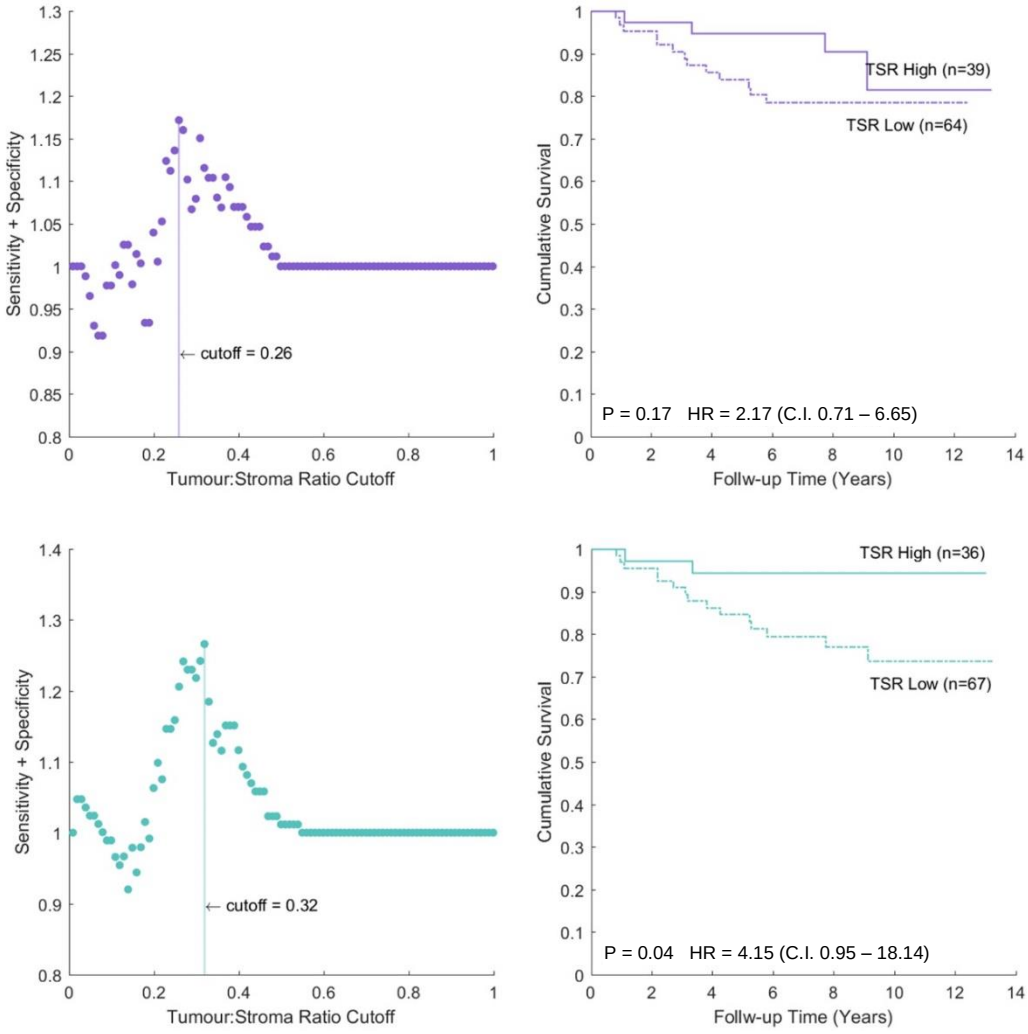
KM survival curves for arm 1 cases stratified into equally sized groups using ordered TSR generated by Algorithm J

Cases were ordered by ascending TSR, and grouped into equal sized groups. The legends on the graphs show the upper cut-offs for each group.
Top: Curves for 3-group cancer specific survival (left) and overall survival (right)
Bottom: Curves for 5-group cancer specific survival (left) and overall survival (right)



Cropped confusion matrices for arm 1, comparing log-rank p-values between three and five patient groups

