A Perceptual Colour Separation Methodology for Automated Quantification of Ki67 and Hematoxylin Stained Digital Histopathology Images

by

Peter Morreale

A Thesis presented to The University of Guelph

In partial fulfilment of requirements for the degree of
Master of Applied Science in Biological Engineering

Guelph, Ontario, Canada

© Peter Morreale, April, 2018.
ABSTRACT

A PERCEPTUAL COLOUR SEPARATION METHODOLOGY FOR AUTOMATED QUANTIFICATION OF KI67 AND HEMATOXYLIN STAINED DIGITAL HISTOPATHOLOGY IMAGES

Peter Morreale
University of Guelph, 2018

Advisor: Dr. Robert Dony
Co-Advisor: Dr. April Khademi

This thesis is focussed on the development of a colour separation methodology for quantification of histopathology stains Ki67 and hematoxylin. Traditional methods require tedious manual evaluations prone to inter and intra-observer variability due to inherent subjectivity. Automatic algorithms have been proposed as the solution to manual challenges, however many algorithms are sensitive to wide staining variability common among histopathology images. This thesis proposes a perceptual colour separation framework that models colour information in a way that mimics the human visual systems ability to identify colour content. It is an automatic framework that was evaluated against 30 manually labelled canine mammary TMA core images. The average difference between manual and automatic PI estimates was 3.25%, with a Kappa similarity statistic of 0.86 across four Ki67 cut-off levels, and a linear correlation coefficient of 0.93. Validation studies show the colour separation framework achieves robust results across various Ki67 staining levels.
Acknowledgments:

I would like to thank all those that contributed to this work both through direct influence and external support. I would like to thank my advisor Dr. April Khademi, for both academic and personal guidance given at each stage of this work. Your insight always inspired further thoughts and ideas that could be explored, which allowed me to really develop my understanding of the research material and to consider more possibilities than I could imagine on my own.

Thank you Emily Brouwer for creating the manually identified dataset which took tremendous amounts of time and concentration to count each cell individually. Your time and energy was very much appreciated and this work was only possible due to the contributions you made. I would also like to thank Geoffrey Wood for working with us to supply and generate the required images used in the dataset.

Thank you Dr. Robert Dony for supporting this work first as an advisory committee member and later as my Guelph advisor. Your support allowed me to continue working on my thesis and I would also like to thank you for taking the time to review it during the final stages of development.

I owe so much to my partner Heather, who was both supportive when I was discouraged and uplifting when I achieved a milestone. Your enthusiasm for Canada's natural landscape and our camping trips were a major source of relief while pursuing research, I love you very much. Thank you to my family for providing me with both financial and emotional support, which allowed me to pursue my academic career and manage a work/life balance while I was focussed on research activities.
Contents

Chapter 1: Introduction ........................................................................................................1
  1.1 Motivation ..................................................................................................................2
  1.2 Scope and Challenges ...............................................................................................6
  1.3 Outline of Proposed Solution ....................................................................................8
  1.4 Thesis Outline .........................................................................................................9
  1.5 Contributions ........................................................................................................10

Chapter 2: Background .....................................................................................................11
  2.1 - Materials ...............................................................................................................11
    2.1.1 Hematoxylin and Eosin .......................................................................................11
    2.1.2 Immunohistochemical Stain Ki67 ......................................................................12
    2.1.3 Tumour Acquisition and Specimen Preparation ..............................................13
    2.1.4 Scoring and Grading Systems ..........................................................................16
    2.1.5 Digital Image Generation ..................................................................................18
    2.1.6 Image Quality Challenges .................................................................................19
  2.2 - Colour Spaces for Colour Separation ....................................................................21
    2.2.1 The Red, Green, Blue (RGB) Colour Space .....................................................22
    2.2.2 The Hue, Saturation, Value (HSV) Colour Space .................................................23
    2.2.3 The Luminance, a Chrominance, b Chrominance (L*a*b*) Colour Space ........24
  2.3 - Existing CAD Algorithms and Challenges ............................................................26
    2.3.1 - Pre Processing ...............................................................................................26
    2.3.2 - Stain Separation .............................................................................................27
    2.3.3 - Nuclei Segmentation ......................................................................................32

Chapter 3: Proposed Framework .......................................................................................36
  3.1 - The Human Visual System and L*a*b* Considerations for IHC Analysis ............38
3.1.1 - b* Chrominance Colour Differentiation Characteristics ........................................40
3.1.2 - Perceptual Linearity and b* Chrominance Membership ....................................44
3.2 - Image Processing Methods ..........................................................................................46
  3.2.1 - Colour Separation Threshold ..................................................................................46
  3.2.2 - Determining an adaptive threshold T .................................................................49
  3.2.3 - Pre-Processing ........................................................................................................53
  3.2.4 - Nuclei Detection .....................................................................................................55
3.3 - Validation Methods .......................................................................................................58
  3.3.1 - Nuclei Detection Performance ..............................................................................60
  3.3.2 - Sensitivity, Precision, F1 Score, Effective Rate ..................................................61
  3.3.3 - Cell Radius Estimator ............................................................................................62
  3.3.4 - Adaptive Colour Threshold T ................................................................................64
  3.3.5 - Proliferation Index Accuracy ..................................................................................64
  3.3.6 - Pearson Correlation Coefficient and Kappa Similarity Statistic .........................65
Chapter 4: Validation Results and Parameter Selection ........................................................67
  4.1 - Manual Dataset Generation .......................................................................................67
  4.2 - Nuclei Detection .........................................................................................................70
    4.2.1 - Validation Metric Considerations ......................................................................70
    4.2.2 - Cell Radius Estimate Performance ..................................................................71
  4.3 - Colour Separation .......................................................................................................73
    4.3.1 - Colour Separation Under Various Threshold Conditions ....................................74
    4.3.2 - PI Performance for Colour Thresholds ...............................................................76
    4.3.3 - F1 Score for Colour Separation .......................................................................76
    4.3.4 - Effective Rate for Colour Separation Thresholds ...............................................79
    4.3.5 - Threshold Performance Summary .....................................................................80
  4.4 - Pearson Coefficient and Kappa Similarity .................................................................81
4.5 - Comparison to Other Works .................................................................83

Chapter 5: Summary and Conclusions ...........................................................86

Bibliography ....................................................................................................89
List of Tables

Chapter 2

Table 2.1 - Overall performance trade-off of common nuclei segmentation ..........................35

Chapter 3

Table 3.1 - A visualization of how the b* chrominance colour channel represents
RGB vectors for the range of colours observed in a basic colour wheel image. .................. 42

Table 3.2: Summary of Nuclei Detection Algorithm Steps with Cell Radius Estimator ......56

Table 3.3 - Kappa Agreeability Between Automatic and Manual Methods. .................... 65

Chapter 4

Table 4.1 - Determined Threshold Values ........................................................................... 75

Table 4.2 - Proliferation Index Performance Summary for Colour Separation Thresholds.............................................................................................................. 76

Table 4.3 - F1 Score Summary for Colour Separation Conditions .................................... 79

Table 4.4 - Effective Rate Summary for Colour Separation Conditions ......................... 80

Table 4.5 - Performance Metrics for Balanced Threshold Conditions ............................ 81

Table 4.6 - Comparison Table for Kappa Similarity Calculation........................................ 83

Table 4.7 - Performance Metrics of Similar Works for Ki67 Quantification ...................... 84

Table 4.8 - ImmunoRatio Vs Proposed Work on Thesis Dataset ....................................... 85
List of Figures

Chapter 1

Figure 1.1 - Stain variations due to scanner and stain variability ........................................7
Figure 1.2 - General image analysis framework. ......................................................................8
Figure 1.3 - Proposed algorithm framework ...........................................................................9

Chapter 2

Figure 2.1 - Hematoxylin and eosin stained images .................................................................12
Figure 2.2 - Tissue microarray core stained with Ki67 (brown) and hematoxylin (blue) ..........13
Figure 2.3 - Core needle biopsy procedure ..............................................................................14
Figure 2.4 - Visualization of organized TMA cores ..................................................................16
Figure 2.5 - Morphological features associated with the Nottingham grading system .. ....17
Figure 2.6 - Image artifacts of TMA cores stained with Ki-67 and Hematoxylin .................21
Figure 2.7 - RGB Ki67 and hematoxylin stained image and associated RGB colour channels ........................................................................................................................................................................................................................................22
Figure 2.8 - The HSV colour space mapped to a cylinder .........................................................24
Figure 2.9 - VMF filtering applied to a single hematoxylin stained nuclei ..............................27
Figure 2.10 - RGB optical density and projected plane scatter plots .....................................29
Figure 2.11 - Stain separation achieved through colour deconvolution with associated quantization errors ................................................................................................................................................................................30
Figure 2.12 - Ki67 and hematoxylin variability, nuclear boundaries ........................................32

Chapter 3

Figure 3.1 - Diagram of proposed method illustrating parameter selections at each step in the framework. ..................................................................................................................38
Figure 3.2 - RGB colour wheel and b* channel equivalent......................................................40
Figure 3.3 - Multiple colour distributions of b* chrominance channel ..................................43
Figure 3.4 - All b* intensities observed over all images in the validation dataset plotted on the range of all possible b* intensities from the RGB colour gamut ..................................45
Figure 3.5 - Colour distribution characteristics of the b* channel ...........................................46

Figure 3.6 - Comparison of pathologist labelled manual data for images containing wide stain variability. .................................................................48

Figure 3.7 - Histogram comparison for images with high, low, and mixed expression levels ..........................................................50

Figure 3.8 - Balanced threshold determination without stop (Left) and with stop condition (Right) ........................................................................51

Figure 3.9 - Adaptive threshold comparison for images with high hematoxylin (T = 0), high Ki67 (T = 3.26), and mixed (T = 0.99) expression levels. ..................................................52

Figure 3.10 - Grayscale confidence images H(x,y) and K(x,y) with colour equivalent images obtained by applying a binary image mask to the sample RGB image ............53

Figure 3.11 - Nuclei detection procedure on hematoxylin confidence image H(x,y) ..........57

Figure 3.12 - Final nuclei seed detections for both hematoxylin and Ki67 confidence images imposed on a specific nuclei region ........................................58

Figure 3.13 - Flow diagram of validation procedures and parameter determination after stain separation. .................................................................59

Figure 3.14 - Manually obtained nuclei image with green markers for Ki67 and magenta markers for Hematoxylin ..................................................60

Figure 3.15 - Precision versus Sensitivity trade-off curve ...................................................63

Chapter 4

Figure 4.1 - Cropping of single TMA core from larger TMA block section for the purpose of manual counting in Sedeen Viewer .................................68

Figure 4.2 - Example zoomed in region of TMA image edge for manual nuclei marker placement in ImageJ .................................................................69

Figure 4.3 - F1 score values for each image in the dataset. ..................................................70

Figure 4.4 - F1 performance for arbitrarily selected cell radius estimates and the adaptive approach for radius estimation. ........................................72

Figure 4.5 - ROC curve for sensitivity and precision trade-off performance for Ki67. ....73

Figure 4.6 - ROC curve for sensitivity and precision trade-off performance for Hematoxylin. ..............................................................................73
Figure 4.7 - F1 scores are compared for both Ki67 and hematoxylin stain content using the automatic cell radius estimator (Adaptive) and various arbitrarily selected radius estimates for the colour separation threshold methods. ...............................................................78

Figure 4.8 - Automatic vs manual PI estimates where each data point represents a single image in the dataset. .................................................................82
Chapter 1

Introduction

Cancer is the leading cause of deaths in Canada and is defined by rapid uncontrolled cell division that can exist undetected for long periods of time before symptoms appear. For women, breast cancer has the highest incidence rate of cancers in Canada where an expected 1 in 9 females will develop breast cancer in their lifetime [1]. As a result of an aging population and the higher incidence rate associated with older women, new cases of female breast cancer are expected to increase 55% by 2028 to 2032 if current trends continue [1]. An increase in cancer rates could inflate the economic costs associated with disease treatment as well as more personal costs such as lost wages, which was estimated $3.18 billion in 2009 among patients, caregivers, and parents [2].

Despite the increase in breast cancer incidence rates, mortality rates have been declining which is often attributed to improved screening procedures and improved treatment options [3]. Early detection and better understanding of disease characteristics allows for the development of more accurate diagnosis procedures and better predictions for patient response to therapy. To further understand disease progression and survivability, implementation of various histopathology stains and computer aided methodologies are becoming increasingly common in a pathologist's daily workflow.

Formation of a diagnosis involves the staining of tumour tissue with histopathological stains, typically hematoxylin and eosin (H&E), and visual evaluation through an optical microscope by a pathologist. While H&E are the most widely used stains for determining a diagnosis, other stains are used for identifying disease biomarkers correlated to patient survival and tumour aggressiveness allowing for the formation of a prognosis. Among these additional stains, immunohistochemical stain Ki67, counterstained with hematoxylin, is currently the most widely used assay for measuring and monitoring tumour proliferation showing potential in predicting disease survival, recurrence, and response to various treatment options [4][5][6]. Prediction of patient response, recurrence, and survival is key to further understanding disease progression and ensuring patients receive proper and effective care.

For Ki67 analysis, a pathologist manually identifies positive and negatively stained nuclei to generate a measure of proliferation activity used to classify tumour aggressiveness and form a patient prognosis. Since this framework relies on the skill and experience of a pathologist to manually identify
disease features, it is inherently subjective, observer dependent, and time consuming. The increased availability of digital images for histopathology analysis has opened up the possibility of applying computerized methods to combat the challenges of manual analysis. As a result, computer aided diagnosis algorithms have been proposed as a solution to the difficulties associated with manual methods [7] as it is a more objective, reproducible, and a faster approach to histopathology analysis. The shift to implementation of automatic computer analysis algorithms in a pathologist's daily workflow will assist the pathologist in making a more informed and accurate prognosis and allow reliable statistical comparisons between previously processed data.

Despite the promise of computer assisted analysis methods, current methods still poses various limitations such as the requirement of user defined parameters and limited consideration of the colour content characteristics available in histopathology images. This thesis investigates the challenges and limitations associated with manual and automatic nuclei detection methods and proposes a fully automatic image analysis framework for Ki67 quantification that mimics the human perception of colour content. An automatic framework utilizing colour content can result in a closer comparison with manual data, while maintaining robust analysis for variable stain levels.

1.1 Motivation

To quantify the disease characteristics of breast cancer, pathologists perform evaluations on tissue samples obtained through either a biopsy procedure or a surgical sample from a removed tumour. These evaluations rely on visual assessment techniques and apply standardized scoring systems specific to the suspected disease type and histopathological stains applied to the tissue sample. These standardized systems are intended quantify specific disease indicators through visual observation leading to a final cancer diagnosis. The most common breast cancer scoring method is the Nottingham [8] grading system which generates a histologic grade based on a pathologist's assessment of tumour tubule formation, nuclear pleomorphism, and mitotic count. The histologic grade is often the main determining metric for a final diagnosis, however in an effort to better model disease characteristics, predict patient response to therapy, and obtain disease free survival statistics additional characteristics and staining methods must be considered.

In addition to diagnosis, a prognosis allows for the monitoring and prediction of tumour activity over time. A breast cancer prognosis can be determined through application of immunohistochemical (IHC) stain Ki67 counterstained with hematoxylin and results can be compared against other patient outcomes. Ki67 is currently the most widely used assay for measuring and monitoring tumour proliferation, showing potential in predicting disease survival, recurrence, and response to various treatment options [4], [5], [6]. Quantification of Ki67 is used to characterize tumour aggressiveness via a
proliferation index (PI) metric which can be used to identify ideal patient treatment options and predict patient survivability. The PI is determined via visual assessment by a pathologist through an optical microscope and represents the percentage of positively stained cells (Ki67) in the sample opposed to negatively stained healthy cells (hematoxylin). Inclusion of a prognosis, in addition to an initial diagnosis, through application of Ki67 allows practitioners to more objectively determine what course of treatment is available to patients as well as inform patients of their recovery probability.

Ki67 has previously been evaluated as a predictive marker for breast cancer and a metric for modelling survival statistics. In [6], tumour samples were collected from 129 patients with primary breast cancer with a median follow up period of 42 months. Patient overall survival and disease free survival were compared to the various levels of Ki67 proliferation rates which were considered low for samples with less than 20% expression and high proliferation activity was associated with a rate greater than 20% [6]. Higher disease free survival was observed for lymph node-negative patients with a low Ki67 PI. For patients in a nuclear estrogen receptor negative subset the PI also predicted a higher probability of tumour relapse in samples showing high Ki67 content. Similarly, a lower disease free survival probability was observed for postmenopausal women with high Ki67 values opposed to a higher disease free survival probability for women in the same subset with low Ki67 values [6]. Based on the above observations, it was shown that patients with slow proliferating tumours had a significantly higher probability of survival than those with more aggressive tumours. By using Ki67 as a predictive indicator of survival statistics, medical professionals may be able to better predict patient outcome and provide more accurate determinations of tumour characteristics.

In addition to the prediction of long term survival statistics, Ki67 has been considered as a determining variable for selection of appropriate patient treatments and as a method of measuring response to therapy. In [9], chemotherapy response for 152 patients with breast cancer was evaluated by examining Ki67 expression levels before and after treatment and prior to surgical mastectomy. It was observed that the median positive percentage rate of Ki67 per tumour decreased from 16% to 8% across patients from initial biopsy to mastectomy. In this case, Ki67 was used to evaluate patient response to therapy suggesting it could be implemented to track changes in tumour activity over time. In some cases Ki67 expression was unchanged for patients who did not respond to treatment whereas Ki67 expressions were lower in patients with complete response to chemotherapy treatment. Since reduction of Ki67 expression is more often observed in good responders to chemotherapy, [10] explored the use of Ki67 as a predictive marker of early breast cancer recurrence by examining the proliferation activity before and after treatment. Ki67 expressions in biopsy and surgical specimens were examined for 116 patients with residual invasive disease in the breast. The mean Ki67 expression value after neoadjuvant chemotherapy...
was higher in metastatic patients than in non-metastatic patients. It was also observed that tumours with higher Ki67 expression in the surgical specimen tended to be associated with early metastatic development. This study suggests that the use of Ki67 for tracking tumour development and response to therapy may be able to predict early metastatic development and recurrence potential.

Despite the predictive and prognostic potential of Ki67, application in regular clinical practice is limited due to numerous sources of variability and staining inconsistencies. As mentioned, Ki67 expression levels are determined through the identification of a PI and can be determined by a pathologist implementing various techniques. One approach is a fast scanning method where a pathologist simply estimates the number of positive nuclei in a sample through fast visual observation. This method allows a pathologist to process a large number of samples in a short amount of time, however it is not an accurate measurement of proliferative activity and is not acceptable for clinical practice where a measured PI could be used to predict patient treatment or outcome. Another approach is to individually count large numbers of positive and negative nuclei either in a specific region of interest or by splitting a sample into smaller quadrants. The PI is then calculated based on the total nuclei counts. While this approach obtains more reliable and accurate results it is extremely tedious and time consuming for slides that may contain many thousands of stained nuclei. Therefore manual counting is not possible in a pathologist's already demanding workflow and the fast scanning method lacks the attention to detail required for an accurate prognosis. Another barrier to clinical implementation is the inter and intra-observer variability found among pathologists and different labs performing PI estimation regardless of established standardized counting procedures. Despite the barriers to clinical implementation, Ki67 has been viewed as one of the most robust IHC biomarkers for obtaining measures of proliferation activity in clinical practice and could be implemented in clinical trials with further standardization criteria [4].

To utilize the predictive potential of Ki67 while solving the challenges of manual PI estimation, digital histopathology analysis and computer aided diagnosis algorithms have become increasingly relevant in disease quantification and are expected to play an increasingly significant role in a pathologist's daily workflow as more robust and efficient algorithms are developed [11], [12]. Existing algorithms have shown they are capable of attaining fast and objective results comparable to manual segmentation methodologies or PI estimates [13], [14]. Since computational algorithms are created based on mathematical formulae and specific criteria they are more reproducible and objective than individual pathologist assessment. Automated algorithms allow a pathologist to produce more objective prognostic results and can be applied to large scale studies so statistical analysis methods do not have to rely on the judgement of multiple pathologists and institutions.
While automatic image analysis algorithms provide solutions to the problems associated with manual Ki67 analysis, existing algorithms possess various limitations that can inhibit their effectiveness in a clinical or practical setting. Common challenges include the use of computationally expensive iterative methods for stain separation, the requirement of user defined parameters for accurate analysis, and reliance on grayscale methods that do not consider the colour information available in digital histopathology images. Many automatic image analysis algorithms incorporate a colour deconvolution (CD) [15] method in the framework such as in [16], [17], [18]. CD relies on the Beer-Lambert law of light absorption, however Ki67 staining can produce non-linear light scattering properties that will be shown and discussed in following sections. Since stain separation is typically one of the first steps in image analysis algorithms, inconsistencies present during colour deconvolution may greatly impact final results by decreasing reproducibility and accuracy. Other methods usually involve different iterative approaches, such as learning classifiers and clustering algorithms, for nuclei segmentation and stain separation as in [19], [20], [21]. Iterative approaches often require large or repetitive computations which negatively impact processing time on large high resolution histopathology images. An additional common dependency with existing algorithms is reliance on user defined features or parameters which introduces subjectivity into the model resulting in reduced repeatability and robustness. With the above challenges, it is currently difficult to accurately and objectively model stain content due to the wide variability common among chemically stained tissue samples.

