Aligning Knowledge Concepts to Whole Slide Images for Hirschsprung's Disease Classification

Youssef Megahed
MASc, Data Science, Analytics, and Artificial Intelligence
Carleton University
Ottawa, Canada
youssefmegahed@cmail.carleton.ca

Atallah Madi
MASc, Electrical and Computer Engineering
Carleton University
Ottawa, Canada
atallahmadi@cmail.carleton.ca

Rowan Hussein
MCS, Computer Science Concentration Applied Artificial Intelligence
University of Ottawa
Ottawa, Canada
rhuss060@uottawa.ca

***Abstract*—Hirschsprung's Disease (HD) is a congenital condition marked by the absence of ganglion cells in the colon, leading to bowel obstruction. Accurate classification of myenteric plexus regions in histopathological whole slide images (WSIs) is critical for effective diagnosis and surgical intervention. While conventional deep learning models like Vision Transformers (ViTs) have demonstrated strong performance, they lack the interpretability and domain-aligned reasoning of clinical decision-making. In this study, we propose a novel framework that integrates expert-derived textual concepts into a CLIP-based vision-language model to guide plexus classification. Using expert-validated prompts generated by large language models and encoded with QuiltNet, our approach aligns clinically relevant semantic cues with visual features. Experimental results show that although the baseline ViT model achieved slightly higher classification accuracy (87.17% vs. 83.93%), the proposed model demonstrated superior discriminative capability with an AUC of 91.76%. These findings highlight the potential of multi-modal learning in histopathology and underscore the value of incorporating expert knowledge for more clinically relevant model outputs**

***Index Terms*—Hirschsprung's disease, classification, histopathology, deep learning.**

I. Introduction

A. Scientific Problem and Importance

HIRSCHSPRUNG disease (HD) is a congenital birth defect, involving the malformation of specific nerve cells within the colon causing bowel obstruction. HD has a global prevalence rate of 1 in 5000 infants and can be fatal if left untreated. The clinical hallmark of HD is the absence of ganglion cells within the colon tissue. These ganglion cells should be found within myenteric plexus regions, within muscularis propria section of the colon (Fig. 1 and Fig. 10). A common surgical intervention is the pull-through procedure, which removes the aganglionic section of the colon and joins the healthy portion to the anus.

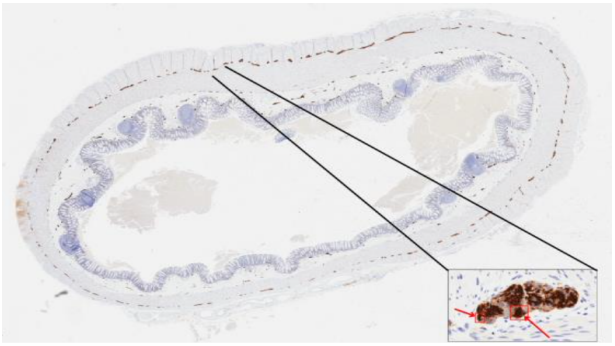

Fig. 1. Whole slide image of a cross-section of the colon. The zoomed-in portion shows a plexus region with ganglion cells indicated with red arrows [13].

Discerning healthy and aganglionic sections of the colon is performed by pathologists who visually assess histopathology images of the colon. More quantitative analyses, such as counting of the ganglion and assessment of their spatial distribution, may lead to increased surgical success and better patient outcomes by establishing more obvious criteria for discerning healthy and aganglionic sections of the colon; however, such quantitative assessment would be time-consuming for pathologists and would increase healthcare costs. In addition, manual assessment is prone to inter- and intra-rate variability since different pathologists would label the cells as ganglion cells differently [1], [2]. This motivates the need for a computational tool that can assist or automate the quantification of ganglion cells to support a more consistent and efficient diagnostic

process.

B. Related Literature Review

Attempts have been made to automate the analysis of histopathology whole slide images (WSIs) of the colon, where the detection of ganglion cells was divided into a three-stage problem [3]–[6], [13]: (1) segmenting the muscularis, (2) within the muscularis, segmenting the plexus regions, and (3) within the plexus regions, identifying ganglion cells.

In prior works for stage (2), a shallow machine learning approach known as k-means clustering was used and it achieved a Precision score of approximately 73.8%, Recall of 85.9%, a Ganglia Inclusion Rate (GIR) of 99.2% [6]. GIR is the percentage of all ganaglia cells that are found within the segmented plexus region. This is a valuable measure, as not all plexus segmentation errors are detrimental. If the segmented plexus still contains all the ganaglia cells, the segmentation can be considered acceptable because, ultimately, we are interested in identifying the ganglion cells, which are found in these plexus regions.

Recently, deep learning approaches have demonstrated a state-of-art performance in medical image analysis, including computational pathology [13]. Convolutional neural networks (CNNs) are one of the popular deep-learning approaches that are used for stage (1) with 81.9% Precision, and Recall of 96.2% [5], [6]. However, Vision Transformers (ViTs) are gaining popularity. ViTs can detect long-range patterns more efficiently than CNNs, which have narrow receptive fields at each layer. ViTs are non-hierarchical — they retain the spatial dimensions across layers, rather than gradually reducing the spatial dimensions via pooling. This feature of ViTs makes adapting them to segmentation effective and simple (i.e., by placing a single linear layer atop patch representations) [7]. ViTs have been applied to histopathology images for object detection [9], classification [10], and segmentation [11]. In [24], ViT model is applied to stage (2) which has shown better results than K-means clustering with 84.2% Precision, Recall of 94.8%, GIR of 99.7%.

