CSI 5386

Natural Language Processing

Project Proposal

CLIP-based Vision-Language Model for Automated Hirschsprung's Disease Diagnosis

Authors – Group 9:

Atallah Madi – 300131314

Rowan Hussein – 300381055

Youssef Megahed – 300460672

Report Date:

February 22nd, 2025

Introduction

Hirschsprung's disease is a congenital defect characterized by the absence of ganglion cells within the myenteric plexus regions of the colon's muscularis propria. Accurate diagnosis of this condition is crucial for effective treatment, which typically involves the removal of the aganglionic section of the colon and reconnection to the anus through a surgical intervention known as the pull-through procedure [1][4]. Traditionally, pathologists diagnose Hirschsprung's disease by visually assessing histopathology images of the colon. However, manual assessment is time-consuming, costly, and prone to inter- and intra-rater variability. These challenges can lead to inconsistencies in diagnosis and treatment planning, potentially affecting patient outcomes.

Quantitative analyses, such as counting ganglion cells and assessing their spatial distribution, may increase surgical success and improve patient outcomes by establishing clear criteria for discerning healthy and aganglionic sections [2][4]. However, such assessments would be highly time-consuming for pathologists and significantly increase healthcare costs. Additionally, manual assessment is susceptible to inter- and intra-rater variability, further complicating the diagnostic process.

To address these challenges in this study, we propose a methodology that augments a CLIP-based vision-language model with expert-derived concepts to guide the classification of plexus regions within the muscularis propria layer. The process begins with expert concept induction, where relevant medical literature and expert texts detailing the histopathological characteristics of plexus regions are curated from reputable sources. A large language model will then extract and refine detailed textual descriptors—such as "dense interstitial arrangement," "distinctive staining patterns," and "specific cellular morphology"—to capture the nuanced expert knowledge critical for identifying plexus regions. The proposed method will be evaluated using precision and recall. This will ensure that the classification accurately captures all critical plexus regions for effective Hirschsprung's disease diagnosis.

Dataset

The proposed study will utilize a dataset similar to that employed in recent Hirschsprung's disease research [3]. The dataset comprises 30 whole slide images (WSIs) from 26 patients acquired at 20x magnification using high-resolution digital scanners shown in Figure 1. This dataset is from a closed source and is only accessible to the members found on the approved Research Ethics Boards (REB) list. Each WSI is accompanied by three sets of annotations: ***(1) Muscularis Propria Segmentation**,* Manually delineated boundaries of the muscularis layer; ***(2) Plexus Regions**:* Roughly segmented regions within the muscularis that likely contain ganglion cells; and ***(3) Ganglion Cells:*** Annotations that indicate the presence and spatial distribution of ganglion cells, accompanied by confidence levels. Preprocessing will include colour normalization using the Macenko method to mitigate staining variability and tiling WSIs into overlapping sub-images (224 x 224 pixels) for both training and inference.

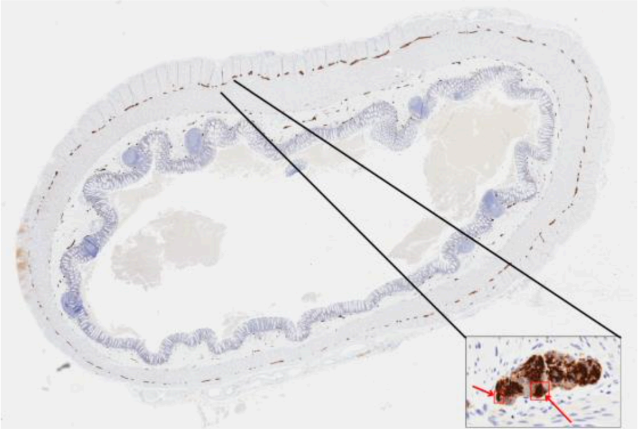

Figure 1: Whole slide image of a cross-section of the colon. Zoomed-in portion shows a plexus region with ganglion cells indicated with red arrows [4].

Current Methodologies

Existing approaches for the automated analysis of Hirschsprung's disease typically fall into two categories: ***K-Means Clustering Models*** or ***ViT-based Models***.