Further development into automatic image analysis algorithms for automated Ki67 detection algorithms can contribute to solving the mentioned challenges with manual scoring methods and assist a pathologist in their daily workflow. With robust automatic methods, implementation on large scale medical datasets and images may become possible and development of colour descriptive models may be applicable to other histopathological stains and disease diagnosis/prognosis procedures beyond breast cancer. Contribution in the area of automatic image analysis of Ki67 for breast cancer can have a significant impact on cancer treatment in society and potentially extend to other disciplines of colour image analysis.

As a result of the potential benefits associated with application of Ki67 in breast cancer analysis and the stain quantification challenges with current automatic algorithm methodologies, this thesis focuses on the development and investigation of existing colour models and algorithms for histopathological stain separation as well as nuclei detection and PI estimation. The following section describes the scope of the proposed investigation and the associated challenges with developing the proposed methodology.
1.2 Scope and Challenges

Typical computer aided frameworks for Ki67 analysis are compared against manually labelled ground truth data obtained by one or more pathologists which posses various challenges. Due to the variability between pathologist evaluations, datasets are ideally labelled by 2 or more pathologists to obtain more consistent results for validation purposes. Unfortunately, labelling by 2 or more pathologists is often not possible due to the demanding time commitments and tediousness of producing manually labelled datasets. Despite the validation dependence on pathologist labelled images, most methods do not consider the fundamental decision making process performed by a pathologist via the human visual system's (HVS) perception of staining colour and intensity. Existing methods often generalize the histopathology images by treating every image similarly if not identical to the next, while a pathologist is able to identify over or under-stained images and adjust their criteria for positive and negatively labelled cells on an image by image basis. Therefore it is within the scope of this thesis to develop a novel colour separation method that removes the current dependency on CD. The colour separation method will be applied to quantify Ki67 content by generating an image specific PI after application of pre-processing, nuclei detection, and colour separation procedures.

While image analysis algorithms are able to calculate and discretize information much better than their human counterpart, validation of results remain dependent on human professionals and can negatively affect the reproducibility of an automatic method. In addition to the dependence on pathologist labelled data, automatic methods are often validated on individually obtained datasets that may not be reflective of what is observed in routine clinical practice. In a practical setting, scanned slides may contain imaging variations depending on vendors, may suffer from improper staining procedures, and can contain many levels of stain concentrations which affect colour characteristics. A pathologist is able to make up for these inconsistencies by shifting their decision making criteria. For example, if a tissue sample was over-stained with Ki67 content so there is a light brown tint across the image, a pathologist may raise their criteria for what is considered positive Ki67 staining content to account for this discrepancy. Although a pathologist can actively shift their criteria for counting procedures, manual count for large scale datasets suffer from inter and intra-observer variability because manual counting of several thousand cells is error prone and time consuming, especially for large tissue microarray (TMA) studies [22]. An example of potential stain/colour variations as a result of preparation and scanning methods are shown in Fig. 1.1. Crafted datasets on the other hand may be composed of images with very little stain variation or images that are specifically selected for their clear cell morphology. This variation in testing sets can result in learned classification algorithms tailored to a specific dataset or hand crafted methods that are inconsistent when applied to stained images with wide staining variability.
Another major challenge with automatic stain separation in histopathology images is that there are no ideal levels of colour content for specific stains. For example in magnetic resonance imaging (MRI) data there are distinct gray levels associated with various anatomical structures that can be manipulated to obtain accurate results. In histopathology analysis however the addition of colour content and various stain types add many potential dimensions for nuclei representation when considering multiple colour spaces that contain no ideal expected range of intensities. Therefore automatic methods often rely on CD which utilizes pre-measured stain vectors for stain separation which are effective for H&E but are not as consistent and reliable for some Ki67 stained images [23]. In addition to CD, clustering methods may use various colour spaces to characterize regions that correlate to Ki67 and hematoxylin. This sort of generalization often does not hold for images containing wide stain variability as positive identification of IHC stains like Ki67 require a nuclei be stained brown where a darker brown is not necessarily more indicative of positivity over a lighter stained nuclei.

Computational efficiency must be considered for an image analysis algorithm as histopathology images generate very large high resolution images containing up to billions of pixels. If unchecked, iterative methods relying on a pixel by pixel classification scheme can spend a significant amount of time processing just one image due to the amount of data in the image. As a result, each step of a proposed method must be designed to minimize computational requirements, while also incorporating enough information to form adequate analysis. In combination with large image sizes, the ability to obtain numerous histopathology images for research purposes is difficult as medical data must follow strict ethical guidelines and a pathologist must generate manually labelled ground truth information.

Implementation of an automatic algorithm for Ki67 quantification that considers the decision making criteria performed by a pathologist can have distinct advantages over other proposed methods. It would also provide an alternative to the CD method, which can be inconsistent for IHC stained images. Since colour characteristics are considered independent of observed images, the method would not rely on
stain determination generated from expensive clustering or learning methods. In this work, an algorithm that models the HVS is developed that mimics the current trusted human evaluation methods. By identifying colour characteristics similarly to a human observer the algorithm can obtain more consistent results for images containing various staining characteristics. A method robust to stain variation is also more likely to maintain robust results for other datasets. Additionally, computational requirements can be reduced by modeling colour content without expensive iteration requirements.

1.3 Outline of Proposed Solution

This thesis investigates the design and development of a colour separation method and colour image analysis methods that perform PI estimation in Ki67 images. The first portion of this thesis evaluates the existing methodology used for quantification of Ki67 as well as available colour processing techniques and various colour spaces commonly utilized. It is typical for many histopathology analysis algorithms to be constructed around a general image analysis framework observed in Fig 1.2. The image datasets are usually obtained through the use of a whole slide image (WSI) scanner at specific levels of cell magnification which will be discussed in greater detail in the following literature review section.

![Figure 1.2 - General image analysis framework.](image)

Since image quality is largely dependent on staining procedures and often contains significant stain variability, image processing algorithms implement pre-processing steps in an attempt to reduce variability or to condition the images in a way that is helpful to the successive steps in the algorithm. In order to account for various sources of image acquisition variability, this thesis applies a vector median filter (VMF) to smooth small colour variations that occur as a result of staining variability and static noise that can occur as part of the image acquisition process.

As previously mentioned, determination of a PI is required to describe the aggressiveness of a tumour and form a final prognosis. Many automatic methods for nuclei segmentation/detection have been proposed for H&E stained images where all nuclei are identified with H&E stained stromal tissues are considered background content. Unlike H&E images, Ki67 and hematoxylin both stain nuclear content when specific indicators are present. Therefore, nuclei content must either be separated into positive and negative stain content or a final classification must be performed to after nuclei segmentation to assign an appropriate staining class. This thesis proposes an initial background subtraction followed by a fully automatic stain separation method before implementation of a modified nuclei detection algorithm
originally proposed for H&E stained images. The steps of the proposed algorithm can be observed in Fig 1.3.

![Proposed algorithm framework](image)

**Figure 1.3 - Proposed algorithm framework.**

Extensive validation studies are performed to observe the performance of the proposed method including evaluations of individual nuclei detection for each stain type, percentage differences comparison between the automatically determined PI and manually labelled ground truths, and calculated statistical similarities metrics between the automatic and manual results.

**1.4 Thesis Outline**

The remaining thesis content is organized into the following chapters. Chapter 2 describes various background content identifying the use of histopathological stains for disease quantification including slide preparation and staining procedures. The typical image acquisition process will be described highlighting the types of samples used in histopathology analysis as well as the associated challenges and impact on the design and development of image analysis algorithms. Finally, a review of commonly used colour spaces in histopathology analysis is performed and prior image analysis algorithms are presented along with the specific advantages and barriers to successful implementation for Ki67 quantification and PI estimation.

Chapter 3 describes the specific components of the proposed framework in the proposed methodology consisting of pre-processing steps, stain separation, colour space histogram correction, and final nuclei detection for each stain type. Chapter 4 discusses the methods of experimental validation used to evaluate the proposed method. The performance of the nuclei detection for each stain type, PI estimation, similarity statistics, and selection of parameters are described with relevant figures and mathematical equations. Chapter 5 presents the experimental results of the proposed method and draws
comparisons to other Ki67 quantification algorithms and manual data. The thesis concludes in Chapter 6 which contains a thesis summary and conclusion.

1.5 Contributions

- Novel colour separation methodology to separate IHC (DAB) stains and hematoxylin that is robust to stain concentration, colour variability due to differing stain vendors, and scanner vendors.
- Investigation of colour space characteristics and representations for variable images.
- Evaluation of several colour threshold techniques to determine optimal colour separation.
- Design of an automated nuclei diameter estimation method to automatically estimate the size of nuclei in an image for each stain.
- Design of an image analysis framework for Ki67 PI calculation which performs better than open source implementations, and results in an average 3% error rate when compared to manual counts.
- Design of a nuclei validation framework to compare automatic and manual seed locations and overall counts.
Chapter 2

Background

This chapter provides the background material associated with the design and development of the image analysis framework for Ki67 and hematoxylin stained images. It begins by discussing the pathology slides themselves as preparation procedures are extremely relevant to overall digital image quality. The stains used to visualize cells, tissues and biomarkers, slide preparation procedures, scoring systems, digital pathology, and image quality issues are discussed in detail. It continues by defining common colour spaces used in histopathology analysis and how they relate to the development of the colour separation method. Finally, an extensive literature review of the current state-of-the-art algorithms is presented.

2.1 - Materials

As mentioned, this thesis proposes an automatic Ki67 quantification algorithm for digital histopathology images. Before an algorithm can be applied to a digital image, a tumour sample must be collected, stained, and imaged. The following section describes the fundamentals and application of histopathological stains, tissue sample collection, preparation, and staining procedures that allow for the generation of a high resolution image required for computer aided diagnosis (CAD) analysis. The methods of scoring and grading stained tumour characteristics are also discussed as they the pathologist's tool for determining manual ground truth information.

2.1.1 Hematoxylin and Eosin

H&E have long been used for histopathology analysis due to its effectiveness at highlighting the structural characteristics of cell nuclei and surrounding tissues. Hematoxylin binds to proteins in the cellular nuclei resulting in a blue colour and eosin binds to structural proteins in the stromal tissue appearing pink. These stains were introduced more than a century ago, however they have remained the primary tool for identification of breast cancer due to the ability to capture cellular characteristics [24]. As a result of its consistent use in histopathology, diagnosis procedures rely heavily on H&E analysis with particular emphasis on the determination of a histologic grade. Malignant cells in breast cancer are commonly associated with abnormal morphological features such as enlarged nuclei, poor tubule structure, and high mitotic activity [25]. Fig 2.1 illustrates the application of H&E on both benign and malignant tumour specimens. As previously mentioned, application of H&E alone is not sufficient for
predicting a prognosis and correlating disease characteristics to long term survival. With application of hematoxylin with the addition of Ki67, proliferative activity can be observed and quantified.

![Figure 2.1 - H&E stained images. Benign (Left) and malignant (Right) tissue specimens.](image)

### 2.1.2 Immunohistochemical Stain Ki67

Uncontrolled cell proliferation is a hallmark of cancer which has resulted in many methodologies for detection and quantification in order to better understand disease progression, treatments, and recovery potential. Proliferation activity can be assessed through application of IHC stain Ki67 which has become the most widely used assay for measuring and monitoring tumour proliferation [4]. The Ki67 protein is present at all stages of the cell cycle with varying levels of intensity. Overall, levels of Ki67 are low during the initial stages of the cell cycle and progressively increase until reaching a maximum at mitosis [26]. Ki67 is not expressed during the cell resting phase and can pose a proliferative marker for any human cell type [27]. The presence of the Ki67 protein during cell division and the ability to detect its expression with IHC staining has resulted in its continued study as a proliferative biomarker of prognostic significance. The Ki67 stain binds to cells undergoing cell division resulting in brown coloured nuclei associated with cancerous cells due to the large increases in cell replications within cancerous tissue. Since the Ki67 protein is present in all cells during proliferation, it may be implied that the application of Ki67 staining on normal breast tissue would result in some percentage of healthy cells stained as Ki67 positive. However, application of the IHC stain on normal breast tissue only results in a very low expression with only a < 3% positive rate [26]. As a result, the use of IHC stain Ki67 represents a reliable indicator for identifying active cancerous cells in a tissue sample and its prognostic value has been noted by multiple reviews and studies [4], [28], [29], [30].
To determine a PI, the number of stained nuclei in an image are to be determined. As mentioned, the PI is defined by the overall percentage of positively stained nuclei. Since the PI requires the identification of both healthy and cancerous nuclei, tissue samples are counter stained with hematoxylin which stains all nuclei structures blue. Determination of these positive and negative counts lack adequate standardization and are prone to observer variability at various expression levels limiting the clinical use of Ki67 in pathology labs for the purpose of breast cancer [31], [32]. To combat observer variability, the use of CAD algorithms for proliferation determination may be useful for future studies of the prognostic capabilities of Ki-67. A histopathology image of nuclei containing both Ki-67 and hematoxylin staining can be observed in Fig 2.2.

![Histopathology image of nuclei containing both Ki-67 and hematoxylin staining](image)

**Figure 2.2 - Tissue microarray core stained with Ki67 (brown) and hematoxylin (blue).**

### 2.1.3 Tumour Acquisition and Specimen Preparation

Once a tumour has been detected, a biopsy procedure is performed to obtain a tissue sample that must undergo preparation procedures before tumour features can be evaluated. Previously, the "gold standard" for retrieving breast tumour samples was to undergo a surgical biopsy procedure, however an increase in screening procedures has resulted in many more cases of breast tumours requiring diagnostic follow-up. As a result, less invasive needle core biopsies are becoming the most utilized method for
obtaining multiple tumour samples used in diagnosis and prognosis procedures [33]. The core needle is inserted into the target breast region, sometimes guided by ultrasound or other imaging techniques, where multiple samples are taken for analysis. An example of the procedure can be observed in Fig 2.3. Following sample acquisition, the specimen is stained with histopathological chemicals to highlight cell structures or in the case of Ki67, proliferative activity.

For longer term studies beyond an initial diagnosis, tumour samples must be preserved after removal from the patient since tissue ischemia, which is the degradation of DNA, RNA, and proteins, begins almost immediately [35]. The tissue samples can be preserved by freezing fresh samples in liquid nitrogen, immersion in special agents to stabilize nucleic acids, or fixation of the proteins to preserve tissue architecture [35]. Long term preservation is achieved through the creation of formalin-fixed paraffin-embedded (FFPE) tissue sections that allow for stable storage and can be used for testing and evaluation of different histopathology stains. FFPE samples are created by first preserving the tumour tissue in a diluted formaldehyde solution that causes cross-linking of proteins and preserves tissue structure. This process is referred to as formalin fixation and the amount of time it takes for fixation to occur depends on the degree of formalin penetration into the tissue which is on average
1mm per hour [36]. Therefore, fixation occurs more rapidly in smaller tumour samples which would limit protein degradation and ensure proper structure preservation. Once the sample has been formalin-fixed it can be embedded in paraffin wax which allows the samples to be thinly sliced for effective histochemical staining and observation. Since FFPE samples can be stored for long periods of time they are optimal for use in long term studies of cancer treatment effectiveness and for correlating patient survivability to past diagnosis. Unfortunately, tissue processing is not a universally standardized process and the type of fixative, processing times and temperatures, and types of paraffin vary considerably among institutions [37].

Although FFPE sections can be assembled over long periods of time to generate larger medical datasets, the aforementioned preparation variability results in less reliable specimens that inhibit the validity of stain evaluation studies. The creation of tissue microarray (TMA) cores have been presented as an effective way to simultaneously stain and evaluate many individual tumours from different tissue donors. TMAs are created by extracting 0.6-3mm core samples from different FFPE tumour samples and seating them into a recipient paraffin blocks in a positionally encoded format. The assembled TMA block can then be sliced into ~5µm sections where a single block can generate as much as 100-300 sections [38]. The implementation of TMAs ensure staining consistency across samples and reduces procedural variability in the staining process because all tumour samples are stained simultaneously following the same protocols. Unfortunately TMAs are not commonly used in clinical practice because they require careful construction and are not time appropriate for a functioning pathology lab. However, TMA cores are an effective tool for CAD algorithm evaluation because they include various levels of tumour grade, ensure stain consistency, and can show up to 1000 tumour samples within a single paraffin block [39]. An example of a TMA stained with Ki-67 and hematoxylin can be seen in Fig 2.4.
2.1.4 Scoring and Grading Systems

The following section outlines the application methods for manually scoring and grading histopathology images. The associated challenges with percentage estimations and observer variability are discussed and reproducibility is investigated.

Pathologists evaluate and characterise their observations by utilizing standardized scoring and grading systems which vary depending on the suspected pathology of the patient as well as the histopathological stain used on the tumour sample. The most common scoring system for breast cancer grading is the Nottingham scale which is a modification of the Bloom-Richardson scoring and grading system. The Nottingham technique determines a histologic grade through H&E staining and evaluation of three morphological features: the percentage of tubule formation, the degree of nuclear pleomorphism, and an accurate mitotic count using a defined cell area [25]. This modification was designed to reduce the subjectivity of pathology reports and has become the primary tool for pathologists grading breast cancer biopsies. Lookup tables for evaluating the morphological features described in the Nottingham method as well as the associated numerical scoring can be seen in Fig 2.5. The lookup tables rely on the skill and experience of a pathologist to evaluate the expression of these morphological features and results are usually consistent for high or low levels however observer variability is most often observed at moderate expression levels.
Figure 2.5 - Morphological features associated with the Nottingham grading system. The left table describes the features a pathologist would reference in percentages. The right table shows the breakdown of the scoring based on the number of mitotic counts observed in a defined tumour area and microscope magnification. The final scores for each feature would then be used to determine a final histologic grade of one to three for the tumour [25].

Similarly to the histologic grade, evaluation of Ki67 requires an estimation or calculation of stained features. Scoring methods for IHC stains like Ki67 are generally based on the percentage of tumour cells stained by the antibody. For example, a pathologist can examine a stained section using a light microscope at 40X magnification, using a 10X10 graticule, and the Ki67 score is defined as the percentage of tumour cells with positive nuclear staining [31]. This example would require individually counting about 1000 tumour cells which is impractical for clinical settings due to time restrictions and productivity requirements, so some pathologists may estimate the percentage of nuclei staining [31]. By estimating percentage of staining through a fast scanning technique, pathologists increase the subjectivity and variability in their reports but it may be unavoidable due to the number of samples they are required to evaluate daily. To increase concordance among pathologists for PI estimation, standardization protocols are continuously proposed and suggested. Despite these efforts manual PI methodologies remain inherently subjective, time consuming, and sensitive to inter and intra observer variability among pathologists for various stain concentrations [4], [40], [41]. In addition to observer variability, selection of appropriate cut-off thresholds for positive stain characterization remain uncertain and must be standardized before Ki67 can be introduced in common clinical practice [32]. Cut-off thresholds evaluating predictive and prognostic potential of Ki67 are variable, however it is most commonly separated into low (<10%), intermediate (11-30%), and high (>30%) levels of proliferation activity [29].

Even with the difficulties associated with scoring proliferation using Ki67, it was still found to serve as a good prognostic indicator of disease free survival rates as well as having an association with baseline and immediate response hormone therapy or chemotherapy [29].
Despite the strict guidelines and methodologies for scoring tumour characteristics there are still cases of diagnostic uncertainty which historically have been settled by obtaining additional opinions of other pathologists and selecting the most agreed upon diagnosis [42]. While this method may slightly improve overall diagnosis results, it still relies on additional subjective observations. Beyond the second opinions of a pathologist's colleagues, it was not practical to standardize a single protocol over all laboratories because of different methods of specimen procurement [42]. More recently standardized procedures in the form of training protocols have been implemented in an attempt to ensure consistency across laboratories, however studies comparing diagnoses still result in questionable agreement [43]. A review of inter and intra-observer variability for atypical ductal hyperplasia (borderline benign/malignant breast cancer) found only 11% of cases had complete agreement among nine pathologists, and 47% of cases had seven or more in agreement [43]. Another studied the variability of the Gleason scoring and grading system for male prostate cancer cases. The inter and intra-observer reproducibility was high for classic Gleason patterns however the decision boundary between a Gleason pattern of 3 and 4 resulted in only fair reproducibility [44]. This is significant result because the difference between these Gleason patterns determines patient treatment, so with only fair reproducibility some patients may receive ineffective and potentially harmful treatment options. In a Ki67 reproducibility study [45] 100 breast cancer TMAs were evaluated by eight separate laboratories. While intralaboratory reproducibility was high, interlaboratory reproducibility was only moderate. Interlaboratory variability also increased when the institutions performed the Ki67 staining on site opposed to evaluation of previously prepared and stained cores. Some of the institutional variability was attributed to different staining protocols, selection of sample hot spot locations for counting, and use of discrete percentage estimates when determining a PI.

