Interpretable histopathology-based prediction of disease relevant features in Inflammatory Bowel Disease biopsies using weakly-supervised deep learning

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Crohn's Disease (CD) and Ulcerative Colitis (UC) are the two main Inflammatory Bowel Disease (IBD) types. We developed deep learning models to identify histological disease features for both CD and UC using only endoscopic labels. We explored fine-tuning and end-to-end training of two state-of-the-art self-supervised models for predicting three different endoscopic categories (i) CD vs UC (AUC=0.87), (ii) normal vs lesional (AUC=0.81), (iii) low vs high disease severity score (AUC=0.80). We produced visual attention maps to interpret what the models learned and validated them with the support of a pathologist, where we observed a strong association between the models' predictions and histopathological inflammatory features of the disease. Additionally, we identified several cases where the model incorrectly predicted normal samples as lesional but were correct on the microscopic level when reviewed by the pathologist. This tendency of histological presentation to be more severe than endoscopic presentation was previously published in the literature. In parallel, we utilised a model trained on the Colon Nuclei Identification and Counting (CoNIC) dataset to predict and explore 6 cell populations. We observed correlation between areas enriched with the predicted immune cells in biopsies and the pathologist's feedback on the attention maps. Finally, we identified several cell level features indicative of disease severity in CD and UC. These models can enhance our understanding about the pathology behind IBD and can shape our strategies for patient stratification in clinical trials.

Keywords: Weakly supervised learning, self-supervised learning, attention maps, IBD.

© 2023 Anonymous.

1. Introduction

Inflammatory bowel diseases (IBD) are chronic, relapsing-remitting inflammatory disorders of the gastrointestinal (GI) tract. Crohn's disease (CD) and Ulcerative colitis (UC) represent the two main types of IBD (Baumgart and Carding, 2007). Both result from a complex interplay of several factors, including abnormal immune responses, genetics, microbiome and environmental triggers (Endo et al., 2009; Baumgart and Carding, 2007). CD can affect any region of the gut, and is characterised by segmental mucosal ulceration, transmural inflammation, fissures, fibrosis, and stricture formation. In contrast, UC principally involves the colon, and manifests as continuous mucosal inflammation extending from the rectum proximally, with variable extent, and a more superficial inflammatory infiltrate. CD and UC share similar symptoms, but are pathophysiologically distinct diseases (Shanahan, 2001) however, in approximately 10% of cases the observed clinical and pathological features are insufficient to reliably distinguish between the two diagnoses, and such patients are diagnosed with IBD-unclassified (IBD-U).

IBD is diagnosed based on a combination of clinical presentation, radiographic, endoscopic and histopathological findings. Defining the extent and severity of inflammation in IBD can influence treatment decisions and support prediction of the patient's prognosis. Histopathology is a key tool in diagnosing IBD, assessing disease activity at the microscopic level, and identifying dysplasia (suboptimally controlled IBD carries a long-term risk of colorectal cancer), and is therefore evolving as a treatment target. Several histopathological indices have been developed for IBD (Reddy et al., 2019; Abbas et al., 2019; Popa et al., 2020), but their relative complexity hinders the development of AI algorithms without large-scale labour-intensive manual annotations by pathologists. We, therefore, aimed to develop tools to assist decoding the complex histopathological disease features for both CD and UC. This will inform understanding of disease pathology and patient stratification to support clinical trial recruitment strategies.

In this paper, we address the challenging problem of whole slide image (WSI) classification applied to weakly-labeled IBD patients' biopsies. We demonstrate that applying self-supervised learning coupled with weakly-supervised learning to Haematoxylin & Eosin (H&E)-stained WSIs can accurately identify disease type, macroscopic tissue appearance, and endoscopic score in IBD. We trained two latest state-of-the-art architectures, including Dual-Stream Multiple Instance Learning (DSMIL) (Li et al., 2021a) and Hierarchical Image Pyramid Transformer (HIPT) (Chen et al., 2022), and explored two training strategies - fine-tuning and end-to-end training. The high attention regions identified by the models were confirmed through histopathological interpretations by a pathologist (H.A.) to contain epithelial and stromal morphological/structural features consistent with inflammation.