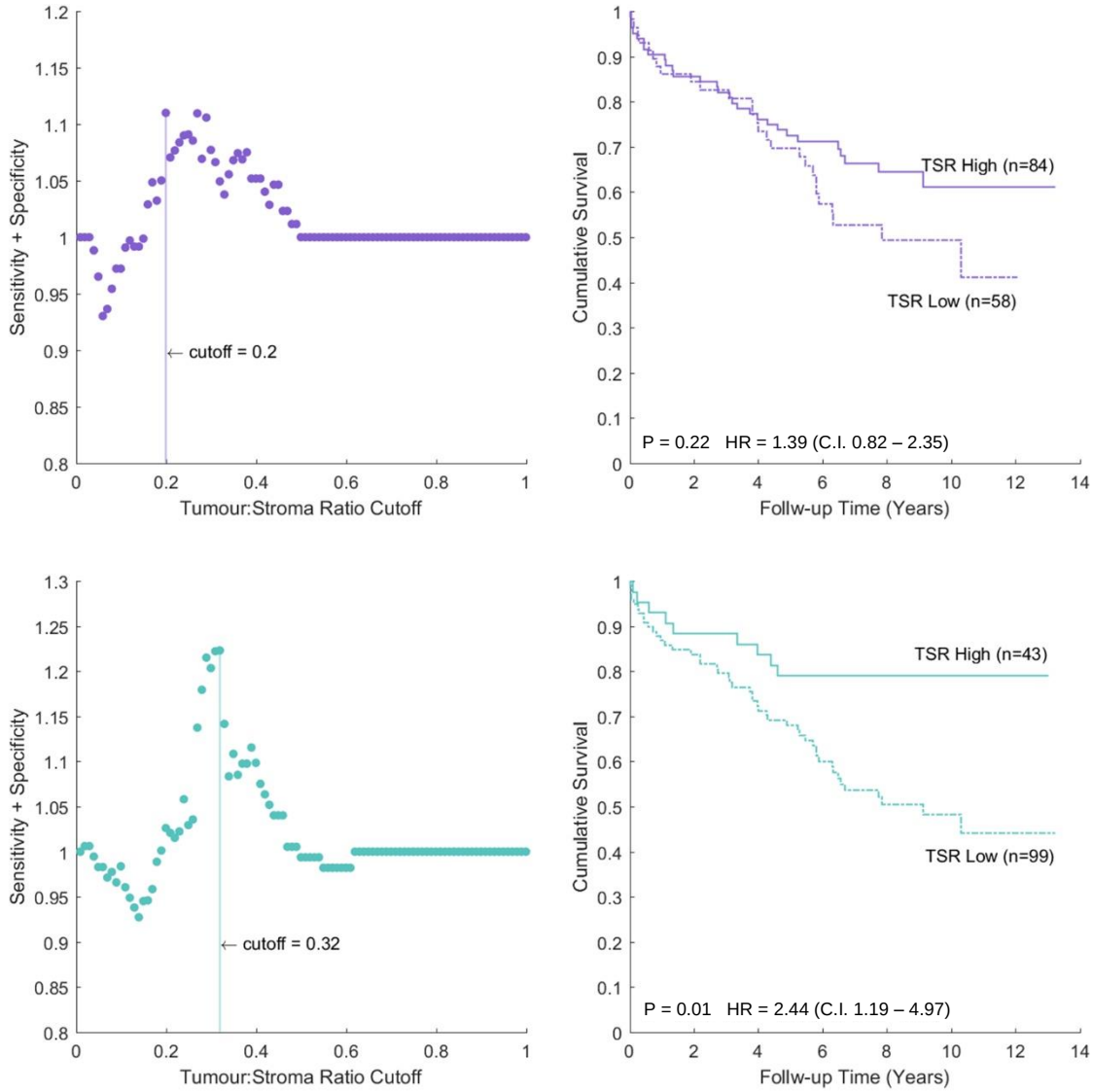
The confusion matrices display log-rank p values for three (top) and five (bottom) pairwise group comparisons. The matrices are cropped along the diagonal to avoid repetition.
Left: Significance values for cancer specific survival
Right Significance values for overall survival



Modified ROC curves and KM survival curves for cancer specific survival using pathologist and machine generated TSRs

Top Left: Modified ROC curve for pathologist-generated ratios (peak value at TSR = 0.26)
Top Right: KM survival curves for pathologist-generated ratios

Bottom Left: Modified ROC curve for algorithm-generated ratios (peak value at TSR = 0.32)
Bottom Right: KM survival curves for algorithm-generated ratios