These automated methods have shown promising results and have raised the state-of-art limits; however, they rely solely on visual features extracted from histopathology images. This image-centric approach does not fully align with how pathologists typically classify Hirschsprung's disease. In clinical practice, diagnosis is inherently a multi-modal and interpretive process that involves visual identification of ganglion cells and contextual information such as tissue architecture, patient history, and spatial relationships between anatomical structures. This limitation can lead to model misattribution, where the system makes decisions based on spurious correlations rather than clinically relevant features. A well-known example of this is a model trained to differentiate between dogs and wolves, which mistakenly learned to associate the presence of snow in the background with wolves (Fig. 2) [12]. Similarly, models analyzing histopathology images may rely on irrelevant visual cues rather than the true indicators of disease. This gap underscores the need for systems that incorporate domain-specific knowledge and mimic the nuanced reasoning employed by human experts.

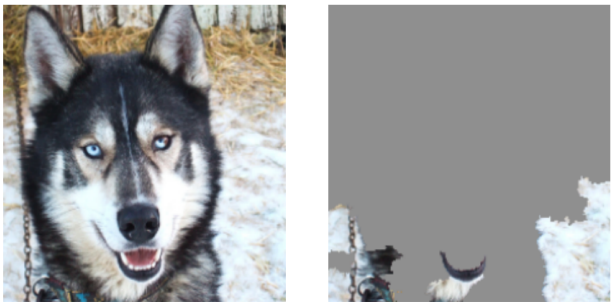

Fig. 2. Image of Husky dog that was classified as a wolf with snow being the main attribution [12].

C. Research Objectives

This study 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. This will ensure that the classification captures all critical plexus regions for effective Hirschsprung's disease diagnosis and that the process is similar to the human expert diagnosis rational.

All the methods discussed in sub-section (Related Literature Review) rely on pixel-level segmentation, known as semantic segmentation. In contrast, the current approach frames the problem as a classification task, where correctly classifying plexus regions leads to improved identification of all ganglia cells within them, achieving this study's main objective (HD diagnosis), and goals which are:

- To create a ViT-based baseline model for benchmarking and comparison against the proposed vision-language approach.
- To extract domain-specific textual concepts from reputable medical literature using large language models (e.g., GPT-4o and DeepSeek-R1).
- To integrate textual descriptors with visual features extracted from whole slide images (WSIs) using a CLIP-based model, aligning image patches with clinically meaningful expert knowledge.
- To evaluate the effectiveness of the proposed method using standard classification metrics such as precision, recall, F1-score, and AUC.

II. METHODOLOGY

A. Data Overview

The proposed study will utilize a dataset that was employed in recent Hirschsprung disease research [6], [13]. The dataset consists of 30 WSIs from 26 patients diagnosed with Hirschsprung disease. Images were acquired by the Children's Hospital of Eastern Ontario (CHEO) using the digital scanner Aperio Scan Scope CS at 20× resolution (0.50 µm/pixel).The size of WSIs ranged from 17,983×17,602 to 65,940×34,997 pixels, with three colour channels. This dataset is from a closed source and is only accessible to the members found on the approved Research Ethics Boards (REB) list. Therefore, only the author Y. Megahed had access to the original dataset in its raw format, while the other authors were allowed to access the data only after processing it into a tile format (224x224).

Each WSI is associated with three manually annotated ground truth images: (i) the muscularis propria, (ii) the myenteric plexus regions (i.e., a visually noticeable amount of tissue around the plexus regions are also included) shown in Fig. 3, and (iii) the ganglion cells (i.e., parts of the ganglion cell may be excluded and/or areas around the cell may be included, as it can be difficult to visually identify the exact border of a ganglion cell). A confidence level accompanies each ganglion cell annotation. A high confidence level indicates a high certainty that the annotated object is a ganglion cell, which can be seen in Fig. 10. A low confidence level indicates a belief that the annotated object is a ganglion cell, but there is some uncertainty.

B. Data Preprocessing

There are several preprocessing steps that were applied to the entire dataset to ensure consistency and to generate suitable inputs for model training and evaluation of the classification task.

First, colour normalization is performed using the Macenko method; a widely adopted stain normalization technique in computational pathology [8]. This will eliminate the colour variation found in each WSI which results in reducing inter-slide and inter-patient staining variability caused by the different histological staining process or scanning conditions.

Next, WSIs were downsampled from 20x to 5x. This lower resolution helps to save on the computational costs, and allows the model to capture an entire plexus region, given that 20x is zoomed in and captures only parts of it. As a result, the model will have a complete view of the plexus regions in a sufficient resolution for accurate predictions. Then, overlapping tiles are extracted in a size of 224 x 224 pixels from each WSI to create compatible inputs for the deep learning models that will be experimented with. The tiles were extracted with a fixed stride to ensure contextual continuity and adequate coverage, especially at region boundaries.

Lastly, classification labels were derived from the manually annotated ground truth segmentation maps (Fig. 3). Each tile was labelled as "plexus" if at least one pixel overlapped with the annotated myenteric plexus regions; otherwise, it was labelled as "no plexus." This binary classification approach allows the model to distinguish between plexus and non-plexus regions within the muscularis propria layer.

