

Automated Identification of Myenteric Ganglia in
Histopathology Images for
the Study of Hirschsprung's Disease

by

Jinu Kurian

A thesis submitted to the Faculty of Graduate and Postdoctoral
Affairs in partial fulfillment of the requirements for the degree of

Master of Applied Science

in

Biomedical Engineering

Carleton University

Ottawa, Ontario

© 2020, Jinu Kurian

Abstract

Automating digital pathology processing to identify ganglia in seromuscular biopsies from patients with Hirschsprung's disease (HSCR) may be able to provide objective measures and minimize evaluation time for expert pathologists. This adjunctive tool has the potential to improve surgical success for treatment. With thirty patient images, we proposed and evaluated an image processing pipeline that identifies colon tissue structures containing myenteric ganglia. This is an undertaking that is the first of its kind. From whole slide images (WSIs) of calretinin-stained colon sections, we initially segmented the muscularis propria using a convolutional neural network (CNN) resulting in a mean inclusion of myenteric plexus of 95.96%. Then, colour thresholding identified myenteric plexus regions with a mean inclusion of ganglia of 99.2%. Finally, within these ideal search spaces, we segmented and classified ganglia candidates to achieve an overall mean precision and recall of 64.8% and 80.2%, respectively. Preliminary results encourage further development of these algorithms.

Acknowledgements

I am fortunate and blessed to have some incredible pillars of support. Thank you to my thesis supervisors, Prof. Adrian Chan and Prof. Eranga Ukwatta, for their guidance and help with my work. Thank you to Dr. Dina El Demellawy for her incredible patience with annotating, imparting clinical knowledge, and permission to work with the images in this dataset. Thank you, as well, to Dr. Joseph de Nanassy for his humorous and meticulous approach to manual annotations. I would also like to express my appreciation to Carleton University staff (including Ryan Taylor, Shawn Stansel, Heather MacDonald, and many more) whose time and contributions were vital to the completion of this thesis.

To my dear family: Ammay, Appa, for loving and always finding ways to show it; Vavs, thanks for tolerating the mountains in the living room and the gourmet dinners you let me poach; RAK, vacuum hauls for massive spiders; Ammachi & Appacha, for your prayers and generally making my heart swell with your voices.

To my community: Canadian candle, thank you for visiting just so we could be sickly together with butternut squash soup, Netflix, and photos of an apricot. Solange, thanks for grabbing my hand as I was slipping into the abyss to be eaten by bears. I always had so much fun with you around! Raincoats too long and tough love, hearts for giving it to me straight even if I didn't always appreciate it at first. My left hand (yes, yes, you're my right hand), for making smiles easy. My right hand (I'm ambidextrous, so does it matter?), always around with late-night dessert when you thought I needed it. Midnight walks and soup dumplings, for obliging to late requests. Nerd, thank you for the chance in rate control and answering endless questions. Part of the solution only clicked after I began copying you. Oberon, thank you for accepting another misfit and dealing with all that came up over the years. From ground zero until now, you have been an unwavering cornerstone. Thank you is not enough for the times I have walked away feeling a thousand times better.

Table of Contents

Abstract.....	i
Acknowledgements.....	ii
List of Tables	v
List of Figures.....	vi
List of Abbreviations	xii
Chapter 1: Introduction.....	1
1.1 Motivation & Thesis Objectives	2
1.2 Contributions.....	4
1.3 Thesis Outline	6
Chapter 2: Background & Literature Review.....	7
2.1 Clinical Problem.....	7
2.2 Anatomical Background.....	10
2.3 Histopathologically-Defined Colon	13
2.4 Characterizing HSCR via histopathological image processing methods	21
2.5 Conclusions	26
Chapter 3: Clinical Data Set and Manual Annotations.....	27
3.1 Study Subjects and Image Acquisition.....	27
3.1.1. Tissue Preparation prior to Image Acquisition	27
3.1.2. Histopathological Stain.....	28
3.1.3. Image Data	31
3.2 Manual Annotations	33
3.2.1 Staining Errors and Irregularities.....	37
3.2.2 Manual Segmentation Criteria for Regions of Interest.....	39
3.3 Conclusions	43
Chapter 4: Evaluation of Muscularis Propria Segmentation Algorithms	44
4.1 Methods.....	45
4.1.1 k-means Clustering Segmentation	45
4.1.2 Convolutional Neural Network (CNN)-based Segmentation	47
4.1.3 Algorithm Evaluation Methods.....	50
4.2 Results	53
4.3 Analysis & Discussion	62

4.4	Conclusion.....	65
Chapter 5:	Segmentation of Myenteric Plexus Regions	66
5.1	Methods.....	67
5.1.1	Colour Segmentation in RGB Space.....	67
5.1.2	Algorithm Evaluation Methods.....	71
5.2	Results	74
5.3	Analysis & Discussion	76
5.4	Conclusion.....	83
Chapter 6:	Identification of Ganglia	84
6.1	Methods.....	85
6.1.1	Ganglia Candidate Selection.....	85
6.1.2	Feature Selection.....	88
6.1.3	Classification.....	91
6.1.4	Algorithm Evaluation Methods.....	93
6.2	Results	95
6.3	Analysis & Discussion	101
6.3.1	Performance of Ganglia Candidate Segmentation	101
6.3.2	Ganglia Candidate Classification.....	104
6.3.3	Overall Ganglia Identification	109
6.3.4	Inter-observer Agreement	111
6.4	Conclusion.....	113
Chapter 7:	Pipeline Performance for Identifying Myenteric Ganglia	114
Chapter 8:	Summary & Conclusions	120
8.1	Main Conclusions.....	120
8.2	Study Limitations	121
8.3	Future Work	122
8.3.1.	Improving Search Spaces.....	123
8.3.2.	Improving Ganglia Identification within Myenteric Plexus Regions	123
8.3.3.	Expanding Pipeline Output to Include Clinically Relevant Features	123
References.....		124

List of Tables

Table 3.1: The image content within an SVS file for this dataset.	33
Table 3.2: High-level Information of Image Dataset.	36
Table 3.3: Ganglia identification criteria for calretinin-stained histopathological Images, as described by expert pathologist, Observer One.	42
Table 4.1: Evaluation Metrics from k -means Clustering Segmentation.	56
Table 4.2: Evaluation Metrics from CNN-based Segmentation.	59
Table 5.1: Plexus Segmentation Results.	75
Table 6.1: Comparing Feature Selection Methods.	89
Table 6.2: Equations used to determine Classification Metrics.	95
Table 6.3: Ganglia Detection Rate from the Segmentation Stage of Ganglia Candidates	98
Table 6.4: Classification Metrics reported as percentages for Precision, Recall, Matthew's Correlation Coefficient (MCC), and Area under the Receiver Operator Curve (AUROC). Skew is the class ratio comparing number of negative candidates to positive candidates. Confidence intervals (as indicated by '±') were derived over $n = 1000$ bootstrap iterations.	99
Table 6.5: Overall performance of Ganglia identification for a single bootstrap iteration.	100
Table 6.6: Total ganglia in subset of four images by three observers.	111
Table 7.1: Summary of Pipeline Performance for the Identification of Myenteric Ganglia. Stage 1 and Stage 2.	115
Table 7.2: Summary of Pipeline Performance for the Identification of Myenteric Ganglia. Stage 3.	116

List of Figures

Figure 1.1: Proposed image processing pipeline	3
Figure 2.1: (a) Healthy colon: Proximal Colon includes the Ascending Colon and Transverse Colon. The Distal Colon refers to the region including the Descending Colon, Sigmoid Colon, and Rectum; (b) Functional obstruction and inflamed colon affected by Hirschsprung's disease	8
Figure 2.2: “Layers of GI tract”, Anatomy & Physiology, Openstax, Rice University, [2013] [21]	11
Figure 2.3: “A diagram of the transition zone illustrates the distribution of ganglion cells and extrinsic innervations from the proximal (PROX) euganglionic bowel to the distal (DIST) aganglionic segment. (A) Longitudinal section of full-thickness bowel wall, including muscularis propria (mp), submucosal (sm), and mucosa (mu). (B) Illustration of asymmetric circumferential distribution of ganglion cells at different points along the transition zone. (C) Transverse sections at different sites along the segment of bowel illustrated in (B) to show regional and partial circumferential nature of various neuropathological findings” [5] {Reused with Copyright Permission}	16
Figure 2.4: a) 3-dimensional reconstruction of myenteric plexus (green) in a normal colon (control case). An extrinsic nerve (blue) is shown joining the plexus. The upper portion denotes the proximal side of the intestine; the lower part denotes the distal side. Bar = 500µm. b) Myenteric plexus in a normal ileum. Note that the mesh is markedly coarser than in the colon. c) Patient case 1 – the aganglionic zone of the sigmoid colon. Only hypertrophic extrinsic nerves (blue) form a sparse network. d) Patient case 1 –the transitional zone contiguous to the aganglionic zone. Some isolated ganglia are seen (arrows). e) The transition zone 1 cm proximal to the area displayed in (d). Compared with the normal ileum from (b), the fibers are strikingly hypoplastic. Hypertrophic extrinsic nerves also seen and join the myenteric plexus (arrows). f) Patient case 2 – the transition zone of the sigmoid colon. Hypoplastic bundles form an irregular structure, as in patient case 1. [43] {Reused with Copyright permission}	19
Figure 2.5: “Pixel count by image processing and analysis. The top row shows unedited captured images of colonic mucosa with calretinin immunostaining at ×200. The middle row shows segmentation. The submucosa was subtracted by an editing program. The	

space and crypt are segmented and the mucosa remains as the area of black. The bottom row shows brown-colored pixels within the mucosa.” [52] {Reused with Copyright Permission} 23

Figure 3.1: Visualizing Plexus Regions and Ganglia under H&E and AchE IHC stain. These images are from two different sources, and not directly comparable to each other.

a) Enteric nerve plexus showing ganglion cell in H&E stain. Magnification (x400) [56].

b) AChE stain at high power magnification showing two ganglia (arrows) in the submucosa. These are negatively stained nuclei of the ganglion cell [57]. These are Open Access Licensed images. 29

Figure 3.2: Calretinin Stain highlight myenteric plexus regions and ganglion from surrounding muscularis propria at 15.4x apparent magnification 30

Figure 3.3: Colon cross-section with portions spread across the slide 32

Figure 3.4: Annotation layers for one of the images in the dataset. Green ink outlines the muscularis propria; magenta contours myenteric plexus; yellow are ganglia with high certainty; cyan are refer to objects with low certainty of being classified as ganglia 34

Figure 3.5: Image 4 presents false positive staining errors of Calretinin 37

Figure 3.6: Image 9 has minor staining errors localized to a small region of the muscularis. Manual annotations help to distinguish the staining error from the myenteric plexus regions. 38

Figure 3.7: Image 13 does not contain staining errors. Rather, we see almost no expression of brown chromogen to highlight myenteric plexus regions. 38

Figure 3.8: Manual annotations for myenteric plexus simply outlines the region 40

Figure 3.9: Ganglia, as identified by expert pathologist, Observer One. Yellow contours are ganglia with high degree of certainty, whereas cyan contours indicate low certainty.

a) reveals an object classified as potentially a ganglion in cyan, due to its relatively small size and blurry edges b) despite the lighter intensity, expert pathologist still marks a large, smooth object as a ganglion c) although clustered together, these two objects are separated by a thin fine line, and therefore marked as two distinct ganglia 42

Figure 4.1: The thumbnail for image 9 is shown in the right bottom corner. The green box indicates the region zoomed to 400µm scale for the labelled snapshot. The manual annotations of the muscularis propria are shown by green delineations. The dotted black

line highlights the intermuscular septum. Note the difference in granularity between the two muscle layers due to opposing orientation of the blue-stained support cells.....	44
Figure 4.2: Depiction of k-means clustering segmentation method. Operations are reported in pixels at the downsampled resolution of (2 μ m/pixel)	47
Figure 4.3: Architecture of CNN model [12].....	49
Figure 4.4: Indicating level of class-imbalance for identifying the muscularis propria, based on manual segmentations.....	53
Figure 4.5: Comparing spread of satisficing and optimizing metric results between the two segmentation methods.....	55
Figure 4.6: Visualizing the Muscularis Propria k-means Segmentation for images 1-15. Black outlines indicate the manual delineations of the muscularis propria, whereas the transparent blue shading depicts the segmentation by the k-means algorithm. Regions highlighted in red are myenteric plexus not captured by the algorithm's segmentation ..	57
Figure 4.7: Visualizing the Muscularis Propria k-means Segmentation for images 16-30. Black outlines indicate the manual delineations of the muscularis propria, whereas the transparent blue shading depicts the segmentation by the k-means algorithm. Regions highlighted in red are myenteric plexus not captured by the algorithm's segmentation. .	58
Figure 4.8: Visualizing the Muscularis Propria CNN Segmentation for images 1-15. Black outlines indicate the manual delineations of the muscularis propria, whereas the transparent blue shading depicts the segmentation by the k-means algorithm. Regions highlighted in red are the plexus regions not captured by the algorithm's segmentation.	60
Figure 4.9: Visualizing the Muscularis Propria CNN Segmentation for images 16-30. Black outlines indicate the manual delineations of the muscularis propria, whereas the transparent blue shading depicts the segmentation by the k-means algorithm. Regions highlighted in red are the plexus regions not captured by the algorithm's segmentation.	61
Figure 4.10: Depicting the raw output from the k-means clustering (k = 4) for five images with poor segmentation results. Shaded regions shows structures which fall into a cluster. The light red shading indicates the cluster with the second largest area.	63
Figure 5.1: Red contours outline some myenteric plexus regions magnified to 100 μ m scale from image 9. These regions spotted fibrous dendritic tissue, light blue glial cells, and ganglia.	66

Figure 5.2: Munsell chips used as materials in most color-naming studies, obtained from [80].	67
Figure 5.3: Flowchart of Myenteric Plexus Segmentation method	70
Figure 5.4: Increasing the Minimum Size Filter from 20-pixel ($20\mu\text{m}^2$) to 60-pixel ($60\mu\text{m}^2$) objects by increments of five pixels. This plot shows average precision and average recall across thirty images at each size filter. The z-axis plotted average ganglia inclusion, but remained relatively consistent at 97-98%, and therefore was not shown here. All thresholds are based on downsampled images.	71
Figure 5.5: Evaluation cases for myenteric plexus segmentation. Generally, we count a true positive (TP) detection if the Intersection overlap is at least $\text{Th} = 30\%$. If there is no overlap, then we may call this a false negative (FN) or false positive (FP). Counts for edge cases are also illustrated.	73
Figure 5.6: Visualizing the spread of results for plexus segmentation in the dataset. The smaller subfigure visualizes the segmentation results with respect to only the optimizing metrics.	74
Figure 5.7: Image 22. Orange highlights the manual segmentation. The dark blue outlines are the computer-generated plexus segmentations. Green stars highlight included ganglia.	76
Figure 5.8: Segmentations with low recall (a) Plexus Segmentation Results of Image 6. Orange highlights the manual segmentation. The dark blue outlines are the algorithm-generated plexus segmentations. Green stars highlight included ganglia. (b) This is a magnified segment from the window in a). (c) This is a magnified segment of the original image at 100- μm resolution from the window in b), with manual delineations outlining the plexus regions. The faint colour of the plexus is barely detectable to the human eye. d) Manually outlined plexus regions from image 7. e) Manually outlined plexus regions from image 13.	78
Figure 5.9: Image 3. Blue manual delineations outline the muscularis propria. The red box highlights a submucosal plexus region within the manual segmentation of the muscularis propria, containing five ganglia. The automated plexus segmentation detected this submucosal plexus, an error that will add to the false positive count in the final step of the image-processing pipeline.	81

Figure 5.10: Image 26. (a) Image of the plexus segmentation results. Orange highlights the manual segmentation. The dark blue outlines are the computer-generated plexus segmentations. Green stars highlight included ganglia. (b) The original image. (c) Magnified to 600um scale bar resolution from the blue window in b). Plexus regions are outlined, but staining errors of folds and the background muscularis are also visible. These cause the excessive false positive detection as shown in a) 82

Figure 6.1: a) Image 9 from dataset b) Red bounding box contains a myenteric plexus region magnified to 100µm scale resolution c) Digitally magnified, the yellow bounding box contains a single ganglion cell. 84

Figure 6.2: Histogram of Ganglia Sizes Sampled from Manual Segmentations Across the Dataset..... 86

Figure 6.3: Flowchart depicting Ganglia Candidate Detection Post-processing Steps 87

Figure 6.4: Visualizing Spread of Data for Selected Features in Weka. Red indicates datapoints belonging to ganglia class of high certainty; cyan is ganglia class of low certainty; dark blue are false candidates 90

Figure 6.5: Visualizing the spread of ganglia candidate classification results for this dataset. Size of datapoint is reflective of the relative confidence level over n=1000 iterations..... 97

Figure 6.6: Image 20 with poor segmentation recall and precision. Green outlines indicate missed manual (M) segmentations. Red (manual segmentations) and blue (algorithm segmentations) outlines indicate ganglia candidates that have been detected. 103

Figure 6.7: Image 28 where a merge error exists. Red outlines are manual segmentations. Blue outlines are the algorithm's segmentation. Missed manual segmentations include object #7M and #10M. Merged object is #103. 103

Figure 6.8: Image 9 where the classifier performed with high recall and high prevalence-corrected precision. However, poor precision and MCC may be due to the heavy staining of myenteric plexus regions. Blue circles indicate false positives. Green circles are true positives, as according to manual labels. The bottom window shows two false positives arising from a tissue fold in the upper corner. 105

Figure 6.9: Green circles are true positives, blue circles highlight false positives, and red circles are false negatives as according to manual labels. Purple circles highlight missing

segmentation errors in magnified windows. Note that these circles simply act as identifiers and do not show the actual segmentation, simply for easier viewing. a) image 20 – high classification performance; b) Image 24 – low classification recall; c) image 12 – poor classification precision and MCC scores; d) image 30 – mediocre performance all around. 108

Figure 6.10: Spread of overall performance metrics labelled according to each image in this dataset. Dotted lines at 70% & 75% precision and recall help to identify high performance in upper right quadrant..... 109

Figure 6.11: Image 14 110

Figure 6.12: Subjective differences in ganglia identification. a) Ganglia identified by Observer One. b) Ganglia identified by expert Observer Two (black arrows highlight difference) 112

Figure 7.1: Ganglia Identified in Image 14 after full pipeline run. This image had excellent ganglia performance in Chapter 6, with ideal initiating search spaces. Green circles are true positives, red circles are false negatives, and blue circles are false positives. 119

Figure 7.2: In Image 30, we see an excessive number of submucosal ganglia identified (blue circles) along the circumference. Green circles indicate true positives and red circles are false negatives..... 119

List of Abbreviations

AChE – acetylcholinesterase

ADASYN – adaptive synthetic sampling method

AG – algorithm-generated

AUROC – area under the receiver operating curve

CART – classification and regression tree

CHEO – Children’s Hospital of Eastern Ontario

CMBEC - Canadian Medical and Biological Engineering Conference

CNN – convolutional neural network

DISBE – Digital Image Subtraction Blue Enhancement

ENS – Enteric Nervous System

GB – gigabytes

GI – gastrointestinal tract

H&E – hematoxylin and eosin

HSCR – Hirschsprung’s disease

HSV – hue, saturation, value

IHC – immunohistochemistry

ICC – Interstitial Cells of Cajal

JPEG – Joint Photographic Experts Group

kD – kiloDalton

kNN – k -Nearest Neighbour

LDA – Linear Discriminant Analysis

LOO – leave one out

MB – megabytes

MCC – Matthew’s correlation coefficient

MP – muscularis propria

MS – manually segmented

MSF – minimum size filter

PC – pixel count

REB – Research Ethics Board

ReLU – rectified linear unit

RGB – red, green, blue

ROI – region of interest

SMOTE – Synthetic Minority Over-sampling TEchnique

SNR – signal-to-noise ratio

SVM – support vector machine

TCA – total colonic aganglionosis

TIFF – Tagged Image File Format

TZ – transition zone

WSI – whole slide images

Chapter 1: Introduction

Hirschsprung's disease (HSCR) is a congenital condition describing the underdevelopment of the enteric nervous system (ENS), typically within the large intestine. It is characterized by an absence of nerve cells (referred to as ganglia) in colonic tissue, the lack of which prevents innervation for bowel movements [1]. This can lead to symptoms of constipation, distending stomachs, enterocolitis, and death [2]. Affecting 1 in 5000 live births ([3], [4]), HSCR is the highest incident gut motility disorder among infants, with treatment requiring a morbid pull-through surgery to remove and bypass the diseased part of the colon [1]. Pathologists are consulted with serial biopsy samples throughout and after treatment in attempts to ensure the effective removal of the diseased colon. However, enigmatic standards in assessing histopathological images of colonic seromuscular samples poses subjectivity in these examinations [5]. This subjectivity undoubtedly introduces a source of error and is suspected of contributing to failures in the surgical treatment of HSCR; namely with residual aganglionic bowel [6]. Post-operative complications can include enterocolitis (in 17-50% of cases), incontinence (for 1%), and reoccurring constipation (for 10%) [7]. Nearly 10% of patients require colostomy [8], and 10-20% require a redo-surgery as described by studies undertaken in [2], [9], [10]. Devising a standardized scoring rubric for the histopathology assessment requires further research and examination of numerous colon segments, for which manual evaluation remains time-consuming and subjective. Therefore, we propose an adjunctive tool for the digital pathology assessment using image processing methods, to enable further knowledge expansion and computer-aided standardization of the seromuscular evaluation for HSCR.

1.1 Motivation & Thesis Objectives

Pathologists acknowledge that a method to standardize the evaluation of colon cross-sections affected by Hirschsprung's disease may improve surgical outcomes for patients [5]. This thesis involves the research and development of computer image analysis methods to objectively inspect histopathological whole slide images (WSIs) of the colon. We are examining the feasibility of identifying and characterizing myenteric ganglia, visible within digitized scans of Calretinin-stained colon tissue specimens and responsible for innervating the intestinal muscle. Potential impact includes reducing subjectivity in future HSCR seromuscular assessments by enabling pathologists to establish objective criteria for "healthy" colon tissue. This tool could not only save time but also improve surgical success for patients undergoing treatment for HSCR, and therefore, improve their quality of life.

In large histopathology scans of the colon, only a small portion of the image contains relevant data to ganglionosis. By limiting processing to regions with meaningful data, we reduce computational load, evaluation time, and false positives-type errors. For this purpose, we proposed an image processing pipeline that begins by segmenting 1) the muscularis propria, that contains 2) myenteric plexus regions, encapsulating 3) ganglion cells, our primary ENS structure of interest. Figure 1.1 summarizes this pipeline while outlining chapter content for this thesis.

The objectives of the thesis are as follows:

1. Refine and evaluate existing algorithms to segment the muscularis propria, the colonic muscles containing structures of interest. The results of which act as search spaces for the second objective.
2. Develop and evaluate the performance of an algorithm to segment myenteric plexus regions, the results of which act as a search space for the third objective.
3. Develop and evaluate the performance of an algorithm to identify and characterize ganglion cells encapsulated within myenteric plexus regions.

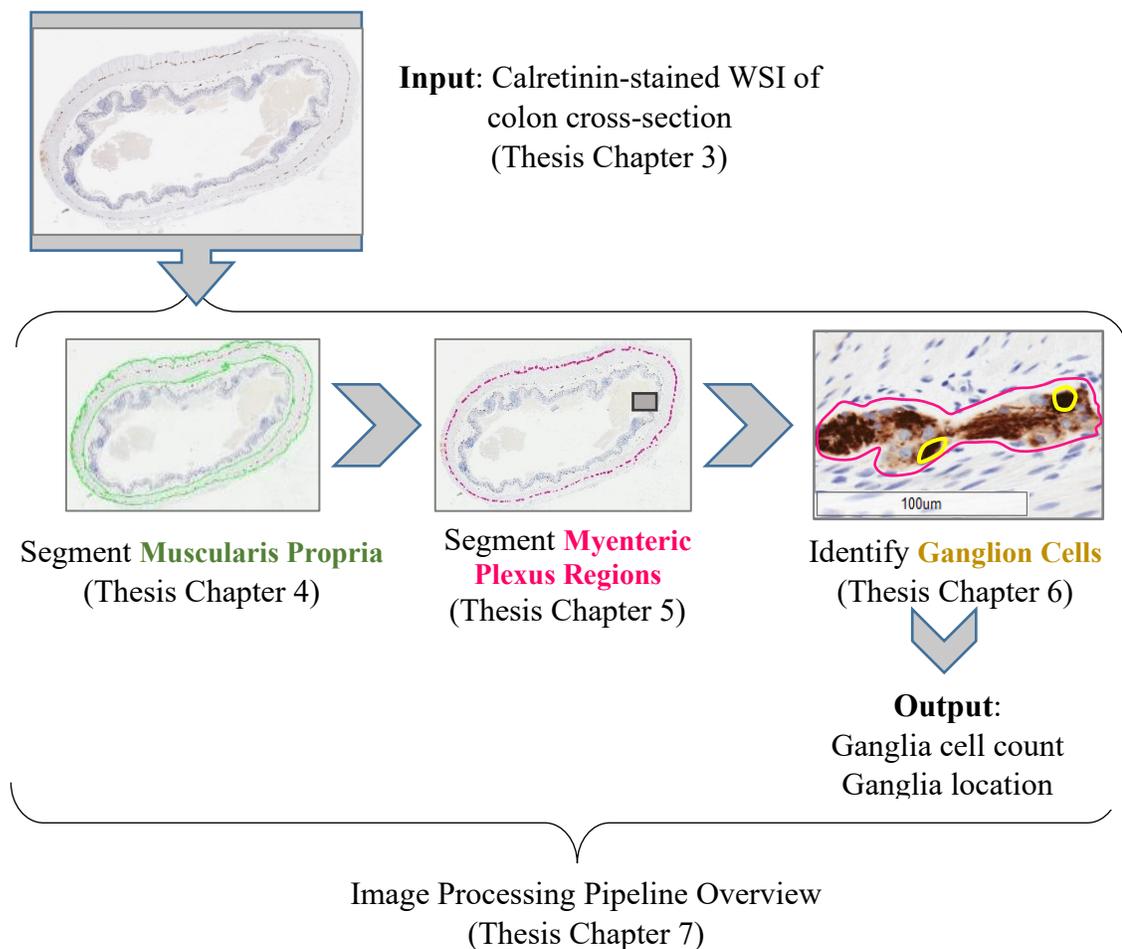


Figure 1.1: Proposed image processing pipeline

1.2 Contributions

The work presented in this thesis culminated into the four major contributions described in the following paragraphs below:

Contribution 1: Refinement of Muscularis Propria Segmentation Algorithms

Preliminary results of muscularis propria segmentation using a k -means algorithm, developed by Marco T.K. Law and Victoria Madge, were presented in [11] at the 41st Canadian Medical and Biological Engineering Conference (CMBEC41) by the thesis author. Recognizing the room for improvement, this work led to development of the muscularis segmentation using a Convolutional Neural Network (CNN)-based algorithm by Conor McKeen and Fatemeh Zabihollahy. These results were presented at the SPIE Medical Imaging Conference 2019 [12]. I modified and extended post-processing for the original algorithms to maximize myenteric plexus inclusion, while optimizing region-based precision and recall for the muscularis propria segmentation. With a plexus inclusion rate of $95.9\% \pm 5.8\%$ and a Dice score of $89.3\% \pm 4.5\%$, I implemented the modification of McKeen *et al.*'s CNN-based algorithm [12] to initialize the image processing pipeline proposed in Section 1.1.

Contribution 2: Automated Segmentation of Myenteric Plexus regions

Pathologists assess seromuscular colon sections by the presence and distribution of ganglia existing in pockets between the muscle layers of the muscularis propria. To minimize computational resources required, we narrowed the search space for ganglia to these pockets, hereby referred to as myenteric plexus regions. Using colour thresholding methods, I developed an automated segmentation of brown-stained plexus regions within

the manually labelled muscularis propria. Across the dataset, this algorithm achieved an average precision of $73.8\% \pm 16.7\%$ and an average recall of $85.9\% \pm 14.0\%$ when evaluated against manually identified myenteric plexus regions. The average ganglia inclusion rate (i.e., percentage of manually identified ganglia) was $99.2\% \pm 1.8\%$.

Contribution 3: Automated Identification of Ganglia Objects

Using an intensity-based threshold, I isolated candidate objects for ganglia from the plexus region search spaces for each image. Then, machine learning techniques were applied to identify ganglia across all candidate objects. I created a feature set for ganglia identification using expert pathologist criteria and labelled data as guidelines. Following this, I trained an LDA classifier to sort ganglia candidate objects. Using manually segmented plexus regions as the search spaces, we achieved an overall ganglia identification precision of $64.8\% \pm 21.7\%$ and average recall of $80.2\% \pm 12.2\%$ across thirty images.

Contribution 4: Expert labelled data for a large, rare dataset

Given the incidence rate of HSCR, the colon sections imaged in this dataset are collected over a time span of more than twenty years, thus creating a valuable dataset. These samples are scanned at 0.5 microns per pixel, to generate colour images of high resolution; the original images within this dataset exceed $17,000 \times 17,000 \times 3$ pixels. Due to the sheer size of histopathology images, labels generally tend to be simplistic (e.g., “HSCR-affected” or “healthy”) or limited in number. Instead, I created manual outlines at the original scanning resolution for the three major structures of interest (i.e., muscularis propria, myenteric plexus regions, and ganglia) across the entire dataset, with input for

each individual label from expert pathologists. I reviewed annotations with expert pathologists, encouraging self-assessment of their certainty with some labels. By keeping the annotations in-house, we ensured high quality labels. This database of high-quality maps allowed us to begin defining target performance for clinical use. We could also take advantage of supervised learning models where access to a large amount of labelled data is indispensable. Finally, beyond the scope of this thesis, lies future development of complex supervised learning algorithms where we aim to improve and expand upon our current results. Therein too, these labels become invaluable.

1.3 Thesis Outline

This thesis contains eight chapters. Chapter 1 briefly introduces the research purpose, states the engineering objectives in relation to the clinical problem and summarizes the contributions made towards solving this problem. Chapter 2 provides greater detail of relevant clinical background and a literature review highlighting various contributions made by global research groups in understanding and characterizing Hirschsprung's disease from both clinical and engineering perspectives. Chapter 3 presents the dataset explored in this thesis, as well as a description of the manual segmentations providing labelled data. Chapter 4 discusses the refinement of existing algorithms segmenting the muscularis propria, quantitatively determining the method most suitable as the first step in the image processing pipeline. This is followed by a description and evaluation of the myenteric plexus segmentation algorithm in Chapter 5. The automated identification of ganglia, which is the last step in the image processing pipeline, is discussed in Chapter 6. Finally, Chapter 7 summarizes the results across the entire image processing pipeline, and Chapter 8 states conclusions and areas for future work.

Chapter 2: Background & Literature Review

In this chapter I provide an overview of existing literature related to the objectives presented in Chapter 1. It begins with a brief description of Hirschsprung's disease (HSCR), the clinical diagnosis and treatment, as well as the shortcomings of current practices with a focus on the role of histopathology. From there, we explore the anatomical setting of the enteric nervous system (ENS), specifically within the large intestine. The section following will highlight some of the clinical findings pertinent for the standardization of HSCR treatment from a histopathology perspective. Finally, we will briefly review existing histopathology image processing methods for HSCR and summarize their insight for identifying affected tissue.

2.1 Clinical Problem

Among infants and children, HSCR is the most common gut motility disorder, with an incidence rate of one in five thousand live births [3], [4]. It is also the most visible and characterized congenital defect of the ENS [13]. Stunted neural development in the large intestine results in a narrowed segment of the bowel, typically affecting regions of the distal colon [7]. As shown in Figure 2.1(b) without neural-mediated propulsive motility, the bowel remains contracted and functional obstruction occurs.

Milder cases of HSCR concern older children and adults with symptoms of chronic constipation and impaired growth from malnutrition [7]. For newborns, HSCR is suspected if the infant has not passed meconium within 24-48 hours [7]. Other symptoms include vomiting and high fever. Colonic obstruction can distend the patient's abdomen, increasing the risk for enterocolitis (intestinal inflammation from bacterial infection) and fatal

perforation [14]. Immediate treatment begins with a course of antibiotics, intravenous saline, stomach decompression using a nasogastric tube, and bowel decompression via rectal catheter [8].

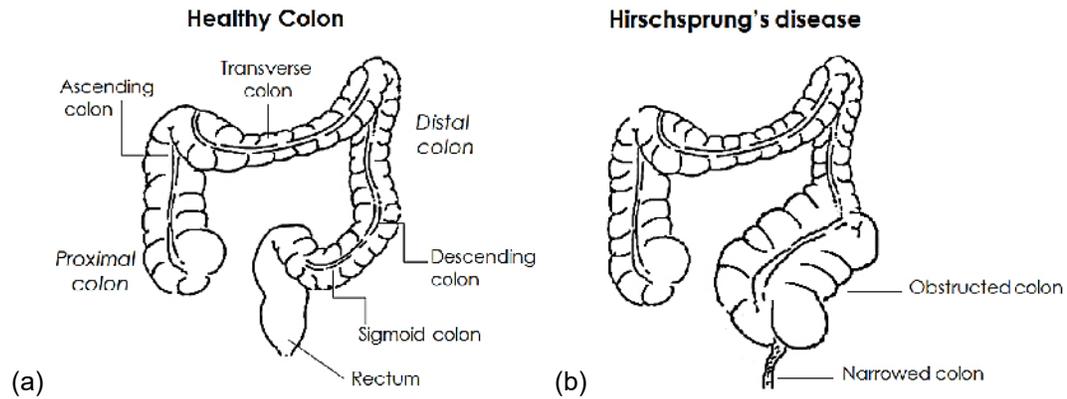


Figure 2.1: (a) Healthy colon: Proximal Colon includes the Ascending Colon and Transverse Colon. The Distal Colon refers to the region including the Descending Colon, Sigmoid Colon, and Rectum; (b) Functional obstruction and inflamed colon affected by Hirschsprung's disease

HSCR is assessed by rectal suction biopsy, where submucosal colon tissue is vacuum-sampled at distal lengths from one to five centimeters from the rectum [7]. The absence of neural cells, hereby referred to as ganglia, confirms suspicions of HSCR and provides an estimate of the length for the affected region. This is the gold standard for HSCR diagnosis. (Note: Although radiology is routinely employed, it is unreliable as a standalone modality and not an option for neonates. Only histopathological examinations can determine the extent of aganglionosis [8]). Approximately 80% of patients have short-segment HSCR, where the aganglionic segment is restricted to the rectosigmoid region [8].

Current treatment for HSCR requires a surgical pull-through procedure where the estimated region of aganglionic colon is resected, along with a minute portion of healthy ganglionated colon. The remaining purportedly functional intestine is anastomosed with the anal sphincter [7]. Pathologists perform serial frozen, or paraffin, -section seromuscular

examination (i.e., sliced biopsies of the excised colon) under microscope magnification, to document the presence of ganglion cells in the proximal margin of the resected colon. This evaluation looks to confirm the complete removal of aganglionic sections and euganglionosis within the patient's remaining intestine.

However, as mentioned in Chapter 1, enigmatic grading criteria introduces subjectivity to this clinical histopathology assessment. Time constraints and individual experience will undoubtedly influence the final judgement calls. Unfortunately, HSCR-like symptoms continue to persist post-surgery for a subpopulation of patients [2], [9]. For infants, reoperation of this highly morbid procedure is ill-advised due to the increased chances of enterocolitis, permanent colostomy, and other associated complications [15]. Morbidity and mortality risks significantly increase as aganglionosis extends past the sigmoid region, as in the cases for long-segment HSCR, total colonic aganglionosis (TCA), or even reaching the distal small intestine [16].

Subjectivity in the initial proximal margin seromuscular assessment may lead to unidentified residual aganglionic regions in the resected colon. This is the putative cause of persistent gut dysmotility in post-operative HSCR patients. Instead of euganglionic intestine, the initial proximal margin may reflect the transition zone (TZ) between histologically normal and aganglionic bowel. The TZ is anatomically defined as the region where euganglionic intestine inclines towards aganglionosis. As summarized in Section 2.3, a perusal of literature indicates that diverse histopathological indexes have been used to define the TZ; yet validated and reproducible diagnostic criteria have not been established.

Incidentally, one may question why greater lengths of the intestine are not resected, to remove the guesswork from this procedure. Unfortunately, resection of the colon involves a trade-off. Excessive removal of colon tissue can result in short intestines as the child approaches maturity. Ultra-short intestines can cause malnutrition, chronic dehydration, impaired immunity, and other concerns leading to a lower quality of life [17]. Hence, correctly identifying the transition zone region is optimal.

By examining historic records and samples for retrospective study, we may be able to help establish objective grading criteria for pathological examination of Hirschsprung's disease. Furthermore, an objective evaluation method based on automated analysis for ganglion identification may improve the outcome of surgical treatment, reduce the need for re-operation, save time, lower financial costs, and aid in the further understanding of HSCR and ENS development.

2.2 Anatomical Background

The large intestine (conjointly referred to as the large bowel, or the rectum and colon) is a major component of the digestive system and an integral element for metabolic processes in the human body. Although primarily known for its functions regarding water retention, sodium conservation and passage of waste materials for excrement, the healthy colon is also responsible for the metabolism of proteins; biotransformation of toxins and xenobiotics; systemic immune responses and more[14]. In verity, the colon serves as a substrate to a multifarious ecosystem of countless microbes, providing the largest sampling base for mass antibody production [14]. Recent studies have explored further symbiotic relations between the host and the colonic commensal bacteria via the ENS [18]–[20]. Sometimes referred to as the second brain of the human body, the ENS is a remarkably

complex structure embedded throughout the gastrointestinal (GI) tract [20]. It is partially composed of interconnected ganglia which appear as two concentric rings within the gut wall, running from the esophagus to the anus. The ENS is capable of autonomously controlling visceral functions, such as digestion and mood regulation through the release of various neurotransmitters (including serotone and dopamine) [20]. In conjunction with the vagus nerve, the ENS also controls GI function by providing innervation for peristalsis of the gastrointestinal wall [17].

Within the large intestine, the ENS processes span across all the major, histologically distinct regions: (a) mucosa, (b) submucosa, (c) muscularis propria, and (d) serosa [21]. Main components are shown below in Figure 2.2.

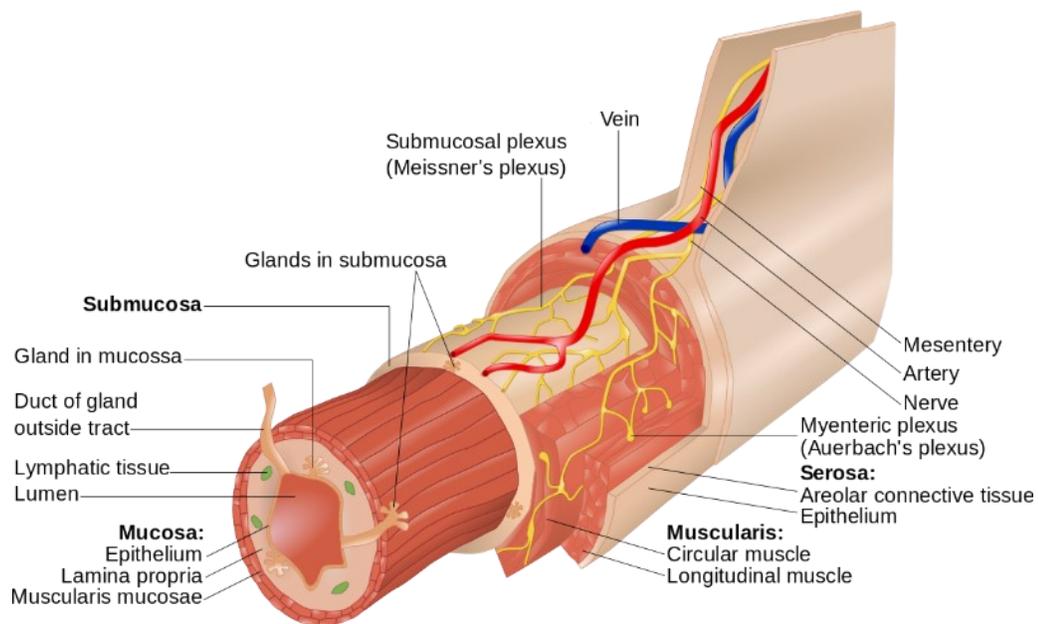


Figure 2.2: "Layers of GI tract", Anatomy & Physiology, Openstax, Rice University, [2013] [21]

Enteric ganglia (nerve cell bodies), neuropil (the fibrous dendritic network), and supportive glial cells intertwine to form plexus regions in both the muscularis propria and

the submucosa [22]. The muscularis propria consists of an inner circular muscle layer and an outer longitudinal muscle layer, with branches of myenteric plexus regions present at the intermuscular septum (the seam between the two muscle layers) [23]. The neural innervation of the muscularis propria allows for peristalsis, which are the wave-like contractions providing the propulsive forces within the colon [14].