We further validated the performance and utility of these models by leveraging a model trained on the publicly available Colon Nuclei Identification and Counting (CoNIC) dataset (Graham et al., 2021). The model can segment and classify various types of cells in the biopsies that we used for comparison with the attention maps produced by the weakly-supervised models and eventually confirmed these with the pathologist. To our knowledge, this is the first work on exploring and validating weakly-supervised methods to associate endoscopic appearance with H&E morphology for large-scale IBD biopsies.

2. Related work

WSI classification using only slide-level labels is commonly framed as a multiple instance learning (MIL) problem (Augustine, 2022) in which image patches are mapped to fixed-length embeddings followed by aggregation by an operator, for example max-pooling. However, these MIL methods suffer from several limitations that have been recently addressed by the research community. For example, MIL methods are prone to misclassifications in the case of an unbalanced number of positive instances when using a simple aggregation operation such as max-pooling (Li et al., 2021a). This can be addressed by utilising attention-based aggregation where each instance is given a specific attention weight during training (Tomita et al., 2019). Additionally, MIL methods can ignore the correlation among different instances across the WSI, which can be mitigated by using transformers to consider morphological and spatial information (Shao et al., 2021) as well as generating attention maps for improved interpretability. Attention map interpretability was further improved by incorporating non-local attention into a dual-stream MIL (DSMIL) architecture to calculate an attention weight for each patch (Li et al., 2021a). MIL models often use fixed patch-based features extracted by a CNN or only features extracted by a fine-tuned model since end-to-end training is expensive and time-demanding for large slides. These sub-optimal extracted features can be improved by incorporating a multi-resolution feature fusion mechanism to leverage varied-size tissue features such as glands vs cells (Li et al., 2021b). Recently, HIPT (Chen et al., 2022) extended multi-resolution feature fusion by introducing hierarchical pre-training, enabling a bottom-up aggregation of tissue features from cells to tissue morphology.

Following these advancements, we leverage DSMIL (Li et al., 2021a) and HIPT (Chen et al., 2022) due to their superior performance over the aforementioned methods on large public H&E classification datasets such as The Cancer Genome Atlas (TCGA). DSMIL achieves strong performance by combining self-supervised pre-training using SimCLR (Chen et al., 2020) with a dual-stream attention-based MIL aggregator. Since both local and global contexts are important for WSI classification, DSMIL takes patches extracted at multiple magnifications with the same resolution as input. This is disadvantageous since fine details are lost at low magnification. By contrast, HIPT extracts patches at a fixed high magnification but with a relatively large resolution, such that fine and coarse-grained details are captured at the same resolution.

In this paper, we predict clinically relevant IBD endoscopic categories including disease types, macroscopic appearances, and endoscopic scores from H&E biopsies which, to our knowledge, has not been addressed previously. Prior work mostly focused on electronic health records (Reddy et al., 2019), genomics, metagenomics (Abbas et al., 2019), biological and clinical parameters for predicting endoscopic scores in UC patients (Popa et al., 2020). Recently a fully supervised approach using H&E images to predict remission and the Nancy Histological Index scores for UC was investigated by Najdawi et al. (2022). In contrast, we consider both CD and UC and use weakly supervised methods with macroscopic patient level labels, without any detailed pathologist annotations at the microscopic level, which are inexpensive relative to the supervised method.

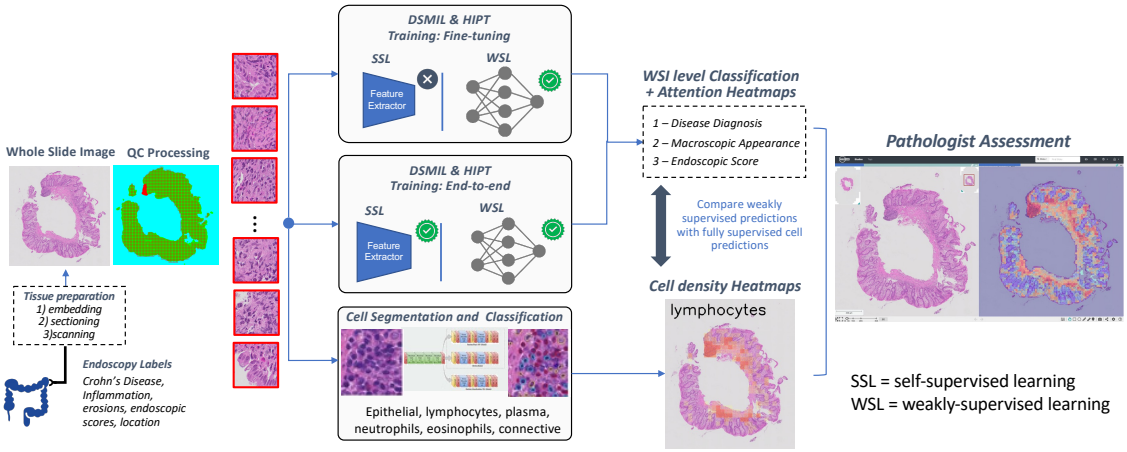

Figure 1: Overview of the implemented pipeline.