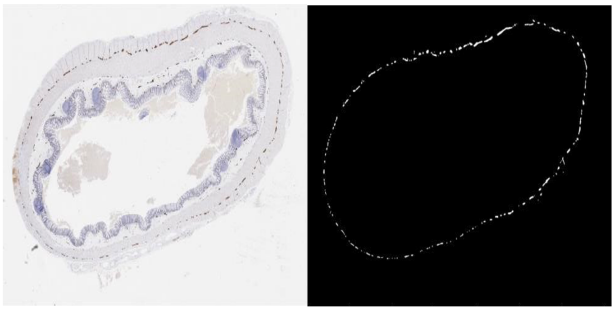

Fig. 3. WSI sample and its corresponding ground truth masks for plexus region

C. ViT Model (Baseline)

The baseline in this study is a pretrianed Vision Transformer model published by Google through the Hugging Face Transformers library (google/vit-base-patch16-224). This model adheres to the original ViT architecture shown in Fig. 4 [14], and is trained on the large-scale ImageNet-1k dataset (1.2 million images) [16]. Pretrained models are widely adopted in medical imaging tasks due to their ability to capture rich and transferable visual features, reducing the need for training large models from scratch, especially since ViTs are considered data-hungry models and require large datasets for training [14]. Given that, the model will be fine-tuned to leverage the pretrained knowledge while adapting the model to the domain-specific features present in histopathology images.

The model processes input images of size 224×224 pixels, which were generated during the preprocessing pipeline. Each image is divided into a grid of 16×16 non-overlapping patches, resulting in 14×14 = 196 patches [14]. Each patch is flattened and passed through a linear projection layer, transforming into a 768-dimensional embedding. A special classification token (CLS) is prepended to the sequence of patch embeddings [14]. The positional embeddings are added to the input sequence to encode spatial information lost during flattening, which maintains spatial context across patches.

The embedded sequence is processed through a stack of 12 Transformer encoder blocks. Each block contains a

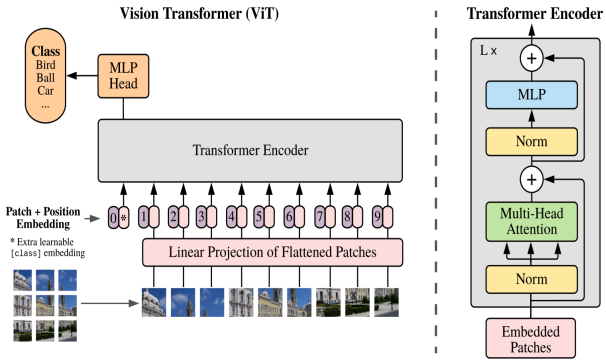

Fig. 4. Vision Transformer architecture [14].

multi-head self-attention mechanism (with 12 attention heads) and a feed-forward network, followed by layer normalization and residual connections (Fig. 4) [14]. These layers enable the model to capture long-range dependencies and interactions across image patches, which is particularly valuable in histopathological contexts where spatial patterns span across multiple cellular structures. The output corresponding to the CLS token is passed through a final linear classification head to produce logits for the binary classification task.

This ViT-based model serves as a foundational benchmark for the study as it represents a purely vision-based classification strategy without incorporating domain-specific knowledge or multi-modal reasoning (no LLM incorporation). As such, it provides a strong reference point for our proposed model performance to be compared to.

D. CLIP Model (Proposed)

The proposed framework (Fig. 5) builds upon recent advancements in Vision-Language Models (VLMs), particularly those based on the CLIP (Contrastive Language-Image Pretraining) architecture, to align domain-specific expert knowledge with histopathological image representations. The proposed approach aims to emulate human reasoning computationally by integrating expert-derived textual concepts with CLIP-based multi-modal embeddings (Fig. 6) [15].

At the core of this framework is the use of Large Language Models (LLMs), such as GPT-4 and DeepSeek-R1, to extract and refine expert-level textual descriptions from curated medical literature. Unlike traditional LLM applications that act as knowledge databases, here, LLMs are employed as reasoning engines to induce disease-specific concepts relevant to plexus identification [15], [17]. Using LLMs as a reasoning machine mitigates key limitations of open-ended LLM prompting, such as hallucination (i.e., generating factually incorrect or unsupported claims) [18], by grounding outputs in structured expert knowledge (Fig. 6).

The instance-level expert concepts and bag-level expert class prompt for Hirschsprung's Disease were extracted from trusted scientific sources of medical literature (e.g., pathology textbooks, peer-reviewed journals) found in [4], [5], [21], [22]. These concepts and prompts were extracted using four different LLMs (GPT-4o, GPT-o3-mini-high, DeepSeek-R1, and Grok-3). After comparative testing of multiple LLMs, DeepSeek-R1 generated the most clinically coherent and contextually accurate outputs, as validated by Dr. Adrian Chan (our expert) during the prompt induction process.

Recognizing that medical literature may not fully capture the morphological complexity of all diseases, the model also learns purely data-driven concept prompts. These are initialized as learnable embeddings and optimized through training, providing complementary features that augment expert-derived knowledge [15]. For instance, in challenging classification tasks (e.g., EBV subtyping), the ConcepPath framework demonstrated up to a 4% performance gain by integrating learned concepts alongside expert priors [15].

After generating expert concepts, the QuiltNet model (pre-trained on the Quilt-1M dataset) is applied to embed both the image patches and concept prompts into a shared latent space. This model is a CLIP-based pathology model trained on over 1 million histopathology image-text pairs is used [19]. The extracted tiles (224×224 pixels) are encoded into visual embeddings, while the textual phrases are embedded via a text encoder. The resulting representations allow direct calculation of similarity scores between visual features and expert-derived features [15], [20]. This similarity is measured using cosine similarity as it quantifies the angular closeness between their respective embeddings in the shared latent space [15]. This alignment provides interpretability; the similarity scores between patches and concepts are visualized as similarity maps, offering insights into the model's decision-making process.