Healthy functional bowel receives innervation both intrinsically and extrinsically. Intrinsic innervation is controlled by ganglia found in myenteric and submucosal plexus regions [24]. There are three distinct plexus structures in the GI tract [1], which can be differentiated by size and immunohistochemical coding:

- 1) The myenteric plexus (also known as Auerbach's plexus) lies within the intermuscular septum of the intestine. These plexus regions provide motor innervation to both muscle layers, as well as secretomotor innervation to the mucosa. Myenteric plexus regions have both sympathetic and parasympathetic nervous system inputs [17].
- 2) The Meissner's plexus, is found in the innermost layer of the submucosa and innervates the smooth muscle of the muscularis mucosae. It has only parasympathetic input [17].
- 3) Henle's plexus (or Schabadasch's plexus) refers to the outermost layer of plexus within the submucosa, nearest the circular muscle. It too, has only parasympathetic input [17].

Extrinsic innervation of the intestine are projections from autonomic and sensory ganglia. Extrinsic nerves penetrate the muscularis propria alongside blood vessels and merge with myenteric plexus regions [25]. Except for the terminal rectum, it is abnormal to observe more than one or two extrinsic nerves in plexus regions within a transverse colon section [25]. This thesis will focus on intrinsic innervation, retaining any findings related to extrinsic innervation as potential for future work.

2.3 Histopathologically-Defined Colon

Although Hirschsprung's disease may have a physiological impact on patients without clear morphological indications [26], euganglionic intestine is managed surgically as though it is functionally normal. As mentioned in Section 2.1 Clinical Problem, contemporary surgical methods to manage HSCR involve a pull-through surgery where aganglionic colon and the adjacent transition zone (the anatomically-defined region where intestine inclines towards aganglionosis) is removed.

The pull-through surgery makes the following inherent assumptions:

1. The transition zone (TZ) can be resolved intraoperatively
2. Inadequate removal of the TZ increases the likelihood of post-operative complications
3. The contiguous bowel proximal to the TZ is sufficiently functional [5]

In recent years, Dr. Raj P. Kapur and associates from Seattle Children's Hospital and University of Washington have published several studies describing the human bowel affected by HSCR from a pathology perspective, to correlate surgical outcome with histopathological findings. These studies include [5], [27]–[32]. Prior to their work, most reports describing the pathology of HSCR relied on subjective impressions where observer and sampling bias influenced the clinical interpretation [5]. Some of these older studies even drew conclusions based on redo-resections, without accounting for neuroanatomical changes or idiosyncratic responses from the initial surgery, such as [9]. To establish universal grading standards for the diagnosis and treatment of Hirschsprung's disease, Kapur *et al.* aimed to create well-controlled studies describing clinical features of the transition zone. Some of their findings have been summarized in this section as a reference,

and to provide a deeper understanding of the current state for clinical seromuscular assessments of HSCR.

A number of studies (including but not limited to [23], [26], [33]–[37]) also aimed to identify the TZ using novel criteria beyond ganglionosis and extrinsic innervation, such as the distribution of immunoreactive Interstitial Cells of Cajal (ICCs) and the presence of eosinophilic ganglioneuritis. Although not the focus of this thesis, we do want to acknowledge that HSCR-related symptoms can still occur despite euganglionic presence [26]. This is an uncommon phenotype (or not as well-studied), and likely occurs due to other ill-functioning elements of the ENS which are required for peristalsis. Nevertheless, this is research conducted by global experts in HSCR pathology, and their findings ([10], [17], [26], [38]) may undoubtedly prove to be useful for our future work.

To our knowledge, only a limited number of studies have been conducted thus far using histological methods with the intent to formally map the boundaries of the TZ based on ganglionosis. Three of these publications studied sub-circumferential aganglionosis. This term describes the distal end of the TZ, where ganglia along portions of the colon circumference extend further caudally than in other portions. Therefore, in a cross-section slice, some sections appear to be aganglionic while other sections appear to be euganglionic; the distribution of ganglia along the circumference noticeably varies in density.

The first two of these works include Gheradi [39] and White & Langer [40], where authors mapped the distribution of submucosal and myenteric ganglia in closely spaced, H&E-stained transverse sections. White & Langer divided the colon circumference from these transverse sections into eight sectors, counting ganglia in each sector for distances

ranging between 12-24mm proximal to the most distal ganglion cell [40]. Their publication suggests hypoganglionosis (minimal presence of ganglia) flanks aganglionic sections, but an insufficient length of proximal bowel fails to confidently establish this point. Although the H&E-stained approach used by these two groups can be applied intraoperatively to analyze frozen sections, neither of these studies were designed to discriminate healthy bowel from the proximal TZ. Nevertheless, both demonstrated segments of sub-circumferential aganglionosis up to 3cm in length.

The third publication by Swaminathan & Kapur [41] studied myenteric neuronal densities along the full colon circumference. They examined transverse sections along the length of resected short-segment specimens from six HSCR cases, using Hu C/D immunostaining. A graded increase and eventual plateau in myenteric ganglia density along the circumference was observed as transverse sections moved proximally from aganglionic colon. Amongst the small sample, the slope in changing ganglia density varied between patients, but suggested a TZ length ranging from two to five centimeters.

Unfortunately, due to inadequate controls or histopathological references for intraoperative procedures, the results of the studies described above are difficult to translate into clinical practice. To combat this, Kapur & Kennedy [5] examined the proximal margin of frozen resected colon tissue sections from HSCR cases using a conservative approach to define healthy bowel. The tissue had to exclude: obvious subcircumferential aganglionosis, hypoganglionosis, hypertrophic submucosal nerves, and intestinal neuronal dysplasia type B. Figure 2.3 is a depiction from Kapur & Kennedy (2012) of the ideal resected colon tissue, illustrating aganglionic tissue and the transition zone.

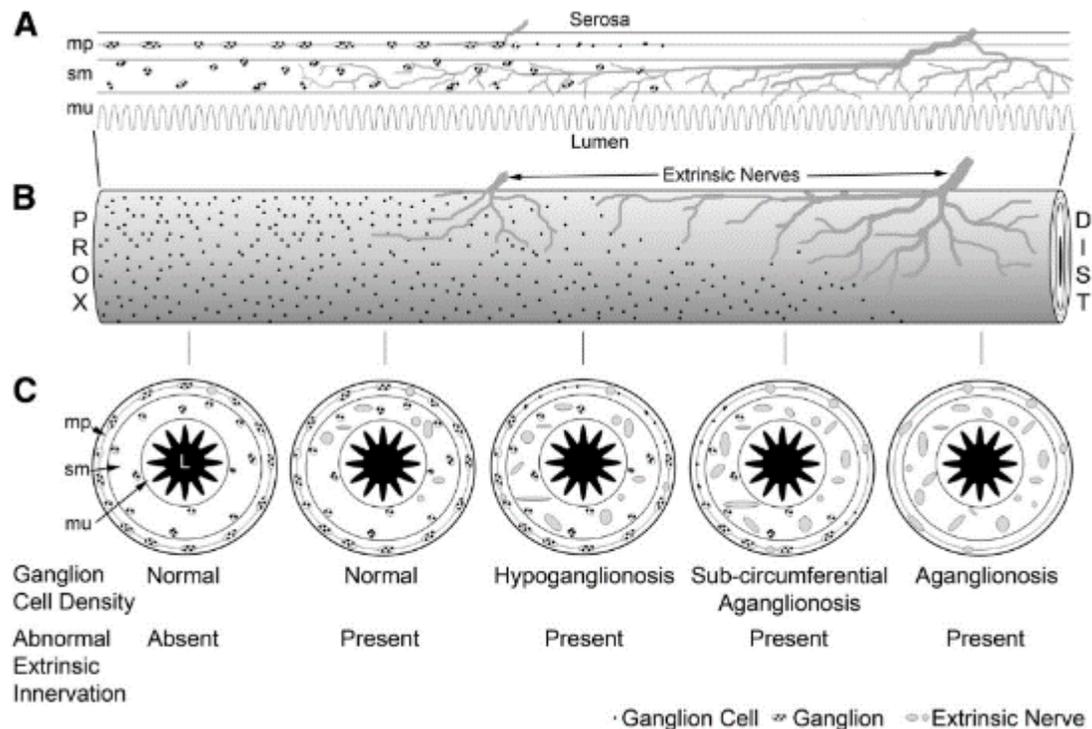


Figure 2.3: “A diagram of the transition zone illustrates the distribution of ganglion cells and extrinsic innervations from the proximal (PROX) euganglionic bowel to the distal (DIST) aganglionic segment. (A) Longitudinal section of full-thickness bowel wall, including muscularis propria (mp), submucosa (sm), and mucosa (mu). (B) Illustration of asymmetric circumferential distribution of ganglion cells at different points along the transition zone. (C) Transverse sections at different sites along the segment of bowel illustrated in (B) to show regional and partial circumferential nature of various neuropathological findings” [5] {Reprinted by permission from Elsevier: Elsevier, Seminars in Pediatric Surgery, (Transitional zone pull through: surgical pathology considerations, Raj P. Kapur, Amy J. Kennedy), © 2012}

Figure 2.3 is based on findings from two studies examining sub-circumferential aganglionosis over 23 HSCR resection specimens [5]. One of the indicators of the transition zone, is a string of single ganglion cells without surrounding neuropil scattered between the muscle layers of the muscularis propria [31]. Submucosal and myenteric ganglia presence are generally aligned in the proximal-distal distributions along the colon [31]. However, submucosal euganglionosis tends to extend further caudal than myenteric (the opposite can still occur), despite some overlapping regions of sub-circumferential aganglionosis [5]. Although myenteric hypoganglionosis is easier to determine, it can also extend several centimeters proximal to the last segment of sub-circumferential

aganglionated colon [31]. Clinically, Boman *et al.* also recognized the advantages of using myenteric ganglia as the intraoperative criterion for colon resection length [42]. Furthermore, Kapur & Kennedy mentioned that “[m]oderate or mild hypoganglionosis is difficult to recognize without considerable experience or formal neuronal counts with appropriate normative data. Less severe forms of sub-circumferential aganglionosis, in which one or more gaps between ganglia is slightly or moderately enlarged, may be difficult to diagnose without formal measurements and appropriate age, gender, and location-specific controls.” [5]. Ganglion density along the colon is inversely correlated with patient age. The myenteric plexus network becomes significantly less dense within the first 3-4 years of postnatal development [38].

This density is also dependent on location within the intestine, as discovered by Miura *et al.* in 1996 when they examined HSCR effects using histopathology and computerized 3D reconstruction methods [43]. Since digital scanners were yet to be commonplace, this group manually traced serial slices of stained tissue specimens with a projector and uploaded these traces to a computer for reconstruction. Researchers used resected intestine samples from two patients (6-month old males) with HSCR to build a 3D reconstruction for visualizing the structural change of myenteric plexus nerve branches along the intestine, particularly within the transition zone. An autopsy of colon tissue with no evidence of gastrointestinal abnormalities from a third patient served as the control for this study.

These reconstructed images revealed plexus as a finely woven network extending along the intermuscular septum. In the control case (Figure 2.4a and Figure 2.4b), ganglia are connected by nerve bundles in a relatively regular mesh structure, where the large

intestine consists of a more densely woven network with bundle diameters of ~70um and the small intestine is made of coarser meshes with bundle spanning ~50um in diameter. These diameters may correspond to the size of plexus regions found in a cross-section colon slice. Reconstructed images of the completely aganglionic regions (Figure 2.4c), for both patient cases revealed only hypertrophic bundles of extrinsic nerves in sparse and irregularly shaped networks, running mainly along the intermuscular septum but within one of the muscle layers in some regions. In the transition zones, reconstructed images for patient case 1 (total colonic aganglionosis) revealed thinning and increasingly irregular neural meshes near the aganglionic zone, with disrupted and open-ended branches in several locations. The researchers also observed isolated ganglia without connections to the neural network mesh. Extrinsic nerves also ran into the intermuscular septum to join myenteric plexus regions. The TZ was thought to be 1 cm or longer. With the second patient case (sigmoid colon aganglionosis), the plexus structure resembled the distal margin of the TZ in patient case 1, except only 3 mm in length. For both patient cases, 3D reconstruction confirmed the intramuscular presence of some ganglia in the TZ, indicating that hypoplastic plexus strays from the intermuscular septum in the muscularis propria. Miura *et al.* (1996) concluded that the range of bowel resection needs to account for myenteric ganglia density and sizes of nerve diameters, ideally adding a “grade of plexus hypoplasia” to seromuscular biopsy evaluation criteria. Although these researchers did not investigate sub-circumferential aganglionosis, their findings are consistent with those described by Kapur & Kennedy (2012) nearly two decades later.

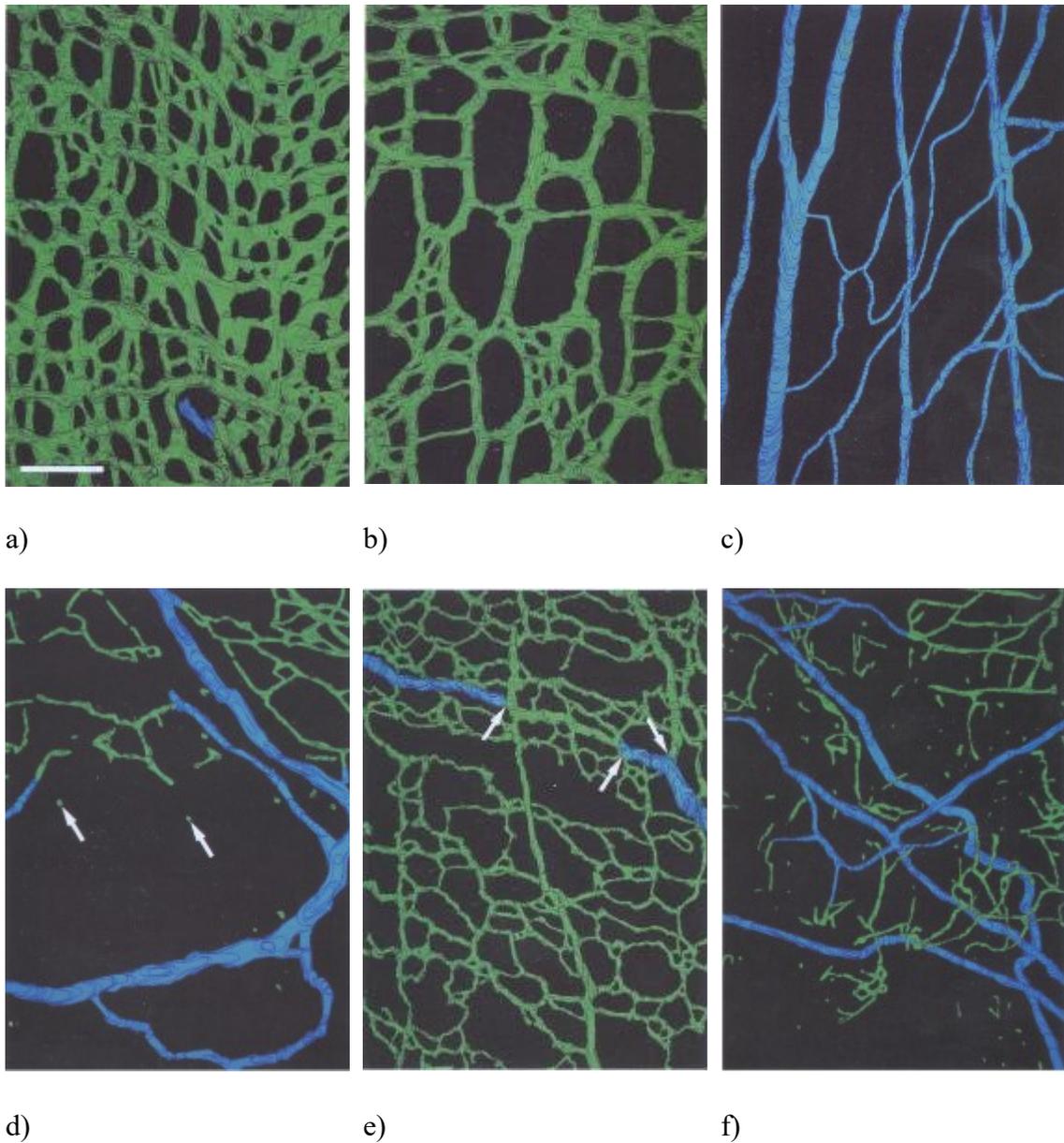


Figure 2.4: “a) 3-dimensional reconstruction of myenteric plexus (green) in a normal colon (control case). An extrinsic nerve (blue) is shown joining the plexus. The upper portion denotes the proximal side of the intestine; the lower part denotes the distal side. Bar = 500 μ m. b) Myenteric plexus in a normal ileum. Note that the mesh is markedly coarser than in the colon. c) Patient case 1 – the aganglionic zone of the sigmoid colon. Only hypertrophic extrinsic nerves (blue) form a sparse network. d) Patient case 1 –the transitional zone contiguous to the aganglionic zone. Some isolated ganglia are seen (arrows). e) The transition zone 1 cm proximal to the area displayed in (d). Compared with the normal ileum from (b), the fibers are strikingly hypoplastic. Hypertrophic extrinsic nerves also seen and join the myenteric plexus (arrows). f) Patient case 2 – the transition zone of the sigmoid colon. Hypoplastic bundles form an irregular structure, as in patient case 1.” [43] {Reprinted by permission from Elsevier: Elsevier, Journal of Pediatric Surgery, (The structure of the transitional and aganglionic zones of Auerbach's plexus in patients with Hirschsprung's disease: A computer-assisted three-dimensional reconstruction study, Hiromitsu Miura,Ryoji Ohi,Shang Wen Tseng,Tohru Takahashi), © 1996}

The studies described in this chapter section provide valuable insight into the pathophysiology of HSCR-affected bowel. We learned that ganglia presence in the submucosa does not necessarily correlate to ganglia presence in myenteric plexus, and healthy colon requires euganglionosis within both regions. We also learned that it is insufficient to base pathology evaluations on lateral seromuscular biopsies. Transverse cross-sections are required since the density of myenteric ganglia around the colon axis is relevant to determine the extent of sub-circumferential aganglionosis. Determining the appropriate levels of myenteric ganglia density at the resection level is also heavily dependent on patient age, the location within the intestine, and maybe even gender. Yet, these are all conclusions drawn from computerized morphometric analyses where a human operator is still required to perform manual visual assessments and measurements on screen. Since manual examinations are heavily time-consuming, studies are limited to a small number of HSCR cases and are not directly applicable in the intraoperative setting. Clinicians need to account for many variables when making intraoperative decisions or post-operative analysis on seromuscular evaluations. Here we see the potential gains an automated image processing method could provide for HSCR evaluation. Automating the information that can be extracted from the colon cross-sections has the potential to not only minimize observer bias and significantly reduce time for assessment, but also provide a greater amount of data for expert analysis in the operative context when a resection level needs to be chosen. More data may allow expert pathologists to improve interpretations and create objective standards for the results of the automated image analysis. Therefore, we consider the measurements and descriptions from this chapter section relevant to developing and evaluating algorithms appropriate for examining colon cross-sections for

HSCR. These findings help to provide clear targets for algorithm development, ensuring that important edge cases can eventually be addressed (e.g. identifying single-cell ganglions without surrounding neuropil). Since this thesis proposes data extraction using image processing methods, the last section of Chapter 2 will highlight studies using said method for HSCR cases.

2.4 Characterizing HSCR via histopathological image processing methods

Digital histopathology incorporates the process of acquiring whole slide images (WSI) of tissue specimens in a format to be shared and interpreted by experts in the field. Popularized within the last decade due to technological advances of cameras and computational capacity, digital pathology scanners capture high-resolution images at magnifications comparable to microscopes. These images can be viewed on a computer screen with automated image analysis tools and software allowing pathologists to remotely collaborate and evaluate with greater consistency and speed than what was ever achievable in the past [44]. This is exemplified by the computerized morphometric studies described in Section 2.3. Now, digital pathology has evolved to include image processing: a method that has become increasingly widespread for cancer diagnosis, including those of the colon [45]. This has allowed for development of algorithms generalized for processing large images with various immunohistochemistry stains [46]–[48]. Recently, deep learning methods have proven to be effective for automating histopathology image analysis [49]–[51]. In our research, we use digital pathology to enable knowledge expansion and potential computer aided standardization of the seromuscular evaluation for HSCR. Although ganglia can be found in other tissue structures, such as the brain or ear canal, publications describing ganglia identification in these WSIs seem to have little algorithmic relevance to

our work due to the vast differences from these specimens and ours. Therefore, to understand the current state-of-art, this chapter closes by discussing the few existing publications that utilize histopathology image processing and machine learning methods to study HSCR.

Published in 2018, Najjar *et al.* tested the utility of quantifying mucosal calretinin immunostaining by image processing and analysis as an adjunctive tool to define the transition zone boundaries in HSCR-affected colon [52]. Although not clinically used to measure the TZ, studies have shown the presence of mucosal calretinin-immunoreactive nerve fibers excludes aganglionosis [29]. The authors hypothesized that calretinin-stained mucosal nerve fibers gradually increase throughout the TZ, beginning with complete absence in aganglionated colon and ending with uniform density in ganglionated tissue. Resected colon specimens from fifteen pull-through cases of HSCR between 2010 and 2015 were used for this pilot study, where all patients were under the age of two and mostly neonates. From each specimen, multiple cross-sections were prepared at 2-3mm intervals longitudinally. The circumference of the colon cross-section was divided and imaged at 200x magnification (excluding sections containing large folds or artefacts). Researchers defined the beginning of the TZ arbitrarily as the point at which mucosal calretinin-stained nerve fibers were first detected. Continuing with the assumption that the TZ is generally less than five centimeters in short-segment HSCR, the euganglionic region was defined as the zone that is at least five cm proximal to the beginning of the transition zone.

Researchers used an image editor to manually delete the submucosa and any obvious artefacts. In MATLAB, Calretinin-stained brown pixels were then identified in the remaining tissue as those where the red channel had greater intensity than both green and

blue in RGB space, yet with a lightness value less than 100 in L*a*b* colour-space. Remaining objects had to be greater than 50 pixels in size to be identified as a mucosal nerve fiber. The slide background and crypt regions of the colon tissues were removed using k-means clustering, to leave only the mucosa region in the image. Finally, authors defined pixel count (PC), as the percentage of brown pixels within the mucosa, as the metric used to evaluate mucosal nerve fiber presence in each image. Figure 2.5 depicts evaluation results in MATLAB for three different images.

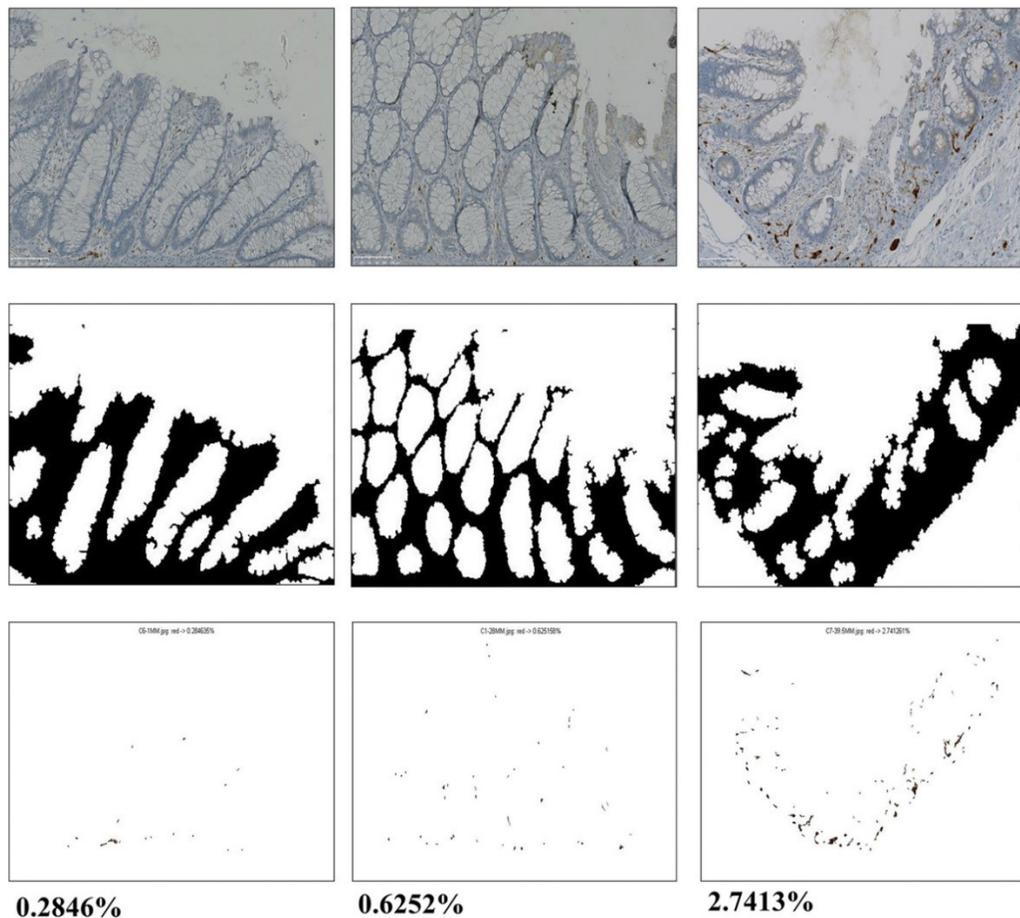


Figure 2.5: “Pixel count by image processing and analysis. The top row shows unedited captured images of colonic mucosa with calretinin immunostaining at $\times 200$. The middle row shows segmentation. The submucosa was subtracted by an editing program. The space and crypt are segmented and the mucosa remains as the area of black. The bottom row shows brown-colored pixels within the mucosa.” [52] {Reprinted by permission from Georg Thieme Verlag KG: Georg Thieme Verlag KG, *European Journal of Pediatric Surgery*, (Image Processing and Analysis of Mucosal Calretinin Staining to Define the Transition Zone in Hirschsprung Disease: A Pilot Study, Saleh Najjar, Sangtae Ahn, Israel Kasago, Chunlai Zuo, Kavita Umrau, Sanaz Ainechi, Christine Whyte, Christine E. Sheehan, Suzanne M. Homan, Hwajeong Lee), © 2019}

To validate the algorithm, five pathologists independently reviewed 45 randomly selected images and visually rated them on a 3-tiered ordinal scale of 1 (low), 2 (intermediate), or 3 (high) to describe the level of mucosal calretinin staining. Scores were then compared with pixel count using the student t-test for each pathologist. For all five pathologists, images with a visual score of 1 had lower average brown pixel counts than images scored as 2 or 3, indicating some level of agreement between the two measures. Researchers also plotted PC for each image against distance from the rectal distal end of the colon, to quantify mucosal nerve presence along the resection. The Pearson's correlation coefficient for pixel count versus distance along the TZ, was found to be positive in ten out of thirteen (77%) of the cases with a mean of $r = 0.78$ (ranging between 0.66 and 0.94). Additionally, the first appearance of ganglion cells (as determined by human expert) was evaluated in ten cases. The most common sequence was the detection of mucosal calretinin stain (beginning of the TZ) followed by submucosal ganglia, and then myenteric ganglia.

To our knowledge, this is the first journal publication to use image processing methods to analyze images of immunohistochemistry (IHC)-stained HSCR colon tissue. With this technique, Najjar *et al.* (2018) successfully quantified calretinin IHC stain in the mucosa and found it to correlate with expert pathologist visual interpretation with statistical significance. While mucosal calretinin staining does gradually increase in the TZ, substantial variation of staining around the colon circumference (even in euganglionic colon) prevents this metric from acting as an unequivocal measure for HSCR evaluation. We also wonder if structures, aside from nerve fibers, may be detected by Calretinin stain in the mucosa. Nevertheless, although not a completely automated process, authors were

able to show image processing and analysis techniques as a useful tool for characterizing serial colon slices in HSCR studies.

Soon afterwards, Schilling *et al.* published machine learning methods to differentiate digital pathology images of HSCR-affected colon from healthy colon based on the submucosal region [53]. From tissue blocks and suction biopsies collected between 2013 and 2016 across 31 specimens (12 HSCR-positive and 19 HSCR-negative; ages ranging from 31 days to 31 years, median = 5 years), their dataset consisted of 75 Calretinin-stained images, 86 of MAP2-stained images, 58 S100 β -stained images, and 88 GLUT1-stained images. These numbers lead us to believe tissue samples have been divided across slides. Using Definiens Developer Tissue Studios, researchers manually marked the submucosa, and then employed built-in stain isolation algorithms to detect regions with IHC chromogen. For Calretinin stained images, the main features extracted include: 1) Percentage Area of IHC marker, 2) Average brown chromogen intensity, 3) Area of brown chromogen stain in SI units, and 4) Number of marker areas. With Python's scikit library, they built a voting ensemble classifier consisting of logistic regression, support vector machine (SVM), decision tree, and random forest. After applying stratified re-sampling of the minority class and normalizing data to unit norm and variance, authors trained on 56% of the data, using 19% for hyperparameter development. 10-fold cross validation on the test data revealed 83% precision and 87.5% recall. Automated diagnosis of "diseased" or "not diseased" was successful for calretinin and MAP2, but ineffective for S100 β and GLUT1 -stained images. However, authors recognized that simply evaluating aganglionosis is an over-simplification of expert diagnoses. For instance, these detection and discrimination methods are not viable for marking the transition zone. Researchers also

expressed concerns with variability among images due to differences in brightness, contrast, and resolution of digital scanners. Their future focus remains on developing unsupervised training methods without the need for large pre-labelled training data.

2.5 Conclusions

Based on the literature review and summaries presented in this chapter, various researchers have observed structural data relating to the colon, such as ganglion count and spread in myenteric plexus regions, to be characteristic of and correlate to the severity of HSCR-affected intestine. However, manual identification is time-consuming, tedious, and subject to high observer variability. This thesis proposes an automated image analysis approach for extracting structural information from digital histopathological images of excised colon specimens, upon which an objective measure for universal scoring rubrics for HSCR may be derived. From extensive literature searches, very few studies were found using image processing to analyze histopathology for HSCR cases. The results achieved by Najjar *et al.* (2018) with quantifying mucosal nerve fibers imply that image processing methods can be used to quantify plexus regions in the submucosa and myenteric regions with calretinin immunostaining. Their conclusions also provide an expectation to quantify image analyses to patient outcome. Likewise, the work completed by Schilling *et al.* (2019) on submucosal image classification confirms that Calretinin-stained images will perform well with machine learning techniques to discriminate HSCR-affected tissue from healthy colon. Through this review, we have also confirmed that our work presents novel and unique contributions since myenteric ganglia identification examines a grading standard which is extensively used in clinical practice but has yet to be quantified using automated image processing and machine learning methods to allow for more complex analysis.

Chapter 3: Clinical Data Set and Manual Annotations

In this chapter, I describe the images used in our dataset, briefly providing explanations for tissue specimens, IHC staining and acquisition. I also describe the manual annotations used to label our data for evaluation purposes and supervised learning methods.

3.1 Study Subjects and Image Acquisition

Given the incidence rate of HSCR, the colon sections imaged in this dataset are collected over a time span of more than twenty years, thus creating a valuable dataset. Our image dataset includes 30 histopathology images of colonic tissue obtained from 26 infant patients diagnosed with HSCR at CHEO, the Children's Hospital of Eastern Ontario. To address the discrepancy in the current clinical assessments in determining criteria for healthy colon, colon tissue cross-sections were taken from the proximal end of resected colon during an HSCR pull-through surgery. Theoretically, since the tissue was obtained after clinical assessment, these samples should be representative of healthy and functional intestine. The CHEO Research Ethics Board (REB) approved the study protocol. Guardians provided written and informed consent for collection and usage of patient tissue data. The Carleton University REB approved of the secondary use of these images, which have been de-identified, for this research.

3.1.1 *Tissue Preparation prior to Image Acquisition*

Paraffin blocks (of 3-4 μm in thickness) from the ganglionic intestinal segments of past HSCR cases were stained automatically for consistency with the Leica BONDMAX from Leica Microsystems (IL, USA) and Leica BONDIII from Leica Biosystems (IL,

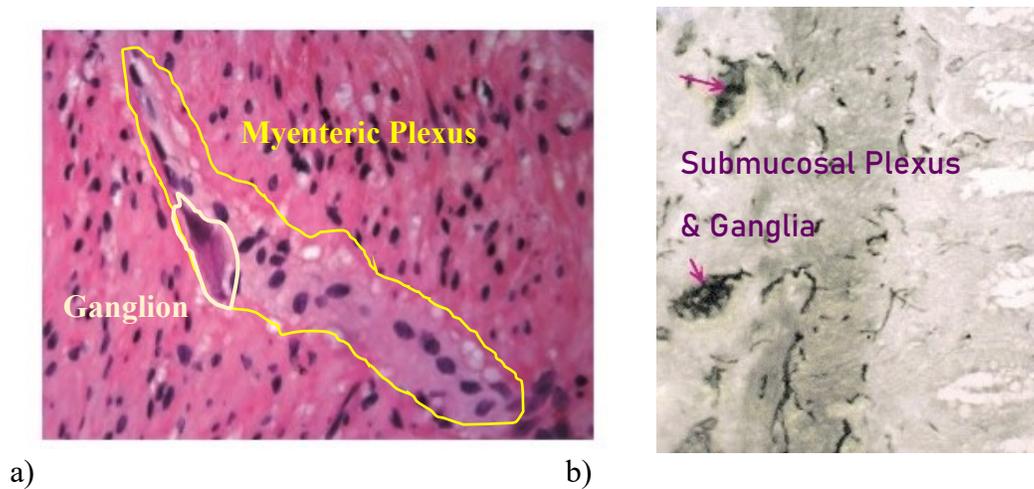
USA), using immunohistochemistry (IHC) systems for Calretinin stain. The IHC-stained slides were scanned using the Aperio ScanScope CS (Model SC) by Aperio Technologies (CA, USA) at an effective magnification of 200 \times , permitting a resolution of 0.5 $\mu\text{m}/\text{pixel}$.

3.1.2. *Histopathological Stain*

In clinical practice, HSCR paraffin blocks are commonly stained with hematoxylin and eosin (H&E) due to ease of use and fast staining times. However, there are multiple IHC stains that can be used to assess colon tissue for HSCR in research settings. These include but are not limited to acetylcholinesterase (AChE), HU C/D, MAP2, synaptophysin; all of which highlight different cell components. In Figure 3.1a), one can see hematoxylin stains nucleic acids, including those present in smooth muscle cells, glial cells, and ganglion cell bodies, with a deep-blue and purple hue [54]. Eosin non-specifically stains acidophilic structures, such as most proteins in the cytoplasm and extracellular matrix [54]. Hence, the muscularis propria and surrounding tissue appear pink. On the other hand, as depicted in Figure 3.1b), acetylthiocholine reacts to the increased AChE presence in hypertrophied nerves of the hypoganglionic colon to precipitate copper ferrocyanide in high-AChE activity regions [55].

AChE negative staining was historically the favoured method to highlight HSCR-affected tissue. However, it assumes normal nerve cells do not have sufficient AChE expression for staining to occur [55]. In reality, the interpretation of AChE stain results is complex due to its dependence on multiple factors, such as a positive correlation in stain intensity with increasing patients' age [55]. This staining method is also a time-sensitive and complicated procedure [56]. Instead, the clinical gold standard of H&E stained images allows pathologists to identify ganglia using the presence of a nucleolus as a definitive

marker for the nerve cell. If the cross-section of the paraffin section cuts through the center of the ganglion body, the structure is included within the pathologist’s count. Yet, determining plexus regions and ganglia is a difficult task due to poor distinguishing features between these structures and the enveloping muscularis propria [57]. This is particularly true for immature ganglia.



a) b)
Figure 3.1: Visualizing Plexus Regions and Ganglia under H&E and AchE IHC stain. These images are from two different sources, and not directly comparable to each other. a) Enteric nerve plexus showing ganglion cell in H&E stain. Magnification (x400) [57]. b) AChE stain at high power magnification showing two ganglia (arrows) in the submucosa. These are negatively stained nuclei of the ganglion cell [58]. These are Open Access Licensed images.

Over the past few decades, Calretinin IHC staining has been steadily gaining popularity for its greater ability to discern plexus regions and ganglia from the remaining tissue, as captured in Figure 3.2. Calretinin is a 29 kiloDalton (kD), vitamin D-dependent protein that attaches to calcium and is involved with neural signaling [59]. It provides a reliable adjunctive test in the evaluation of rectal suction biopsies for HSCR compared to the use of AChE [60]. This has been observed by numerous studies, including [57], [60]–[63], due to Calretinin stain’s relatively simple “positive-or-negative” interpretation. Moreover, the increased distinction between ganglia from muscularis propria in terms of hue and pixel intensity provides an easier platform for computer-aided identification,

particularly in comparison to H&E stained images. Using monoclonal antibodies that bind to epitopes on Calretinin, and a complimentary chromogen, immunohistochemistry (IHC) staining allows for the visual expression of Calretinin found in enteric nerves [64]. This biomarker distinguishes plexus regions in IHC-stained sections by pigment that appears brown to the human eye. Therefore, Calretinin-stained images were used for our study since it presents stronger discriminant to the human eye (and likely computers) than H&E stain. Simpler interpretation via visual inspection also allows for easier evaluation and error analysis of our image processing algorithms.

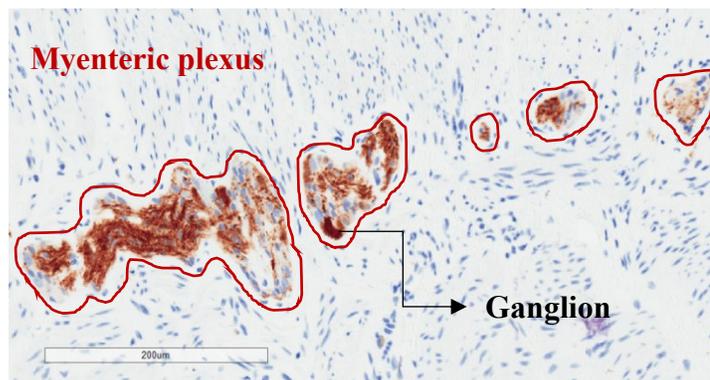


Figure 3.2: Calretinin Stain highlight myenteric plexus regions and ganglion from surrounding muscularis propria at 15.4x apparent magnification

However, we do realize that the intensity of Calretinin stain is dependent upon multiple factors and this creates variations across samples. Automated IHC staining systems attempt to minimize differences arising from chromogen concentration and binding time. However, intra and inter-image variations between the staining of plexus regions can also be affected by changes in biomarker concentration, improper deparaffinization during sample preparation, or weak binding of antibodies to background tissue that mimic the target receptor sites [65]. False-negative staining can also occur in the ganglionic colon due to early ischemic necrosis of tissue [66]. Intraoperatively, H&E

is still preferred over Calretinin due to significantly faster IHC staining procedures [66]. Yet, Calretinin stain is more ideal for post-operative analysis and retrospective study.