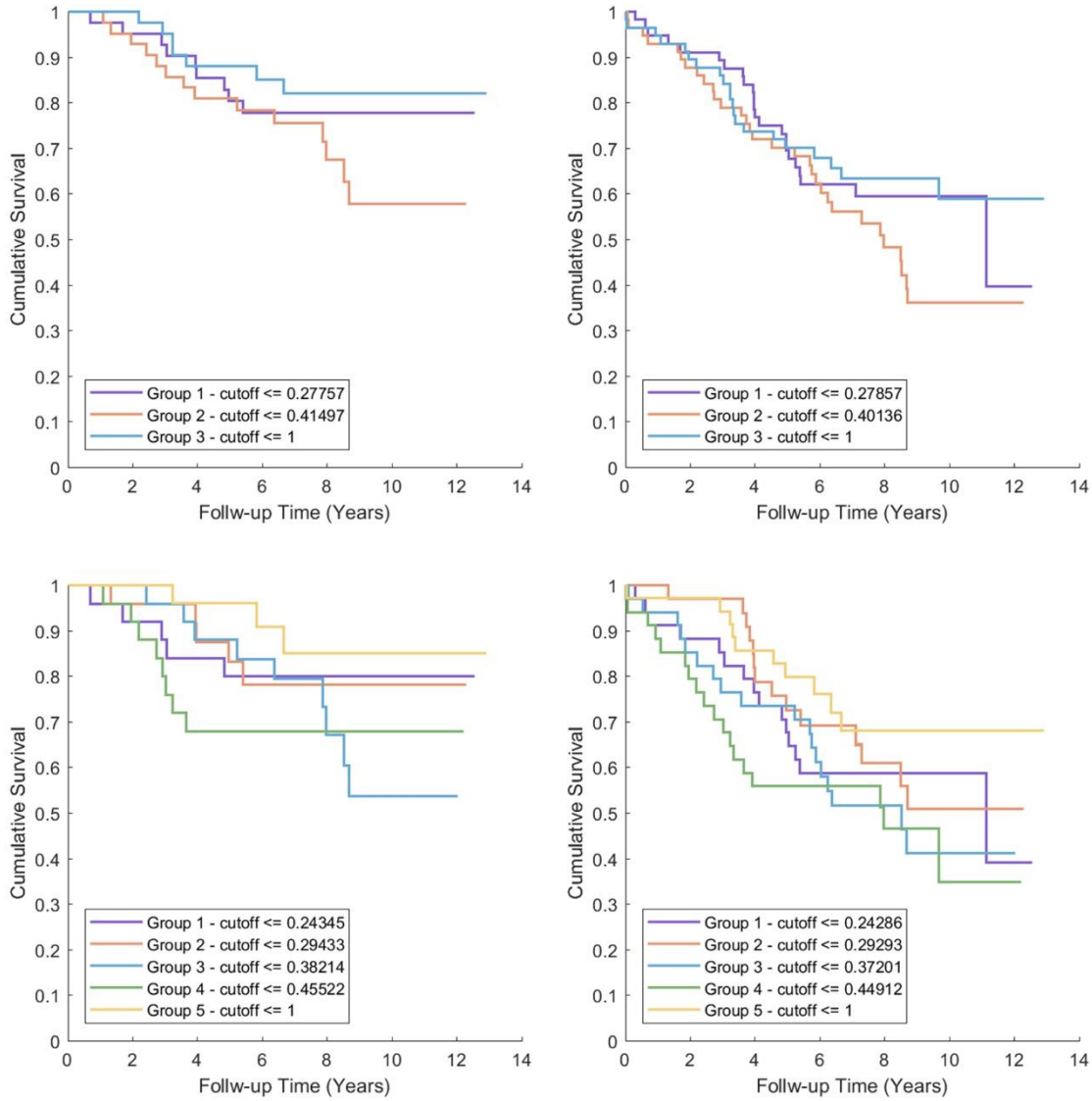


Modified ROC curves and KM survival curves for overall survival using pathologist and machine generated TSRs

Top Left: Modified ROC curve for pathologist-generated ratios (peak value at TSR = 0.20)
Top Right: KM survival curves for pathologist-generated ratios