The model then employs a two-stage (Fig. 5), concept-guided hierarchical aggregation mechanism consisting of a Concept-Level Aggregation followed by a Bag-Level Aggregation. In the first stage, instance-level features (i.e., patch embeddings) are aggregated into concept-specific bag-level representations [15]. This is achieved through a similarity-based attention mechanism guided by both expert-derived and data-driven instance-level concepts. In the second stage, these concept-specific features are further aggregated into a comprehensive slide-level representation [15]. This is done by measuring the correlation between the concept-level embeddings and the bag-level expert class prompts, effectively capturing higher-level diagnostic context across the whole slide.

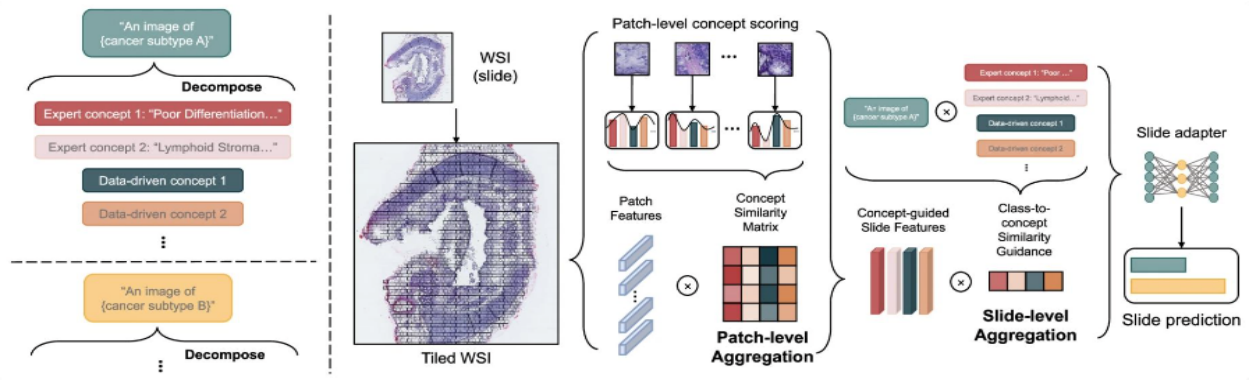

Fig. 5. ConcepPath framework decomposes a specific complex WSI analysis task into multiple subtasks of scoring patch-level concepts [15].

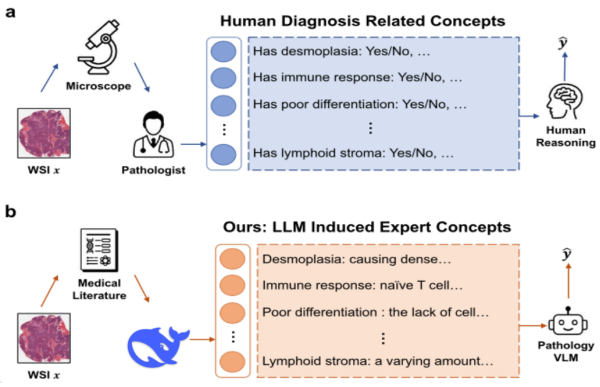

Fig. 6. a) The real clinical processes. b) Utilizing a large language model to induce expert concepts related to diagnosis from medical literature [15].

To mitigate domain shifts between the pre-trained encoders and the downstream histopathology classification task, slide adapters are introduced (Fig. 5) [15]. These are learnable bottleneck layers and they are inserted between the aggregated slide representation and the classification head [22]. Slide adapters serve two primary functions. First, they adapt the feature distribution to better match the target domain, and second, they enable residual-style feature blending that enriches the original embedding with task-specific signals. This lightweight adjustment mechanism improves generalization without requiring full fine-tuning of the pre-trained vision-language backbone [15].

III. Model Training and Validation

A. Training Setup

For training both the baseline Vision Transformer (ViT) model and our proposed CLIP-based ConcePath model, we utilized a consistent training setup to ensure fair comparison and robust evaluation of the model performances. We used a 5-fold cross-validation, splitting the 30 WSIs into groups of 6. Within each fold, five groups (24 WSIs) were used to train the model, 0.5 group (3 WSIs) was used for validation, and 0.5 group (3 WSIs) was used to test the model (classifying plexus). The test group was rotated in each fold, so half of the WSIs were eventually tested. WSIs were tiled into overlapping smaller sub-images of size 224×224 pixels. From each WSI in the training set, 500 tiles were sampled (250 with plexus, 250 with no plexus pixels), resulting in a total of 15,000 tiles.

The baseline ViT model was initialized with weights pre-trained on ImageNet-1k via masked autoencoding. The model was fine-tuned using AdamW optimizer, configured with an initial learning rate of 5e-4 and weight decay of 1e-4. A cosine learning rate scheduler was applied, featuring a warm-up period of five epochs, after which the learning rate was gradually reduced until training completion. Each training cycle comprised 20 epochs with a batch size of 64 patches.

The training for our proposed CLIP-based ConcePath model involved a two-stage hierarchical aggregation approach to leverage visual data and DeepSeek-R1-derived expert concepts effectively. We utilized the QuiltNet vision-language model, explicitly employing the ViT-B/32 image encoder and GPT/77 text encoder. The training incorporated instance-level and bag-level expert concepts, represented through 16 learnable tokens, alongside one instance-level concept for each target classification class. The ConcePath model employed an Adamw optimizer with a learning rate set at 1e-4, optimized through patient-level five-fold cross-validation to ensure robust model generalization. Data augmentation techniques were applied during training, including random rotations, horizontal and vertical flips, and scaling for both models.

We froze the pre-trained image and text encoders during training to retain generalizable features learned from the extensive QuiltNet dataset. We only trained the newly introduced learnable embeddings for instance-

level expert and data-driven concepts. Additionally, we utilized slide adapters as bottleneck layers to mitigate potential domain shifts between the QuiltNet training data and our specific downstream histopathology analysis task, blending adapted features with the original embeddings.

