UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Novel Vision-AI Techniques for Morphological Discovery in System Biology

Permalink

https://escholarship.org/uc/item/0xh5w6wb

Author

Nanda, Amitash

Publication Date

2023

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Novel Vision-AI Techniques for Morphological Discovery in System Biology

A Thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Electrical Engineering (Intelligent Systems, Robotics, and Control)

by

Amitash Nanda

Committee in charge:

Professor Debashis Sahoo, Chair
Professor Bill Lin, Co-Chair
Professor Xiaolong Wang
Professor Pengtao Xie

2023

Copyright

Amitash Nanda, 2023

All rights reserved.

The Thesis of Amitash Nanda is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

I dedicate this creative endeavor to Dr. Debashis Sahoo and Dr. Dharnidhar Dang, whose unwavering support and guidance helped me to complete this task. I am grateful for their encouragement and for clearing all my technical and non-technical doubts throughout the research work. Also, to all Boolean Lab members, friends, and family who supported, motivated, and inspired me to complete this task.

EPIGRAPH

“When you put together open medicine, open science, open access, open source, and open data—Open5—all sorts of new channels of research activity become available, and existing ones become exponentially more powerful.”

—Eric Topol from “The Patient Will See You Now: The Future of Medicine is in Your Hands” [1]

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Debashis Sahoo for his support as the chair of my committee. Through multiple drafts and many doubt-clearing sessions, his guidance has proved to be invaluable. In particular, to my advisor Debashis Sahoo and co-advisor, Bill Lin, for their support, guidance, and valuable feedback throughout my master's thesis work.

Additionally, I would like to express my sincere gratitude to Xiaolong Wang and Pengtao Xie for the service of my committee members.

Finally, I would like to acknowledge Dr. Dharanidhar Dang, Atishna Samantaray, Aman Deep Kaur, Christopher Duroiu, Ajaya Mohan, Aadya Daita, and other members of Boolean Lab for the support that helped me in an immeasurable way.

Chapter 2, in full, is currently being prepared for submission for publication of the material. Behroozikhah, Mahdi*; Khandelwal, Soni*; Nanda, Amitash*; Dang, Dharanidhar; Samantaray, Atishna; Prabhudesai, Arya; Sahoo, Debashis. The thesis author was the primary investigator and author of this material.

Chapter 3, in full, has been submitted for publication of the material as it may appear in PNAS, 2023, Dang, Dharanidhar*; Nanda, Amitash*; Carpinelli, Sophia; Tindle, Courtney; Sinha, Saptarshi; Ghosh, Pradipta; Sahoo, Debashis. The thesis author was the primary investigator and author of this material.

VITA

2015-2019	B.Tech. in Electronics and Instrumentation, CET Bhubaneswar, India
2019-2021	Software Engineer, Accenture Labs, Bengaluru, India
2021-2023	M.S in Electrical Engineering (Intelligent Systems, Robotics, and Control), University of California San Diego
2021-2023	Graduate Student Researcher, University of California San Diego
2022	Machine Learning Research Intern, Teradata US, San Diego
2023	Software Engineering Intern, Teradata US, San Diego

PUBLICATIONS

Vedula, R., **Nanda, A**., Gochhayat, S. S., Hota, A., Agarwal, R. R., Rai, S. K., Mahapatra, S., Swain, K. K., & Das, S. K. (2018d). *Computer Vision Assisted Autonomous Intra-Row Weeder*. https://doi.org/10.1109/icit.2018.00027.

Nanda, A., Ahire, D. (2022). An Autonomous Intelligent System to Leverage the Post-harvest Agricultural Process Using Localization and Mapping. In *Smart Innovation, Systems and Technologies* (pp. 507–516). https://doi.org/10.1007/978-981-19-0011-2_45.

Nanda, A., Swain, K. K., Reddy, K. S., & Agarwal, R. R. (2020b). *sTransporter: An Autonomous Robotics System for Collecting Fresh Fruit Crates for the betterment of the Post Harvest Handling Process*. https://doi.org/10.1109/icaccs48705.2020.9074439.

Dang, D., **Nanda, A**., Lin, B., Sahoo, D. (2022, August 6). *NeuCASL: From Logic Design to System Simulation of Neuromorphic Engines*. arXiv.org. https://arxiv.org/abs/2208.03500.

Vedula, R., **Nanda, A**., Swain, K. K., Das, S. K., & Mohanty, M. N. (2020). Plant Sustainability Monitoring Using Unmanned Aerial Vehicle. In *Springer eBooks* (pp. 1175–1183). https://doi.org/10.1007/978-981-15-1420-3_128.

Dang, D., **Nanda, A**., Lin, B., Sahoo, D. (2023). PhoTen: A Novel Silicon Photonics based Real Time Learning Accelerator, ICCD, Washington DC, USA. (Unpublished).

ABSTRACT OF THE THESIS

Novel Vision-AI Techniques for Morphological Discovery in System Biology

by

Amitash Nanda

Master of Science in Electrical Engineering (Intelligent Systems, Robotics, and Control)

University of California San Diego, 2023

Professor Debashis Sahoo, Chair
Professor Bill Lin, Co-Chair

Morphological study in system biology provides a broader perspective of understanding biological systems' structure, form, and organization. Nowadays, incorporating state-of-the-art novel vision-AI techniques revolutionizes this study and could accelerate the feature extraction process and lead to groundbreaking discoveries. The design of novel computer vision-based Deep Learning algorithms enables the development of predictive models, which helps in studying disease progression, developing personalized medicines, drug testing, organ replacement, etc. This thesis presents novel procedures and techniques to extract features from confocal and histopathological images to study organoid culture and colorectal cancer. I have successfully created a unique dataset of Crohn's disease patient-derived organoids (PDOs) and normal colon tissue samples from mice and humans. Organoids need rigorous rapid imaging for continuous monitoring over a long period. Therefore, it is challenging for scientists to process and verify the data manually. Our developed first-of-its-kind novel organoid mining engine process provides a real-time investigation of organoids. The developed model accurately locates, quantifies, tracks, and classifies human colon organoids without expert intervention. Histopathology image analysis is the key to diagnosing colon cancer by focusing on cell morphology and tissue structures. A pathologist takes images from the interest section of the tissue and prepares them for further analysis. The traditional method involves hand-crafted feature extraction followed by classical image processing techniques. I have introduced an original U-shaped crypt segmentation model using novel vision-AI on colon tissue, revealing a new gene expression pattern on the glandular epithelium cells.

CHAPTER 1. Introduction

1.1 Thesis Outline

This thesis is structured into three chapters. Each chapter addresses the unique objective outlined in this thesis research study (below). *Chapter 1* begins with an introduction to system biology and overviews its scope and significance. Further, it explores the role of computational methods in understanding biological systems. The chapter highlights the importance of morphological analysis as a significant component of systems biology, providing the groundwork for incorporating vision-AI-based technologies. Further, it introduces various types of images typically studied in system biology and focuses mainly on confocal and histopathological images. *Chapter 2* centers on the application of vision-AI in the morphological analysis of colon crypts and its study for colorectal cancer, emphasizing the significance of the CDX2 marker. The chapter highlights techniques such as gland and cell/nuclei segmentation and outlines the method for handling gene-expression data using Boolean implication relationships. It introduces a novel approach for the staining process and a vision-AI colon-crypt segmentation model. The chapter concludes with the revelation that the above-said techniques aid in novel discoveries from the morphological data verified by experts. *Chapter 3* uses vision AI to explore the morphological analysis of organoids associated with Chron's disease. It provides a comprehensive introduction to organoids derived from Crohn's disease patients and underscores its importance in biological findings and current gaps in organoid research. Also, it talks about the innovative computer-vision-based deep learning model for organoid counting and classification. The chapter ends with a discussion about the proposed approach's contribution to discoveries in morphological studies validated by experts in the respective field.

1.2 Systems Biology

Systems biology is a holistic approach in biomedical research deciphering the complex biological systems at the cell, tissue, or organism level. Moreover, this research focuses on all the components of an organism and the interaction among them, considering all as one system rather than just individual parts[2]. It's a paradigm shift from the reductionist biology of the past, where the emphasis was primarily on deconstructing complex biological systems into their simple component. Instead of focusing entirely on individual genes or proteins, it studies the complex network of interactions that helps in disease research. Scientists use these interconnections to develop effective therapeutic strategies leading pathways for precision medicine, finding new disease biomarkers, gene profiling, drug targets, and several other treatments. Systems biology has produced significant health science advancement and led to numerous discoveries in biomedical research[3]. It is a collaborative field involving disciplines like biology, computer science, mathematics, engineering, bioinformatics, etc.

1.3 Computation in Systems Biology

Predicting the outcome of an observable phenomenon is the basis of natural science. However, such predictions in biology are challenging because of the complex systems of living organisms. A single cell comprises many molecules that undergo numerous biochemical reactions influenced by enzymes, drugs, and variations in nutrition. Therefore, it is impossible for scientists to track all the bio-chemical processes considering the complexity of the biological systems, and it requires computational approaches along with experimental research[4]. Hence to calculate the effects of cellular functions, it has become necessary to develop computational models.

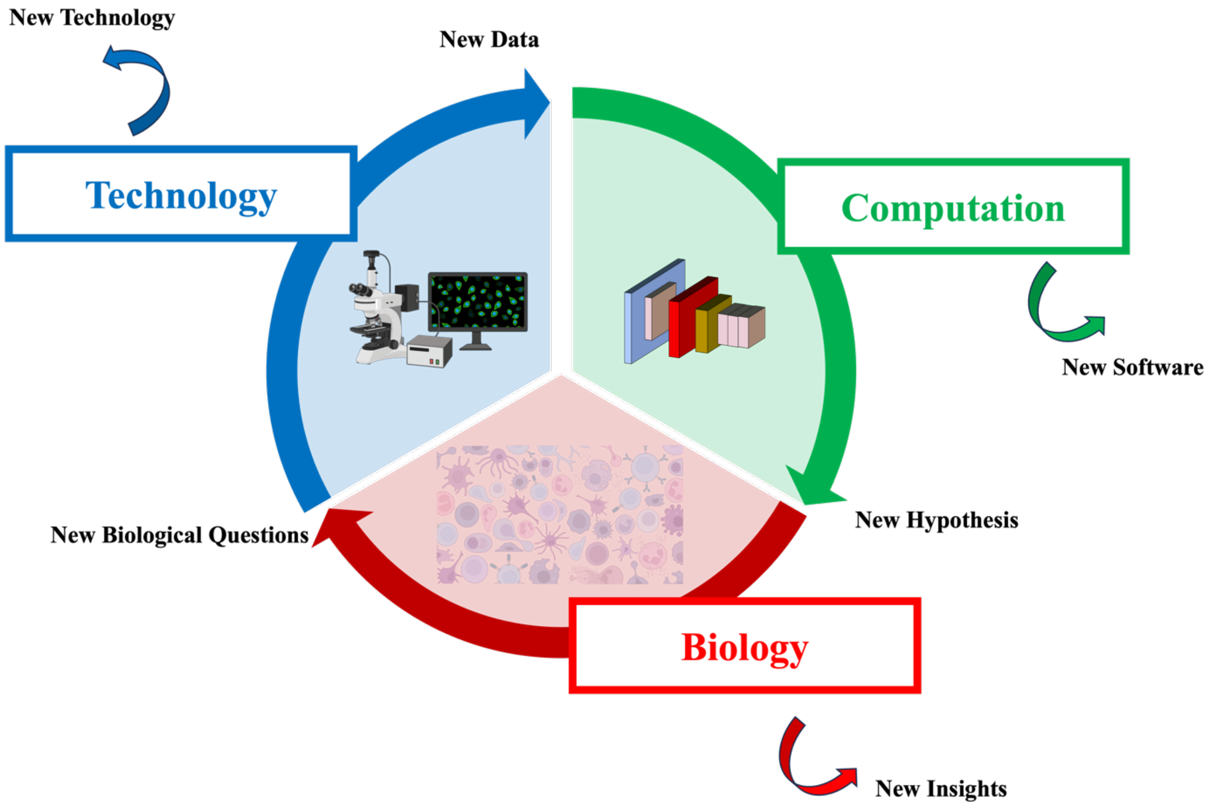


Figure 1.1: Computational Biology Overview.

These models help visualize trends, perform precise component interaction calculations, and predict system behavior. Simulating entire cellular systems provides a deeper understanding of the system, calculates accurate medication dosage for patients, and identifies potential vulnerabilities in harmful pathogens for drug development. Process algebra is a computational method biologist have recently explored for modeling and analyzing biological systems. They are powerful tools that provide an unambiguous formal specification of interactions and synchronizations between concurrent processes. PAs often serve as the intermediate model and are translated into other computational models like differential equations or Markov models[5]. Boolean networks are a commonly used computational method that oversimplifies the complexity of biological systems by disregarding the intermediate states. It is widely used for analyzing the robustness and stability of genetic regulatory networks[6]. Fig 1.1 displays the use of biology, technology, and computation for novel discovery. Fig 1.2 displays the need for vision-AI framework in morphological study.

1.4 Morphological Study in Systems Biology

In systems biology, morphological analysis is the study that involves the investigation of the structure, form, and spatial organization of biological systems. This study usually ranges from the molecular to organism level and helps to understand the function and behavior through the system's organization and arrangement. Some standard techniques are used to understand morphology, such as microscopy for observing cellular structure or detailed structural analysis through computer imaging. Morphological studies are essential in systems biology to understand the system's overall behavior through individual components. Studying the morphology of cells helps us understand their function in tissue. Moreover, specific studies help in disease processes like changes in cellular morphology, which can indicate disease progression or the drug's effect on cells. Morphological analysis in systems biology is a complex process and involves numerous steps, as shown in Fig 1.3. The first step is to input data which could be biological samples like tissue, cells, organs, etc. The next step is to capture detailed images of the abovesaid biological samples through microscopy, tomography, radiology, etc. A pre-processing technique is used further to enhance the image quality and remove noise. The next step is identifying and quantifying morphological features of interest in the images, such as size, shape, texture, etc. These extracted features are used to identify patterns, classify samples, and make predictions based on extracted features using statistical analysis, machine learning, or vision AI. Finally, the results obtained during the analysis are used for interpretation in a biological context.

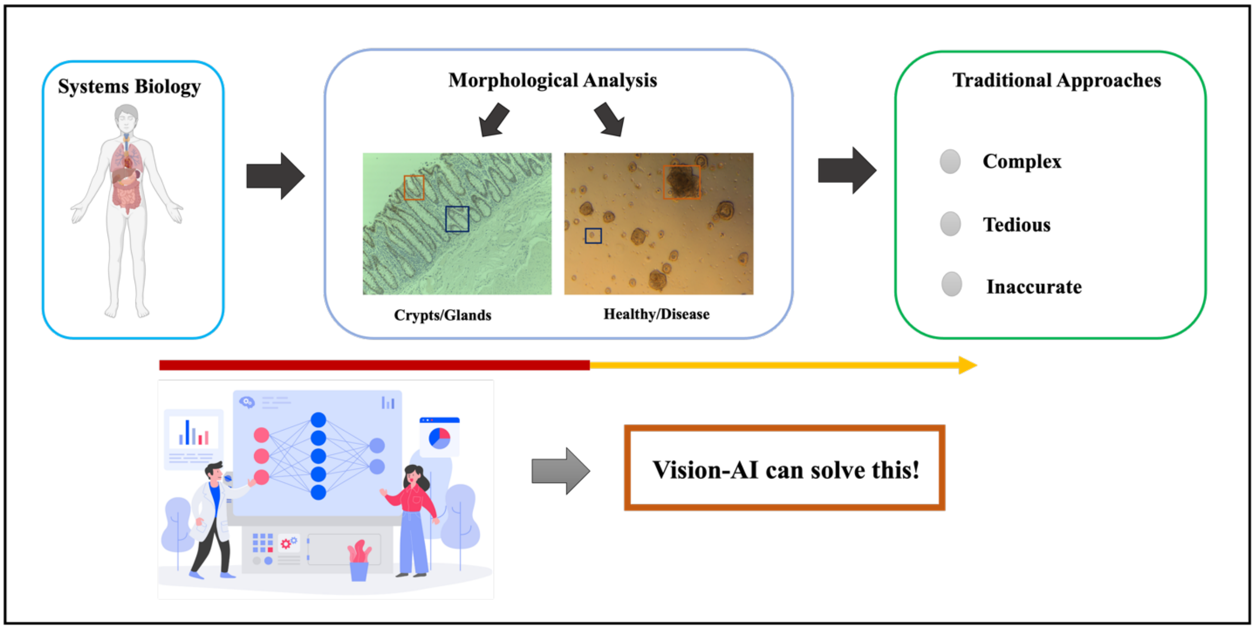


Figure 1.2: Morphological Analysis in Systems Biology.

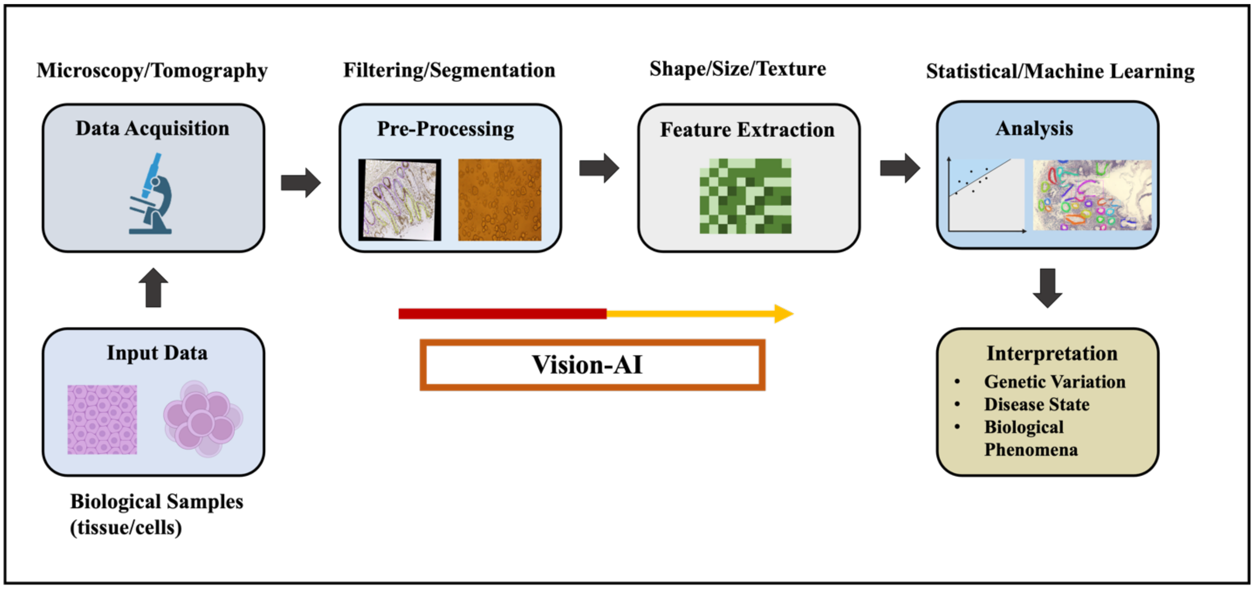


Figure 1.3: Importance of Vision-AI in the morphological analysis of the biological samples.

1.5 Different types of Images used in Biology for Computational Analysis

Microscopy: These are images obtained using different microscopy techniques. Examples include bright-field microscopy, fluorescence microscopy, confocal microscopy, electron microscopy, and light sheet microscopy. Microscopy images provide a large amount of information about cellular and subcellular structures, protein localization, molecular interactions, and tissue organization. Microscopic imaging is pivotal in advancing our knowledge of biological systems and human health. Pathologists routinely employ microscopy to examine tissue samples and identify abnormal cellular structures that are crucial indicators of various diseases, including cancer, infectious diseases, and genetic disorders. This imaging technique is also used to examine the morphology, composition, and crystal structure of biological samples[7].

X-Ray Crystallography Images: X-Ray Crystallography Imaging determines the three-dimensional structural design of a specified set of molecules. It is primarily used for both proteins and nucleic acids. The images obtained through X-Ray crystallography represent the electron density distribution within the crystal and provide valuable insights into the atomic arrangement of molecules. X-Ray crystallography has been used extensively to study the three-dimensional structures of nucleic acids, such as DNA and RNA. These studies have revealed valuable insights into the structural features and interactions within nucleic acid molecules, including DNA double helices, RNA folding motifs, and RNA-protein complexes. Understanding this structure and interaction is essential for deciphering processes such as DNA replication, transcription, and translation. X-Ray crystallography has also contributed to creating comprehensive structural databases, providing information for understanding protein evolution, function, and interaction networks[8].

Cryo-Electron Microscopy: Cryo-Electron Microscopy, also known as Cryo-EM, is a technique used to determine the three-dimensional structure of macromolecules like proteins and complexes at near-atomic resolution. Cryo-EM images are obtained by freezing samples in a thin layer of vitreous ice and imaging this frozen sample using an electron microscope. Cryo-EM has emerged as a powerful tool for structure-based drug discovery and design. High-resolution cryo-EM structures of macromolecular targets can provide valuable information for developing small molecule inhibitors or therapeutics that precisely target specific regions or binding sites. Techniques such as single-particle cryo-EM, cryo-electron tomography, and focused ion beam milling have expanded the range of biological samples and complexes that can be studied. Moreover, developments in detector technology and image processing algorithms have enabled higher-resolution structural determination and improved data interpretation. Cryo-electron microscopy has revolutionized the field of structural biology by providing high-resolution insights into the structures and dynamics of biological macromolecules and complexes[9].

Medical Imaging: A wide range of medical imaging techniques are available to clinicians, including computed tomography (CT). Magnetic imaging (MRI), positron emission tomography (PET), ultrasound, etc. Each of these techniques produces images of organs, tissues, and physiological processes within the human body. These images are used in computational biology to study anatomical structures, disease progression, and treatment responses. Techniques such as PET, SPECT, and optical imaging can be used to track the distribution, metabolism, and interaction of targeted molecular probes or tracers. Molecular imaging plays a crucial role in studying molecular pathways, drug development, and personalized medicine. Functional magnetic resonance imaging (fMRI) measures the changes in blood flow and oxygenation levels in the brain.

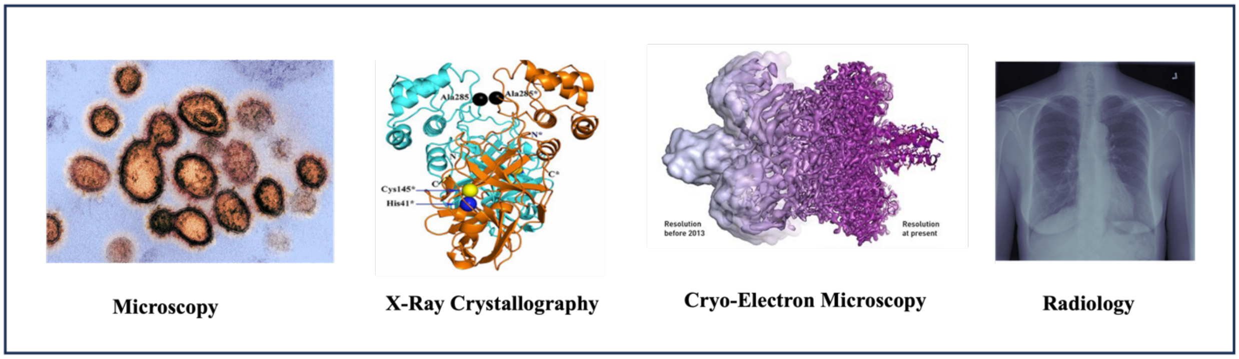


Figure 1.4: Different types of images used in computational biology[11-14].

fMRI enables researchers to map brain activity to study cognitive processes, neural networks, and brain disorders. Positron emission tomography (PET) and single-photon emission computed tomography (SPECT) are functional imaging techniques that assess organ function, metabolic activity, and receptor binding. Techniques such as MRI, CT, ultrasound, and positron emission tomography (PET) enable the visualization of pathological changes, tumor growth, organ dysfunction, and other disease-related abnormalities[10]. Fig 1.4 shows different types of images used in computational biology.

1.6 Previous Work: Vision-AI in Medical Image Data

Vision-based Deep Learning methods have been very effective for various medical diagnostic tasks, surpassing medical professional's performance. This work implements multiclass classification on pulmonary diseases based on the NIH chest X-rays sample dataset. Further, chest radiographs, a 2D high-resolution greyscale medical image is used to detect Pneumonia. We performed a comparative analysis of detecting Pneumonia using different image classification models (custom CNN, VGG-16, ResNet-50). Further used model interpretability methods like SHAP [15] value analysis to justify the classification outcomes.

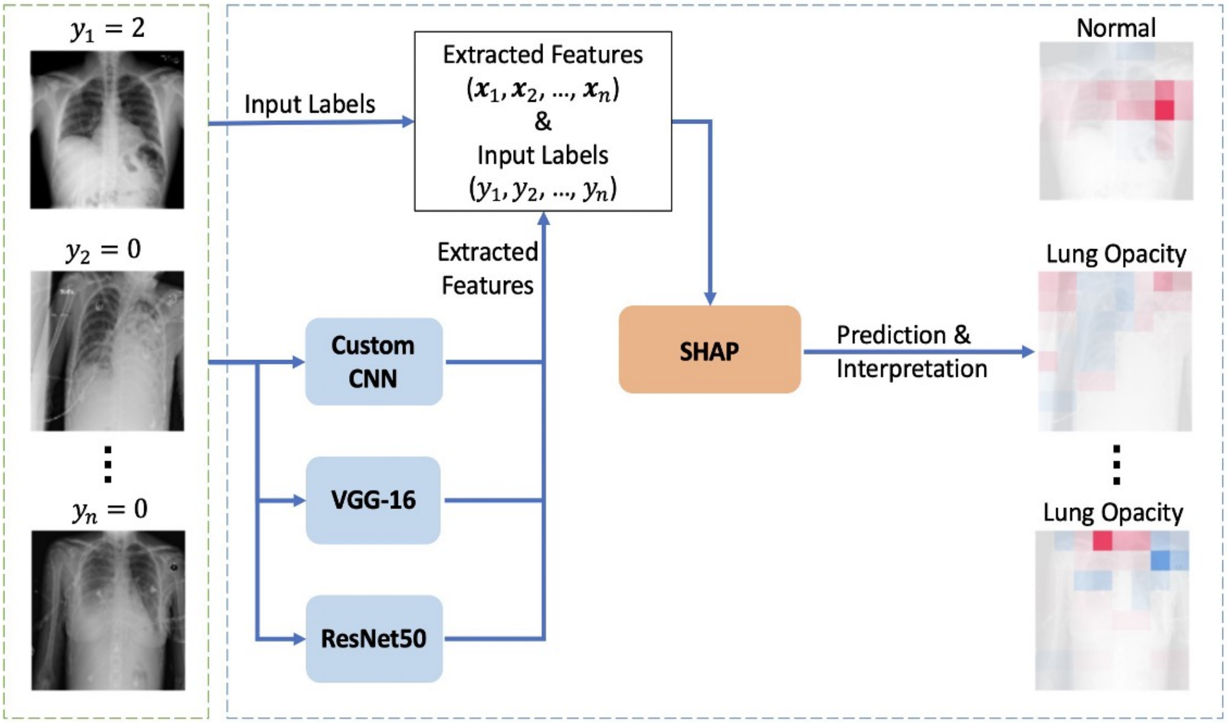


Figure 1.5: Overview of the research[16].