2.1.5 Digital Image Generation

Digital pathology images were first used as a sharing tool for pathologists and were obtained by mounting digital cameras on optical microscopes. Current methods for digitizing, storing, and evaluating histopathology images utilize WSI scanners which can be coupled with viewers to simulate slide viewing by a conventional microscope [46]. These scanners generate high resolution colour images by inputting a glass slide containing a stained tumour sample and produce whole slide digital images as an output. This automates all intermediate steps such as localization of tissue and focus plane segmentation [47]. WSI scanners introduce the possible acquisition of large medical datasets that can anonymously store patient data utilized in research studies and thus improve overall patient care and outcomes. The potential for large medical datasets provide large promise to the industry of computer aided diagnosis (CAD) of histopathology images because medical procedures can be optimized to ensure a patient is receiving appropriate treatment or lead to the early detection of pathologies through study and clinical trials [48]. In
addition to developing treatment predictions for patients, whole slide imaging can also serve as an important teaching tool for training pathology students. The digital images can be also be readily stored, shared, and be discussed in a classroom or laboratory setting.

Despite the potential of WSI scanners, there are still challenges that must be addressed before their implementation in clinical workflows becomes the new standard. Generated images from WSI scanners tend to have large memory requirements ranging from 200MB to 5GB for a single image file [49]. For large scale implementation of WSI scanners, infrastructure would have to be created to handle large scale storage requirements of patient images. If data storage structures were utilized with WSI scanners via internet connections they would require high speed connections and the security of patient data would need to be considered. Long term storage of digital images often includes "lossy" compression techniques that sacrifice image data to obtain smaller memory requirements. If compression is utilized it may reduce resolution and make it difficult to confirm past diagnoses or link results to survival statistics. Even with lossy compression techniques images are around 350MB and for a medium sized pathology laboratory this may result in requiring 500GB a day in image storage [46]. Memory solutions have been proposed to combat the memory requirement issues such as a tiered memory system where images are stored in a short term bank and after a period of time are compressed and sent to a long term memory bank. Methods such as this would allow for the archiving of large images and still retain recent information for quick retrieval and sharing.

In radiology, standardization of magnetic resonance imaging (MRI) machines has been achieved through the implementation of Digital Imaging and Communications in Medicine (DICOM) imaging procedures. This dictates many requirements of image generation procedures such as file format, documentation procedures for patient data, scanner used as well as characteristics, and storing procedures. This sort of standardization is currently not implemented for WSI scanners due to their recent development so there are many competing vendors each with their own unique imaging formats and software packages [49]. This assumed as a temporary challenge because once WSI scanners become more adopted as a laboratory standard, regulations will be required by governing bodies of medical labs to facilitate data sharing and storage procedures.

2.1.6 Image Quality Challenges

There are a number of image artifacts that can affect output image quality, many of which can be associated with occurrences during slide preparation. WSI scanners produce 2D high resolution images from a very thin sample that has three dimensions, so thicker samples may result in blurred nuclei regions. This artifact occurs because the focal plane on the scanner can auto focus on nuclei near the surface of the sample and nuclei near the bottom are imaged out of focus. This can be overcome by multi-
plane scanning along multiple z axes (depth) and reconstruction of the image [12]. Multi-plane scanning however increases the time to scan a single slide and also creates very large image that require more computer memory to store. Multi-plane scanning improves image quality, however it also increases the time it takes to scan the image. Therefore thinner tissue sections, about three micrometers, require less time to generate a higher quality image than slices seven micrometers thick. Thinner sections are associated with better image quality and the reduced number of focal points required improved canning time by 10-15% [50].

Other image artifacts associated with image quality include water or bubbles on the slides, folding of TMA cores, staining variability, or even pen marks used when processing and tracking TMA cores [11]. Water droplets and bubbles throw the image scanner out of focus because the regions within the bubble or water droplet scatter light and result in blurred regions where nuclei cannot be quantified correctly or accurately. Pen marks can also affect scanners focus, because they may be on top of the glass slide and if only a single z-plane was used during the scanning process the nuclei would appear blurred and out of focus. Folding of the TMA cores can also occur during slide preparation and result in overlapping and unclear nuclear boundaries. Sometimes a loss of a TMA core can occur if the sample undergoes necrosis and as a result sinks in the paraffin block and is not included in the slide and cannot be imaged. TMA cores that are lost are typically not included in pathologist analysis or image quantification through CAD systems. Staining procedures are designed to be as standard as possible however huge variability exists whether it is a result of older samples or staining amounts used. Samples can appear very light and translucent or very dark and contain a lot of stain bleeding into other tissues. Examples of some of these artifacts can be seen in Fig 2.6. These artifacts are all a result of challenges with slide preparation and illustrate the idea of "garbage in garbage out" processing procedures.
2.2 - Colour Spaces for Colour Separation

In the following section, colour spaces available to image analysis algorithms for automatic histopathology analysis will be described as well as the challenges associated with their implementation for Ki67 and hematoxylin stained images.
2.2.1 The Red, Green, Blue (RGB) Colour Space

The RGB colour space is the most commonly used colour space used in image capture technologies and display devices. It follows an additive colour model that consists of three colour channels where each represent the red, green, and blue primary colours. The gamut is the range of potential colours that can be observed and each channel is typically limited to the integer range \([0, 255]\) where zero corresponds to no content and 255 is the maximum colour value content. Since the RGB space is a mixture of the three channels there are up to \(255^3\) possible colour combinations that represent the gamut of potential observable colours. Pixel values are obtained by including the three colour channels and colours vary greatly if even one channel input is manipulated. For example an RGB colour vector value of \([0, 0, 0]\) corresponds to black and \([255, 255, 255]\) is visualized as white. Similarly a value of \([100, 100, 100]\) represents a medium intensity gray level because each colour channel is present at identical intensities. For the purpose of colour conversions or computational requirements, the RGB colour space can also be normalized to the range \([0, 1]\) and is functionally the same as the \([0, 255]\) range. Colour content is observed when there are different intensity values from each channel and because it is represented as an additive colour model where the higher the intensity value from a channel, the more it will dominate the overall observed colour.

Since histopathology images are generated by a high resolution scanner, colours are recorded in the RGB format. For image processing, individual channels in an RGB image can be utilized for specific purposes, such as the blue channel which can be used to segment hematoxylin nuclei from Ki67 stained nuclei. However even though a stain like hematoxylin is predominantly blue and can be associated with the blue channel, other non-blue stains may still contain significant blue intensities even though they do not appear through visual observation. In Fig 2.7 the blue colour channel is shown for a Ki67 and hematoxylin stained image where significant blue content is observed in lighter stained Ki67 nuclei. This property makes it difficult to identify staining content using only one colour channel that does not appropriately differentiate colour content.

![Figure 2.7 - RGB Ki67 and hematoxylin stained image and associated RGB colour channels.](image-url)
2.2.2 The Hue, Saturation, Value (HSV) Colour Space

Humans have difficulty perceptually understanding the outcome of mixed primary tristimulus values like the RGB space, so describing perceived colours with specific mixtures of RGB values is difficult. Since humans perceive colour with more subjective descriptors such as luminosity (brightness), hue (chrominance), and saturation of colour, multiple colour spaces have been proposed that quantify colour using these characteristics [51]. The HSV colour space is one of the more popular perceptual models and describes colour using three channels. The hue channel describes chromal content over a 360° range where 0° corresponds to red, 120° is green, 240° is blue, and 360° wraps back around to red. The saturation and value channels lie on the range [0, 1] and represent the relative strength and brightness of a specific colour defined by its hue. The HSV space is be obtained by applying transformation equations to the RGB colour space. The associated equations are seen below in equations (2.1) - (2.6).

\[
hue = \begin{cases} 
0^\circ & \text{if } C = 0 \\
60^\circ \times \frac{G-B}{C} \mod(6), & \text{if } M = R \\
60^\circ \times \frac{B-R}{C} + 2 & \text{if } M = G \\
60^\circ \times \frac{R-G}{C} + 4 & \text{if } M = B 
\end{cases}
\]

(2.1)

\[
saturation = \begin{cases} 
0 & \text{if } M = 0 \\
\frac{C}{M} & \text{if } M \neq 0 
\end{cases}
\]

(2.2)

\[
Value = M
\]

(2.3)

where,

\[
M = \max(R, G, B), \\
m = \min(R, G, B), \\
C = M - m
\]

(2.4-2.6)

The R, G, B values are normalized to the range [0, 1] and represent each of the three RGB colour channels.

Similar to the RGB colour space, the three channels are used together to describe the perceived colour at a single pixel in the format [H, S, V]. As mentioned, saturation represents the "strength" of a
colour and luminance is related to the brightness of a colour, meaning maximum saturation values correspond to "pure" colours, while a zero level luminance results in black. For example, with a hue value of 0° a colour with maximum saturation would be perceived as a "pure" red while a colour with 0° hue and 0 saturation will appear white even though the hue is associated with red. Manipulating the value channel using the above example, an HSV value of [0,1,0] is black and [0,1,1] is "pure" bright red. These properties can be visualized in Fig 2.8, which illustrates a theoretical shape of the colour model.

![HSV colour space mapped to a cylinder](image_url)

**Figure 2.8 - The HSV colour space mapped to a cylinder** [52].

The more perceptual characteristics of the HSV channel allow for better understanding of histopathology stain colour representations however it still poses challenges for image analysis algorithms. The hue channel is effective at quantifying stains with large differences in colour content, however since it is defined on a circular range it is difficult to determine appropriate regions for stain colour distributions. Therefore to apply a threshold methodology for stain separation two thresholds must be selected. The first is used to shift the hue channel to define the colour range of interest and then another must be applied to effectively separate stains. Since there are no expected or ideal hue colours representing Ki67 or hematoxylin the selection of two thresholds is difficult for images containing significant variability.

### 2.2.3 The Luminance, a Chrominance, b Chrominance (L*a*b*) Colour Space

The L*a*b* colour space was designed to model the human visual system and its ability to perceive and differentiate colours. The L* channel represents luminance which is the same as the value channel in the HSV colour space and also lies on the range [0, 1]. The a* and b* channels are used to describe and differentiate chromal content where a* differentiates red and green, while b* differentiates...
yellow and blue. Unlike the RGB channel, which uses mixtures of primary colours, the a* and b* channels describe colour using opposites. This takes advantage of a perceptual concept that a colour could not be described as "reddish green" or "yellowish blue". The exact scaling and expected ranges for the a* and b* colour channels are not clearly defined since they possess a larger colour gamut than the RGB model, but they are usually on the range [-128, 128] where -a* is green, +a* is red, -b* is blue, and +b* is yellow. Furthermore, a value of 0 in both the a* and b* channels result in a grayscale intensity depending on the L* channel where L* = 1 is white and L* = 0 is black.

To acquire a L*a*b* image, the recorded RGB image must first be converted into a XYZ colour model where the L*a*b* colour space can then be derived from the XYZ tristimulus values through non-linear relations defined by equations (2.7) - (2.10) [51].

\[
L^* = \begin{cases} 
116 \times \sqrt[3]{\frac{Y}{Y_W}} - 16 & \text{if } \frac{Y}{Y_W} > 0.008856 \\
903.3 \times \frac{Y}{Y_W} & \text{if } \frac{Y}{Y_W} \leq 0.008856
\end{cases}
\] (2.7)

\[
a^* = 500 \times \left( f\left(\frac{X}{X_W}\right) - f\left(\frac{Y}{Y_W}\right) \right)
\] (2.8)

\[b^* = 200 \times \left( f\left(\frac{Y}{Y_W}\right) - f\left(\frac{Z}{Z_W}\right) \right)
\] (2.9)

where,

\[f(x) = \begin{cases} 
\sqrt[3]{x} & \text{if } x > 0.008856 \\
7.787x + \frac{16}{116} & \text{if } x \leq 0.008856
\end{cases}
\] (2.10)

Where \(X_W, Y_W, Z_W\) are the tristimulus values of a reference white point and the non-linear response to colour perception is modeled by a cubic root relation. The L*a*b* space was developed to model the HVS perception of colour where the difference between two L*a*b* values coincide with a Euclidean "just noticeable difference" in colour content by a human observer. As a result the non-linear mapping of the colour space generates a perceptually linear composition.

For Ki67 analysis, the L*a*b* space is advantageous for colour processing and stain separation because stain identification by a pathologist relies heavily on colour content in addition to morphological features. Therefore, since the L*a*b* space is modelled after the human visual system it should be able to model stain content similarly to a pathologist and could achieve more consistent results than similar
methods using one of the other colour spaces. Additionally, an automatic Ki67 quantification algorithm proposed in [21] examined the implementation a clustering framework using multiple colour spaces. It was found that the results obtained using the L*a*b* colour space obtained the best results compared to manually labelled images including PI estimation. Further use of the L*a*b* channel characteristics for automatic Ki67 analysis are described in Chapter 3.

2.3 - Existing CAD Algorithms and Challenges

The multiple sources of variability in histopathology analysis pose great challenges to automatic analysis algorithms for stain separation and Ki67 quantification. Since staining variability is common, proposed algorithms must be robust to variations in staining content. Following the general image analysis framework, pre-processing procedures are implemented in an attempt to reduce image variability before stain separation, nuclei detection, and PI estimation procedures. The following sections examine proposed methodologies associated with each step in the image analysis framework.

2.3.1 - Pre Processing

Pre-processing methods are often used to standardize images that are variable in appearance or contain image artifacts. In histopathology analysis there are a number of proposed methods for image normalization and are often focussed on improving nuclei representations and normalizing colour characteristics. Due to the variability of histopathology images as a result of staining procedures and slide preparation, pre-processing is often a requirement before nuclei segmentation methods are applied.

Simple pre-processing methods include the use of digital filters, which are composed of a small kernel or mask that is applied to an image for a desired smoothing effect. Application of a filtering kernel to a grayscale image is relatively simple since only gray level intensities within a pixel neighbourhood are considered. Since colour images are represented with three dimensions, where each pixel contains 3 colour intensity values, standard grayscale filtering techniques cannot be applied directly to the colour image without some modification. The simplest way to apply a grayscale filter, such as a moving average filter, on a colour image is to independently filter each of the colour channels. This allows for smoothing of the colour image, however since each colour channel is represented by different intensities this smoothing may not have the intended effect. For example if a smoothing filter was applied to an RGB image by independently filtering each channel then the resultant RGB vectors may be shifted from the original proportions. It is possible that an RGB vector [100, 102, 95] becomes something similar to [102, 97, 99] which would change the colour proportions of the observed pixel. This is non-desirable effect that may result in pixel misclassifications when performing a stain separation method that relies on consistent colour characteristics.
There are filtering methodologies that consider colour content by considering the 3 colour channels at a pixel location as a vector. One of these filtering techniques is the vector median filter (VMF). The VMF filter calculates a mean colour vector for each image pixel using the surrounding pixels in a pre-defined neighbourhood. A small neighbourhood results in more subtle smoothing effects and a large neighbourhood obtains greater effects because it includes more pixels to contribute to an overall mean. The Euclidean distance between each pixel vector and the mean vector are measured and the results are ordered in ascending or descending order where a defined number of samples are removed. The final result of the filtering process is the mean colour vector obtained by averaging the remaining ordered vectors. A VMF is useful for histopathology images because it is able to smooth the inconsistencies associated with cancerous cell morphology and noise characteristics that can occur during image generation. An example of the smoothing effects achieved using a VMF can be seen for one nuclei stained with hematoxylin in Fig 2.9. It can be noted that by increasing the pixel neighbourhood greater smoothing effects are achieved, however by using a 5x5 neighbourhood nuclei edge content is also smoothed. If a VMF uses a larger window size areas of dense nuclei clusters can be mixed together posing challenges for nuclei detection algorithms.

![VMF filtering applied to a single hematoxylin stained nuclei. The 3x3 and 5x5 values are referring to the pixel neighbourhood used in each implementation.](image)

There are a number of other colour image conditioning methods for image normalization however an in depth analysis is beyond the scope of this thesis, so an extensive review of pre-processing methods will not be performed. It is sufficient to note the different colour representations and understand that the main purpose of pre-processing is to standardize the appearance of the histopathology images for more accurate results.

### 2.3.2 - Stain Separation

Stain separation may occur before or after nuclei segmentation and for Ki67 it is used to separated the nuclei into positive and negative classes. Pre-processing methods are implemented to
minimize the staining variability and can greatly improve performance, however there is still a wide range of possible staining intensities that can be observed image to image. For H&E images, identification of hematoxylin can be incorporated into a nuclei segmentation scheme because all the nuclei will be identifiable with hematoxylin. This is not the case for images with Ki67 and the increased variability makes stain separation and nuclei identification more difficult for CAD algorithms.

Colour deconvolution (CD) proposed in [15] has dominated as a method for stain separation and nuclei segmentation methodologies due to its effectiveness at separating hematoxylin stained objects (nuclei) from the eosin counterpart (stroma) based on their approximate stain concentrations. It is also applicable to Ki67 analysis and has been implemented for stain quantification and PI estimation [16], [17], [18]. CD is dependent on the Beer-Lambert (BL) law of absorption which characterizes each pure stain in an image by a specific optical density (OD) vector of light in the red, green, and blue (RGB) intensity channels [15]. The OD of each channel is expressed by the equation:

\[
OD_c = -\log_{10}\left(\frac{I_c}{I_{0,c}}\right) = A \cdot c_c
\]

(2.11)

Where, \(OD\) is the optical density of a detection channel \(c\), \(I_c\) is the intensity of light detected after passing the specimen, \(I_{0,c}\) is the intensity of light entering the specimen, and \(A\) is the amount of stain with absorption factor \(c\). Variations in staining intensity as perceived in the RGB channels are non-linear due to different absorption characteristics of specific stains [53], however the OD conversion is performed to linearize this relationship. These vectors are used to unwrap the stain concentrations in H&E stained images on a per pixel basis. As mentioned, CD has been applied for IHC analysis with promising results, however high concentrations of Ki67 staining exhibit non-linear properties as a result of light scattering characteristics [23]. This is in contrast to H&E stains which can be modeled considering only light absorption characteristics. Since DAB (Diaminobenzidine) stains, like Ki67, are not true absorbers of light they do not follow the BL law [54] and cannot be accurately represented in the OD colour space.

To visualize the problematic effects associated with the conversion of RGB images containing high Ki67 concentrations three images containing low, medium, and high stain concentrations were converted to OD space and visualized with scatter plots. The results of the visual experiment are observed in Fig 2.10 which demonstrates the effects of the light scattering properties that manifest at higher stain concentrations in the OD colour space. At low and intermediate stain levels Ki67 is linearly separable in the OD space, however for an image with higher concentrations the light scattering effects result in inaccurate representation and a non-linear decision boundary. To further visualize this relationship the OD space was projected onto a plane of best fit through principle component analysis where the quantization error is more pronounced.
Figure 2.10 - (Left) Sample images of low, intermediate, and high Ki67 stain concentrations after vector median filtering. (Center) RGB optical density space for respective stain levels. The high Ki67 image appears to contain quantization error at high stain levels due to light scattering and absorption characteristics of Ki67. (Right) Projected plane obtained via principal component analysis on the RGB OD colour space. High Ki67 concentrations result in non-linear properties, while the low and intermediate stain levels remain linearly separable.

Further implementation of CD also requires the selection of pure stain vectors to generate an OD deconvolution matrix. This matrix represents the RGB channels for each stain type present in the image, where each row in the matrix corresponds to a specific stain. The deconvolution matrix is then used to separate the stains into density images that represent the absorption characteristics of each stain. There are various methods for selection or calculation of appropriate stain vectors including application of previously measured values, manual measurement, or automatic calculation of stain vectors by analyzing stain content. Pre-measured values should in theory be suitable for image analysis, however implementation using these values can lead to variable or sub-optimal results. A Ki67 quantification algorithm in [18] implemented CD with predefined stain vectors for DAB and hematoxylin stain separation, but found the resultant complimentary image contained stain content from both DAB and hematoxylin. Appropriate stain vectors were then determined though manual measurements and testing of parameter selections. Other methods [55], [16] have recognized CD results vary for different staining and image protocols and have independently adjusted the originally proposed parameters to values better...
suited to their datasets. This suggests pre-measured stain vectors do not generalize well and are sensitive to the wide variability common in histopathology images. To remove subjectivity introduced by manual selection of stain vectors, automatically determined vectors are preferred. Automatic stain vector estimation was proposed in [56], where H&E stains were separated in OD space by projecting pixels onto a plane of best fit through singular value decomposition. While this method is effective for H&E analysis, high levels of Ki67 do not transfer well onto a plane of best fit due to the light scattering effects mentioned previously and shown in Fig 2.10. Results of Ki67 and hematoxylin separation using pre-measured stain vectors in CD were obtained using ImageJ and are observed in Fig 2.11 where the nuclei with dense Ki67 intensities are not fully represented in the Ki67 channel image due to quantization error within the OD colour space.