This structured and detailed training setup enabled a thorough comparative analysis between our proposed ConcePath methodology and the ViT-based baseline, laying a strong foundation for subsequent evaluation and interpretation of model performance.

B. Hyperparameter Tuning

Hyperparameter tuning was systematically performed to optimize model performance. The baseline ViT model underwent tuning primarily for learning rate, batch size, and epochs, employing a grid search approach within specified ranges. The optimal learning rate was found to be 5e-4, with a batch size of 64 and a training duration of 20 epochs (we tested with low epoch sizes for computational and time constraints).

For the proposed ConcePath model, additional hyperparameters required tuning, such as the number of data-driven concepts (n_ddp), the orthogonal ratio, and the number of learnable context tokens (n_ctx). We conducted grid searches and evaluations based on validation set performance, ultimately selecting 8 data-driven concepts, an orthogonal ratio of 2, and 16 learnable context tokens. These choices were guided by performance metrics such as validation loss, accuracy, macro/micro F1 score, and AUC values, as shown in Fig. 7.

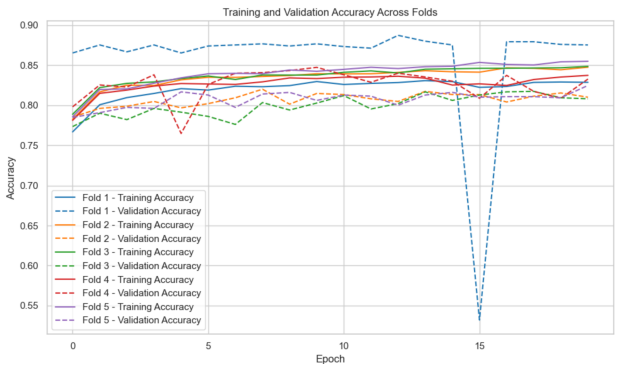

Fig. 7. Training and validation accuracy curve across 5 folds.

C. Evaluation Metrics

To comprehensively evaluate model performance in plexus region classification tasks, we utilized several standard classification metrics shown through equations (1) - (6): *accuracy, recall, specificity, F1-score (micro and macro), and area under the receiver operating characteristic curve (AUC)*. These metrics provide a balanced view of model behaviour across class distributions and are particularly relevant when addressing class imbalance in classification tasks.

Let TP (true positives), TN (true negatives), FP (false positives), and FN (false negatives) be defined in the context of 224x224 tile classification, where positive refers to pixels predicted as belonging to the target class (e.g., plexus), and negative otherwise.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3}$$

$$\text{F1}_{\text{micro}} = \frac{2 \cdot \sum TP}{2 \cdot \sum TP + \sum FP + \sum FN} \tag{4}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{F1}_{\text{macro}} = \frac{1}{C} \sum_{i=1}^{C} \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{6}$$

where C is the number of classes and AUC refers to the area under the ROC curve, which summarizes the trade-off between sensitivity and specificity across thresholds. A higher AUC indicates better discrimination capability between classes.

IV. Results and Interpretation

A. Model Performance

The performance metrics for both the baseline ViT-B16 model and the proposed QuiltNet model are summarized in Table I. Overall, the ViT-B16 baseline model achieved higher scores across most metrics with an accuracy of 87.17%, recall of 85.23%, specificity of 89.12%, F1 Micro and Macro scores around 87%, and an AUC of 87.17%. On the other hand, the proposed QuiltNet model achieved slightly lower performance in terms of accuracy, recall, specificity, and F1 scores (around 83–84%) but showed significant improvement in the AUC metric, achieving 91.76%.

Confusion matrices provided further insights into the model performances (Fig. 8 and Fig. 9). The ViT-B16 model demonstrated a stronger ability to correctly classify both plexus and non-plexus regions, with fewer false positives (408) and false negatives (554). Conversely, the QuiltNet model showed higher rates of misclassification, particularly in false negatives (741 cases), indicating it incorrectly classified a larger number of plexus regions as non-plexus. Despite this, the higher AUC of QuiltNet

TABLE I
COMPARISON OF MODEL PERFORMANCE METRICS BETWEEN VIT-B16 (BASELINE) AND QUILTNET (PROPOSED).

Models	Accuracy (%)	Recall (%)	Specificity (%)	F1 Micro (%)	F1 Macro (%)	AUC (%)
ViT-B16 (Baseline)	**87.17**	**85.23**	**89.12**	**87.17**	**87.14**	87.17
QuiltNet (Proposed)	83.93	83.68	84.40	83.93	83.86	**91.76**

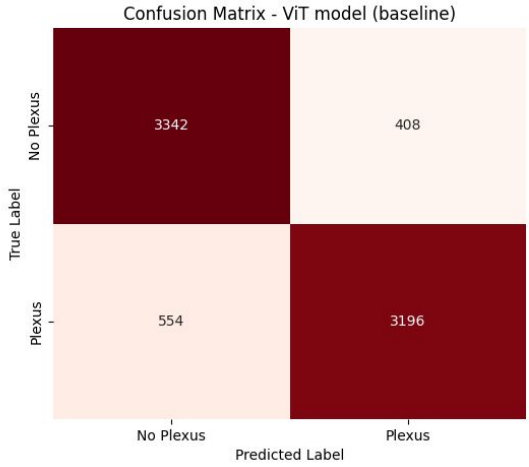

Fig. 8. ViT model (baseline) confusion matrix.