3. Methodology

The proposed methodology is summarised in Figure 1. First, we used an image quality control (QC) algorithm and extracted patches for training DSMIL and HIPT to predict 3 clinically relevant endoscopic categories and generate visual attention maps. A cell classification model was also applied to all slides and the individual cell predictions were aggregated into heatmaps for comparison with the weakly-supervised attention maps. Finally, attention maps and cell heatmaps were visualised in HALO Link for discussion with the pathologist. More details on these steps are in the subsections below.

3.1. Patch Extraction and QC

To process gigapixel WSIs using deep learning models, we first split each WSI into many small, non-overlapping image patches using the Histolab (Marcolini et al., 2022) and CLAM (Lu et al., 2021) Python packages. We trained an image quality control (QC) model based on DenseNet (Huang et al., 2017) to automatically identify and remove regions from the WSIs with imaging and/or tissue artefacts such as overstaining, tissue folds, debris, out-of-focus areas and variations in contrast and hue markings, that were introduced during slide preparation and digitisation. The QC model provided readouts at the slide-level, identifying how much of the slide was rejected as a proportion of the tissue. These readouts were used to exclude slides containing large areas of rejected tissue (e.g. >50%). The QC model was also integrated into the patch extraction pipeline to save individual patches containing >50% accepted tissue (see Appendix A).

3.2. Self-Supervised Pre-training and Weakly-Supervised Classification

In the following sections, we briefly outline how we distinguish self-supervised pre-training and weakly-supervised classification in DSMIL and HIPT - further details are provided in Section 4 and Appendix A.

Self-supervised Pre-training. We only consider the self-supervised representation learning components of DSMIL and HIPT. For DSMIL, this involves the SimCLR component, while for HIPT, this involves the $\text{ViT}_{16} - 256$ and $\text{ViT}_{256} - 4096$ components, which learn to output embeddings at the cell and patch-level, respectively. These self-supervised components, once trained, are used to generate patch embeddings for all WSIs, which are subsequently used for training the weakly-supervised classification components.

Weakly-supervised classification. We only consider the weakly-supervised components of DSMIL and HIPT for predicting 3 endoscopic categories. For DSMIL the attention-based MIL aggregator, and for HIPT the $\text{ViT}_{4096} - \text{WSI}$ component were considered.

3.3. Interpretation of Model Predictions

Following weakly-supervised training, we generated visual attention maps, which facilitated optimal collaboration with the pathologist to interrogate the models' predictions.

DSMIL attention maps are visualised using the per-patch attention weights of MIL aggregator as a heatmap overlay on the WSI (Li et al., 2021a). A weight value close to 1.0 indicates that the patch contributes heavily to the final prediction compared with a patch that has a score close to 0.0. We generated these attention maps for all WSIs in our dataset.

HIPT attention maps are a natural part of HIPT due to its transformer backbone (Chen et al., 2022). The attention maps in the original paper demonstrated that they can highlight unique cancer-relevant tissue morphologies. Therefore, we also sought to understand what IBD-relevant tissue morphologies HIPT could learn.

HALO Link Integration was used to make the attention map review process interactive and straightforward for the pathologist. We integrated the generated attention maps with HALO Link (Labs, 2022). This allowed the pathologist to view the attention maps and WSIs side by side, provide comments and make annotations. We asked the pathologist to describe the critical histopathological features for each slide including identification of inflammatory regions cellular composition, and morphological/structural features of the epithelium and stroma. We compared these comments with the attention maps to interpret features learnt by the models.