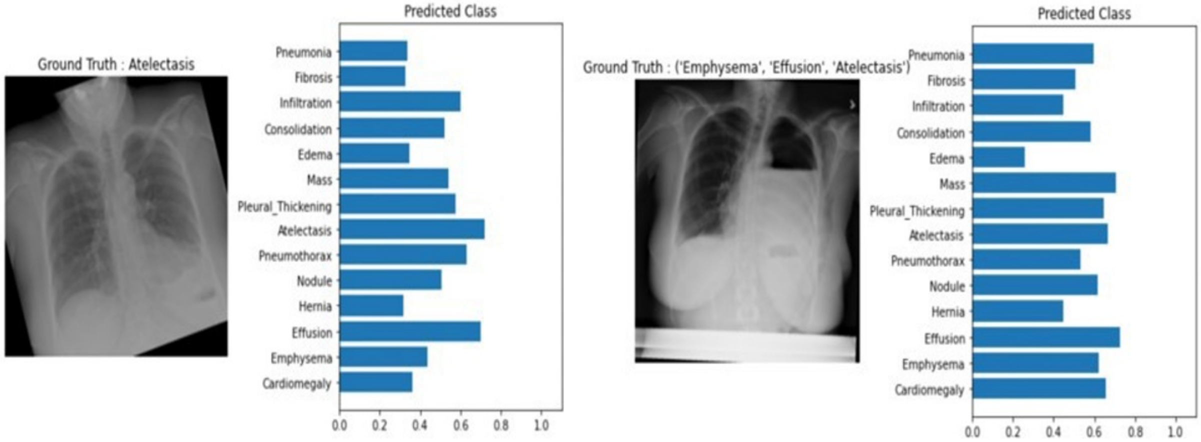


Figure 1.6: Model prediction on sample images[16].

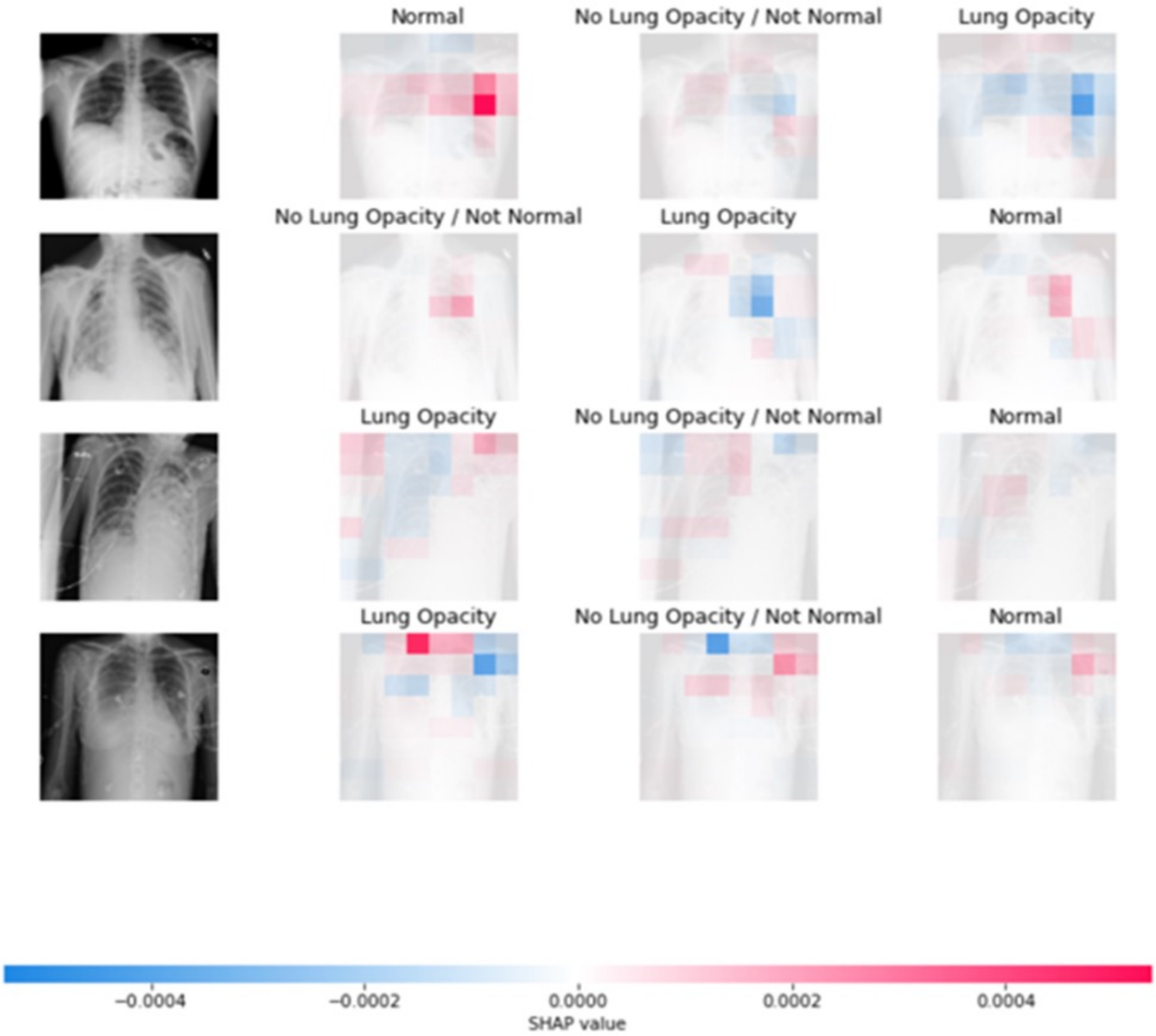


Figure 1.7: SHAP values for predicted model[16].

1.7 Target Image Data: Confocal and Histopathological Images

Confocal Microscopic Images[17] : Morphological investigations of confocal microscopic images involve analyzing and interpreting the structural characteristics, spatial organization, and morphological features of samples. These investigations give insight into the cellular and tissue morphology and the spatial relationships between different structures within the sample. Confocal microscopy allows for high-resolution imaging of cells and subcellular structures. Morphological investigations examine the shape, size, and arrangement of cells and organoids within the specimen. This analysis can reveal details about cellular morphology, such as the presence of specific organelles, cell shape changes, or cellular interactions. Confocal microscopy also provides a means to study the organization and architecture of tissues.

Morphological investigations assess the spatial arrangement of cells, tissues, and extracellular matrix components within the sample. This analysis can provide insights into tissue integrity and alterations associated with diseases or experimental conditions. Confocal microscopy enables the visualization of different structures labeled with various fluorescent probes or dyes.

Morphological investigations analyze the co-localization patterns and spatial relationships between the different fluorescent labels. This analysis can reveal interactions, proximity, or co-localization of specific molecules, proteins, or cellular components within the sample. Morphological investigations can also involve quantitative analysis of confocal microscopic images. This may include measuring various morphological parameters, such as cell or organoid shape descriptors, size, spatial distribution, or density of specific structures. Quantitative morphometry can provide objective and numerical measurements, allowing for comparisons between different samples or experimental conditions. Ultimately, morphological analysis of confocal images plays a crucial role in understanding cellular and tissue biology, characterizing disease processes, and evaluating experimental interventions. These investigations provide valuable insights into the morphological changes, spatial organization, and structural features within biological samples, contributing to our understanding of normal and pathological conditions.

Histopathological Images[18] : Histopathological imaging involves examining and analyzing tissue samples to study the microscopic features and changes associated with diseases.

It is an essential technique used in pathology and medical research to diagnose and understand diseases while guiding treatment decisions. Histopathological images are obtained through samples taken from biopsies or surgical resections. These samples are either embedded in paraffin wax or frozen, sliced into thin sections using a microtome, mounted on glass slides, and subjected to specific staining to enhance tissue visibility and highlight specific cellular components and structures. The most common staining technique is hematoxylin and eosin (H&E) staining. H&E staining provides information about tissue architecture, cellular morphology, and overall tissue composition. Another popular staining technique, Immunohistochemistry (IHC) staining, utilizes specific antibodies to detect and visualize the presence or absence of specific proteins or biomarkers in the tissue. Other staining techniques, such as special stains and fluorescent stains, are employed to assess specific tissue components or detect specific cellular abnormalities.

In recent years, digital imaging technologies have been increasingly used in histopathology. Whole-slide imaging (WSI) systems capture high-resolution digital images of the entire tissue section, enabling digital viewing and analysis. These images allow for the analysis of the morphology, organization, and distribution of cells and tissues in the sample to identify any pathological changes or disease-specific features. Histopathological image analysis can automate tasks such as cell counting, morphological measurements, tissue segmentation, and broader pattern recognition. These analyses can aid in identifying specific features, quantifying biomarkers, and correlating histological findings with clinical outcomes. Histopathological imaging is crucial in diagnosing diseases, understanding disease progression, and guiding treatment decisions in various medical specialties. It provides valuable insights into the cellular and tissue changes associated with diseases, facilitating the development of targeted therapies and personalized medicine approaches.

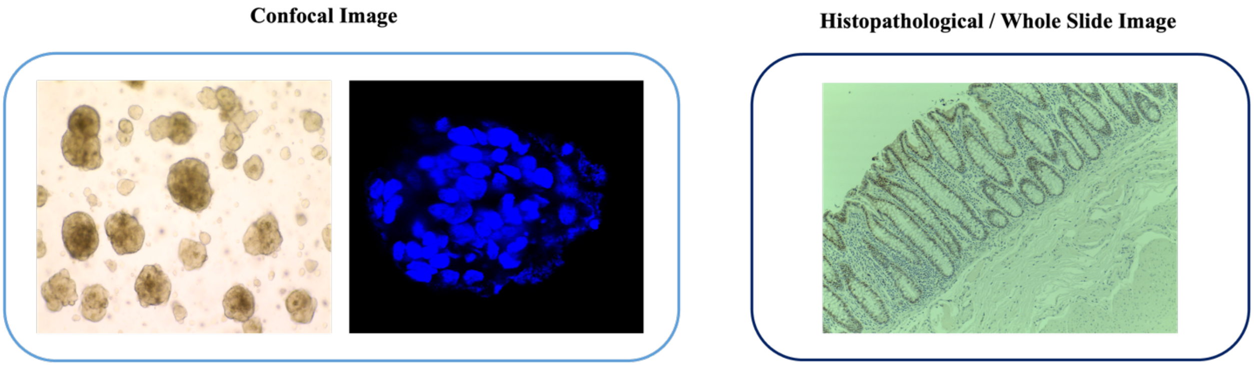


Figure 1.8: Crohn’s disease organoids and whole slide normal colon tissue image.

Fig 1.8 shows the confocal and histopathological images used in this research. Crohn’s disease confocal images are collected from the patient derived organoids, while whole slide images are collected from the normal colon tissue.

CHAPTER 2. Morphological Analysis of Colon Crypt using Vision-AI

2.1 Abstract

Colorectal cancer (CRC), or colon cancer, is the second most common cancer diagnosed in men and women annually in the United States. CDX2 is a crucial biomarker for colorectal cancer, and a thorough understanding of its expression pattern within the colon crypts is a significant factor in refining the diagnostic procedures and therapeutics for the disease. In addition, leveraging histopathological image analysis as a key tool for colon cancer diagnosis by focusing on cell morphology and tissue structures enhances the ability to detect this common malignancy. The traditional analysis involved hand-crafted feature extraction followed by applying classical computer vision methods. However, recent advancement in vision-based deep-learning techniques has improved digital pathology. In this research, I proposed that CDX2 is not expressed uniformly in every cell of the colon epithelium and is low in stem cells, which is associated with high-risk colorectal cancer. Additionally, it is believed that CDX2 is low at the bottom of the crypt and can be used as a biomarker for differentiation in colorectal cancer. A new staining process is adopted, which demonstrates the differential expression of CDX2 in colon crypts. I performed several experiments throughout the investigation to validate the proposed hypothesis. I introduced an original work for gland instance segmentation using novel mask-RCNN and state-of-the-art yolo-based architectures, which reveals a new gene expression pattern on the glandular epithelium cells.

2.2 Introduction

Colorectal cancer (CRC), also known as colon or rectal cancer, represents a significant health threat, as it stands as the second leading cause of cancer-related mortality in the United States. According to the worldwide study in 2020[19], CRC caused 930 thousand deaths out of more than 1.9 million cases. The study by researchers from the International Agency for Research on Cancer (IARC) indicated that by 2040 the numbers might increase to 3.2 million cases per year with an increase of 73% deaths which accounts for about 1.6 million deaths per year[20]. According to the American Cancer Society, in 2023, approximately 153,020 individuals will be diagnosed with CRC, and 52,550 will die[21]. Moreover, there has been a noticeable rise in the incidence of cases in the younger generation. The incidences of colorectal cancer vary region-wide worldwide, with higher rates being observed in developed countries like North America, Europe, and Australia. However, the deaths associated with colorectal cancer are also increasing in developing countries like India, China, etc[22]. Early detection and diagnosis of colorectal cancer can improve the chances of survival and save millions of lives. Pathologists can reach any suspiciously identified area through colonoscopy and collect samples for further examination. To facilitate easy identification of target area methods like inking are adopted, and to enhance visualization, different staining techniques like Hematoxylin and Eosin(H&E) staining, Immunohistochemistry (IHC), Periodic Acid-Schiff (PAS) staining, etc. are used. Though manual pathological practices have been followed for ages, digital pathology can be more beneficial; for instance, it can eliminate human-induced noises[23].

Digital Pathology (DP) is the process of converting a physical histopathology slide into a high-resolution digital image or Whole Slide Image (WSI), which can range in size from 200MB to 10GB[24]. This conversion provides efficient compression, storage, sharing, and viewing the

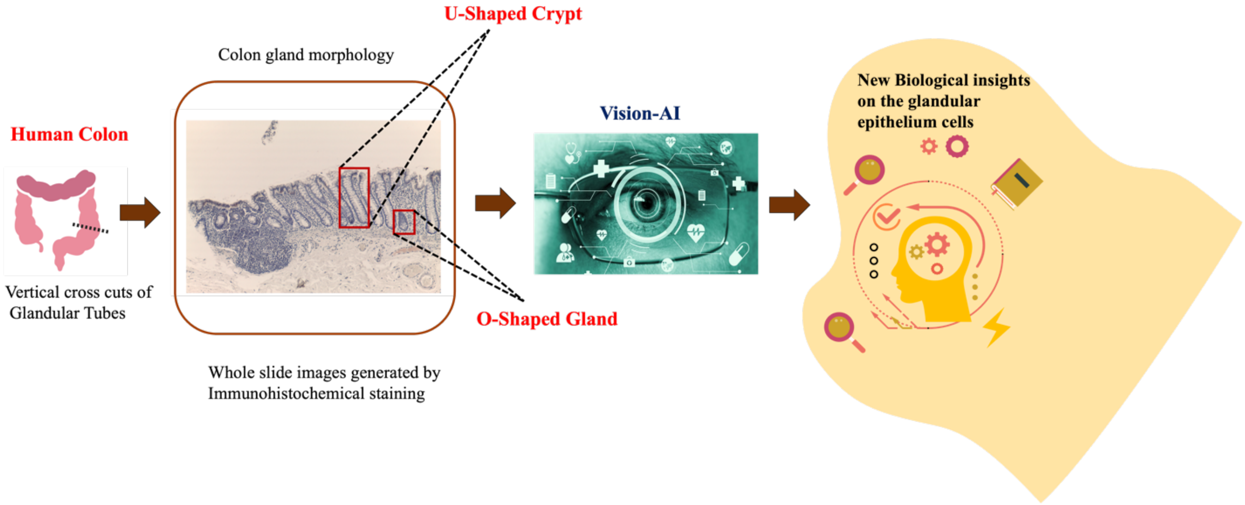

U-Shaped Crypt
Colon gland morphology
Human Colon
Vertical cross cuts of Glandular Tubes
Vision-AI
New Biological insights on the glandular epithelium cells
O-Shaped Gland
Whole slide images generated by Immunohistochemical staining

Figure 2.1: Application of vision-AI in the morphological analysis of colon crypts.

scanned slides on any platform. Utilizing DP can enhance efficiency and accuracy while decreasing operational costs, reducing data biases, and decreasing manual labor[25-26]. Traditional approaches involve extracting hand-crafted features from tissue structures or cells using classical image processing techniques[27]. DP aims to predict and characterize cancer prognosis; however, the conventional methods fall short due to poor feature selection, staining bias, and lack of generalization, leading to undesirable results. The recent advancement in Deep Learning and state-of-the-art architectures have significantly influenced biomedical image analysis, offering immense improvements in the field.

There has been a significant advancement in vision-AI using Deep Learning. Some tasks include image classification, object detection, image segmentation, etc. Analyzing any problem in Deep Learning involves a set of standard steps; preparing the data, which involves annotating the desired objects, splitting into the train, test, and valid; selecting a pre-existing model trained on a large dataset such as ImageNet or Coco; Adjusting the selected network to suit our use case by transfer learning or fine-tuning; followed by dataset evaluation and hyper-parameter optimization to enhance model accuracy. Several professionals and researchers working with medical data have thought about starting with ImageNet pre-trained weights. Medical data like X-ray differs from ImageNet data, as an X-ray image in grayscale, while the ImageNet model is trained on RGB images. Therefore, the features extracted from ImageNet might not necessarily apply to all medical images. Even medical datasets are small, as they involve huge costs in generating each cohort. The size limitation necessitates freezing most of the neural network layer to avoid overfitting. Despite the abovementioned challenges, the medical data trained on ImageNet pre-trained weights have achieved human-level accuracy[28]. Hence, with proper data selection, annotation, and pre-processing, selecting recently advanced architecture and fine-tuning on ImageNet pre-trained weights with error analysis can significantly provide outstanding results. In colon tissue, novel vision-AI can solve three major tasks instant or semantic segmentation on nuclei level and glandular areas, tissue classification (adenoma, cancerous), and detection[29].

The colon is the most extended segment of the large intestine and performs a crucial function in our body's utilization and processing of food. Uncontrolled cellular growth in the colon or rectum leads to colorectal cancer, as shown in Fig 2.2. In the initial stage, colon cancer is confined to the inner lining of the colon; as the progression of the disease, cancer infiltrates the colon's layers, extends to nearby structures, and further spreads to other organs. Glands constitute a fundamental component of the colon, and the epithelium of the colon glands contains morphologically and biochemically identifiable mature cell types. These include absorptive Enterocytes, mucus-secreting Goblet cells, Paneth cells, and undifferentiated crypt Stem cells[30] as shown in Fig 2.3. Intestinal crypts are small tubular recesses in the epithelial lining of the colon and small intestine and serve as the home for stem cells, which are vital for the maintenance and repair of the epithelial layer.

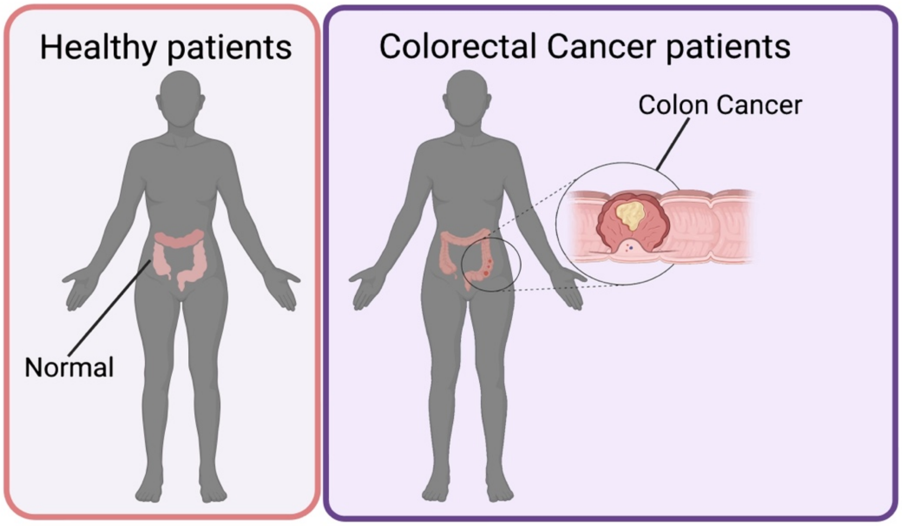


Figure 2.2: Colorectal cancer patient's colon due to uncontrolled cellular growth.

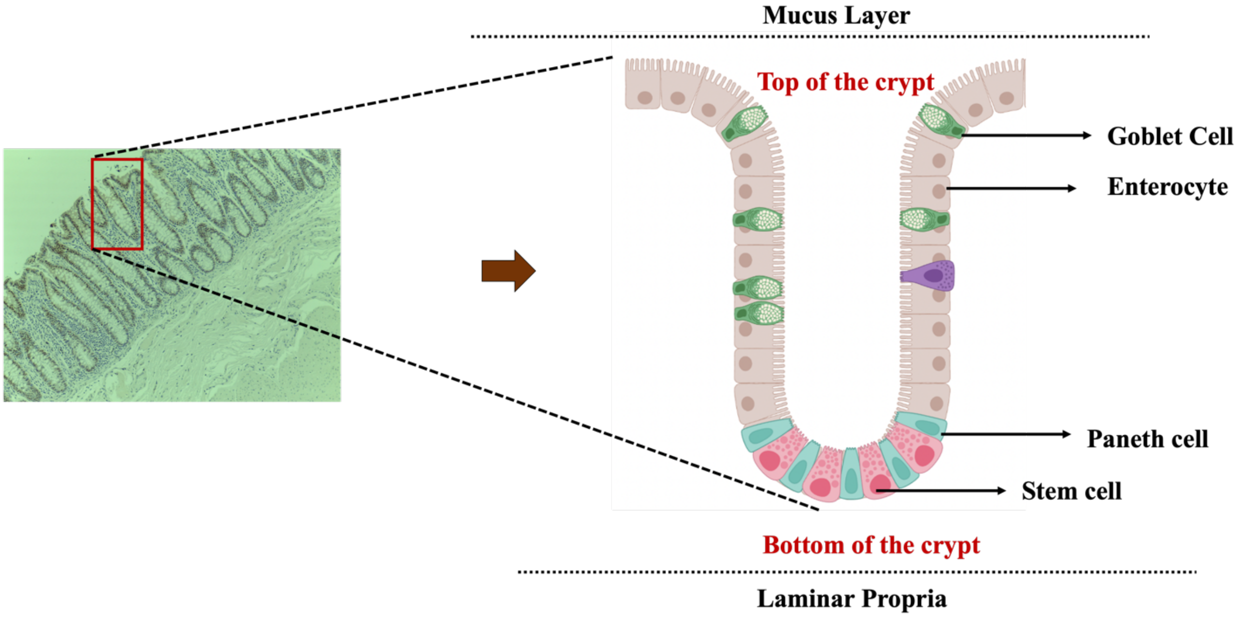


Figure 2.3: Glands constitute a fundamental component of the colon.

One important application of deep learning in colon tissue is gland segmentation[31]. Glands are the essential part of the colon, and the epithelium of the colon glands contain morphologically and biochemically identifiable mature cell types that include absorptive Enterocytes, mucus-secreting Goblet cells, Paneth cells, undifferentiated crypt Stem cells[32]. While numerous studies focus on stem cell characterization in the small intestine, research on colon crypt characterization remains limited. It is well known that crypt epithelial cells are heterogeneous, and cell purification strategies are a big limitation. Gene expression patterns specific to each cell type provide information about the differentiation states, and in Colorectal cancer (CRC), some of them become strongly prognostic[33], which Immunohistochemistry (IHC) technique can do.

IHC is a technique used by pathologists to test for the presence of clinically important biomarkers in tumors. It generally involves identifying specific antigens present in the tissue sample and staining them with corresponding antibodies. This process results in a microscopic slide that typically displays two colors, visually representing the target gene expression. Studies using Immunohistology on normal colon crypts have demonstrated that ALCAM (also known as CD166 Antigen) exhibits high expression at the bottom of the crypts and low expression at the top. Therefore, it is considered a good biomarker for intestinal stem cells and Paneth cells[34]. Furthermore, our previous study identified a strong Boolean implication relationship between CDX2 and ALCAM; "$CDX2\ low\ => ALCAM\ high$"[35].This means when CDX2 expression is low in the bulk tumor tissue, ALCAM is high in those samples. CDX2 is a protein-coding gene and a member of the caudal-related homeobox transcription factor family, which plays an essential role in regulating cell differentiation and development. CDX2 is typically expressed in the epithelium. In the epithelium layer, the cells are interconnected vis junction cells and act as a glue to prevent individual cells from separating when touched. The luminal surface of the colon, when viewed in 3D, appears mountain like structure that expands the surface areas and allows for increase food and water absorption. Stem cells are located at the bottom of these crypt-like structure. If all stem cells are eliminated from the colon, the entire crypt structure collapses.