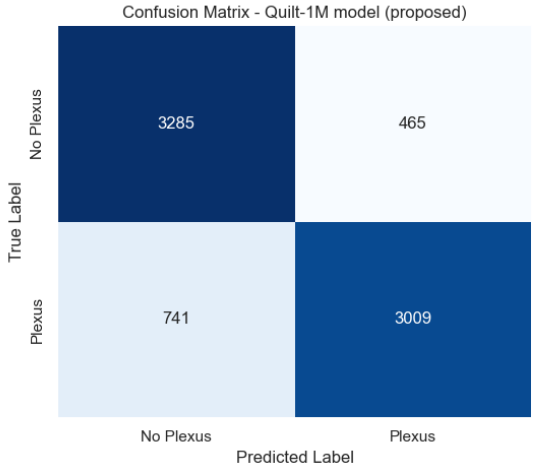

Fig. 9. QuiltNet model (proposed) confusion matrix.

suggests better overall discriminative power across different threshold values, highlighting its potential in tasks that prioritize sensitivity-specificity trade-offs.

Impact of Prompt Diversity on Performance: A contributing factor to the QuiltNet model's lower performance metrics, compared to the ViT-B16 baseline, is the utilization of only a single prompt per class during training. This limitation restricts the model's exposure to the full spectrum of intra-class variability, potentially impacting its ability to generalize to diverse clinical scenarios. In the current implementation, cosine similarity is employed to map textual descriptions to the predictions of image tiles. However, using only one prompt per class constrains the model's capacity to capture the nuanced variations within each class, which is particularly critical in histopathological contexts where tissue morphology can vary significantly. Despite this constraint, the QuiltNet model's performance approaching that of the ViT-B16 baseline underscores the promise of integrating textual descriptors into vision-language models. It suggests that with increased prompt diversity, the QuiltNet model could potentially surpass traditional vision-only models in performance, which is very exciting.

B. Connecting Model Outputs to Goals

The primary objective of our study was to enhance the classification of plexus regions in colon histopathology images to assist in the accurate and efficient diagnosis of Hirschsprung's Disease (HD). Although the baseline ViT model demonstrated strong classification performance overall, the proposed QuiltNet model's superior AUC indicates a robust ability to distinguish between plexus and non-plexus regions across varying decision thresholds. This capability is particularly beneficial in clinical scenarios where adjusting decision thresholds is essential to balance sensitivity and specificity according to diagnostic requirements.

The integration of expert-derived textual descriptors into the QuiltNet model introduces a layer of interpretability and aligns the automated analysis more closely with the clinical reasoning processes employed by pathologists. This multimodal approach reflects real-world diagnostic practices, where contextual and morphological understanding significantly impacts diagnostic accuracy.

Despite the lower performance scores compared to ViT-B16, QuiltNet's interpretability and its higher discriminative capability (as indicated by AUC) make it a valuable tool for augmenting pathologist expertise, particularly in borderline or ambiguous cases where clinical judgment relies heavily on nuanced contextual and visual information.

V. DISCUSSION

A. Scientific and Clinical Implications

This study introduces a novel approach to the classification of plexus regions within the muscularis propria

layer in histopathological images of the colon, specifically targeting the diagnosis of Hirschsprung's disease (HD). By integrating expert-derived textual concepts into a CLIP-based vision-language model, we aim to emulate the nuanced reasoning employed by pathologists, who rely on both visual cues and contextual knowledge during diagnosis.

The incorporation of domain-specific knowledge into the model addresses a critical gap in existing automated diagnostic tools, which often depend solely on visual features. This multimodal strategy enhances the model's interpretability and aligns its decision-making process more closely with clinical practices. Consequently, the proposed method holds the potential to assist pathologists in making more accurate and consistent diagnoses, ultimately improving patient outcomes.

Furthermore, the successful application of this approach to HD diagnosis suggests its broader applicability to other medical imaging tasks where expert knowledge plays a pivotal role. By demonstrating the feasibility and benefits of integrating textual concepts into image analysis, this work lays the groundwork for future advancements in computational pathology and the development of more sophisticated diagnostic tools.

B. Limitations

Despite the promising results, several limitations should be acknowledged that caused ViT to slightly outperform the CLIP model:

Limited Prompt Diversity: During training, only one prompt per class was utilized. This constraint has likely restricted the model's ability to capture the full range of intra-class variability, potentially impairing its generalizability across diverse clinical scenarios. This limitation arises from the architectural logic in the CLIP-based ConcePath implementation, where the computation of `slide_features_` requires strict dimensional alignment between slide-level and patch-level text features:

```
slide_features_ = torch.bmm(
    slide_text_features.unsqueeze(1),
    patch_text_features.reshape(-1, self.
        num_patch_prompt_, embedding_len).
        transpose(1, 2),
).squeeze(1)
```

Due to this formulation, the number of patch-level prompts must be consistent across all classes. However, since slide-level prompts are defined as a single prompt per class—per the model's original implementation—this inherently restricts prompt diversity. Modifying the number of patch-level prompts per class to increase variability would break the required dimensional agreement for batch matrix multiplication (`torch.bmm`). Although we contacted the ConcePath authors to clarify whether this was an intentional design decision or an oversight, I have not received a response in over a month.

Limited Model Documentation: The CLIP-based model employed is relatively new, and comprehensive documentation detailing its architecture, training procedures, and debugging is lacking. This scarcity of information poses challenges for reproducibility and hinders a thorough understanding of the model's behaviour.

Dataset Constraints: The dataset comprises 30 whole slide images (WSIs) from 26 patients, which may not encompass the full heterogeneity present in the broader patient population. Additionally, access to the raw dataset is restricted, limiting opportunities for external validation and collaborative research. Expanding the dataset further is extremely costly and time-consuming, as the annotation process requires detailed and expert-level input from pathologists. In particular, delineating the ground truth for the three intestinal layers (muscle, plexus regions, and ganglion cells) is a labour-intensive task that demands significant manual effort and domain expertise. Although we possess 15 additional WSIs to the 30 used ones, these slides remain unlabeled due to the high cost of expert annotation and were therefore excluded from training and evaluation. This underscores a key challenge in digital pathology: the scarcity of high-quality, labelled data, especially in tasks requiring fine-grained, region-specific annotations.