Although not explored further in this thesis, we recognize that different types of ganglia form part of an integrated circuit within the ENS [67]–[69]. Therefore, depending on the staining method, certain ganglia types may be highlighted more than others and some types may play a larger role in gastrointestinal motility [70]. For instance, in the guinea pig ileum, Calretinin as a biomarker has been shown to highlight nearly 30% of total myenteric ganglia [71]. Although these proportions are suspected to differ in human models, researchers are investigating the capability of stains for visualizing human intrinsic gastric neurons [67], [72], [73]. Nevertheless, a decrease in Calretinin stain intensity is recognized as an indicator of aganglionosis [74]. Therefore, regardless of the proportion of ganglia captured by this IHC stain, quantifying the change of Calretinin stain along serial cross-sections may still be adequate in determining transition zone boundaries. Expected neuronal counts, however, may need to be adjusted according to the IHC staining process.

3.1.3. Image Data

The ideal image depicts a colon cross-section in a doughnut shape from a slide specimen that is 3-4 microns thick. However, due to tissue delicacy during the preparation of the slides, many of the specimens in this study do not achieve this ideal shape for imaging. Thus, as seen in Figure 3.3, some regions of the colon sections are separated and spread across the slide to allow the a whole specimen to fit. In some cases, a cross-section may be split and scanned across multiple slides. Within this image dataset, two of the patients have a colon cross-section scanned and imaged across three slides. Thus, despite

having a population of 26 patients, 30 images were used for this research project. The images were anonymized, and patient information was not revealed.

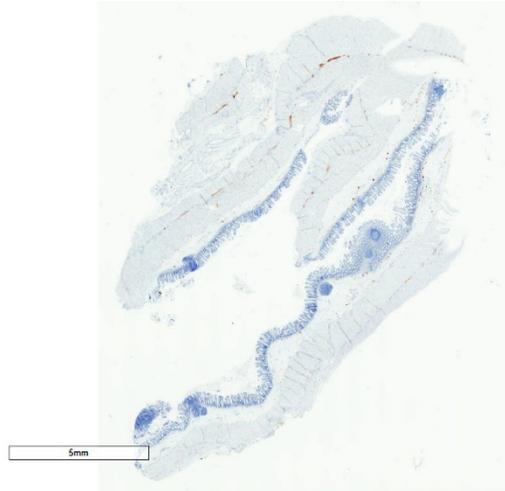


Figure 3.3: Colon cross-section with portions spread across the slide

The Aperio ScanScope CS (Aperio Technologies, CA, USA) digital scanner creates high resolution images of prepared tissue slides. Within the dataset, the smallest image consists of $17983 \times 17602 \times 3$ pixels, or 906 Megabytes (MB). Conversely, the largest image in this dataset has $53951 \times 44014 \times 3$ pixels or 6.6 Gigabytes (GB) in magnitude. In terms of size, the median within this dataset is 3.3 GB and corresponds to two images. The first of these has dimensions of $39963 \times 29900 \times 3$; the second image has $37965 \times 30676 \times 3$ pixels. The sheer size of the pixel data makes loading and analyzing these images, a memory-intensive task.

Images are saved in SVS graphics format, a semi-proprerty and standard file format for pathological digital imaging. It consists of multi-tiled Tagged Image File Format (TIFF) images, storing the content as described below in Table 3.1 using JPEG2000 compression. The tiles are linked between levels, which allows for loading and analyzing the image at a lower resolution and connecting regions of the lower resolution image to the higher quality

base image. The primary advantage is the reduction in memory load when viewing or processing the image.

Table 3.1: The image content within an SVS file for this dataset.

Tile	Content	Size
1	Base Image	Full resolution image
2	Thumbnail	Variable ratio*
3	Reduced image by 2^2	1:4
4	Reduced image by 2^4	1:16
5	Slide Label	Variable ratio*
6	Entire microscope slide with delineations to show crop designating the base image	Variable ratio*

(*image ratio is at a minimum of 1:20, if not greater)

3.2 Manual Annotations

Ground truths for our defined regions of interest (i.e., muscularis propria, myenteric plexus, and ganglia) do not exist. Instead, manual delineations, drawn by the author, act as a surrogate to the ground truth. Hereby referred to as “Observer One”, an expert pediatric and perinatal pathologist at CHEO with over a decade of experience reviewed the annotations and identified ganglia. The thesis author created these annotations using Aperio ImageScope 12.3.3 (Leica Biosystems, Hesse, Germany) at the maximum resolution (i.e. $20\times$ apparent magnification OR $200\times$ effective magnification) in four separate layers:

1. Layer one corresponds to the muscularis propria
2. Layer two corresponds to the myenteric plexus regions
3. Layer three are outlines of ganglia, with a degree of high certainty
4. Layer four are outlines of ganglia, with a low degree of certainty

Figure 3.4 depicts the annotation layers for one of the patient images, with magnified window to show the plexus regions and ganglia annotations in greater detail.

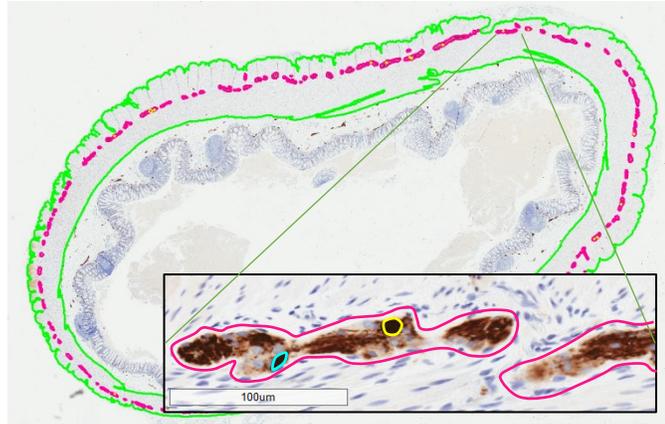


Figure 3.4: Annotation layers for one of the images in the dataset. Green ink outlines the muscularis propria; magenta contours myenteric plexus; yellow are ganglia with high certainty; cyan are refer to objects with low certainty of being classified as ganglia

Using MATLAB 2017a (Mathworks Inc, MA, USA), binary masks encompassing the regions outlined by the manual delineations were created for each of the thirty images in this dataset. A binary mask exists for each corresponding annotation layer. Table 3.2 summarizes the image dataset used, as well as the number of annotations per layer for each image. These high-resolution masks allow for quantifiable evaluation of automated segmentation methods and to identify areas for algorithmic improvement.

Admittedly, this is an unusual approach for histopathology image processing. Given the sheer sizes of WSIs, histopathology labels generally tend to be simplistic (e.g. “HSCR-affected” or “healthy”) or very few in number. However, for the training and development of complex models with supervised learning, access to a large amount of labelled data is indispensable. Albeit a slow process, by keeping the annotations in-house, we ensured high quality labels. For instance, when drawing the muscularis propria, submucosal plexus were excluded when possible. The thesis author also reviewed ganglia

annotations with expert pathologists, asking them to clarify their decisions as they selected objects. This not only encouraged a feedback loop with experts to consider their level of certainty, but also provided us with insight into their thought process. By building this database of high-quality maps, we could begin defining target performance for algorithms. This extends to methods beyond the scope of this thesis.

Table 3.2: High-level Information of Image Dataset.
 Note that the count of relevant structures (e.g. number of muscularis propria segments) is based on manual annotations.

No.	Filename	Original Dimensions	Number of Muscularis Propria Segments	Number of Myenteric Plexus Regions	Number of Ganglia according to Expert One	
					High Certainty	Low Certainty
1	'S00-1910'	15985 × 26214 × 3	4	84	4	2
2	'S03-3178 D2'	47956 × 22563 × 3	2	65	5	2
3	'S04-52'	37965 × 30676 × 3	2	85	16	9
4	'S07-1465' **	65940 × 34997 × 3	2	193	32	6
5	'S14-580'	39963 × 29900 × 3	2	192	93	8
6	'S97-2054'	35967 × 22200 × 3	1	147	4	2
7	'S01-18'	53951 × 42214 × 3	3	224	10	4
8	'S02-410'	35967 × 29909 × 3	4	103	23	5
9	'S02-484'	35957 × 25026 × 3	1	148	32	11
10	'S03-2391'	29972 × 34930 × 3	4	106	19	2
11	'S03-3178 D3'	33969 × 16219 × 3	1	48	3	1
12	'S03-3178 D4'	17983 × 42662 × 3	1	71	1	5
13	'S04-1840'	53951 × 38617 × 3	3	136	41	16
14	'S04-910'	43960 × 28740 × 3	1	168	30	7
15	'S07-1808'	37965 × 35303 × 3	2	145	82	16
16	'S08-2215'	45958 × 32591 × 3	1	180	51	8
17	'S09-2723'	39963 × 32729 × 3	2	52	24	8
18	'S09-2909'	33969 × 28628 × 3	2	84	44	14
19	'S11-1760'	53951 × 44014 × 3	2	262	26	16
20	'S14-1715'	17983 × 17602 × 3	1	64	10	4
21	'S14-2038'	57947 × 26788 × 3	2	172	72	24
22	'S14-3414'	25976 × 16561 × 3	1	112	53	1
23	'S15-1442'	25976 × 29797 × 3	2	131	86	10
24	'S15-1518'	21980 × 35457 × 3	7	118	79	18
25	'S16-1197 B1'	51953 × 34637 × 3	2	52	51	18
26	'S16-1197 B2'	43960 × 29125 × 3	1	105	32	5
27	'S16-1197 B3'	39963 × 39667 × 3	2	93	38	9
28	'S16-1415'	33969 × 40447 × 3	2	55	22	2
29	'S16-1467'	33969 × 22330 × 3	4	89	20	2
30	'S16-567'	41962 × 29385 × 3	2	135	146	11

**major staining errors are present in the image

3.2.1 Staining Errors and Irregularities

Although automated IHC systems reduce variance in staining arising from procedural slide preparation, some staining outliers and anomalies exist in our dataset. Localized false positive stains, although more common, tend to be minor errors and should have little affect on our segmentation results. In our dataset, the most extreme case of minor staining errors is shown in Figure 3.6. However, where we expect staining discrepancies to affect the results and analysis of our algorithms is for image 4, as shown in Figure 3.5. Here, false positive staining of Calretinin outlines the submucosa across the entire cross-section. We suspect this type of staining error will poorly skew the precision of the algorithm-generated myenteric plexus and ganglia segmentation. Finally, one last irregularity we want to point out are images with faint cytoplasmic myenteric plexus staining. In Figure 3.7, we see that image 13 barely expresses brown chromogen in plexus regions. This may be an indication of aganglionosis, as opposed to poor staining.

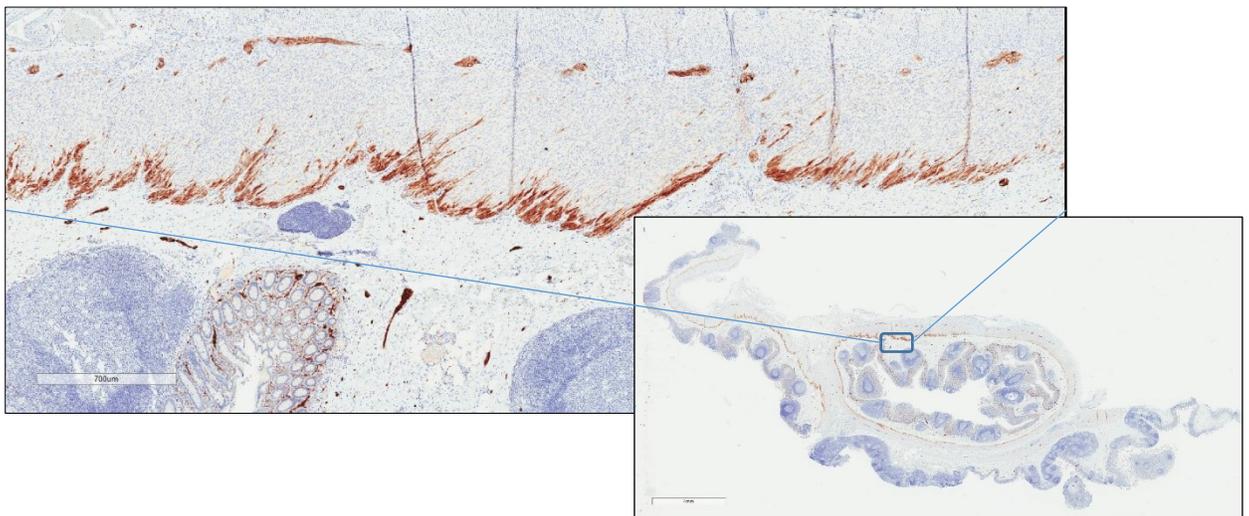


Figure 3.5: Image 4 presents false positive staining errors of Calretinin

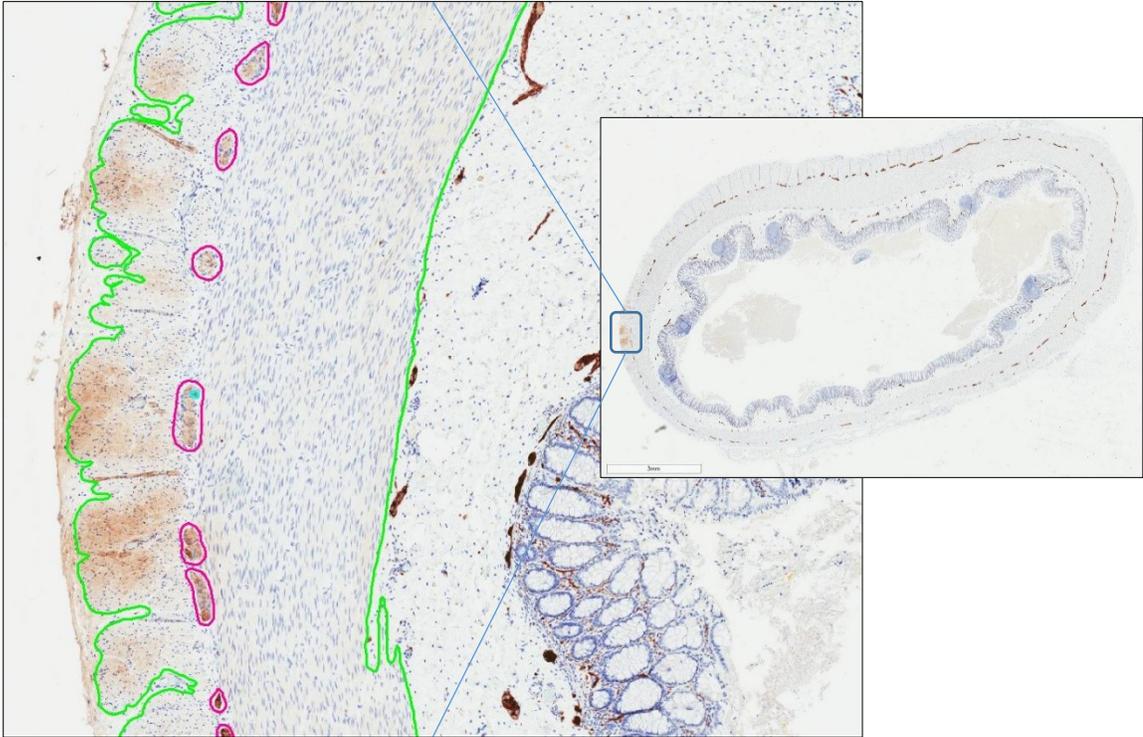


Figure 3.6: Image 9 has minor staining errors localized to a small region of the muscularis. Manual annotations help to distinguish the staining error from the myenteric plexus regions.

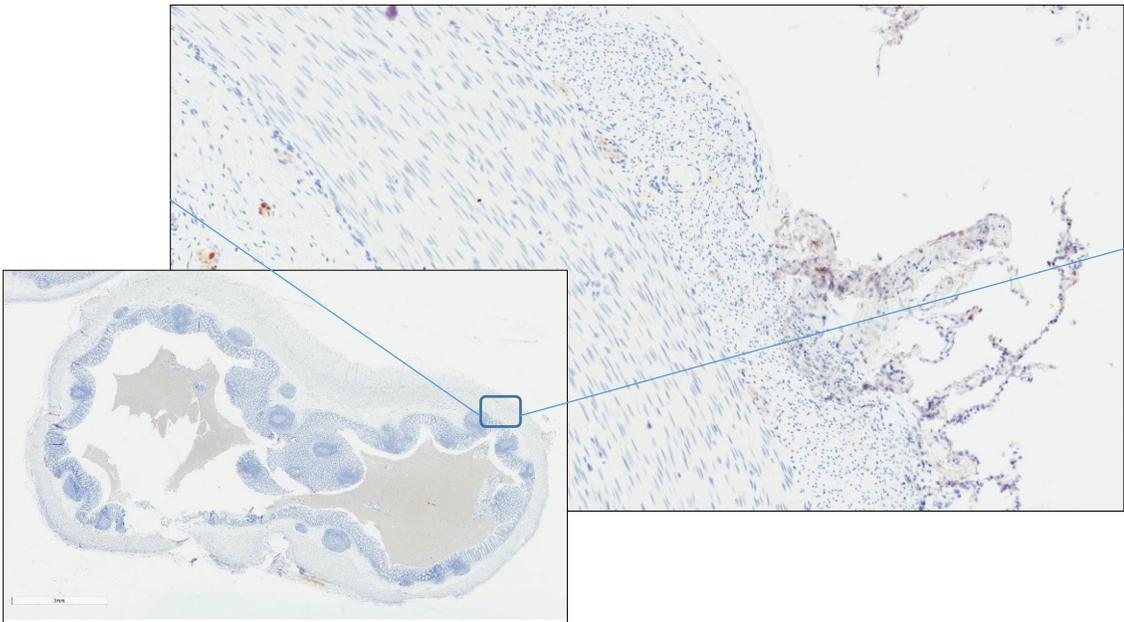


Figure 3.7: Image 13 does not contain staining errors. Rather, we see almost no expression of brown chromogen to highlight myenteric plexus regions.

3.2.2 *Manual Segmentation Criteria for Regions of Interest*

3.2.2.1 *Muscularis Propria*

The muscularis propria consists of two muscle layers: the inner circular muscle and the outer longitudinal muscle. The two regions are histologically distinguishable due to differences in orientation and structure. Similar to changes in grain direction between plywood layers, the smooth muscle cells have opposing orientation between the two muscle layers. In addition, folds of tissue at regular intervals are often present in longitudinal muscle. The heavy density of smooth muscle cells, relative to the density within the submucosa, allow for the ease in differentiation between the muscle layers and the structural submucosa. This is depicted in Figure 4.1. When feasible, Henle's plexus regions were excluded from the areas designated to the muscularis propria, to reduce the likelihood of false positives in the plexus identification step (as described in Section 2.2, these are submucosal plexus bordering the circular muscle layer of the muscularis propria. An example is shown in Figure 5.9.). Likewise, segmentation of structures similar in appearance to the muscularis propria, such as the serosa, were excluded when possible. Uncertainties within muscularis propria segmentation arise when blood vessels are present near the submucosa and muscularis propria borders. Distinguishability can also be challenging if the tissue has deformed during the slide preparation.

The time taken to annotate the muscularis propria ranged between two to twelve hours. This measure depended heavily on the size of the image, difficulty discerning between boundaries, as well as the response time of ImageScope particularly when annotations became large in number and length. The manual segmentations for the muscularis propria were reviewed by Observer One and adjusted according to their recommendations.

3.2.2.2 Myenteric Plexus Regions

Within the seam between the inner circular muscle layer and the outer longitudinal muscle layer of the muscularis propria, are where nerve branches typically grow. These neural pockets are referred to as myenteric plexus regions. When stained with Calretinin, myenteric plexus regions appear brown in colour with intensity varying with uptake of the stain. If the colour contrast between plexus and muscularis propria is low, as in the case for image 13, then ambiguity in the manual annotations arise. In some cases, faint connections may exist between plexus regions, which increases obscurity on the size of the plexus region. Thus, for the sake of consistency, manual annotations loosely outline the regions containing a myenteric plexus. Figure 3.8 illustrates myenteric plexus regions, with manual annotations highlighted by magenta contours. Depending on the number of plexus regions present, an hour (on average) would be dedicated to creating these contours for each image.



Figure 3.8: Manual annotations for myenteric plexus simply outlines the region

3.2.2.3 *Ganglia*

Ganglia (plural of ganglion) are the neural cell bodies in charge of activating muscle contraction. For this image dataset, Observer One identified ganglia within myenteric plexus regions in all thirty images according to the criteria described in Table 3.3: Ganglia identification criteria for calretinin-stained histopathological Images, as described by expert pathologist. These are anecdotal descriptions based on Observer One's expert training and experience. It should also be noted that ganglia between folds of the longitudinal muscle were not accounted for. Shattered plexus regions (i.e., areas where the delicate tissue has been torn), and by default their containing ganglia, were also not included in the manual annotations. In addition, a large ganglion-like object would not be counted as multiple ganglia, unless there was a clear distinction between the cell bodies.

Some exceptions to the rules regarding colour and shape do exist. For instance, an object of lighter brown may be identified as an immature ganglion so long as the three other criteria are met [75]. Likewise, a non-circular object may be strongly suspected as a ganglion if assumed that slide preparation distorted its appearance. Difficulty in discerning ganglia from surrounding neuropil also increases when the stain has been strongly absorbed by both structures. Thus, a second set of manual annotations are included for a class of objects stated as "potential ganglia". Expert pathologists had a higher degree of uncertainty for classifying these objects as ganglia. However, since these structures met some of the defining features, experts did not ignore these objects altogether. Figure 3.9a depicts a case where pathologists identified an object as a "potential" ganglion cell. The time taken for Observer One to identify ganglia could range from 3-25 min for a single image. Since this

is a mentally tiring task, ganglia annotations for the entire dataset were spread over 4 separate days.

Table 3.3: Ganglia identification criteria for calretinin-stained histopathological Images, as described by expert pathologist, Observer One

Characteristics	Criteria
Colour	Darker in stain than surrounding plexus region
Shape	Round nuclei with smooth and defined edges
Texture	Consistent and smooth in texture
Size	Large enough to contain at least 1.5-2 glial cells, relative to those present within the same plexus region

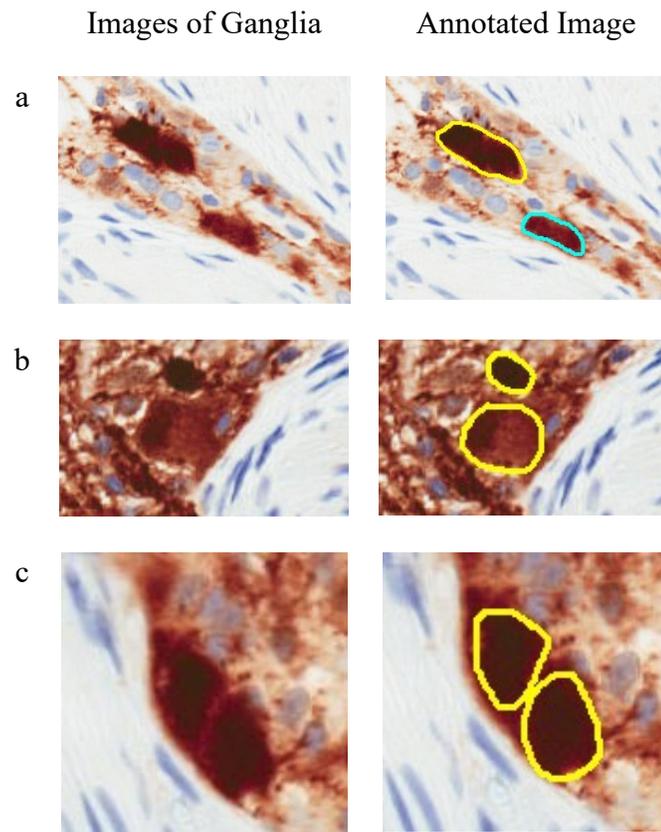


Figure 3.9: Ganglia, as identified by expert pathologist, Observer One. Yellow contours are ganglia with high degree of certainty, whereas cyan contours indicate low certainty. a) reveals an object classified as potentially a ganglion in cyan, due to its relatively small size and blurry edges b) despite the lighter intensity, expert pathologist still marks a large, smooth object as a ganglion c) although clustered together, these two objects are separated by a thin fine line, and therefore marked as two distinct ganglia

3.3 Conclusions

The dataset in this thesis consists of 30 digitized images of Calretinin-stained cross-sections obtained from the proximal end of 26 colon resections after HSCR pull-through procedures. I implemented input from an expert pathologist to use as *a priori* information for algorithm development. Additionally, the manual annotations I describe in this chapter serve analogous to ground truth and is necessary for evaluating the algorithms that segment the muscularis propria, myenteric plexus regions, and finally the encasing ganglia. In the remainder of this thesis, I describe the methods used to identify these histopathological structures of interest.

Chapter 4: Muscularis Propria Segmentation

The muscularis propria (MP) is the powerhouse for intestinal peristalsis. Within the MP, the inner circular and the outer longitudinal muscle layers encapsulate myenteric plexus regions that contain the ganglia responsible for intrinsic innervation of the intestine (Figure 4.1). To detect these enteric ganglia as a quantifiable measure for automated HSCR seromuscular evaluation in Calretinin-stained WSIs, we proposed an image-processing pipeline initiated by the segmentation of the muscularis propria. In this chapter, I apply two different segmentation methods to identify the muscularis propria across the dataset and determine their suitability for initiating the image processing pipeline.

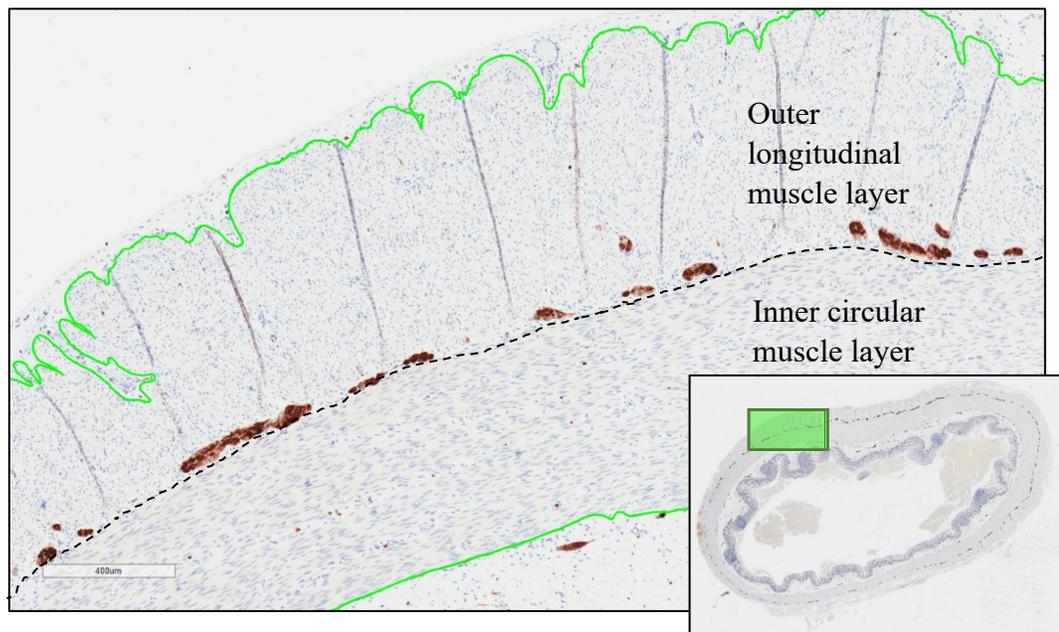


Figure 4.1: The thumbnail for image 9 is shown in the right bottom corner. The green box indicates the region zoomed to 400μm scale for the labelled snapshot. The manual annotations of the muscularis propria are shown by green delineations. The dotted black line highlights the intermuscular septum. Note the difference in granularity between the two muscle layers due to opposing orientation of the blue-stained support cells.

In 2016, Law *et al.* published a pilot study on a single image, applying histogram equalization and empirically derived thresholds in HSV colour-space to identify the muscularis propria as a single connected component [76]. Building upon this idea, Kurian *et al.* employed *k*-means clustering, an unsupervised method to identify the muscularis propria by colour [11]. From there, McKeen *et al.* attempted to improve automated results by utilizing a convolutional neural network (CNN) to extract features from the microstructure of muscularis propria tissue [12]. To the best of our knowledge, these are the only three publications describing algorithms to identify the muscularis propria in Calretinin-stained WSIs. However, since a generalizable model is required to extract muscularis propria regions across varying images, we analyzed the ability of the *k*-means algorithm and a modification to the CNN algorithm to achieve this task across a larger dataset, and identified areas for improvement.

4.1 Methods

4.1.1 *k*-means Clustering Segmentation

Images in this dataset (as described in Chapter Chapter 3:) contain four main structures that are easily distinguishable to the human eye: 1) the background, 2) the muscularis propria, 3) the submucosa, and 4) the mucosa. In terms of classification, most pixels fall under classes belonging to one of these four physical structures. The muscularis segmentation method presented by Kurian *et al.* in [11] uses *k*-means clustering to sort image pixels by colour and texture, to one of these four objects.

This segmentation is conducted on an image downsampled by a factor of 2^4 , to give a resolution of $2\mu\text{m}/\text{pixel}$ from the original resolution ($0.5\mu\text{m}/\text{pixel}$). This not only reduces

computational complexity, but also provides a primary low-pass filter to remove fine details and blur smaller structures such as plexus regions and support cells. Median filtering with a 10×10 pixel kernel ($20 \mu\text{m} \times 20 \mu\text{m}$) provides an edge-preserving low-pass filter on the image. A Gaussian filtering with a standard deviation of six pixels ($12 \mu\text{m}$) was then applied an additional non-uniform low pass filter. These serve as preprocessing steps to minimize noise and average the colour information across large structures, allowing for the background to appear white, the mucosa as purple, the submucosa as near white, while the muscularis propria region blurs to lavender tones. Textural information is encoded by appending the local standard deviation in a 3×3 ($6 \mu\text{m} \times 6 \mu\text{m}$) neighbourhood to the image structure as a fourth dimension. From here, k -means clustering with four clusters, attempts to differentiate the four main histopathological structures. As experimentally determined, the cluster containing the second largest number of pixels tends to correspond with the muscularis propria. However, in a few cases where this assumption does not hold true and provides poor results, the correct cluster was chosen manually (as shown in Figure 4.10). A series of morphological operations on the selected cluster act as the post-processing step for the final segmentation. Since the muscularis is a large structure, small objects are first eroded from the binary image and the remaining larger structures are further dilated. A closing operation removes random branching edges while an additional erosion step is used to prevent the segmentation from including the submucosa. Objects smaller in area than 25% of the largest object are removed from the binary mask. A final dilation is conducted on the relatively cleaned mask, to capture as much of the muscularis propria as possible. Filling holes in the remaining structures completes the post-processing. Figure 4.2 provides a visual overview of this k -means clustering segmentation method.

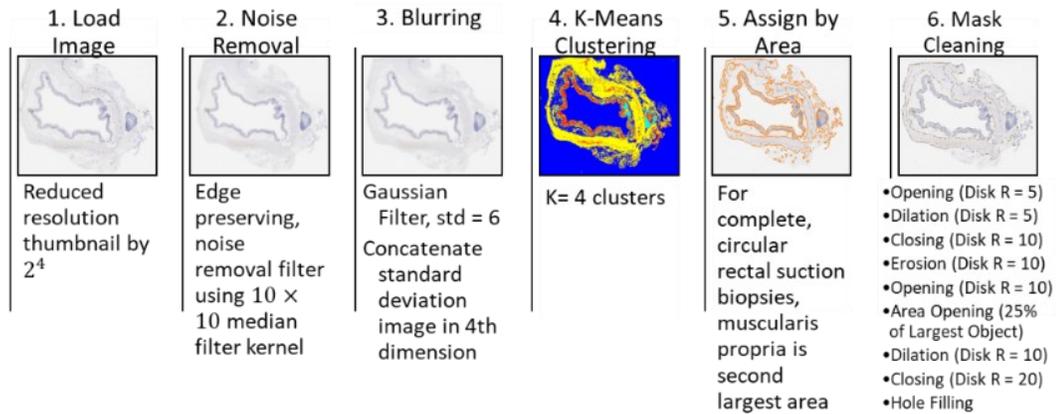


Figure 4.2: Depiction of k-means clustering segmentation method. Operations are reported in pixels at the downsampled resolution of ($2\mu\text{m}/\text{pixel}$)

4.1.2 Convolutional Neural Network (CNN)-based Segmentation

Another method of segmenting the muscularis propria is described in [12], where McKeen *et al.* recognized the density and orientation of the support cells in the MP provide textural information that differentiates this structure from all other objects in the image. To utilize this feature, a CNN model was constructed with the architecture depicted in Figure 4.3. This CNN model consists of three convolutional and pooling layers, ending with two hidden layers of a fully connected (FC) neural network of 256 nodes. The convolutional layers are responsible for feature extraction using 3×3 kernels, where each layer consecutive layer has 64, 128, and 256 kernels, respectively. The pooling layer selects the maximum value from the convolutional layer output using a sliding 2×2 kernel. The FC layer contains rectified linear unit (ReLU) and sigmoid decision functions for feature selection. Applying dropout after the pooling layers minimizes the likelihood of overfitting. Further details on training can be found in [12].

The CNN was trained on six images from the dataset presented in Chapter 3: These six images, identified in Table 4.3, were selected to represent variability in the test set, such

as colour, size, position, sparsity, formation, and presence of staining errors. Given the sheer amount of data on the muscularis within a single image, more was not required for training. The remaining 24 images served as the test set, unobserved by the CNN model.

Training began by downsampling all the images by a factor of twelve ($1.73\mu\text{m}/\text{pixel}$ resolution) to reduce computational costs. Image intensity values were normalized within the entire dataset between zero and one for each colour channel. For the six training images, patches of $13\times 13\times 3$ window sizes around each pixel acted as the input for the CNN model. These patches provided contextual information for each pixel. Manual segmentations of the muscularis propria (section 3.2) served as labels for each image patch in the training dataset (e.g., if the center pixel of the patch was labelled as part of the muscularis propria, the entire corresponding image patch was labelled as muscularis). To provide a balanced dataset and lessen computational costs, several patches labelled as not belonging to the muscularis propria were randomly discarded, such that the total number of training patches was 3,400,000. Binary cross entropy acted as the loss function for the CNN model and accuracy with the manual segmentations provided the training evaluation metric.

The segmentation of the muscularis propria for the test dataset first required extracting $13\times 13\times 3$ patches for each pixel from a test image. The CNN model classified each patch on whether it belonged to the muscularis propria. These prediction labels created an 8-bit intensity image highlighting the MP regions. Post-processing, as described in [12], produced a final binary map for the muscularis propria segmentation.

Although the algorithm [12] provided high fidelity segmentations compared to the manual delineations (as quantified by evaluation metrics like the Sørensen-DICE

Coefficient), holes exist in these binary maps that often correspond to myenteric plexus regions. Therefore, we made modifications to the post-processing of the original algorithm to allow these masks to act as appropriate search spaces for myenteric plexus regions and their containing ganglia. First, an intensity threshold of 170 was applied to the 8-bit prediction labels to create a binary mask where white pixels are generally associated with the muscularis propria. Then, an opening operation removed connected components less than 250 pixels ($750\mu\text{m}^2$) in size at this reduced image resolution. This step decreased the density of pixels belonging to objects unattached to the typically larger MP regions. Next, a morphological closing operation, using a 30-pixel radius disk (area size of $8482.3\mu\text{m}^2$) as the structuring element, filled in small gaps and erased random edge branches while maintaining objects' relative size. Post-processing continued with the removal of small, sparsely segmented objects (i.e., false positives) smaller than 10000 pixels (0.03mm^2), and filling of holes smaller than 30000 pixels (0.09mm^2) in a large connected component. A morphological dilation using a 10-pixel radius disk (area size of $942.5\mu\text{m}^2$) expands the estimated region of the muscularis propria, to maximize the inclusion of myenteric plexus regions found at the boundaries. Finally, an averaging filter smooths the bumps artificially created by the structuring element of the closing and dilation operations. Thresholds for all these operations were determined experimentally.

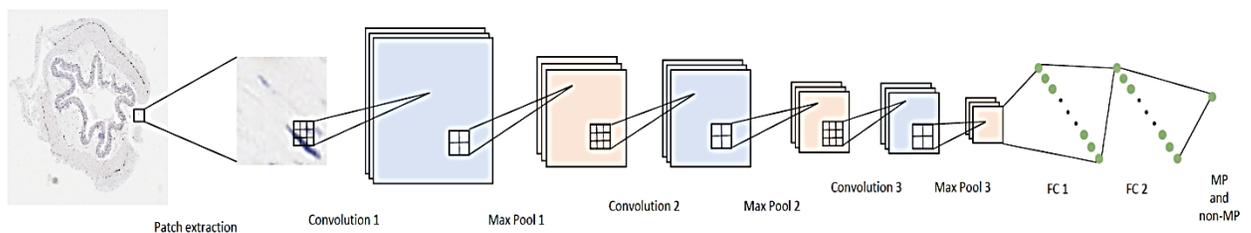


Figure 4.3: Architecture of CNN model [12]

4.1.3 Algorithm Evaluation Methods

To assess the performance of the segmentation methods, evaluation metrics that can quantitatively discern the instances of failure should be chosen. Here we wanted to evaluate the ability of the k -means clustering model and the CNN model to segment the muscularis propria across this image dataset. Manual segmentations of the muscularis propria, drawn at the original image resolution (i.e., $0.5 \mu\text{m}/\text{pixel}$) for each image in this dataset, acted as surrogate of the ground truth. Accordingly, algorithm-generated segmentations of the muscularis propria were compared to its corresponding manual segmentation for evaluation. Since the algorithm-generated muscularis propria segmentation was conducted on downsampled images, the binary masks of the manual segmentation were also downsampled to correctly match the algorithm's image size.

Multiple metrics allowed to evaluate different aspects of the segmentation and help to narrow down the classifier best suited for our purposes. However, effective practice for comparing classifiers requires certain evaluation metrics to be prioritized according to the goal of the project. For the objectives of this thesis work, the aim of the muscularis propria segmentation is to provide ideal search spaces for ganglia located within the myenteric plexus regions. Therefore, the muscularis propria segmentation must primarily maximize myenteric plexus inclusion. This captures the criterion for the satisficing evaluation metric (i.e., the minimal acceptable threshold that the classifier should simply meet). Myenteric plexus inclusion describes the percentage of myenteric plexus regions included in the muscularis propria segmentation, from the number present in the manual segmentation (Equation 4.1).

$$Plexus\ Inclusion = \frac{\#\ of\ plexus\ regions\ in\ algorithm\ generated\ MP\ segmentation}{\# \ of\ plexus\ regions\ in\ manual\ MP\ segmentation} \times 100\% \quad (4.1)$$

Optimizing the segmentation to the exact boundary of the muscularis propria is secondary, but still necessary to exclude false positives (e.g., preventing submucosal plexus regions proceeding as search spaces for the second step in the overall image processing pipeline). This describes the purpose of the optimizing evaluation metric (i.e., the metric to optimize after the satisficing metric has been met).

A region-based metric is perhaps more than satisfactory for this purpose. Accuracy or total error (the percentage of misclassified pixels) were not considered since the muscularis propria does not fit a balanced class scenario. Figure 4.4 reveals the extent of this class imbalance by graphing the percentage of pixels corresponding to the manual segmentations of the muscularis propria for each image in the dataset. The following region-based similarity measures were considered for the evaluation of the muscularis propria segmentation as optimizing metrics:

- a) Sørensen-DICE Coefficient: also known as the F1-score, this near-metric is the harmonic mean between precision and recall and ranges from zero to one

$$DICE = \frac{2|S_{manual} \cap S_{automated}|}{|S_{manual}| + |S_{automated}|} \quad (4.2)$$

- b) Jaccard-Tanimoto Index: also ranging from zero to one, this similarity metric is described by

$$Jaccard = \frac{|S_{manual} \cap S_{automated}|}{|S_{manual} \cup S_{automated}|} = \frac{DICE}{2 - DICE} \quad (4.3)$$

Here S_{manual} refers to the binary mask of the manually segmented muscularis propria, and $S_{automated}$ refers to the binary mask of the algorithm-generated segmentation of the

muscularis propria. Both the Sørensen-DICE coefficient and the Jaccard-Tanimoto index are useful measures when there is a notable class imbalance in the dataset, as is the case for this application. These Tversky indices reach a value of one when perfect similarity occurs and zero when similarity does not exist. Although tempting to assume the two are functionally equivalent due to their monotonic and rank-equivalent relationship (see Equation 4.3), Sørensen-DICE provides a lesser penalty for dissimilarity (i.e., outliers) and missing data than the Jaccard-Tanimoto index [77]. Therefore, when quantifying the comparative performance between classifiers, the preference for using the Jaccard-Tanimoto index emerges if greater resolution is required between data points for analysis.