A **cell prediction model** for detection and classification of six cell types trained on the publicly available annotated CoNIC H&E dataset (Graham et al., 2021) was used for understanding the cell populations. The CoNIC dataset was part of a grand medical challenge with the aim of predicting six types of cells on H&E slides, including epithelial cells, neutrophils, lymphocytes, plasma cells, connective cells, and eosinophils. We aggregated the cell predictions into statistics per patch and then visualised these as the heatmaps (Appendix E). The heatmaps were compared with the attention maps produced by DSMIL and HIPT and several representative examples were reviewed by the pathologist. Finally, several human interpretable features (HIFs) were calculated from the readouts and correlated with the endoscopic scores for both CD and UC - Appendix E.

4. Experimental Setup

We included 1397 H&E-stained biopsies from 418 CD and 218 UC patients enrolled in a multi-centred longitudinal Study of a Prospective Adult Research Cohort with IBD (SPARC IBD). The samples reside in the IBD Plexus database, provided with informed consent by

the Crohn's and Colitis Foundation (Crohn's and Foundation, 2022). The total tissue area in 1397 biopsies is ≈7500µm^2 whereas the TCGA-lung resection dataset used to train DSMIL in the original work (Li et al., 2021a) has a total area of ≈65000µm^2 across 1054 slides, making SPARC IBD a relatively challenging dataset for classification.

We used 3 clinically-relevant labels from the SPARC IBD dataset that were acquired from endoscopy to define weakly-supervised classification tasks. These included: i) disease diagnosis (CD vs UC), ii) macroscopic tissue appearance (normal vs lesional), and iii) CD and UC-specific endoscopic severity scores (low vs high score) - see Appendix A. We used 4 NVIDIA A100 GPUs for training.

The QC model was trained on 19 WSIs with patches of 1280x1280 pixels at 40x and was then applied to all SPARC IBD slides. We excluded 11 slides with >50% rejected tissue and then DSMIL and HIPT were trained on patches extracted at 40x (Appendix A).

In training, we initially used DSMIL and HIPT models, which were pre-trained on the TCGA dataset and fine-tuned on our dataset, in order to assess the transferability of knowledge from TCGA cancer resections to IBD biopsies. In fine-tunning, the weights of the self-supervised components of both models (SimCLR in DSMIL and ViT_{16-256} & $\text{ViT}_{256-4096}$ in HIPT) were frozen and only the weakly-supervised classification components (MIL aggregator in DSMIL and $\text{ViT}_{4096-\text{WSI}}$ in HIPT) were trained to predict the weakly-supervised classification tasks in SPARC IBD. We labelled the models that were fine-tuned on SPARC IBD as DSMIL-F and HIPT-F.

We also explored training DSMIL and HIPT end-to-end on SPARC IBD. The self-supervised and weakly-supervised components of both models were trained on SPARC IBD from scratch to compare with the performance of fine-tuned models mentioned previously. We labeled the models that were trained end-to-end on SPARC IBD as DSMIL-E2E and HIPT-E2E. See Appendix A for more information on the experimental designs and intermediate results. In self-supervised and weakly-supervised training, we closely followed the training settings in the original papers (Li et al., 2021a; Chen et al., 2022).

In weakly-supervised training, we performed 5-fold cross validation, stratified on patients and ensuring that distributions of disease diagnosis, macroscopic appearance, biopsy location and the target label were consistent across all training and testing splits (more details in Appendix A). The same cross-validation splits were used for both DSMIL and HIPT models in all experiments. We used area under the receiver-operator characteristic curve (AUROC) to measure predictive performance. For further evaluation, the visual attention maps for all slides were generated and compared with the cell prediction heatmaps from the CoNIC model. Finally, the attention maps were shared with the pathologist via the HALO Link tool to verify if the models focused on disease-relevant features.

5. Results and Discussion

Model Performance. In Table 1 we compare the performance of end-to-end and fine-tuned DSMIL and HIPT models for predicting 3 weakly-supervised tasks in SPARC IBD. HIPT-E2E significantly outperforms DSMIL-E2E across all tasks (two-tailed t-test: diagnosis - $p<0.0001$, macroscopic appearance - $p<0.0005$, endoscopic scores (CD) - $p<0.05$, endoscopic scores (UC) - $p<0.0001$). Furthermore, we found that E2E training on SPARC IBD was the optimal strategy for both DSMIL and HIPT training across all tasks. Po-