![Original Image](image1.png) ![Hematoxylin Channel](image2.png) ![Ki67 Channel](image3.png)

Figure 2.11 - Stain separation achieved through colour deconvolution with associated quantization errors.

Alternative methods for IHC stain separation have been proposed using iterative computational approaches with promising results. A method for automatic IHC analysis was proposed in [19] where a support vector machine was used to separate Ki67 and hematoxylin stain content in addition to final nuclei locations and classification. In [20] an iterative dictionary learning approach was used for classification of tissue into probability maps of each stain type. The dictionary learning algorithm automatically updated an image dictionary from the training data which was then used for final implementation on the test images. One major concern with these implementations of iterative learning methods is that they can have large computational requirements and require partitioning of the dataset into training and testing data. Learned models that rely on training data are susceptible to over training on specific datasets and may not generalize well to images containing wide stain variability common in histopathology images. The use of training data can be feasible if there is a large number of images available in a dataset representing all possible stain concentrations and image variability. Unfortunately, datasets are often limited in their size due to difficulties associated with obtaining medical image data.
Obtaining samples for histopathology analysis research is often required to follow extensive ethics standardization procedures in addition to obtaining specific patient permissions for use of their tissue samples. These challenges often limit the size of image datasets which cannot fully encapsulate the range of staining possibilities resulting in under trained models. Both of the above learning methods also did not consider alternative colour space representations which could provide a more expansive feature space for classification.

Another popular iterative method for stain separation is the k-means, also referred to as c-means, clustering algorithm. Simple implementation on a grayscale image is achieved by minimizing the cost function defined in equation (2.12).

$$J_s = \sum_{i=1}^{c} \sum_{l=1}^{q} u_{il}^m \cdot v_i^2$$  \hspace{1cm} (2.12)

Where $v_i$ is the cluster location of the $i$th cluster, $u_{il}$ represents the membership of gray value $l$ with respect to cluster $i$. The $q$ denotes the number of gray levels in the image defined as 256 for a grayscale 8 bit image, and $c$ is the number of clusters in the image often representing the background, and stain types. The membership matrix for each cluster as well as the updated cluster location are iteratively updated using the equations:

$$u_{il} = \frac{|x_l - v_i|^{2\frac{m-1}{m}}}{\sum_{j=1}^{c} |x_l - v_i|^{2\frac{m-1}{m}}}$$  \hspace{1cm} (2.13)

$$v_i = \frac{\sum_{l=1}^{q} u_{il}^m \cdot x_l}{\sum_{l=1}^{q} u_{il}^m}$$  \hspace{1cm} (2.14)

Where $m$ is a weighting exponent on each fuzzy membership, $u_{il}$ is the fuzzy membership of gray level $l$ to the cluster $i$, $x_l$ represents the pixels at gray level intensity $l$, and $v_i$ is the intensity location of cluster $i$. Equations 13 and 14 are iterated until some stopping criteria is met or if an iteration limit is defined by a user.

The above k-means equations only consider gray level intensities for pixel clustering, which means that for colour implementation the k-means algorithm must consider more dimensions. Colour based application of the k-means algorithm for stain separation using the blue and hue colour channels was proposed in [57] for 20 stained images. As described previously, for Ki67 images the blue channel does not effectively differentiate the stain content as brown also contains significant amounts of blue intensity. Also since the hue channel is also a circular representation of colour it must be shifted for consistent differentiation between colours. Despite these challenges the algorithm was able to achieve
comparable results to another Ki67 quantification algorithm. In [21], multiple colour spaces were considered for stain separation by k-mean clustering, however only five images were considered for determination of the optimal colour space which limits the validity of their colour space selection. While both methods were able to incorporate the use of colour content, clustering methods often rely on manual initialization estimates and contain large processing requirements for histopathology images containing up to billions of pixels.

2.3.3 - Nuclei Segmentation

To obtain a robust PI estimation a nuclei detection algorithm must be implemented. Many nuclei segmentation techniques focus on gray-level information that has difficulty identifying overlapping nuclei, dense cell clusters, and stain variability. An example of these challenges are observed in Fig 2.12. In the following section, a few dominant approaches to nuclei segmentation are explored with advantages and disadvantages of proposed methods. These approaches are pixel based segmentation, active contour modelling, and watershed segmentations. A performance summary of the three methodologies can be observed in Table 2.1 at the end of the current section.

![Image](image_url)

Figure 2.12 - (Left) Dense cell clusters make it difficult to differentiate nuclear boundaries and some of the nuclei are overlapping. (Right) Light staining has resulted in membrane staining as well as a darker nucleoli which can result in poor nuclei segmentation.

Pixel based segmentation techniques often focus on the intensity value of each pixel where specific intensities are assigned to a specific class such as nuclei, background, and stromal content. An example of a pixel based segmentation technique already discussed is the c-means clustering algorithm which can be modified to generate a fuzzy membership which describes how similar a particular pixel intensity is to a certain cluster. The term "fuzzy" refers to the likelihood of a pixel belonging to a certain
cluster. As a result, it is expected that a pixel within a nuclei would have a higher membership value to the nuclei cluster than the other clusters so when segmentation is performed the algorithm will retain the pixels with the highest probability of being nuclei. Difficulties for this method arise when considering an image with light staining as in Fig 2.12 where the white background content within the nuclear membrane would result in segmentation "holes" as the more white pixels would be more similar to background content than nuclei content. Variations of this method have been proposed in [58], [59] which both discuss improving performance of the original method by incorporating information from neighbouring pixel values to better represent border regions where the pixels change from one object to another. Including neighbourhood information could help in defining nuclei boundaries and improve segmentation performance, however it may not completely eliminate the sensitivity to lightly stained nuclei and densely clustered nuclei. Another issue with the fuzzy c-means method that [59] improves upon is processing time which is often overlooked in nuclei segmentation where a single image contains thousands of nuclei. It is important that the algorithm performs segmentation in a timely manner because pathology labs process many slides in a single day and an excessively slow CAD algorithm can hinder instead of assist in the workflow. Some other pixel based methods exist with similar challenges as those mentioned above. These include probabilistic methods such as a Bayesian classifier for the purpose of overlapping nuclei separation in [60] or Otsu's segmentation method [61] performed with adaptive thresholding techniques in [62]. These pixel-based methods obtain nuclei segmentations by assigning pixels to a cluster or classification which is not ideal for obtaining specific nuclei counts. If a densely packed nuclei cluster is segmented as one object all of the nuclei within the cluster cannot be counted without additional processing. Similarly, misclassified pixels may be mistaken for a nuclei which would skew the overall counts. Pixel based segmentation methods also tend to be susceptible to image inconsistencies such as intensity variability and overlapping nuclei which is common for histochemically stained images.

Watershed segmentation methods are widely implemented in proposed methods at various points along the image analysis framework. They can be used to boost performance of segmentation schemes or implemented as a standalone nuclei segmentation algorithm. Watershed models segment objects by comparing and thresholding intensity values on a grayscale image. In a grayscale image, pixel intensity values can be thought of as the depth dimension where a value of zero 0 is the bottom of a water basin and a value of 255 has the maximum height of the basin. Cellular nuclei would have a relatively consistent intensity value so the watershed algorithm is able to differentiate those intensities from the background or the rest of the image. It is a very good method for segmenting cell nuclei in clusters because it can recognize these localized maxima and minima regions representing the nuclei and the background content in between. The watershed segmentation method often leads to over and under segmentation because the intensities can vary depending on factors such as the histopathological stain
variability or stain mixing. To combat this problem, methods have been proposed to include marker information in the watershed algorithm so that it can include information about the locations of the nuclei. Markers can be manually counted by a pathologist or computed using other techniques that can identify nuclei locations as in [63], [64]. Adding the additional processing step results in better segmentation accuracy and reduces the amount of over segmentation common to the original method. While watershed and marker controlled watershed methods perform well for dense cell clusters, they do not consider overlapping nuclei. Since the watershed method is particularly good at the particular task of separating cell nuclei it is used in many algorithm techniques either as an initialization method or in subsequent intermediate steps in the image analysis workflow, but is not suitable for the primary nuclei segmentation method without further processing.

Active contour (AC) modelling focuses on the shape of individual nuclei instead of evaluating each pixel intensity independently. Sometimes these methods are performed as an additional step after a basic segmentation method in order to identify and separate overlapping nuclei or dense cell clusters. A basic example of this method is region growing which functions by randomly or actively selecting a point in the image and then incorporating neighbouring pixels of similar intensities. When the growing region reaches a significantly different intensity value it identifies an edge and eventually an image mask is created with a positive region and a negative region which can be applied to the original image to segment one region or the other. More complex methods exist that involve tracing boundaries of specific nuclei to solve the issue of overlapping nuclei as in [65] which utilizes a watershed segmentation method to perform a rough segmentation before tracing boundaries. Other methods involve fitting a shape prior to segmented nuclei to identify boundaries and separate clusters or overlapping nuclei [66]. Shape priors can be understood as a generic shape such as an ellipse that is considered the average shape of nuclei in a histopathology image. These priors can then be mapped to nuclei since they should be the only objects with an elliptical shape. Malignant nuclei however can often be inflated and deformed so that they are no longer tightly bound circular nuclei within a membrane. Additionally some nuclei may be stretched or compressed and not appear circular causing an active contour algorithm to ignore these. Shape priors may also need to be defined by the user, in this case a pathologist, which is problematic because CAD systems should be completely automatic to speed up the pathology workflow and remain objective. Another issue that may arise using shape priors is that an algorithm can become biased to shapes associated with the priors which may not obtain consistent results if not evaluated on a large dataset. While active contour methods are very good at segmenting overlapping nuclei and cell clusters, they often require or rely on an initial segmentation using other methods [67], [65], [68]. The problem with implementation of other methods is that active contouring rarely exists as a standalone nuclei segmentation scheme.
Each of the above methods are based on grayscale content and have their own advantages and disadvantages. It can then be questioned as to why an algorithm cannot be proposed that utilizes all of the above methods to overcome the normal challenges with nuclei segmentation. A combined method was proposed in [69] which consisted of initial segmentation by spatial FCM and thresholding, fuzzy contour mapping to an optimal boundary, and the determination of concavity points followed by an enhanced watershed method leading to final segmentation results. While this method may achieve good segmentation accuracy, problems such as computing time and computational efficiency are ignored.

Table 2.1 - Overall performance trade-off of common nuclei segmentation methods.

<table>
<thead>
<tr>
<th></th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pixel-Based Methods</strong></td>
<td>Suitable for images with well defined nuclei.</td>
<td>Susceptible to image variability and dense nuclei clusters.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Does not identify overlapping nuclei.</td>
</tr>
<tr>
<td><strong>Active Contour Modelling</strong></td>
<td>Accurate methods require shape priors which introduce bias.</td>
<td>Usually requires initialization segmentation.</td>
</tr>
<tr>
<td></td>
<td>Good at identifying overlapping nuclei</td>
<td>Can be computationally expensive</td>
</tr>
<tr>
<td><strong>Watershed Methods</strong></td>
<td>Good at clustered object separation</td>
<td>Basic method often results in over segmentation.</td>
</tr>
<tr>
<td></td>
<td>Marker controlled methods improve over segmentation.</td>
<td>Marker methods may include input requirements.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensitive to noise.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Does not consider overlapping nuclei.</td>
</tr>
</tbody>
</table>
Chapter 3

Proposed Framework

Prognosis estimation for breast cancer samples stained with Ki67 rely on accurate and repeatable PI determinations to predict patient survival and response to treatment. Determination of a PI through manual counting techniques are laborious and require copious amounts of pathologist time which is not realistic for routine clinical implementation. Conversely, manual fast scanning methods are inaccurate and do not provide a reliable metric that can be utilized for modelling disease progression over time. As a whole, manual methods are inherently prone to some level of inter-observer and intra-observer variability, particularly for intermediate stain levels, and can cast doubt on the use of therapeutic guidelines based on subjective Ki67 scoring [41].

To obtain more objective PI estimates, Ki67 quantification algorithms may be used, which have typically been implemented using various methods of stain separation and nuclei identification. Stain separation can be performed before or after nuclei segmentation, however it is the defining step in an algorithm that separates Ki67 content from hematoxylin content and the resultant performance greatly impacts PI estimation. Automatic stain separation is often achieved using colour deconvolution [15] or colour space clustering methods [57], [21]. Implementation of CD for Ki67 stain separation was performed in [55], [17], [18], and achieved good results, however to obtain these results the proposed methods had to modify the CD framework to achieve better stain quantification. Recall that the CD framework relies on either standardized or individually determined stain deconvolution vectors that model the staining content for hematoxylin and IHC stains. Due to the wide range of staining variability between images, standardized colour vectors are not appropriate for clinical datasets where considerations must be made for sub optimally stained images in order to acquire PI results. In [18], standardized deconvolution vectors were initially implemented and led to stain separation challenges. To account for the variability in the image dataset, the authors manually selected more appropriate stain vectors through parameter optimization and result comparisons. Similarly, CD based methods proposed in [55] and [16] modified the pre-recorded deconvolution vectors to better suit the images in their dataset. Modifications to the standardized stain vectors are usually performed by manually measuring stain vectors at pixel locations where only one stain is present, or by independently staining separate samples with only one stain and recording the resultant "pure" colour vector. Either way, manual selection of appropriate stain vectors can be used to increase algorithm performance on prepared datasets but this is unacceptable for routine
clinical practice as these manual modifications rely on the experience and subjective opinion of the user, leading to inconsistent results that are likely sub optimal.

In addition the problems associated with determining robust deconvolution vectors, recall that CD methods rely on the Beer-Lambert law of absorption [54], as described in section 2.3.2, which does not hold for darker IHC stain concentrations as they contain light-scattering characteristics [23]. As previously described in Chapter 2, the non-linear characteristics associated with darkly stained Ki67 content results in quantization errors when the RGB colour vectors are converted into the optical density space. As a result, colour deconvolution is acceptable for images containing low to intermediate staining levels, but images containing higher IHC concentrations result in segmentation "holes" within nuclear boundaries, this effect is shown in Fig 2.11.

Other methods of stain separation include more iterative or learning approaches which often do not consider the merit of utilizing alternative colour spaces for stain separation. Learned models often rely on the ability of an algorithm to iteratively learn morphological or colour patterns for stained images across a dataset. In these cases, the dataset images are usually input as the standard RGB format and alternative colour representations, such as the HSV space and the L*a*b* spaces, may be more appropriate. A more recent clustering method in [21], determined that a k-means clustering technique in the L*a*b* space obtained the best segmentation results compared to other colour spaces. The possible L*a*b* colour characteristics that led to this result was not explored. Therefore it is unknown if this method of clustering in the L*a*b* space would be consistent for images outside of the evaluated dataset. A method proposed in [57], utilized the blue and hue colour spaces for clustering analysis but did not appear to consider the circular characteristics of the hue colour space or suggest a shifting of values to normalize the hue channel. Since a specific hue value coincides with a degree indicator that lies on a circular range, it becomes difficult to isolate an acceptable positive or negative range of colour characteristics that can be correlate to a specific stain type. These challenges arise because the RGB colour space represents colours with different R, G, and B proportions where large discrepancies can exist due to relative intensity levels.

Despite the fact that all automatic quantification methods are validated on manually identified data, existing methods seem to ignore the identification process performed by a pathologist when labelling individual nuclei. While CAD algorithms are more objective, trusted clinical prognoses are still obtained manually by pathologists. Since pathologist obtained results remain the gold standard for prognosis generation and stain identification, it can be suggested that the human visual system’s perception of colour content is a defining characteristic behind staining analysis. Pathologist identified data is able to achieve results across the wide range of possible colour intensities and combinations
common in digital histopathology images. As many pre-defined algorithms are often unable to consistently adapt to the variability present in clinical images such as over staining, under staining, or morphological variations this thesis aims to mimic how the human visual system perceives brown and blue content for a wide range of hematoxylin and Ki67 staining levels. By specifically considering the colour characteristics of the $b^*$ chrominance channel, the proposed methodology achieves a fully automatic stain separation method that obtains comparable results to the manually identified ground truths for a wide range of staining variability. More specifically, this thesis utilizes the perceptual based colour characteristics of the $L^*a^*b^*$ colour space to imitate the physiological response to the brown and blue colours associated with Ki67 and hematoxylin stains. The following method is also a fully objective framework that does not require user defined parameters, learned models, or pre-defined stain vectors like in CD.

In this chapter, we will first present the novel colour separation framework for automated hematoxylin and Ki67 separation, followed by the other processing pieces of our pipeline to quantify PI, as shown in Fig 3.1. The framework includes a pre-processing phase to limit the small inconsistencies within the images, a background separation to remove irrelevant background details, colour separation to separate blue and brown colour content, before application of a final nuclei detection scheme. Another novelty of this work is the addition of an automated nuclei radii estimator described in section 3.2.4.

![Diagram of proposed method illustrating parameter selections at each step in the framework.](image)

**Figure 3.1** - Diagram of proposed method illustrating parameter selections at each step in the framework.

### 3.1 - The Human Visual System and $L^*a^*b^*$ Considerations for IHC Analysis

The simplest model of colour perception in the HVS includes a trichromatic stage followed by an opponent stage. The trichromatic stage consists of the stimulation of three independent short-, medium-, and long-wavelength cone receptors in the human eye [70]. The cone stimulation response represents the trichromatic stage of human vision where colour is initially perceived as a mixture of cone stimulation levels. Since each cone is not distinct to a specific wavelength of light, there exist overlapping regions
where various levels of stimulation can occur for multiple cones simultaneously. This property led to the generation of tristimulus colour representations, which are approximations of how the human perception of colour is generated. These approximations are not always consistent with the human visual system as noted by the perceptual inconsistencies of the RGB colour space discussed in Chapter 2. Following the trichromatic stage, the opponent stage refers to neural circuitry effects that compare the outputs of two or more cone types when there is overlapping spectral sensitivities. These differences in cone output were initially obtained through psychophysical measures of human observers which identified that the opponent colour pairs, red/green and blue/yellow, could not exist in the same colour representation [70].

The L*a*b* colour space models the opponent process theory described above by defining the -a*/+a* colour channel on a red/green axis and the -b*/+b* channel on a blue/yellow axis. The associated L* is used to describe the relative brightness of the observed colours in the a* and b* channels. Since the positive identification of Ki67 stain content is not dependent on the "lightness" or "darkness" of staining intensity, the L* channel is excluded from the proposed framework. The L* channel also introduces variability into the proposed stain separation model due to the overall staining intensity differences on an image by image basis. The a* (red/green) channel does not contribute to stain representation as it is unable to differentiate blue hematoxylin content from more brown Ki67 content. As a result, this work focuses on the b* channel characteristics which can used to differentiate brown Ki67 and blue hematoxylin staining content for the purpose of stain separation. The b* chrominance channel representation of colour content can be visualized in Fig 3.2, where an RGB colour wheel is translated into the b* chrominance channel intensity image. The b* chrominance colour wheel has been normalized to the range [0, 1] for viewing purposes, so the negative blue region is observed at 0 and the positive yellow region is observed at 1, where b* = ~0.5 is the division region where no colour content is observed.
Figure 3.2 - RGB colour wheel and b* channel equivalent. Note the b* channel effectively differentiates the blue and yellow content, however red and green content possess similar intensity levels.

3.1.1 - b* Chrominance Colour Differentiation Characteristics

One may consider the fact that Ki67 content is brown while the hematoxylin content is blue, so it may be questioned as to how the b* chrominance channel which differentiates blue and yellow is able to model the staining characteristics in a blue and brown image. When considering RGB value representations it can be stated that pure yellow is defined by the RGB vector [255, 255, 0] and pure blue is represented by the RGB vector [0, 0, 255]. Brown is observed when there are high mixtures of red and green with a low representation in the blue channel. In a brown colour the red channel is often greater than the green and a more yellow colour is observed similarly when green is larger than red. In both cases of brown and yellow, the red and green channels are both greater in magnitude than the blue channel. As a result, brown pixels associated with Ki67 staining are visualized in the +b* region of the b* chrominance colour channel and blue is in the -b* region. Following the opponent process theory, the b* channel models Ki67 content based on the differences between the red/green and blue colour channels. To examine this characteristic, the colour wheel image in Fig 3.2 was divided into ten equally spaced bins over the b* channel range and the correlated RGB values were averaged for each bin. The results of this procedure are observed in Table 3.1. As the b* channel bins increase from the negative blue to positive yellow region, each bin has a decreasing amount of blue content and an increasing amount of red and green content with the exception of bin 4. Bin 4 shows an artificially decreased blue representation because the RGB values were obtained by averaging the colour vectors. The colour representation in bin 4 is the result of a significant amount of pure red and green content where the respective RGB vectors have minimal blue content and as a result the average blue colour content is "pulled" lower than expected around the b* = 0.5 threshold observed in the image. The colour distribution Table 3.1 provided a rough approximation of how b* chrominance channel represents colour content which provided additional
motivation into investigating if the b* chrominance channel could describe colour properties observed in IHC histopathology images.