Additionally, we found precision and recall as attractive measures. Although region-overlap measures can provide estimates for differentiating between excellent and poor segmentation, it is information extracted from reduced dimensionality. Precision and recall provide greater quantifiable insight between under-segmentation and over-segmentation. Precision, as shown in Equation 4.4, describes the number of segmented pixels correctly belonging to the region outlined by the manual delineations. Recall, as defined by Equation 4.5, quantifies the proportion of the manual segmentation identified by the algorithm’s segmentation of the muscularis propria. Here true positives refer to the pixels identified belonging to the muscularis propria, false positives refer to pixels of the segmentation which fall outside of the MP, and false negatives are missed pixels when compared to the manual segmentations.

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (4.4)$$

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (4.5)$$

High recall, but lower precision is typical of an over-segmentation. High precision but lower recall describes an under-segmentation. Low precision and low recall can occur if a region other than the muscularis propria is segmented. Lastly, perfect overlap between the manual and algorithm-generated segmentation is characterized by both precision and recall valued at 100%. Although perfect overlap is ideal, a slight over-segmentation is acceptable when myenteric plexus inclusion is maximized. However, an under-segmentation with full myenteric plexus inclusion is preferred since this minimizes the likelihood of including submucosal plexus regions in the MP segmentation. Therefore, we ultimately optimized evaluation to improve recall and precision, desiring a scenario with high precision and slightly lower recall.

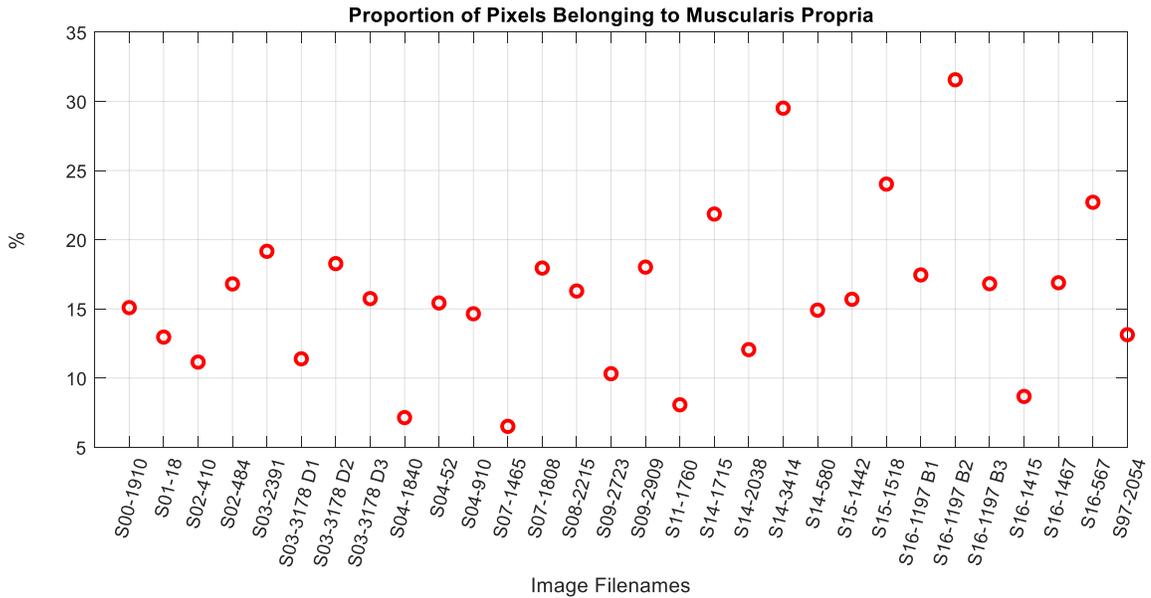


Figure 4.4: Indicating level of class-imbalance for identifying the muscularis propria, based on manual segmentations

4.2 Results

As shown in Table 4.1 and Table 4.2, both the CNN segmentation and the *k*-means clustering segmentation achieved high performances for most of the images in this dataset

when concerning the satisficing metric of myenteric plexus inclusion. The median values for myenteric plexus inclusion are comparable at 98.4% and 96.7%, respectively. However, as captured by the average percentage of plexus inclusion, there are greater instances of outliers with the k -means segmentation. The CNN-based algorithm achieved $96.0\% \pm 5.8\%$ average plexus inclusion over the test set of 24 images, whereas the k -means clustering performed at an average of $77.4\% \pm 32.2\%$ over the entire dataset of 30 images ($75.4\% \pm 34.4\%$ over the 24 images for direct comparison with the test set by the CNN model). Despite complete plexus inclusion for more than a quarter of the dataset (i.e., eight out of thirty), and greater than 90% plexus inclusion in another eight images, the k -means clustering method performs dismally (i.e., less than 30%) on four of the images, with less-than-mediocre results on another three images. Figure 4.6 and Figure 4.7 provide visualization of these results, and reasons for poor myenteric plexus inclusion for some of these images are explored in Section 4.3.

Concerning the optimizing metric, the CNN algorithm performed with an average precision of $81.9\% \pm 7.8\%$ and average recall of $96.1\% \pm 6.8\%$. The high fidelity of the algorithm-generated segmentations with the manual delineations can be seen in Figure 4.8 and Figure 4.9. Comparatively, the k -means clustering algorithm provided segmentations with an average precision of $70.6\% \pm 17.1\%$ and average recall of $78.9\% \pm 32.1\%$. Figure 4.5 visually compares the performance of these two muscularis propria segmentation algorithms. We can see when both algorithms have correctly identified the MP, there is a predisposition for over-segmentation. However, even among the segmentations with high plexus inclusion, the k -means model tends towards greater over-segmentation relative to the CNN-generated binary masks. This is also confirmed by the comparable median recall

produced by both algorithms (k -means at 97.3% and CNN at 98.2%) but the large difference in median precision for both models, where k -means has a median precision of 74.0% but the CNN-model produces a median precision of 86.3%. We can also grasp from these figures that the CNN algorithm is more robust (as confirmed by the lower standard deviation in the evaluation metrics). This is displayed in Figure 4.5 by the tighter cluster of red data points inclining towards ideal overlap (100% recall and precision), while maximizing plexus inclusion.

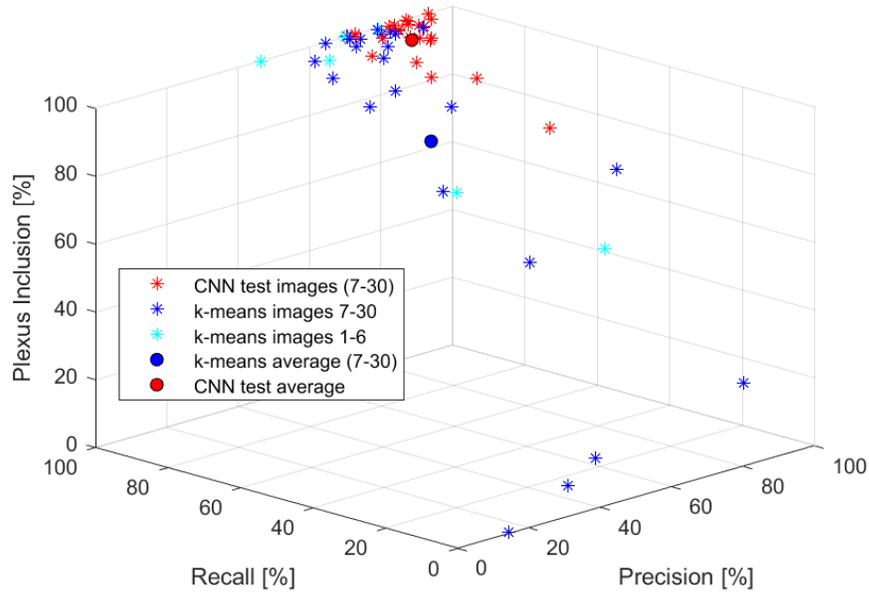


Figure 4.5: Comparing spread of satisfying and optimizing metric results between the two segmentation methods

For the purposes of reporting standard region-overlap measures in the image processing field, we have also included DICE and Jaccard scores. The CNN model performed with an average test Sørensen-DICE score of $89.3\% \pm 4.5\%$ and an average test Jaccard-Tanimoto index of $80.9\% \pm 7.0\%$. Comparatively, the k -means clustering method performed with an average Sørensen-DICE score of $70.3\% \pm 27.1\%$ and an average Jaccard-Tanimoto index of $59.2\% \pm 25.1\%$.

Table 4.1: Evaluation Metrics from k -means Clustering Segmentation

<i>No.</i>	<i>Filename</i>	<i>Jaccard [%]</i>	<i>DICE [%]</i>	<i>Precision [%]</i>	<i>Recall [%]</i>	<i>Plexus Inclusion [%]</i>	<i>Detected # of Plexus Regions</i>
1	'S00-1910' *	40.9	58.0	85.6	43.9	49.4	42 out of 85
2	'S03-3178 D2'	77.5	87.4	78.2	98.9	100.0	64 out of 64
3	'S04-52'	69.7	82.1	69.7	99.9	100.0	85 out of 85
4	'S07-1465'	45.9	62.9	46.0	99.6	100.0	193 out of 193
5	'S14-580'	48.0	64.9	65.2	64.6	66.2	131 out of 198
6	'S97-2054'	61.7	76.3	62.8	97.3	95.9	141 out of 148
7	'S01-18'	11.7	20.9	51.8	13.1	7.1	16 out of 224
8	'S02-410'	77.3	87.2	77.9	99.0	100.0	103 out of 103
9	'S02-484'	81.8	90.0	82.1	99.5	97.9	143 out of 146
10	'S03-2391'	71.4	83.3	72.3	98.2	99.1	105 out of 106
11	'S03-3178 D3'	76.4	86.6	78.1	97.3	92.2	47 out of 51
12	'S03-3178 D4'	70.2	82.5	70.2	100.0	100.0	71 out of 71
13	'S04-1840' *	1.9	3.7	32.7	2.0	8.2	11 out of 135
14	'S04-910'	69.2	81.8	70.4	97.5	97.6	162 out of 166
15	'S07-1808'	62.6	77.0	75.6	78.5	59.1	78 out of 132
16	'S08-2215'	68.6	81.4	69.4	98.3	100.0	178 out of 178
17	'S09-2723'	63.7	77.9	63.9	99.5	100.0	52 out of 52
18	'S09-2909'	81.8	90.0	81.9	99.9	93.7	74 out of 79
19	'S11-1760'	58.4	73.7	63.9	87.1	85.0	215 out of 253
20	'S14-1715'	0.7	1.5	15.1	0.8	0.0	0 out of 64
21	'S14-2038'	79.0	88.3	81.0	97.0	98.3	169 out of 172
22	'S14-3414'	15.3	26.6	95.6	15.4	15.2	17 out of 112
23	'S15-1442'	64.5	78.4	65.1	98.6	89.6	120 out of 134
24	'S15-1518'	71.3	83.2	75.9	92.0	84.6	99 out of 117
25	'S16-1197 B1'	80.4	89.2	85.2	93.6	100.0	52 out of 52
26	'S16-1197 B2'	58.7	74.0	86.0	64.9	39.1	41 out of 105
27	'S16-1197 B3'	51.3	67.8	97.2	52.1	67.0	59 out of 88
28	'S16-1415'	61.2	75.9	61.3	99.7	95.5	42 out of 44
29	'S16-1467'	65.8	79.4	79.1	79.6	82.8	72 out of 87
30	'S16-567'	78.6	88.0	79.2	99.1	98.5	129 out of 131
Mean		58.9	70.7	70.6	78.9	77.4	N/A
Std. Dev.		22.8	24.2	17.1	32.1	32.2	N/A
Median		65.1	78.9	74.0	97.3	96.8	N/A
Summary of Images 7-30 for comparison with CNN test data							
Mean		59.3	70.3	71.3	77.6	75.4	N/A
Std. Dev.		25.1	27.1	18.1	34.2	34.4	N/A
Median		67.2	80.4	75.8	97.1	92.9	N/A

* refers to an image where the k -means cluster was manually selected prior to post-processing

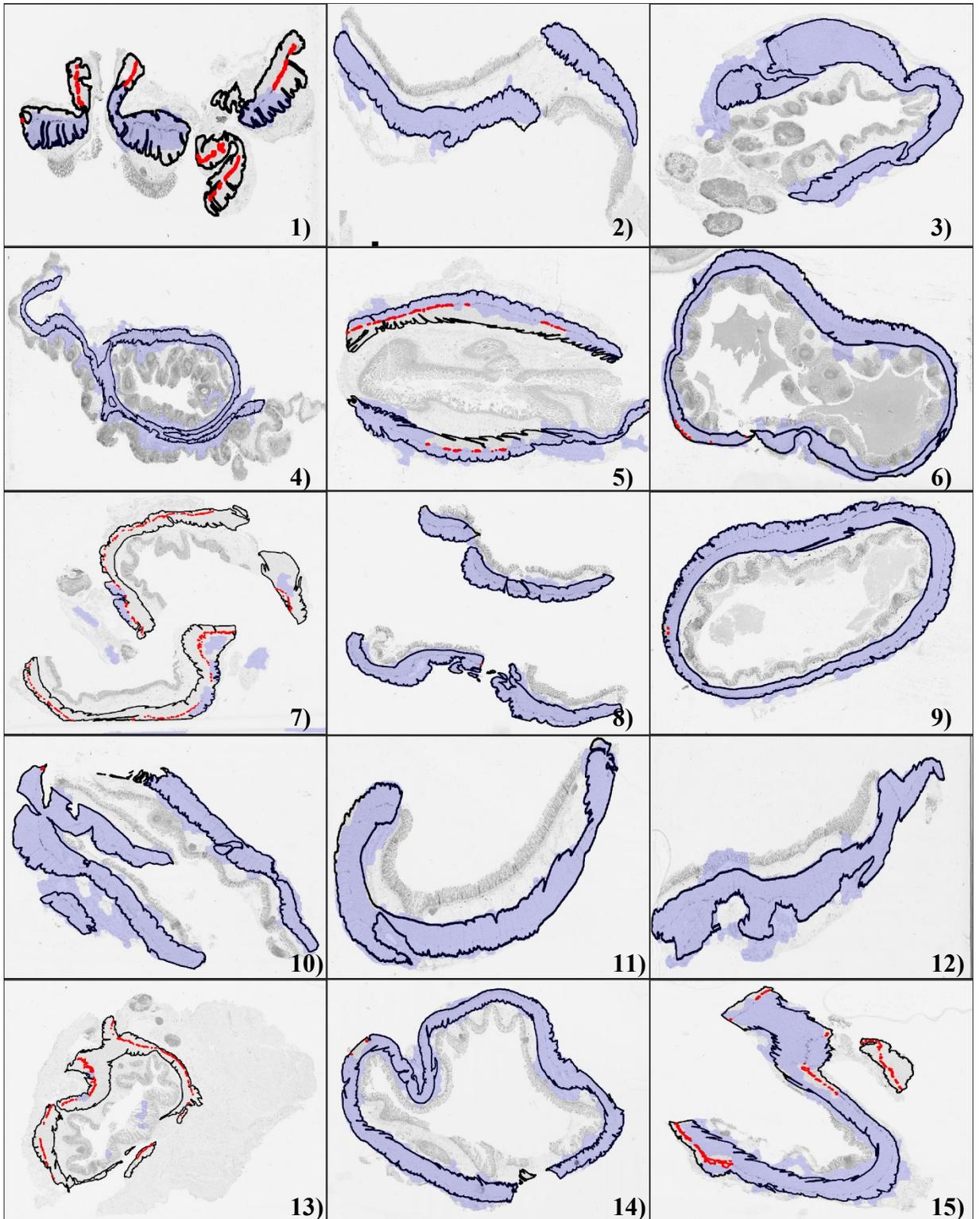


Figure 4.6: Visualizing the Muscularis Propria k-means Segmentation for images 1-15. Black outlines indicate the manual delineations of the muscularis propria, whereas the transparent blue shading depicts the segmentation by the k-means algorithm. Regions highlighted in red are myenteric plexus not captured by the algorithm's segmentation

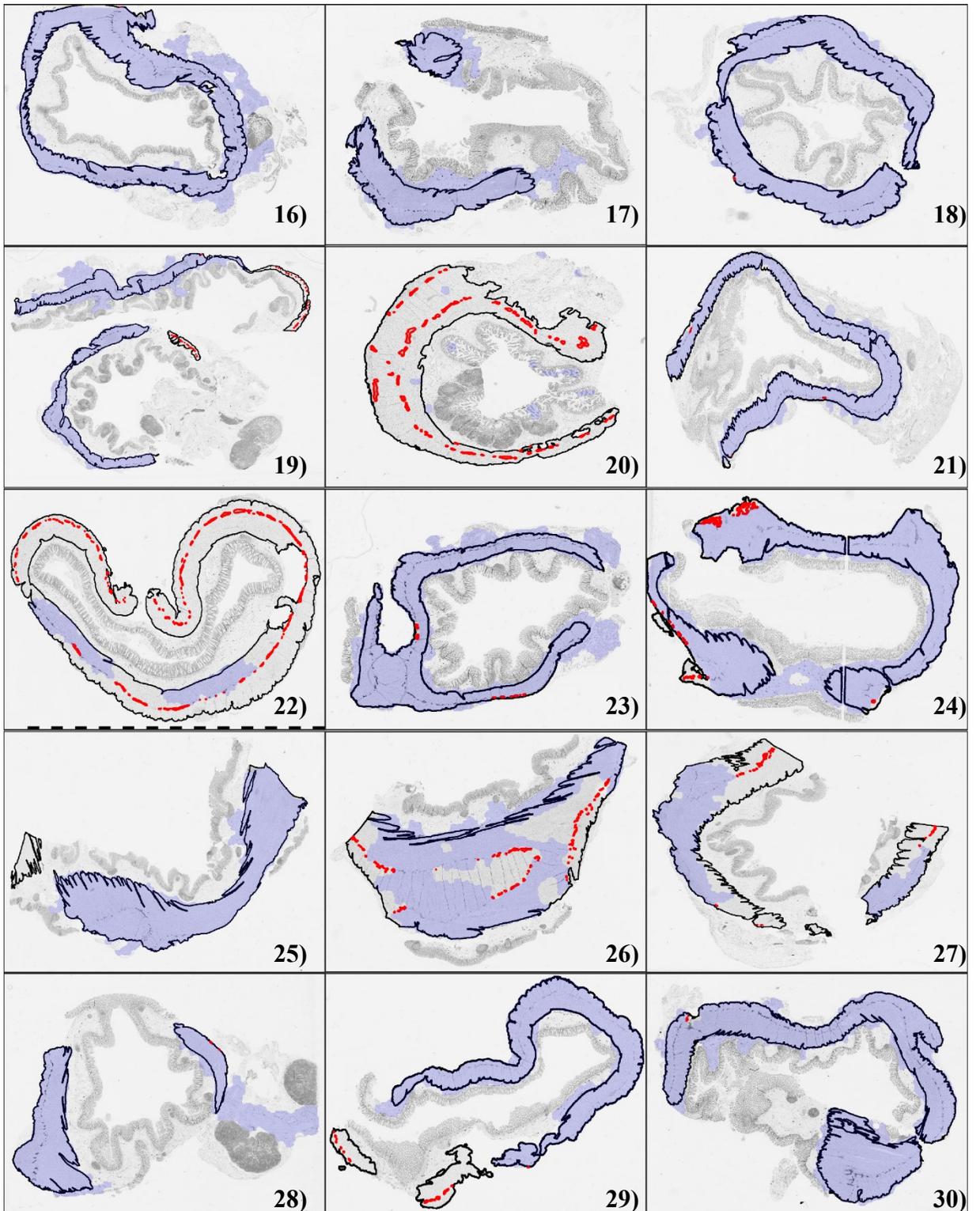


Figure 4.7: Visualizing the Muscularis Propria k-means Segmentation for images 16-30. Black outlines indicate the manual delineations of the muscularis propria, whereas the transparent blue shading depicts the segmentation by the k-means algorithm. Regions highlighted in red are myenteric plexus not captured by the algorithm's segmentation.

Table 4.2: Evaluation Metrics from CNN-based Segmentation

No.	Filename	Jaccard [%]	DICE [%]	Precision [%]	Recall [%]	Plexus Inclusion [%]	No. of Detected Plexus Regions
1	'S00-1910'	67.9	80.9	69.5	96.6	97.6	83 out of 85
2	'S03-3178 D2'	76.6	86.8	78.1	97.6	100.0	64 out of 64
3	'S04-52'	77.9	87.6	78.9	98.5	100.0	85 out of 85
4	'S07-1465'	64.9	78.7	65.0	99.7	100.0	193 out of 193
5	'S14-580'	67.2	80.4	68.8	96.6	99.5	197 out of 198
6	'S97-2054'	72.3	83.9	73.4	98.0	95.9	142 out of 148
Training Set Mean		71.1	83.0	72.3	97.9	98.9	N/A
Training Std. Dev.		5.3	3.6	5.5	1.2	1.7	N/A
Training Set Median		70.1	82.4	71.5	97.8	99.7	N/A
7	'S01-18'	84.9	91.9	87.3	96.9	99.1	222 out of 224
8	'S02-410'	81.7	89.9	82.6	98.7	100.0	103 out of 103
9	'S02-484'	82.4	90.3	82.4	100.0	99.3	145 out of 146
10	'S03-2391'	84.0	91.3	87.4	95.6	99.1	105 out of 106
11	'S03-3178 D3'	77.3	87.2	77.3	99.9	92.2	47 out of 51
12	'S03-3178 D4'	86.7	92.8	87.0	99.5	98.6	70 out of 71
13	'S04-1840'	61.3	76.0	91.7	65.0	77.0	104 out of 135
14	'S04-910'	72.5	84.1	72.6	99.9	100.0	166 out of 166
15	'S07-1808'	86.0	92.5	86.4	99.5	100.0	132 out of 132
16	'S08-2215'	82.8	90.6	86.2	95.6	96.1	171 out of 178
17	'S09-2723'	71.4	83.3	71.9	99.2	100.0	52 out of 52
18	'S09-2909'	86.8	93.0	87.4	99.2	94.9	75 out of 79
19	'S11-1760'	76.1	86.5	77.7	97.4	98.0	248 out of 253
20	'S14-1715'	78.5	87.9	91.3	84.8	85.9	55 out of 64
21	'S14-2038'	76.0	86.4	81.6	91.7	91.3	157 out of 172
22	'S14-3414'	83.5	91.0	84.1	99.2	98.2	110 out of 112
23	'S15-1442'	83.9	91.3	84.4	99.3	97.8	131 out of 134
24	'S15-1518'	80.1	88.9	86.0	92.1	85.5	100 out of 117
25	'S16-1197 B1'	87.5	93.3	90.5	96.3	100.0	52 out of 52
26	'S16-1197 B2'	92.7	96.2	92.9	99.8	100.0	105 out of 105
27	'S16-1197 B3'	86.1	92.5	89.5	95.8	94.3	83 out of 88
28	'S16-1415'	69.9	82.3	70.1	99.5	100.0	44 out of 44
29	'S16-1467'	81.2	89.6	86.7	92.8	96.6	84 out of 87
30	'S16-567'	87.3	93.2	87.5	99.8	99.2	130 out of 131
Test Set Mean		80.9	89.2	81.9	96.2	96.0	N/A
Test Set Std. Dev.		7.0	4.5	7.7	6.8	5.8	N/A
Test Set Median		82.6	90.5	86.3	98.9	98.4	N/A

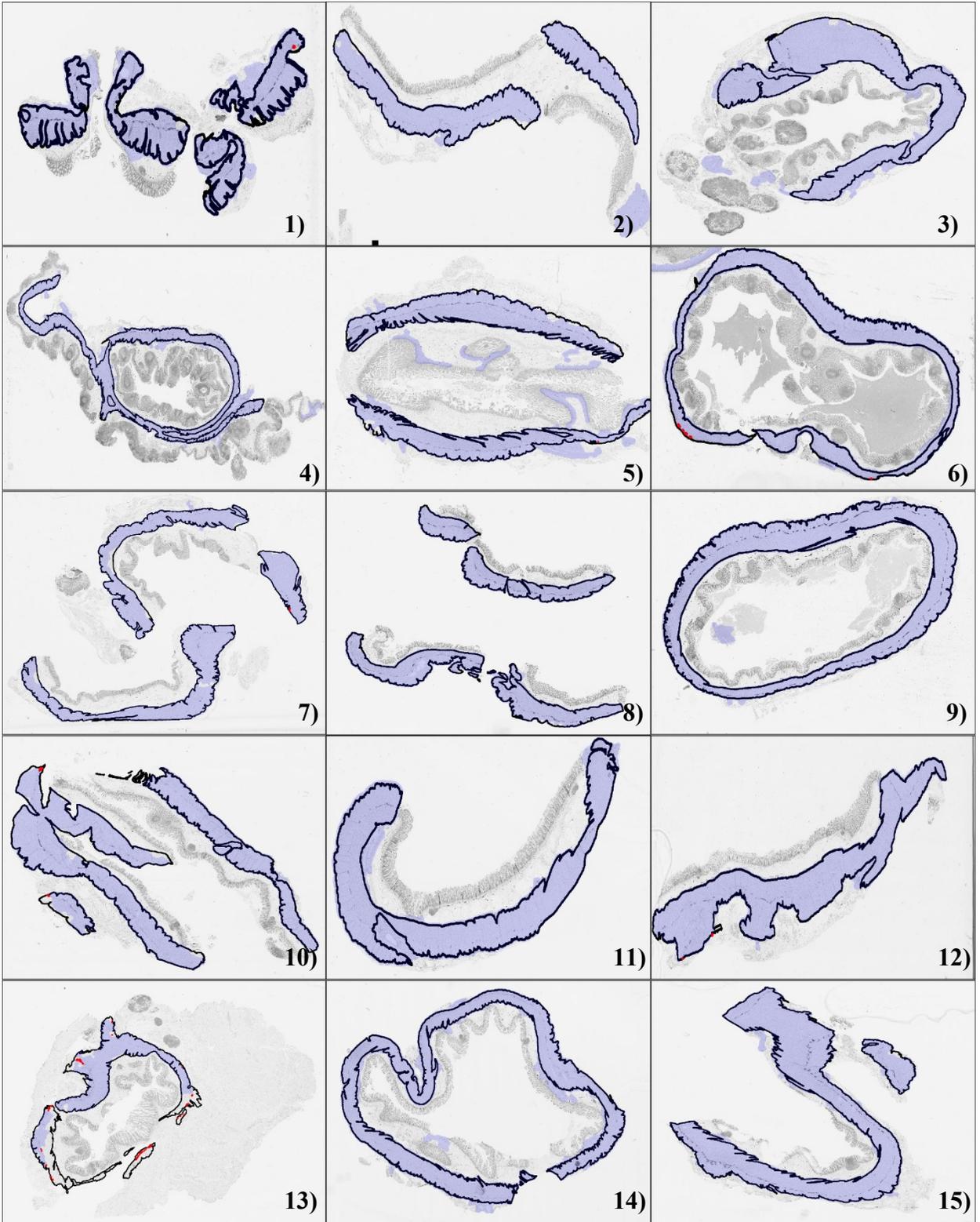


Figure 4.8: Visualizing the Muscularis Propria CNN Segmentation for images 1-15. Black outlines indicate the manual delineations of the muscularis propria, whereas the transparent blue shading depicts the segmentation by the k-means algorithm. Regions highlighted in red are the plexus regions not captured by the algorithm's segmentation

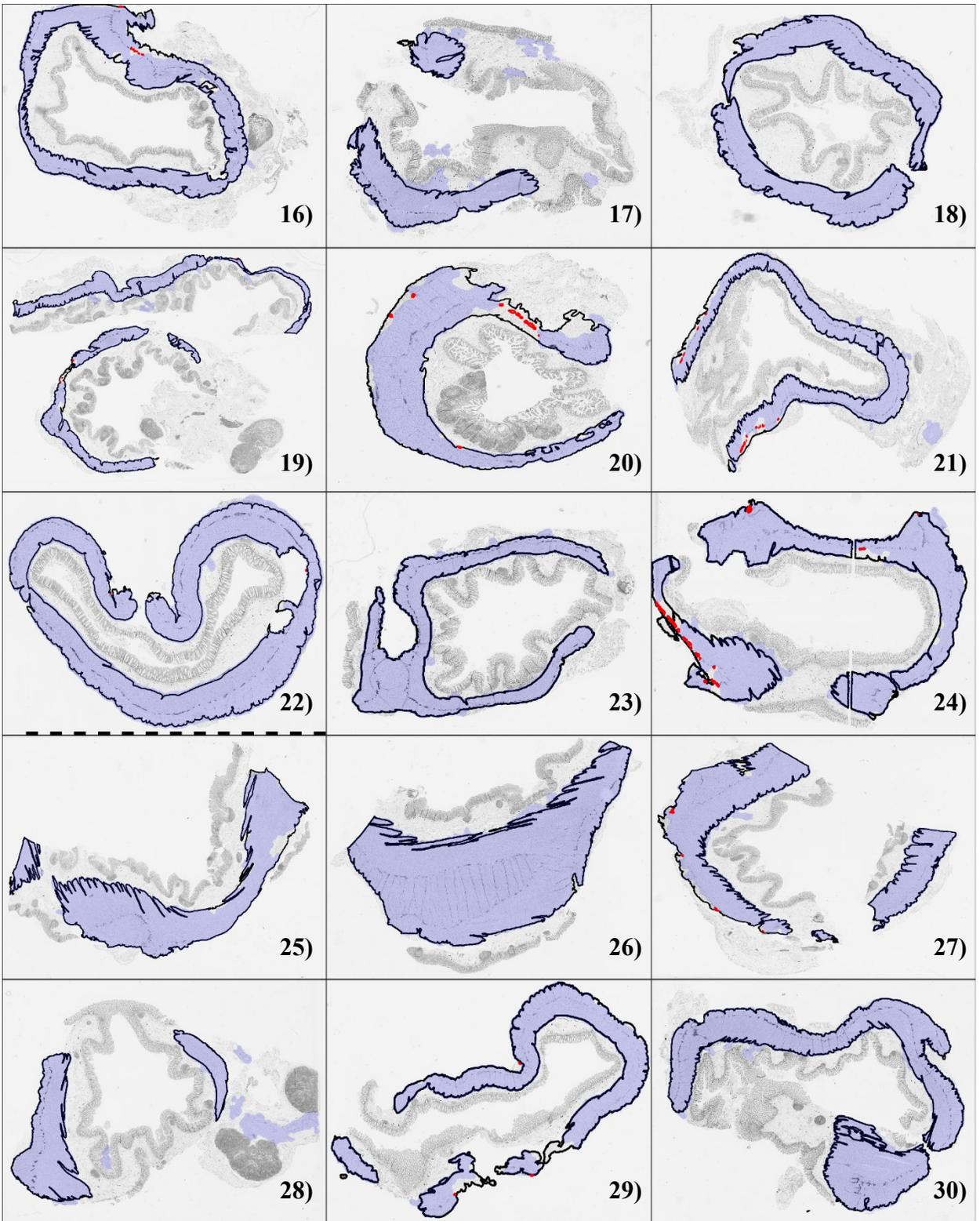


Figure 4.9: Visualizing the Muscularis Propria CNN Segmentation for images 16-30. Black outlines indicate the manual delineations of the muscularis propria, whereas the transparent blue shading depicts the segmentation by the k-means algorithm. Regions highlighted in red are the plexus regions not captured by the algorithm's segmentation.

4.3 Analysis & Discussion

The results presented in the previous section revealed several instances where the k -means clustering algorithm failed to segment the muscularis propria, skewing the average performance towards decreased scores. In Figure 4.10, we highlighted the four k -means clusters associated with the images where the automated segmentation failed. From this depiction, the muscularis propria is clearly identified by at least one cluster, although not necessarily associated with the second largest area. This implicates that across the entire dataset of thirty images, the k -means clustering algorithm successfully created clusters identifying the muscularis propria. However, since k -means clustering is an unsupervised learning approach which does not require training (and therefore has reduced computational complexity and costs), this method can generate clusters with increased variability in the structures it identifies. Clusters containing the muscularis propria, also included parts of the mucosa or highlighted some pixels from the submucosa to varying degrees. Since post-processing of morphological operations was empirically determined on a random sample of images, these extensive series of operations did not account for the variability between selected clusters and were found to be unsuitable for all images. Hence, the k -means clustering method failed to adequately segment the muscularis propria for a set of images and require further improvement for automated performance to be comparable to the CNN method. These refinements can include introducing known constraints on the clusters based on experimental extraction of data to reduce variability (turning this into a semi-supervised learning approach), and/or further generalizing the post-processing method.

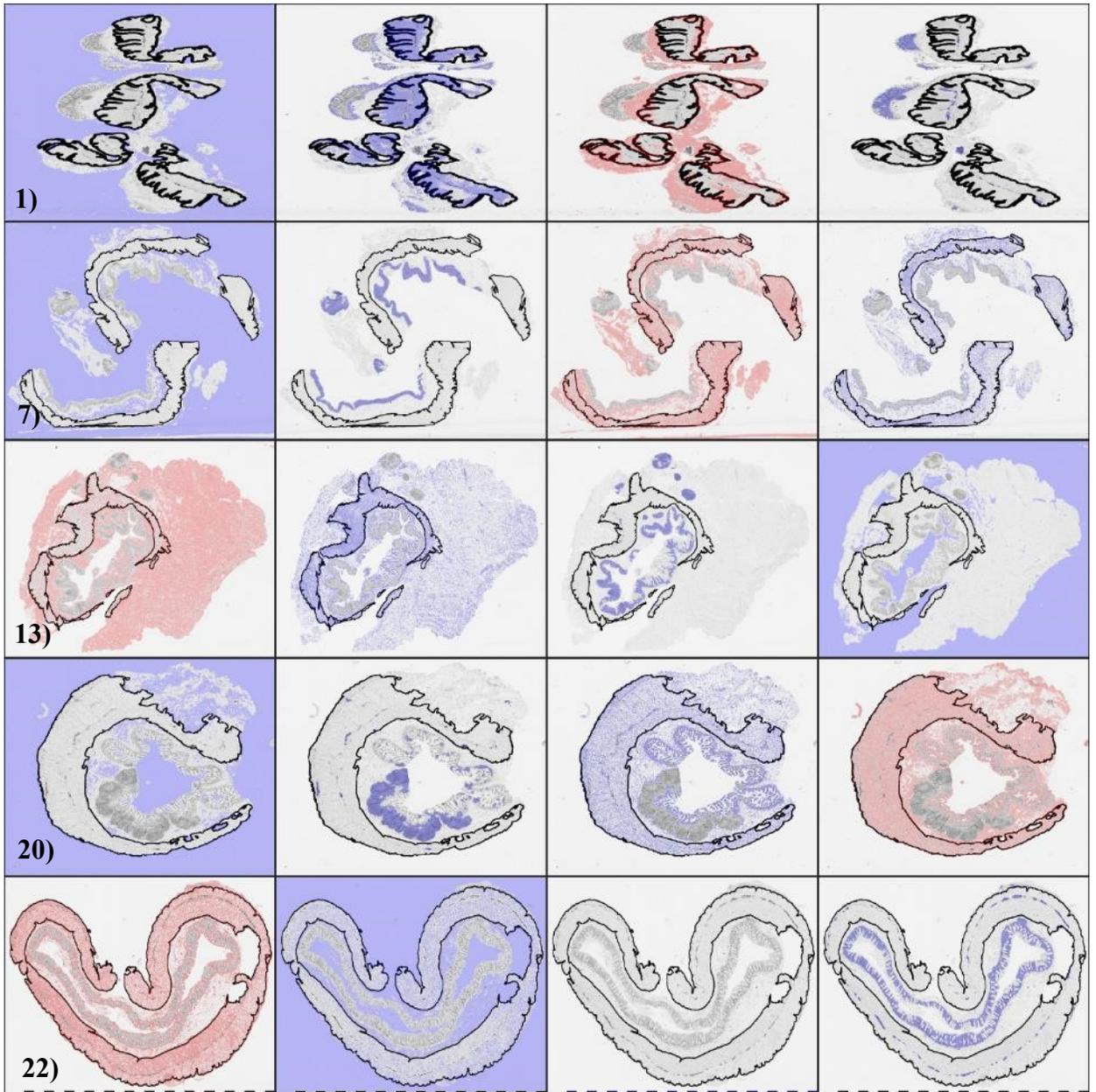


Figure 4.10: Depicting the raw output from the k-means clustering ($k = 4$) for five images with poor segmentation results. Shaded regions shows structures which fall into a cluster. Light red shading indicates the cluster with the second largest area, which does not always correspond to the muscularis propria.

The introduction of experimentally derived constraints on automated cluster selection (e.g., limits on RGB values relating to the muscularis propria) could also help resolve errors for images where the second largest cluster is not the most appropriate criterion for highlighting the muscularis propria. However, it would also be easy for the human operator to intervene as necessary and manually choose the correct cluster, which

was the semi-automated solution for the two images in this dataset (i.e., as illustrated in Figure 4.10, image 1 and image 13 have larger submucosal regions relative to the muscularis propria). User intervention could also allow for an adequate selection of the muscularis propria region by manually highlighting areas in the image through mouse input. This was the method used by both Schilling *et al.* [53] and Najjar *et al.* [52] to identify the submucosa and mucosa, respectively.

The CNN method also struggled with segmenting the muscularis propria while maximizing myenteric plexus inclusion for certain images. The postprocessing of the CNN model included a final dilation to capture these plexus regions near the boundaries of the muscularis propria. However, this could not account for all the images in the dataset, such as the cases for images 20 and 24. Here, myenteric plexus regions were found at the outer edges of the muscularis propria, either due to tissue tearing or abnormalities from specimen preparation. Some images, however, were found to have outlier characteristics from the rest of the dataset and proved to be difficult to segment for both the CNN and *k*-means methods. Such is the case for image 13. Although not ideal for a fully automated system, manual intervention could easily permit the human user to expand certain regions where the initial automated segmentation has missed myenteric plexus regions.

Despite these identified shortcomings, the modified CNN-based model outperformed the *k*-means algorithm on both the satisficing and optimizing metrics for the segmentation of the muscularis propria on this image dataset. The modifications to the CNN segmentations consistently provided higher plexus inclusion across the dataset. This is integral to the second part of the image processing pipeline where myenteric plexus regions are identified. Although both segmentations methods could achieve high recall for

many of the images, the CNN model proved to be more robust and could create segmentations of higher precision by at least 10% on average. This translates to a lower likelihood of including submucosal plexus regions in the muscularis propria segmentation. Although not further investigated in this thesis, a more accurate segmentation of the muscularis is valuable for determining the sub-circumferential distribution of ganglia. Additionally, there were also multiple cases where the *k*-means algorithm failed to segment the MP region. Based on the images where the MP was correctly segmented, the *k*-means model has the potential to achieve similar segmentation to the CNN-model. However, criteria for selecting the correct cluster needs to be refined and post-processing steps need further investigation.

4.4 Conclusion

This concludes the part of the thesis that addresses segmentation of the muscularis propria in Calretinin-stained histopathology colon images. In this chapter, we identified that a modification to McKean *et al.*'s CNN method [12] improved muscularis propria segmentation by nearly 20% for precision, recall and inclusion of myenteric plexus from the segmentations created by Kurian *et al.*'s *k*-means method [11]. Therefore, to achieve this thesis research objective, the CNN-based segmentations were used to proceed to the next step in the image processing pipeline, where myenteric plexus regions will be identified. This algorithm could segment the muscularis propria with higher fidelity to the manual segmentations displaying an average precision of $81.9\% \pm 7.8\%$, average recall of $96.1\% \pm 6.8\%$, and average plexus inclusion of $96.0\% \pm 5.8\%$ over the test set of 24 images.