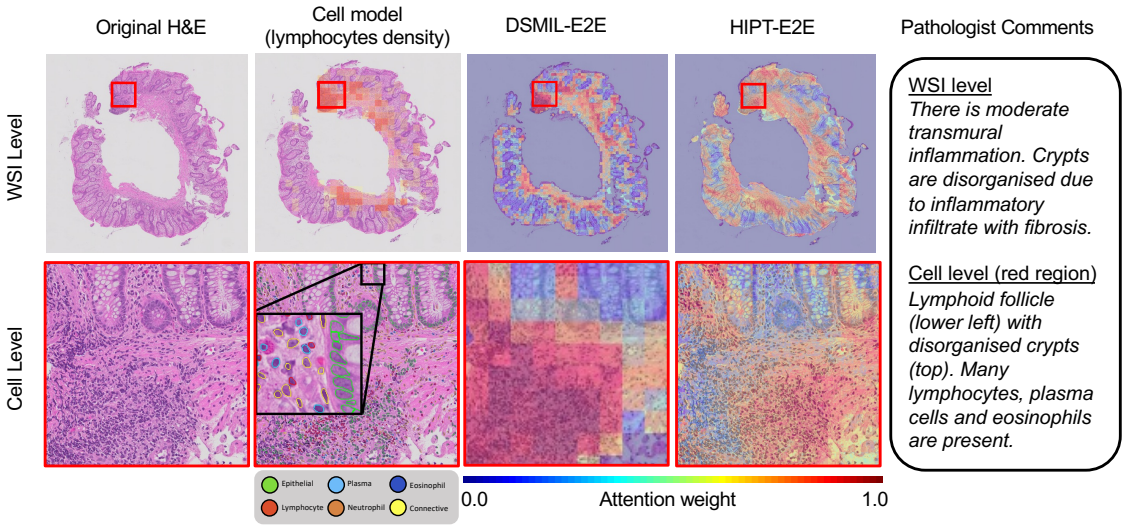

Figure 2: H&E image, lymphocytes density by the cell level model, attention maps from two models, at WSI level and cell level resolutions, and pathologist comments.

Table 1: Mean AUROC ± standard error (5-fold cross validation) of trained models across different prediction tasks.

Model	**Disease Diagnosis**	**Macroscopic Appearance**	**Endoscopic Score (CD)**	**Endoscopic Score (UC)**
DSMIL-F	0.656±0.007	0.522±0.008	0.528±0.016	0.592±0.007
DSMIL-E2E	0.692±0.010	0.750±0.006	0.740±0.009	0.634±0.017
HIPT-F	0.825±0.017	0.780±0.012	0.766±0.026	0.788±0.034
HIPT-E2E	**0.865±0.019**	**0.814±0.008**	**0.786±0.017**	**0.802±0.014**

Table 2: Performance of DSMIL-E2E and HIPT-E2E trained on clinically-relevant subsets of SPARC IBD for predicting macroscopic appearance.

Model	**All data**	**Just CD**	**Just UC**	**Just Ileum**	**Just Colon**
DSMIL-E2E	0.750±0.006	0.689±0.010	0.808±0.013	0.586±0.010	0.775±0.019
HIPT-E2E	0.814±0.008	0.804±0.010	0.837±0.015	0.739±0.032	0.823±0.036

tentially, the embeddings learned by E2E models through self-supervised pre-training on SPARC IBD were more useful for downstream prediction tasks than pre-trained embeddings

from TCGA. This difference can be attributed to the differences between IBD and cancer morphologies as well as the difference between TCGA resections vs IBD biopsies.

Attention Maps. Both DSMIL and HIPT generated informative attention map overlays that could reveal disease relevant features in IBD (Figure 2). With the pathologist's support, we confirmed numerous cases where both DSMIL-E2E and HIPT-E2E accurately and precisely localised areas of inflammation in the tissue (Figure 2 and Appendix C).

Cell Prediction Model. The cell-based model correctly identifies multiple cell types at the WSI level, which we confirmed with the pathologist for several WSIs (for example Figure 2). We also explored HIFs (Appendix E) and observed that neutrophil features in UC along with eosinophil and lymphocyte features were the most indicative of the disease severity, which is reflected in the literature (Alhmoud et al., 2020). However, the cell level correlations must be explored further to understand the model's quantitative performance. As this is a labor-intensive and time-consuming task for pathologists, we suggest combining weakly-supervised learning with supervised models is a more efficient approach.