Figure 3.3 further examines the colour differentiating characteristics of the b* chrominance channel on various levels of Ki67 staining. Similar to the colour bins in Table 3.1, various RGB histopathology images were converted into the L*a*b* colour space where the b* channel was used to organize colour content. The RGB colour vectors were ordered according to their b* chrominance equivalent and were averaged over a range of 20 colour bins. Each bin represents the average RGB colour vector observed for the b* intensity range of each bin, which is defined in the title of each bin in Fig 3.3. The specific images selected for Fig 3.3 were chosen to illustrate the colour descriptive characteristics of the b* channel at various stain concentration levels and colour distributions. By evaluating the b* chrominance values on an image specific range it is possible to form a robust determination of staining content. The image specific range refers to the minimum and maximum range of b* intensities observed in each individual image, which allowed for a more adaptive colour comparison for images with lighter or darker staining characteristics. This effect can be observed in the last image in Fig 3.3 which contained very dark Ki67 expression levels at low b* intensities as a result of different scanning conditions. Despite the low expression levels the blue and brown staining content is effectively differentiated across the b* chrominance bins. The effect described for bin 4 in Fig 3.2 was not observed in the IHC stained histopathology images since there are no pure red or green colour values at any staining level that could influence the colour distribution. The apparent ability to separate blue and brown colour content associated with histopathology stains under various scanning conditions and concentration levels was the main motivational characteristic for developing a colour separation method that could be generalized to a wide range of clinical image datasets.
Table 3.1 - A visualization of how the b* chrominance colour channel represents RGB vectors for the range of colours observed in a basic colour wheel image.

<table>
<thead>
<tr>
<th>Bin</th>
<th>B Channel (mean)</th>
<th>Colour Wheel (B Chrominance)</th>
<th>Colour Wheel (RGB)</th>
<th>R (mean)</th>
<th>G (mean)</th>
<th>B (mean)</th>
<th>RGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-46.8</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td>56</td>
<td>111</td>
<td>184</td>
<td><img src="image3.png" alt="RGB" /></td>
</tr>
<tr>
<td>2</td>
<td>-32.8</td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td>107</td>
<td>142</td>
<td>198</td>
<td><img src="image6.png" alt="RGB" /></td>
</tr>
<tr>
<td>3</td>
<td>-18.8</td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td>170</td>
<td>160</td>
<td>197</td>
<td><img src="image9.png" alt="RGB" /></td>
</tr>
<tr>
<td>4</td>
<td>-4.7</td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td>185</td>
<td>150</td>
<td>99</td>
<td><img src="image12.png" alt="RGB" /></td>
</tr>
<tr>
<td>5</td>
<td>9.3</td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td>195</td>
<td>171</td>
<td>157</td>
<td><img src="image15.png" alt="RGB" /></td>
</tr>
<tr>
<td>6</td>
<td>23.3</td>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image17.png" alt="Image" /></td>
<td>190</td>
<td>165</td>
<td>124</td>
<td><img src="image18.png" alt="RGB" /></td>
</tr>
<tr>
<td>7</td>
<td>37.3</td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
<td>172</td>
<td>163</td>
<td>86</td>
<td><img src="image21.png" alt="RGB" /></td>
</tr>
<tr>
<td>8</td>
<td>51.3</td>
<td><img src="image22.png" alt="Image" /></td>
<td><img src="image23.png" alt="Image" /></td>
<td>200</td>
<td>188</td>
<td>75</td>
<td><img src="image24.png" alt="RGB" /></td>
</tr>
<tr>
<td>9</td>
<td>65.4</td>
<td><img src="image25.png" alt="Image" /></td>
<td><img src="image26.png" alt="Image" /></td>
<td>230</td>
<td>209</td>
<td>53</td>
<td><img src="image27.png" alt="RGB" /></td>
</tr>
<tr>
<td>10</td>
<td>79.4</td>
<td><img src="image28.png" alt="Image" /></td>
<td><img src="image29.png" alt="Image" /></td>
<td>247</td>
<td>227</td>
<td>25</td>
<td><img src="image30.png" alt="RGB" /></td>
</tr>
</tbody>
</table>
Figure 3.3 - Multiple colour distributions of b* chrominance channel. The bins represent the average RGB colour associated with each b* range divided into an arbitrary number of equally spaced bins. Regardless of stain content, the colour separation method is able to separate stain content at a zero threshold where the R and G components become greater than the blue stain resulting in brown content.
3.1.2 - Perceptual Linearity and b* Chrominance Membership

Perceptual linearity refers to the visual characteristic where a human observing the range of b* chrominance values from pure blue to pure yellow should note a consistent rate of change from one colour to the other. Since the HVS perceives colour on a non-linear scale, the b* channel is defined by a cubic root relation, seen in equation 2.9, to mimic this relationship. Therefore the numerical boundaries around pure blue and pure yellow are non-linear which can potentially influence algorithm quantification of Ki67 and hematoxylin. These non-linear boundary values are visualized in Fig 3.4 which displays the entire range of possible b* values obtained by converting all possible RGB vectors in a simulated 4096x4096 image into their ordered b* space equivalent. The 4096x4096 image was generated by assembling a matrix containing all the possible RGB colour vectors that can be represented on a [0,255] integer scale. RGB implementation began as [0,0,1], [0,0,2], [0,0,3] and continued until [255,255,255] was achieved which resulted in a total of $256^3$ RGB colour vector combinations. These vectors were then converted into the L*a*b* colour space and organized using the b* chrominance channel. The b* channel was sorted from -b* to +b* to generate the curve in Fig 3.4 representing the total possible b* range that could be obtained using RGB vectors. The red data points observed on the b* intensity curve are all of the b* intensities observed over the entire histopathology image dataset which were obtained using the RGB conversion framework described above. While the full b* chrominance curve is non-linear, the range of values observed in the dataset have an approximate linear relationship over the range [-42.27, 51.29] which is used to describe the respective levels of hematoxylin and Ki67 stain content.
Figure 3.4 - All b* intensities observed over all images in the validation dataset plotted on the range of all possible b* intensities from the RGB colour gamut.

While Fig 3.3 illustrates the perceptual range of b* channel intensities through colour bins, a more quantitative representation can be shown by ordering the unique b* intensities observed in the sample images shown in Fig 3.3. As observed in Fig 3.4, the range of b* intensities that occur in the image dataset are mostly located on the linear region of all possible b* intensities, however the curves generated in Fig 3.5 appear non-linear. These curves appear non-linear because unlike the range of values in Fig 3.4, only the unique b* intensities that are observed in each sample image are used to generate the b* chrominance curves in Fig 3.5. Since the b* intensities are ordered according to the observed image specific unique values and not the full range of values, the numerical jump between neighbouring data points is more significant and pronounced resulting in a sigmoidal relationship. The sigmoidal boundaries represent the regions of higher colour content compared to the more even and linear intermediate regions. This property results in a sigmoidal curve unique to each image where the various levels of b* chrominance are represented in a way similar to a fuzzy membership function. Observing Fig 3.5, each sigmoid generated from the sample images represent the ordered unique b* intensities present in each image where the curved extremity regions represent a greater rate of change from b* values from the "strong blue" to "strong brown" regions and the linear regions show a more steady rate of change. These
curves are completely dependent on the current image and by only considering the range of b* values within the image instead of all possible values, colour characteristics can be considered adaptively for variable staining levels. As a result, the organized b* unique intensity levels can be used to develop a robust colour differentiation membership regardless of colour or scanner variability. Also note that while b* = 0 represents no colour content, it does not appear suitable as a general threshold for every image.

![Ordered b* Intensities Unique to Each Image](image)

Figure 3.5 - The colour distribution characteristics of the b* channel are observed more quantitatively by ordering the unique b* intensities observed in each of the above sample images where the corresponding RGB colour to each unique value is represented on the curve.

### 3.2 - Image Processing Methods

#### 3.2.1 - Colour Separation Threshold

Based on the similar RGB expressions of brown and yellow, it can be suggested that the higher intensity of the red and green content in brown Ki67 staining allows the b* channel to approximate a linear model of fuzzy membership for hematoxylin versus Ki67 stain content. However, the wide stain variability observed for histopathology images as well as the fact that there are no pre-determined values for where "ideal" hematoxylin and Ki67 stain content should be located on the b* channel pose a barrier to generation of a single mathematical membership. Given an ideal setting where stains must pass specific intensity value thresholds before they can be considered as positive stain content, a general model would be applicable to every stained image and obtain good results. In clinical practice the wide variability between images requires independent consideration of overall staining characteristics. Recall that b* = 0
represents no colour content and should therefore represent a natural threshold between the negative blue region and the positive brown region. While this may work in theory, observation of the manually labelled images showed significant manual overlap for images with high Ki67 expression levels. This overlap is examined in Fig 3.6 where the manually placed markers for Ki67 and hematoxylin were isolated and the average $b^*$ values were obtained for each marker and its respective stain type. The isolated $b^*$ values were then plotted on the range of $b^*$ values observed in the image. Fig 3.6A demonstrates an image containing a large amount of both hematoxylin and Ki67 staining content while not suffering from over staining. For this image it was found that 92% of the markers labelled as Ki67 positive occurred at $b^* > 0$ and 99% of the hematoxylin labelled markers were $b^* < 0$. However, marker consistency was lower for images suffering from over staining, such as in Fig 3.6B which had 99.6% of the Ki67 markers $b^* > 0$, but only 59% of markers labelled as hematoxylin were in the $b^* < 0$ region. This suggests that depending on the overall staining observed in an image, a pathologist may slightly shift their colour perception criteria for what counts as positive Ki67 content or negative hematoxylin content, which inhibits the use of a globally standardized threshold for stain separation in the $b^*$ channel.
Figure 3.6 - Comparison of pathologist labelled manual data for images containing wide stain variability. (A) Represents an image containing a large amount of both stains without over staining. (B) Represents an image with over stained Ki67 expression.

Since it is difficult to generate a standardized threshold, an adaptive approach can be performed on an image by image basis for stain separation to generate two intensity images representing respective confidence levels of each stain type independently. These confidence images for each image in the dataset can represented by the following:

$$H(x, y) = \begin{cases} \frac{(-b(x, y) + T)}{b_{\min}} & \text{for } b(x, y) < T \\ 0 & \text{for } b(x, y) \geq T \end{cases}$$  \hspace{1cm} (3.1)
\[
K(x, y) = \begin{cases} 
0 & \text{for } b(x, y) < T \\
\frac{(b(x, y) - T)}{b_{\text{max}}} & \text{for } b(x, y) \geq T 
\end{cases}
\] (3.2)

Where \(H(x, y)\) and \(K(x, y)\) are the generated confidence level intensity images for hematoxylin and Ki67 respectively, \(b(x, y)\) is the \(b^*\) chrominance channel image with pixel locations \((x, y)\), \(b_{\text{min}}\) is the minimum \(b^*\) intensity and \(b_{\text{max}}\) is the maximum \(b^*\) intensity in the image, and \(T\) is an adaptive threshold value that segments the \(b^*\) channel image into separate stain images. The resultant confidence level images for each stain are on a \([0, 1]\) scale where a value of 0 is associated with zero or extremely low stain colour content and a value of 1 represents the strongest stain colour content observed in the image.

3.2.2 - Determining an adaptive threshold \(T\)

Due to staining variability, pathologists appear to adaptively change their definition of what is considered positive and negative Ki67 stain content on an image by image basis to counteract inconsistencies with staining protocols and maximize the amount of useful information for diagnostic decision making. In order to model this perceptual shifting of stain content, this thesis proposes an adaptive threshold to correct for the variable decision making parameters inherent in a pathologist determined PI. The \(b^*\) chrominance channel is used to model the variations on an image by image basis which is defined by the image specific range in \(b^*\) values. For hematoxylin dominant images the natural threshold \(T(b^*) = 0\) is suitable for colour separation because there is a more compact range of \(b^*\) values in contrast to images that contain high Ki67 expression and are much more variable. Therefore an adaptive threshold is proposed to perform colour separation for images with over staining and to increase concordance with manual scoring. A modified balanced histogram approach was designed to shift the initial threshold to a \(b^*\) intensity value that more closely resembles a pathologist's altered perceptual decision making criteria. Similarly to the local mean adaptive threshold method used for background subtraction, most algorithms of this type perform optimally on histograms that contain two distinct classes or distributions present in every image. Unfortunately, since there is a more linear transition from blue hematoxylin content to brown Ki67 content there are rarely two distinct classes. Some images also contain a high amount of only one stain type when there are low amounts of identifiable cancerous nuclei or when they suffer from Ki67 over staining. Sample histograms are shown in Fig 3.7 which demonstrate the absence of two reliable class distributions for images containing primarily high hematoxylin and Ki67 content as well as an image with a high expression of both.
Figure 3.7 - Histogram comparison for images with high, low, and mixed expression levels. Hematoxylin contains a more defined range while Ki67 has a wider distribution across b* intensities.

The proposed modified balanced histogram approach first defines an initial threshold at the bin representing the zero chrominance boundary b* = 0 and iteratively shifts the threshold left or right by one bin depending on the amount of staining content on either side of the current threshold. This process is modelled for the $i^{th}$ iteration by the following equations:

\[ T_{bin_i} = \begin{cases} T_{bin_i} + 1 & \text{if } A_i \geq B_i \\ T_{bin_i} - 1 & \text{if } A_i < B_i \end{cases} \quad (3.3) \]

\[ A_i = \sum_{x=T_{bin_i}}^{N_i} h(x) \quad , \quad x = [T_{bin_i}, T_{bin_i} + 1, \ldots, N_i] \quad (3.4) \]

\[ B_i = \sum_{x=M_i}^{T_{bin_i}} h(x) \quad , \quad x = [M_i, M_i + 1, \ldots, T_{bin_i}] \quad (3.5) \]

and,

\[ M_i = M_i + 1 \quad \text{if } A_i \geq B_i \quad (3.6) \]

\[ N_i = N_i - 1 \quad \text{if } A_i < B_i \quad (3.7) \]

Where, $h(x)$ is the histogram for the current image, $x$ is a single bin in the histogram, $T_{bin_i}$ is the histogram threshold bin for the $i^{th}$ iteration of the algorithm, $A$ and $B$ are the sum of histogram counts on either side of the threshold $T_{bin}$, $N_i$ is the upper boundary threshold for number of bins included in the $A$ calculation, and $M_i$ is the lower boundary for bins included in the $B$ variable summation. For each
iteration, the threshold is shifted by one bin in the direction of the larger summation and simultaneously a single boundary bin is removed from the "larger" side. In the unmodified implementation the threshold shift is away from the larger summation, however the lack of two distinct classes results in a consistently repeated threshold direction until the threshold exceeds the histogram boundary and an error is generated.

The original implementation also iterates over the entire histogram until $T_{bin_t} = M_i = N_i$, where the total number of algorithm iterations is equal to the number of bins in the histogram. The proposed modified version instead incorporates a stop condition to ensure reliable threshold selection. The stop criteria is determined by tracking the current $T_{bin_i}$ for each iteration and recording the local min and max values associated with a change in threshold shift direction. When a change in direction is detected, the algorithm is halted if a change in direction has occurred within the last 10 iterations. This stop condition prevents meaningless iterations and prevents continued iterating back and forth over the same small binned region until all bins are exhausted. The stop condition is shown in Fig 3.8 which illustrates the perceptual shifting effect that can occur for images with one dominant stain type. The final threshold $T$ for colour stain separation is determined as the average of the last two recorded local maximum and minimum. Final thresholds for images with various staining levels are observed on the $b^*$ histograms in Fig 3.9.

![Figure 3.8 - Balanced threshold determination without stop (Left) and with stop condition (Right). Application of a stop condition to the balanced threshold selection framework eliminates oscillating shifting effect that occurs for many TMA core images as a result of Ki67 histogram effects.](image-url)
Figure 3.9 - Adaptive threshold comparison for images with high hematoxylin ($T = 0$), high Ki67 ($T = 3.26$), and mixed ($T = 0.99$) expression levels.

Fig 3.10 illustrates the confidence images $H(x,y)$ and $K(x,y)$ with the respective colour separated images after application of the adaptive balanced histogram threshold $T$. In order to better compare the validity of the proposed adaptive threshold approach, experimental analysis is performed on various methods of threshold selection. The different threshold conditions that are compared are described in section 3.4.4.
3.2.3 - Pre-Processing

**Colour Denoising**

While it is extremely difficult to normalize histopathology images in a way that standardizes the observed stain characteristics, some pre-processing methodologies can be applied to improve the performance of automatic algorithms. In the proposed framework a vector median filter is applied to images before additional processing. The VMF is able to smooth/reduce some noise patterns common to image capture systems such as grainy Gaussian noise while maintaining the original colour characteristics in the image. After smoothing, a local mean adaptive threshold method [71] is applied to isolate stained nuclei content from lighter background content. The background subtraction method was performed on the L* channel as the stained nuclei content was composed of darker luminance values compared to the lighter background content. The thresholding method was able to roughly segment stained nuclei content apart from very light staining associated with exterior stromal content. This method was preferred over similar methods such as Otsu's method [61] as it was more inclusive of light stain content which resulted in less segmentation holes within nuclear boundaries.

![Figure 3.10 - Grayscale confidence images $H(x, y)$ and $K(x, y)$ with colour equivalent images obtained by applying a binary image mask to the sample RGB image.](image)
A VMF is applied to smooth colour variations while maintaining some of the colour characteristics. As previously described in Sec 2.3.1, the vector median filter generates an average colour vector for each pixel by averaging the values observed in a defined pixel neighbourhood. The mean colour vector for a pixel at location \((x, y)\) is defined by:

$$\mathbf{w}_{mean}(x, y) = \left[ \frac{\sum_{i=1}^{n} R_i}{n}, \frac{\sum_{i=1}^{n} G_i}{n}, \frac{\sum_{i=1}^{n} B_i}{n} \right] \quad \text{for } i = [1, 2, \ldots, n] \quad (3.8)$$

Where \(i\) represents each pixel located within the pixel neighbourhood, \(n\) is the number of pixels within the neighbourhood (for a 3×3 window \(n = 9\)), and \(R_i, G_i, B_i\) are the pixel intensity values for each RGB colour channel at pixel \(i\). Next, the Euclidean distance between the mean colour vector and each pixel within subset \(i = [1, 2, \ldots, n]\) are determined using the following formula.

$$D_i = \sqrt{(R_i - w_{mean\; R})^2 + (G_i - w_{mean\; G})^2 + (B_i - w_{mean\; B})^2} \quad (3.9)$$

The distance measures are then ordered into a descending or ascending array where a specified number of pixels are removed from the subset and a final colour vector is generated by taking the mean of the remaining pixels in the subset. The number of pixels removed from the ordered distance subset are determined by a pre-set parameter referred to as the alpha trim parameter. The other pre-set parameter for application of a VMF filter is the window size that defines the pixel neighbourhood and directly impacts the strength of the smoothing effect performed by the filter.

**Background Subtraction**

In addition to filtering, an automatic adaptive thresholding method [71] is applied to the grayscale equivalent to eliminate background content and isolate nuclei. Like many automatic thresholding segmentation techniques the local mean adaptive method separates image characteristics into two separate classes based on the observed gray level intensities over the image. It is an iterative approach where an initial threshold is determined simply by taking the mean of the grayscale image. The threshold is applied to generate two grayscale regions and a new threshold is determined by:

$$T_{new} = \frac{(\text{mean } R_1) + (\text{mean } R_2)}{2} \quad (3.10)$$

Where \(R_1\) and \(R_2\) are the two grayscale regions obtained through application of the previous threshold \(T\). This procedure is performed until the difference between the new threshold and the previous threshold is below some minimal value. Despite being an iterative approach, the local mean threshold method converges to a final \(T\) quickly because it is performed on a grayscale image and is not computationally demanding. Since the images in the dataset and many other histopathology images contain a significant
portion of whitespace, the final threshold is effective at eliminating the very light staining content around nuclei while remaining inclusive of lighter stained regions within the nuclear boundary.