Chapter 5: Segmentation of Myenteric Plexus Regions

Myenteric plexus, also referred to as Auerbach's plexus, are bundles of intrinsic nerves that branch along the intestinal intermuscular septum. These consist of ganglia and neuropil (a term describing the dense interwoven network of glial cells, axons and neural dendrites). In a Calretinin-stained histopathological cross-section, myenteric plexus regions appear as regularly dispersed brown-coloured pockets between the longitudinal and circular muscle layers of the muscularis propria. Figure 5.1 below highlights representative regions in a sub-circumferential portion of the cross-section. In this chapter, we present an algorithm to identify the myenteric plexus regions while utilizing the manual segmentation of the muscularis propria. This acts as the second step in the image-processing pipeline identified within the goals of this thesis.

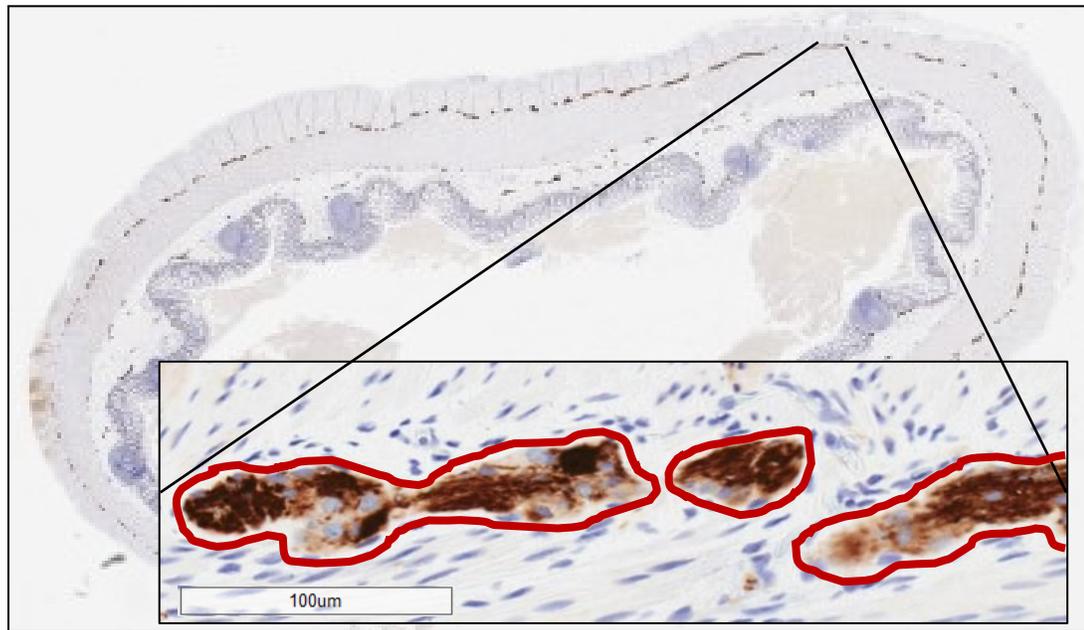


Figure 5.1: Red contours outline some myenteric plexus regions magnified to 100µm scale from image 9. These regions spotted fibrous dendritic tissue, light blue glial cells, and ganglia.

5.1 Methods

5.1.1 Colour Segmentation in RGB Space

To identify myenteric plexus regions, we propose a method to segment regions that appear brown within the muscularis propria. Brown, which is a composite colour, can refer to multiple hues ranging from red to yellow with low saturation or luminance [78], as seen in Figure 5.2. In optics, brown is a dark colour relative to a lighter background [79]. To isolate these brown-stained plexus regions, we must determine how computers depict variations of “brown” in a digitized image. In RGB colour-space, brown pixels are typically a mixture of strong input from the red channel with some input from green and minimal signal from the blue channel. Within the muscularis propria, the main structures include blue glial cells, brown plexus, and the near-white muscularis propria background tissue. Out of these, only the brown plexus has a strong red channel signal. This idea draws upon the Digital Image Subtraction Blue Enhancement (DISBE) method utilized in [11] and coincides with the method used by Najjar *et al.* to identify brown-stained pixels of mucosal nerve fibers in Calretinin-stained colon images [52]. Therefore, converting the images to another colour space was deemed unnecessary.

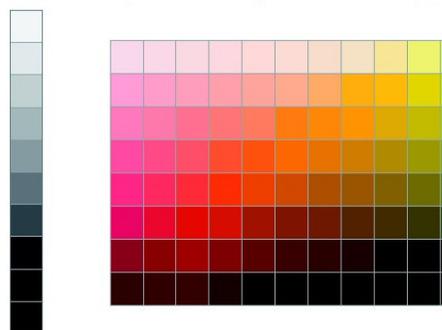


Figure 5.2: Munsell chips used as materials in most color-naming studies, obtained from [80]. Darker shades of these red, orange, and yellow hues appear as brown. (Open Access License Image)

Presented in Figure 5.3 is a flowchart describing the steps to segment myenteric plexus regions. It begins with downsampling the original image by 4 times ($1\mu\text{m}/\text{pixel}$ resolution) to reduce some computational costs. At this resolution, the smallest manually identified plexus is twenty pixels in area size ($20\mu\text{m}^2$). This also corresponds to the size of a small ganglion cell at the same resolution. Next, empirically derived colour thresholds are applied to isolate brown chromogen stain while reducing background noise pick-up from the muscularis propria. If the red and green channels are similar in value (i.e., red having a greater intensity than green by less than ten for an 8-bit image), then the pixels likely correspond to near-white colours and therefore, can be removed. From the remaining pixels, blue pixels can be removed if the red channel provides a greater signal than the blue channel by a threshold intensity difference of at least five in an 8-bit image. These threshold values were experimentally derived. The remaining pixels were masked with the segmentations of the muscularis propria to exclude purple regions from the mucosa and reduce the inclusion of submucosal plexus. We have summarized this procedure in the equations below.

$$BM_{RG}[i,j] = \begin{cases} 1, & \text{if } (I_R[i,j] - I_G[i,j]) > 10 \\ 0, & \text{otherwise} \end{cases} \quad (5.1)$$

$$BM_{RB}[i,j] = \begin{cases} 1, & \text{if } (I_R[i,j] - I_B[i,j]) > 5 \\ 0, & \text{otherwise} \end{cases} \quad (5.2)$$

$$BM_{\text{Brown chromogen}} = BM_{RG} \cap BM_{RB} \cap BM_{\text{muscularis}} \quad (5.3)$$

Where I_R, I_G, I_B are the matrices designated to the red, green, and blue channels making up the whole slide image, I . Equations 5.1 and 5.2 apply thresholds at each pixel

location, $[i,j]$, to create two different binary maps, BM . $BM_{muscularis}$ refers to the binary mask of the muscularis propria segmentation, which sets our search space for myenteric plexus. When we apply a Boolean intersection operation between these three different masks (Equation 5.3), the resulting binary map can isolate the brown chromogen-stained regions within the muscularis propria in image I . After greyscaling the resultant image of only brown pixels, faint background noise is minimized with a median filter using a kernel of 5×5 pixels and an Otsu threshold.

In the post-processing of the binary myenteric plexus mask, we wanted to remove objects associated with lightly stained muscularis or false positive stains of brown chromogen. Of the former, numerous small objects exist in the binarized image. In an attempt to isolate myenteric plexus regions, we applied an area-based Minimum Size Filter (MSF) to remove falsely detected background objects. We know from Miura *et al.*'s measurements that myenteric plexus nerve diameters in healthy colon are estimated to be $70\mu\text{m}$ in diameter [43]. Although the specimens in our dataset are obtained from the proximal end of resected colon, they may not necessarily be healthy tissue. In unhealthy colon, Miura *et al.* determined that plexus nerve diameters shrink in size. In our dataset, the smallest manually identified plexus is approximately twenty pixels ($20\mu\text{m}^2$) in size, which is also the size of a small, single ganglion cell. However, as seen in Figure 5.4, our plexus detection performs abysmally in terms of precision with a MSF threshold at $20\mu\text{m}^2$, likely due to a large quantity of small false positives. By increasing the MSF, we could reduce false plexus detection. As expected, this results in a trade-off between precision and recall. Ultimately, our goal is to keep plexus regions as search spaces for ganglia, while reducing the likelihood of false ganglia detections in the next stage of our image processing

pipeline. Therefore, an MSF threshold of forty pixels ($40\mu\text{m}^2$) was selected to maintain high ganglia inclusion scores across the dataset, while reducing the rate of false plexus detection. Although we may not detect some single ganglion objects, this MSF still provides a large tolerance for detecting smaller plexus regions that may be present in unhealthy colon tissue. Objects too large to be considered a plexus region (i.e., greater than the maximum size identified in the dataset) and likely corresponding to a staining error were also removed if they exceeded 90,000 pixels (0.09 mm^2). Finally, a dilation with a structuring element of a thirteen-pixel disk radius ($\sim 531\ \mu\text{m}^2$ in disk area) creates the final binary mask containing plexus regions and therefore, the search spaces for ganglia.

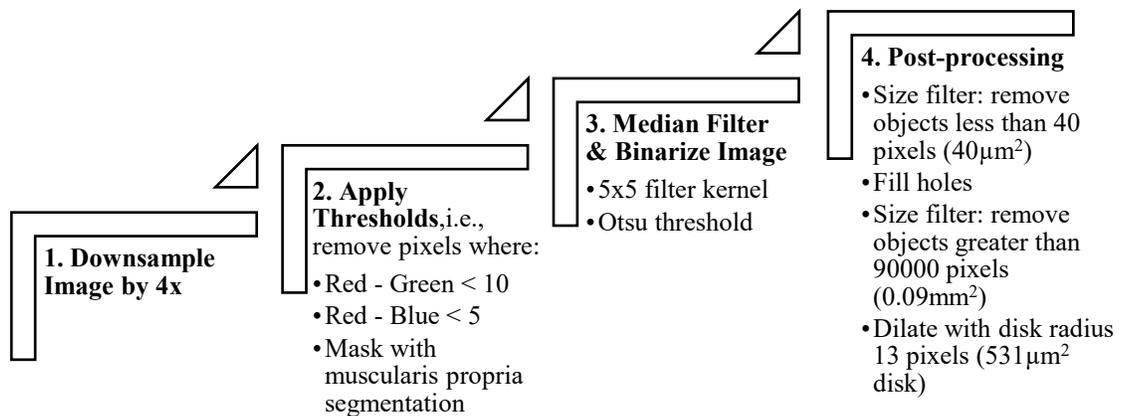


Figure 5.3: Flowchart of Myenteric Plexus Segmentation method

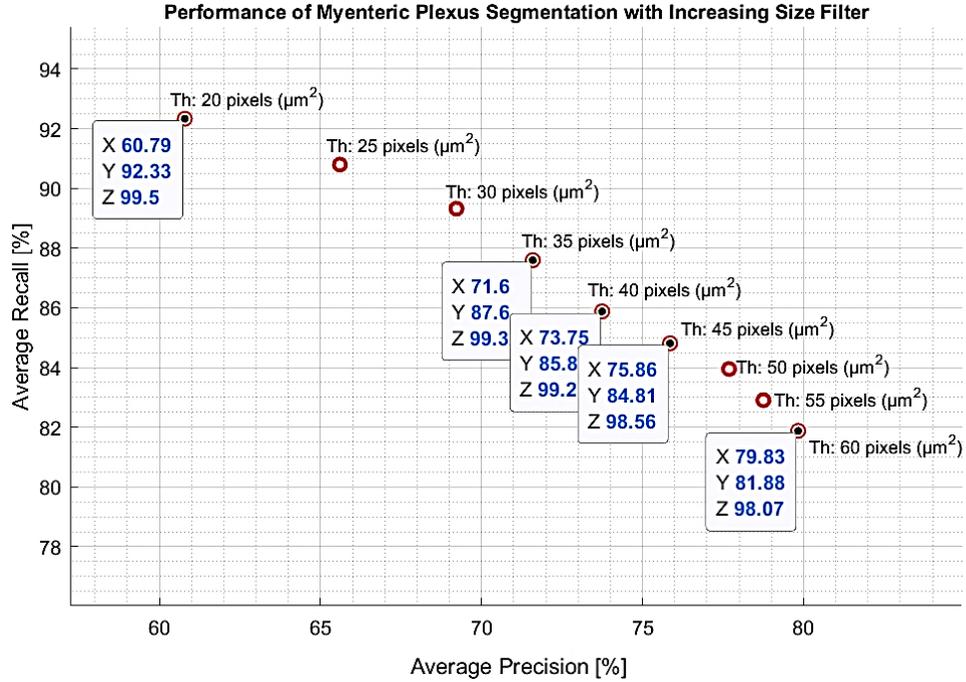


Figure 5.4: Increasing the Minimum Size Filter from 20-pixel ($20\mu\text{m}^2$) to 60-pixel ($60\mu\text{m}^2$) objects by increments of five pixels. This plot shows average precision and average recall across thirty images at each size filter. The z-axis plotted average ganglia inclusion, but remained relatively consistent at 97-98%, and therefore was not shown here. All thresholds are based on downsampled images.

5.1.2 Algorithm Evaluation Methods

Due to its fibrous structure and varying levels of stain absorption, the edges of myenteric plexus regions can be ambiguous and hard to define in some instances. Therefore, manual delineations of the myenteric plexus regions were loosely drawn at the original image resolution (i.e., $0.5\mu\text{m}/\text{pixel}$), with the aim of simply containing the plexus region within the boundaries. This was acceptable since the goal is to ensure ganglia are included in the final proposed search spaces, not necessarily accurately segmenting each plexus region. Hence, standard image evaluation metrics, such as Jaccard, would not be appropriate for assessing the performance of the segmentation algorithm. However, do note that when properly stained, myenteric plexus regions tend to be quite distinct in colour relative to the surrounding muscularis propria. This lends to the credence that if the algorithm does correctly identify a plexus region, then the segmentation should completely

encompass the structure. (As in the previous chapter, since the algorithm-generated plexus segmentation was conducted on downsampled images, the binary masks of the manual segmentation were also downsampled.)

Evaluation methods, instead, began by identifying objects within an image, as determined by connected components with a 4-neighbourhood connectivity. If an image object produced by the algorithm-generated segmentation intersected with at least thirty percent of an object in the manual segmentation, then it counted as a true positive. This intersection threshold is generous (and slightly arbitrary since lowering the threshold only increases the reported precision and recall metrics), but the purpose is to solely identify whether even part of the plexus region has been detected. We know that perfect, or even near-perfect, overlap will not exist between the algorithm-generated regions and the manual delineations. All objects in the manual segmentation that did not overlap with objects in the algorithmic segmentation, counted as false negatives. Objects present in the automated segmentation but missing in the manual segmentation, were considered as false positives. In edge cases where multiple objects in one mask overlap with one object in the other, the manual segmentation was always used as reference. For instance, if the algorithm identified multiple objects for one manually segmented plexus region, where the combined area exceeded the intersection threshold, then this counted as one true positive. If the algorithm identified one object when manually two were outlined, then two true positives are included in the evaluation. Instances where a larger algorithm object may intersect with a true plexus region and a staining error, were still counted as true positives (unless these detected objects exceeded the maximum size filter of 90,000 pixels or 0.09mm^2). Figure 5.5 provides a visual summary of these evaluation cases.

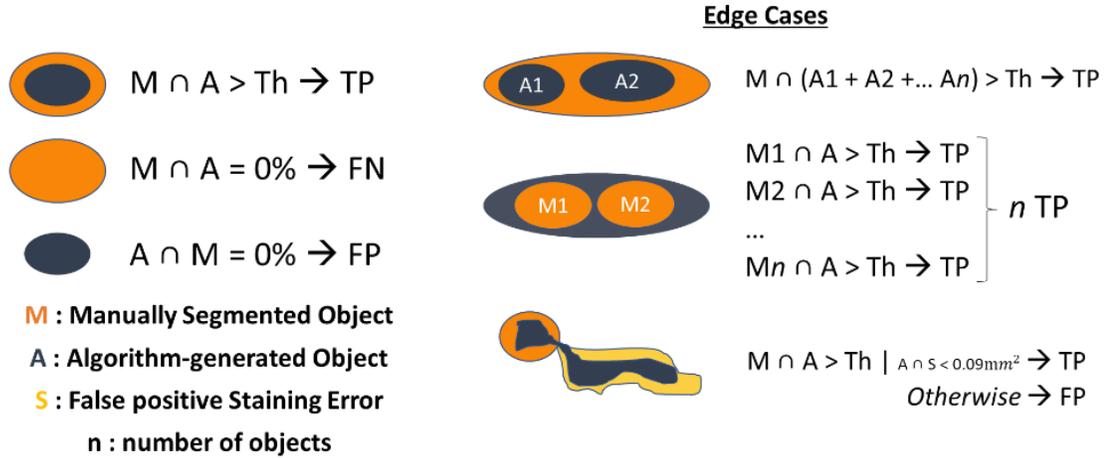


Figure 5.5: Evaluation cases for myenteric plexus segmentation. Generally, we count a true positive (TP) detection if the Intersection overlap is at least $Th = 30\%$. If there is no overlap, then we may call this a false negative (FN) or false positive (FP). Counts for edge cases are also illustrated.

From this, precision and recall could be determined. Precision, as shown in Equation 4.4, describes the proportion of objects correctly identified out of all the objects detected. Recall, as defined by Equation 4.5, quantifies the percentage of manually segmented plexus regions identified by the automated plexus segmentation.

Additionally, ganglia inclusion served as the satisfying metric for evaluating the plexus segmentation algorithm, with the aim to include all objects identified as ganglia by the primary expert pathologist. Here, ganglia inclusion is reported as a percentage reflecting true positive rate. It simply reflects the proportion of ganglia included in the plexus detection compared to the number of ganglia in the manual segmentation. High plexus recall guarantees high ganglia inclusion. Choosing object-based size filters (from Figure 5.4) aimed to optimize precision while maintaining high ganglia inclusion. This precision-recall trade-off ensured most plexus objects were correctly identified, while reducing the likelihood of introducing false ganglia-like object detection in the next segment of the image-processing pipeline.

5.2 Results

As shown in Table 5.1, the plexus segmentation algorithm performs with high ganglia inclusion at $99.2\% \pm 1.8\%$ across all the images in the dataset, satisfying the initial evaluation requirement. The optimizing evaluation metrics for detecting objects containing myenteric plexus regions are also listed in Table 5.1, with average precision at $73.8\% \pm 16.7\%$ and average recall at $85.9\% \pm 14.0\%$. Figure 5.6 plots the performance of the algorithm, as quantified by the optimizing metrics, for each image in the dataset. Reasons for varying performance for myenteric plexus segmentation are explored in Section 4.3.

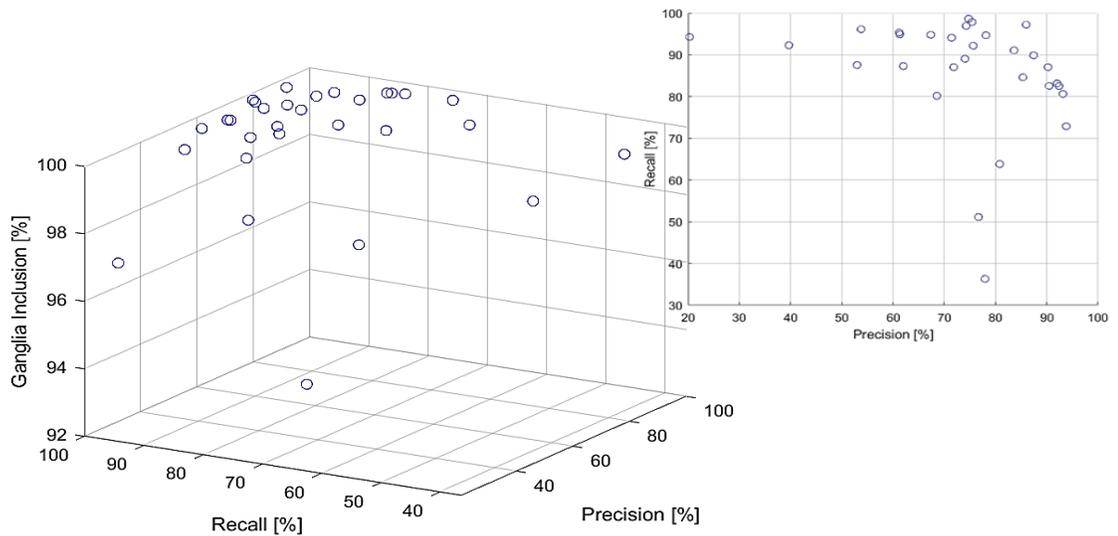


Figure 5.6: Visualizing the spread of results for plexus segmentation in the dataset. The smaller subfigure visualizes the segmentation results with respect to only the optimizing metrics.

Table 5.1: Plexus Segmentation Results

<i>No.</i>	<i>Filename</i>	<i>Precision [%]</i>	<i>Recall [%]</i>	<i>Ganglia Inclusion [%]</i>
1	'S00-1910'	92.0	83.1	100.0
2	'S03-3178 D2'	74.0	89.1	100.0
3	'S04-52'	71.8	87.1	92.0
4	'S07-1465'	53.0	87.6	100.0
5	'S14-580'	61.3	94.9	100.0
6	'S97-2054'	77.9	36.3	100.0
7	'S01-18'	80.8	63.8	100.0
8	'S02-410'	92.4	82.5	100.0
9	'S02-484'	86.0	97.2	100.0
10	'S03-2391'	68.6	80.2	100.0
11	'S03-3178 D3'	75.4	97.9	100.0
12	'S03-3178 D4'	62.0	87.3	100.0
13	'S04-1840'	76.7	51.1	98.3
14	'S04-910'	93.8	72.9	100.0
15	'S07-1808'	78.1	94.7	99.0
16	'S08-2215'	87.4	89.9	100.0
17	'S09-2723'	39.7	92.3	100.0
18	'S09-2909'	74.7	98.7	98.3
19	'S11-1760'	93.2	80.6	100.0
20	'S14-1715'	75.6	92.2	100.0
21	'S14-2038'	90.5	82.6	99.0
22	'S14-3414'	83.6	91.1	100.0
23	'S15-1442'	74.3	97.0	100.0
24	'S15-1518'	85.3	84.6	100.0
25	'S16-1197 B1'	53.8	96.2	100.0
26	'S16-1197 B2'	20.3	94.3	97.3
27	'S16-1197 B3'	71.4	94.1	100.0
28	'S16-1415'	61.2	95.4	100.0
29	'S16-1467'	90.2	87.1	95.5
30	'S16-567'	67.4	94.8	96.8
Mean		73.8	85.9	99.2
Std. Dev.		16.7	14.0	1.8
Median		75.5	89.5	100.0

5.3 Analysis & Discussion

As shown in Table 5.1, the myenteric plexus segmentation performed well on several images in this dataset, with the magnitude of errors attributed to smaller and faintly stained objects. The manual plexus segmentations for this dataset include all objects of varying shades of brown within the intermuscular septum, showing characteristics of a different blue glial cell pattern from the muscularis propria. However, many of these plexus candidates do not contain ganglia. Therefore, despite varying plexus segmentation performance, we could still maintain high ganglia inclusion across the entire dataset. Figure 5.7 displays the myenteric plexus segmentation of image 22 as an example of the successful segmentations within this dataset.

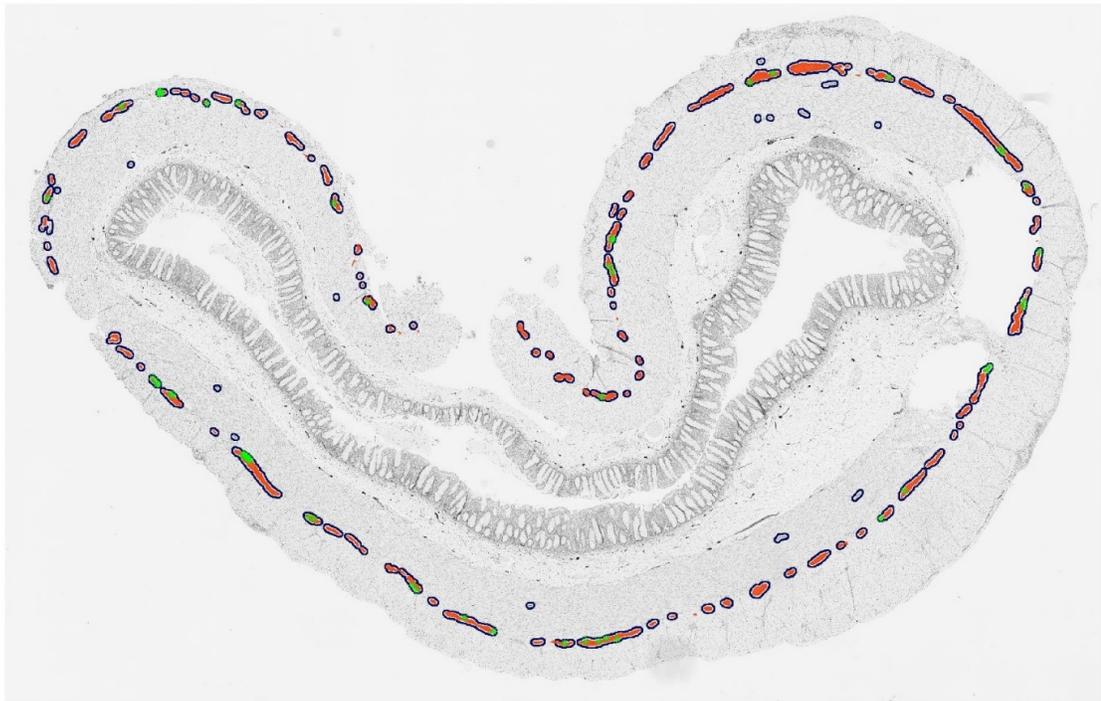


Figure 5.7: Image 22. Orange highlights the manual segmentation. The dark blue outlines are the computer-generated plexus segmentations. Green stars highlight included ganglia.

In three instances, the algorithm did not detect all manually segmented myenteric plexus regions and performed with low recall values. These images included 6, 7, and 13, which have recalls of 36.30%, 63.84% and 51.11%, respectively. As shown in Figure 5.8a-c, the chromogen weakly stained the myenteric plexus in image 6 and many regions appear almost indistinct to the human eye. As mentioned in Section 3.2.1, this image is an outlier within our dataset due to its low staining intensity for plexus and ganglia. However, we suspect it may not be an uncommon phenomenon in larger datasets, particularly if the brown chromogen staining intensity is associated with aganglionosis. Interestingly, for image 7, fibers within plexus regions still appear with high contrast but the entire region is not shaded brown. This likely introduces some ambiguity with some objects. Image 13 is similar to image 6 in that plexus regions are faintly stained, but ganglia appear with high contrast (which is an ideal combination for assessment by pathologists). For these sparse and faint staining results, this colour-based algorithm may detect a plexus region as multiple smaller objects. Therefore, we run the risk of not detecting true plexus regions and losing small ganglia due to the MSF. However, within this dataset, since precision and ganglia inclusion scores are high for these three images, the low recall performance does not prevent these segmentations acting as satisfactory search spaces for the next step in the image processing pipeline.

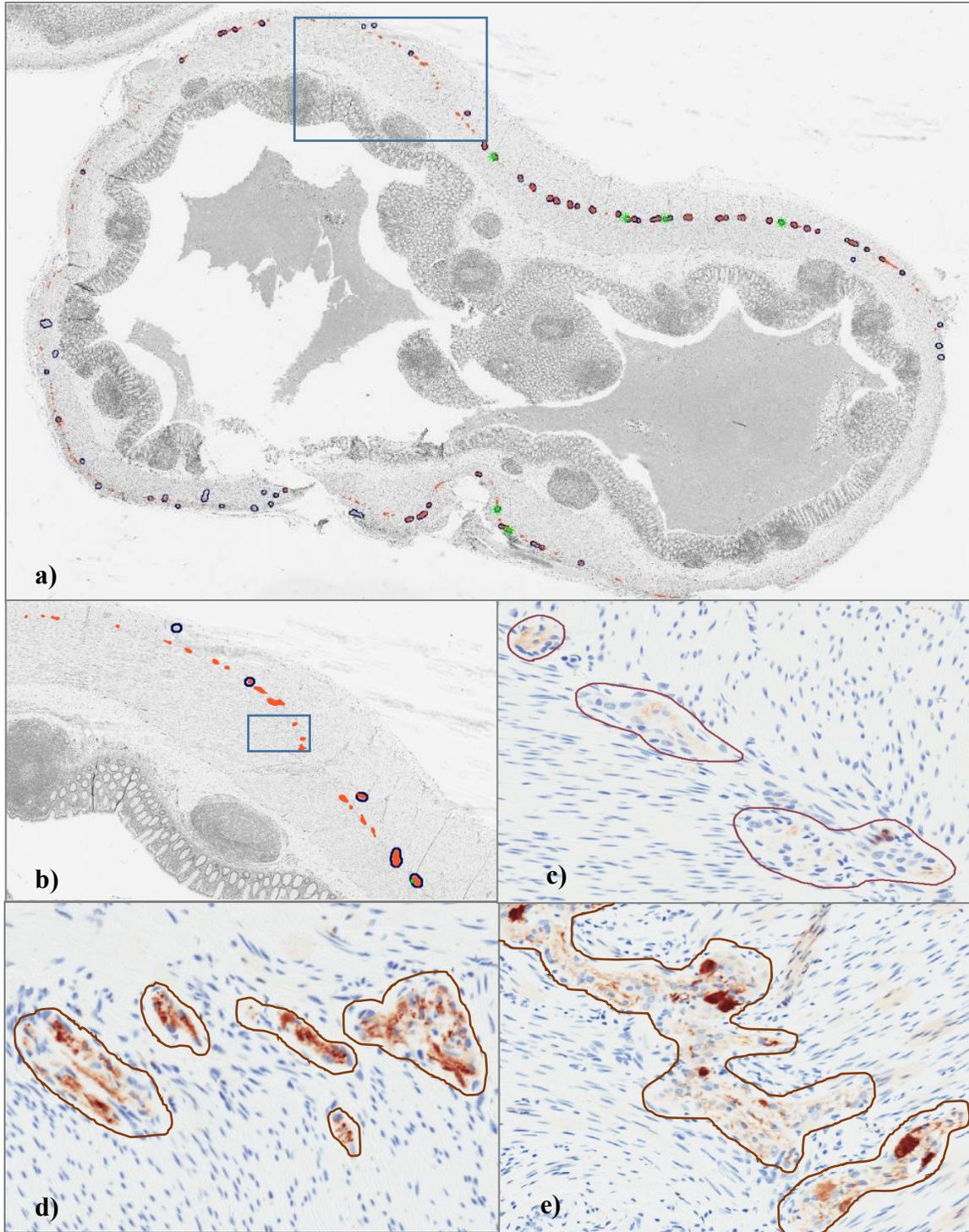


Figure 5.8: Segmentations with low recall (a) Plexus Segmentation Results of Image 6. Orange highlights the manual segmentation. The dark blue outlines are the algorithm-generated plexus segmentations. Green stars highlight included ganglia. (b) This is a magnified segment from the window in a). (c) This is a magnified segment of the original image at 100- μ m resolution from the window in b), with manual delineations outlining the plexus regions. The faint colour of the plexus is barely detectable to the human eye. (d) Manually outlined plexus regions from image 7. (e) Manually outlined plexus regions from image 13.

Plexus segmentations with low precision include images 4, 5, 10, 12, 17, 25, 26, 28 and 30. (These segmentations also meet the satisficing criterion with high ganglia inclusion and recall scores.) Out of these, a high false positive detection rate is expected for image 4 due to heavy staining errors present around the muscularis propria (as mentioned in Chapter 3). Yet unexpectedly, image 26 had the greatest high false positive count in the dataset. In Figure 5.10, we show the detection of objects where myenteric plexus regions are not present. For many cases, the false positives are due to an excessive number of staining errors that lightly shade folds or regions of the muscularis propria as brown. However, this is the poorest performing image (precision 20.29%). All other plexus segmentations with poor precision scores, are still reasonable (precision > 50%, save for image 17 which has a precision of 39.67%). These false positive objects are often small and due to minor staining errors. Since the evaluation provides equal weighting to each detected object, whether small or large, the presence of multiple small false detections negatively skews precision values. This is particularly true for images with few true myenteric plexus regions to begin with. Therefore, a poor precision score does not always reflect poor segmentation across the dataset. Although there is room for improvement, we did not explore other avenues of reducing false positives because it was uncertain if staining errors and falsely detected “plexus regions” contain objects similar in shape or intensity to true ganglia. However, we are aware introducing these false positive regions as search spaces for the next pipeline step can potentially induce false ganglia detection.

The errors that require careful consideration are those of the type that appear in Figure 5.9. Submucosal plexus regions usually encase ganglia. As mentioned in Chapter 2, within the transition zone the myenteric plexus regions may be devoid of ganglia despite

their presence in the submucosa. Thus, to prevent improper counts, it is imperative to exclude submucosal plexus from the segmentation. Currently, the inclusion of submucosal plexus regions is dependent on the segmentation of the muscularis propria. Future work may consider methods to refine the second step of the image-processing pipeline for improved capability of removing submucosal plexus regions.

Another consideration for future work includes the ability to detect plexus regions consisting of a single ganglion. The presence of isolated ganglion cells without surrounding neuropil is characteristic of the transition zone. The plexus segmentation algorithm currently filters out objects that are larger in area than single, small ganglion cells. For this dataset, such an assumption is acceptable due only a few isolated ganglia. However, for future iterations, this is a scenario that should be accounted for.

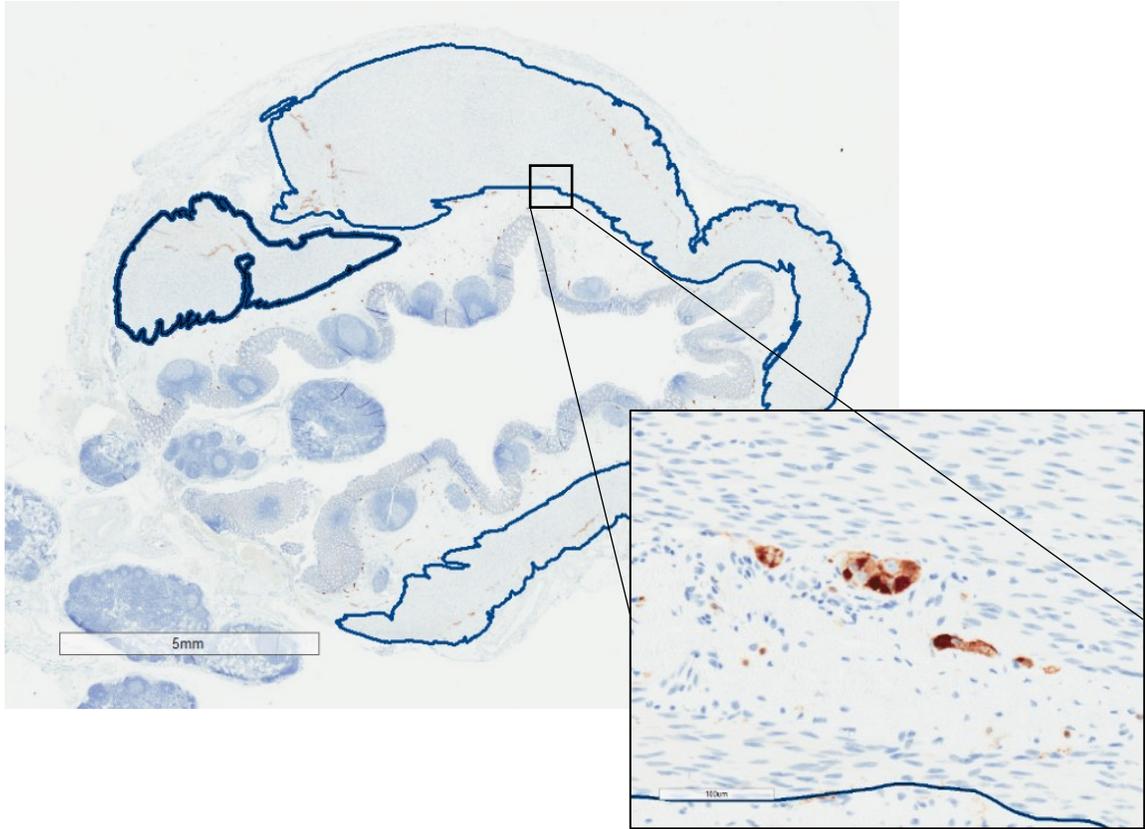


Figure 5.9: Image 3. Blue manual delineations outline the muscularis propria. The red box highlights a submucosal plexus region within the manual segmentation of the muscularis propria, containing five ganglia. The automated plexus segmentation detected this submucosal plexus, an error that will add to the false positive count in the final step of the image-processing pipeline.

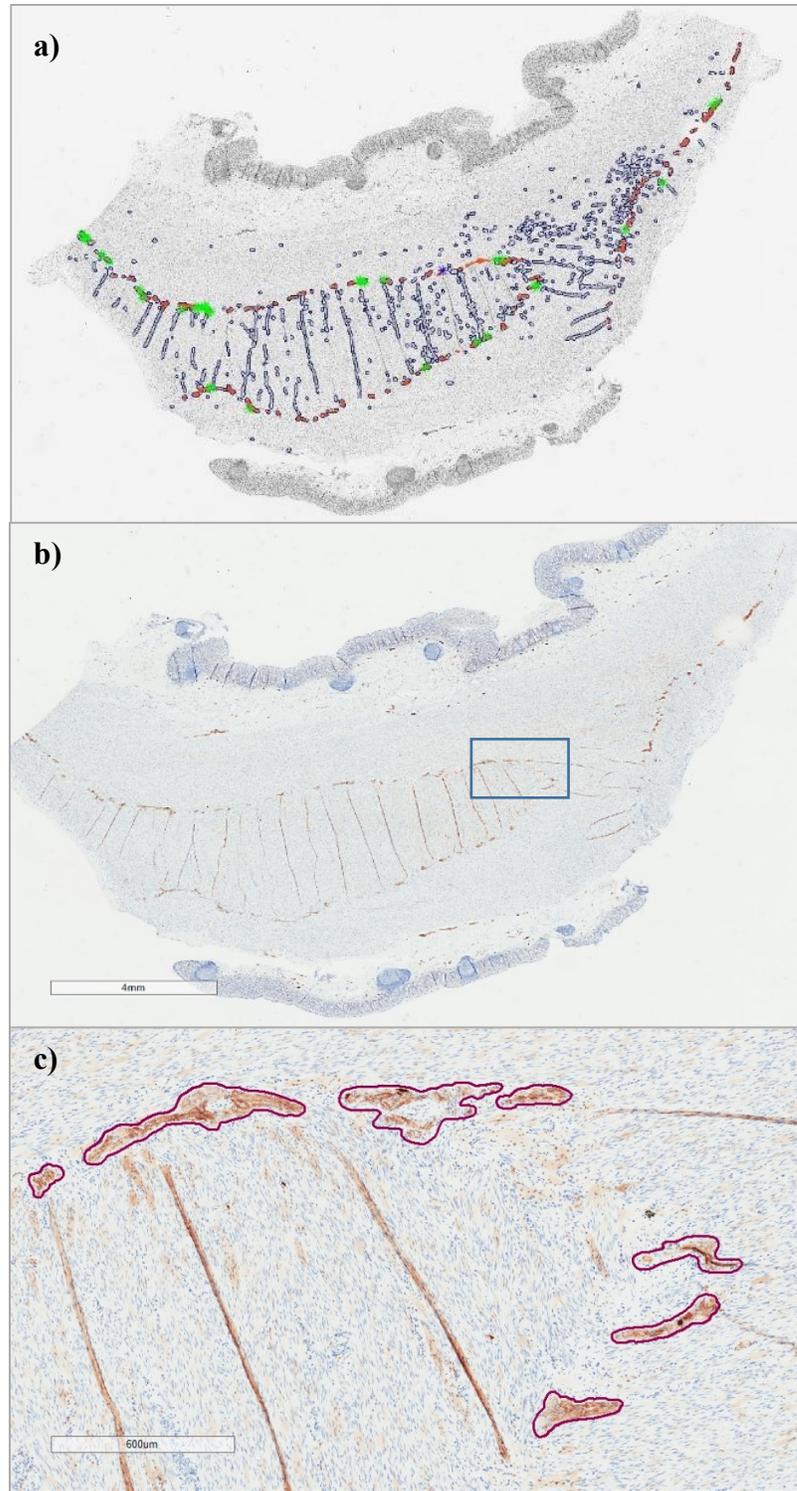


Figure 5.10: Image 26. (a) Image of the plexus segmentation results. Orange highlights the manual segmentation. The dark blue outlines are the computer-generated plexus segmentations. Green stars highlight included ganglia. (b) The original image. (c) Magnified to 600um scale bar resolution from the blue window in b). Plexus regions are outlined, but staining errors of folds and the background muscularis are also visible. These cause the excessive false positive detection as shown in a)

5.4 Conclusion

In this chapter, I describe the performance of a colour-detection algorithm for identifying myenteric plexus regions in Calretinin-stained histopathology colon images. The segmentation is the second part of the overall image-processing pipeline, providing a search space for myenteric ganglia. A quantitative evaluation and error analysis were also conducted by using manual segmentations as the reference for object overlap. The algorithm achieved an average precision of $73.8\% \pm 16.7\%$, an average recall of $85.9\% \pm 14.0\%$, and an average ganglia inclusion of $99.2\% \pm 1.8\%$ over the dataset of thirty images, when we utilized manual segmentations of the muscularis propria as the initiating search space. Missing plexus regions include smaller area objects, as well as regions where the brown chromogen of Calretinin stain was weakly projected. Images with greater instances of staining errors understandably overestimates the number of myenteric plexus regions present since the algorithm uses a colour-based detection method. However, given the near-perfect ganglia inclusion scores, we assumed these errors are not completely detrimental for the next step of the image-processing pipeline. Nevertheless, counting false positives as search spaces may skew the detection of total myenteric ganglia, particularly if submucosal plexus regions are included. This potential for error is captured in Chapter 7.