Learning histological features with endoscopy labels. We demonstrated that endoscopic labels can allow H&E models to learn disease features at the microscopic level (Table 1). Then we explored both false positive and false negative WSIs from DSMIL-E2E in predicting macroscopic appearance. The WSIs were evaluated by the pathologist blinded to the macroscopic classifications who identified several cases where slides classified as "normal" in endoscopy labels had extensive inflammation at the microscopic H&E level (Figure 6). This "misclassificaiton" corresponds to a previously observed tendency of histological presentation in UC to be more severe than endoscopic presentation (Mitsuishi, 2019). DSMIL-E2E was still able to localise areas of inflammation in these slides where endoscopic and histopathology finding determined by the pathologist differed.

Applications. The trained models can be used to automatically rank biopsies by disease severity allowing pathologists to save time by prioritising biopsies with higher disease activity. This is an important consideration since the majority of biopsies received by pathologists are within normal limits. For biopsies that are not within normal limits, the attention maps can be used for guiding the pathologist during microscopic evaluation of potentially malignant areas. For clinical applications, we explored training DSMIL and HIPT using only CD or only UC biopsies as well as exclusively ileum or colon biopsies for predicting macroscopic appearance since disease features can present differently within these subsets (Table 2). We found that model performance can be improved by training on just UC and just colon subsets, while performance decreases when training on just ileum, suggesting that this is a more difficult task due to fewer ileum biopsies. We consider these types of modifications critical in applying these models to clinical trials and will be further explored in future works.

6. Conclusion and Future Work

We demonstrated that weakly supervised learning can produce accurate models trained on H&E biopsies with macroscopic labels only. We observed potential in the publicly available CoNIC dataset for IBD data. However, further quantitative and qualitative assessment of its generalisability is needed. We plan to validate the trained SPARC IBD models with the corresponding attention maps on IBD clinical trial datasets as well as enrich the training dataset with more IBD data.

Acknowledgments

Acknowledgments withheld.

References

Mostafa Abbas, John Matta, Thanh Le, Halima Bensmail, Tayo Obafemi-Ajayi, Vasant Honavar, and Yasser El-Manzalawy. Biomarker discovery in inflammatory bowel diseases using network-based feature selection. *PloS one*, 14(11):e0225382, 2019.

Tarik Alhmoud, Anas Gremida, Diego Colom Steele, Imaneh Fallahi, Wael Tuqan, Nina Nandy, Mahmoud Ismail, Barakat Aburajab Altamimi, Meng-Jun Xiong, Audra Kerwin, and David Martin. Outcomes of inflammatory bowel disease in patients with eosinophil-predominant colonic inflammation. *BMJ Open Gastroenterology*, 16(7):1–7, 2020.

Tanya N. Augustine. Weakly-supervised deep learning models in computational pathology. *eBioMedicine*, 81:104117, 2022.

Daniel C Baumgart and Simon R Carding. Inflammatory bowel disease: cause and immunobiology. *The Lancet*, 369(9573):1627–1640, 2007.

Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Crohn's and Colitis Foundation. SPARC IBD, 2022. URL https://www.crohnscolitisfoundation.org/research/current-research-initiatives/ibd-plexus/publications.

Katsuya Endo, Hisashi Shiga, Yoshitaka Kinouchi, and Tooru Shimosegawa. Inflammatory bowel disease: Ibd. *Rinsho byori. The Japanese journal of clinical pathology*, 57(6):527–532, 2009.

Simon Graham, Mostafa Jahanifar, Quoc Dang Vu, Giorgos Hadjigeorghiou, Thomas Leech, David Snead, Shan E Ahmed Raza, Fayyaz Minhas, and Nasir Rajpoot. Conic: Colon nuclei identification and counting challenge 2022. *arXiv preprint arXiv:2111.14485*, 2021.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Indica Labs. HALO Link, 2022. URL https://indicalab.com/halolink/.

Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021a.

Jiayun Li, Wenyuan Li, Anthony Sisk, Huihui Ye, W Dean Wallace, William Speier, and Corey W Arnold. A multi-resolution model for histopathology image classification and localization with multiple instance learning. *Computers in biology and medicine*, 131: 104253, 2021b.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

Alessia Marcolini, Ernesto Arbitrio, and Nicole Bussola. Histolab, 2022. URL https://github.com/histolab/histolab.

Takehiro Mitsuishi. Correlation between histological findings and endoscopic findings in patients with ulcerative colitis: Basal plasmacytosis is an important finding suggesting active inflammation. *JGH Open*, 3(2):100–104, 2019. doi: https://doi.org/10.1002/jgh3.12111. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jgh3.12111.