3.2.4 - Nuclei Detection

Accurate PI estimates are more reliable and accurate when generated using individual nuclei counts instead of expression level estimations using region parameters such as number of pixels or visually estimated features. In this work, a nuclei detection algorithm proposed in [72] was modified for implementation on the confidence images $H(x, y)$ and $K(x, y)$ to determine individual nuclei counts for each stain type. The nuclei detection scheme was chosen as it is able to identify overlapping nuclei structures and quantify the number of individual nuclei in an image opposed to the aforementioned contour modelling and watershed segmentation procedures. Segmentation of nuclear boundaries does not contribute to the final PI estimate, so there is no need to include additional processing schemes for this purpose. The main contribution to nuclei detection in this work is the introduction of an automatic cell radius estimator that automatically determines the cell radius parameter required for generation of the iterative voting image. Prior to this addition, the cell radius estimate was a user defined parameter which introduces subjectivity into the model and may lead to biased or inconsistent results. Thus the objectivity of the proposed framework is preserved with implementation of the cell radius estimator. A summary of the nuclei detection steps can be found in Table 3.2. This implementation also excludes the sum-weighted voting image procedure as it did not improve results and introduced more expensive computations.

The cell radius estimate was determined automatically through a number of steps. First, the unique binary objects in a confidence image $H(x,y)$ or $K(x,y)$ were identified. The area of these objects in terms of number of pixels were assembled into a sorted histogram where the top and bottom 20% of values were removed. This trimming was required because the binary image may classify large nuclei clusters as a singular unique object and similarly a single isolated pixel may also be identified as a unique object. The remaining object areas are then averaged using the mean to determine a rough area estimate for the "average" nuclei in the image. A final radius is determined by applying the equation for the area of a circle where the area is given as the mean value obtained in the previous step. Since the nuclei detection is performed on each stain confidence image independently the algorithm is able to apply separate radius estimates for hematoxylin nuclei and Ki67 stained nuclei. This is an advantageous characteristic because cancerous nuclei often exude different morphological features such as enlarged or poorly defined nuclear boundaries and by applying separate nuclei radius estimates it is ensured that a single radius estimate is not incorrectly generalized to all nuclei.
Table 3.2: Summary of Nuclei Detection Algorithm Steps with Cell Radius Estimator

1. Generate cell radius estimate \( r \) for nuclei in image.
   - Create binary image mask and find pixel area of unique objects.
   - Generate a sorted histogram of object areas.
   - Trim top and bottom 20% of objects.
   - Determine mean radius \( r \) in image by approximating objects with area of a circle \( A = \pi r^2 \).

2. Calculate gradient image \( G(x, y) \) with associated direction.

3. Record pixels \( \epsilon(x, y) \) with gradient magnitude greater than threshold \( R \).

4. Generate iterative vote image \( V(x, y) \) by determining center of cone area with radius \( r \) and applying Gaussian kernel.

5. Apply mean shift algorithm to determine final nuclei seed locations.

The gradient image \( G(x, y) \) is obtained via application of the Sobel kernel with associated gradient magnitude \( |G(x, y)| \) and direction \( \theta(x, y) \) values. Nuclei staining in the confidence images is variable, so weaker gradient magnitude intensities below a threshold \( R \) are not considered for calculation of the voting image. The voting image \( V(x, y) \) is iteratively generated by adding a Gaussian kernel to a cone center location for each gradient magnitude value \( |G(x, y)| > R \). The Gaussian kernel was defined by:

\[
g(u, v, \sigma) = \frac{1}{2\pi \sigma^2} \exp \left( -\frac{u^2 + v^2}{2\sigma^2} \right)
\]

Where \( (u,v) \) are the kernel coordinates defined by the kernel size and \( \sigma \) is the standard deviation of the distribution. Locations for the application of the Gaussian kernel are determined by imposing a cone on each pixel within the gradient magnitude subset. The cone direction is determined by the gradient direction and the kernel is applied to the voting image at cone center location \( V(\bar{x}_c, \bar{y}_c) \). The locations of the cone center for a pixel \( |G(x, y)| > R \) are determined by the equations:

\[
\bar{x}_c = x - \frac{(r_{min} + r_{max}) \times \sin(\theta(x, y))}{2}
\]

\[
\bar{y}_c = y + \frac{(r_{min} + r_{max}) \times \cos(\theta(x, y))}{2}
\]

Where \( r_{min} = \frac{r}{2} \) and \( r_{max} = 1.5r \), \( r \) is the automatically determined cell radius value, and \( \theta(x, y) \) is the gradient direction at pixel \( (x, y) \). The voting image generation procedure is defined for one pixel iteration as:

\[
V(\bar{x}_c, \bar{y}_c) = V(\bar{x}_c, \bar{y}_c) + g(u_1: u_2, v_1: v_2)
\]
Where $V(\mu_x, \mu_y)$ is the cone center indices and $\mu_{x1}, \mu_{xmax}, \mu_{y1}, \mu_{ymax}$ represent the kernel boundaries equal to the size of the Gaussian kernel $g(u, v)$.

In [72], additional voting image processing procedures were implemented to refine the voting image distribution, however for this implementation these additional steps were not included as they failed to impact algorithm performance. Once the final voting image $V(x,y)$ is defined, a maximum shift algorithm was implemented to determine final nuclei centers. The maximum shift algorithm considers each pixel intensity in the voting image and shifts the window to the maximum intensity observed. The window is continually shifted to a new maximum value until there are no greater pixel intensities observed and a final nuclei center location is recorded at the last defined maximum value. This procedure is applied to all the pixels in the voting image without replacement, which means that if a pixel was evaluated at any time it will not be considered again for the maximum shift algorithm. A visualization of the main procedural steps in the nuclei detection framework are observed in Fig 3.11 with final results illustrated in Fig 3.12 with the manual seed counts.

![Figure 3.11 - Nuclei detection procedure on hematoxylin confidence image $H(x,y)$ where cone position represented for one pixel location, $V(x,y)$ is the voting image, and the final seeds obtained by the maximum shift algorithm. The cone placement image was shown in red for optimal viewing of the cone placement location.](image-url)
3.3 - Validation Methods

Validation methodologies were designed for parameter selection where algorithm performance evaluation for each step in the proposed automatic framework was evaluated. The manual dataset generation procedures and known challenges are discussed as well as the parameter selection methods for nuclei detection, colour stain separation, and PI estimation comparisons. The proposed validation methods first aim to examine the performance of the nuclei detection scheme and the automatic PI. The secondary aim of the validation methods section is to evaluate the automatic cell radius estimator and adaptive colour stain threshold value $T$.

The colour stain separation threshold is performed prior to the nuclei seed detection in the proposed framework, however the quality of stain separation must be evaluated as a function of PI estimates because the manual data consists of nuclei counts which do not allow for pixel based comparisons. As a result, the automatic nuclei detection performance must also be evaluated prior to final PI comparisons. Nuclei detection is performed independently on each stain separated confidence images, so it is possible to evaluate both the seed detection method and cell radius estimator for each stain separately. Nuclei detection is evaluated using a seed validation framework that compares the automatic and manual markers within a pixel neighbourhood to calculate performance metrics for sensitivity, precision, effective rate, and F1 score. The F1 score is generated using sensitivity and precision metrics and represents the overall performance of nuclei detection for each stain type, while the effective rate
compares the total nuclei counted in each colour separated image. The nuclei detection metrics are used to compare the performance between Ki67 and hematoxylin across all of the dataset images. The F1 score and effective rate are also used to compare the colour separation influences on nuclei detection when different thresholds are applied for colour separation. Automatic cell radius estimation performance is evaluated by comparing the above metrics against a manual range of arbitrarily selected cell radius values. The associated performance comparisons can then be evaluated to examine the validity of the automatic nuclei detection framework.

The primary metric for colour threshold selection are the image specific PI estimates resulting from the automatic colour separation framework. The PI is the final output of the algorithm where the most significant validation metrics are determined. Automatic and manual PI comparisons are performed directly through an average difference calculation, however other validation metrics include the Pearson correlation coefficient and the Kappa similarity statistic. The Pearson and Kappa metrics are used to compare the results of the current work against similar methods in order to better prove the validity of the perceptual colour separation method. A flow diagram illustrating the nuclei seed validation and final PI comparisons are shown in Fig 3.13, where the final validation metrics used in the algorithm evaluations are shown in red.

![Flow Diagram](image_url)

**Figure 3.13** - Flow diagram of validation procedures and parameter determination after stain separation. Nuclei seed detection is performed on each confidence image and resultant PI estimates are compared with the manual data.
3.3.1 - Nuclei Detection Performance

The nuclei detection algorithm is a fully automatic method for nuclei seed generation applied to each confidence images $H(x,y)$ and $K(x,y)$ so validation results of the nuclei detection framework are directly related to the quality of the colour stain separation threshold. The following section outlines the procedures applied to compare each automatically detected seed location to the manual equivalent.

Each automatic seed location represents the stain positive nuclei identified in the respective confidence images. To compare the automatic seeds to the manual equivalent a seed validation procedure is proposed to test for the presence of a manual seed within a neighbourhood window. Since the manual seed markers were identified with different colour markers, green for Ki67, and magenta for hematoxylin, a binary marker image could be generated for each manual marker type. The RGB colour vector associated with each stain marker were identified and a respective binary image map was created which consisted of values of 1 where seeds were detected in the colour image. Therefore the binary images shown in Fig 3.14 were able to store the manual seed locations for each stain type and could be compared to the automatically determined seed locations.

![Manual Seeds](image1)

![Hematoxylin Markers](image2)

![Ki67 Markers](image3)

Figure 3.14 - Manually obtained nuclei image with green markers for Ki67 and magenta markers for Hematoxylin. The respective binary image masks for the respective stain markers used in automatic seed window validation are also shown.

Since nuclei detection was performed on each confidence image, a single TMA core image from the dataset underwent nuclei detection and marker validation twice, once for each stain type. Each final seed location obtained by the automatic algorithm was compared to the binary image mask representing the manual seed locations. It was expected that the number of seeds detected with the automatic method would not match the number of seeds within the manual dataset, so the pixel neighbourhood around the coordinates of each automatic seed were compared to the manual binary image mask. If there was a manual seed located within the circular pixel neighbourhood around the automatic seed location a true
positive (TP) count was determined. The diameter of the circular pixel neighbourhood window was equal to 2x the estimated cell radius for each respective stain type. The manual seed that was correlated to the automatic seed was then removed from the manual binary map so the following automatic seeds could not be paired with an already accounted for manual seed. If there are two or more detected manual seeds within the detection window only the closest manual marker was counted as the true positive and removed from the binary image map. If no manual seed was detected within the detection window then a false positive (FP) count was determined. False negative (FN) markers were determined by quantifying the remaining manual nuclei markers not correlated to an automatically detected seed. Interestingly, for nuclei detection algorithms it is not possible to determine a true negative (TN) count because only the presence of a nuclei may be evaluated. This partially limits the breadth of validation metrics applicable to nuclei detection schemes.

3.3.2 - Sensitivity, Precision, F1 Score, Effective Rate

Nuclei seed detection validation is evaluated using the TP, FP, and FN counts obtained from the window validation procedure. The non-existence of TN counts inhibits the calculation of commonly used metrics such as specificity and accuracy, so performance metrics applicable to the TP, FP, and FN values are used. To evaluate nuclei detection performance, sensitivity, precision, F1 score, and the effective rate are calculated for each stain type. Sensitivity evaluates the automatic algorithm's ability to correctly detect nuclei center locations when compared to the manual seeds. The sensitivity calculation is given by:

\[ S = \frac{TP}{TP + FN} \]  

(3.15)

A sensitivity value of 1 correlates to a 100% success rate where all of the manually determined seeds have been correctly identified by the automatic method and a value of 0 would mean that not a single automatic seed is valid.

Precision evaluates how many automatically determined seeds were relevant to the manual seed locations. If the automatic algorithm placed seed markers all over the nuclei image in a tight grid pattern, it may achieve 100% sensitivity, however it would result in a very low precision value since there is a large number of automatic seeds that are not associated with a manual seed (FP). The precision equation is defined as:

\[ P = \frac{TP}{TP + FP} \]  

(3.16)

Where a value of 1 is an 100% success rate where every single automatic seed location was coincident with a manual seed.
A high precision value may be obtained if the automatic method correctly identifies only 10 nuclei while returning 0 FP, however this would result in a poor sensitivity value as thousands of nuclei would be unaccounted for. By measuring both sensitivity and precision the automatic algorithm performance can be modelled as a trade-off between these values where optimal performance results in a high $S$ and a high $P$. This sensitivity and precision trade-off can be modelled by the calculation of a total F1 score given as:

$$F1 = \frac{2 \times S \times P}{S + P}$$  \hspace{1cm} (3.17)

Where $S$ is sensitivity and $P$ is precision. By incorporating both sensitivity and precision the F1 score is a more robust measure of overall nuclei detection performance. An F1 score of 1 is obtained if both sensitivity and precision are 100% successful, but if there is a larger difference the resultant F1 score is heavily influenced by the smaller value. For example values of $S=0.8$ and $P=0.2$ result in an F1 score of 0.32.

In addition to the F1 score, an effective rate metric was calculated to observe how similar the total counts are between the total automatic and manual counts. The equation for the effective rate is given by:

$$ER = \frac{\text{Total \ # \ Automatic \ Seeds}}{\text{Total \ # \ Manual \ Markers}}$$  \hspace{1cm} (3.18)

Where $ER$ is the effective rate and simply models the difference between the total nuclei counts. An $ER$ value of 1.1 indicates that the automatic method detected 10% more seed locations than those in the manual counts. A value less than 1 indicates that the automatic method detected less seeds. While the F1 score takes into account the difference in counts with the sensitivity and precision trade-off, the $ER$ is a more comprehensive comparison of total counts.

### 3.3.3 - Cell Radius Estimator

The cell radius estimator was implemented to fully automate the cell detection procedure by eliminating the requirement of user defined parameters. Since the cell radius estimator greatly affects the nuclei counts in a confidence image, it is important that the automatically determined radius is robust and reliable. The cell radius estimator is an algorithm parameter value so it must be evaluated by examining the nuclei detection performance metrics. The validity of the adaptive cell radius estimator can be determined through a performance comparison of other possible cell radius estimates that may be chosen by a user. In a typical experimental design this comparison is made though the generation of a receiver operating characteristic (ROC) curve where the x-axis is defined by the false positive rate (1-specificity) and the y-axis is the true positive rate (sensitivity). The optimal parameter selection is then determined by the location of the plotted points within this space that represent the sensitivity/specificity values for a
single parameter estimate. However, recall that since there is no valid test for a true negative during nuclei seed detection a traditional ROC curve is impossible. Instead an alternative ROC curve can be generated for precision vs sensitivity where the automatically determined cell radius estimator can be compared to possible user estimates. A sample of the precision/sensitivity trade-off curve can be seen in Figure 3.15 where an ideal value of 100% sensitivity and 100% precision is located in the top right corner.

![Precision vs Sensitivity](image)

**Figure 3.15 - Precision vs Sensitivity trade-off curve. The top left star represents a perfect score of 100% for both sensitivity and precision.**

The cell radius estimates can be represented in terms of pixels or µm (1 pixel = 0.5µm) depending on image resolution and user preference. It can be expected that a smaller radius estimate leads to more automatically determined seeds on a confidence image, causing higher sensitivity and lower precision, and a large radius estimate would result in both lower sensitivity and precision. Selection of a small radius estimate would affect the Gaussian kernel placement because it may not reach the center of the nuclei leading to multiple seed locations within one large nuclei. It can also expected that a significantly large radius estimate would cause the Gaussian kernel to "over shoot" the ideal nuclei center location, which may result in final seed locations outside of the nuclear boundaries. Therefore there must exist a reasonable cell nuclei estimate that results in optimal sensitivity and precision trade-off. This value is expected to occur at the "knee" of the alternative ROC curve where a maximum sensitivity and precision are achieved at a specific cell radius estimate.
3.3.4 - Adaptive Colour Threshold $T$

Similar to the cell radius estimate, the adaptive threshold $T$ for stain separation must be evaluated to ensure the perceptual modelling concept is applicable to a wide range of Ki67 stained images. The goal of the adaptive colour threshold is to minimize the overall difference between manual and automatic PI estimates while achieving a more objective and fast determination. As a result, the validity of the adaptive threshold will be confirmed by examining the average PI differences for each image for 5 cases: $T=b^*=0$, $T=$Otsu's method, $T=$adaptive method without any stop criteria, $T=$adaptive method with a stop condition, and a final case where $b^*$=adaptive method with stop condition for $T>0$ and $b^*=0$ for $T<0$.

In order to examine the relationship between the adaptive threshold and the nuclei detection methodology the F1 score performance is graphed against the cell radius estimator and a range of possible cell radius estimates for each of the above threshold scenarios. In addition to the F1 score, additional comparisons are generated using similar figures for the difference in PI estimates and the effective rate. Results of this procedure are found in Chapter 5 where each graph is compared and described in depth.

3.3.5 - Proliferation Index Accuracy

The following section describes the performance evaluation logic and methodologies for automatic PI estimates and the comparison to manually obtained results. As previously mentioned the PI is the proportion of positively stained Ki67 nuclei within a region of interest or over an entire TMA core image. It is defined by:

$$PI = \frac{\# \text{ Ki67 Nuclei}}{\text{Total \# Nuclei}} \times 100\% \quad (3.19)$$

Where the number of Ki67 nuclei are obtained from the image $K(x, y)$, and the total number of nuclei are the summed results of the nuclei detected in $(H(x, y) + K(x, y))$. The final PI estimates can then be compared against the results of the manually labelled images.

Since the PI estimation is determined by nuclei counts obtained by each of the confidence images, one can assume the nuclei detection performance is closely tied to the overall PI estimate. To a certain extent this may be true, however if the consistency of the nuclei detection scheme is relatively constant the automatic PI estimate should remain relatively unchanged. For example if the automatic nuclei detection framework detects 30% more nuclei than the manual method in both the Ki67 and hematoxylin confidence images, then the overall PI should still resemble the manual PI, assuming a good colour separation, because the count proportions between confidence images remain consistent. Therefore the quality of an automatic PI estimate must be evaluated as a relation to the manual PI.
3.3.6 - Pearson Correlation Coefficient and Kappa Similarity Statistic

Comparisons between the automatically determined PI estimate and the manual estimate may be performed by applying various methods. The simplest of which is simply plotting the automatic vs manual PI estimates and calculating a Pearson coefficient. The Pearson coefficient linearly evaluates the correlation between two variables where a value of 1 is complete positive correlation between data, a value of 0 is no correlation, and a value of -1 is complete negative correlation. The Pearson coefficient \( r \) is determined by the equation:

\[
 r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}
\]  

(3.20)

Where \( n \) is the sample size of the dataset (\( n=30 \)), \( x \) and \( y \) represent the PI estimates for the automatic and manual methods for the \( i^{th} \) image, and \( \bar{x}, \bar{y} \) are the sample means for each respective PI set.

The Kappa similarity statistic is a measure of agreeability between the results of multiple observers or methods. It is commonly applied to medical data as a measure of agreeability among medical professionals where standardized procedures are performed to determine final prognosis/diagnosis. A Kappa value is obtained by assembling results of two or more observers in an agreeability matrix or table where the data is split into classifications or data subsets. In this work, the PI estimates are split into four subsets of 0-10, 10-20, 20-30, and 30+ expression levels. Recall that reliable cut-off values for Ki67 are not defined but are usually separated into low (<10%), intermediate (11-30%), and high (>30%) levels of proliferation activity [29]. Since the difficulty in Ki67 classification occurs for intermediate staining regions, the middle region was further separated to create the four subsections in order to properly evaluate intermediate staining levels. An example of an agreeability table is observed in Table 3.3 where each count represents the PI comparisons between the manual and automatic methodologies.

<table>
<thead>
<tr>
<th>Manual PI Estimate</th>
<th>Automatic PI Estimate</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>0-10</td>
<td>10-20</td>
</tr>
<tr>
<td>10-20</td>
<td>e</td>
<td>f</td>
</tr>
<tr>
<td>20-30</td>
<td>i</td>
<td>j</td>
</tr>
<tr>
<td>30+</td>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>Total</td>
<td>( m_0 )</td>
<td>( m_1 )</td>
</tr>
</tbody>
</table>

Kappa
In the above example, if counts were only located in the diagonal a, f, k, z vector then there would be a perfect agreement and a Kappa value of 1. The Kappa statistic is calculated based on the difference between the observed agreement and how much agreement would be expected to be present by chance alone, which is the expected agreement [73].

The observed agreement is calculated by:

\[ p_o = \frac{a+f+k+z}{n} \]  

(3.21)

Where \( p_o \) is the observed agreement, \( a, f, k, \) and \( z \) are the numerical values associated with each respective cell, and \( n \) is the total number of samples evaluated.

The expected agreement is defined by:

\[ p_e = \left( \frac{n_0}{n} \right) \times \left( \frac{m_0}{n} \right) + \left( \frac{n_1}{n} \right) \times \left( \frac{m_1}{n} \right) + \left( \frac{n_2}{n} \right) \times \left( \frac{m_2}{n} \right) + \left( \frac{n_3}{n} \right) \times \left( \frac{m_3}{n} \right) \]  

(3.22)

Where \( p_e \) is the expected agreement and the \( m \) and \( n \) values represent the summation of each column and row respectively.