Chapter 6: Identification of Ganglia

This chapter describes the final stage of the image processing pipeline to identify ganglia within myenteric plexus regions. These nerve cell bodies relay the signals responsible for the innervation of colonic muscle. By automating the process to identify enteric ganglia, we can offer pathologists a tool to analyze digital colon cross-sections efficiently, as well as an aide in the standardization of seromuscular evaluation of HSCR-affected colon. With the extraction of size, colour, and texture-based features, we explore classical machine learning techniques to distinguish ganglia from the remainder of the myenteric plexus region. Figure 6.1 below provides an example of a ganglion cell detected within a myenteric plexus region.

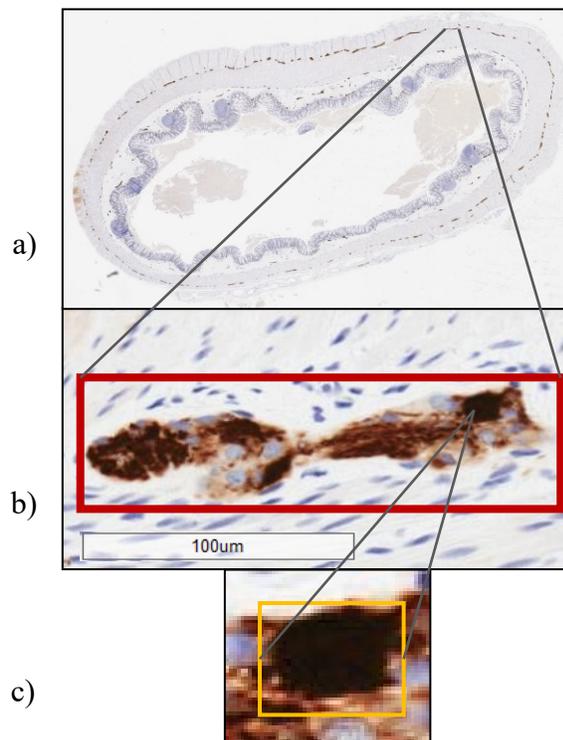


Figure 6.1: a) Image 9 from dataset b) Red bounding box contains a myenteric plexus region magnified to 100µm scale resolution c) Digitally magnified, the yellow bounding box contains a single ganglion cell.

6.1 Methods

We implemented a two-step approach to identify ganglia within myenteric plexus regions. First, an intensity-based method was employed to segment candidate objects within manually labeled myenteric plexus regions. Then, we proceeded to extract and select features from these candidates for cross-validated classification, using manual labels of ganglia for training. Lastly, an evaluation of the overall precision and recall of this two-step approach was conducted to determine algorithm-identified myenteric ganglia.

6.1.1 *Ganglia Candidate Selection*

Within myenteric plexus regions, enteric ganglia containing Calretinin are typically expressed darker in brown colour relative to its surrounding. Although other features are also considered for Calretinin-stained images, pathologists use the deep colour of ganglia as the most obvious distinguishing characteristic. Based on this knowledge, we selected ganglia candidates using percentile intensity thresholds. Within a single image, all brown pixels among the myenteric search spaces were identified using the colour thresholds described by Equations 5.1 and 5.2. Amongst these, we isolated pixels with a greyscale intensity less than or equal to the fifth percentile (i.e., the darkest 5%) and re-mapped these pixels onto a binary mask. The intensity threshold was experimentally derived and chosen to include as many candidates that corresponded to manually segmented ganglia as possible, while attempting to remove the dendritic-like extensions attached to these specific candidates. This trade-off also considered the number and shape of false ganglia candidates when decreasing intensity thresholds, with the aim of improving class-imbalance ratios for subsequent classification.

Due to the sheer number of false positives in this binary mask, some post-processing was required. Objects less than 80 pixels ($20 \mu\text{m}^2$) in size were removed, as the smallest manually segmented ganglion in this dataset exceeds 100 pixels ($25 \mu\text{m}^2$) at the original image resolution (i.e., $0.5\mu\text{m}/\text{pixel}$). From here, objects were filtered into three categories based on size, using estimates derived from sampling ganglia among the manual segmentations (as shown in the histogram from Figure 6.2). A opening morphological operation, with disk radius of three pixels (disk area of $7.01 \mu\text{m}^2$), was applied to objects less than 1000 pixels ($250 \mu\text{m}^2$) in size. This was to remove dendritic-like extremities that remained thinly attached to ganglia objects after the intensity thresholding, to reduce the likelihood of misclassification. The unintended effects result in some smoothing of false candidates within this same size category.

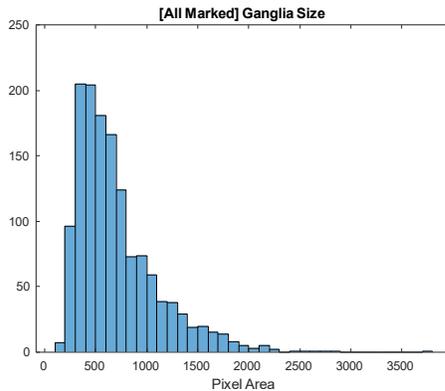


Figure 6.2: Histogram of Ganglia Sizes Sampled from Manual Segmentations Across the Dataset

Since objects greater than 1000 pixels ($250 \mu\text{m}^2$) in size were more likely to contain multiple ganglia in proximity or have extensions that cannot be removed with small, simple morphological operations, we extended post-processing steps to address these instances. We aimed to incrementally increase the extent of separating ganglia clusters with increasing candidate object size. Although the size categories are slightly arbitrary, they prevent over-segmentation of smaller objects (which would induce a greater number of

false negatives). If objects were between 1000-1500 pixels ($250\text{-}375\mu\text{m}^2$) in size, we applied a Chan-Vese active contours model [82] with boundary regularization (i.e., a contraction bias of 0.3 where a value of 1 is the maximum) on the contrast-enhanced [83] colour sub-image. The original binary ganglia candidate object acted as the point of initialization for the sub-image. Then, we applied an opening with disk radius of three pixels (disk area of $7.01\ \mu\text{m}^2$) and removed any connected components smaller than 30 pixels ($7.5\ \mu\text{m}^2$) in size. For objects larger than 1500 pixels ($375\ \mu\text{m}^2$), the contraction bias of the active contour was increased to 0.65. If the object exceeded 2000 pixels ($500\ \mu\text{m}^2$), the region-based active contours was followed by a Euclidean-distance watershed transform [84], using the extreme-most distances to increase the chances of separating ganglia clusters without over-segmentation. Objects smaller than 30 pixels ($7.5\ \mu\text{m}^2$) were filtered out. From here, we stitched sub-images to create the refined ganglia candidate mask. Figure 6.3 summarizes the post-processing in flowchart form.

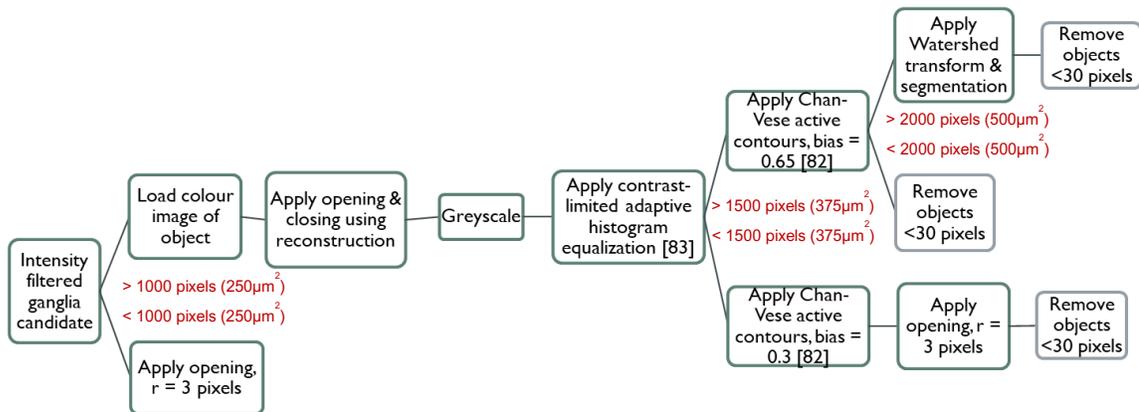


Figure 6.3: Flowchart depicting Ganglia Candidate Detection Post-processing Steps

6.1.2 Feature Selection

As described in Chapter 3, ganglia in these Calretinin-stained images are usually dark brown in colour, round, and smooth in texture (i.e., changes in gradient are hardly noticeable). Therefore, we considered features that could capture and quantify these characteristics, namely relating to shape, colour, and texture. Ideally, these quantitative descriptors can separate true ganglia from background plexus, either from individual attributes or a combination of features. Given that the shape of the feature vector defines the space in which data is situated and plays a role in how decision boundaries are formed, we know that simply increasing the number of features can increase sparsity in feature space, as well as model complexity (i.e., the number of parameters defining the model). This can lead to complex decision boundaries that do not generalize well to future data [85]. Feature selection methods were briefly explored to determine which quantitative descriptors could efficiently lend to the distinction between true ganglia and false candidates. Table 6.1 lists all the features extracted from ganglia candidates, as well as a summary of chosen attributes from various feature selection methods. Feature selection was conducted on the entire dataset using 10-fold cross-validation, and z-score standardized. We recognize this approach may not yield the most optimal solution in feature space, but rather aimed for a reasonable one instead.

Based on methods to maximize variance while reducing redundancy [86], five features for ganglia candidate ranked highly. These, along with features found from a correlation-based feature selection method [87], provided a total of nine attributes selected as input for classification: 1) area in pixels, 2) minimum feret diameter, 3) standard deviation of the saturation channel in HSV colour space, 4) average of the value channel

in HSV space, 5) standard deviation of the value channel in HSV space 6) average b* channel value in L*a*b colour space, 7) gradient contrast, 8) gradient homogeneity, and 9) median gradient. Additionally, 10) maximum feret diameter and 11) circularity were included. Although the results in Table 6.1 do not indicate reasons for inclusion, an initial error analysis of experiments revealed a large subset of false positives that we tried to reduce by including these two features.

Table 6.1: Comparing Feature Selection Methods

Feature Type	No.	Features Extracted	Feature Selection Methods		Selected Features
			Max. Variance Min. Redundancy (MATLAB)	CfsSubset Evaluator & BestFirst Search (Weka)	
Shape	1	'Area'	Rank 1	100%	X
	2	'Eccentricity'			
	3	'Convex Area'		40%	
	4	'Circularity'		0%	X
	5	'Euler Number'			
	6	'Equivalent Diameter'		40%	
	7	'Solidity'			
	8	'Perimeter'		40%	
	9	'Maximum Feret Diameter'		0%	X
	10	'Minimum Feret Diameter'		100%	X
Colour	11	'Mean Intensity'			
	12	'Std. of Intensity'			
	13	'Mean Hue in HSV space'			
	14	'Std. of Hue in HSV Space'			
	15	'Mean Saturation in HSV space'			
	16	'Std. of Saturation in HSV Space'	Rank 2		X
	17	'Mean Value in HSV space'		100%	X
	18	'Std. of Value in HSV Space'		80%	X
	19	'Mean a* channel value in L*a*b space'			
	20	'Std. of a* in L*a*b Space'			
	21	'Mean b* channel value in L*a*b space'		50%	X
	22	'Std. of b* in L*a*b Space'			
Texture	23	'Gradient Contrast'	Rank 3	70%	X
	24	'Gradient Correlation'			
	25	'Gradient Energy'			
	26	'Gradient Homogeneity'	Rank 4	100%	X
	27	'Median Gradient'	Rank 5	100%	X

Figure 6.4 depicts the data spread according to class labels by comparing each feature to one another. As is the case with most real-life data, we do not see clear separation between our classes in two dimensions. However, certain features (like median gradient) and combinations (such as 11) circularity versus 1) area), seem to better distinguish true ganglia from false candidates.

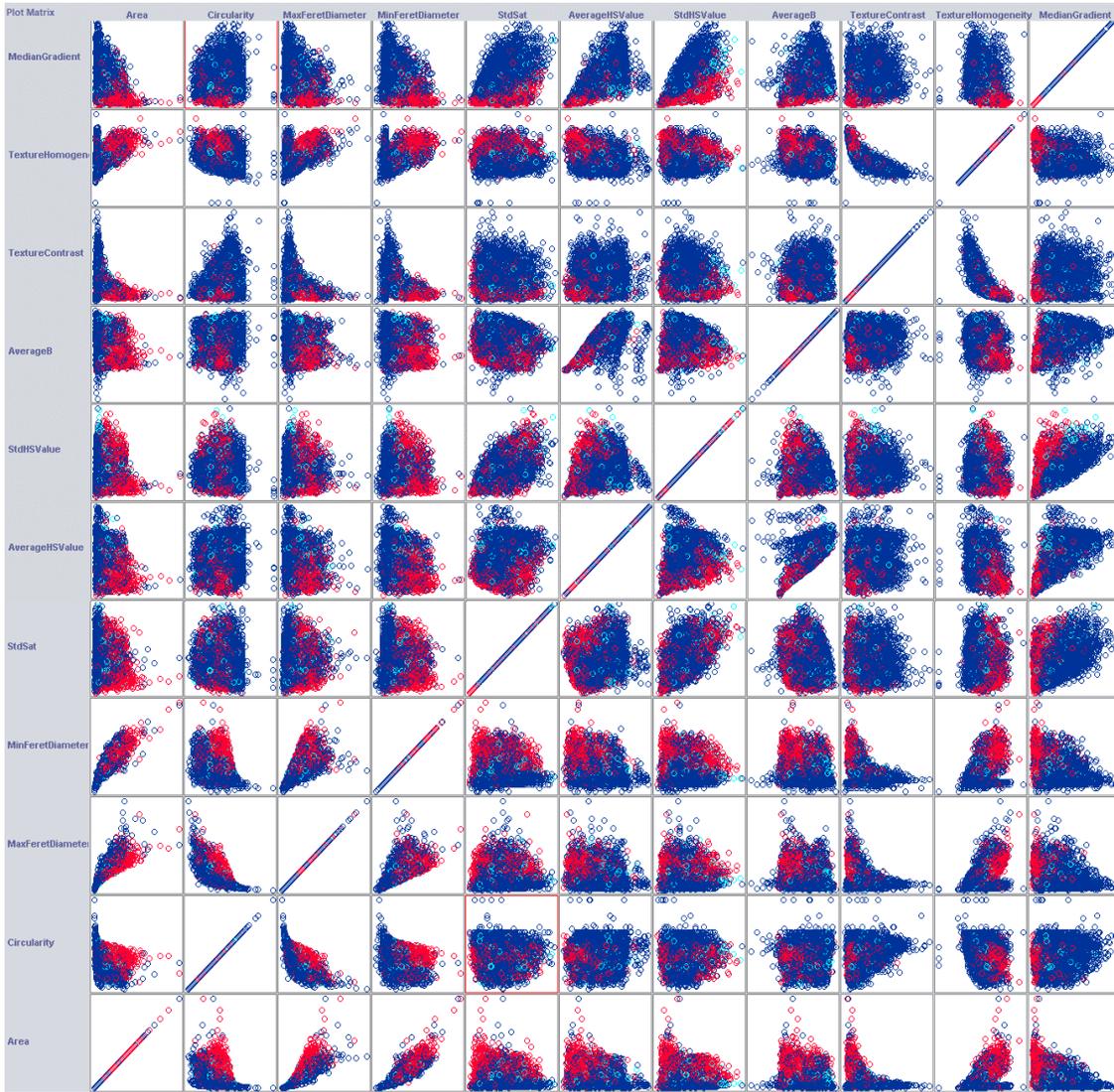


Figure 6.4: Visualizing Spread of Data for Selected Features in Weka. Red indicates datapoints belonging to ganglia class of high certainty; cyan is ganglia class of low certainty; dark blue are false candidates

6.1.3 *Classification*

Ganglia candidates were labeled according to object overlap with manual segmentations. Objects could correspond to ganglia of high certainty, ganglia of low certainty, or neither. Although there are three labels, only two of those classes truly exist: ganglia or not ganglia. Ganglia of low certainty are manual segmentation labels referring to objects that have some resemblance to a ganglion structure, to one extent or another, but not enough to make it unequivocal. Often objects with this label are dissimilar to each other as well. We can also see from Figure 6.4 that this set of data points does not have its own set of clusters but is, rather, interspersed between the other two classes. However, we do note that severe skew in the data may be contributing to this observation. Therefore, misclassification of ganglia with low certainty is not equivalent to misclassification of the other two classes. Instead, it acts as an error buffer, a free-for-all region. We could go so far as to say that should a classifier claim an object with this low-certainty label as a ganglion, or as not a ganglion, then it will not count as a misclassification. This simplifies our multi-class problem to a binary-class one, allowing training to be conducted on only expert labels of ganglia with high certainty and those not belonging to the ganglia class.

To quantify an estimate of the combined discriminatory power between classes within the selected feature space, we applied linear discriminatory analysis (LDA) as a classification technique in this initial investigation phase. Although other classifiers (such as CART, Random Forest, SVM, kNN, and shallow artificial neural networks) were briefly explored, these classical methods were seen to provide similar results. Succinctly, we recognize that “[t]he quality of a model for a given classification problem is often a function of the degree of agreement between the characteristics of the data and the

underlying assumptions of the model” [85]. The LDA is advantageous due to its relative simplicity; computational costs are lowered and the use of a classifier that generates simpler decision boundaries means an increased feature dimensional space can be defined without generally overfitting.

To estimate classification performance, we applied leave-one-out (LOO, or more appropriately leave-one-image-out) cross-validation. Since each image in our dataset has a varying number of manually segmented ganglia, and pathologists analyze images independently from each other, we are interested in quantifying performance across each case. It is true that without a separate test set, we cannot measure the degree of overfitting. However, given how drastically different performance could vary between images, choosing a test set from our dataset of thirty images that is representative and calibrated to actual probabilities is impractical. It would provide an unreliable prediction of actual performance.

As mentioned earlier, the number of candidates that do not correspond to ganglia can be excessive. For many of the images in this dataset, we see a large class imbalance between true ganglia candidates (with high and low certainty) and false candidates. To address this skew during training, we opted to use random undersampling of the false candidate class for computational efficiency. Bootstrapping afterwards accounts for the variance that arises from various false candidates. Although SMOTE [88] and ADYSYN [89] are methods to adjust the positive minority class and provide greater stability, the thesis author noted similar results to the random undersampling and chose this method due to its lower computational runtimes.

6.1.4 Algorithm Evaluation Methods

We defined a ganglion as ‘detected’ when the ganglion candidate is larger than 30% of the manually segmented ganglion that it overlaps with. This threshold catches over-segmented candidate objects as a detection error and prevents it from actualizing as a classification error (We wanted to measure the extent to which classification can correctly identify ganglia while ensuring other objects are not counted as such). After segmentation, we tabulated the number of detected ganglia (true positives) out of the total number of manually labelled ganglia, separately for both high certainty and low certainty labels. This also allowed us to determine precision and recall (i.e., detection rate) of the segmentation algorithm. For edge cases where multiple ganglia are captured in one merged object, we used the manual label as the reference. The number of merges were also tallied.

For the classification stage, scalar metrics, such as accuracy, precision and recall are appropriate. However, the first two are negatively influenced by class imbalance. Within our dataset, ganglia candidate skew exists to varying degrees across the images. Therefore, to be able to compare performance between images, we needed to correct for prevalence. Equation 6.1 and Equation 6.4 describe how recall and prevalence-corrected precision were determined [90].

Additionally, area under the receiver-operating curve (AUROC) values represent the degree of separability between classes. It states how well the model can distinguish between true ganglia and false candidates. AUROC is resistant to variance and bias by skew, making it advantageous for comparing models [91]. We included it here for comparison with future models and iterations. However, like prevalence-corrected precision, AUROC can mask poor classifier performance mainly because it does not

account for true negatives [92]. Equation 6.5 describes the trapezoid method of calculating AUROC in terms of true positive rate and false positive rate [90].

Thus, Matthew's correlation coefficient (MCC) was also examined for each image to confirm performance. MCC is a discretization of the Pearson correlation coefficient. It is a measure to describe the extent to which binary classifier predictions are linearly correlated to actual labels. A value of -1 indicates negative correlation, meaning predictions disagree with labels; 0 is no correlation, indicating predictions seem to be random; +1 is strong positive correlation stating that predictions match labels (i.e., zero false positives or false negatives). Although it is unnecessary to maximize MCC values to obtain high precision and recall, we wanted to see if classification can provide a strong positive correlation and is not random. MCC is also resistant to biases arising from skew but less likely to mask poor classifier performance due to its symmetric properties [90], [92]. Equation 6.6 provides the formula for MCC.

Table 6.2 lists all the equations applied to calculate the classification metrics. An estimate of the confidence level for these metrics were determined using bootstrapping with $n = 1000$. (The value for n was chosen to be large enough to provide higher confidence without excessive computational times.) In edge cases, such as merged objects, the manual segmentations were used to determine counts.

Table 6.2: Equations used to determine Classification Metrics.

Recall/ True Positive Rate	$R = TPR = \frac{TP}{TP + FN}$	(6.1)
False Positive Rate (FPR)	$FPR = \frac{FP}{TN + FP}$	(6.2)
Precision	$P = \frac{TP}{TP + FP}$	(6.3)
Prevalence- corrected Precision	$pcP = \frac{TP}{TP + c \cdot FP}$	(6.4)
Area Under the Receiver Operating Curve (AUROC)	$AUROC = \frac{TPR - FPR + 1}{2}$	(6.5)
Matthew's Correlation Coefficient (MCC)	$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$	(6.6)

TP = true positive; candidate corresponding to the ganglion label, is classified as such
 TN = true negative; candidate corresponding to the not-ganglion label, is classified as such
 FP = false positive; candidate corresponding to the not-ganglion label, is misclassified as a ganglion
 FN = false negative; candidate corresponding to the ganglion label, is misclassified as not a ganglion
 c = prevalence (ratio of positive class to negative class)

6.2 Results

We evaluated the ganglia identification algorithm for:

- 1) its ability to segment candidates from plexus regions,
- 2) ability for LDA classifier to separate ganglia candidates from true ganglia and background noise, and finally,
- 3) its overall performance on each image.

Shown below in We also tallied the number of instances when ganglia remained merged within a single candidate object. This reflects areas of improvement for segmentation post-processing.

is the recall, or detection rate, for ganglia (objects as defined by expert observer #1) using the candidate-segmentation algorithm described in Figure 6.3. Across the images in

the dataset, detection average for ganglia objects marked with high certainty reaches $89.8\% \pm 10.9\%$. As expected, the average detection rate is lower for ganglia objects labelled with low certainty, sitting at $77.1\% \pm 24.9\%$. This brings total detection of manually segmented ganglia to $87.8\% \pm 11.5\%$. Precision is also reported, with an overall performance at $20.3\% \pm 14.8\%$. However, since the objective of this segmentation step was simply to generate ganglia candidates that can be filtered by a classifier, recall became the metric to optimize. We also tallied the number of instances when ganglia remained merged within a single candidate object. This reflects areas of improvement for segmentation post-processing.

For the next step in this algorithm, performance metrics were tabulated in Table 6.4. The LDA classifier achieves an average recall of $92.2\% \pm 6.3\%$ across all the images in the dataset. We corrected precision for negative class prevalence to allow fair comparison between images and found an average dataset value at $89.5\% \pm 7.2\%$. Average MCC sits at $69.6\% \pm 14.11\%$ indicating that there is a strong (and for some images, a very strong) positive correlation between classifier labels and manual segmentations. Figure 6.5 plots the performance of the classifier, as quantified by recall, prevalence-corrected precision, precision and MCC, for each image in the dataset.

Finally, we computed the overall performance of ganglia identification. Average dataset recall sits at $80.2\% \pm 12.2\%$, with average precision at $64.8\% \pm 21.7\%$ for ganglia objects of high and low certainty labels. (Similarly, we found an average recall of $82.1\% \pm 11.4\%$ and an average precision of $60.9\% \pm 24.8\%$ for ganglia objects with only high certainty labels.) We conducted an error analysis and examined reasons for varying detection and classification performance in Section 4.3.

LDA Binary Classification Metrics: Differentiating True Ganglia from Background Plexus

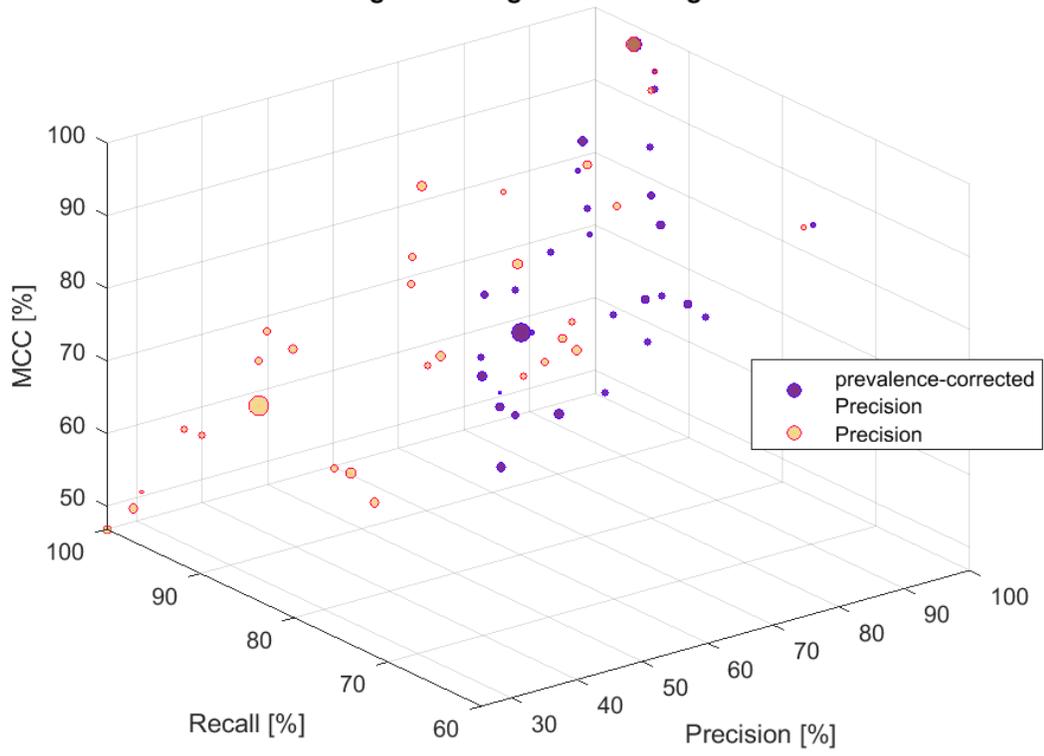


Figure 6.5: Visualizing the spread of ganglia candidate classification results for this dataset. Size of datapoint is reflective of the relative confidence level over n=1000 iterations.

Table 6.3: Ganglia Detection Rate from the Segmentation Stage of Ganglia Candidates

No.	Filename	Ganglia of High Certainty					Ganglia of Low Certainty					Total	
		Total No.	Detected No.	No. of Merges	Recall [%]	Precision [%]	Total No.	Detected No.	No. of Merges	Recall [%]	Precision [%]	Recall [%]	Precision [%]
1	'S00-1910'	4	3	0	75.0	2.7	2	2	0	100.0	1.8	83.3	4.5
2	'S03-3178 D2'	5	4	0	80.0	4.5	2	2	0	100.0	2.3	85.7	6.8
3	'S04-52'	16	14	0	87.5	8.8	9	7	0	77.8	4.4	84.0	13.1
4	'S07-1465'	32	31	0	96.9	14.0	6	6	0	100.0	2.7	97.4	16.7
5	'S14-580'	93	84	2	90.3	37.2	8	4	0	50.0	1.8	87.1	38.9
6	'S97-2054'	4	4	0	100.0	8.5	2	2	0	100.0	4.3	100.0	12.8
7	'S01-18'	10	10	1	100.0	6.3	4	4	0	100.0	2.5	100.0	8.9
8	'S02-410'	23	22	0	95.7	11.1	5	3	0	60.0	1.5	89.3	12.6
9	'S02-484'	32	32	0	100.0	9.6	11	10	1	90.9	3.0	97.7	12.6
10	'S03-2391'	17	13	0	76.5	15.7	2	2	0	100.0	2.4	79.0	18.1
11	'S03-3178 D3'	3	3	0	100.0	5.3	1	1	0	100.0	1.8	100.0	7.0
12	'S03-3178 D4'	1	1	0	100.0	1.0	5	5	0	100.0	5.1	100.0	6.1
13	'S04-1840'	41	32	1	78.1	65.3	16	5	0	31.3	10.2	64.9	75.5
14	'S04-910'	30	30	0	100.0	32.3	7	7	0	100.0	7.5	100.0	39.8
15	'S07-1808'	82	82	1	100.0	24.4	16	15	1	93.8	4.5	99.0	28.9
16	'S08-2215'	51	49	0	96.1	11.3	8	8	0	100.0	1.9	96.6	13.2
17	'S09-2723'	24	22	0	91.7	18.8	8	4	0	50.0	3.4	81.3	22.2
18	'S09-2909'	44	36	2	81.8	13.4	14	7	0	50.0	2.6	74.1	16.0
19	'S11-1760'	26	25	1	96.2	10.3	16	12	3	75.0	5.0	88.2	15.3
20	'S14-1715'	10	5	0	50.0	6.2	4	2	0	50.0	2.5	50.0	8.6
21	'S14-2038'	72	66	5	91.7	25.5	24	10	0	41.7	3.9	79.2	29.3
22	'S14-3414'	53	48	6	90.6	16.5	1	1	0	100.0	0.3	90.7	16.8
23	'S15-1442'	85	74	7	87.1	31.6	10	4	0	40.0	1.7	83.0	33.3
24	'S15-1518'	79	74	3	93.7	16.1	18	18	0	100.0	3.9	94.8	20.0
25	'S16-1197 B1'	51	50	4	98.0	33.3	18	14	0	77.8	9.3	92.8	42.7
26	'S16-1197 B2'	32	25	0	78.1	6.5	5	3	0	60.0	0.8	75.7	7.3
27	'S16-1197 B3'	38	33	0	86.8	17.3	9	7	0	77.8	3.7	85.1	20.9
28	'S16-1415'	22	21	1	95.5	11.8	2	1	0	50.0	0.6	91.7	12.4
29	'S16-1467'	20	18	0	90.0	19.2	2	2	0	100.0	2.1	90.9	21.3
30	'S16-567'	146	126	2	86.3	27.3	11	4	0	36.4	0.9	82.8	28.2
Mean					89.8	17.1				77.1	3.3	87.5	20.3
Std. Dev.					10.9	13.3				24.9	2.4	11.5	14.8
Median					91.7	13.7				84.3	2.6	88.7	16.4

Table 6.4: Classification Metrics reported as percentages for Precision, Recall, Matthew's Correlation Coefficient (MCC), and Area under the Receiver Operator Curve (AUROC). Skew is the class ratio comparing number of negative candidates to positive candidates. Confidence intervals (as indicated by '±') were derived over n = 1000 bootstrap iterations.

No.	Filename	Class Skew (negative) positive	Ganglia Objects with High Certainty and Low Certainty Candidates marked as Positive					No. of Low Certainty Candidates labelled as:	
			Recall [%]	Prevalence- Corrected Precision [%]	Precision [%]	MCC [%]	AUROC [%]	Positive	Negative
1	'S00-1910'	21.2	100 ± 0.0	87.8 ± 0.9	25.4 ± 1.6	46.7 ± 1.8	93.0 ± 0.6	2 ± 0	0 ± 0
2	'S03-3178 D2'	13.8	99.8 ± 2.0	85.0 ± 0.6	29.2 ± 0.9	48.9 ± 1.5	91.1 ± 1.1	2 ± 0	0 ± 0
3	'S04-52'	6.6	89.6 ± 1.8	94.9 ± 0.6	73.1 ± 2.4	77.9 ± 1.9	92.3 ± 0.9	6.5 ± 0.5	0.5 ± 0.5
4	'S07-1465'	5.0	89.8 ± 1.1	88.1 ± 0.3	59.9 ± 0.7	66.9 ± 0.9	88.8 ± 0.6	5 ± 0	1 ± 0
5	'S14-580'	1.6	93.0 ± 0.1	99.0 ± 0.8	98.5 ± 1.3	93.2 ± 1.0	96.1 ± 0.4	2.9 ± 0.3	1.1 ± 0.3
6	'S97-2054'	6.8	98.1 ± 6.8	85.6 ± 5.4	45.4 ± 8.9	60.4 ± 8.5	90.4 ± 4.2	1.4 ± 0.5	0.6 ± 0.5
7	'S01-18'	11.2	98.9 ± 2.7	96.6 ± 0.1	72.0 ± 0.9	82.9 ± 1.7	97.8 ± 1.3	4 ± 0	0 ± 0
8	'S02-410'	6.9	96.0 ± 0.0	93.2 ± 0.5	66.4 ± 1.8	76.6 ± 1.2	94.5 ± 0.3	3 ± 0	0 ± 0
9	'S02-484'	7.1	100.0 ± 0.0	82.6 ± 0.7	39.9 ± 1.2	56.1 ± 1.3	89.5 ± 0.5	9 ± 0	0 ± 0
10	'S03-2391'	4.2	99.7 ± 1.4	82.7 ± 0.8	53.4 ± 1.5	64.8 ± 1.6	89.4 ± 0.8	2 ± 0.1	0 ± 0.1
11	'S03-3178 D3'	13.3	100.0 ± 0.0	85.5 ± 0.2	30.8 ± 0.4	50.6 ± 0.3	91.5 ± 0.1	1 ± 0	0 ± 0
12	'S03-3178 D4'	15.5	100.0 ± 0.0	90.3 ± 0.0	37.3 ± 1.1	57.6 ± 0.9	94.6 ± 0.0	4.9 ± 0.3	0.1 ± 0.3
13	'S04-1840'	0.4	92.7 ± 1.5	72.1 ± 0.4	86.6 ± 0.2	60.7 ± 2.6	78.5 ± 0.9	4 ± 0.1	1 ± 0.1
14	'S04-910'	2.0	93.6 ± 0.0	99.9 ± 0.3	99.9 ± 0.6	95.1 ± 0.5	96.8 ± 0.1	1 ± 0	6 ± 0
15	'S07-1808'	2.6	95.7 ± 0.3	91.2 ± 0.4	79.9 ± 0.7	82.3 ± 0.6	93.3 ± 0.2	9 ± 0	5 ± 0
16	'S08-2215'	7.2	88.7 ± 1.7	91.6 ± 0.6	60.4 ± 2.0	68.8 ± 1.6	90.3 ± 0.8	3 ± 0	5 ± 0
17	'S09-2723'	4.3	89.7 ± 2.4	80.0 ± 1.1	48.1 ± 1.8	55.2 ± 1.9	83.6 ± 1.1	0 ± 0	4 ± 0
18	'S09-2909'	6.1	94.7 ± 0.1	91.7 ± 0.5	64.6 ± 1.3	74.1 ± 1.0	93.1 ± 0.3	4 ± 0	3 ± 0
19	'S11-1760'	7.6	89.3 ± 0.0	86.1 ± 0.7	44.9 ± 1.5	56.8 ± 1.3	87.5 ± 0.4	4 ± 0	5 ± 0
20	'S14-1715'	10.6	95.5 ± 6.6	100.0 ± 0.2	99.8 ± 1.7	97.4 ± 3.8	97.7 ± 3.3	2 ± 0	0 ± 0
21	'S14-2038'	3.0	84.8 ± 0.7	86.3 ± 0.6	67.4 ± 1.2	66.5 ± 1.1	85.7 ± 0.4	3 ± 0.1	7 ± 0.1
22	'S14-3414'	5.6	82.7 ± 1.1	92.1 ± 0.4	67.7 ± 1.3	69.8 ± 0.9	87.8 ± 0.5	1 ± 0	0 ± 0
23	'S15-1442'	2.3	90.3 ± 2.1	71.5 ± 0.4	52.2 ± 0.5	49.9 ± 1.6	77.1 ± 0.8	3 ± 0	1 ± 0
24	'S15-1518'	5.2	76.5 ± 0.9	99.6 ± 0.1	98.1 ± 0.5	84.5 ± 0.6	88.1 ± 0.4	3.5 ± 0.6	14.5 ± 0.6
25	'S16-1197 B1'	1.9	89.3 ± 1.2	93.1 ± 0.6	87.8 ± 0.9	82.4 ± 1.1	91.3 ± 0.6	5.0 ± 0.3	9.0 ± 0.3
26	'S16-1197 B2'	13.8	100.0 ± 0.0	93.2 ± 0.4	49.9 ± 1.7	68.0 ± 1.3	96.4 ± 0.2	1 ± 0	2 ± 0
27	'S16-1197 B3'	4.3	91.7 ± 0.7	96.5 ± 0.6	86.8 ± 2.1	86.7 ± 1.3	94.2 ± 0.4	3.2 ± 0.4	3.8 ± 0.4
28	'S16-1415'	7.5	99.7 ± 1.1	87.5 ± 0.5	48.4 ± 1.1	64.2 ± 1.2	92.8 ± 0.6	1 ± 0	0 ± 0
29	'S16-1467'	3.9	83.8 ± 1.5	91.0 ± 0.7	71.9 ± 1.6	71.3 ± 1.5	87.7 ± 0.8	1 ± 0	1 ± 0
30	'S16-567'	2.7	85.2 ± 0.7	89.0 ± 0.6	75.2 ± 1.1	71.9 ± 1.0	87.3 ± 0.5	1.3 ± 0.5	2.7 ± 0.5
	Mean.	6.8	92.2 ± 1.3	89.5 ± 0.7	64.0 ± 1.5	69.6 ± 1.6	90.6 ± 0.8	3.1 ± 0.1	2.4 ± 0.1
	Std. Dev.	4.8	6.3 ± 1.8	7.2 ± 0.9	21.9 ± 1.5	14.1 ± 1.5	5.0 ± 0.9	2.2 ± 0.2	3.4 ± 0.2
	Median	5.9	93.3 ± 1.0	90.6 ± 0.6	65.5 ± 1.2	68.4 ± 1.3	91.2 ± 0.6	3 ± 0	1 ± 0

Table 6.5: Overall performance of Ganglia identification for a single bootstrap iteration.
 Abbreviations include false negatives (FN) from segmentation (Seg.), classification (Clas.), true positives (TP), and false positives (FP)

No.	Filename	Ganglia of High & Low Certainty							Ganglia of High Certainty						
		Seg. FN	Clas. FN	Total FN	TP	FP	Recall [%]	Precision [%]	Seg. FN	Clas. FN	Total FN	TP	FP	Recall [%]	Precision [%]
1	'S00-1910'	1	0	1	5	13	83.3%	27.8%	1	0	1	3	13	75.0%	18.8%
2	'S03-3178 D2'	1	0	1	6	15	85.7%	28.6%	1	0	1	4	15	80.0%	21.1%
3	'S04-52'	4	2	6	19	6	76.0%	76.0%	2	2	4	13	6	75.8%	67.6%
4	'S07-1465'	1	3	4	34	23	89.5%	59.6%	1	3	4	29	23	87.9%	55.8%
5	'S14-580'	13	6	19	79	0	80.6%	100.00%	9	6	15	76	0	83.5%	100.0%
6	'S97-2054'	0	1	1	4	3	80.0%	57.1%	0	1	1	3	3	71.8%	45.9%
7	'S01-18'	0	0	0	13	5	100.0%	72.2%	0	0	0	9	5	100.0%	64.3%
8	'S02-410'	3	1	4	24	11	85.7%	68.6%	1	1	2	21	11	91.3%	65.6%
9	'S02-484'	1	0	1	41	63	97.6%	39.4%	0	0	0	32	63	100.0%	33.7%
10	'S03-2391'	4	0	4	15	12	78.9%	55.6%	4	0	4	13	12	76.5%	52.0%
11	'S03-3178 D3'	0	0	0	4	9	100.0%	30.8%	0	0	0	3	9	100.0%	25.0%
12	'S03-3178 D4'	0	0	0	6	10	100.0%	37.5%	0	0	0	1	10	100.0%	9.6%
13	'S04-1840'	20	2	22	33	5	60.0%	86.8%	9	2	11	29	5	72.5%	85.3%
14	'S04-910'	0	2	2	29	0	93.5%	100.0%	0	2	2	28	0	93.3%	100%
15	'S07-1808'	1	4	5	89	23	94.7%	79.5%	0	4	4	80	23	95.2%	77.7%
16	'S08-2215'	2	7	9	46	33	83.6%	58.2%	2	7	9	43	33	82.7%	56.6%
17	'S09-2723'	6	2	8	20	20	71.4%	50.0%	2	2	4	20	20	83.3%	50.0%
18	'S09-2909'	15	2	17	36	18	67.9%	66.7%	8	2	10	32	18	76.2%	64.0%
19	'S11-1760'	5	3	8	25	30	75.8%	45.5%	1	3	4	21	30	84.0%	41.2%
20	'S14-1715'	7	0	7	7	0	50.0%	100.0%	5	0	5	5	0	50.0%	100.0%
21	'S14-2038'	20	9	29	55	27	65.5%	67.1%	6	9	15	52	27	77.6%	65.8%
22	'S14-3414'	5	7	12	37	17	75.5%	68.5%	5	7	12	36	17	75.0%	67.9%
23	'S15-1442'	16	5	21	66	58	75.9%	53.2%	10	5	15	63	58	80.8%	52.1%
24	'S15-1518'	5	18	23	56	1	70.9%	98.2%	5	18	23	53	1	69.6%	98.1%
25	'S16-1197 B1'	5	6	11	47	7	81.0%	87.0%	1	6	7	42	7	85.7%	85.7%
26	'S16-1197 B2'	9	0	9	26	28	74.3%	48.1%	7	0	7	25	28	78.1%	47.2%
27	'S16-1197 B3'	7	3	10	33	5	76.7%	86.8%	5	3	8	30	5	78.8%	85.6%
28	'S16-1415'	2	0	2	21	23	91.3%	47.7%	1	0	1	20	23	95.2%	46.5%
29	'S16-1467'	2	4	6	15	6	71.4%	71.4%	2	4	6	14	6	70.0%	70.0%
30	'S16-567'	27	18	45	107	35	70.4%	75.4%	20	18	38	106	35	73.6%	75.1%
Mean							80.2%	64.8%						82.1%	60.9%
Std. Dev.							12.2%	21.7%						11.4%	24.8%
Median							79.5%	66.9%						80.4%	64.1%

6.3 Analysis & Discussion

While observing expert pathologists identify ganglia in Calretinin-stained images, we noticed they marked an object as a ganglion with high likelihood if it appeared dark brown in colour, with a relatively uniform texture, and rounded, defined edges. Since segmenting ganglia using intensity was the easiest method to generate candidates, we passed on the task of separating candidates based on size-, shape-, colour- and texture-based features to the classification stage. As seen from the results, the algorithm to identify ganglia operates with relatively acceptable recall scores (when compared to our expert manual segmentations). However, precision has undoubtedly, room for improvement. Here, we conduct a qualitative error analysis based on the metrics presented in Section 6.2.