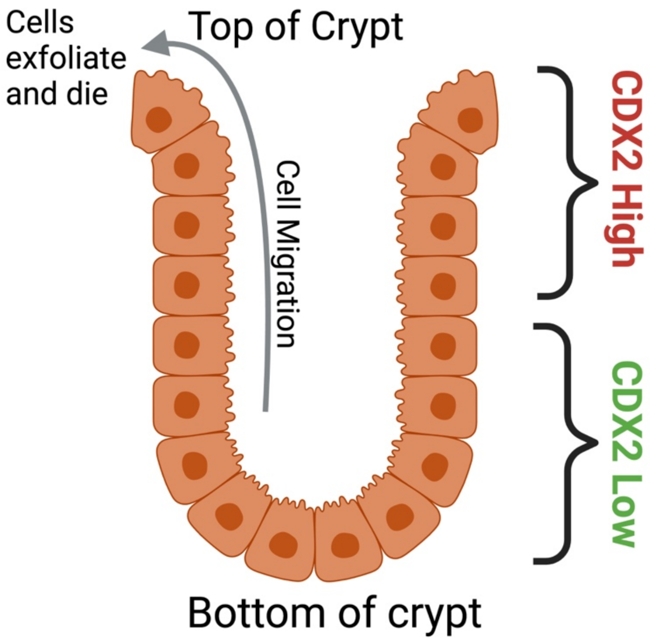


Figure 2.4: Epithelium Gland of Colon Tissue showing the top and bottom of the crypt.

This results in the death of the epithelial lining. Scientists started deleting various genes, and upon deleting CDX2 lead to the collapse of the epithelium. Therefore, CDX2 expression has been studied extensively and is now recognized as crucial for defining the identity of colon tissue. Earlier CDX2 is believed to be a Diagnostic biomarker, but Piero Dalerba, Debashis Sahoo, et al. (2016) proposed CDX2 as a prognostic biomarker in stage II and stage III colon cancer, states that CDX2 can be a prognostic biomarker. From the above discussed CDX2 ALCAM relationship we hypothesize that the stem and progenitor cells of the colon tissue that are ALCAM marker of normal high have low levels of CDX2 and is tightly linked to the state of differentiation in colon tissue. Having these two facts together, does it mean that CDX2 also has a differential expression along the crypts? Fig 2.4 shows the schematic of epithelium gland of colon tissue.

2.2.1 Gland Segmentation

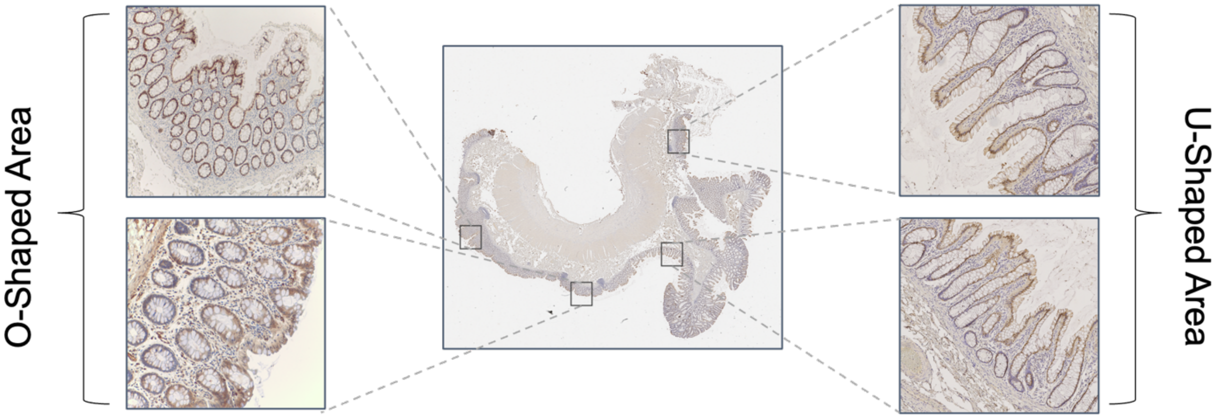
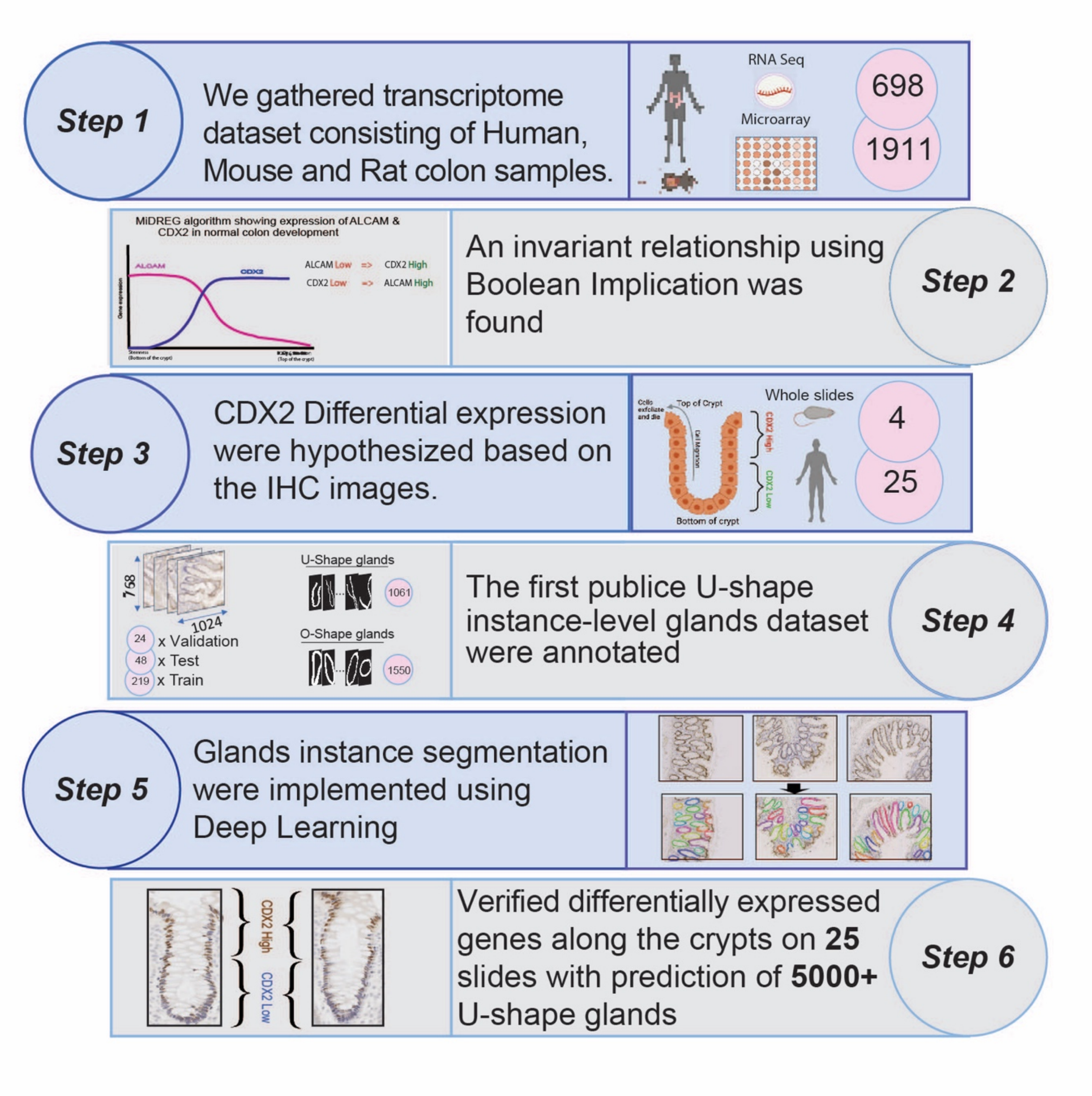


Figure 2.5: Effects of different cuts on 3D tabular glands when we project then in 2D.

In Colorectal cancer, glands morphology and structure play a crucial role in cancer grading[36-37]. The colon glands are tabular-shape epithelium layer that are spread all over the outside layer of the colon tissue[38]. A normal tissue can have more than millions of glandular objects[39]. The mechanism for their cell regeneration, forms a pipeline from the bottom, to the top where the last cell at the top usually dies and release from the tissue. This circular usually happens every two weeks which is one of the fastest regeneration processes in the body where any corruption along the process can lead to cancer formation. As discussed above the epithelium layers contains Goblet cells and Absorptive cells for absorbing nutrients and water or secreting the enzymes and mucus[40]. The inner part of the gland mostly contains the stem cell and Paneth cell. The small intestine and large intestine (colon) both have these glands, however the large intestine doesn't have the Paneth cells. In this study I focus on the morphological analysis of the glands called crypts. Each colonic crypt, depending on the cutting across or parallel to the long-axis, can either have a O-shaped or U-shaped images as shown in Fig 2.5. Nevertheless, most of the published works in this area are using the O-shaped images.

The appearance of histological object like gland as you can see in the Fig 2.3 varies in their size, structural shape, and boundaries. Studies showed that different stages of cancers lead to different outcome of gland morphology, thus H&E images are great resource to predict the cancer degree specifically adenocarcinoma, the most common type of cancer in colon tissue. Focusing on these objects, requires isolating them from the rest of the tissue, which is get done by semantic and instance segmentation methods. Fig 2.6 shows the overall steps followed in this research.



Step 1 We gathered transcriptome dataset consisting of Human, Mouse and Rat colon samples.

Step 2 An invariant relationship using Boolean Implication was found

Step 3 CDX2 Differential expression were hypothesized based on the IHC images.

Step 4 The first publice U-shape instance-level glands dataset were annotated

Step 5 Glands instance segmentation were implemented using Deep Learning

Step 6 Verified differentially expressed genes along the crypts on **25** slides with prediction of **5000+** U-shape glands

Figure 2.6: Individual steps involved during the entire study.

2.3 Methods

2.3.1 Gene Expression Data

We used available microarray and RNA-Seq datasets in NCBI Gene Expression Omnibus (GEO) database.

Table 2.1: Gene Expression Analysis using Boolean Implication and visualized in Hegemon[41].

Data Type	n (Sample)	Platform
Microarray	GSE42069 (n = 90)	Affymetrix Human Genome Homo Sapiens
Microarray	GSE120699 (n = 4)	Affymetrix Human Gene 1.0 ST Array Homo Sapiens
Microarray	GSE45134 (n = 6)	Affymetrix Human Gene 1.0 ST Array Homo Sapiens
RNA-Seq	GSE135460 (n = 30)	Illumina NextSeq 500 Homo Sapiens
RNA-Seq	GSE106378 (n = 9)	Illumina HiSeq 2000 Homo Sapiens
RNA-Seq	GSE162633 (n = 8)	Illumina HiSeq 2500 Homo Sapiens

2.3.2 Boolean Implication Relationship

The expression values of each gene were ordered from low to high, and a rising step function was computed to define a threshold by the StepMiner algorithm in the individual data set. If the assigned threshold for a gene was t, then expression levels above t + 0.5 were classified as high, and the expression levels below t - 0.5 were classified as low. Expression levels between t - 0.5 and t + 0.5 were classified as intermediate. A previously published BooleanNet algorithm was performed to determine Boolean Implication relationships between genes[42]. Briefly, the

BooleanNet algorithm searches for at least one sparsely populated quadrant in a scatterplot between two genes. The intermediate expression values were ignored by the BooleanNet algorithm. There were six possible scenarios: one of the four quadrants was sparse (four asymmetric Boolean implications), and two diagonally opposite quadrants were sparse (Equivalent and Opposite Boolean implications). Using this method, the Boolean relationship between ALCAM and CDX2 expression has been found in the bulk tissue dataset.

2.3.3 Immunohistochemistry Staining procedure

Immunohistochemical staining is a technique used to detect specific proteins in tissue sections using antibodies that bind to these proteins. The tissue sections are first fixed and embedded in paraffin wax or frozen in liquid nitrogen. Then, the sections are treated with primary antibodies that bind to the target protein, followed by secondary antibodies conjugated to a detection system, such as fluorescent dyes or enzymes. The resulting staining pattern can be visualized using a microscope and can provide information on the distribution and abundance of the target protein in the tissue sample. For our purposes, we wish to gain information on the distribution and expression pattern of the CDX2 gene along the colon crypt. Further, we aim to develop an understanding of the differential expression of the CDX2 along the top and bottom of the crypt.

Fig 2.7 shows the schematic representation of a novel IHC staining procedure introduced in this research. More specifically, we have introduced a new method for the antigen retrieval part of the overall staining process. The standard method used in practice has been pressure cooking at a pH of 9.0. In this research, we have proposed using a new boiling method (pH 9.0) which succeeds in showing the differential expression of the CDX2 gene along the top and bottom of the crypt.

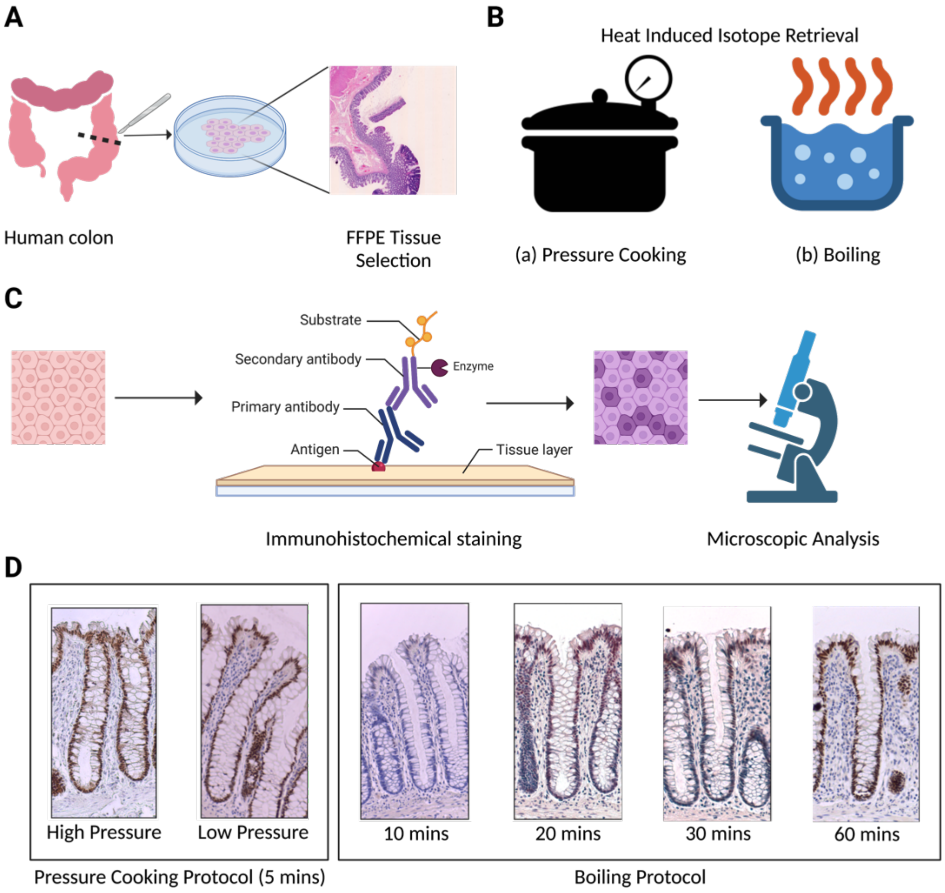


Figure 2.7: New staining procedure introduced and adopted in this research.

We begin the staining procedure with the FFPE selection of the tissue from the colon crypt. To contrast the differences between the standard (pressure cooking) and new (boiling) antigen retrieval procedures, we use both approaches parallelly and compare the final stained slide segments. Once antigen retrieval from the tissue is completed using either pressure cooking or boiling, IHC staining is performed on them. More specifically, after the H2O2 incubation and blocking, we incubate with the CDX2-88 primary antibody, followed by incubation with the conjugated secondary antibody. Finally, the substrate is added for color development, and lastly, we counterstain with hematoxylin. At the end of this process, we get the stained tissue slides which are then analyzed under a microscope.

2.3.4 Data Preparation

We have used different staining slides from human and mouse colon samples for which standard protocol for gene targets has been followed. Specifically, we stained different slides in svs and Czi format for CDX2, KRT20, CAI, MUC2, and SLC26A3. The colon glands appear as U or O-shape in each slide depending upon whether they are cut horizontally or vertically. To get more U-shapes, vertical cross-cuts have been applied to each slide. Normal crypt regions have been extracted using QuPath software[43], and the region of interest (ROI) has been resized to zero padding and has been applied wherever necessary. Then Reinhard normalization was applied to the images so that they have the same color spectrum. After this, for the training procedure, 1061 U-crypts and 1550 O-glands were annotated using Gimp software from 291 slide images and external datasets (Warwick QU and CRCHistoPhenotypes). These images were split into train, test, and validation following the general rule in 219 train, 48 test, and 24 validation images. In order to provide more data for the training and improve model accuracy, image augmentation and color distortion were applied to the images. We also used the Roboflow pipeline to annotate 1308 O-glands and 928 U-crypts resized to 640*640, with 192 training, 55 validation, and 28 testing images. Fig 2.8, Fig 2.9, and Fig 2.10 shows the dataset generation and preparation for model training.

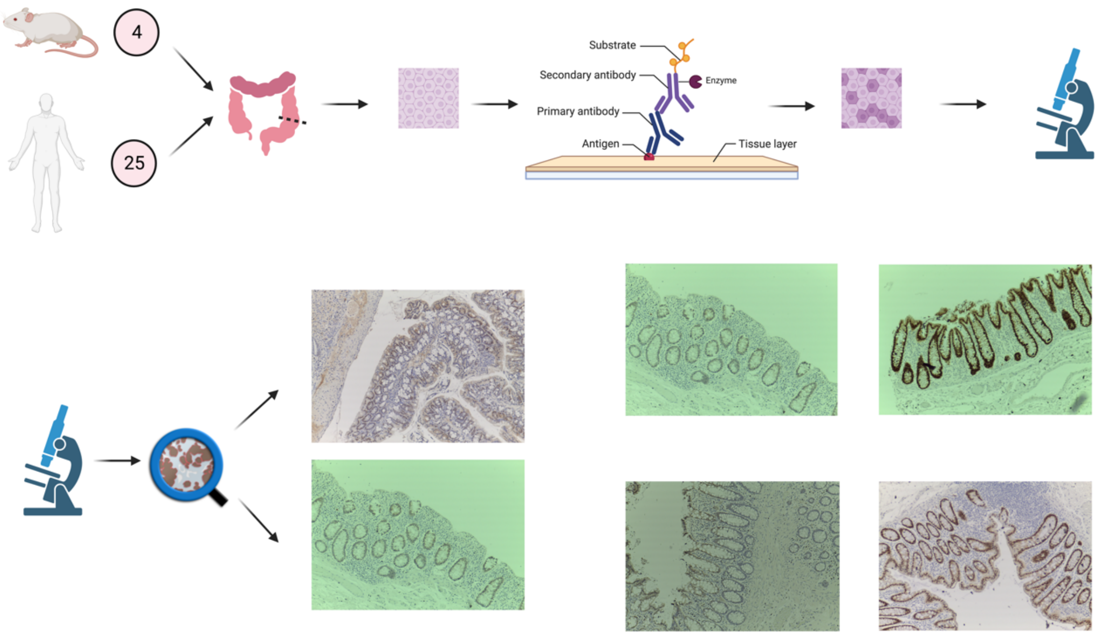


Figure 2.8: The schematic representation of whole slide image data generation.

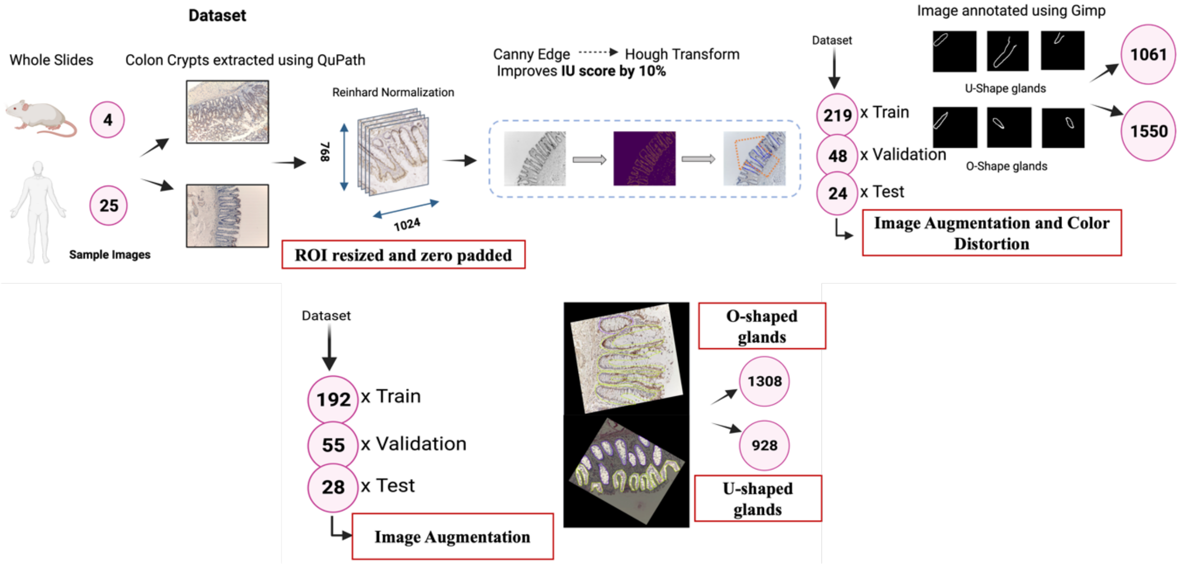


Figure 2.9: The schematic representation of dataset preparation.

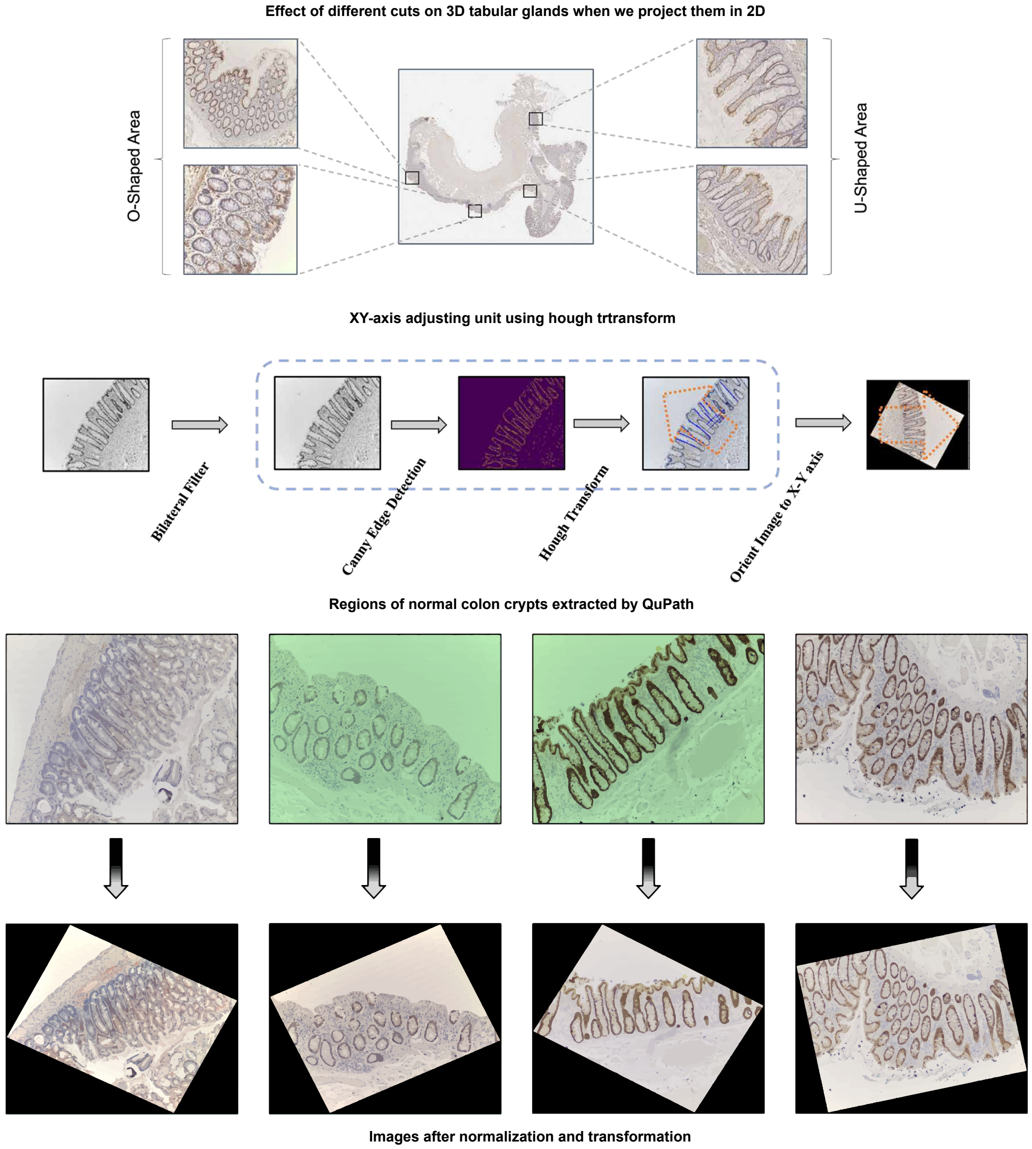

Effect of different cuts on 3D tabular glands when we project them in 2D
O-Shaped Area
U-Shaped Area
XY-axis adjusting unit using hough trtransform
Bilateral Filter
Canny Edge Detection
Hough Transform
Orient Image to X-Y axis
Regions of normal colon crypts extracted by QuPath
Images after normalization and transformation

Figure 2.10: The schematic representation of data pre-processing.