K-means Clustering Models are considered a traditional shallow learning method with k-means clustering to segment the plexus regions. In prior work, this method achieved a Precision score of approximately 73.8%, a Recall of 85.9%, and a GIR of 99.2% [3][4]. Although computationally inexpensive, the k-means approach often fails to capture complex tissue morphology due to its reliance on unsupervised clustering without domain-specific guidance. Furthermore, more recent studies have employed Vision Transformers (ViTs)(ViT-based Models), which divide images into patches and use self-attention mechanisms to capture local and global contextual information. The ViT approach has improved over K-Means by achieving a Precision score of around 84.2%, a Recall of 94.8% and almost a perfect GIR (99.7%) [6]. However, despite these strong performance metrics, the ViT model's reliance solely on pixel-level features can limit its ability to integrate domain-specific expert knowledge regarding plexus regions' appearance and subtle variations.

While these baselines have pushed the state-of-the-art in detection accuracy, neither explicitly incorporates the rich, nuanced expert knowledge available in the literature. This gap motivates the proposed methodology, which seeks to fuse the strengths of CLIP-based vision-language models with expert linguistic priors to achieve a more interpretable and precise identification of plexus regions.

Proposed Methodology

The proposed approach will augment a CLIP-based vision-language model with expert-derived concepts to guide the identification of plexus regions. The process will begin with expert concept induction, where relevant medical literature and expert texts that detail the histopathological characteristics of plexus regions will be curated from reputable sources. A large language model, such as *GPT-4o* or *DeepSeek-R1*, will then be employed to extract and refine detailed textual descriptors—examples include phrases like "dense interstitial arrangement," "distinctive staining patterns," and "specific cellular morphology"—that capture the nuanced expert knowledge critical for identifying plexus regions.

Following the generation of these expert concepts, the methodology will leverage the power of CLIP-based feature extraction. WSIs will be divided into patches/tiles and processed through a CLIP-based model (e.g. MedClip), encoding visual data and inducing textual descriptors into a shared embedding space. This alignment will ensure that the image features are meaningfully correlated with the expert-driven language, allowing for a more informed detection process. A two-stage aggregation process will then be employed: initially, patch-level similarity scores will be computed by comparing each image patch with the induced expert concepts, generating detailed similarity maps that highlight potential plexus regions; subsequently, these patch-level predictions will be hierarchically aggregated into comprehensive slide-level identification, effectively integrating both local details and global context [5].

The model will be fine-tuned for training using a combination of standard classification losses—such as Crossentropy loss—and additional alignment losses designed to encourage the image features to match the expert-derived concepts closely. At the same time, the model will learn other data-driven representations that take in more morphological details which are not explicitly covered by the expert inputs and, therefore, robust and whole feature extraction. Grad-CAM will be applied as the explainable AI (XAI) method to enhance interpretability and generate heatmaps that visually indicate the regions corresponding to the expert concepts. These Grad-CAM visualizations will allow clinicians to verify that the model focuses on the appropriate anatomical areas, providing an interpretable and clinically meaningful output. The performance will be evaluated using precision, recall, ensuring that the detection accurately captures all critical plexus regions for an effective and automated Hirschsprung's disease diagnosis.

References

[1] S. Lotfollahzadeh, M. Taherian, and S. Anand, "Hirschsprung disease," in *StatPearls*, StatPearls Publishing, 2023.

[2] A. J. Demetris et al., "Intraobserver and interobserver variation in the histopathological assessment of liver allograft rejection," *Hepatology*, vol. 14, no. 5, pp. 751–755, 1991.

[3] J. A. Kurian, "Automated Identification of Myenteric Ganglia in Histopathology Images for the Study of Hirschsprung's Disease," *Carleton University Institutional Repository*, 2021.

[4] Megahed, Y., Fuller, A., Abou-Alwan, S., Demellawy, D. E., & Chan, A. D. (2024). Segmentation of Muscularis Propria in Colon Histopathology Images Using Vision Transformers for Hirschsprung's Disease. *arXiv preprint arXiv:2412.20571.*

[5] W. Zhao, Z. Guo, Y. Fan, Y. Jiang, F. Yeung, and L. Yu, "Aligning knowledge concepts to whole slide images for precise histopathology image analysis," *npj Digital Medicine*, vol. 7, no. 1, Dec. 2024.

[6] Megahed, Y., Fuller, A., Abou-Alwan, S., Demellawy, D. E., & Chan, A. D. (2025). Full Ganglia Identification Pipeline Using Vision Transformers for Hirschsprung's Disease. In preparation for ACM Transactions.