6.3.1 Performance of Ganglia Candidate Segmentation

Parameters to segment ganglia candidates were optimized to retain objects manually labeled with high certainty. Hence, many of the images in this dataset scored with high recall as intended. Yet, precision is highly dependent on staining. Within myenteric plexus regions, if we treat the pixel intensity of ganglia as our signal of interest and the intensity of background plexus region as noise, then our signal-to-noise ratio (SNR) varies greatly from image to image. From our experimentation, simple and standard normalization methods cannot sufficiently address this variation. Therefore, images with high segmentation precision tends towards higher ganglia-to-background-plexus SNR, in terms of pixel intensity. However, given the sheer number of false candidates generated at the 5% intensity threshold, precision scores across the dataset are, understandably, poor. The sole exception is image 13, where segmentation precision is as high as 65% (compared to the dataset median at 13.7%) due to high SNR.

On the other hand, image 20 has both poor precision and recall of true ganglia candidates. Here, as shown in Figure 6.6, we can see examples of two errors that the current segmentation algorithm does not account for (errors that are not limited to this one image). The first is identifying ganglia based on local regional variations versus global variations in stain. The darkest 5% intensity threshold is a global threshold and catches most ganglia across the dataset. However, some objects that do not fall within this threshold can still be marked as a ganglion by pathologists, based on local regional intensity levels. For instance, objects 2 and 13 from Figure 6.6 are a light brown compared to detected objects 59 and 67, and this is likely the reason why the first two were not captured as ganglia candidates in the segmentation.

The second type of error relates to the human subjectivity introduced in the manual labels. Objects labelled as ganglia may not be true ganglia. Upon second glance to Figure 6.6, we can see that objects 9 and 66 are fibrous, and their status as ganglia becomes questionable. Therefore, recall and precision may not be capturing the ground truth for image 20. Given the ambiguity that arises with stain, and the subjectivity of human raters, this error is hard to address in an algorithm without determining inter- and intra-observer agreement.

Another set of errors we see are merged objects. In healthy regions of the colon, ganglia often appear in tight clusters. Boundaries between individual cells in clusters are recognized by the naked eye as gradients towards thin white slivers. Although watershed transforms and active contours were applied to address these cases, they were not completely effective, as shown for object 103 in Figure 6.7. Here, for image 28, we can see clean segmentation of true ganglia candidates, with the majority captured for subsequent

classification. However, the trade-off for improving segmentation and reducing merged objects, includes losing smaller ganglia (such as object 7M and 10M).

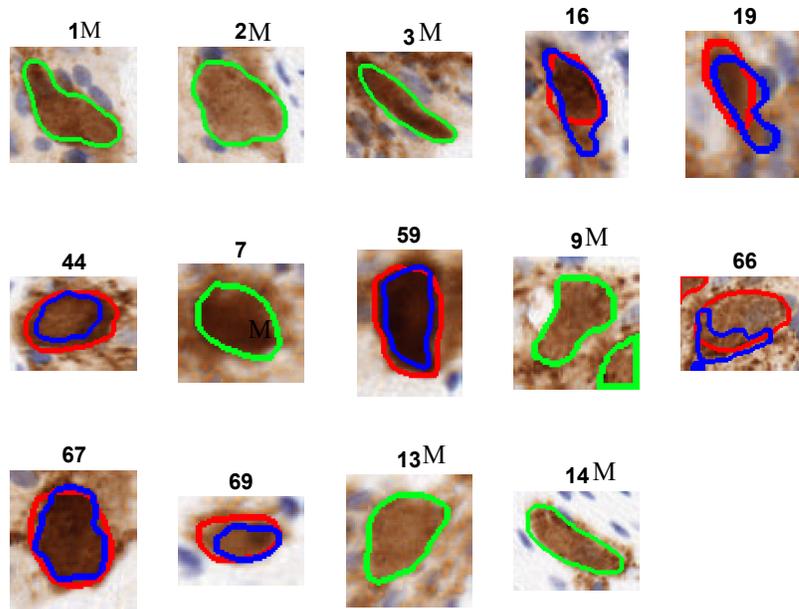


Figure 6.6: Image 20 has poor segmentation recall and precision. Green outlines missed manual (M) segmentations. Red (manual segmentations) and blue (algorithm segmentations) outlines detected ganglia candidates.

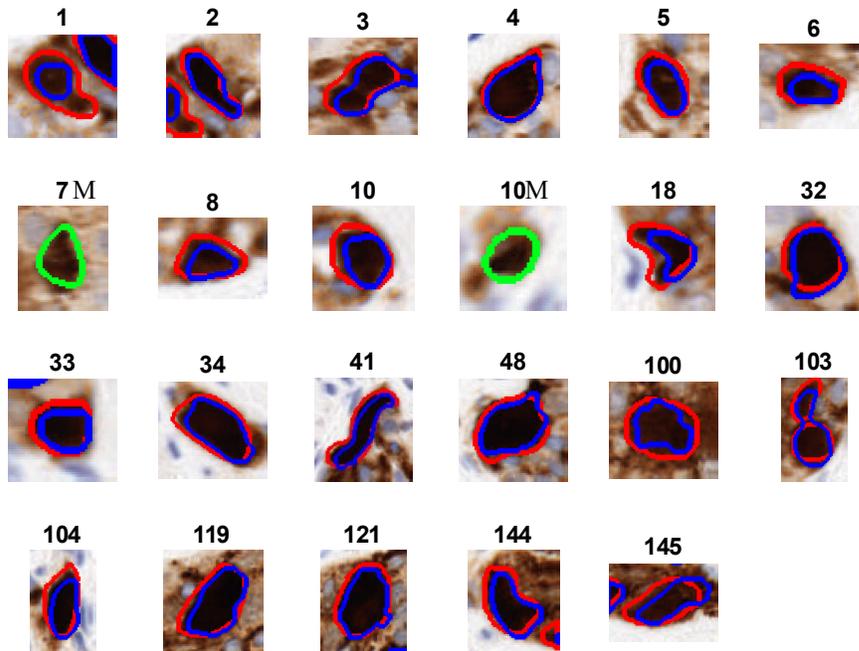


Figure 6.7: Image 28 where a merge error exists. Red outlines are manual segmentations. Blue outlines are the algorithm's segmentation. Missed manual segmentations include object #7M and #10M. Merged object is #103.

6.3.2 *Ganglia Candidate Classification*

From Table 6.4, we can see that the LDA classifier performs with high recall across the dataset, with precision suffering except for a handful of cases. To compare classifier performance between images, we corrected for the varying class imbalance that negatively impacts this metric. Prevalence-corrected precision across the dataset is relatively high, indicating that issues are not due to poor training, but rather inherent with the data provided.

Generating ganglia candidates based primarily on intensity introduced errors for some images that were difficult to resolve, despite segmentation post-processing steps. At the darkest 5% intensity threshold, we have segmented ganglia and some background plexus objects. The pixel intensity for ganglia act as our signal, whereas the pixel intensity for background plexus act as noise. For example, if at this intensity threshold, ganglia have an average intensity value of 30, while background plexus average at 60 (out of 255), the SNR is still high enough for intensity to act as a discriminant. If ganglia and background plexus both have similar average intensities, then even normalization methods cannot discriminate between these two object types. At that point, we are heavily relying on other features for discrimination. However, extracting objects solely based on intensity directly influences other features as well. Segmentation post-processing, although adds some spatial information regarding intensity boundaries, is dependent on how intensity varies. This is not necessarily the same as contextual information, which helps pathologists discriminate ganglia from background plexus in these more ambiguously stained images. Figure 6.8 shows an image where strong staining results in poor SNR, decreasing classification precision.

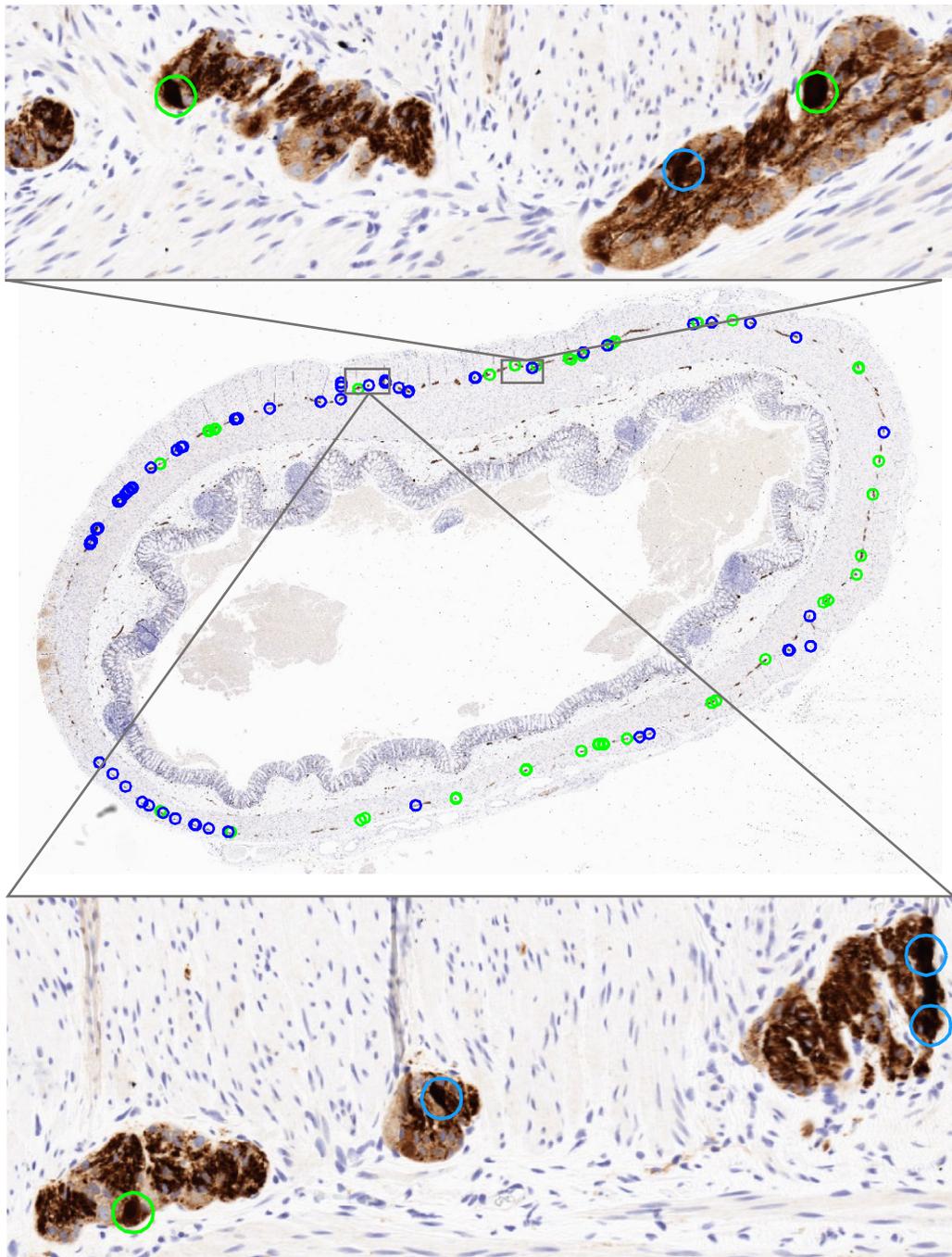


Figure 6.8: Image 9 where the classifier performed with high recall and high prevalence-corrected precision. However, poor precision and MCC may be due to the heavy staining of myenteric plexus regions. Blue circles indicate false positives. Green circles are true positives, as according to manual labels. The bottom window shows two false positives arising from a tissue fold in the upper corner.

In Figure 6.9, we highlighted images with varying classification performances for one bootstrap iteration. Subfigure a) displays results for Image 20 where, despite the large class imbalance, it performs with the highest overall classification scores in this dataset. The LDA could identify nearly all instances of marked ganglia, with little to no errors. This lies in contrast with this image's poor ganglia detection performance in the segmentation stage (as shown by Figure 6.6). Regardless, the performance here is surprisingly impressive since there are very few remaining true ganglia objects, and the classifier did not include false positives as with the general trend for this dataset. We speculate high separability between background plexus and the remaining true ganglia objects exists, i.e., high intensity SNR allowed for distinguishable features between objects of the two classes. Visual inspection of plexus regions seems to confirm this thought process.

Image 24, on the other hand, has one of the poorest instances of classifier recall at approximately 76%. However, precision and MCC remain high. In Figure 6.9 b), we see that although intensity SNR is relatively high, the post-processing from the segmentation stage may have changed some of the features of some true ganglia candidates, such that their characteristics are dissimilar to the other true objects (e.g. too small in size). Another reason for low recall could be subjectivity from manual labels, but this is hard to confirm without determining extent of inter- and intra- observer agreement.

Figure 6.9 c) displays the classification results for image 12, where we see high recall, prevalence-corrected precision, and even AUROC. However, the latter two cannot mask the poor precision and MCC performance. This is a difficult image, since there are only a handful of true positives (all of which are identified) but a great class imbalance favouring negative candidates. Although we expect false positives for this case scenario,

the current performance is unacceptable for clinical use, should manual labels hold true. In the magnified window for this image, we can see that the false positives may potentially be rated as true ganglia by another expert pathologist.

Finally, Figure 6.9 d) shows classification results for image 30. This is an image where class imbalance is relatively low (especially compared to the other four images discussed in this section), and so we expect good results like other images with low skew. However, scores across the board are mediocre with recall around 85%, precision at 75%, and MCC at 72%. There is certainly room for improvement, and the errors found here are a combination of the ones already mentioned (i.e., poor SNR and possible discrepancies in labelling).

What we see from these examples with varying performance, are two main contributors to error. The first is poor ganglia-to-background SNR. But the second may likely be arising from human subjectivity since we can see visualize high SNR. As mentioned in the segmentation analysis, manual labels are not ground truths. However, without intra- and inter-observer agreement, one cannot validate if these are indeed human-based errors. Although not shown here, another minor contributor to classification errors (such as elongated false ganglia candidates) likely arise from proximity to decision boundaries. These are difficult to eliminate without overfitting.

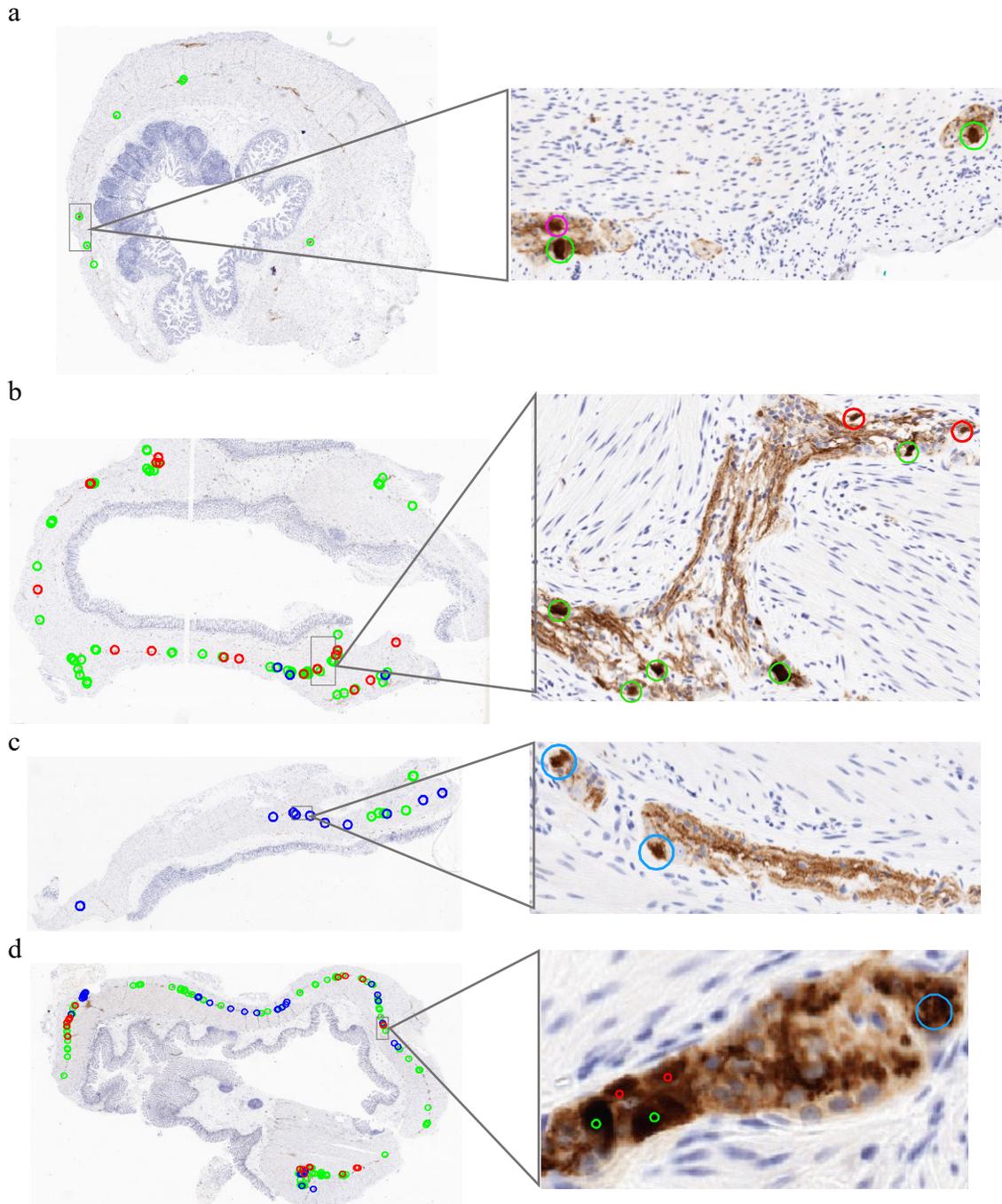


Figure 6.9: Green circles are true positives, blue circles highlight false positives, and red circles are false negatives as according to manual labels. Purple circles highlight missing segmentation errors in magnified windows. Note that these circles simply act as identifiers and do not show the actual segmentation, simply for easier viewing. a) image 20 – high classification performance; b) Image 24 – low classification recall; c) image 12 – poor classification precision and MCC scores; d) image 30 – mediocre performance all around.

6.3.3 Overall Ganglia Identification

Overall ganglia identification is influenced by both the segmentation and the classification steps. In general, we can see that segmentation introduced false negatives, whereas classification mainly introduced false positives. We can also see that although the LDA classifier performs well (as measured by recall and prevalence-corrected precision), overall ganglia identification performance suffers from a class-imbalance burden, errors introduced at the segmentation stage creating indiscernible features, and uncertainty regarding manual labels.

Figure 6.10 shows that despite acceptable recall scores, many of the images have poor precision scores for ganglia identification. This is insufficient for clinical research and diagnostic purposes. We know from the literature search conducted in chapter 2 that spread and count of ganglia is directly reflective of tissue health. Tendency towards greater false positives inflates scores and paints the situation in a better light. Therefore, algorithmic refinements and further expert input are still required.

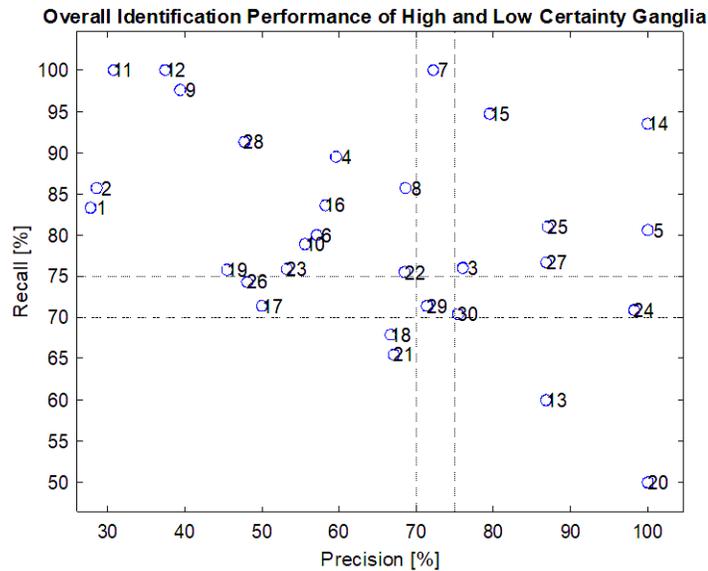


Figure 6.10: Spread of overall performance metrics labelled according to each image in this dataset. Dotted lines at 70% & 75% precision and recall help to identify high performance in upper right quadrant.

Until now, we examined errors in our ganglia identification algorithm and discovered areas for improvement. However, we also want to highlight reasons for excellent performance. Figure 6.11 illustrates overall ganglia identification results for image 14, which had zero segmentation errors and only two false negatives in the classification stage, despite the number of true positives present. There is very high agreement here between the algorithm and manual labels. Upon further inspection, we see consistent staining across the image and almost no ambiguity between ganglia and background plexus. This high SNR resulted in good segmentation, which also enabled excellent classification performance. This is, of course, using only labels from one expert pathologist. However, given the image conditions, we may expect little observer disagreement.

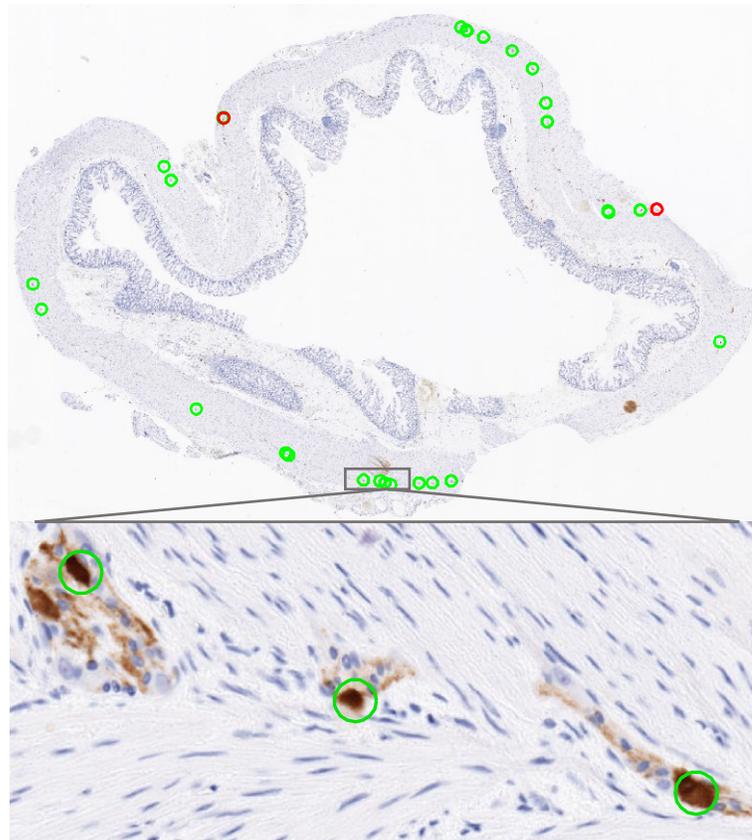


Figure 6.11: Image 14

6.3.4 Inter-observer Agreement

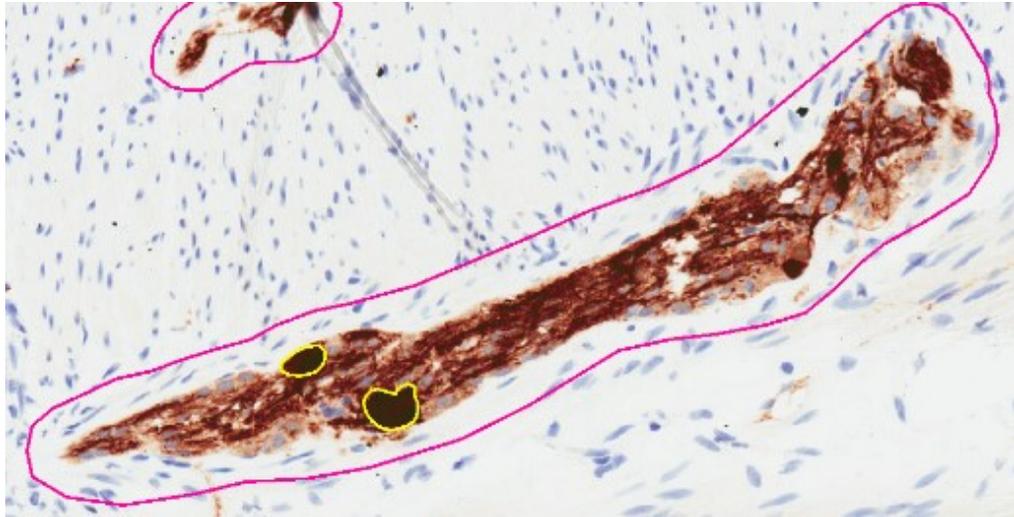
In Section 6.3.2 Ganglia Candidate Classification, we postulated that some of the classification errors may be due to biases from the manual labels, since the task of identifying ganglia is subject to human error. To provide evidence for this hypothesis, a subset of four images were reviewed and marked for ganglia by another expert pediatric and perinatal pathologist at CHEO (henceforth described as “Observer Two”) with over two decades of experience and training. Using this subset of four images, we noted inter-observer variability for ganglia identification between Observer One and Observer Two, as well as the author who acted as the non-expert Observer Three. Figure 6.12 shows an example of the subjective differences that may arise with ganglia identification. Table 6.6 also tallies the total number of ganglia identified across the four images between the three observers. Although this does not reveal inter-observer variability within each plexus region, it does provide some insight into the extent of agreement between experts.

Table 6.6: Total ganglia in subset of four images by three observers

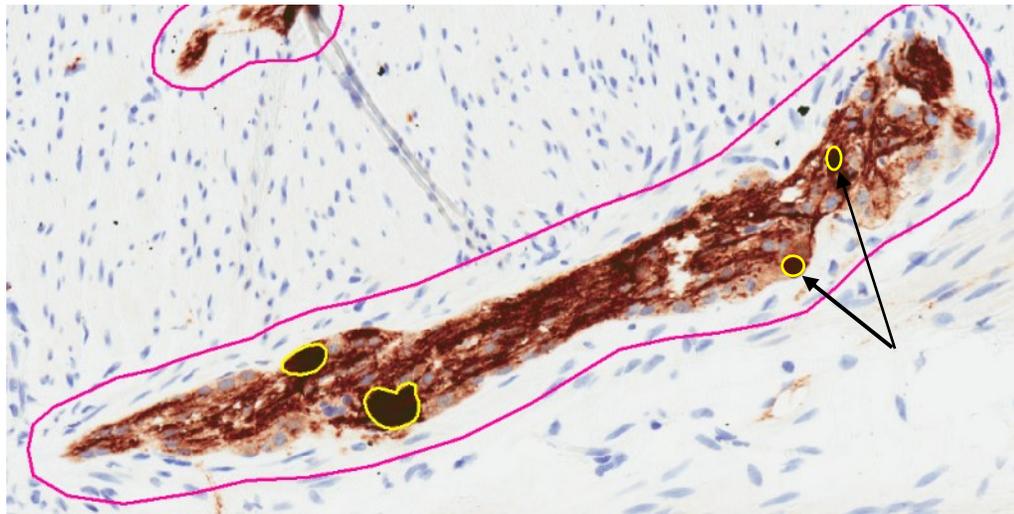
No.	Image	Observer 1: Expert Pathologist		Observer 2: Expert Pathologist		Observer 3: Master’s Student	
		No. of certain ganglia	No. of uncertain ganglia	No. of certain ganglia	No. of uncertain ganglia	No. of certain ganglia	No. of uncertain ganglia
9	'S02-484'	32	11	48	10	50	12
15	'S07-1808'	82	16	114	11	91	16
28	'S16-1415'	22	2	19	4	15	8
30	'S16-567'	146	11	90	25	107	20

We see the extent of agreement fluctuates between images. Perhaps greater variability correlates with the time taken for identifying ganglia. Therefore, difficulty in discerning ganglia from surrounding plexus regions may increase interobserver variability. Interestingly, despite differences across all four images, both expert pathologists noted only

image 9 for its difficulty in ganglia identification, which the results of our automated algorithm also showed (see Figure 6.8). For future work, inter- and intra-observer agreement should be investigated, and variability should be quantified (perhaps via kappa scores) for more reliable ganglia data labels.



a



b

Figure 6.12: Subjective differences in ganglia identification. a) Ganglia identified by Observer One. b) Ganglia identified by expert Observer Two (black arrows highlight difference)

6.4 Conclusion

In this chapter, we described an algorithm for identifying ganglia within myenteric plexus regions in Calretinin-stained histopathology colon images. This is the final step of the overall image processing pipeline described in this thesis. We began by segmenting ganglia candidates using an intensity-based method, and then proceeded to extract features and classify candidates according to expert labels. The algorithm achieved an average precision of $64.8\% \pm 21.7\%$ and an average recall of $80.2\% \pm 12.2\%$ over the dataset of thirty images, directly using manually segmented myenteric plexus regions as the initiating search space. Using the ideal muscularis propria and myenteric plexus segmentation for search spaces allows us to isolate the issues with the ganglia identification algorithm, without introducing other confounding errors. Variances in performance were primarily attributed to differences in staining concentrations between images. This contributed many candidates that were difficult to parse from true ganglia based on shape, texture, and colour features alone. We also recognized that some errors may be easier to identify when inter- and intra-observer agreement scores are included. The current cross-validated metrics reported in this chapter are to be solely treated as an estimate of achievable performance. Potentially incorporating contextual information to remove reliance on intensity-based methods, as well as utilizing anomaly detection methods to avoid severe class imbalances, may improve ganglia detection in future iterations. Future outputs could also incorporate a degree of certainty in ganglia identification, instead of solely marking objects with binary labels.

Chapter 7: Pipeline Performance for Identifying Myenteric Ganglia

In Chapter 1, we described the image processing pipeline used to identify myenteric ganglia. It begins by identifying search spaces for ganglia to minimize computational time and costs on these large images. The muscle layers of the 1) muscularis propria sandwich 2) myenteric plexus regions which contain our objects of primary interest: 3) ganglion cells. The work presented in previous chapters describe algorithms that were initiated with ideal search spaces. We wanted to minimize confounding factors that may skew our evaluation metrics, while being able to isolate and identify the errors associated with each individual step.

In this chapter, we ran the entire pipeline end-to-end to examine complete functionality. Although we do not conduct an in-depth analysis of systematic error, we provide a brief discussion to identify areas of improvement. The results for stage 1 (muscularis propria segmentation) and stage 2 (myenteric plexus region identification) are summarized in Table 7.1. We tabulate the overall precision and recall scores for stage 3 (ganglia identification) after initiating with both algorithm-generated (AG) and manually-segmented (MS) search spaces in Table 7.2.

Table 7.1: Summary of Pipeline Performance for the Identification of Myenteric Ganglia. Stage 1 and Stage 2. Abbreviations made for Precision (Prec.), Manual Segmentation (MS), and Algorithm-Generated (AG). Training metrics are reported separately from test metrics whenever the AG muscularis propria is used as a search space. All values are reported in percentages (%).

No	Image	Stage 1: AG Muscularis Propria Segmentation (Ch 4)				Stage 2: AG Myenteric Plexus Segmentation					
						MS Muscularis (Ch 5)			AG Muscularis		
		DICE	Prec.	Recall	Plexus Inclusion	Prec.	Recall	Ganglia Inclusion	Prec.	Recall	Ganglia Inclusion
1	S00-1910'	80.9	69.5	96.6	97.6	92.0	83.1	100.0	64.5	83.1	100.0
2	S03-3178 D2	86.8	78.1	97.6	100	74.0	89.1	100.0	41.6	89.1	100.0
3	S04-52'	87.6	78.9	98.5	100	71.8	87.1	100.0	42.5	87.1	100.0
4	S07-1465'	78.7	65.0	99.7	100	53.0	87.6	100.0	52.6	83.4	89.5
5	S14-580'	80.3	68.8	96.6	99.5	61.3	94.9	100.0	35.3	94.9	100.0
6	S97-2054'	83.9	73.4	98.0	95.9	77.9	36.3	100.0	51.5	36.3	100.0
Training Mean		83.0	72.3	97.85	98.85				48.0	79.0	98.2
Training Std		3.6	5.5	1.2	1.69				10.4	21.4	4.3
Training Median		82.4	71.5	97.79	99.75				47.0	85.2	100.0
7	'S01-18'	91.9	87.3	96.9	99.1	80.8	63.8	100.0	64.1	62.9	100.0
8	'S02-410'	89.9	82.6	98.7	100.0	92.4	82.5	100.0	58.6	82.5	100.0
9	'S02-484'	90.3	82.4	100.0	99.3	86.0	97.2	100.0	58.5	97.2	100.0
10	S03-2391'	91.3	87.4	95.6	99.1	68.6	80.2	100.0	50.0	79.2	100.0
11	S03-3178 D3'	87.2	77.3	99.9	92.2	75.4	97.9	100.0	42.9	95.7	100.0
12	S03-3178 D4'	92.8	87.0	99.5	98.6	62.0	87.3	100.0	45.5	85.9	100.0
13	'S04-1840'	76.0	91.7	65.0	77.0	76.7	51.1	98.3	54.0	40.0	89.5
14	'S04-910'	84.1	72.6	100.0	100.0	93.8	72.9	100.0	56.7	71.7	100.0
15	'S07-1808'	92.5	86.4	99.5	100.0	78.1	94.7	99.0	47.9	94.7	99.0
16	'S08-2215'	90.6	86.2	95.6	96.1	87.4	89.9	100.0	56.8	87.1	100.0
17	'S09-2723'	83.3	71.9	99.2	100.0	39.7	92.3	100.0	16.7	92.3	100.0
18	'S09-2909'	92.9	87.4	99.2	94.9	74.8	98.7	98.3	41.0	100.0	98.3
19	'S11-1760'	86.4	77.7	97.4	98.0	93.2	80.6	100.0	67.9	79.4	100.0
20	'S14-1715'	87.9	91.3	84.8	85.9	75.6	92.2	100.0	68.0	79.7	85.7
21	'S14-2038'	86.7	81.6	91.7	91.3	90.5	82.6	99.0	44.6	78.5	93.8
22	'S14-3414'	91.0	84.1	99.2	98.2	83.6	91.1	100.0	57.3	91.1	100.0
23	'S15-1442'	91.2	84.4	99.3	97.8	74.3	97.0	100.0	53.1	96.9	100.0
24	'S15-1518'	88.9	86.0	92.1	85.5	85.3	84.6	100.0	42.9	77.8	100.0
25	'S16-1197 B1'	93.3	90.5	96.3	100.0	53.8	96.2	100.0	34.5	94.2	100.0
26	'S16-1197 B2'	96.2	92.9	99.8	100.0	20.3	94.3	100.0	20.2	96.2	100.0
27	'S16-1197 B3'	92.5	89.5	95.8	94.3	71.4	94.1	100.0	51.7	90.6	97.9
28	'S16-1415'	82.3	70.1	99.5	100.0	61.2	95.4	100.0	29.8	97.7	100.0
29	'S16-1467'	89.6	86.7	92.8	96.6	90.2	87.1	95.5	56.5	87.1	95.5
30	'S16-567'	93.2	87.5	99.8	99.2	67.4	94.8	96.8	34.8	94.8	96.8
Test Mean		89.2	81.9	96.2	96.0	74.3	87.4	99.5	48.1	85.6	98.1
Test Std		4.5	7.8	6.8	5.8	17.6	11.6	1.16	13.6	13.5	3.7
Test Median		90.5	86.3	98.9	98.4	76.2	91.6	100.0	50.8	88.8	100.0
Overall Mean						73.7	85.9	99.2			
Overall Std						16.7	14	1.8			
Overall Median						75.5	89.5	100			

Table 7.2: Summary of Pipeline Performance for the Identification of Myenteric Ganglia. Stage 3..
Abbreviations made for Manually Segmented (MS), and Algorithm-Generated (AG). Training metrics are reported separately from test metrics whenever the AG muscularis propria is used as a search space. All values are reported in percentages (%). Estimates for confidence intervals over n = 1000 iterations are indicated by '±'.