Annotation Variability: The ground truth annotations, particularly for ganglion cells, involve a degree of subjectivity and may vary significantly between annotators. Defining the precise borders and full extent of each intestinal layer is a challenging task, even for experienced pathologists, due to the gradual transitions between layers and the heterogeneous appearance of tissue across different regions and patients. This inherent ambiguity leads to inter-observer variability, where different experts may annotate the same WSI differently. In our dataset, for instance, ganglion cell masks are provided with two levels of annotation certainty or confidence, high-certainty and low-certainty masks, reflecting the varying degrees of confidence in the identification of these structures, as shown in Fig. 10. While this richer annotation scheme adds valuable nuance, it also introduces an additional layer of complexity for the model, which must learn to generalize from data that includes uncertain or inconsistent labels. This annotation variability can introduce noise into the training process and may limit the upper bound of achievable model performance.

Computational Resources: The high-resolution nature of whole slide images (WSIs), often comprising millions of pixels per image, necessitates substantial computational resources for both preprocessing and analysis. Even basic operations such as stain normalization can become computationally intensive at this scale. For example, applying Macenko normalization—a widely used technique in histopathology image preprocess-

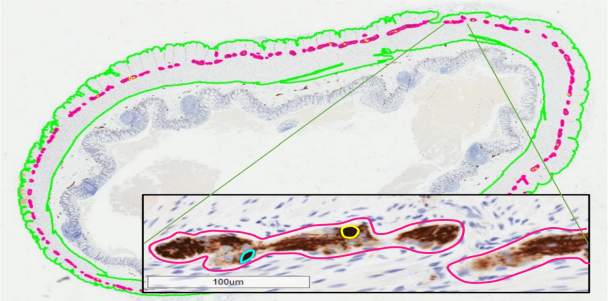

Fig. 10. Annotated histological section with the muscularis propria (green), myenteric plexus (pink), high-confidence ganglia (yellow), and low-confidence ganglia (cyan). The inset shows a magnified region with a 100µm scale bar [6].

ing—required us to first downsample the WSIs from 20× to 10× magnification to fit the memory, despite conducting the process on a high-memory machine with 96GB of RAM. This highlights the considerable resource demands inherent to working with WSIs, where memory consumption and processing time rapidly scale with image resolution. Such constraints pose practical challenges to the scalability and real-time applicability of the proposed method, particularly in resource-limited environments such as clinical settings or large-scale deployments.

C. Lessons Learned

Throughout the development and evaluation of the proposed method, several key insights emerged:

Complexity of Defining Anatomical Boundaries: Accurately delineating anatomical layers' borders and full extents within WSIs is inherently challenging. The gradual transitions between tissue types and the heterogeneous appearance across different regions can lead to significant inter-observer variability. Even experienced pathologists may interpret and annotate these regions differently, which underscores the need for standardized annotation protocols and consensus-building mechanisms to enhance data consistency.

Impact of Annotation Certainty on Model Learning: The presence of annotations with varying levels of certainty, such as high-certainty and low-certainty ganglion cell masks, introduces additional complexity into the training process. Models must learn to generalize from data that includes uncertain or inconsistent labels, which can affect performance. Developing strategies to handle uncertain annotations effectively is crucial for improving model robustness.

Scalability Challenges in Preprocessing High-Resolution Images: WSIs often comprise millions of pixels, resulting in substantial computational demands for preprocessing tasks like stain normalization. For instance, applying Macenko normalization required downsampling images from 20× to 10× magnification, even when utilizing a machine with 96GB of RAM. This highlights the need for more efficient preprocessing algorithms or hardware acceleration to manage large-scale image data effectively.

Limitations Due to Restricted Access to Raw Datasets: Restricted access to raw datasets limits opportunities for external validation. This constraint hinders the ability to benchmark models across diverse datasets and impedes the generalizability of findings. Encouraging data sharing and establishing common repositories can facilitate broader validation efforts and accelerate advancements in the field.

Necessity for Diverse and Extensive Training Data: The limited diversity of prompts and the relatively small dataset used in this study highlight the necessity for more extensive and varied training data. A broader dataset can capture a wider range of variability within each class, improving model robustness and generalizability to diverse clinical scenarios.

Resource Constraints Affecting Real-Time Applicability: The substantial computational resources required for processing high-resolution WSIs may limit the real-time applicability of the proposed method in resource-constrained settings. Exploring algorithmic optimizations and efficient processing techniques is necessary to address this limitation and facilitate practical deployment.

Importance of Multimodal Integration for Enhanced Interpretability: Incorporating expert-derived textual concepts into the model can improve performance and also enhance interpretability. This multimodal integration facilitates a better understanding of the model's decision-making process, which is vital in clinical applications where explainability is crucial.

D. Future Directions

Building upon the findings and addressing the identified limitations, future research should focus on the following areas:

Expanding Prompt Diversity: Developing a more comprehensive set of prompts per class can capture a wider range of features and variations, enhancing the model's ability to generalize across diverse cases.

Enhancing Model Transparency: Collaborating with the model developers to obtain detailed documentation and insights into the model's architecture and training processes will facilitate better understanding and reproducibility.

Augmenting the Dataset: Acquiring additional WSIs from a broader patient cohort, including diverse demographics and disease presentations, will improve the model's robustness and applicability.