A final Kappa statistic is calculated by defining the difference between the observed and expected values.

\[ K = \frac{p_o - p_e}{1 - p_e} \]  

(3.23)
Chapter 4

Validation Results and Parameter Selection

In the following chapter experimental results are evaluated for nuclei detection and colour separation for various parameter selections. Nuclei detection is primarily evaluated through F1 score comparisons, where the cell radius estimator is evaluated by examining the sensitivity/precision trade-off curves for arbitrarily selected cell radius estimates representing possible user defined parameters. Further F1 score comparisons are determined for various colour separation threshold experiments. Before validation metrics are compared, considerations must be made to ensure consistency across the dataset images. The effects of high and low staining content on the validation metrics are explored and their impact on validation consistency is corrected by calculating the performance metrics as a function of total detected cells instead of an image to image average.

The adaptive thresholding method is compared against Otsu's method for foreground and background separation to determine the validity of the proposed framework under various conditions. The proposed balanced colour threshold method is evaluated for five possible thresholding operations: Otsu's method, the zero chrominance threshold $T=0$, the balanced adaptive approach, the balanced approach with inclusion of a stopping criteria to limit perceptual shifting in images, and a final method where the balanced threshold with stopping criteria is only applied when the final $T>0$. The performance results are compared as a function of F1 score, effective rate, and the mean PI difference between the automatic and manually determined ground truth data.

4.1 - Manual Dataset Generation

The proposed framework was implemented on 30 canine mammary TMA cores obtained from the Ontario Veterinary College (OVC) at the University of Guelph. Tumour tissues from various canines were assembled into TMA blocks as in Fig 2.4 where each core is an individual tumour sample. Samples were sliced off the TMA block with a tissue thickness of 5µm and before staining with hematoxylin and Ki67. The TMA images were generated using an unknown WSI scanner at 20X magnification with a pixel resolution of 0.5µm by 0.5µm. Many TMA specimens in the sliced array images suffered from various quality challenges such as folded over cores, total core loss, necrotic tissue sections, severe stain inconsistencies, or other image generation artifacts. In order to assemble a useable dataset for automatic image analysis individual cores were cropped from the full block images, using Sedeen viewer [74], to
approximately 1450X1450 non-compressed tiff images. Fig 4.1 illustrates the cropping procedure performed in Sedeen image viewer.

Figure 4.1 - Cropping of single TMA core from larger TMA block section for the purpose of manual counting in Sedeen Viewer [74].

Each image underwent manual nuclei identification procedures performed by a DVM, DVSc candidate from the OVC department of Pathobiology. Counting protocols were performed through the use of ImageJ software [75], which allowed for the placement of nuclei seed identifiers while counting was performed. For each cropped TMA image, the pathologist zoomed in on a nuclei region so nuclei markers could be placed with a mouse cursor for each stain type. Once the region was complete, a new region was examined until the entire TMA core was manually labelled and identified. Fig 4.2 illustrates the cell counter tool used un ImageJ for nuclei seed marker placement. Once the final marker placements were complete for an image, the total marker counts were used to calculate a manual PI estimate.
Since each image contained thousands of nuclei and multiple hours were required to complete one image, there are marker inconsistencies within the dataset as a result of the laborious and tedious nature of manual counting. Regions associated with the border regions of the TMA images often contained necrotic or degraded nuclei, so manual counting was avoided and no markers were placed despite the presence of potentially viable nuclei. Another inconsistency within the dataset as a result of manual counting is that some markers were placed outside or near nuclear boundaries which poses validation challenges for the nuclei detection scheme. It is also possible that valid nuclei in the TMA images were simply missed and could also result in inconsistent nuclei detection results. The above challenges are an example of some of the potential issues that can inhibit algorithm performance validation on a manual dataset, however the determined PI estimate is based on total nuclei counts and is a more robust metric of algorithm performance than simply nuclei detection.
4.2 - Nuclei Detection

4.2.1 - Validation Metric Considerations

The F1 score is an invaluable metric for evaluating the nuclei detection performance as it incorporates both precision and sensitivity. It is determined for each confidence image so an F1 score for each stain type is obtained which can be compared to evaluate the effects of staining differences or different parameter selections. However, before the F1 score can be used to compare nuclei detection performance, considerations must be made to ensure accurate and valid comparisons are performed. In order examine the F1 score results for various staining patterns, it was plotted for each image and stain type against the manual PI estimates. Results are shown in Fig 4.3, where each data point represents the F1 score of each stain for a single image.

![Figure 4.3 - F1 score values for each image in the dataset. Each manual PI estimate results in two data points, one for each stain type for a particular image.](image)

Observing Fig 4.3, it is apparent that F1 scores for Ki67 are poor for images with expression levels under 5% while hematoxylin scores are fairly consistent until more significant levels of Ki67 are observed. Recall that the F1 score is generated as a function of TP, FP, and FN nuclei counts which is greatly affected by the lower sensitivity or precision values. The lower expression levels of Ki67 in the 1-5% range contain only small amounts of positively stained nuclei and may suffer from small sample sizes
when calculating an F1 score. For example, if an image contains a very low Ki67 expression level it may contain 30 positively stained nuclei and 3000 negatively stained hematoxylin nuclei. In this scenario, if 35 Ki67 nuclei and 2900 hematoxylin nuclei are detected by the automatic algorithm then the percentage difference in nuclei counts is ~17% for Ki67 but only ~3% for hematoxylin even though there was a larger nuclei discrepancy for hematoxylin. Therefore if the overall F1 score is determined by simply taking the mean of every image then it would not be a representative metric for algorithm evaluation since images containing low expression levels are not necessarily good indicators of the algorithm performance. This same adverse effect may be observed for hematoxylin F1 scores at higher Ki67 levels.

In order to ensure final F1 scores are robust to the sample sizes that change from image to image, the final F1 score for each stain is composed of total nuclei counts over the entire dataset. The TP, FP, and FN counts are recorded for each image during nuclei window validation procedures and a final score is calculated based on these counts. By recording the nuclei counts for each image and summing the total the nuclei sample size represents each individually detected nuclei over the 30 image dataset ensuring a large sample size and a more robust F1 calculation. This procedure is also performed for the effective rate, sensitivity, and precision calculations to ensure validation consistency and ensure appropriate comparisons are performed.

4.2.2 - Cell Radius Estimate Performance

The cell radius estimator was implemented in order to achieve a fully automatic framework which is a major requirement in the proposed methodology. In the original nuclei detection framework a user defined estimate of the average cell radius was determined manually before nuclei detection was performed. This introduced unwanted subjectivity and variability where consistent results could not be achieved. By developing a method for automatic radius estimation, the proposed method does not rely on user input to obtain a final analysis. Recall, the automatic estimator approximates the average object size in the image as a function of area and calculates a rough radius estimate by applying the equation of a circle. The cell radius estimator generates an estimate independently for each stain type since colour separation occurs prior to nuclei detection. Therefore an automatic estimate is determined for each stain type independently on an image by image basis allowing for some algorithm flexibility where nuclei size can vary between images. To evaluate the validity of the proposed estimator, various arbitrarily defined radii were selected and applied to the dataset for a colour threshold $T=0$. Unlike the automatic method, the arbitrarily chosen values were consistent across all images in the dataset. Resultant F1 scores are observed in Fig 4.4.
The adaptive approach obtained F1 scores of 0.73/0.69 for hematoxylin and Ki67 respectively, while the best F1 scores were obtained for cell radius estimates of 2µm for hematoxylin and 3µm for Ki67. These ideal estimates resulted in an F1 score of 0.74 and 0.72 respectively. It is expected that the ideal radius estimate for each stain would differ as cancerous nuclei positively stained would on average appear larger than their healthy counterpart. Regardless, the automatic cell radius estimate is able to achieve comparable results to the ideal values associated with each stain. Despite the small performance difference, the automatic approach is preferred over the need for user defined estimates because in a clinical setting a user would not be able to run various implementations in order to find the best result. Therefore the probability that a user selects a less than ideal cell radius estimate for at least one stain type increases and as shown in Fig 4.4 where a single pixel difference in the cell radius estimate has fairly large effects on the F1 score performance.

Recall that the F1 score is generated using both sensitivity and precision where the final F1 score leans towards the lower metric. Beyond a simple F1 score comparison, sensitivity and precision ROC curves are shown in Fig 4.5 and 4.6 where sensitivity and precision are plotted for the previously selected arbitrary radius estimates. The ROC curves were generated by plotting the precision and sensitivity values used in the calculation of the F1 score for both the manually determined and automatic estimates. Where the ideal trade-off is expected to be located in the “knee” of the ROC curve.
For both ROC curves it is clear that the automatic cell radius estimator is able to achieve an almost ideal trade-off between sensitivity and precision performance. This result is consistent with the above F1 score experimentation as the automatic estimator may not achieve the most ideal trade-off, but is able to consistently score close to the ideal estimate. As a result, the ROC curves further suggest the automatic method is able to achieve more consistent results than a single subjective user defined value.

### 4.3 - Colour Separation

Typical image analysis algorithms perform pixel based comparisons when regions of interest are segmented into two classes. In this work, colour separation performance is focussed on the measured PI values obtained from the nuclei detection framework and the manual markers. The variability of histopathology images and manual counting procedures pose challenges for objective validation.
procedures because they are less exact and are prone to subjectivity and user error as discussed in previous chapters. The following section evaluates the adaptive stain thresholds by measuring the PI differences over the dataset and comparing the resultant effects on the F1 scores.

4.3.1 - Colour Separation Under Various Threshold Conditions

The difference between the automatic and manual PI estimates for each image are the defining metric for testing the overall performance of the proposed framework. The PI value is required for clinical analysis and must be comparable to the manual counts for a wide range of staining variability. Since the perceptual shifting attained through automatic threshold determinations may have a significant effect on the final PI estimates, the adaptive balanced threshold methodology is evaluated with varying threshold decision making criteria where the difference between the automatic PI and manual PI estimates for each image are determined. The chrominance threshold \( b^* = 0 \) is applied as the control implementation and three different balanced threshold iterations are evaluated to determine if application of a perceptual threshold improves algorithm performance. To incorporate similar methods, the Otsu threshold method is also applied to better evaluate the colour separation methods.

In the first adaptive iteration, the balanced threshold method is applied without the stop condition discussed in chapter 3. For this implementation, the balanced histogram threshold often resulted in a \( T > 0 \) because the threshold would oscillate and shift towards the positive \( b^* \) region while the negative \( b^* \) region was minimized during \( T \) shifts. With an average threshold greater than zero, it is possible to examine if \( T = 0 \) is an optimal base estimate despite being the numerical value in the \( b^* \) channel where there is no defined colour content.

In the second adaptive iteration, the balanced histogram correction method is applied with the stop condition. This aims to eliminate the threshold drift that occurs as the histogram approaches a final threshold when all the bins are removed. The threshold shifting effect was observed in Fig 3.8 for a TMA core image with and without the aforementioned stop condition. As a result, images that are predominantly hematoxylin will now have a balanced threshold in the negative region and the opposite should be true for Ki67. Overall, this implementation results in increased sensitivity to Ki67 staining content for images dominated by hematoxylin and decreased for those containing higher Ki67 levels.

The final adaptive implementation evaluated the application of the balanced threshold with stop condition for only those images with a final \( T \) greater than 0. For this implementation, sensitivity to Ki67 staining levels will be unchanged for images containing predominantly hematoxylin content and will decrease staining sensitivity to Ki67 for images with high Ki67 expression levels.
Finally, the Otsu method was applied to the b* chrominance channel to evaluate the effectiveness of the proposed method when compared to similar threshold techniques.

The results of the balanced histogram threshold selection for each separate implementation are observed in Table 4.1 where the final threshold for each image in the dataset is shown. It is apparent that the application of a stop condition greatly affects the final $T$ for many images. The average threshold magnitude is decreased through application of a stop condition, so the degree of perceptual threshold shifting is reduced. This limits the potential negative effects of a larger b* magnitude threshold selection because the large changes in perceptual shifting may result in poor stain separation. It should be noted that image 24 contains significant Ki67 staining over the entire image which can be observed in Fig 3.6B. After implementation of a stop condition, image 24 represents the greatest shift into the positive b* chrominance space which is expected since it contains the highest overall Ki67 staining content. Therefore after application of a stop condition, it appears the balanced threshold selection is appropriately responding to overall staining content observed in the dataset images as there are also negative thresholds associated with hematoxylin dominated images. In contrast the Otsu method results in large shifts in both the negative and positive b* chrominance directions, suggesting that the method does not perform well when the histograms are not clearly bimodal where an image lacks two large histogram distributions for each stain.

### Table 4.1 - Determined Threshold Values

<table>
<thead>
<tr>
<th>Image</th>
<th>Balanced Threshold Selection</th>
<th>No Stop Threshold</th>
<th>Stop Condition</th>
<th>Threshold for $T&gt;0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Otsu</td>
<td>2.79</td>
<td>0.94</td>
<td>0.94</td>
</tr>
<tr>
<td>2</td>
<td>Otsu</td>
<td>2.20</td>
<td>1.39</td>
<td>1.39</td>
</tr>
<tr>
<td>3</td>
<td>Otsu</td>
<td>3.55</td>
<td>-0.84</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>Otsu</td>
<td>2.45</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>5</td>
<td>Otsu</td>
<td>1.10</td>
<td>-1.24</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>Otsu</td>
<td>4.44</td>
<td>-0.50</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>Otsu</td>
<td>4.63</td>
<td>-0.60</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>Otsu</td>
<td>2.86</td>
<td>-1.89</td>
<td>0.00</td>
</tr>
<tr>
<td>9</td>
<td>Otsu</td>
<td>5.08</td>
<td>1.61</td>
<td>1.61</td>
</tr>
<tr>
<td>10</td>
<td>Otsu</td>
<td>4.36</td>
<td>-0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>11</td>
<td>Otsu</td>
<td>3.35</td>
<td>-0.39</td>
<td>0.00</td>
</tr>
<tr>
<td>12</td>
<td>Otsu</td>
<td>6.32</td>
<td>2.95</td>
<td>2.95</td>
</tr>
<tr>
<td>13</td>
<td>Otsu</td>
<td>2.10</td>
<td>-1.82</td>
<td>0.00</td>
</tr>
<tr>
<td>14</td>
<td>Otsu</td>
<td>2.69</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>15</td>
<td>Otsu</td>
<td>2.79</td>
<td>0.94</td>
<td>0.94</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Image</th>
<th>Balanced Threshold Selection</th>
<th>No Stop Threshold</th>
<th>Stop Condition</th>
<th>Threshold for $T&gt;0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>Otsu</td>
<td>3.24</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>17</td>
<td>Otsu</td>
<td>3.24</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>18</td>
<td>Otsu</td>
<td>3.24</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>19</td>
<td>Otsu</td>
<td>3.24</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>20</td>
<td>Otsu</td>
<td>3.24</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>21</td>
<td>Otsu</td>
<td>3.24</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>22</td>
<td>Otsu</td>
<td>3.24</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>23</td>
<td>Otsu</td>
<td>3.24</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>24</td>
<td>Otsu</td>
<td>3.24</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>25</td>
<td>Otsu</td>
<td>3.24</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>26</td>
<td>Otsu</td>
<td>3.24</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>27</td>
<td>Otsu</td>
<td>3.24</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>28</td>
<td>Otsu</td>
<td>3.24</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>29</td>
<td>Otsu</td>
<td>3.24</td>
<td>1.01</td>
<td>1.01</td>
</tr>
<tr>
<td>30</td>
<td>Otsu</td>
<td>3.24</td>
<td>1.01</td>
<td>1.01</td>
</tr>
</tbody>
</table>
4.3.2 - PI Performance for Colour Thresholds

The PI is the desired output of the proposed methodology which can be used to assist a pathologist in making more informed prognostic predictions. Therefore it is important that the parameter framework that results in the most consistent and accurate PI is determined. To directly compare the PI estimates for each colour separation threshold, the overall mean PI difference between the automatic and manual data was compared for each colour threshold method across and arbitrary range of cell radius estimates. The results of the current experimentation are summarized in Table 4.2 with ideal scores highlighted in yellow and the best score highlighted with green.

Table 4.2 - Proliferation Index Performance Summary for Colour Separation Thresholds

<table>
<thead>
<tr>
<th>Cell Radius Estimate</th>
<th>T=0</th>
<th>Otsu</th>
<th>Rolling No Stop</th>
<th>Rolling Stop</th>
<th>Rolling Stop (T&gt;0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptive</td>
<td>3.72</td>
<td>11.15</td>
<td>4.15</td>
<td>3.92</td>
<td>3.25</td>
</tr>
<tr>
<td>2</td>
<td>4.58</td>
<td>9.18</td>
<td>3.76</td>
<td>4.01</td>
<td>3.52</td>
</tr>
<tr>
<td>2.5</td>
<td>4.39</td>
<td>8.59</td>
<td>3.73</td>
<td>3.05</td>
<td>3.50</td>
</tr>
<tr>
<td>3</td>
<td>3.87</td>
<td>8.03</td>
<td>3.17</td>
<td>3.38</td>
<td>2.98</td>
</tr>
<tr>
<td>3.5</td>
<td>3.61</td>
<td>7.43</td>
<td>3.24</td>
<td>3.12</td>
<td>2.83</td>
</tr>
<tr>
<td>4</td>
<td>3.78</td>
<td>7.71</td>
<td>2.76</td>
<td>3.49</td>
<td>3.07</td>
</tr>
</tbody>
</table>

Observing the values in Table 4.2, it is apparent that for PI determination the ideal threshold is the balanced histogram approach including the rolling stop condition method and T>0 for all scenarios, with one exception. It is unclear as to why the arbitrarily selected radius estimate of 4µm achieves the best PI result, however the rolling stop method for T>0 is by far the most consistent at determining a PI estimate consistent with the manually determined ground truth. Considering the fact that the goal of the algorithm is to automatically determine an accurate PI estimate by considering the perceptual colour characteristics of the human visual system, it can be suggested that the rolling stop threshold T>0 is the ideal threshold for colour separation of Ki67 and hematoxylin stains. The T>0 performance indicates that it is advantageous to only consider the perceptual shifting effect into the positive b* region for images containing high Ki67 concentrations. This may be due to the fact that the hematoxylin staining content is much less variable and consistent compared to the wide range of variability present in the Ki67 staining content. Since detection of positive Ki67 staining content is determined by the presence of brown staining, a pathologist shifts their criteria in over-stained images to account for a consistent brown understaining across the image.

4.3.3 - F1 Score for Colour Separation

In addition to nuclei detection, the F1 score can be used to indirectly evaluate the performance effects of the various colour separation thresholds. The F1 score was compared for each colour threshold
for the same arbitrarily selected cell radius estimates. Depending on the colour threshold method applied, the F1 score may increase or decrease since nuclei may be shifted from one class to another if the perceptual shift in each thresholding method is significant. Therefore the F1 performances must be considered for each threshold criteria. The resultant F1 scores for hematoxylin and Ki67 are observed in Fig 4.7 where each column represents the F1 score obtained for a specific cell radius estimate and colour separation threshold. Observing Fig 4.7, the automatic cell radius estimate obtains comparable results to the higher performing arbitrary values in the 2-4µm range. It can also be observed that the colour separation thresholds presented in this work obtain better F1 scores than Otsu's thresholding method. Since the perceptual shift resulting from the Otsu's method were larger than the rolling methods, it had a greater negative effect on the F1 scores because more nuclei were shifting from one stain type to the other.
Figure 4.7 - F1 scored are compared for both Ki67 and hematoxylin stain content using the automatic cell radius estimator (Adaptive) and various arbitrarily selected radius estimates for the colour separation threshold methods.