No.	Images	Stage 3: Overall Performance of High- and Low-Certainty Ganglia Identification					
		MS Muscularis & MS Plexus (Ch 6)		MS Muscularis & AG Plexus		AG Muscularis & AG Plexus	
		Precision	Recall	Precision	Recall	Precision	Recall
1	S00-1910'	25.4 ± 1.6'	83.3 ± 0.0	18.4 ± 1.0	83.3 ± 0.0	14.8 ± 0.6	83.3 ± 0.0
2	S03-3178 D2	29.1 ± 1.0	85.3 ± 2.5	25.3 ± 1.2	85.7 ± 0.0	19.3 ± 1.2	85.7 ± 0.0
3	S04-52'	73.2 ± 2.3	74.9 ± 1.6	60.0 ± 1.6	82.3 ± 1.9	23.4 ± 0.7	79.6 ± 2.3
4	S07-1465'	59.8 ± 0.7	87.4 ± 1.1	24.7 ± 0.6	92.4 ± 1.1	26.7 ± 1.8	50.9 ± 2.3
5	S14-580'	98.5 ± 1.3	80.8 ± 0.1	96.5 ± 1.7	79.2 ± 0.6	52.4 ± 0.7	78.0 ± 0.8
6	S97-2054'	45.2 ± 8.7	97.9 ± 7.4	34.1 ± 5.7	59.1 ± 5.6	16.5 ± 5.0	52.6 ± 15.1
Training Mean						25.5	71.7
Training Std						13.9	15.7
Training Median						21.3	78.8
7	'S01-18'	72.1 ± 1.0	99.0 ± 2.6	65.4 ± 2.3	93.2 ± 2.8	24.3 ± 0.5	92.4 ± 1.0
8	'S02-410'	66.4 ± 1.8	85.7 ± 0.0	59.3 ± 0.9	82.1 ± 0.0	43.5 ± 1.0	81.9 ± 1.6
9	'S02-484'	40.0 ± 1.2	97.6 ± 0.0	35.6 ± 0.9	97.7 ± 0.0	27.8 ± 1.0	96.9 ± 1.0
10	S03-2391'	53.4 ± 1.4	79.7 ± 1.1	45.3 ± 1.4	64.2 ± 2.5	47.9 ± 1.8	71.3 ± 0.8
11	S03-3178 D3'	30.8 ± 0.3	100 ± 0.0	34.8 ± 1.8	100 ± 0.0	10.8 ± 0.2	100.0 ± 0.0
12	S03-3178 D4'	37.2 ± 1.1	100 ± 0.0	32.2 ± 1.6	83.3 ± 0.0	20.2 ± 1.9	100.0 ± 0.0
13	'S04-1840'	86.6 ± 0.3	59.0 ± 1.0	90.3 ± 0.7	71.5 ± 1.3	54.3 ± 1.4	39.1 ± 1.7
14	'S04-910'	99.9 ± 0.6	93.5 ± 0.0	74.6 ± 1.9	89.9 ± 1.5	42.9 ± 1.5	75.4 ± 3.7
15	'S07-1808'	79.9 ± 0.7	94.7 ± 0.2	70.0 ± 0.8	95.6 ± 0.7	42.4 ± 0.6	86.5 ± 0.0
16	'S08-2215'	60.4 ± 1.9	85.5 ± 1.7	57.5 ± 1.2	88.0 ± 1.5	32.5 ± 0.7	77.8 ± 1.7
17	'S09-2723'	48.1 ± 1.8	70.6 ± 1.9	29.8 ± 0.8	76.7 ± 0.1	9.1 ± 0.3	48.2 ± 0.6
18	'S09-2909'	64.5 ± 1.3	67.9 ± 0.1	62.7 ± 1.6	71.4 ± 0.7	31.2 ± 0.7	68.4 ± 0.8
19	'S11-1760'	44.9 ± 1.5	75.8 ± 0.0	41.0 ± 1.6	75.7 ± 2.2	24.0 ± 0.7	68.3 ± 1.8
20	'S14-1715'	99.8 ± 1.6	48.0 ± 3.2	98.7 ± 4.2	46.4 ± 1.1	98.6 ± 4.9	39.0 ± 1.4
21	'S14-2038'	67.4 ± 1.2	64.7 ± 0.6	64.0 ± 1.3	65.2 ± 0.5	34.3 ± 0.6	54.6 ± 0.7
22	'S14-3414'	67.6 ± 1.3	74.3 ± 1.0	60.5 ± 1.4	74.5 ± 1.2	40.9 ± 0.8	77.6 ± 0.3
23	'S15-1442'	52.2 ± 0.5	73.6 ± 1.7	50.2 ± 0.6	76.4 ± 0.4	34.3 ± 0.8	66.4 ± 1.6
24	'S15-1518'	98.1 ± 0.6	71.6 ± 0.9	88.5 ± 1.3	71.5 ± 1.5	41.5 ± 1.0	54.0 ± 1.9
25	'S16-1197 B1'	87.8 ± 0.9	81.4 ± 1.1	67.7 ± 1.7	74.7 ± 0.9	40.5 ± 1.1	72.6 ± 1.2
26	'S16-1197 B2'	49.8 ± 1.6	74.3 ± 0.0	18.4 ± 2.3	68.6 ± 0.6	16.8 ± 1.7	67.9 ± 1.3
27	'S16-1197 B3'	86.9 ± 1.9	76.8 ± 0.6	87.7 ± 2.1	78.5 ± 1.5	46.7 ± 2.3	47.2 ± 1.3
28	'S16-1415'	48.4 ± 1.1	91.1 ± 1.0	49.1 ± 1.8	90.7 ± 1.5	26.9 ± 0.9	86.9 ± 0.2
29	'S16-1467'	71.8 ± 1.6	75.8 ± 1.3	61.4 ± 2.3	72.7 ± 4.0	19.8 ± 0.7	54.8 ± 2.4
30	'S16-567'	75.2 ± 1.1	70.1 ± 0.6	69.2 ± 1.2	76.0 ± 0.3	35.9 ± 0.4	58.6 ± 1.1
Test Mean		66.2	79.6	58.9	78.5	35.3	70.2
Test Std		20.3	13.7	20.8	12.3	18.0	18.3
Test Median		66.9	76.3	61.0	76.2	34.3	69.8
Overall Mean		64.0	80.7	55.8	78.9		
Overall Std		21.9	12.8	23.2	11.9		
Overall Median		65.5	80.3	59.6	77.6		

In an attempt to maximize myenteric plexus inclusion within the first search space, we accepted that the AG muscularis propria (from Chapter 4) would generally over-segment the region. In terms of metrics, this over-segmentation corresponds to high recall with lower precision across the dataset, as seen in Table 7.1. Slight over-segmentation of muscularis boundaries results in the inclusion of submucosal plexus regions, decreasing the precision of myenteric plexus identification. In fact, we see that recall and ganglia inclusion scores at Stage 2 are comparable, if not identical, when using either the MS or AG muscularis propria. (The exceptions to this trend are images where myenteric plexus regions are missed because they are located at the edge of a segmentation.) However, myenteric plexus precision scores plummeted by an average of 25% over the thirty images. We see from Table 7.2 that combination of AG muscularis and plexus search spaces has a domino effect on the precision of ganglia identification (i.e., a drop of ~30%), whereas overall recall falls by only ~10%. Approximately 10% of the precision drop at Stage 3 can be attributed to the plexus segmentation algorithm, perhaps arising from staining errors. However, the remainder of the fall is likely due to submucosal plexus and ganglia inclusion. Figure 7.1 and Figure 7.2 provide examples of images where we see excessive ganglia identification in the submucosal regions.

When annotating the images, we observed a general trend where submucosal ganglia can have a stronger uptake and expression of brown chromogen than myenteric ganglia. At the third stage in the pipeline, the 5% global intensity threshold applied to generate ganglia candidates for each image, is dependent on the resultant brown chromogen regions within search spaces from the prior stages. For instance, if numerous submucosal plexus regions are included in the 5% global intensity threshold, ganglia candidates will

generally have lower intensity values. Therefore, the algorithm may miss myenteric ganglia with lighter shading, resulting in a higher false negative rate and lower overall recall. Intensity of staining errors can also introduce more segmented candidate objects at the classification stage, which further skews the class imbalance, increasing the potential for more false positive objects.

Without even accounting for the issues regarding ganglia identification as discussed in Chapter 6 (i.e., poor plexus-to-ganglia SNR, and observer bias in manual labels), we can see that improperly identifying myenteric plexus regions causes a cascade of errors, further complicating our goal of identifying myenteric ganglia. The critical step that is causing the current main performance issues is the identification of myenteric plexus regions, namely differentiating these areas from submucosal plexus and staining errors. The largest source of error is an over-segmentation of the muscularis propria. These steps need to be improved in the automation, so that an isolated improvement in stage 3 of the pipeline can be effective.

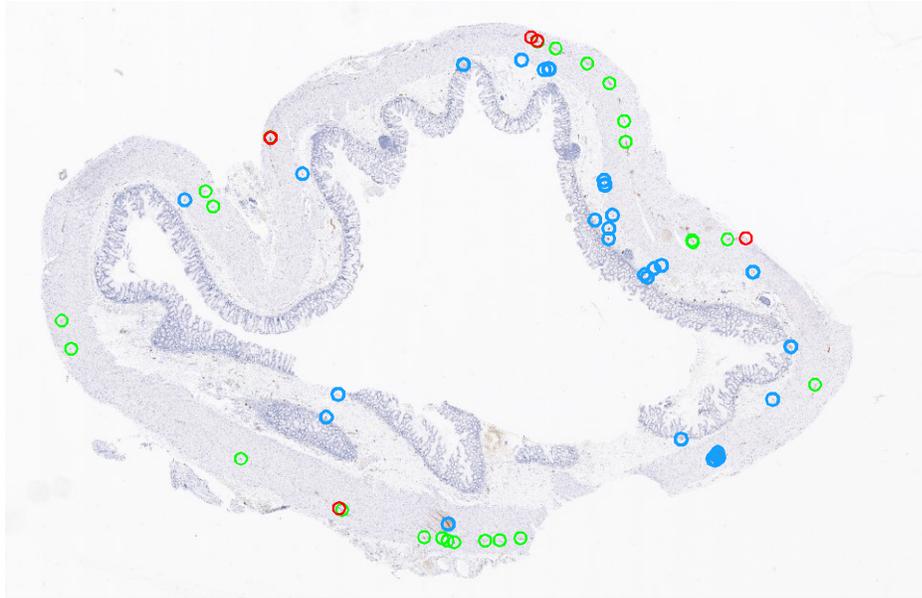


Figure 7.1: Ganglia Identified in Image 14 after full pipeline run. This image had excellent ganglia performance in Chapter 6, with ideal initiating search spaces. Green circles are true positives, red circles are false negatives, and blue circles are false positives.

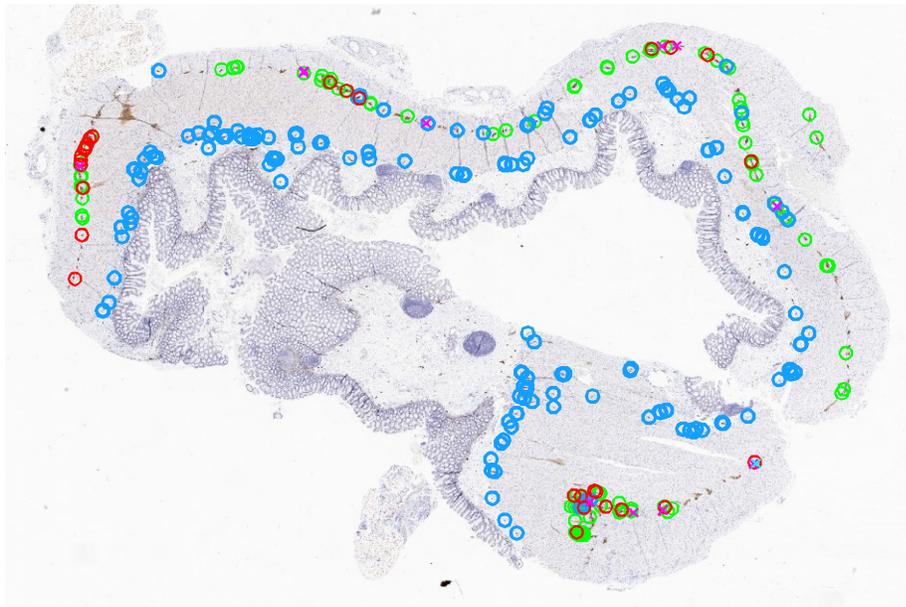


Figure 7.2: In Image 30, we see an excessive number of submucosal ganglia identified (blue circles) along the circumference. Green circles indicate true positives and red circles are false negatives.

Chapter 8: Summary & Conclusions

Our research examines the feasibility of creating an adjunctive tool for pathologists to assess whole slide images of resected colon tissue from patients who have undergone a pull-through surgery in treatment of Hirschsprung’s disease. Although many criteria may exist to evaluate seromuscular samples, pathologists unequivocally examine for ganglia presence and distribution within myenteric plexus regions. Unfortunately, grading criteria for determining the appropriate resection level are enigmatic and objective standards do not exist for pathologists during the seromuscular assessment. This can increase patients’ risk for surgical failure with morbid symptoms, and inadvertently lower their quality of life.

We proposed an adjunctive tool involving an image processing pipeline to examine digital scans of Calretinin-stained colon tissue for ganglia presence. Our method begins with the segmentation of the muscularis, proceeding to the identification of myenteric plexus regions, and finally tallying the containing ganglion cells. This work is the first of its kind to implement automated image processing and machine learning techniques to extract and identify myenteric ganglia from whole slide images, for the study of Hirschsprung’s disease.

8.1 Main Conclusions

In this thesis, we refined McKean *et al.*’s CNN-based muscularis propria segmentations [12] to initiate our image processing pipeline for myenteric ganglia identification. When compared to manual segmentations, modifications to this patch-based

CNN performed with a mean precision of $81.9 \pm 7.8\%$, mean recall of $96.2 \pm 6.8\%$, and mean myenteric inclusion score of $96.0 \pm 5.8\%$, across our dataset.

From there, we utilized colour-thresholding methods to isolate regions expressing brown chromogen in Calretinin-stained WSIs of the colon cross-section. Within manually segmented muscularis propria, this algorithm identified myenteric plexus regions with a mean precision of $73.7 \pm 16.7\%$, mean recall of $85.9 \pm 14.0\%$, and mean ganglia inclusion score of $99.2 \pm 1.8\%$ across thirty images. However, when employing the CNN-based muscularis propria segmentation as a search space for myenteric plexus, mean precision drops to $48.1 \pm 13.6\%$, while high recall and ganglia inclusion are maintained.

Finally, within ideal search spaces, we developed and implemented an intensity-based segmentation followed by an LDA classification to select and sort ganglia candidate objects. This well-established and classical approach to cell-body identification achieved an overall precision of $64.0 \pm 21.9\%$ and recall of $80.7 \pm 12.8\%$ across our dataset. However, when applied with algorithm-generated search spaces, overall ganglia identification performance dropped to a mean precision of $35.3 \pm 18.0\%$, with a mean recall of $70.2 \pm 18.5\%$ due to submucosal plexus inclusion. We certainly see room for improvement in the overall system.

8.2 Study Limitations

This study utilized specimens from 26 patients to create a dataset of 30 images. Although this is a sufficient amount of data for testing stages 1 and 2 of our pipeline, we can only provide an estimate of performance at stage 3. We do not have a separate representative test set to understand the generalizability of our ganglia identification algorithm. We realize a full colon cross-section is required for HSCR clinical assessment

purposes. However, this is not necessarily a limitation for initial algorithm development where our focus remains on tissue structure identification. Therefore, for future training and testing, we could expand our dataset to include images of specimens that may not have a complete cross-section sample. Nevertheless, we do recognize that our data will be limited regardless simply due to the low incidence rate of HSCR.

In Chapter 6, we highlighted a lack of data regarding inter- and intra-observer agreement on ganglia labels. Without this information, we cannot quantify the upper performance bar expected for stage 3 of our pipeline. With data from a single observer, we also build in bias to our algorithm to favour preferences from this observer. Hopefully expanding expert input can provide greater confidence in our manual labels. However, we do recognize that manual annotations are a time-consuming and tedious task. Expert pediatric pathologists who are experienced and trained with these images and samples may be limited in number and availability within our geographical region.

Finally, although the thesis author briefly explored ganglia classification performance of various classifiers, the abilities of ensemble classifiers were not investigated. Therefore, the achievements described in this thesis simply provide a ballpark estimate of performance capabilities.

8.3 Future Work

There are three major areas of focus for immediate future work that would improve our overall pipeline performance. These include precisely identifying the correct search spaces, improving ganglia identification within myenteric plexus regions, and expanding outputs to describe features relevant for clinical assessment.

8.3.1. Improving Search Spaces

Although stage 1 (muscularis propria segmentation) and stage 2 (myenteric plexus identification) of our overall image processing pipeline perform well individually, they hinder myenteric ganglia identification when applied in conjunction with each other. A method to differentiate submucosal plexus from myenteric plexus regions should greatly boost the overall pipeline performance, whether that is achieved through an improved muscularis segmentation or other means (e.g., classifying brown chromogen regions).

8.3.2. Improving Ganglia Identification within Myenteric Plexus Regions

For stage 3 of our pipeline, we suggested including inter- and intra-observer input for more reliable ganglia labels. This will provide informative insight during error analysis of future algorithms. In terms of algorithm development, targeting ganglia objects with anomaly detection methods can avoid class imbalance issues and may be most appropriate since this is truly a one-class problem. Furthermore, deep learning methods may be able to provide contextual information, reducing ambiguity at the classification stage. However, we do recognize ambiguity may be unresolvable for some images. Therefore, future algorithms could potentially output a degree of certainty when labelling an object as a ganglion cell.

8.3.3. Expanding Pipeline Output to Include Clinically Relevant Features

Finally, the purpose of our image processing pipeline is to extract characteristics from the images relevant to the clinical assessment. These include but are not limited to descriptions of ganglia count, density, size, and spread along the muscularis circumference. Therefore, future iterations of our pipeline should incorporate these outputs.

References

- [1] E. Passarge, “Gastrointestinal Tract : Molecular Genetics of Hirschsprung Disease,” *eLS. John Wiley Sons*, no. 3, pp. 1–8, 2011.
- [2] T. A. Lawal, K. Chatoorgoon, M. H. Collins, A. Coe, A. Peña, and M. A. Levitt, “Redo pull-through in Hirschsprung’s disease for obstructive symptoms due to residual aganglionosis and transition zone bowel,” *J. Pediatr. Surg.*, vol. 46, pp. 342–347, 2011.
- [3] A. (Benha U. Nasr, K. J. Sullivan, E. W. Chan, C. A. Wong, and E. I. (Children’s H. of E. O. Benchimol, “Validation of algorithms to determine incidence of Hirschsprung disease in Ontario, Canada: a population-based study using health administrative data,” *Clin. Epidemiol.*, vol. 9, pp. 579–590, 2017.
- [4] E. L. Goldberg, “An Epidemiological Study of Hirschsprung’s Disease,” *Int. J. Epidemiol.*, vol. 13, no. 4, pp. 479–485, 1984.
- [5] R. P. Kapur and A. J. Kennedy, “Transitional zone pull through: surgical pathology considerations,” *Semin. Pediatr. Surg.*, no. 21, pp. 291–301, 2012.
- [6] R. O. Heuckeroth, “Hirschsprung disease - Integrating basic science and clinical medicine to improve outcomes,” *Nat. Rev. Gastroenterol. Hepatol.*, vol. 15, no. 3, pp. 152–167, 2018.
- [7] J. Kessmann, “Hirschsprung’s disease: Diagnosis and management,” *Am. Fam. Physician*, vol. 74, no. 8, pp. 1319–1322, 2006.
- [8] S. Chhabra, R. Harwood, and S. E. Kenny, “Hirschsprung’s disease,” *Surg. (United Kingdom)*, vol. 34, no. 12, pp. 628–632, 2016.
- [9] A. Coe *et al.*, “Reoperation for Hirschsprung Disease: Pathology of the Resected Problematic Distal Pull-Through,” *Pediatr. Dev. Pathol.*, no. 15, pp. 30–38, 2012.
- [10] F. Friedmacher and P. Puri, “Residual aganglionosis after pull-through operation for Hirschsprung’s disease: A systematic review and meta-analysis,” *Pediatr. Surg. Int.*, vol. 27, no. 10, pp. 1053–1057, 2011.
- [11] J. Kurian, M. T. K. Law, V. Madge, A. D. C. Chan, D. El Demellawy, and E. Ukwatta, “Image Processing and Analysis of Histopathological Images Relating to Hirschsprung’s disease,” in *Canadian Medical and Biological Engineering Conference*, 2018, vol. 41.
- [12] C. McKeen, F. Zabihollahy, J. Kurian, A. D. C. Chan, D. E. Demellawy, and E. Ukwatta, “Machine learning-based approach for fully automated segmentation of muscularis propria from histopathology images of intestinal specimens,” in *SPIE Medical Imaging 2019: Digital Pathology Conference Proceedings*, 2019, vol. 10956, no. 0P, p. 23.
- [13] M. D. Gershon and E. M. Ratcliffe, “Developmental biology of the enteric nervous system: Pathogenesis of Hirschsprung’s disease and other congenital

- dysmotilities,” *Semin. Pediatr. Surg.*, vol. 13, no. 4, pp. 224–235, 2004.
- [14] P. G. A. James S. Lowe, “Human Histology,” in *The Yale Journal of Biology and Medicine*, 2015, pp. 221–223.
- [15] A. Arshad, C. Powell, and M. P. Tighe, “Hirschsprung’s disease,” *BMJ*, vol. 345, 2012.
- [16] A. G. Coran and D. H. Teitelbaum, “Recent advances in the management of Hirschsprung’s disease,” *Am. J. Surg.*, vol. 180, no. 5, pp. 382–387, 2000.
- [17] A. M. Holschneider and P. Puri, *Hirschsprung’s Disease and Allied Disorders*, 3rd ed. Berlin: Springer-Verlag Berlin Heidelberg, 2008.
- [18] M. Carabotti, A. Scirocco, M. A. Maselli, and C. Severi, “The gut-brain axis: Interactions between enteric microbiota, central and enteric nervous systems,” *Ann. Gastroenterol.*, vol. 28, no. 2, pp. 203–209, 2015.
- [19] M. Rao and M. D. Gershon, “The bowel and beyond: The enteric nervous system in neurological disorders,” *Nat. Rev. Gastroenterol. Hepatol.*, vol. 13, no. 9, pp. 517–528, 2016.
- [20] E. A. Mayer, “Gut feelings: The emerging biology of gut-"brain communication,” *Nat. Rev. Neurosci.*, vol. 12, pp. 453–466, 2011.
- [21] J. G. Betts *et al.*, *Anatomy and Physiology*. Houston, Texas: OpenStax, 2013.
- [22] C. Ippolito *et al.*, “Quantitative evaluation of myenteric ganglion cells in normal human left colon: implications for histopathological analysis,” *Cell Tissue Res.*, vol. 336, pp. 191–201, 2009.
- [23] M. S. Faussone-Pellegrini, C. Cortesini, and D. Pantalone, “Neuromuscular structures specific to the submucosal border of the human colonic circular muscle layer,” *Can J Physiol Pharmacol*, vol. 68, no. 11, pp. 1437–1446, 1990.
- [24] M. Fu, P. K. H. Tam, M. H. Sham, and V. C. H. Lui, “Embryonic development of the ganglion plexuses and the concentric layer structure of human gut: a topographical study.,” *Anat. Embryol. (Berl.)*, vol. 208, no. 1, pp. 33–41, Apr. 2004.
- [25] T. Uesaka, H. M. Young, V. Pachnis, and H. Enomoto, “Development of the intrinsic and extrinsic innervation of the gut,” *Developmental Biology*, vol. 417, no. 2, pp. 158–167, 2016.
- [26] F. Friedmacher and P. Puri, “Classification and diagnostic criteria of variants of Hirschsprung’s disease,” *Pediatr. Surg. Int.*, vol. 29, no. 9, pp. 855–872, 2013.
- [27] R. P. Kapur, R. C. Reed, L. S. Finn, K. Patterson, J. Johanson, and J. C. Rutledge, “Calretinin immunohistochemistry versus acetylcholinesterase histochemistry in the evaluation of suction rectal biopsies for Hirschsprung Disease.,” *Pediatr. Dev. Pathol.*, vol. 12, no. 1, pp. 6–15, 2009.
- [28] R. P. Kapur and A. J. Kennedy, “Histopathologic Delineation of the Transition

- Zone in Short-Segment Hirschsprung Disease,” *Pediatr. Dev. Pathol.*, vol. 16, pp. 252–266, 2013.
- [29] R. P. Kapur, “Calretinin-Immunoreactive Mucosal Innervation in Very Short-Segment Hirschsprung Disease: A Potentially Misleading Observation,” *Pediatr. Dev. Pathol.*, vol. 17, pp. 28–35, 2014.
- [30] R. P. Kapur, “Submucosal nerve diameter of greater than 40 μm is not a valid diagnostic index of transition zone pull-through,” *J. Pediatr. Surg.*, vol. 51, no. 10, pp. 1585–1591, 2016.
- [31] R. P. Kapur, “Histology of the transition zone in hirschsprung disease,” *Am. J. Surg. Pathol.*, vol. 40, no. 12, pp. 1637–1646, 2016.
- [32] R. P. Kapur *et al.*, “Remodeling of Rectal Innervation After Pullthrough Surgery for Hirschsprung Disease: Relevance to Criteria for the Determination of Retained Transition Zone,” *Pediatr. Dev. Pathol.*, vol. 22, no. 4, pp. 292–303, 2019.
- [33] M. Bettolli *et al.*, “Colonic dysmotility in postsurgical patients with Hirschsprung’s disease. Potential significance of abnormalities in the interstitial cells of Cajal and the enteric nervous system,” *J. Pediatr. Surg.*, vol. 43, pp. 1433–1438, 2008.
- [34] M. S. Faussone-Pellegrini, D. Pantalone, and C. Cortesini, “Smooth muscle cells, interstitial cells of Cajal and myenteric plexus interrelationships in the human colon,” *Acta Anat.*, vol. 139, no. 1, pp. 31–44, 1990.
- [35] T. Taguchi, S. Suita, K. Masumoto, and A. Nagasaki, “An Abnormal Distribution of C-Kit Positive Cells in the Normoganglionic Segment Can Predict a Poor Clinical Outcome in Patients with Hirschsprung’s Disease,” *Eur J Pediatr Surg*, vol. 15, pp. 153–158, 2005.
- [36] A. K. Yadav, K. Mishra, S. Agarwal, and A. Mohta, “Role of Computerized Morphometric Analysis in the Diagnosis of Hirschsprung’s disease,” *Anal. Quant. Cytol. Histol.*, vol. 32, pp. 114–116, 2010.
- [37] H. L. Monforte-Muñoz, I. Gonzalez-Gomez, J. M. Rowland, and B. H. Landing, “Increased submucosal nerve trunk caliber in aganglionosis: A ‘positive’ and objective finding in suction biopsies and segmental resections in Hirschsprung’s disease,” *Arch. Pathol. Lab. Med.*, vol. 122, pp. 721–725, 1998.
- [38] T. Wester, D. S. O’Briain, and P. Puri, “Notable postnatal alterations in the myenteric plexus of normal human bowel,” *Gut*, vol. 44, pp. 666–674, 1999.
- [39] G. J. Gheradi, “Pathology of the ganglionic-aganglionic junction in congenital megacolon,” *Arch. Pathol.*, vol. 69, pp. 520–523, 1960.
- [40] F. V White and J. C. Langer, “Circumferential Distribution of Ganglion Cells in the Transition Zone of Children with Hirschsprung Disease,” *Pediatr. Dev. Pathol.*, vol. 3, pp. 216–222, 2000.
- [41] M. Swaminathan and R. P. Kapur, “Counting myenteric ganglion cells in

- histologic sections: An empirical approach,” *Hum. Pathol.*, vol. 41, no. 8, pp. 1097–1108, 2010.
- [42] F. Boman, R. Sfeir, R. Priso, M. Bonnevalle, and R. Besson, “Advantages of intraoperative semiquantitative evaluation of myenteric nervous plexuses in patients with Hirschsprung disease,” *J. Pediatr. Surg.*, vol. 42, pp. 1089–1094, 2007.
- [43] H. Miura, R. Ohi, S. W. Tseng, and T. Takahashi, “The structure of the transitional and aganglionic zones of Auerbach’s plexus in patients with Hirschsprung’s disease: A computer-assisted three-dimensional reconstruction study,” *J. Pediatr. Surg.*, vol. 31, no. 3, pp. 420–426, 1996.
- [44] “What is Digital Pathology?,” 29-Jun-2017. [Online]. Available: <https://www.leicabiosystems.com/pathologyleaders/what-is-digital-pathology/>. [Accessed: 04-Jul-2018].
- [45] K. Sirinukunwattana, D. R. J. Snead, and N. M. Rajpoot, “A Stochastic Polygons Model for Glandular Structures in Colon Histology Images,” *IEEE Trans. Med. Imaging*, vol. 34, no. 11, pp. 2366–2378, 2015.
- [46] A. Madabhushi and G. Lee, “Image analysis and machine learning in digital pathology: Challenges and opportunities,” *Med. Image Anal.*, vol. 33, pp. 170–175, 2016.
- [47] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, “Histopathological Image Analysis: A Review,” *IEEE Rev. Biomed. Eng.*, vol. 2, pp. 147–171, 2009.
- [48] A. D. Belsare and M. M. Mushrif, “Histopathological Image Analysis using Image Processing Techniques: An Overview,” *Int. J. Signal Image Process.*, vol. 3, no. 4, pp. 23–36, 2012.
- [49] A. Janowczyk and A. Madabhushi, “Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases,” *J. Pathol. Inform.*, vol. 7, no. 1, p. 29, 2016.
- [50] A. Janowczyk, S. Doyle, H. Gilmore, and A. Madabhushi, “A resolution adaptive deep hierarchical (RADHicaL) learning scheme applied to nuclear segmentation of digital pathology images,” *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 6, no. 3, pp. 270–276, 2018.
- [51] X. Pan *et al.*, “Accurate segmentation of nuclei in pathological images via sparse reconstruction and deep convolutional networks,” *Neurocomputing*, vol. 229, pp. 88–99, 2017.
- [52] S. Najjar *et al.*, “Image Processing and Analysis of Mucosal Calretinin Staining to Define the Transition Zone in Hirschsprung Disease: A Pilot Study,” *Eur. J. Pediatr. Surg.*, vol. 29, no. 2, pp. 179–187, 2018.
- [53] F. Schilling *et al.*, “Digital pathology imaging and computer-aided diagnostics as a novel tool for standardization of evaluation of aganglionic megacolon

- (Hirschsprung disease) histopathology,” *Cell Tissue Res.*, vol. 375, pp. 371–381, 2019.
- [54] S. Paxton, M. Peckham, K. Adele, S. Paxton, K. Adele, and M. Peckham, “The Leeds Histology Guide,” 2003. [Online]. Available: https://histology.leeds.ac.uk/what-is-histology/H_and_E.php. [Accessed: 05-Jul-2018].
- [55] S. W. Moore and G. Johnson, “Acetylcholinesterase in Hirschsprung’s disease,” *Pediatr. Surg. Int.*, vol. 21, no. 4, pp. 255–263, Apr. 2005.
- [56] L. Szyllberg and A. Marszalek, “Diagnosis of Hirschsprung’s disease with particular emphasis on histopathology. A systematic review of current literature,” *Prz. Gastroenterol.*, vol. 9, no. 5, pp. 264–269, 2014.
- [57] V. Zuikova, I. Franckevica, I. Strumfa, and I. Melderis, “Immunohistochemical Diagnosis of Hirschsprung’s Disease and Allied Disorders,” *Acta Chir Latv.*, vol. 15, no. 1, pp. 50–57, 2015.
- [58] R. K. Agrawal, N. Kakkar, R. K. Vasishta, V. Kumari, R. Samujh, and K. L. N. Rao, “Acetylcholinesterase histochemistry (AChE) - A helpful technique in the diagnosis and in aiding the operative procedures of Hirschsprung disease,” *Diagn. Pathol.*, vol. 10, no. 28, 2015.
- [59] V. Guinard-Samuel *et al.*, “Calretinin immunohistochemistry: a simple and efficient tool to diagnose Hirschsprung disease,” *Mod. Pathol.*, vol. 22, no. 10, pp. 1379–1384, Oct. 2009.
- [60] D. H. Gonzalo and T. Plesec, “Hirschsprung disease and use of calretinin in inadequate rectal suction biopsies,” *Arch. Pathol. Lab. Med.*, vol. 137, no. 8, pp. 1099–1102, 2013.
- [61] Z. A. Musa, B. J. Qasim, H. F. Ghazi, and A. W. A. K. Al Shaikhly, “Diagnostic roles of calretinin in hirschsprung disease: A comparison to neuron-specific enolase,” *Saudi J. Gastroenterol.*, vol. 23, no. 1, pp. 60–66, 2017.
- [62] N. Rakhshani *et al.*, “Hirschsprung Disease Diagnosis: Calretinin Marker Role in Determining the Presence or Absence of Ganglion Cells,” *Iran. J. Pathol. Iran J Pathol*, vol. 11, no. 114, pp. 409–415, 2016.
- [63] V. Guinard-Samuel *et al.*, “Calretinin immunohistochemistry: A simple and efficient tool to diagnose Hirschsprung disease,” *Mod. Pathol.*, vol. 22, no. 10, pp. 1379–1384, 2009.
- [64] T. K. H. Lim, Kok Hing Wan, Wei Keat Lim, A. H. L. Loh, S. A. Nah, and K. T. E. Change, “Primary Diagnosis of Hirschsprung Disease - Calretinin Immunohistochemistry in Rectal Suction Biopsies, with emphasis on diagnostic pitfalls,” *World J. Pathol.*, vol. 3, no. 3, pp. 14–22, 2014.
- [65] “Overview of Immunohistochemistry (IHC),” *ThermoFisher Scientific*. [Online]. Available: [128](https://www.thermofisher.com/ca/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-</p>
</div>
<div data-bbox=)

protein-methods/overview-immunohistochemistry.html.

- [66] L. Yadav, U. Kini, K. Das, S. Mohanty, and D. Puttegowda, “Calretinin immunohistochemistry versus improvised rapid Acetylcholinesterase histochemistry in the evaluation of colorectal biopsies for Hirschsprung disease.,” *Indian J. Pathol. Microbiol.*, vol. 57, no. 3, pp. 369–75, 2014.
- [67] D. Anetsberger, S. Kürten, S. Jabari, and A. Brehmer, “Morphological and immunohistochemical characterization of human intrinsic gastric neurons,” *Cells Tissues Organs*, vol. 206, pp. 183–195, 2019.
- [68] N. J. Spencer and H. Hu, “Enteric nervous system: sensory transduction, neural circuits and gastrointestinal motility,” *Nat. Rev. Gastroenterol. Hepatol.*, vol. 17, pp. 338–351, 2020.
- [69] A. Brehmer, F. Schrödl, and W. Neuhuber, “Morphological classifications of enteric neurons - 100 years after Dogiel,” *Anat. Embryol. (Berl.)*, vol. 200, pp. 125–135, 1999.
- [70] T. W. Gould, W. A. Swope, D. J. Heredia, R. D. Corrigan, and T. K. Smith, “Activity within specific enteric neurochemical subtypes is correlated with distinct patterns of gastrointestinal motility in the murine colon,” *Am. J. Physiol. - Gastrointest. Liver Physiol.*, vol. 317, no. 2, pp. 210–221, 2019.
- [71] M. Costa, S. J. H. Brookes, P. A. Steele, I. Gibbins, E. Burcher, and C. J. Kandiah, “Neurochemical classification of myenteric neurons in the guinea-pig ileum,” *Neuroscience*, vol. 75, no. 3, pp. 949–967, 1996.
- [72] D. Ganns, F. Schrödl, W. Neuhuber, and A. Brehmer, “Investigation of general and cytoskeletal markers to estimate numbers and proportions of neurons in the human intestine,” *Histol. Histopathol.*, vol. 21, no. 1, pp. 41–51, 2006.
- [73] N. Beuscher, S. Jabari, J. Strehl, W. Neuhuber, and A. Brehmer, “What neurons hide behind calretinin immunoreactivity in the human gut?,” *Histochem. Cell Biol.*, vol. 141, no. 4, pp. 393–405, 2014.
- [74] I. Barshack, E. Fridman, I. Goldberg, Y. Chowders, and J. Kopolovic, “The loss of calretinin expression indicates aganglionosis in Hirschsprung’s disease,” *J. Clin. Pathol.*, vol. 57, no. 7, pp. 712–716, 2004.
- [75] A. M. Kwok, A. B. Still, and K. Hart, “Acquired segmental colonic hypoganglionosis in an adult Caucasian male: A case report,” *World J. Gastrointest. Surg.*, vol. 11, no. 2, pp. 101–111, 2019.
- [76] M. T. K. Law, A. D. C. Chan, and D. El Demellawy, “Color image processing in Hirschsprung’s disease diagnosis,” in *2016 IEEE EMBS International Student Conference (ISC)*, 2016, pp. 1–4.
- [77] A. Gragera and V. Suppakitpaisarn, “Relaxed triangle inequality ratio of the Sørensen–Dice and Tversky indexes,” *Theor. Comput. Sci.*, vol. 718, pp. 37–45, 2018.

- [78] J. C. Maxwell, "On the Theory of Compound Colours, and the Relations of the Colours of the Spectrum," *Sci. Pap.*, vol. 1, pp. 410–444, 1860.
- [79] R. Feynman, "Color Vision," in *The Feynman Lectures on Physics*, C. University, Ed. Addison Wesley Longman, 1964.
- [80] Y. Jraissati and I. Douven, "Delving Deeper Into Color Space," *Iperception.*, vol. 9, no. 4, pp. 1–27, 2018.
- [81] J. Serra, *Image Analysis and Mathematical Morphology*. Orlando: Academic Press Inc., 1983.
- [82] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 266–277, 2001.
- [83] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," in *Graphics Gems*, 1994, pp. 474–485.
- [84] F. Meyer, "Topographic distance and watershed lines," *Signal Processing*, vol. 38, pp. 113–125, 1994.
- [85] R. J. Peace and J. R. Green, "Computational Sequence and NGS-based MicroRNA Prediction," in *Signal Processing and Machine Learning for Biomedical Big Data*, Illustrate., E. Sejdic and T. H. Falk, Eds. CRC Press, 2018.
- [86] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinform. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, 2005.
- [87] M. Hall and L. a Smith, "Feature Selection for Machine Learning : Comparing a Correlation-based Filter Approach to the Wrapper CFS : Correlation-based Feature," *Int. FLAIRS Conf.*, 1999.
- [88] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
- [89] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proceedings of the International Joint Conference on Neural Networks*, 2008, pp. 1322–1328.
- [90] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure To ROC, Informedness, Markedness & Correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.
- [91] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, pp. 861–874, 2006.
- [92] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data - Recommendations for the use of performance metrics," in *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 245–251.