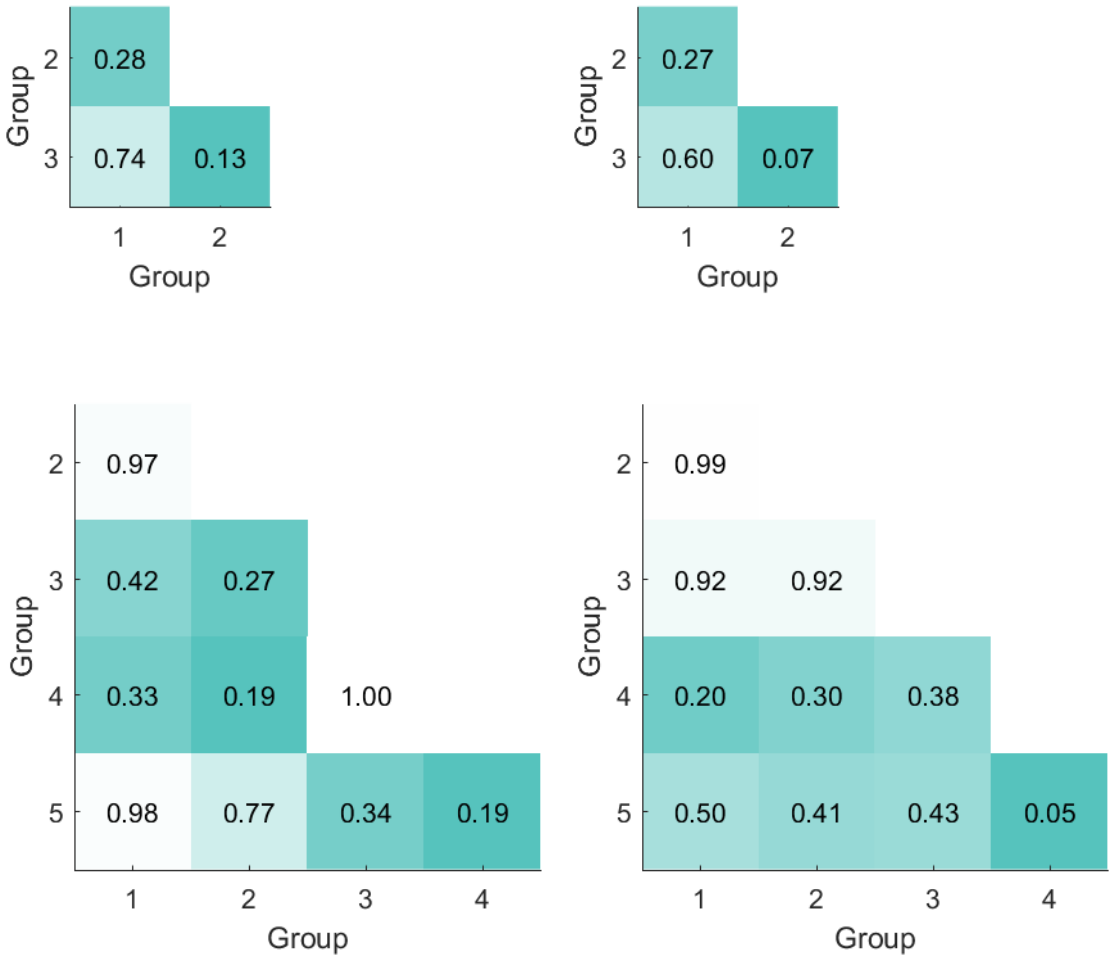
Bottom Left: Modified ROC curve for algorithm-generated ratios (peak value at TSR = 0.32)
Bottom Right: KM survival curves for algorithm-generated ratios