2.3.5 Gland Segmentation

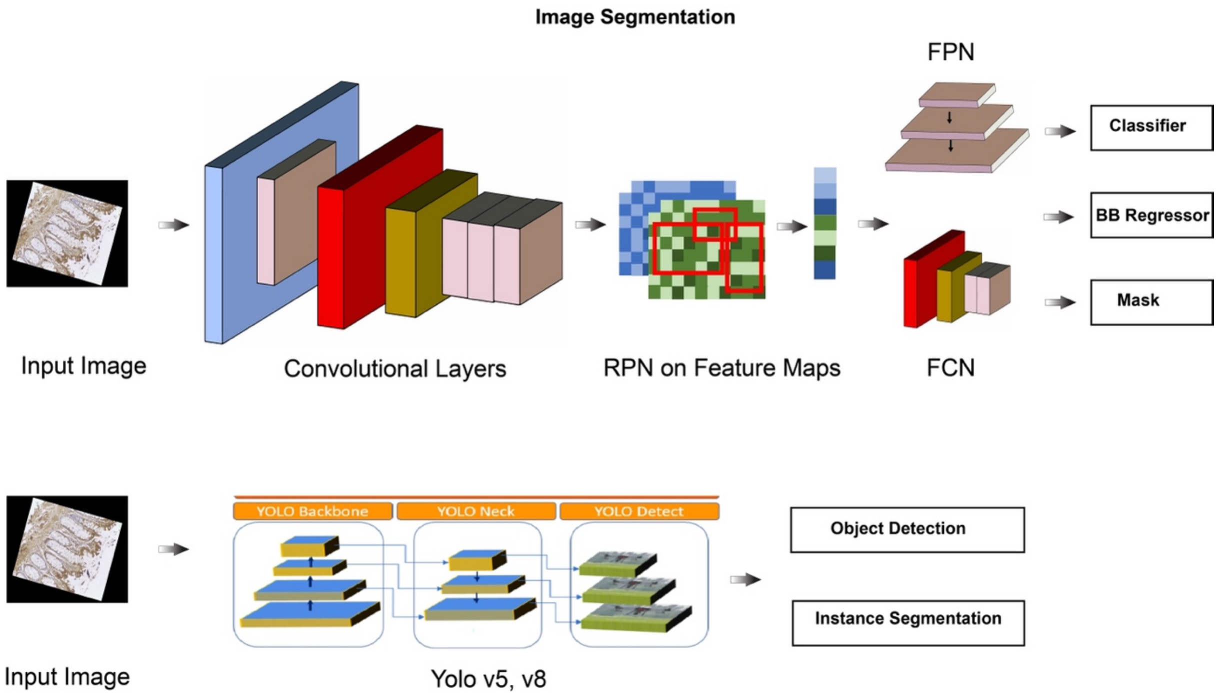


Figure 2.11: The schematic representation of maskRCNN and state-of-the-art yolo model.

Gland segmentation has been performed using the Mask R-CNN[44] model Matterport implementation. I deployed various models such as Resnet50-UNet, Resnet50-segnet, Resnet101, FCN-8, FCN-32, etc. It was found that Resnet50[45] + FPN[46] gave superior performance on our dataset in terms of accuracy and time trade-off. The learning rate has been set to 0.0001 and RPN anchor scales to (32, 64, 128, 256, 512), NMS threshold to 0.4 and min confidence to 0.7. The RPN anchor was unable to find satisfactory bounding boxes due to the presence of arbitrary orientation of crypts. Bilateral filtering was applied to blur the neighboring nucleus and form a line and then the borders were detected using canny edge detector. After this, probabilistic Hough transform has been applied to find straight lines and image is rotated along the fitted line angle to orient all the U-crypts along the same direction. This led to an improvement of 10% IU score for the U-crypt detection. I also used Yolo based models to perform object detection in yolov5 and instance segmentation in yolov8 trained on Coco dataset as shown in Fig 2.11. The state-of-the-art yolo models handle different orientation of the input image and provides improved accuracy than the maskRCNN.

2.3.6 Color Pattern Detection

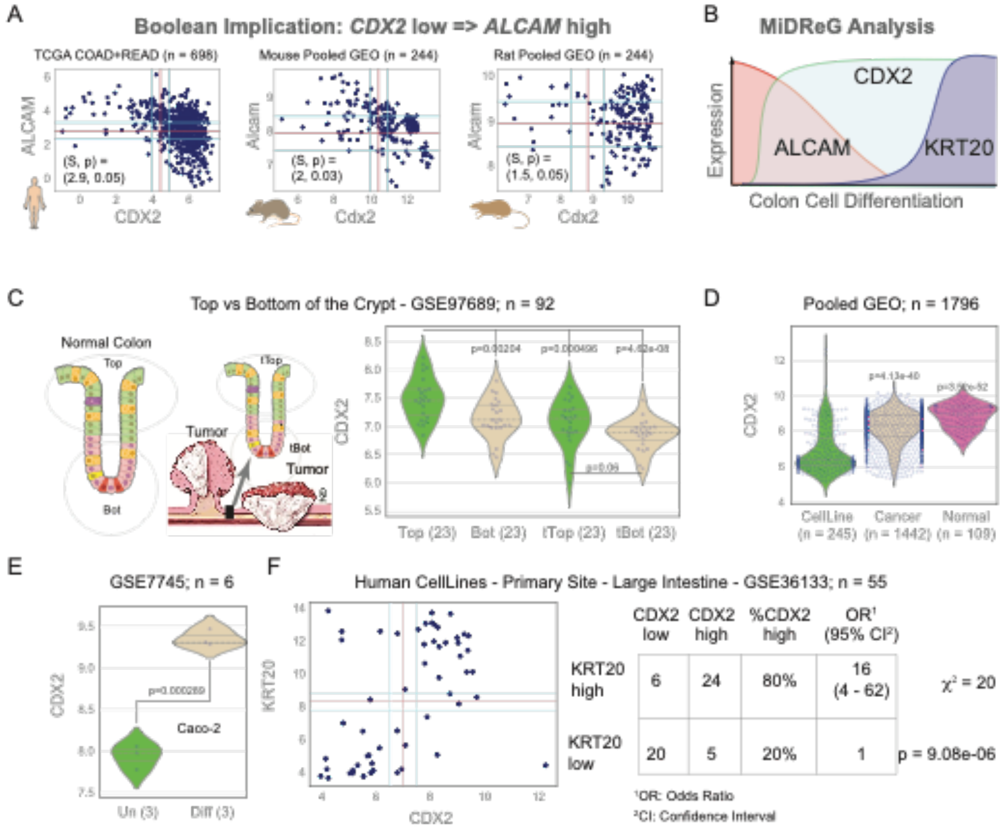
After detecting the crypts, the next step is to detect if there exists any color variation along them. To do this first the U-shaped crypts have been aligned from top to bottom by fitting ellipses and then the color pattern along these aligned crypts were measured. The nucleus is stained blue if it passes through the expression for threshold else it is stained brown. Using HSL color spectrum, each pixel is classified either as blue if Hue value lies in the range of (80, 140) and brown if it either has value between (0, 40) or (150, 180), then the ratio is calculated along the crypt to justify the pattern.

2.4 Result

2.4.1 Boolean Implication Relationship

Scientists widely known that ALCAM has differential expressions along the colon epithelial cells. I used Boolean analysis to search for other biomarkers of colon epithelial differentiation. Various transcriptomic Microarray and RNA-Seq human, mouse, and rat colon tissue datasets are collected from NCBI GEO and normalized. Gene expressions in colon tissue is searched where low expression of that gene implies the high expression of the ALCAM. Candidate genes were ranked according to the availability of clinical-grade diagnostic assays, and then CDX2

Figure 2.12: CDX2 and ALCAM relationship using Boolean Implication.



	CDX2 low	CDX2 high	%CDX2 high	OR[1] (95% CI[2])	
KRT20 high	6	24	80%	16 (4 - 62)	$\chi^2 = 20$
KRT20 low	20	5	20%	1	p = 9.08e-06

[1] OR: Odds Ratio

[2] CI: Confidence Interval

Figure 2.12: CDX2 and ALCAM relationship using Boolean Implication.

had the best score among the candidates. Our previous study, similarly, has shown a strong Boolean relationship between CDX2 and ALCAM[35]. We developed a mathematical model around CDX2 and identified CDX2 low and CDX2 high as distinct differentiation states of colon epithelial cells. Based on this, I hypothesize that the crypt base consisting of stem and progenitor cells may be CDX2 low. This Boolean implication "if CDX2 low then ALCAM high" is true across all human datasets from different platforms- Affymetrix and TCGA RNA-Seq. Based on this strong and robust relationship between CDX2 and ALCAM, I assume that there may be a Boolean invariant relationship between CDX2 and ALCAM in the colon and small intestine tissue that is preserved in both normal and cancer, and across species between human, rat, and mouse.

2.4.2 Immunohistochemical Staining Result

The role or function of CDX2 expression in colon cancer has been studied extensively. Our hypothesis contrasts with the current understanding of the CDX2 patterns in the human colon. Nuclear staining was observed from the bottom to the top of the crypt when the pressure setting was changed from high to low. It is hard to assess the CDX2 expression pattern using the IHC method quantitatively. Our mathematical model gives new insights into the differential expression patterns of CDX2 between stem cells and the differentiated cells of the colon crypt. Our result agrees with the previous CDX2 staining result. CDX2 expression may be high in all human colon epithelial cells compared to other tissue cells. However, the differentiated cells' CDX2 expression may vary between the stem cells, and it is hard to capture that difference using the IHC approach. Therefore, the antigen retrieval step is modified in the IHC technique, and the boiling method is used to analyze the CDX2 stain in a normal colon crypt. Normal crypts were stained from 5 patients using CDX2, and surprisingly, able to capture this difference using IHC. Our CDX2 staining shows that the top of the crypt cells is enriched with CDX2 positive (brown stain), whereas the bottom has CDX2 negative cells (blue stain) in the adjoining normal tissue. Fig 2.13 shows different staining, boiling, vs. pressure cooking with different settings. In boiling protocol with this type of setting, you can see a pattern along the crypt, and the expression changes as we go upward; however, in pressure cooking, we can't see any pattern. Furthermore, more CDX2 negative cells are found in the adenoma, and it is consistent all over the infected region compared to the normal crypts. Both the H&E and IHC DAB stained slides are evaluated in this study. Finally, Fig 2.14 shows all boiling and pressure-cooking samples in one plot. As you can see in boiling, as we increase time, the bottom and top of the crypt are getting the same staining, and you see the same

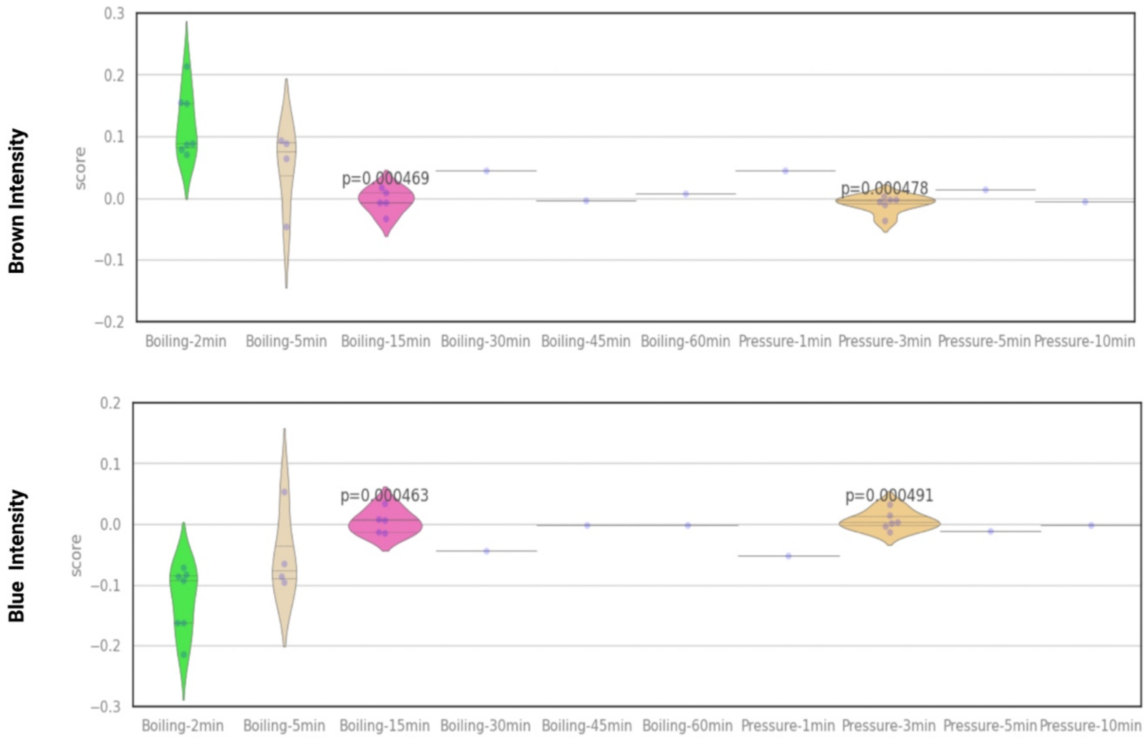


Figure 2.13: Violin plots of all the patient difference between the top and bottom of the crypts.

pattern for pressure cooking. However, in pressure cooking, the differences are not as clear as in the boiling setting. This data replication shows that the finding is promising.

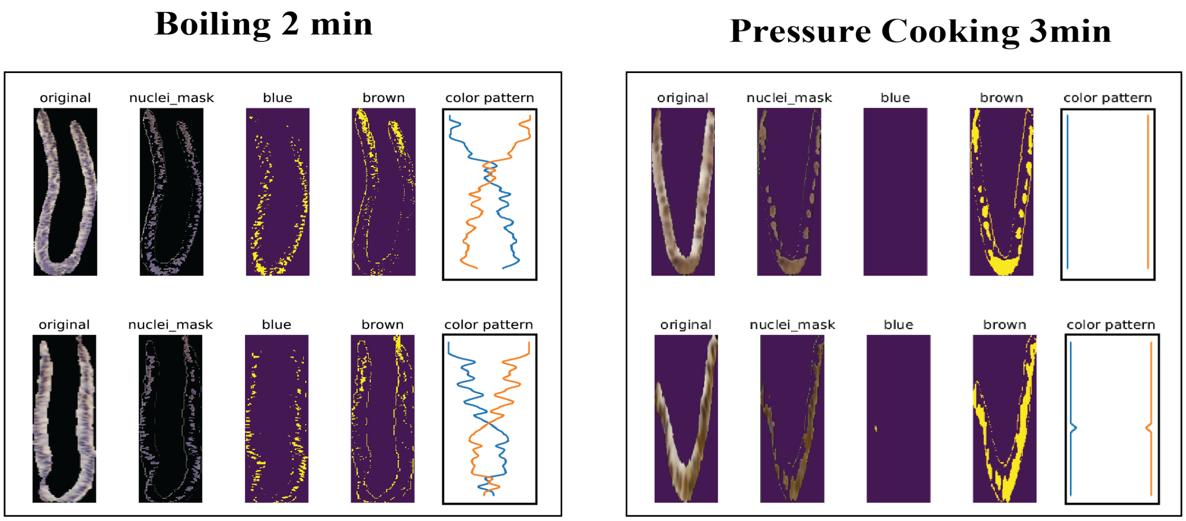


Figure 2.14: Segmented crypt in boiling 2 minutes and pressure cooking 3 minutes.

2.4.3 Implementation Details

The suggested MaskRCNN model was trained on the glandular regions with 90 epochs by the Adam optimizer. The model was trained on 4 NVIDIA GPU 1080 GTX, Nvidia driver v430, Cuda v10.1, and TensorFlow version v2.2. Each epoch took around 15 minutes to finish for the Resnet50 backbone model. I started with an initial learning rate of 10-3 and random weights. Each inference takes about 30 seconds. All validation loss function has decreasing value except the RPN class loss, which seems to overfit after epoch 40th because the two classes will be the same if the O-shape glands get cropped by the bounding box; This makes the network convergence even harder with our limited dataset. The suggested Yolo models were trained on the Google Colab Pro + version. Object detection with the crypt and gland annotation saved in png format is performed using a YOLOv5s model. The original YOLOv5s model was trained on 80 classes; we modified the same and built a custom YOLOv5s model with two classes for our use case. Google Colab used Tesla T4 GPU and trained the model for 100 epochs with a batch size 16. In comparison, the YOLOv8 model was used to perform instance segmentation on the new annotated images. The model was trained for 100 epochs.

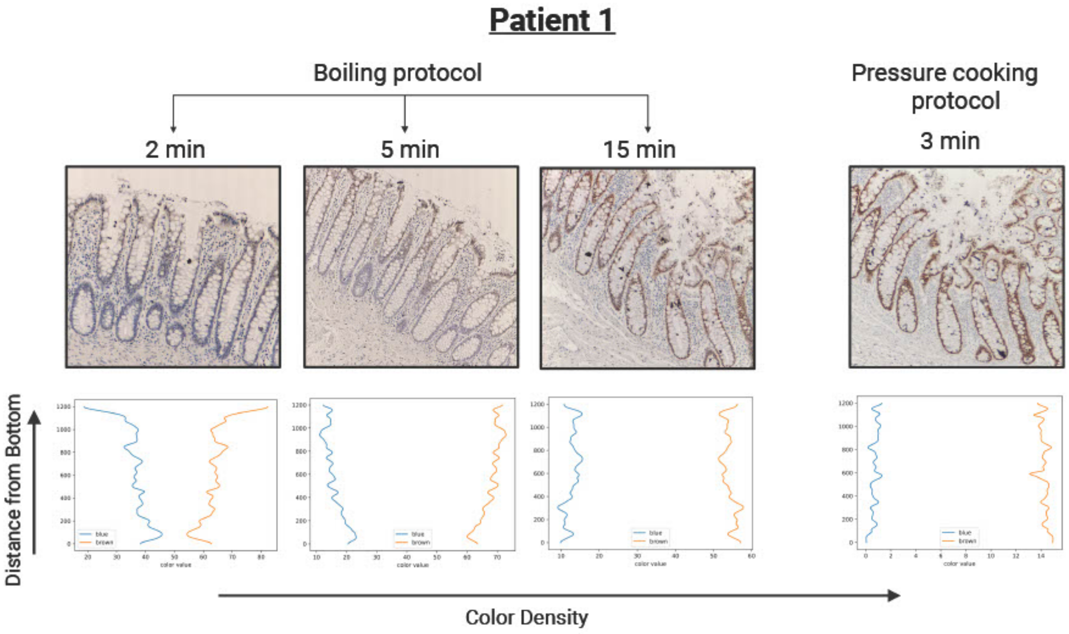


Figure 2.15: Aggregated result for all crypts in patient 1 pattern in different staining protocol.

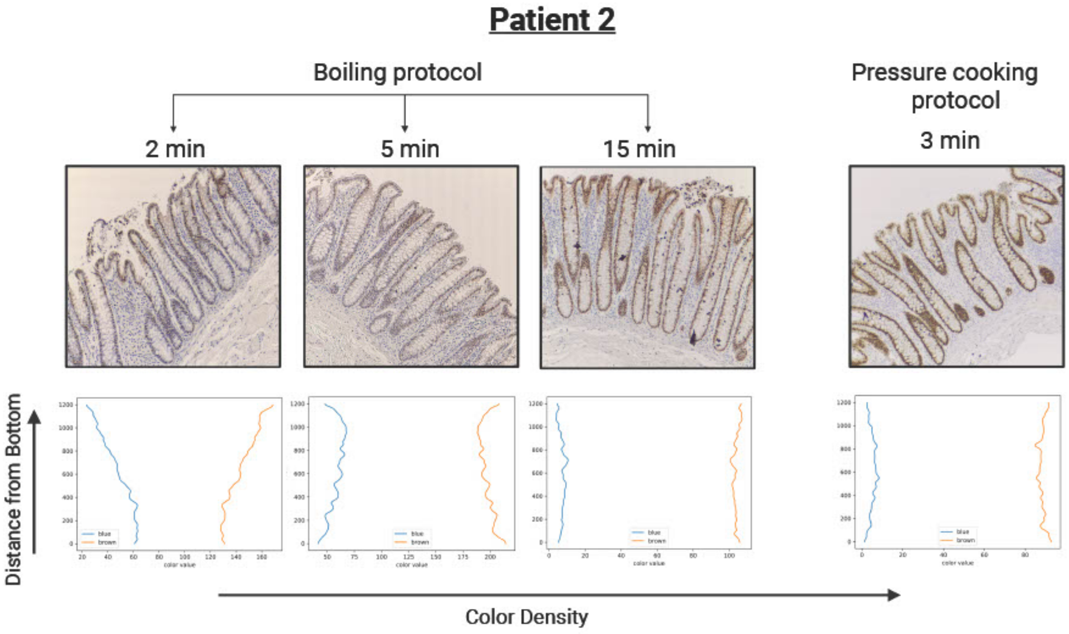


Figure 2.16: Aggregated result for all crypts in patient 2 pattern in different staining protocol.

.

2.4.4 Evaluation and Comparison

In the proposed model using mask RCNN only objects with min confidence and non-maximum suppression of 0.7 and 0.4 were kept. To assess the semantic segmentation performance F1 and intersection over union (IOU) scores were computed. For every detected pixel 3 different classes are assigned: background, U-shape gland, or O-shape gland. Having the correspondent set of pixels as the ground truth, the IOU score is calculated for each class to measure the similarity using:

$$IOU_{score} = \frac{predicted \ \cap groundtruth}{predicted \ \cup groundtruth}$$

Also, the precision and recall for each class was calculated and F1 score was computed using:

$$F_{1_{score}} = 2 \times \frac{precision \ \times \ recall}{precision + recall}$$

A set of models, including Segnet[47] and UNet[48], which are widely used in previous colon segmentation, were compared against the MaskRCNN. Table 2.2 shows the model comparison, which demonstrates all the model's scores for each class. We evaluate each model on different classes with two scores. We also train all models on the normal colon crypt dataset (our dataset) and then test them on our test set and an external dataset (H&E-stained Warwick QU and CRCHistoPhenotypes) to measure how well each model can be generalized. As you can see, FCN and PSPnet[49] didn't perform well on the test. Segnet and Unet needed to work better with others. However, VGG-Net as a backbone had improved the accuracy. Nevertheless, both achieved the best score with Resnet50 as a backbone. Even though these models can segment images from the same distribution as a training set, they perform poorly when we generalize the model with the external dataset. On the other hand, MaskRCNN outperforms all other models in all cases. Furthermore, adding our new XY-alignment element before MaskRCNN inputs increased the accuracy by 5% on U-shape crypts which shows that this technique is useful. However, due to decreased image quality, O-shape gland scores were reduced by 3%. You can see the configuration of MaskRCNN has almost the same score on the external dataset, which means these models are robust to the data source. Fig 2.18 shows the segmentation result in Yolo v5 model with mAP50, and mAP50-95 score as 0.776 and 0.0491 respectively. Fig 2.19 shows the segmentation result in Yolo v8 model for gland and crypt with mAP50 score as 0.937, mAP50-95 score as 0.567 and mAP50 score as 0.748, mAP50-95 score as 0.654 respectively. This shows that the Yolo based model segmentation outperformed maskRCNN. Fig 2.17 shows the result obtained using maskRCNN.

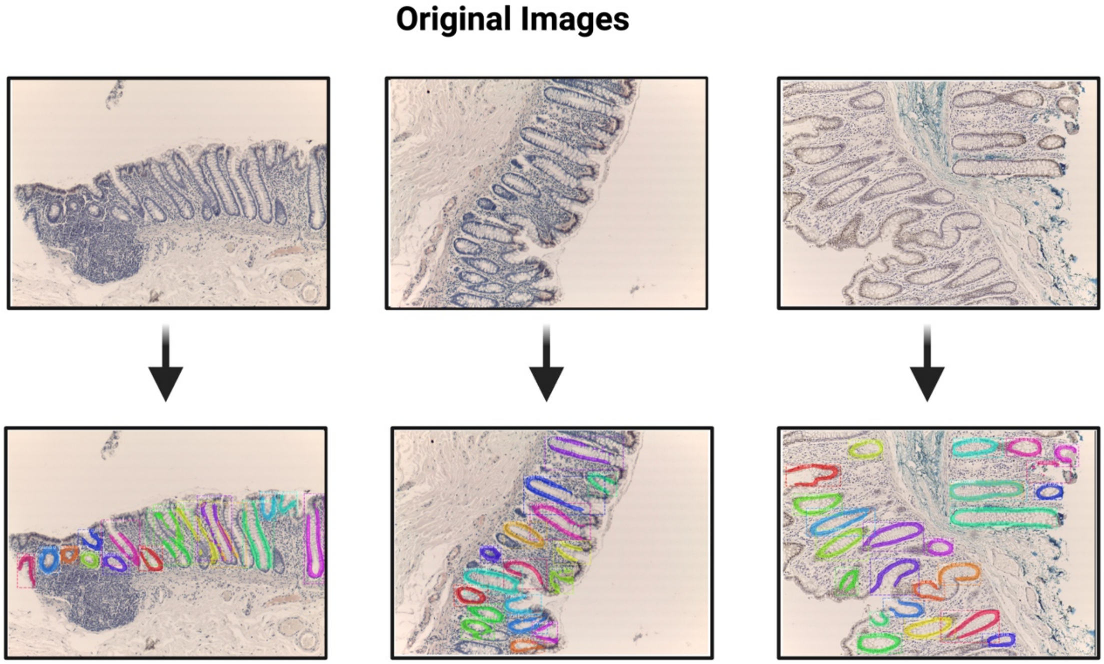


Figure 2.17: Original images with detected mask using novel maskRCNN.