Standardizing Annotations: Implementing standardized annotation guidelines and leveraging consensus among multiple experts can reduce variability and improve the quality of training data.

Optimizing Computational Efficiency: Exploring techniques such as model pruning, quantization, and efficient tiling strategies can mitigate computational challenges and support real-time applications.

Clinical Validation: Conducting prospective studies and clinical trials to evaluate the model's performance in real-world settings will provide valuable insights into its practical utility and impact on patient care.

VI. Conclusion

In this study a multi-modal framework is proposed to classify Hirschsprung's Disease with expert-derived knowledge concepts to emulate clinical diagnose process using a CLIP-based vision-language model. The ViT baseline model demonstrated higher performance across most conventional metrics, however, the proposed model outperformed in AUC, indicating better results in distinguishing plexus regions across various thresholds. Limitations such as restricted prompt diversity and dataset scale currently constrain the full potential of the proposed method. Future work should explore prompt expansion and larger annotated datasets to improve the model's overall performance. In support of reproducibility and further research in this domain, the full implementation of this work, including code, model configurations, and concept prompts, has been made publicly available on our GitHub repository.

References

[1] A. Mukherjee *et al.*, "The placental distal villous hypoplasia pattern: interobserver agreement and automated fractal dimension as an objective metric," *Pediatric and Developmental Pathology*, vol. 19, no. 1, pp. 31–36, 2016.

[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. Kurian *et al.*, "Image Processing and Analysis of Histopathological Images Relating to Hirschsprung's Disease," *CMBES Proceedings*, vol. 41, 2018.

[4] M. T. K. Law, A. D. C. Chan, and D. El Demellawy, "Color image processing in Hirschsprung's disease diagnosis," in *Proc. IEEE EMBS Int. Student Conf. (ISC)*, 2016, pp. 1–4.

[5] C. McKeen, F. Zabihollahy, J. Kurian, A. D. Chan, D. El Demellawy, and E. Ukwatta, "Machine learning-based approach for fully automated segmentation of muscularis propria from histopathology images of intestinal specimens," in *Medical Imaging 2019: Digital Pathology*, vol. 10956, pp. 146–151, Mar. 2019.

[6] J. A. Kurian, *Automated Identification of Myenteric Ganglia in Histopathology Images for the Study of Hirschsprung's Disease*, M.A.Sc. thesis, Carleton University, 2021.

[7] M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[8] M. Macenko *et al.*, "A method for normalizing histology slides for quantitative analysis," in *Proc. IEEE Int. Symp. Biomed. Imaging: From Nano to Macro*, 2009, pp. 1107–1110.

[9] A. Atabansi *et al.*, "Applications of Transformers in Histopathological Image Analysis: A Comprehensive Survey," *Biomedical Engineering Online*, vol. 23, no. 1, 2023. doi: 10.1186/s12938-023-01157-0.

[10] A. Kanadath, J. A. A. Jothi, and S. Urolagin, "CViTS-Net: A CNN-ViT Network with Skip Connections for Histopathology Image Classification," *IEEE Access*, 2024.

[11] L. Hörst *et al.*, "CellViT: Vision Transformers for Precise Cell Segmentation and Classification," *arXiv preprint arXiv:2306.15350*, 2023. [Online]. Available: https://arxiv.org/abs/2306.15350

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, San Francisco, CA, USA, 2016, pp. 1135–1144.

[13] Y. Megahed, A. Fuller, S. Abou-Alwan, D. E. Demellawy, and A. D. Chan, "Segmentation of Muscularis Propria in Colon Histopathology Images Using Vision Transformers for Hirschsprung's Disease," *arXiv preprint arXiv:2412.20571*, 2024. [Online]. Available: https://arxiv.org/abs/2412.20571

[14] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, Jun. 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[15] 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.

[16] "google/vit-base-patch16-224 · Hugging Face," Hugging Face. [Online]. Available: https://huggingface.co/google/vit-base-patch16-224

[17] D. Truhn, J. S. Reis-Filho, and J. N. Kather, "Large language models should be used as scientific reasoning engines, not knowledge databases," *Nature Medicine*, vol. 29, pp. 2983–2984, 2023.

[18] L. Qu *et al.*, "The rise of AI language pathologists: exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification," *NeurIPS*, 2023.

[19] W. O. Ikezogwo *et al.*, "Quilt-1M: One million image-text pairs for histopathology," *NeurIPS*, 2024.

[20] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. 38th International Conference on Machine Learning*, 2021.

[21] K. Rathi, A. Verma, and P. Pingat, "Early Detection and Intervention for Hirschsprung's Disease: A Key to Successful Outcomes," *Clin Med Insights Case Rep.*, vol. 17, p. 11795476241226577, Jan. 2024, doi: 10.1177/11795476241226577. PMID: 38269147; PMCID: PMC10807328.

[22] L. Szylberg and A. Marszałek, "Diagnosis of Hirschsprung's disease with particular emphasis on histopathology. A systematic review of current literature," *Prz Gastroenterol.*, vol. 9, no. 5, pp. 264–269, 2014, doi: 10.5114/pg.2014.46160. Epub Oct. 18, 2014. PMID: 25395999; PMCID: PMC4223113.

[23] P. Gao *et al.*, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, vol. 132, pp. 581–595, 2024.

[24] Y. Megahed, A. Fuller, S. Abou-Alwan, D. E. Demellawy, and A. D. Chan, (2025). Full Ganglia Identification Pipeline Using Vision Transformers for Hirschsprung's Disease. In preparation for ACM Transactions on Multimedia Computing, Communications, and Applications, Special Issue on Advancing Medical Segmentation Through Emerging Deep Learning Architectures and Large Models.