F.2 – Arm 2: patients with postoperative radiotherapy



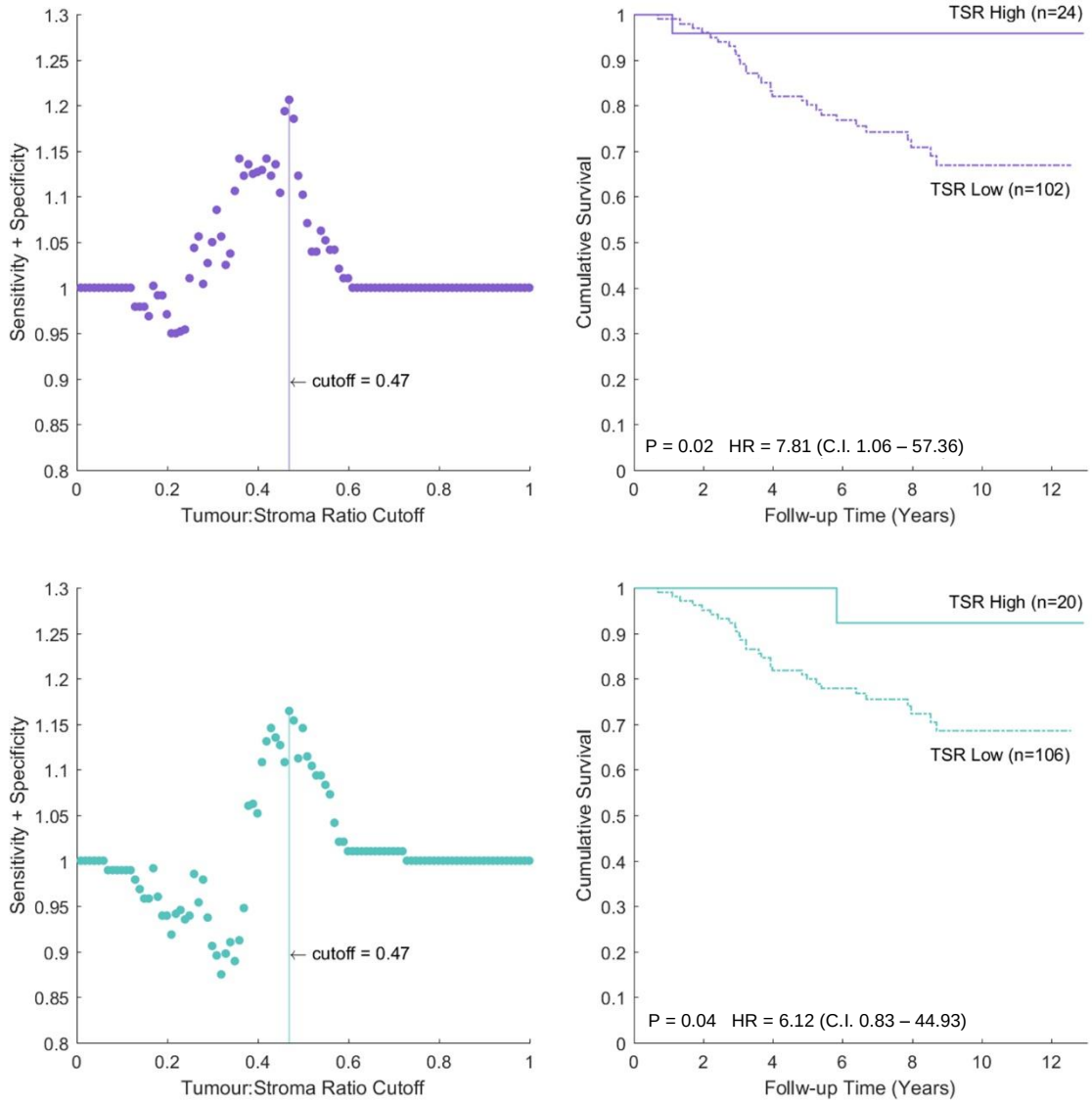
KM survival curves for arm 2 cases stratified into equally sized groups using ordered TSR generated by Algorithm J

Cases were ordered by ascending TSR, and grouped into equal sized groups. The legends on the graphs show the upper cut-offs for each group.
Top: Curves for 3-group cancer specific survival (left) and overall survival (right)
Bottom: Curves for 5-group cancer specific survival (left) and overall survival (right)



Cropped confusion matrices for arm 2, comparing log-rank p-values between three and five patient groups