Fedaa Najdawi, Kathleen Sucipto, Pratik Mistry, Stephanie Hennek, Christina Jayson, Mary Lin, Darren Fahy, Shawn Kinsey, Ilan Wapinski, Andrew H Beck, et al. Artificial intelligence enables quantitative assessment of ulcerative colitis histology. *medRxiv*, 2022.

Iolanda Valentina Popa, Alexandru Burlacu, Catalina Mihai, and Cristina Cijevschi Prelipcean. A machine learning model accurately predicts ulcerative colitis activity at one year in patients treated with anti-tumour necrosis factor α agents. *Medicina*, 56(11):628, 2020.

Bhargava K Reddy, Dursun Delen, and Rupesh K Agrawal. Predicting and explaining inflammation in crohn's disease patients using predictive analytics methods and electronic medical record data. *Health informatics journal*, 25(4):1201–1218, 2019.

Fergus Shanahan. Inflammatory bowel disease: immunodiagnostics, immunotherapeutics, and ecotherapeutics. *Gastroenterology*, 120(3):622–635, 2001.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021.

Naofumi Tomita, Behnaz Abdollahi, Jason Wei, Bing Ren, Arief Suriawinata, and Saeed Hassanpour. Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. *JAMA network open*, 2(11):e1914645–e1914645, 2019.

Appendix A. Supplementary Methods

A.1. QC and Patch Extraction

From applying our QC model to all 1397 WSIs, we found that 6, 11, and 40 slides had, respectively, >75, >50, and >25% rejected areas (Figure 3). Following QC, patches were extracted at 40x for training DSMIL-E2E (224 by 244 resolution) and HIPT-E2E (4096 by 4096 resolution), following the methods of (Li et al., 2021a) and (Chen et al., 2022). 2,054,901 patches were extracted for DSMIL and 13,952 patches were extracted for HIPT.

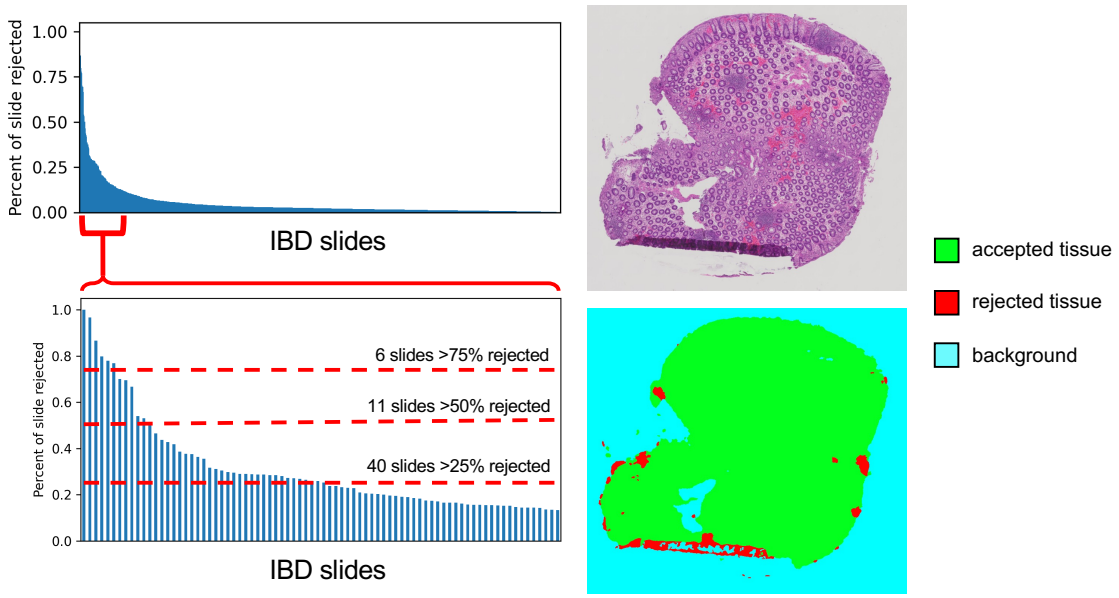

Figure 3: QC summary statistics on SPARC IBD and QC result on an additional slide

A.2. Cross-Validation

5-fold cross validation is performed on in all weakly-supervised classification tasks in SPARC IBD. Splits are performed at the patient level, maintaining the distributions of biopsy location, disease diagnosis and the target label - see Figure 4 for an example for macroscopic appearance. This was done to ensure that models are tested on a representative distribution of biopsies. The same cross validation splits are used for both DSMIL and HIPT experiments to allow them to be compared. For predicting macroscopic appearance, the "erosions/ulcers" and "inflammation" labels were combined into a "lesional" class. For endoscopic score, the median of SES score for CD and modified Mayo endoscopic score for UC were 0, so we treated the negative class as a score of 0 and positive class as score >0.