The F1 score results are further summarized in Table 4.3 which identifies the best results obtained for each colour separation threshold in yellow as well as the best total result highlighted in green. While the best results were obtained by an arbitrary radius estimate, the automatic cell radius estimator was able to achieve the second highest result in each scenario. This further supports the idea that any manually selected radius outside of the ideal value would result in lower scores than the automatic method, which is able to objectively achieve almost ideal results.
Table 4.3 - F1 Score Summary for Colour Separation Conditions

<table>
<thead>
<tr>
<th>Cell Radius Estimate</th>
<th>T=0</th>
<th>Otsu</th>
<th>Rolling No Stop</th>
<th>Rolling Stop</th>
<th>Rolling Stop [T&gt;0]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptive</td>
<td>0.69</td>
<td>0.63</td>
<td>0.68</td>
<td>0.75</td>
<td>0.69</td>
</tr>
<tr>
<td>2</td>
<td>0.62</td>
<td>0.61</td>
<td>0.62</td>
<td>0.71</td>
<td>0.63</td>
</tr>
<tr>
<td>2.5</td>
<td>0.64</td>
<td>0.61</td>
<td>0.63</td>
<td>0.71</td>
<td>0.64</td>
</tr>
<tr>
<td>3</td>
<td>0.72</td>
<td>0.71</td>
<td>0.71</td>
<td>0.76</td>
<td>0.72</td>
</tr>
<tr>
<td>3.5</td>
<td>0.68</td>
<td>0.66</td>
<td>0.67</td>
<td>0.72</td>
<td>0.68</td>
</tr>
<tr>
<td>4</td>
<td>0.69</td>
<td>0.67</td>
<td>0.68</td>
<td>0.71</td>
<td>0.69</td>
</tr>
</tbody>
</table>

The validity of the colour separation methods can also be examined as the ideal colour separation threshold for Ki67 is the balanced threshold with the stop condition. Alternatively, for hematoxylin the balanced threshold with no stop condition achieves the most consistent results. Referring back to the threshold values in Table 4.1, these results are partly due to the average overall perceptual shift determined by each respective method. Recall that the perceptual threshold shift that occurs for the balanced method with no stop condition results in an average $T$ in the positive $b^*$ region. This shift results in more nuclei content in the negative hematoxylin region and as a result a higher F1 score is achieved for hematoxylin, but a lower F1 score is achieved for Ki67. Similarly, the rolling stop condition achieves the ideal result for the Ki67 content because more $T$ values are located in the negative hematoxylin region which increases nuclei representation in the Ki67 colour separated image and decreases it for the hematoxylin colour separated image. This observation is important as the rolling stop method applied for $T>0$ achieves second place results in both the hematoxylin and Ki67 stains suggested it is a more consistent approach that performs a subtle perceptual shift that is not favourable for only one stain type. It also outperforms the natural $T=0$ threshold suggesting that implementation of an adaptive threshold is warranted for increased and more consistent F1 score performance.

4.3.4 - Effective Rate for Colour Separation Thresholds

Beyond the F1 scores, the effective rate was also determined for the same colour separation conditions. Recall that the effective rate is a measure of nuclei count consistency between the automatic and manual counts where an ideal measure is 1 which indicates a complete agreement between the total
number of nuclei counted for each stain type. A effective rate greater than 1 indicates that the automatic method identified more nuclei than the manual counts and an effective rate less than one indicates less nuclei identified. Final results are summarized in Table 4.4 where the ideal values are highlighted using the same criteria listed for Table 4.2.

Table 4.4 - Effective Rate Summary for Colour Separation Conditions

<table>
<thead>
<tr>
<th>Cell Radius Estimate</th>
<th>T=0</th>
<th>Otsu</th>
<th>Rolling No Stop</th>
<th>Rolling Stop</th>
<th>Rolling Stop (T&gt;0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptive</td>
<td>1.094</td>
<td>1.351</td>
<td>1.001</td>
<td>1.115</td>
<td>1.070</td>
</tr>
<tr>
<td>2</td>
<td>1.268</td>
<td>1.373</td>
<td>1.097</td>
<td>1.263</td>
<td>1.229</td>
</tr>
<tr>
<td>2.5</td>
<td>1.456</td>
<td>1.591</td>
<td>1.281</td>
<td>1.448</td>
<td>1.414</td>
</tr>
<tr>
<td>3</td>
<td>1.023</td>
<td>1.127</td>
<td>0.913</td>
<td>1.027</td>
<td>1.004</td>
</tr>
<tr>
<td>3.5</td>
<td>1.220</td>
<td>1.356</td>
<td>1.114</td>
<td>1.221</td>
<td>1.196</td>
</tr>
<tr>
<td>4</td>
<td>1.070</td>
<td>1.135</td>
<td>0.935</td>
<td>1.021</td>
<td>1.002</td>
</tr>
</tbody>
</table>

Unlike the F1 score, there is no clear colour separation method that achieves ideal results over the entire range of arbitrary values. Instead, ideal effective rates are obtained for either one stain or the other where no parameter conditions result in ideal results for both stain types. As a result it is difficult to directly compare the results of the effective rate calculation, however some observations can still be concluded. Similar to the F1 scores, we can observe that the rolling stop method with the \( T>0 \) condition achieves consistent results for both stain types despite not getting the best results. For example, using the adaptive cell radius estimator and the rolling stop \( (T>0) \) data, effective rates of 1.070 and 1.064 are achieved for Ki67 and hematoxylin effectively. Since final PI estimations are dependent on nuclei counts it is important that the proportion between nuclei counts remain similar or consistent. If this is attained, then the PI values should be more similar to the expected results even if overall nuclei counts are higher or lower since the proportional difference between the stained nuclei is the same. This observation can be explored in the following experiment that examines the final PI estimates for each colour separation criteria.

4.3.5 - Threshold Performance Summary

Table 4.4 examines the F1 score, effective rate, and mean PI difference between the automatic and manual results for each balanced histogram implementation using the automatic cell radius estimator.
Observing the F1 scores, there appears to be minimal effects to the nuclei detection performance when various stain separation thresholds are applied. This may occur because the stain separation shifting slightly increases nuclei performance for some images and decreases for others, resulting in a minimal total difference. The effective rate however is more significantly affected by the threshold selection. Recall that the effective rate models the difference between automatic and manual nuclei counts where a value of 1.04 refers to 4% more nuclei detected using the automatic method over the manual. Also recall that the effective rate may be considered for stain separation evaluation because if the effective rate is significantly different for Ki67 and hematoxylin it suggests the proposed framework may consistently detect more nuclei in one stain type, affecting final PI determinations. The difference in the mean effective rates is smallest for the balanced threshold implementation for T>0. This suggests that while the automatic detection method may consistently generate ~7% more nuclei, the final PI estimate will be less affected by the quality of the nuclei detection performance between stain images because the effective rate is similar for both stains.

The most obvious indicator of stain separation performance is the mean PI difference between the automatic and manual estimates. The final PI is determined by the algorithm for each image using the nuclei counts and the difference between the manual ground truth data can be calculated. The smallest mean PI difference is achieved by applying the balanced threshold only for a final T>0. This criteria suggests that a pathologist perceptual shifting occurs only for images containing high Ki67 content. Since the hematoxylin content is observed in a more consistent region across the negative portion of the b* channel, application of a threshold shift into this region does not improve classification of staining content. Conversely there seems to be merit when considering application of a positive b* threshold in the more variable and wide ranged Ki67 staining region.

Table 4.5 - Performance Metrics for Balanced Threshold Conditions

<table>
<thead>
<tr>
<th>Otsu</th>
<th>T=0</th>
<th>T=Adaptive, No Stop</th>
<th>T=Adaptive, Stop</th>
<th>T=Adaptive Stop for T&gt;0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hem</td>
<td>Ki67</td>
<td>Hem</td>
<td>Ki67</td>
</tr>
<tr>
<td>F1 Score</td>
<td>0.68</td>
<td>0.63</td>
<td>0.72</td>
<td>0.69</td>
</tr>
<tr>
<td>Effective Rate</td>
<td>0.899</td>
<td>1.351</td>
<td>1.094</td>
<td>1.036</td>
</tr>
<tr>
<td>Mean PI Difference</td>
<td>11.15</td>
<td>3.72</td>
<td>3.58</td>
<td>3.92</td>
</tr>
</tbody>
</table>

4.4 - Pearson Coefficient and Kappa Similarity

The Pearson coefficient $r$ is a measure of linear correlation between data. In this work it is applied to evaluate the correlation between the automatic and manual PI estimates. Recall an ideal value of 1 indicates complete linear correlation, and a value of 0 is no correlation. In Fig 4.8 the automatic PI values
are plotted against the manual equivalent in order to observe the degree of correlation between the two values. A linear regression line is fit to the data including the outlier and another is fit to the data excluding the outlier information. In each case an R squared value is observed that represents the coefficient of determination where $\sqrt{R^2} = r$ is the Pearson correlation coefficient. A correlation coefficient $r$ of 0.93 and 0.96 are determined for the outlier and non-outlier data respectively. This suggests that there is a significant positive correlation between the manual and automatic PI values despite the presence of a significant outlier. The correlation is fairly consistent because the data points are following a mostly linear relationship over a wide range of PI estimates, which means that the automatic stain estimates are consistent with the manual data for various Ki67 expression levels.

![Automatic vs Manual PI](image)

**Figure 4.8 - Automatic vs manual PI estimates where each data point represents a single image in the dataset.**

The Kappa similarity statistic was calculated by sectioning the PI estimates in the dataset into four Ki67 cut-off value subsets consisting of 0-10, 10-20, 20-30, and 30+. The Kappa represents a measure of agreeability between the algorithm and the manual estimates where a value of 1 is perfect agreement, 0 is attained for random data, and negative values up to -1 represent a less than random agreement. While some of the PI estimates for the automatic algorithm contained larger error compared to the manual values, most of the time they were within the same Ki67 cut-off ranges. The agreeableness between the manual automatic subsets are observed in Table 4.6 where each count represents the concordance between each PI method for a single image. A final Kappa value of 0.86 was calculated which signifies substantial agreement between methods. The Kappa calculation is included in this work as
a useful tool for evaluating inter-observer variability, however the sample size of the dataset bring into question the implications that may occur for large scale clinical implementation.

Table 4.6 - Comparison Table for Kappa Similarity Calculation

<table>
<thead>
<tr>
<th>Manual PI Estimate</th>
<th>Automatic PI Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ki67 Cut-off</td>
<td>0-10</td>
</tr>
<tr>
<td>0-10</td>
<td>7</td>
</tr>
<tr>
<td>10-20</td>
<td>1</td>
</tr>
<tr>
<td>20-30</td>
<td>0</td>
</tr>
<tr>
<td>30+</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
</tr>
</tbody>
</table>

Kappa 0.86

4.5 - Comparison to Other Works

The proposed framework quantifies the amount of Ki67 staining content through the estimation of a PI. Many existing methods have various forms of validation procedures that are specific to the dataset images. For example some methodologies calculate a PI estimation based on percentage of pixels separated into two stain images and the quality of stain separation is determined as a measure of image mask comparisons. In other works nuclei detection methodologies are the main focus and final PI estimates are not determined or evaluated.

The following section compares the results of the proposed framework to a very similar implementation proposed in [21] which was performed on a dataset of 30 images where nuclei counts and final PI estimates were compared to the manual data. Stain separation was achieved through a fuzzy c-means algorithm in the L*a*b* colour space and nuclei counts were generated using a k-means clustering algorithm. Since the full data table was posted in the publication a respective coefficient of correlation \( r \) was calculated to further compare results to the proposed method. In addition to their method, an automatic plug-in ImmunoRatio [76] was also applied to the images in their dataset as a performance comparison between PI values. The results of the ImmunoRatio plug-in were also used to calculate a correlation coefficient \( r \).

The original ImmunoRatio methodology consists of a pixel percentage stain estimation based on the stain separated images obtained through colour deconvolution. Before PI estimation, thresholds were applied to the stain separated images and morphological processing techniques were implemented to isolate nuclei content. The method also required calibration through application of a 50 image training set.
and was evaluated with another 50 image testing set. The linear correlation coefficient \( r \) was calculated for the manual and automatic IHC staining percentages with two outliers in the \( \sim 10-20\% \) region.

For studies performed on larger datasets, automatic methods are often compared to the manual counterpart by calculation of the Kappa similarity statistic. In [17] and [18], colour deconvolution approaches were performed in order to obtain stain separated images followed by application of a threshold for refined stain identification. The results were then compared in an agreeability matrix for image datasets of 1292 and 1320 respectively. In [17], Ki67 staining content was divided into low (0 - 49.6\%), moderate (49.7 - 56.6\%), and high (56.7 - 100\%) staining levels. However in [18], the subsets were determined by the pixel values in the stain separated images which were segmented into four equally spaced regions. The four pixel intensity regions consisted of high positive, positive, low positive, and negative subsets.

The results of the above studies are summarized in Table 4.7 along with the calculated validation metrics for the proposed method. Out of the reviewed methods, the proposed framework outperforms all other methods for the Kappa similarity statistic and achieves the smallest average difference for PI estimations. Since the Kappa statistic functions optimally on large datasets, the Kappa value in this work may not be sufficient to claim that the proposed method performs better than others but it can be suggested that it would obtain competitive results if applied to a much larger bank of images. The first CD approach also applies a weighted Kappa statistic which assigns weighting parameters to each subset, which may lead to a lower overall value due to the potentially stricter evaluation criteria. The ImmunoRatio plug-in was also applied to the dataset in order to effectively compare results to the colour stain separation framework as variability between datasets may be substantial when comparing various methods. Results of this comparison are observed in Table 4.8.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean PI Difference (%)</td>
<td>6.84</td>
<td>16.13</td>
<td>/</td>
<td>/</td>
<td>3.25</td>
</tr>
<tr>
<td>Kappa</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>0.57</td>
<td>0.84</td>
</tr>
<tr>
<td>Correlation ( r )</td>
<td>0.88</td>
<td>0.79</td>
<td>( 0.98 )</td>
<td>/</td>
<td>0.93</td>
</tr>
</tbody>
</table>
The colour stain separation methodology with the histogram correction threshold achieves minimal differences when compared to the manual dataset and greatly outperforms the ImmunoRatio application. The clustering methodology obtained the best results when using the L*a*b* channel, however it seems that no consideration for colour content was made beyond identifying that the L*a*b* space obtained the best results out of those evaluated. By considering the colour differentiating potential of the b* channel, this work achieves a simple but effective method for separation of hematoxylin and Ki67 staining content without computationally expensive 3D clustering procedures. It is also more robust to staining variability as a result of the histogram correction scheme implemented for images with higher PI estimates.

Application of a nuclei detection framework instead of pixel based membership percentages allows the proposed method to generate a more exact PI estimate for images that may contain variably sized nuclei. The ImmunoRatio method applied to the dataset was unable to accurately quantify staining characteristics associated with nuclei locations. Since cancerous cells often contain enlarged structures, images containing high Ki67 content may appear to have higher Ki67 positive percentages for pixel content methods because of the enlarged cell boundaries and increased pixel area.

Table 4.8 - ImmunoRatio Vs Proposed Work on Thesis Dataset

<table>
<thead>
<tr>
<th>Proposed Method</th>
<th>ImmunoRatio</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8.32</td>
<td>3.25</td>
</tr>
<tr>
<td></td>
<td>0.22</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>0.57</td>
<td>0.93</td>
</tr>
</tbody>
</table>
Chapter 5

Summary and Conclusions

This thesis is focused on developing a fully automatic and robust framework for quantification of Ki67 and hematoxylin stained histopathology images. Clinical methods for histopathology analysis involve unrealistic and laborious manual counting procedures or inaccurate fast scanning estimation techniques. Numerous studies examining the level of agreement between these heavily standardized evaluation procedures have found various levels of inter and intra-observer variability. In order to eliminate observer bias and more objectively evaluate Ki67 stained histopathology images, computerized analysis algorithms are generally viewed as the promising solution.

Many automatic quantification methods rely on user defined parameters introducing subjectivity and observer bias into previously objective frameworks. Other methods do not consider the availability of colour information for histopathology analysis and only apply grayscale image processing techniques. The most common processing algorithms for stain separation are colour deconvolution and clustering approaches. Colour deconvolution aims to model the staining concentrations with respect to light absorption characteristics defined by the Beer-Lambert law of absorption, however some IHC stains such as Ki67 contain light scattering properties at high concentrations and result in quantification errors. CD approaches have been implemented for Ki67 quantification with good results, however reliance on pre-measured staining colour vectors increases algorithm sensitivity to staining variations. Methods that implement CD must often manually select or calculate appropriate stain vectors which introduces observer subjectivity and algorithm performance may become trained to a specific dataset. Clustering methods on the other hand are typically computationally expensive as implementation involves iterative approaches where pixels are continuously classified into shifting clusters intended to represent final stain classes.

This work considers the perceptual colour descriptive characteristics of the $b^*$ colour channel within the $L^*a^*b^*$ colour space. The human visual system is extremely effective at differentiating the variable colour content associated with multiple stain types and computerized methods still rely on manually identified ground truth data for validation. Consideration of how the HVS processes colour content, with respect to opponent colour theory, allows the proposed colour separation method to perceptually model Ki67 and hematoxylin stain content across $b^*$ channel intensities. Since the $b^*$
Chrominance channel models the amount of colour content from blue to yellow according to opponent process theory, it is able to model staining characteristics for various staining levels of Ki67 because the positive brown staining is similar to yellow content. Modelling Ki67 colour content is also difficult to perform across images as it appears manual data may contain a shift to the perception of colour content for images over stained with brown Ki67 colour content. By implementing a threshold correction methodology, the proposed work accounts for this perceptual shifting effect and describes staining content more similarly to the HVS.

Positive staining content is modelled by calculation of a PI which represents the percentage of positively stained nuclei within a region of interest or over an entire image. To obtain a robust PI estimate a nuclei detection algorithm was implemented on the colour stain separated images that represent the relative confidence levels of each stain type. A cell radius estimator was included to fully automate the framework and ensure consistent results by eliminating the requirement for user defined parameters. Calculation of a PI index can also be performed as a function of pixel areas, however staining variability can greatly influence the final result leading to unreliable metrics.

Algorithm performance was evaluated through application of various validation methodologies to examine each step in the proposed framework. Nuclei detection was evaluated through calculation of an F1 score value which models the trade-off between sensitivity and precision metrics. F1 scores were found to be inconsistent for images with one dominant stain type due to small sample sizes, so a more robust F1 score was calculated using the total nuclei counts across the entire dataset. Colour separation was examined by experimental comparisons for various balanced threshold conditions where the F1 scores, effective rates, and PI estimates between the automatic and manual data were compared. The linear correlation coefficient $r$, Kappa similarity statistic, and the average difference between the automatic and manual methods were calculated. The Kappa similarity statistic was determined by separating the dataset into four subsets of expression levels (<10, 10-20, 20-30, 30+) and examining the agreement for each image. The PI difference was simply calculated as the mean difference between the manual and automatic PI values across the dataset.

Once ideal nuclei detection and colour separation parameters were determined through experimental performance comparisons, F1 scores of 0.73 and 0.69 were obtained for hematoxylin and Ki67 staining content respectively. Challenges associated with images with a limited stain presence for Ki67 or hematoxylin contributed to lower F1 scores, however the total number of nuclei detected across the dataset was consistent for each stain with an effective score representing approximately 7% more nuclei detected by the automatic framework. The colour separation method obtained a linear correlation coefficient of 0.93, a Kappa value of 0.86, and an average PI difference of 3.25%. These metrics represent
significant linear correlation, significant agreement for the various Ki67 cut-off values, and a minimal difference between the manually obtained results. Due to the limited size of the dataset, the Kappa statistic can be interpreted as a metric showing potential for a large scale implementation, but currently the proposed framework must be implemented on a larger dataset for a more robust and definitive Kappa value. Similarly application of the histogram corrected threshold for stain separation was able to model the higher staining characteristics of this dataset, but large scale implementation is required to confirm these effects.

To better evaluate the accuracy of the proposed algorithm performance, other methods of Ki67 stain quantification were compared. The L*a*b* clustering algorithm in [21] performs validation metrics on a similar sized dataset where the performance of each image was summarized and a mean PI estimate was calculated. In the same work the automatic algorithm ImmunoRatio [76] was also applied to their dataset for comparison purposes. To generate a more direct comparison between algorithm performance in this work, the ImmunoRatio method was also implemented on the current dataset. Since histopathology image datasets may contain wide ranges of staining content and variability, application of the ImmunoRatio algorithm forms a direct comparison to the proposed colour separation method. Since both the manual and automatic PI estimates were presented in the L*a*b* clustering results, a linear correlation coefficient and Kappa statistic were calculated for a more complete comparison. The novel colour separation technique proposed in this work was able to achieve a much smaller average difference in PI estimates, a better linear correlation, and higher agreeability than the other proposed methods. Similarly, the Kappa value was found to obtain competitive results when compared with methods utilizing CD based methodologies (Kappa = 0.57 and 0.84). The CD methods reviewed however were implemented on much larger datasets making a direct comparison of Kappa values difficult.

Overall, the colour separation framework is a fully independent and automatic framework capable of modelling the wide range of staining intensities observed in Ki67 histopathology images. It achieves consistent results across intermediate staining levels typically associated with low agreement for similar manual methods. The colour separation framework also outperforms similar works while maintaining computational simplicity and objective results. This technique is a general approach to perceptual Ki67 stain quantification, but it may also be considered for other brown IHC stains where colour deconvolution sub optimal. It is believed that this work may define a new perspective for quantifying IHC staining content that may serve as a valuable alternative to existing colour separation techniques.
Bibliography


S. Ali and A. Madabhushi, “ACTIVE CONTOUR FOR OVERLAP RESOLUTION USING WATERSHED BASED INITIALIZATION ( ACOReW ): APPLICATIONS TO