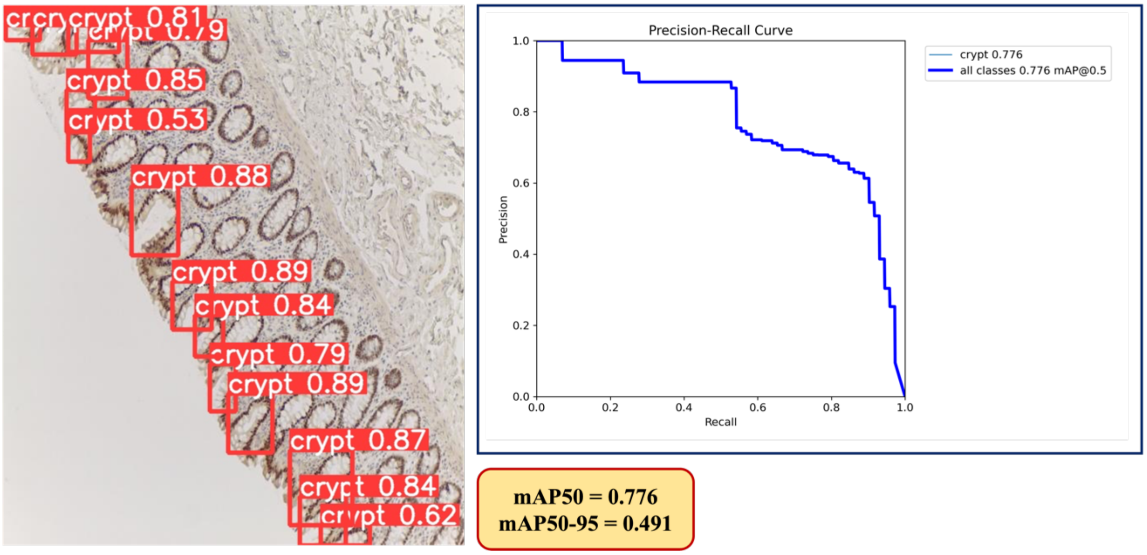


Figure 2.18: Detected segmentation result using Yolo v5.

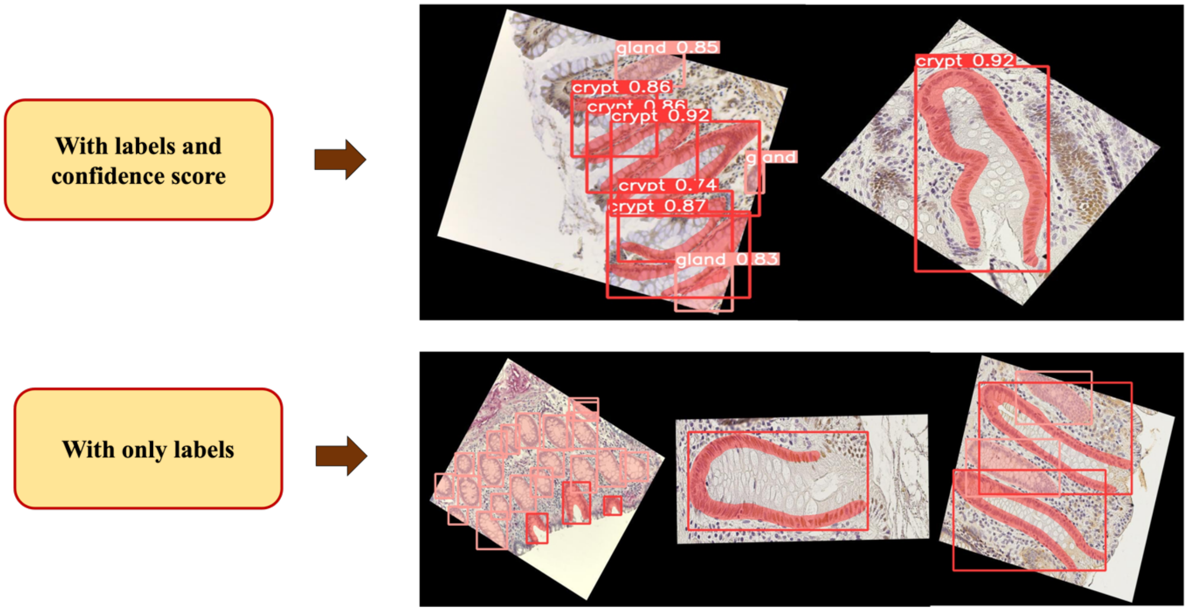


Figure 2.19: Detected segmentation result using Yolo v8.

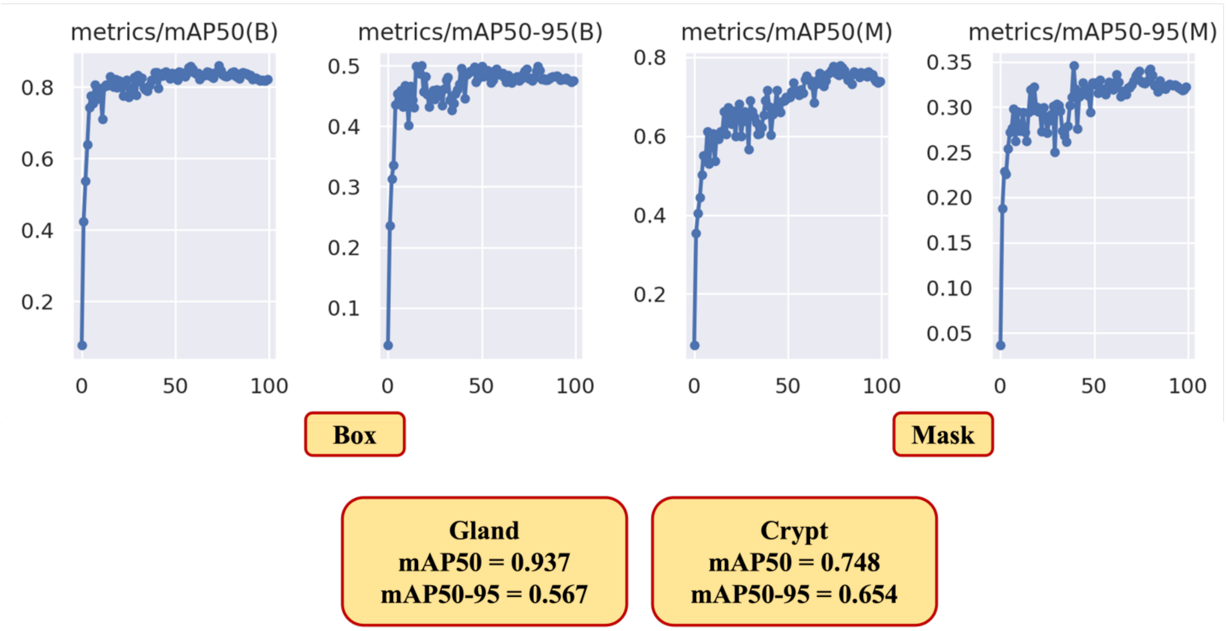


Figure 2.20: Performance metrices for results obtained using Yolo v8 for gland and crypt.

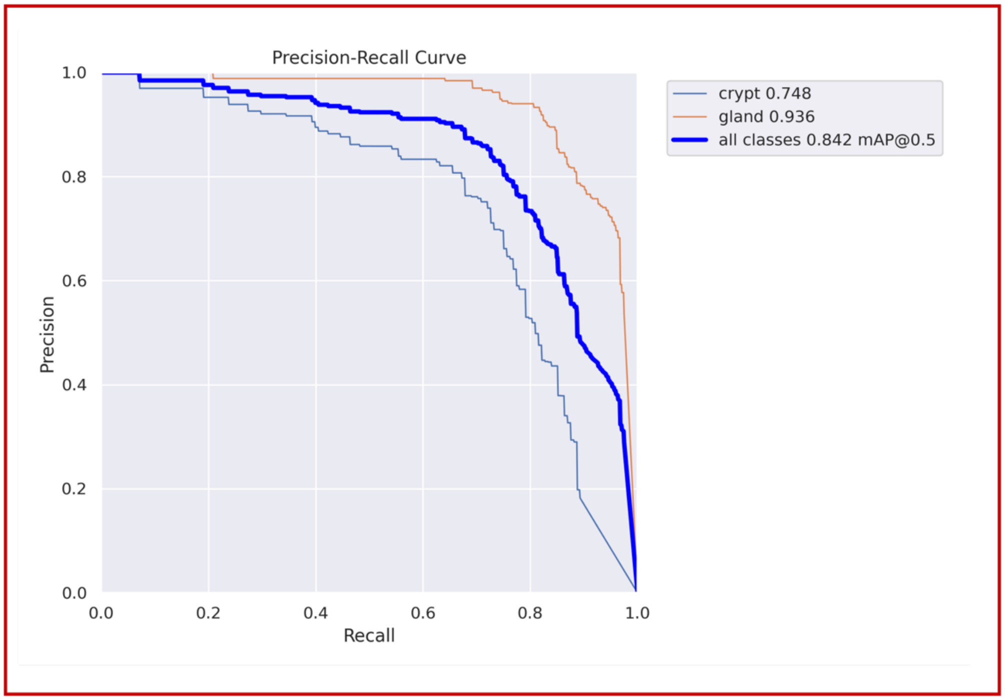


Figure 2.21: Precision and Recall curve using Yolo v8 for gland and crypt box.

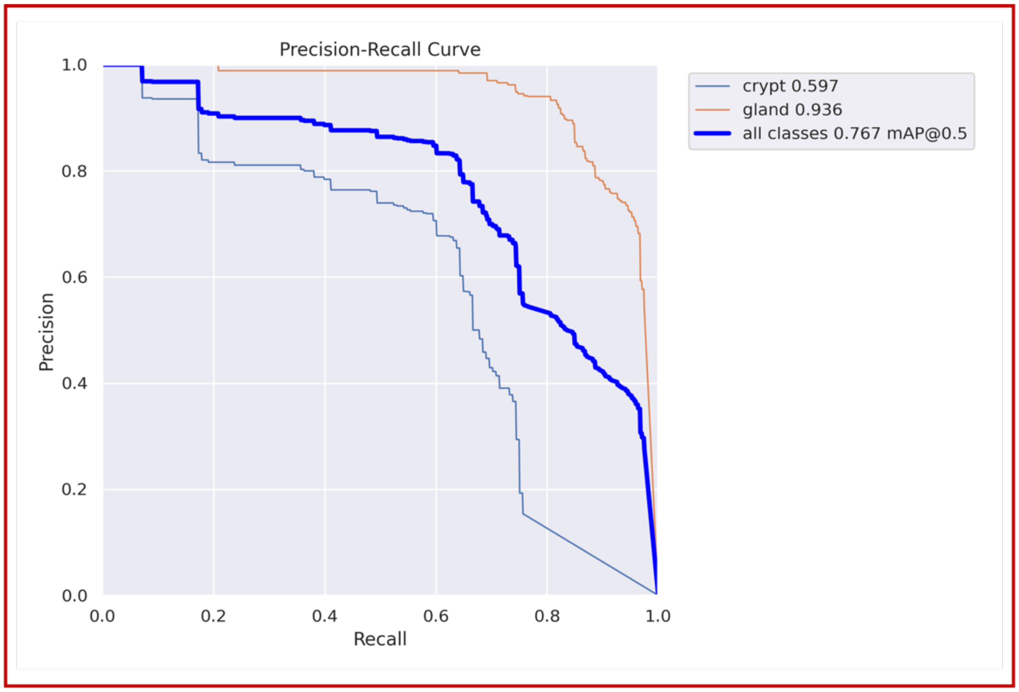


Figure 2.22: Precision and Recall curve using Yolo v8 for gland and crypt mask.

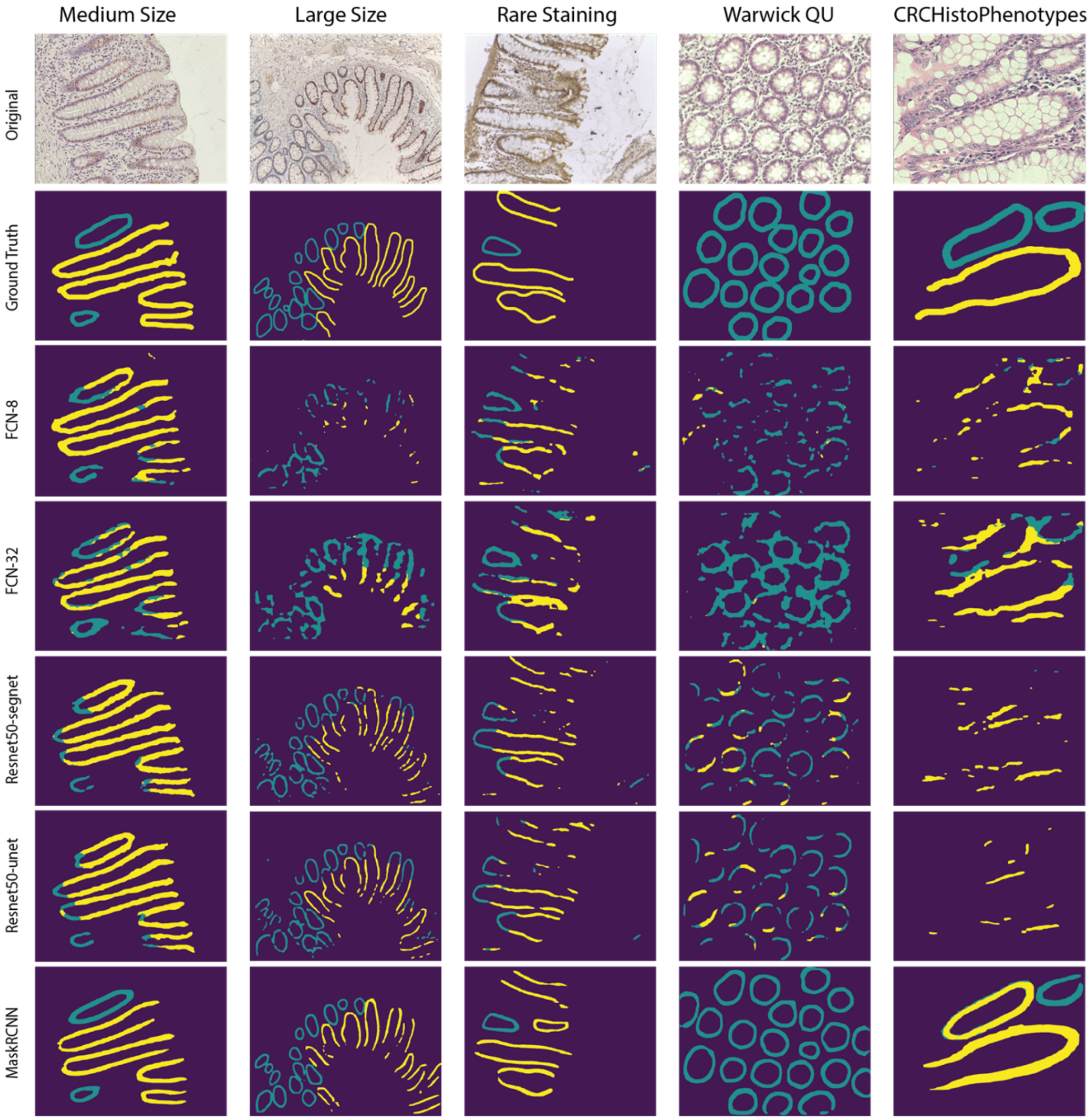
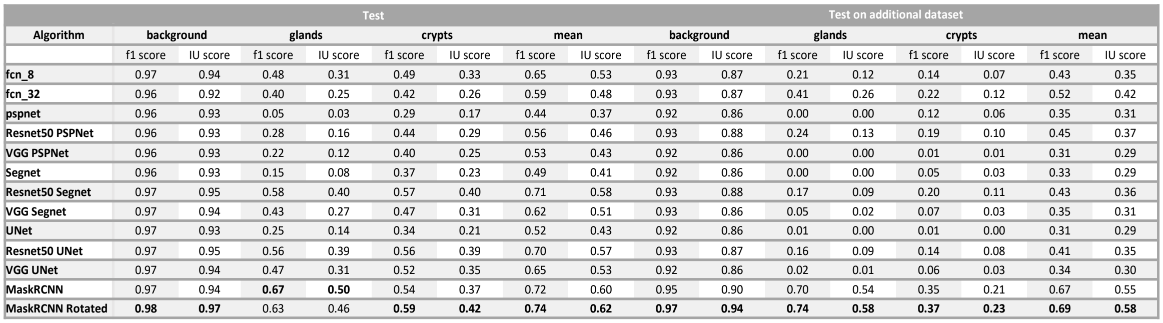


Figure 2.23: Example of the output of semantic segmentation models in different settings.

Table 2.2: Evaluation metrices for all methods shows the MaskRCNN outperforms the rest.



	Test								Test on additional dataset							
Algorithm	**background**		**glands**		**crypts**		**mean**		**background**		**glands**		**crypts**		**mean**	
	f1 score	IU score	f1 score	IU score	f1 score	IU score	f1 score	IU score	f1 score	IU score	f1 score	IU score	f1 score	IU score	f1 score	IU score
fcn_8	0.97	0.94	0.48	0.31	0.49	0.33	0.65	0.53	0.93	0.87	0.21	0.12	0.14	0.07	0.43	0.35
fcn_32	0.96	0.92	0.40	0.25	0.42	0.26	0.59	0.48	0.93	0.87	0.41	0.26	0.22	0.12	0.52	0.42
pspnet	0.96	0.93	0.05	0.03	0.29	0.17	0.44	0.37	0.92	0.86	0.00	0.00	0.12	0.06	0.35	0.31
Resnet50 PSPNet	0.96	0.93	0.28	0.16	0.44	0.29	0.56	0.46	0.93	0.88	0.24	0.13	0.19	0.10	0.45	0.37
VGG PSPNet	0.96	0.93	0.22	0.12	0.40	0.25	0.53	0.43	0.92	0.86	0.00	0.00	0.01	0.01	0.31	0.29
Segnet	0.96	0.93	0.15	0.08	0.37	0.23	0.49	0.41	0.92	0.86	0.00	0.00	0.05	0.03	0.33	0.29
Resnet50 Segnet	0.97	0.95	0.58	0.40	0.57	0.40	0.71	0.58	0.93	0.88	0.17	0.09	0.20	0.11	0.43	0.36
VGG Segnet	0.97	0.94	0.43	0.27	0.47	0.31	0.62	0.51	0.93	0.86	0.05	0.02	0.07	0.03	0.35	0.31
UNet	0.97	0.93	0.25	0.14	0.34	0.21	0.52	0.43	0.92	0.86	0.01	0.00	0.01	0.00	0.31	0.29
Resnet50 UNet	0.97	0.95	0.56	0.39	0.56	0.39	0.70	0.57	0.93	0.87	0.16	0.09	0.14	0.08	0.41	0.35
VGG UNet	0.97	0.94	0.47	0.31	0.52	0.35	0.65	0.53	0.92	0.86	0.02	0.01	0.06	0.03	0.34	0.30
MaskRCNN	0.97	0.94	**0.67**	**0.50**	0.54	0.37	0.72	0.60	0.95	0.90	0.70	0.54	0.35	0.21	0.67	0.55
MaskRCNN Rotated	**0.98**	**0.97**	0.63	0.46	**0.59**	**0.42**	**0.74**	**0.62**	**0.97**	**0.94**	**0.74**	**0.58**	**0.37**	**0.23**	**0.69**	**0.58**

2.5 Discussion

Our previous study discovered a Boolean implication relationship between CDX2 and ALCAM, specifically. However, when we used the pressure-cooking method (specifically, the Nordic QC recommended protocol for CDX2 obtained in run 48) for heat-induced epitope retrieval (HIER) in the immunohistochemical staining procedure, this relationship between CDX2 and ALCAM was not observed. On the contrary, the results from this protocol showed a high CDX2 expression along the entire length of the epithelial cells in the crypt. This indicated that pressure cooking was unsuitable for observing the CDX2 pattern in the normal colon tissue. Our speculation for this behavior is that the combination of pressure and high temperature during pressure cooking might be a limiting factor for CDX2 expression leading to the staining of all cells. Therefore, we introduced a new boiling approach as the antigen retrieval procedure in this study. This method successfully helped us identify the CDX2 low expression at the bottom of the crypt, highlighting the differential expression of the CDX2 along the entire length of the crypt. Hence, this finding demonstrates the effect of temperature on the detection limit of the CDX2 expression. We believe this study may help in improved diagnosis of colorectal cancer and a better tissue organization from the top to the bottom of a normal crypt.

This study introduces a novel deep-learning approach for epithelium segmentation in normal colon tissue slide images. While many advances have been made in glandular structure segmentation within the tissue images, vertical crypt segmentation has yet to see much progress. We address this gap by developing a new algorithm that integrates computer vision and deep learning with our Boolean analysis framework to create a mathematical model of the human colon tissue. Our methodology leverages the characteristics of the tissue slide images with which we perform statistical analysis of the staining patterns. Using computational techniques, we successfully segmented each stained tissue image into vertical colon crypts with top and bottom orientations. Integrating computational approaches and the Boolean framework provides valuable insights into the genetic composition of colorectal cancer. Our mathematical model is based on a set of robust Boolean invariant relationships between genes, the discovery of which is done with the help of existing large-scale cancer datasets. These invariant relationships reveal new insights into the biology of the human tissue. Manual identification and grading of colon cancer through the analysis of biopsy specimens is a tedious and time-consuming task generally undertaken by pathologists. Our computer-aided diagnostic tool will help improve the efficiency and accuracy of this histopathological diagnosis, ultimately leading to the determination of the appropriate treatment for the patient.

2.6 Conclusion and Future work

In this research we have shown that digital pathology can help us to diagnose the colon cancer. We discussed about different approaches in deep learning for nuclei and gland segmentation and how these finding can contribute. Later we explained our new protocol and staining for CDX2. Furthermore, we elaborate on our new model for gland instance segmentation on our new dataset and we compared it with different state-of-the-art methods. Using novel vision-AI developed a first-of-its-kind colon U-shaped crypt segmentation. Also, using the proposed new IHC staining process along with vision-AI model we saw there is a differential expression at the bottom vs top of the colon crypts. We can identify more patient with high risk using this potential biomarker and I belief that this finding can open new doors to identify other features of colon stem cells. Using this research, we can study macrophage polarity in colon crypt.

Acknowledgements

Chapter 2, in full, is currently being prepared for submission for publication of the material. Behroozikhah, Mahdi*; Khandelwal, Soni*; Nanda, Amitash*; Dang, Dharanidhar; Samantaray, Atishna; Prabhudesai, Arya; Sahoo, Debashis. The thesis author was the primary investigator and author of this material.

CHAPTER 3. Morphological Analysis of Organoids using Vision-AI

3.1 Abstract

Organoid cultures are 3D in vitro tissue construct that emulates their corresponding in vivo organ. Organoids' accurate mimicking nature has made them powerful in vitro models to study various aspects of a tissue. Organoids are generally grown in a 3D setup using naturally derived or synthetic extracellular matrices. They are commonly studied by investigating their morphological features and growth characteristics. However, such a practice is very challenging due to the inherent imaging artifacts in organoid images. Recently, very few segmentation techniques have been introduced in the literature to perform localization and quantification of organoids. Unfortunately, no attempts have been made to classify healthy and diseased organoids reliably or to predict ailments in an organoid. This research demonstrates OrgaTuring, an end-to-end deep learning approach that can efficiently locate, quantify, and classify human colon organoids. OrgaTuring can be a completely automated computational framework to investigate thousands of images without expert intervention. OrgaTuring comprises (1) a novel vision-AI pipeline; and (2) a manually labeled human colon organoid image dataset. I have made the deep learning model, inference procedures, image dataset publicly available and a detailed manual for easy adoption.

3.2 Introduction

The COVID-19 pandemic created a global health crisis and further incentivized the development of intelligent medical devices[50]. Devising smart medical devices and enabling them to offer real-time insights would allow early diagnosis and expert interventions. Investigating organoids could facilitate the design of real-time disease-specific smart devices in this context. We present OrgaTuring, a novel deep-learning approach for the automatic detection and classification of organoids. The CNN-based interpretable deep-learning model enables the real-time location, quantification, tracking, and classification of organoids from 2D and 3D images. This research will serve as a steppingstone to creating smart point-of-care devices equipped with mobile healthcare.

Traditional in vitro cultures use primary or immortalized cell lines placed on 2D surfaces. Owing to their 2D nature, these cultures fail to mimic the complex physiological environment of their corresponding tissues and, thus, cannot predict the in-vivo behaviors[51]. These pitfalls have propelled the recent emergence of Organoids, which are multicellular spheroids grown in a 3D culture. Organoids, essentially miniature, self-assembled, and self-replicating versions of tissues, are cultivated from stem cells that are extracted from either normal or pathologically altered samples such as tumor excisions or needle biopsies[52]. The unique capability of organoid technology lies in its ability to encourage the growth of cells that traditionally resist proliferation in vitro while preserving characteristics to in vivo conditions, including complex structural organization, tissue-specific functionalities, and the representation of disease-associated phenotypes. In other words, the organoids recapitulate their parent organ's processes and cellular composition in many regards. Examples include organoids from the gut, pancreas, liver, and many others, which have been

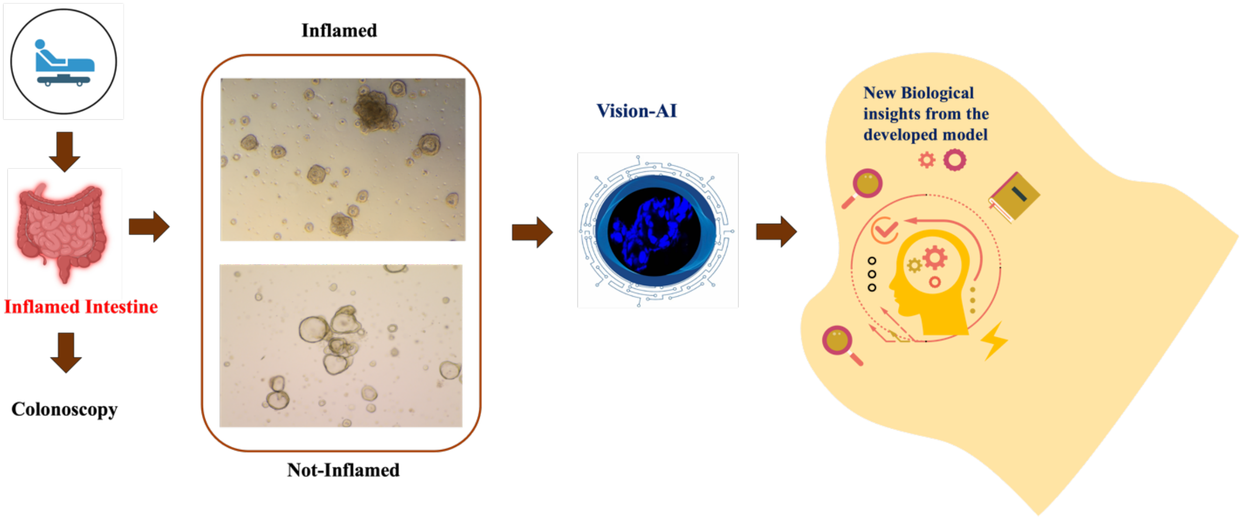


Figure 3.1: Vision-AI importance to observe the morphology of organoids.