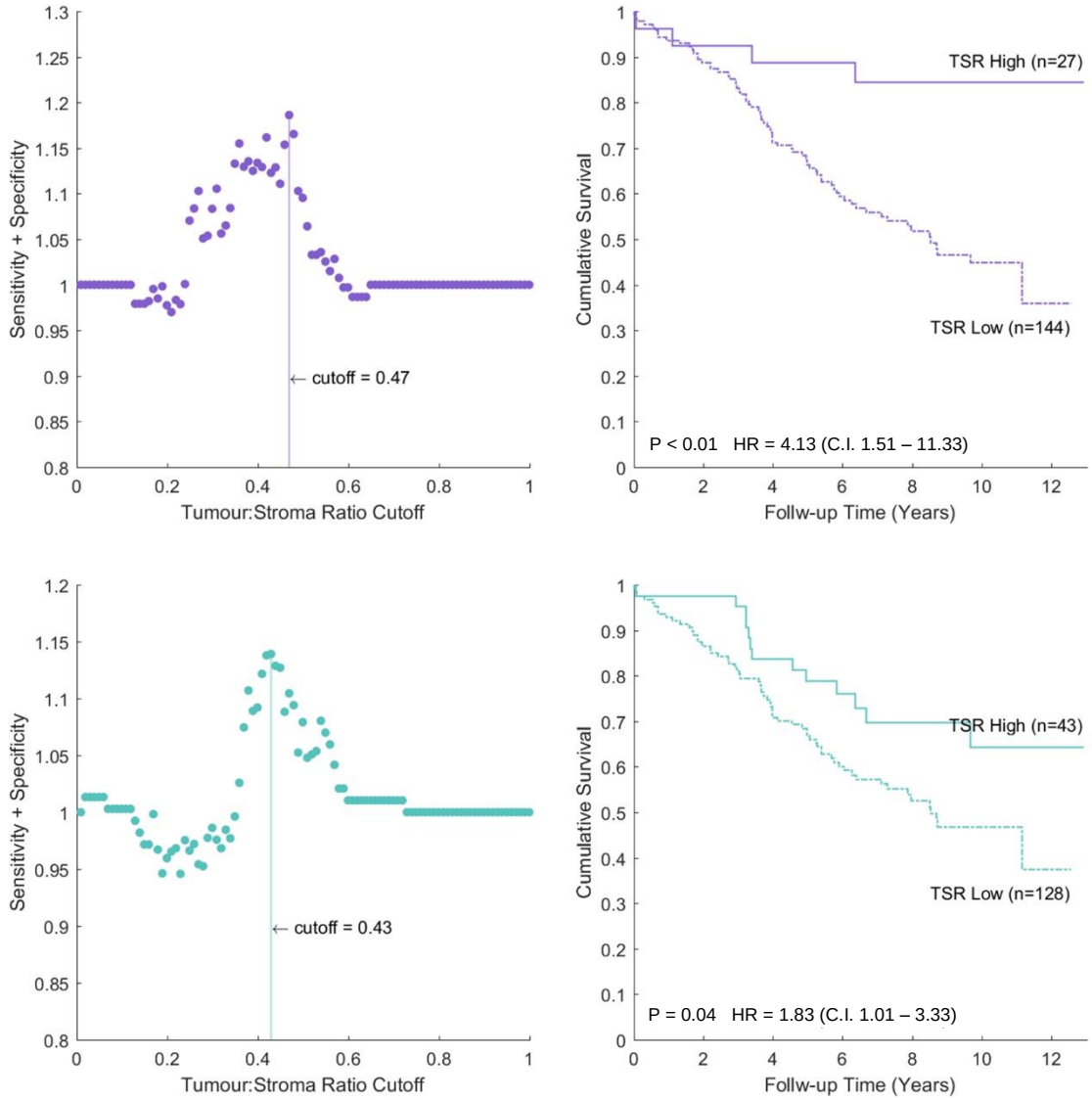
The confusion matrices display log-rank p values for three (top) and five (bottom) pairwise group comparisons. The matrices are cropped along the diagonal to avoid repetition.
Left: Significance values for cancer specific survival
Right Significance values for overall survival



Modified ROC curves and KM survival curves for arm 2 cancer specific survival using pathologist and machine generated TSRs

Top Left: Modified ROC curve for pathologist-generated ratios (peak value at TSR = 0.47)
Top Right: KM survival curves for pathologist-generated ratios

Bottom Left: Modified ROC curve for algorithm-generated ratios (peak value at TSR = 0.47)
Bottom Right: KM survival curves for algorithm-generated ratios



Modified ROC curves and KM survival curves for overall survival using pathologist and machine generated TSRs

Top Left: Modified ROC curve for pathologist-generated ratios (peak value at TSR = 0.47)
Top Right: KM survival curves for pathologist-generated ratios

Bottom Left: Modified ROC curve for algorithm-generated ratios (peak value at TSR = 0.43)
Bottom Right: KM survival curves for algorithm-generated ratios