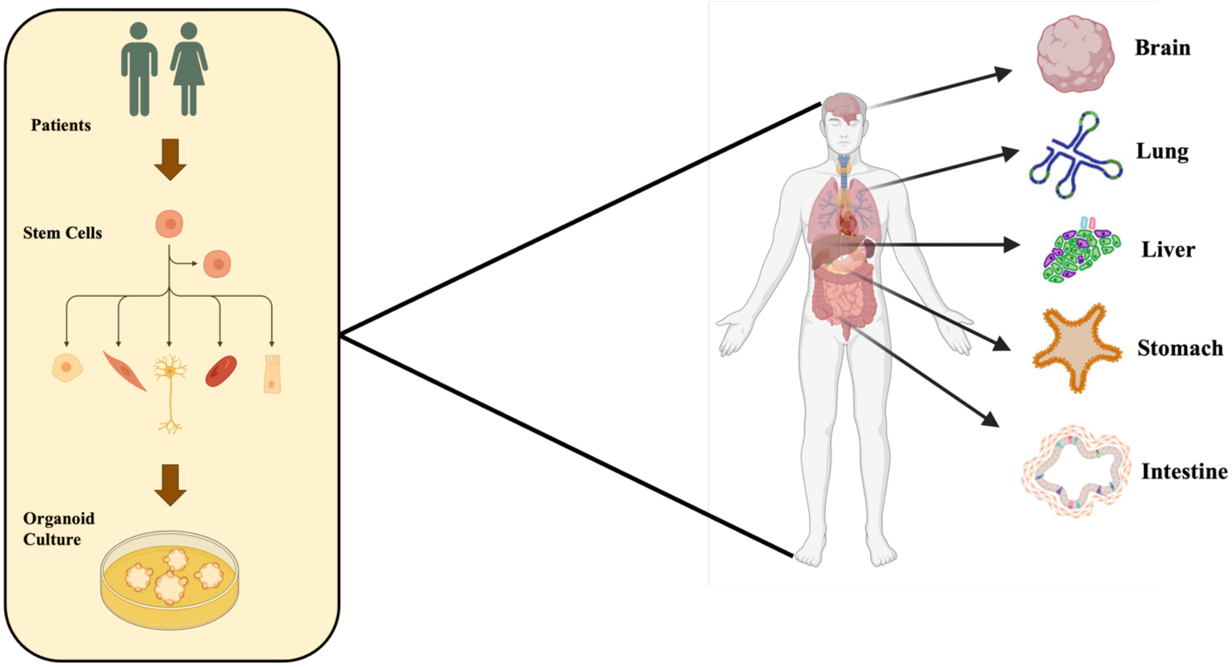


Figure 3.2: The schematic representation of organoids derived from different organs.

widely accepted. This has made organoids a go-to model approach for physiological investigation and drug discovery. Fig 3.1 shows the necessity of using vision-AI in organoid morphological analysis.

While organoids are very promising models for studying human processes and structures, it requires a tedious job to culture an organoid model that accurately mimics the target organ's in vivo functionality and cellular composition as shown in Fig 3.2. To deduce features of an organoid requires continuous monitoring of its growth and ensuring they receive the necessary growth factors and structural support. To this end, morphological (such as their shape, size, spectrum, quantity, and growth rate) understanding of an organoid is paramount in research. At present, the standard protocol to culture organoids is as follows. Single cells or tissue fragments from primary sources are embedded within a three-dimensional matrix derived from Engelbreth-Holm-Swarm (EHS) murine sarcoma[53]. This forms a gel dome on conventional tissue-culture-treated plastic surfaces. Once set, this gel dome is covered with a specially formulated medium composed of small molecules, recombinant proteins, and possibly other supplements specific to tissue type and disease condition. These cells or fragments proliferate throughout the culture period and autonomously arrange themselves into three-dimensional structures. Organoids can be propagated and expanded by removal of the ECM followed by enzymatic and mechanical dissociation. The dissociated organoids are then returned to 3-D culture conditions to continue expanding and re-develop into organoids. For monitoring and investigation, the gel dome is imaged in brightfield. However, these images encounter several imaging pitfalls, which make morphological investigations of organoids very challenging. The pitfalls include large variations in size and shape, overlapping organoids, out-of-focus spheroids, sparsity in organoid distribution, and bad lighting conditions. Each of these images contains hundreds of organoids. Therefore, the underlying pitfalls make the manual investigation of organoids very challenging. To this end, limited computer vision approaches have been proposed, primarily for counting the number of organoids.

3.3 Previous Work

The research community has recently witnessed several discoveries in generating tissues in-vitro from stem cells. Organoids are miniature 3D cultures grown in-vitro that resemble organs. Scientists have successfully cultivated various in-vitro organoid models using cells derived from patients, which can imitate the source organ's physiology. Organoids are being demonstrated as a powerful tool due to their immense potential to promote research in fields such as regenerative medicine, personalized treatment, drug discovery, organ replacement, genetic disorder pathology, etc. As such, organoids can help further research related to tissue morphogenesis, toxicity screening, drug testing, regenerative medicine, and disease modeling and help facilitate a better understanding of the development and physiology of organs.

Live-cell imaging of organoids enables us to study and track growth, apoptosis/necrosis, and movement within the medium. Since the organoids must be monitored rigorously using rapid imaging, it is not feasible to visually interpret and verify the data manually. ML-based algorithms can be adapted into bioimaging pipelines to aid in the real-time processing of organoid image data. Organoids display complex phenotypes and can be difficult to describe using standard features only. Furthermore, due to the culturing medium and the thickness of the samples, the images undergo/suffer from various distortions/imaging artifacts, and thus standard bioimaging tools/pipelines need to be modified to identify organoids in images.

In 2017 Piccinini et al. presented ACC (Advanced Cell Classifier), a user-friendly graphical software package/tool that implements image analysis methods and machine learning algorithms to aid in the mining and exploration of microscopic single-cell image data[54]. ACCv2.0, the current version of ACC, implements methods to analyze and visualize cell data and includes algorithms to find new or rare phenotypes. The latest version of CellProfiler (McQuin et al. (2018)) can identify and quantify biological objects and their morphological features from 2D and 3D images[55]. OrganoSeg, developed by Borten et al. (2018), is an open-source computer vision approach that enables the segmentation, quantification, and filtering of brightfield phenotypes[56]. However, it uses conventional image processing techniques and requires user intervention in the form of parameter tweaking/tuning. In 2019, Kassis et al. presented OrgaQuant, a trained deep CNN that enables the localization and detection of human intestinal organoids in brightfield images[57]. It requires no user intervention, automatically localizes individual organoids and labels them (using a 'bounding box') and can also be used as a clustering tool. OrgaQuant's ability to identify and annotate can be considered at par with that of humans, but it is significantly faster than humans.

In 2021 Gritti et al. developed MorgAna (Machine Learning based Organoid Analysis), a flexible python-based tool that requires minimal coding experience and offers visualization, quantification, and segmentation for 2D images of organoids via a GUI[58]. When focusing on the run time, OrganoSeg and MOrgAna can be considered at par, while CellProfiler takes about twice as long. However, MOrgAna outperforms OrganoSeg and CellProfiler when compared based on accuracy and precision. MOrgAna can also handle bent organoids due to its 'straightening' algorithm that extracts the midline of the organoid and recomputes all the morphological parameters, adding an eccentricity value and the lengths of the major and minor axes as new parameters. However, MOrgAna can only process two-dimensional images. While OrgaQuant and MOrgAna are elegant ML-based bioimaging approaches, they do not offer classification and could overlook potentially valuable morphological information. Visiopharm is a commercial solution based on AI-driven image analysis and tissue mining tools. It has been used on fluorescent cerebral organoid images to count cell[59].

3.4 Material and Methods

3.4.1 Dataset Generation from human subjects

Human Subjects: For generating healthy and CD patient-derived organoids (PDOs), patients were enrolled for colonoscopy as part of routine care for the management of their disease from the University of California, San Diego IBD-Center, following a research protocol compliant with the Human Research Protection Program (HRPP) and approved by the Institutional Review Board (Project ID# 1132632: PI Boland and Sandborn). Histologically normal healthy colon samples were collected from patients presenting for screening colonoscopy or undergoing the procedure for making the diagnosis of irritable bowel syndrome as shown in Fig 3.3. Each participant provided a signed informed consent to allow for the collection of colonic tissue biopsies for research purposes to generate 3D organoids. Isolation and biobanking of organoids from these colonic biopsies were carried out using an approved IRB (Project ID # 190105: PI Ghosh and Das) that covers human subject research at the UC San Diego HUMANOID Center of Research Excellence (CoRE). For all the deidentified human subjects, information, including age, gender, and previous history of the disease, was collected from the chart following the rules of HIPAA. The study design and the use of human study participants were conducted in accordance with the criteria set by the Declaration of Helsinki.

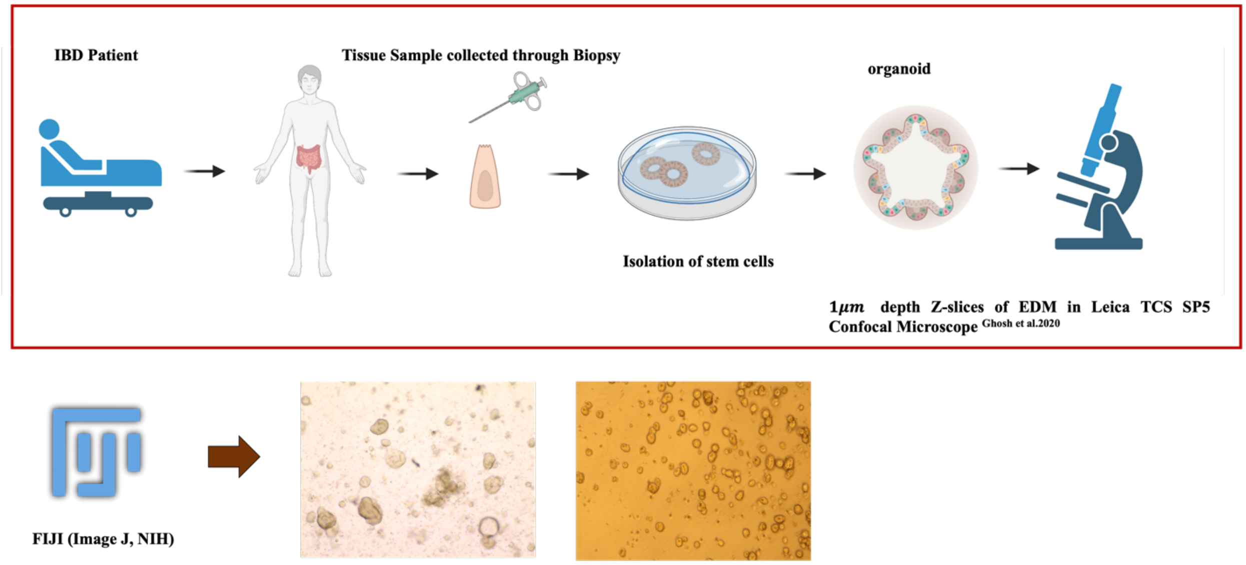

IBD Patient
Tissue Sample collected through Biopsy
organoid
Isolation of stem cells
1μm depth Z-slices of EDM in Leica TCS SP5 Confocal Microscope Ghosh et al.2020
FIJI (Image J, NIH)

Figure 3.3: The schematic representation of organoid images generation.

Isolation of Enteroids from colonic specimens of healthy and Crohn's Disease subjects: Intestinal crypts, comprised of crypt-base columnar (CBC) cells, were isolated from human colonic tissue specimens using the previously published paper (Ghosh et al., 2020; Sahoo et al., 2021; Sayed et al., 2020c; Sayed et al., 2021). In brief, intestinal crypts were dissociated from tissues by digesting with collagenase type I (2 mg/mL solution containing gentamicin 50 µg/mL). The plate was incubated in a CO2 incubator at 37°C, mixing every 10 min with vigorous pipetting in-between incubations while monitoring the release of single epithelial units from tissue structures by light microscopy. To inactivate collagenase, wash media (DMEM/F12 with HEPES, 10% FBS) was added to cells, filtered through a 70 µm cell strainer, centrifuged at 200 g for 5 min, and then the supernatant was aspirated, leaving behind a cell pellet. The number of viable intestinal stem cells was determined by the Trypan Blue Exclusion method using Countess II Automated Cell Counter. Epithelial units were resuspended in Matrigel, and 25 µl of cell-matrigel suspension was added to the wells of a 12-well plate on ice and incubated upside-down in a 37°C CO2 incubator for 10 min, which allowed for polymerization of the Matrigel. After 10 min of incubation, 1000 µL of 50% conditioned media, prepared from L-WRN cells with Wnt3a, R-Spondin and Noggin, ATCC® CRL-3276™ (Miyoshi & Stappenbeck, 2013) with a GI-organoid media cocktail 1 (purchased from HUMANOID CoRE), 10 µM Y27632 and 10 µM SB431542. The medium was changed every two days, and the enteroids were either expanded or frozen in liquid nitrogen for biobanking.

Quantitative assessment of organoid morphology by Imaris: LIF files were first converted into native IMARIS format (.ims). Then a spot filter and surface filter were created. This filter is used as a batch function on all processed images. Finally, a cell object is created where broken fragments of single organoids are stitched together manually. Upon manual completion, specific measurements are exported from IMARIS to GraphPad Prism for further analysis and for visualization as graphs.

Embedding of organoids in HistoGel™: Healthy and CD colonic organoids were embedded in histogel as done previously (Tindle et al., 2021). Briefly, mature organoids after 7 days of culture in 6-Well plates were fixed in 4% PFA at room temperature for 30 min and quenched with 30 mM glycine for 5 min. After washing with PBS, organoids were resuspended in PBS and stained using Gill's hematoxylin for 5 min for ease during embedding in paraffin blocks and visualization during and after sections. Excess hematoxylin was removed, and organoids were resuspended in HistoGel™ and centrifuged at 65°C for 5 min. HistoGel™ embedded organoid pellets were cooled to room temperature and stored in 70% ethanol at 4°C until ready for embedding in paraffin blocks. FFP-embedded organoid sections were cut at a setting of 4 µm thickness and fixed onto microscope slides for H&E staining. Immunofluorescence of FFPE organoids: Sections of FFP-embedded healthy- and CD- PDOs were deparaffinized, rehydrated, and underwent antigen retrieval immersed in Sodium Citrate buffer (pH 6.0) and boiled at 100°C inside a pressure cooker for 3 min. Once sections returned to room temperature, samples were washed in DI water and then permeabilized and blocked for two h using an in-house blocking buffer (2 mg/mL BSA and 0.1% Triton X-100 in PBS), as described previously (Lopez-Sanchez et al., 2014; Tindle et al., 2021). Primary antibodies [see Key Resource Table] were diluted in a blocking buffer and incubated overnight at 4°C. Secondary antibodies were diluted in a blocking buffer and allowed to incubate for two h in the dark. Antibody dilutions are listed in the Supplementary Key Resource Table. ProLong Glass was used as a mounting medium. Coverslips (No.1 thickness) were applied to slides to seal and stored at 4°C until imaged.

Estimation of Paneth: Goblet cell ratio by confocal imaging of cell markers: Fluorescent Z-stack images of lysozyme (a bona-fide marker of Paneth cells) and muc2 (a bona-fide marker of goblet cell) stained organoids were acquired by successive 1 μm depth Z-slices of EDMs in the desired confocal channels of Leica TCS SP5 Confocal Microscope as done previously (Ghosh et al., 2020)[60]. Fields of view that were representative of a given transwell were determined by randomly imaging three different fields. Z-slices of a Z-stack were overlaid to create maximum-intensity projection images; all images were processed using FIJI (Image J) software[61]. All images were processed on ImageJ software (NIH) and assembled into figure panels using Photoshop and Illustrator (Adobe Creative Cloud).

3.4.2 Dataset Preparation for AI-pipeline

This research involved two cohorts of image data from colon organoids—each labeled as two classes having Inflammation and No Inflammation. Cohort 1 comprises 49 Inflamed images and 359 Not Inflamed Images, while Cohort 2 comprises 66 Inflamed and 334 Not Inflamed images. All the images have two types of dimensions, 2048*1536 and 3888*2592, and different color stains. This labeled dataset of Crohn's disease colon organoids is first-of-its-kind. All images were color normalized for organoid counting using Reinhard normalization and then passed through Roboflow's open-source computer vision platform for further pre-processing. Two hundred seventy-nine images were manually annotated for the object detection task using the Roboflow annotation tool.

Further, the images were split into training, testing, and validation set, with the training set oversampled with adding augmented images. Image augmentation, like horizontal and vertical flips, rotation, brightness, etc., was performed to increase the training set. Finally, the training set, validation set, and testing set involved 388, 54, and 31 images, respectively. For the organoid classification task, the images were divided into three different types of groups. First, a simple stratified sampling was performed on each cohort and a combination of cohort 1 and cohort 2 to split the dataset. Second, the images were grouped based on the different zoom sizes available and then followed by stratified sampling on each cohort and a combination of cohort 1 and cohort 2. Third, the images were grouped based on the same type of images, followed by the same stratification process stated above. Fig 3.4, Fig 3.5, Fig 3.6 shows the dataset preparation process.

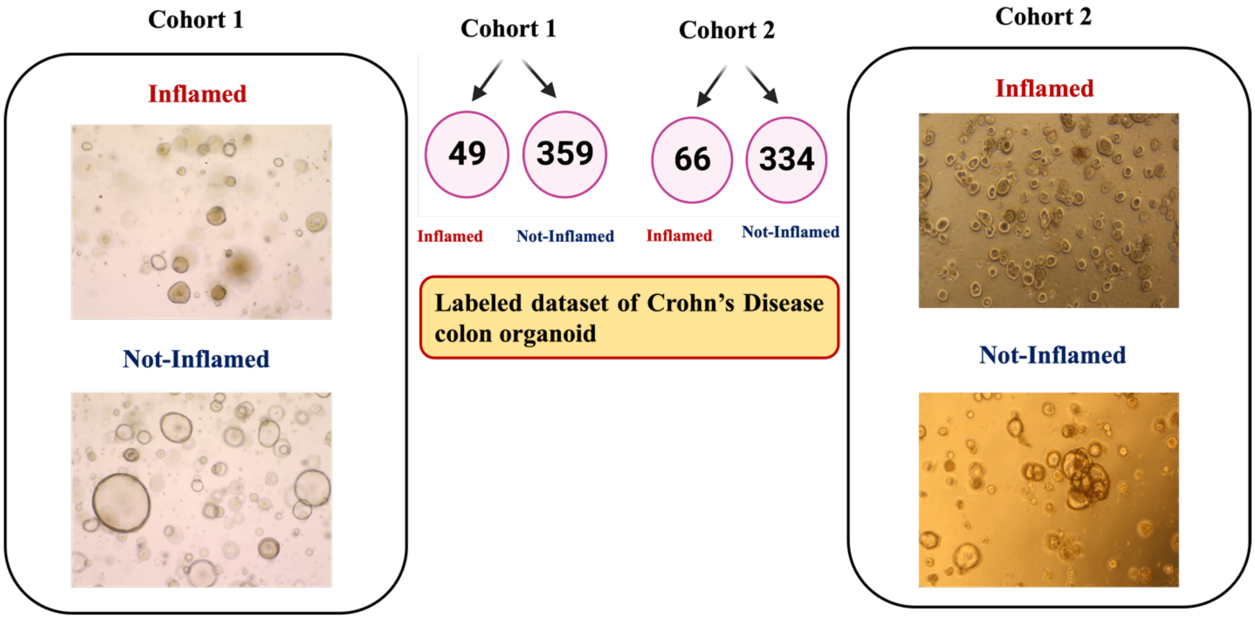


Figure 3.4: Labeled dataset of Crohn's disease colon organoid with each cohort distribution.

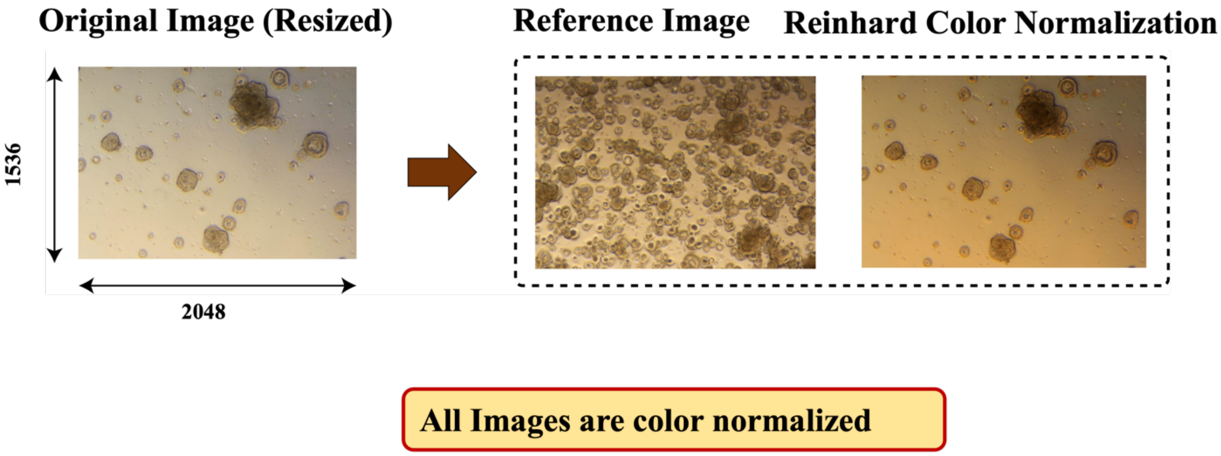


Figure 3.5: Data normalization process using Reinhard color normalization.

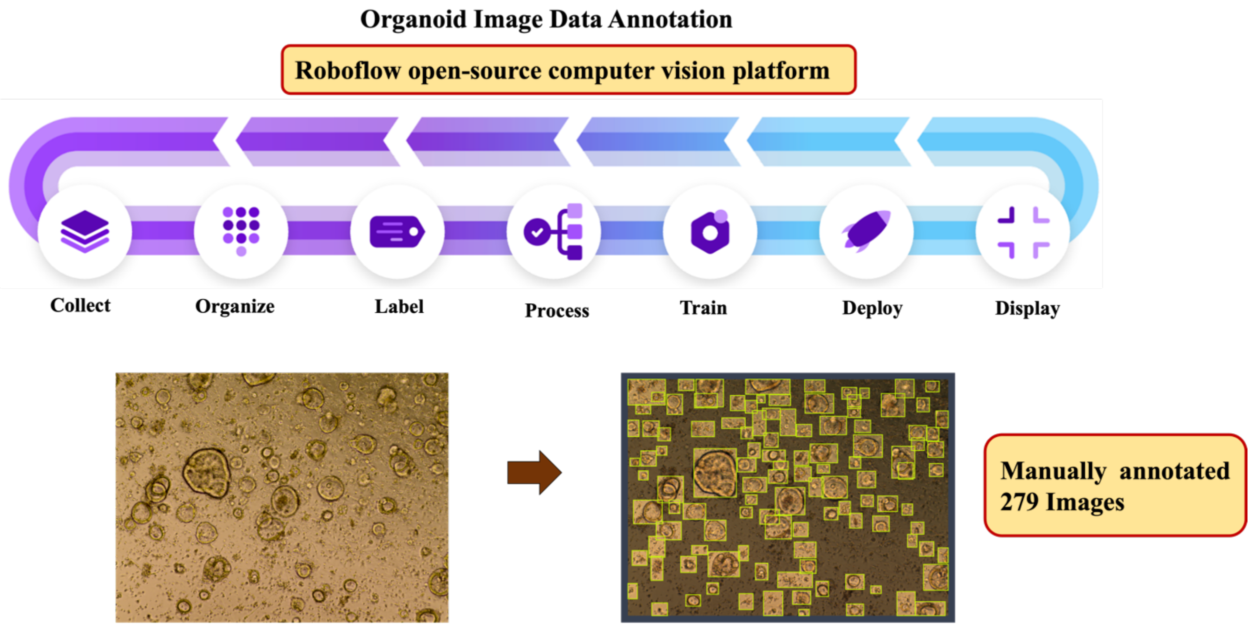

Organoid Image Data Annotation
Roboflow open-source computer vision platform
Collect
Organize
Label
Process
Train
Deploy
Display
Manually annotated 279 Images

Figure 3.6: Data annotation process using Roboflow open-source tool.

3.5 Result and Discussion

3.5.1 OrgaTuring Counting

Roboflow open-source tool was used for the organoid counting task. After generating the dataset, the images were exported in the Yolo v5 format for the model training. Yolo is a state-of-the-art object detection and segmentation architecture. This single-shot detector performs the task of object localization and classification in a single forward pass of the network. Hence, they are faster and simpler models and don't require a separate object proposal stage as in maskRCNN. They achieve this by dividing the image into multiple grids and predicting the multiple bounding boxes and probabilities for each grid cell. The original Yolo v5 model was trained on 80 classes; I created a custom Yolo v5m model for the organoid detection task. The model was trained using Tesla T4 GPU with image size resizing to 1024*1024. The batch size was kept to 16, with the

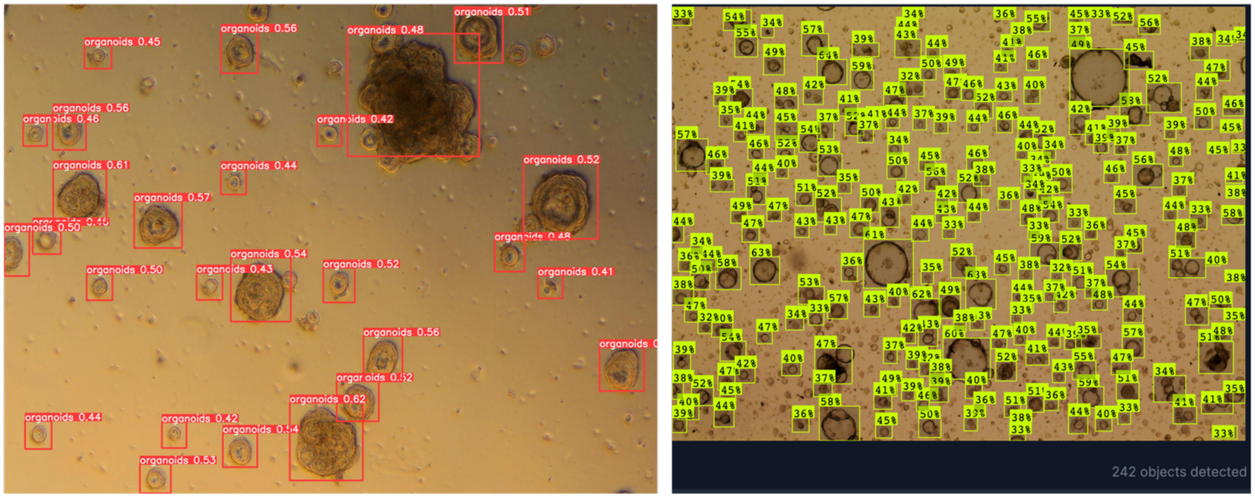


Figure 3.7: Organoid detection and counting using Yolo v5m architecture.

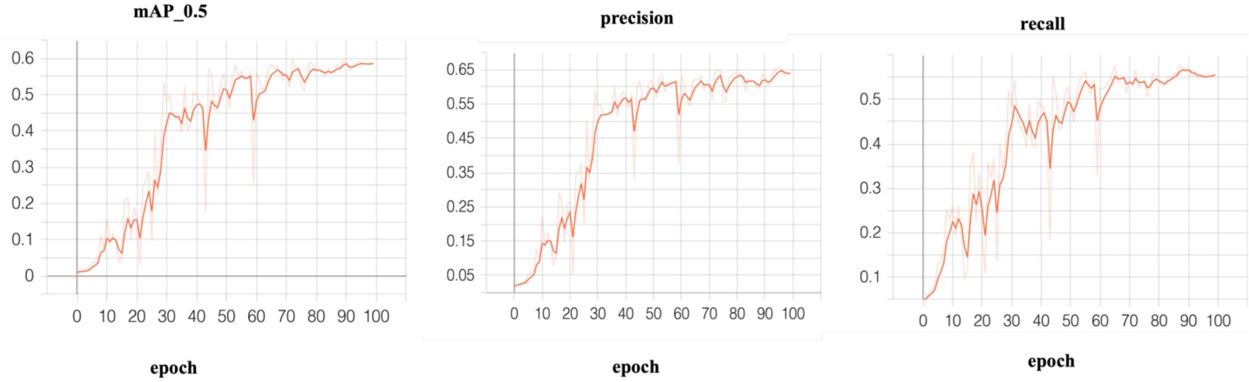


Figure 3.8: Performance metrices for the organoid object detection.

epoch of 100 for one training and 270 for the second training, a learning rate of 0.001, and a weight decay of 0.0005. Fig 3.7, 3.8, 3.9, 3.10 shows the result of the organoid counting, and the model is successfully counting and printing the number of organoids detected in an image. Average precision is used in evaluating the model performance. It represents the area under the curve, higher the curve represents larger areas, and higher average precision. mAP is the average of all the average precision values. It is calculated by fixing the confidence threshold score to 0.5.

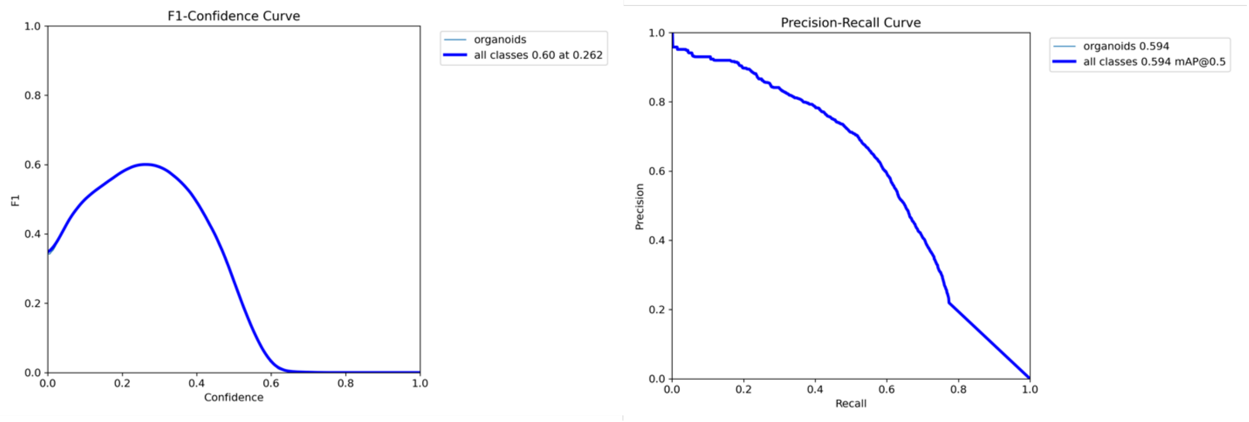


Figure 3.9: F1 confidence score and precision-recall curve for organoid detection.

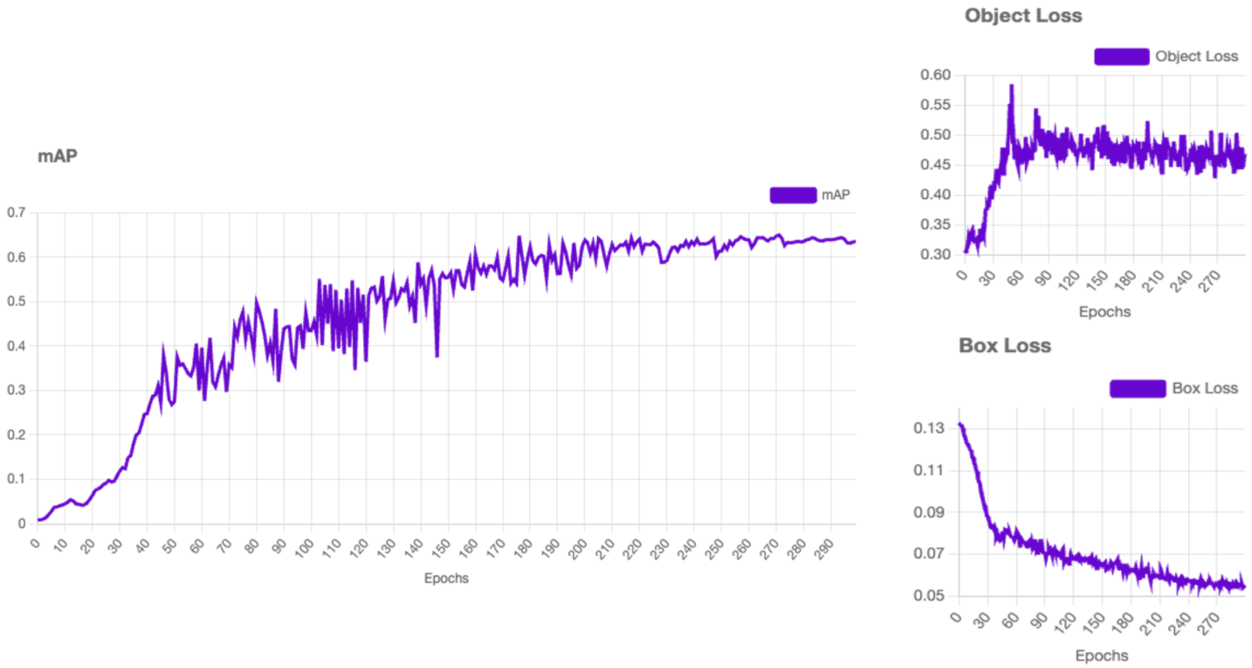


Figure 3.10: Mean average precision curve, object, and box loss for organoid detection.

3.5.2 OrgaTuring Classification

Transfer learning on Medical Dataset is quite a challenging task. With the limited amount of data for medical image classification task, it is difficult to create a new model architecture which is expensive and time consuming. Using ImageNet pretrained architecture has shown expert level classification in previous research in the field of medical image classification. More than 70% of the time involved in this research was understanding the data and pre-processing. With ample amount of research, pretrained architecture like ResNet152v and DenseNet201 were selected for the above process. VGG-16 model was used for the baseline to check for the performance with the datasets. The class in each cohort was highly imbalanced. Class imbalance is a significant issue in classification problems. There are several methods to handle class imbalance, like class weighting, oversampling, under-sampling, data augmentation, transfer learning, etc. We have performed class weighting, data augmentation, and transfer learning techniques to deal with lesser data. When the dataset is imbalanced, the model learns less about the minority class since the model gets fewer data points for the same and does not learn a good representation of the minority class. Hence results in an inaccurate classification model on such datasets. The generated organoid image dataset has a sample ratio of 1:7. This means forcing our algorithm to treat every instance of class 1 as seven instances of class 0. So here, every instance of class Inflamed is seven instances of class Not Inflamed, which means a higher value was assigned to these instances in the loss function.

This task involves the binary classification between Inflamed and Not Inflamed types of images. As the class was highly imbalanced, the normal Binary cross entropy loss would not have provided dedicated results. So, I have used Binary focal cross entropy loss by setting the alpha and gamma values for the best results. Alpha and gamma are hyperparameters that controls the shape of the loss and was adjusted setting different training parameters. Alpha is a balancing factor that handles the class imbalance. If the classes are imbalanced, alpha set to a value less than 0.5 for the majority class and a value greater than 0.5 for the minority class. Gamms is moreover a focusing parameter, that reduces the loss contribution.

ResNet152v and DenseNet201 pretrained ImageNet architectures are used as said previously for the classification task. Then transfer learning is implemented, by setting the ‘include_top’ argument False, final classification layer of each of the model was not included, making the base model suitable for feature extraction. Then weights of the base model were frozen, to keep the learned features from the pre-trained model, and only train a few final layers. The final layer or top layer took the input of the output tensor of the pre-trained backbone and flattened the output. Then new dense layer with 512 units and ReLU activation function were added with dropout layer, to prevent overfitting. A final dense layer was added with a single unit and a sigmoid activation function for binary classification. Further, few of the top layers of the pre-trained model was unfrozen for fine-tuning and trained jointly with the newly added layers. Fig 3.11 and 3.12 shows the model architecture used for this task.

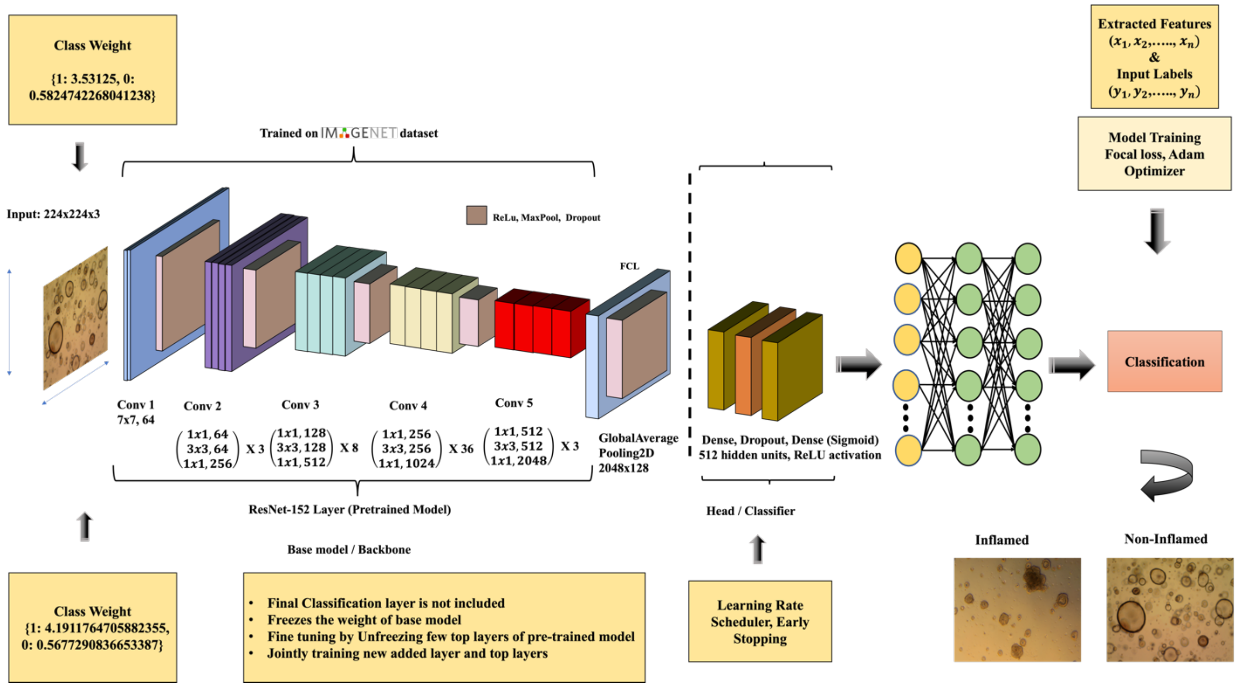

Class Weight
{1: 3.53125, 0: 0.5824742268041238}
Trained on IMAGENET dataset
Input: 224x224x3
ReLu, MaxPool, Dropout
FCL
Conv 1
7x7, 64
Conv 2
Conv 3
Conv 4
Conv 5
1x1, 64 3x3, 64 1x1, 256 X 3
1x1, 128 3x3, 128 1x1, 512 X 8
1x1, 256 3x3, 256 1x1, 1024 X 36
1x1, 512 3x3, 512 1x1, 2048 X 3
GlobalAverage Pooling2D 2048x128
ResNet-152 Layer (Pretrained Model)
Base model / Backbone
Class Weight
{1: 4.1911764705882355, 0: 0.5677290836653387}
Final Classification layer is not included
Freezes the weight of base model
Fine tuning by Unfreezing few top layers of pre-trained model
Jointly training new added layer and top layers
Dense, Dropout, Dense (Sigmoid)
512 hidden units, ReLU activation
Head / Classifier
Learning Rate Scheduler, Early Stopping
Extracted Features
$(x_1, x_2, \ldots, x_n)$
&
Input Labels
$(y_1, y_2, \ldots, y_n)$
Model Training
Focal loss, Adam Optimizer
Classification
Inflamed
Non-Inflamed

Figure 3.11: Organoid classification using ResNet152v pretrained architecture.

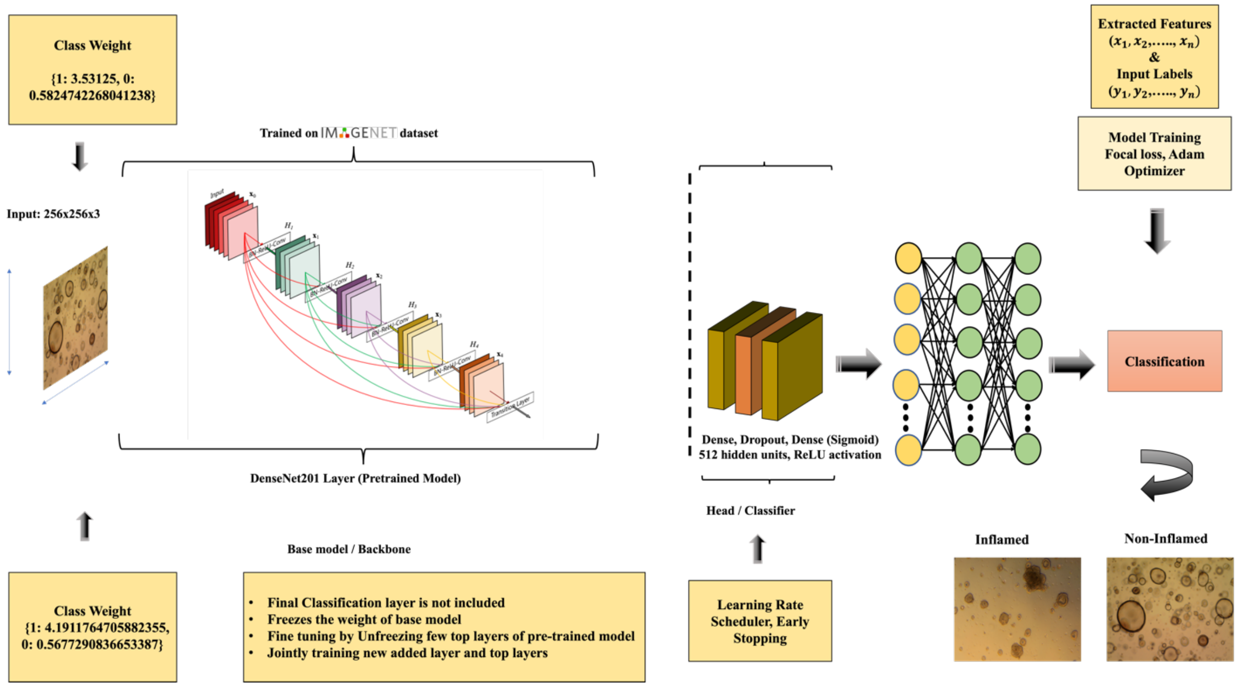

Class Weight
{1: 3.53125, 0: 0.5824742268041238}
Trained on IMAGENET dataset
Input: 256x256x3
Input
BN-ReLU-Conv
Transition Layer
DenseNet201 Layer (Pretrained Model)
Base model / Backbone
Class Weight
{1: 4.1911764705882355, 0: 0.5677290836653387}
Final Classification layer is not included
Freezes the weight of base model
Fine tuning by Unfreezing few top layers of pre-trained model
Jointly training new added layer and top layers
Dense, Dropout, Dense (Sigmoid)
512 hidden units, ReLU activation
Head / Classifier
Learning Rate Scheduler, Early Stopping
Extracted Features
$(x_1, x_2, \ldots, x_n)$
&
Input Labels
$(y_1, y_2, \ldots, y_n)$
Model Training
Focal loss, Adam Optimizer
Classification
Inflamed
Non-Inflamed

Figure 3.12: Organoid classification using DenseNet201 pretrained architecture.

Results (Training only on Cohort 2)

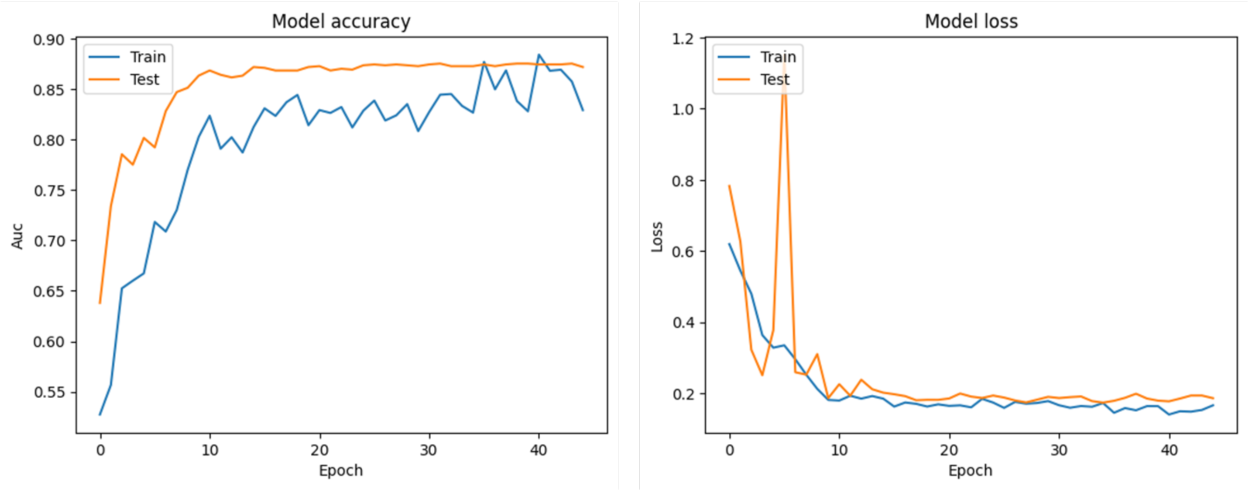
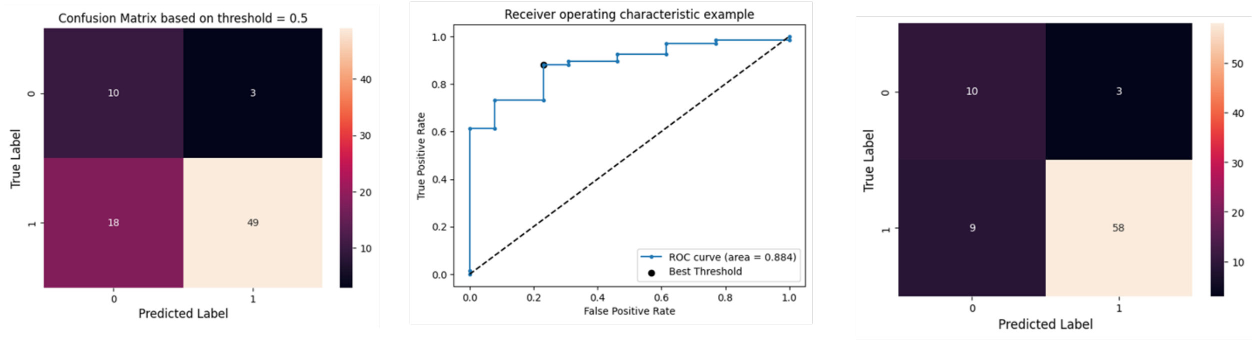


Figure 3.13: Organoid classification model accuracy and loss curve on cohort 2.

Results (Training only on Cohort 2)



0 → Inflamed
1 → Not-Inflamed

Optimal Threshold calculated using G-Mean

Figure 3.14: Organoid Classification confusion matrix on cohort 2.

Results (Training only on Cohort 1)

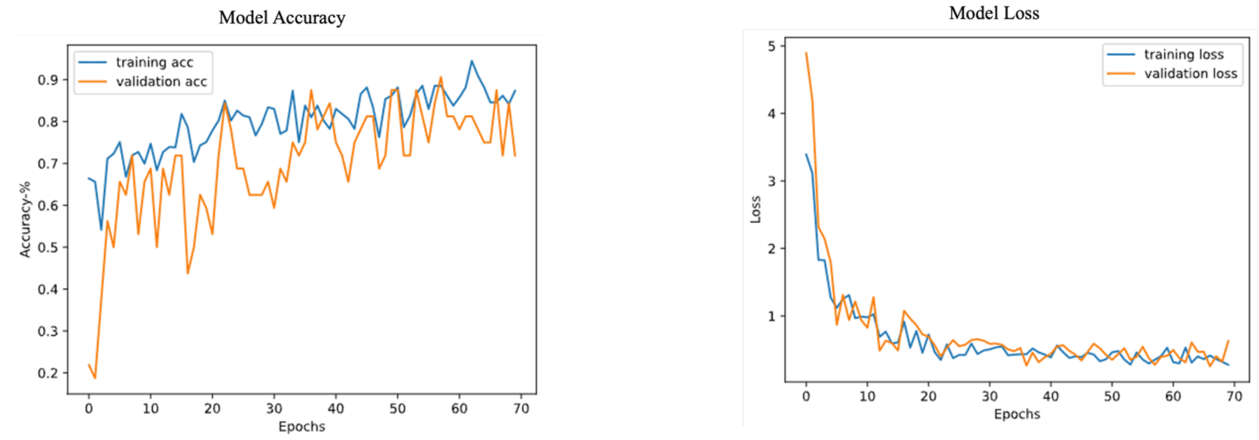
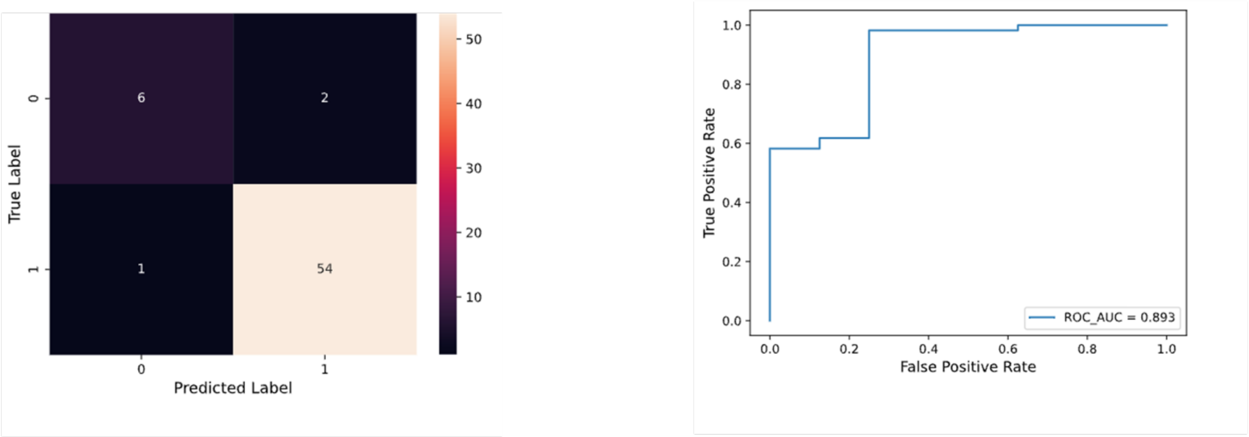


Figure 3.15: Organoid Classification model accuracy and loss curve on cohort 1.

Results (Training only on Cohort 1)



0 → Inflamed
1 → Not-Inflamed

Figure 3.16: Organoid Classification confusion matrix on cohort 1.

Results (Cohort1 + Cohort2)

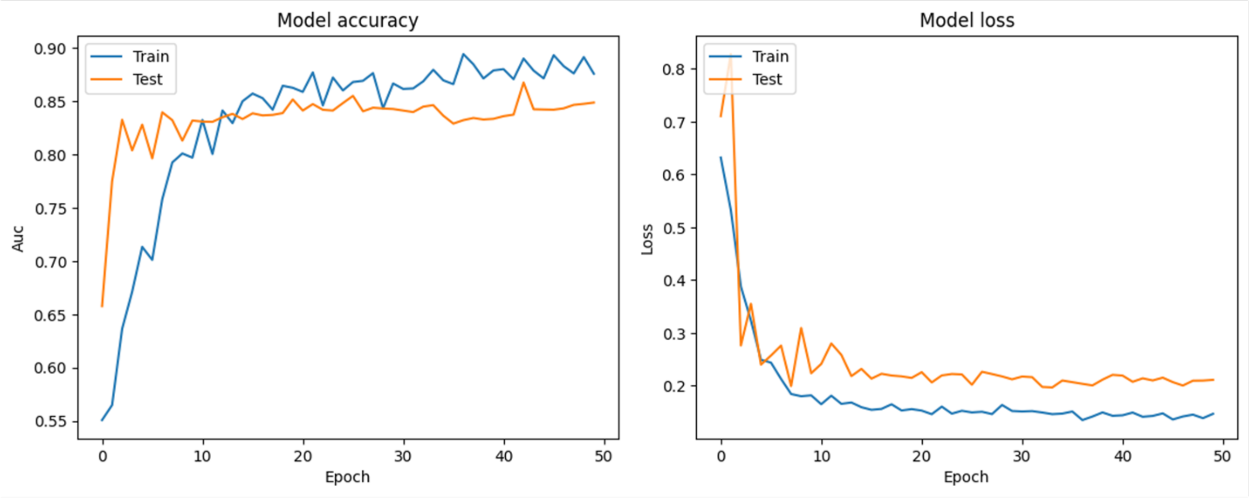
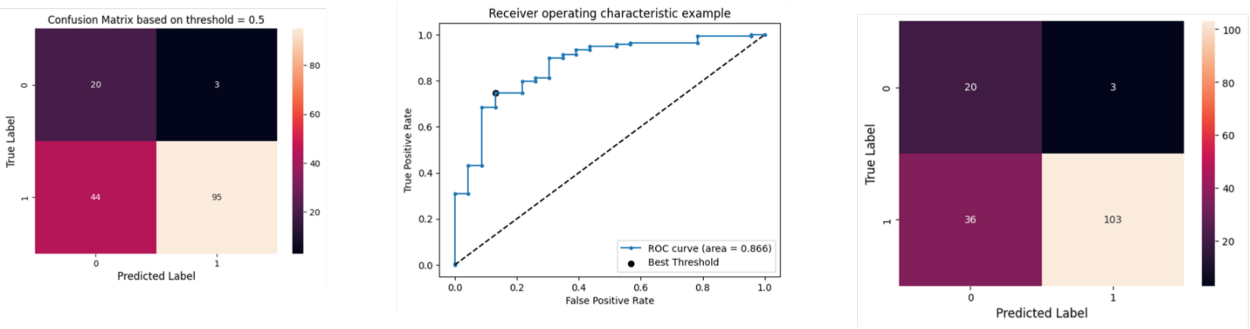


Figure 3.17: Organoid Classification model accuracy and loss curve on both cohorts.

Results (Cohort1 + Cohort2)

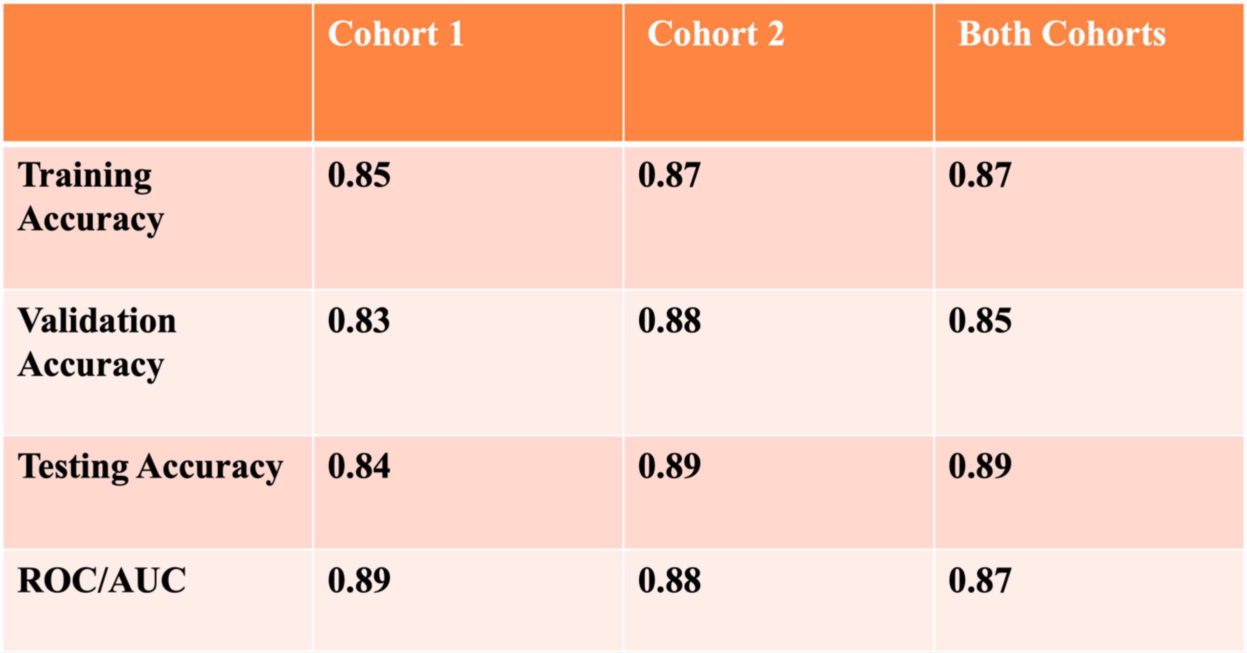


0 → Inflamed
1 → Not-Inflamed

Optimal Threshold calculated using G-Mean

Figure 3.18: Organoid Classification confusion matrix on both cohorts.

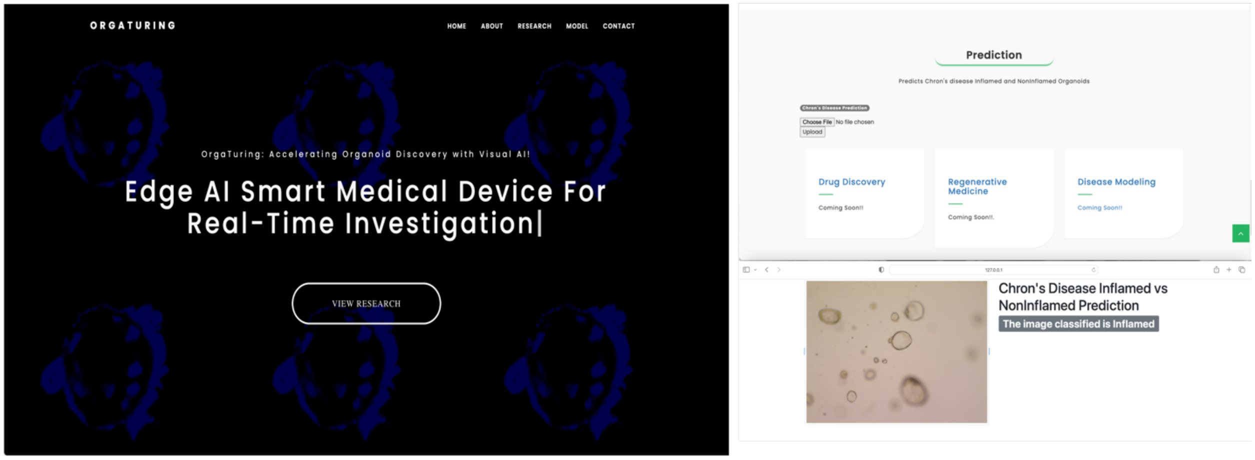
Table 3.1: Evaluation metrics for all the methods.



	Cohort 1	Cohort 2	Both Cohorts
Training Accuracy	**0.85**	**0.87**	**0.87**
Validation Accuracy	**0.83**	**0.88**	**0.85**
Testing Accuracy	**0.84**	**0.89**	**0.89**
ROC/AUC	**0.89**	**0.88**	**0.87**

3.5.3 OrgaTuring API

API for OrgaTuring



OrgaTuring: Accelerating Organoid Discovery with Visual AI!

Edge AI Smart Medical Device For Real-Time Investigation|

Prediction

Predicts Chron's disease Inflamed and NonInflamed Organoids

Chron's Disease Prediction

Choose File No file chosen

Upload

Drug Discovery

Coming Soon!!

Regenerative Medicine

Coming Soon!!.

Disease Modeling

Coming Soon!!

127.0.0.1

Chron's Disease Inflamed vs NonInflamed Prediction

The image classified is Inflamed

Figure 3.19: RESTful web API for OrgaTuring.

OrgaTuring provides a python flask based RESTful web application program interface (API) to upload organoid image data and obtain the prediction in real-time. It is a static webpage to upload images and take the inference quickly with no user intervention and parameter tweaking. The simple, and open-source application allows medical professionals to easily adopt and use the interface. Fig 3.19 shows the designed API landing page.

3.6 Conclusion and Future work

In this research a novel first-of-its-kind AI tool chain to understand organoids morphology is designed and developed. Most of the computational tools and techniques used to study organoids focus only on quantification. Contrary to that, OrgaTuring with its unique AI algorithms goes beyond counting (or quantification). It can also locate, track, and classify organoids w.r.t a variety of meaningful phenotypes. This makes OrgaTuring an all-encompassing AI guided toolbox for rapid organoid discovery. Further, OrgaTuring's real-time nature is a boon to system biologists to predict outcomes in milliseconds without relying on too many expert interventions. Our model bears enough potential to classify thousands of images obtained from different imaging techniques, parameters, and cohorts in real-time. In addition, the deep domain adaptation techniques have been implemented to leverage the classification tasks of cohort divided by molecular and clinical subtypes which has been ignored by the recent literature. Generating organoid images is expensive, tedious, and time-consuming task, which involves lot of resources, and labor. Developing a deep learning model which uses domain adaptation and learn label classification simultaneously is itself a novel task. Through this process the model can predict the unlabeled target data using the source and target data in an adversarial training process. This method can make the current research in organoid domain less expensive and accelerate organoid research towards drug discovery, regenerative medicine, organ replacement, etc.

Acknowledgements

Chapter 3, in full, has been submitted for publication of the material as it may appear in PNAS, 2023, Dang, Dharanidhar*; Nanda, Amitash*; Carpinelli, Sophia; Tindle, Courtney; Sinha, Saptarshi; Ghosh, Pradipta; Sahoo, Debashis. The thesis author was the primary investigator and author of this material.

REFERENCES

1. Topol E. (2016). *The Patient Will See You Now: The Future of Medicine is in Your Hands*. Basic Books.

2. *Systems Biology as Defined by NIH | NIH Intramural Research Program*. (n.d.). https://irp.nih.gov/catalyst/19/6/systems-biology-as-defined-by-nih#:~:text=Systems%20biology%20is%20an%20approach,involves%20taking%20the%20pieces%20apart.

3. Institute for Systems Biology. (2019, December 12). *What Is Systems Biology · Institute for Systems Biology*. https://isbscience.org/about/what-is-systems-biology/.

4. Kitano, H. (2002). Computational systems biology. *Nature*, *420*(6912), 206–210. https://doi.org/10.1038/nature01254.

5. Priami C, Regev A, Shapiro E, Silverman W. Application of a stochastic name-passing calculus to representation and simulation of molecular processes. Information Processing Letters. 2001;80(1):25–31. http://scholar.google.com/scholar?q=Application+of+a+stochastic+name-passing+calculus+to+representation+and+simulation+of+molecular+processes+Priami+2001.

6. Sahoo, D., Dill, D. L., Gentles, A. J., Tibshirani, R., & Plevritis, S. K. (2008). Boolean implication networks derived from large scale, whole genome microarray datasets. *GenomeBiology.com (London. Print)*, *9*(10), R157. https://doi.org/10.1186/gb-2008-9-10-r157.

7. Fischer, S. (2019). An Introduction to Image-Based Systems Biology of Multicellular Spheroids for Experimentalists and Theoreticians. In *Codon Publications eBooks*. https://doi.org/10.15586/computationalbiology.2019.ch1.

8. Smyth, M. J., & Martin, J. (2000). x Ray crystallography. *Journal of Clinical Pathology-molecular Pathology*, *53*(1), 8–14. https://doi.org/10.1136/mp.53.1.8.

9. Milne, J. L. S., Borgnia, M. J., Bartesaghi, A., Tran, E. E. H., Earl, L. A., Schauder, D. M., Lengyel, J. S., Pierson, J., Patwardhan, A., & Subramaniam, S. (2013). Cryo-electron microscopy - a primer for the non-microscopist. *FEBS Journal*, *280*(1), 28–45. https://doi.org/10.1111/febs.12078.

10. Hussain, S., Mubeen, I., Ullah, N., Shah, S. a. A., Khan, B. H., Zahoor, M., Ullah, R., Khan, F. A., & Sultan, M. A. (2022). Modern Diagnostic Imaging Technique Applications and Risk Factors in the Medical Field: A Review. *BioMed Research International*, *2022*, 1–19. https://doi.org/10.1155/2022/5164970.

11. Novel Coronavirus SARS-Cov-2-This scanning electron microscope image shows SARS-CoV-2-also known as 2019-nCoV, the virus that causes COVID-19. Original image sourced from US Government department. The National Institute of Allergy and Infection" by Free Public Domain Illustrations by rawpixel is licensed under CC-BY 2.0.

12. Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., Becker, S., Rox, K., & Hilgenfeld, R. (2020b). Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science*, *368*(6489), 409–412. https://doi.org/10.1126/science.abb3405.

13. Broadwith, P. (2023). Explainer: What is cryo-electron microscopy. *Chemistry World*. https://www.chemistryworld.com/news/explainer-what-is-cryo-electron-microscopy/3008091.article.

14. Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald Summers. (2017), *ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases*, IEEE CVPR, pp. 3462-3471.

15. Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. In *Neural Information Processing Systems* (Vol. 30, pp. 4768–4777). https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

16. Amitashnanda. (n.d.). *GitHub - amitashnanda/Interpretability-in-ChexNet: Implemented CNN-based Deep-Learning model(s) to detect pneumonia from chest X-rays, also Incorporated model interpretability using Sample Handling and Analysis Plan (SHAP). Then used the above metrics to quantify training data based on quality for better model performance and reliability.* GitHub. https://github.com/amitashnanda/Interpretability-in-ChexNet.git.

17. *Confocal Microscopy - Introduction | Olympus LS*. (n.d.). https://www.olympus-lifescience.com/en/microscope-resource/primer/techniques/confocal/confocalintro/.

18. Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., & Yener, B. (2009). Histopathological Image Analysis: A Review. *IEEE Reviews in Biomedical Engineering*, *2*, 147–171. https://doi.org/10.1109/rbme.2009.2034865.

19. *Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN – IARC*. (n.d.). https://www.iarc.who.int/news-events/global-burden-of-colorectal-cancer-in-2020-and-2040-incidence-and-mortality-estimates-from-globocan/.

20. Morgan, E., Arnold, M., Gini, A., Lorenzoni, V., Cabasag, C. J., Laversanne, M., Vignat, J., Ferlay, J., Murphy, N., & Bray, F. (2022). Global burden of colorectal cancer in 2020

and 2040: incidence and mortality estimates from GLOBOCAN. *Gut*, *72*(2), 338–344. https://doi.org/10.1136/gutjnl-2022-327736.

21. Siegel, R. L., Wagle, N. S., Cercek, A., Smith, R. A., & Jemal, A. (2023). Colorectal cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, *73*(3), 233–254. https://doi.org/10.3322/caac.21772.

22. Xi, Y., & Xu, P. (2021). Global colorectal cancer burden in 2020 and projections to 2040. *Translational Oncology*, *14*(10), 101174. https://doi.org/10.1016/j.tranon.2021.101174.

23. Deng, S., Zhang, X., Yan, W., Chang, E. Y., Fan, Y., Lai, M., & Xu, Y. (2020). Deep learning in digital pathology image analysis: a survey. *Frontiers of Medicine*, *14*(4), 470–487. https://doi.org/10.1007/s11684-020-0782-9.

24. Ying, X., & Monticello, T. M. (2006). Modern Imaging Technologies in Toxicologic Pathology: An Overview. *Toxicologic Pathology*, *34*(7), 815–826. https://doi.org/10.1080/01926230600918983.

25. Foran, D. J., Yang, L., Chen, W., Hu, J., Goodell, L., Reiss, M., Wang, F., Kurc, T., Pan, T., Sharma, A., & Saltz, J. H. (2011). ImageMiner: a software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology. *Journal of the American Medical Informatics Association*, *18*(4), 403–415. https://doi.org/10.1136/amiajnl-2011-000170.

26. Xing, F., & Yang, L. (2016). Robust Nucleus/Cell Detection and Segmentation in Digital Pathology and Microscopy Images: A Comprehensive Review. *IEEE Reviews in Biomedical Engineering*, *9*, 234–263. https://doi.org/10.1109/rbme.2016.2515127.

27. Wolberg, W. H., Street, W. N., Heisey, D. M., & Mangasarian, O. L. (1995). Computer-derived nuclear features distinguish malignant from benign breast cytology. *Human Pathology*, *26*(7), 792–796. https://doi.org/10.1016/0046-8177(95)90229-5.

28. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J. S., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056.

29. Janowczyk, A., & Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, *7*(1), 29. https://doi.org/10.4103/2153-3539.186902.

30. Rao, J. N. (2010). *Regulation of Gastrointestinal Mucosal Growth*. NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK54091/.

31. Kainz, P. (2015, November 21). *Semantic Segmentation of Colon Glands with Deep Convolutional Neural Networks and Total Variation Segmentation*. arXiv.org. https://arxiv.org/abs/1511.06919.

32. Granger, D. N. (2011). *Colloquium Series on Integrated Systems Physiology: From Molecule to Function to Disease*. NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK53199/.

33. Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P. S., Rothenberg, M. E., Leyrat, A. A., Sim, S., Okamoto, J., Johnston, D. M., Qian, D., Zabala, M., Bueno, J., Neff, F. N., Wang, J., Shelton, A. A., Visser, B., Hisamori, S., Shimono, Y., Van De Wetering, M., Clevers, H., Clarke, M.F., Quake, S. R. (2011). Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnology*, *29*(12), 1120–1127. https://doi.org/10.1038/nbt.2038.

34. Beck, F. (2004). The role of Cdx genes in the mammalian gut. *Gut*, *53*(10), 1394–1396. https://doi.org/10.1136/gut.2003.038240.

35. Dalerba, P., Sahoo, D., Paik, S., Guo, X., Yothers, G., Song, N., Wilcox-Fogel, N., Forgó, E., Rajendran, P. S., Miranda, S. P., Hisamori, S., Hutchison, J., Kalisky, T., Qian, D., Wolmark, N., Fisher, G. A., Van De Rijn, M., & Clarke, M. (2016b). CDX2 as a Prognostic Biomarker in Stage II and Stage III Colon Cancer. *The New England Journal of Medicine*, *374*(3), 211–222. https://doi.org/10.1056/nejmoa1506597.

36. Nagtegaal, I. D., Odze, R. D., Klimstra, D., Paradis, V., Rugge, M., Schirmacher, P., Washington, K. M., Carneiro, F., & Cree, I. A. (2019). The 2019 WHO classification of tumours of the digestive system. *Histopathology*, *76*(2), 182–188. https://doi.org/10.1111/his.13975.

37. Washington, M. K., Berlin, J., Branton, P., Burgart, L. J., Carter, D. K., Fitzgibbons, P. L., Halling, K., Frankel, W., Jessup, J., Kakar, S., Minsky, B., Nakhleh, R., Compton, C. C., & Members of the Cancer Committee, College of American Pathologists (2009). Protocol for the examination of specimens from patients with primary carcinoma of the colon and rectum. *Archives of pathology & laboratory medicine*, *133*(10), 1539–1551. https://doi.org/10.5858/133.10.1539.

38. Humphries, A., & Wright, N. A. (2008). Colonic crypt organization and tumorigenesis. *Nature reviews. Cancer*, *8*(6), 415–424. https://doi.org/10.1038/nrc2392.

39. Sirinukunwattana, K., Pluim, J. P. W., Chen, H., Qi, X., Heng, P., Guo, Y., Wang, L., Matuszewski, B. J., Bruni, E., Sanchez, U., Böhm, A., Ronneberger, O., Cheikh, B. B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead, D., & Rajpoot, N. M. (2017). Gland segmentation in colon histology images: The glas challenge contest. *Medical Image Analysis*, *35*, 489–502. https://doi.org/10.1016/j.media.2016.08.008.

40. Gibson, P. R., Anderson, R. P., Mariadason, J. M., & Wilson, A. J. (1996). Protective role of the epithelium of the small intestine and colon. *Inflammatory bowel diseases*, *2*(4), 279–302.

41. *Exploring Gene Expression Dataset*. (n.d.). http://hegemon.ucsd.edu/Tools/explore.php?key=colon

42. Sahoo, D., Dill, D. L., Gentles, A. J., Tibshirani, R., & Plevritis, S. K. (2008). Boolean implication networks derived from large scale, whole genome microarray datasets. *GenomeBiology.com (London. Print)*, *9*(10), R157. https://doi.org/10.1186/gb-2008-9-10-r157.

43. Bankhead, P., Loughrey, M. B., Fernández, J. M. G., Dombrowski, Y., McArt, D. G., Dunne, P. D., McQuaid, S., Gray, R. T., Murray, L. J., Coleman, H. G., James, J., Salto-Tellez, M., & Hamilton, P. W. (2017). QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-17204-5.

44. He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). *Mask R-CNN*. https://doi.org/10.1109/iccv.2017.322.

45. He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. https://doi.org/10.1109/cvpr.2016.90.

46. Lin, T. (2016, December 9). *Feature Pyramid Networks for Object Detection*. arXiv.org. https://arxiv.org/abs/1612.03144.

47. Badrinarayanan, V., & Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(12), 2481–2495. https://doi.org/10.1109/tpami.2016.2644615.

48. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1505.04597.

49. Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2016). Pyramid Scene Parsing Network. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1612.01105.

50. Martel, T., & Orgill, D. P. (2020). Medical Device–Related Pressure Injuries During the COVID-19 Pandemic. *Journal of Wound Ostomy and Continence Nursing*, *47*(5), 430–434. https://doi.org/10.1097/won.0000000000000689.

51. Astashkina, A., Mann, B. K., Prestwich, G. D., & Grainger, D. W. (2012). Comparing predictive drug nephrotoxicity biomarkers in kidney 3-D primary organoid culture and immortalized cell lines. *Biomaterials*, *33*(18), 4712–4721. https://doi.org/10.1016/j.biomaterials.2012.03.001.

52. Gilazieva, Z., Ponomarev, A. M., Rutland, C. S., Rizvanov, A. A., & Solovyeva, V. V. (2020). Promising Applications of Tumor Spheroids and Organoids for Personalized Medicine. *Cancers*, *12*(10), 2727. https://doi.org/10.3390/cancers12102727.

53. Kibbey, M. C. (1994). Maintenance of the EHS sarcoma and Matrigel preparation. *Journal of Tissue Culture Methods*, *16*(3–4), 227–230. https://doi.org/10.1007/bf01540656.

54. Piccinini, F., Balassa, T., Szkalisity, A., Molnar, C., Paavolainen, L., Kujala, K., Buzas, K., Sarazova, M., Pietiainen, V., Kutay, U., Smith, K., & Horvath, P. (2017). Advanced Cell Classifier: User-Friendly Machine-Learning-Based Software for Discovering Phenotypes in High-Content Imaging Data. *Cell systems*, *4*(6), 651–655.e5. https://doi.org/10.1016/j.cels.2017.05.012.

55. McQuin, C., Goodman, A. C., Chernyshev, V. S., Kamentsky, L., Cimini, B. A., Karhohs, K. W., Doan, M., Ding, L., Rafelski, S. M., Thirstrup, D., Wiegraebe, W., Singh, S., Becker, T., Caicedo, J. M., & Carpenter, A. E. (2018). CellProfiler 3.0: Next-generation image processing for biology. *PLOS Biology*, *16*(7), e2005970. https://doi.org/10.1371/journal.pbio.2005970.

56. Borten, M. A., Bajikar, S. S., Sasaki, N., Clevers, H., & Janes, K. A. (2018). Automated brightfield morphometry of 3D organoid populations by OrganoSeg. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-017-18815-8.

57. Kassis, T., Hernandez-Gordillo, V., Langer, R., & Griffith, L. G. (2019). OrgaQuant: Human Intestinal Organoid Localization and Quantification Using Deep Convolutional Neural Networks. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-48874-y.

58. Gritti, N., Lim, J. L., Anlas, K., Pandya, M., Aalderink, G., Martínez-Ara, G., & Trivedi, V. (2021). MOrgAna: accessible quantitative analysis of organoids with machine learning. *Development*, *148*(18). https://doi.org/10.1242/dev.199611.

59. Stachowiak, E. K., Benson, C. A., Narla, S. T., Dimitri, A., Chuye, L. E. B., Dhiman, S., Harikrishnan, K., Elahi, S., Freedman, D., Brennand, K. J., Sarder, P., & Stachowiak, M. K. (2017). Cerebral organoids reveal early cortical maldevelopment in schizophrenia—computational anatomy and genomics, role of FGFR1. *Translational Psychiatry*, *7*(11). https://doi.org/10.1038/s41398-017-0054-x.

60. Tindle, C., Katkar, G. D., Fonseca, A.G., Taheri, S., Lee, J., Maity, P., Sayed, I.M., Ibeawuchi, S., Vidales, E., Pranadinata, R.F., Fuller, M., Stec, D.L., Anandachar, M. S., Perry, K., Le, H.N., Ear, J., Boland, B.S., Sandborn, W.J., Sahoo, D., Das, S., Ghosh, P. (2023). A Living Organoid Biobank of Crohn's Disease Patients Reveals Molecular Subtypes for Personalized Therapeutics. *bioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2023.03.11.532245.

61. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J., White, D. K., Hartenstein, V., Eliceiri, K. W., Tomancak, P., & Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nature Methods*, *9*(7), 676–682. https://doi.org/10.1038/nmeth.2019.