A.3. Model Training

The hyperparameters used in self-supervised pretraining and weakly-supervised training for all DSMIL and HIPT models are shown in Table 3. To follow closely the methods of (Li

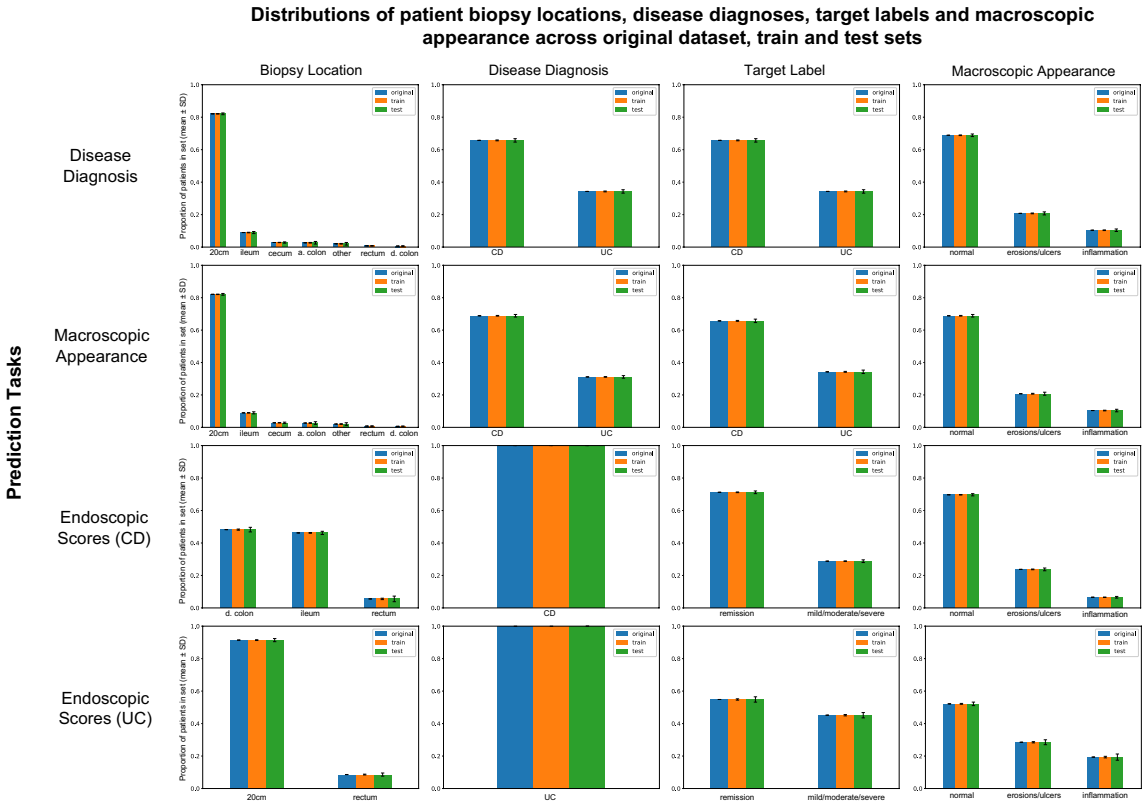

Figure 4: Proportion of patients (n=638) in original, train and test sets across 5 cross validation folds and across biopsy location, disease diagnosis and macroscopic appearance for all tasks. Proportions of patients in each category are kept consistent across all train and test splits.

et al., 2021a) and (Chen et al., 2022), we use the default parameters for both models across the board, but for HIPT we found that self-supervised pretraining converged after 30 epochs and in weakly-supervised learning we use only 1 transformer encoder layer. The same cross validation splits are used for both DSMIL and HIPT in all experiments.

Table 3: Hyperparameters used to train DSMIL and HIPT models.

Model	**Self-supervised pretraining**	**Weakly-supervised training**
DSMIL-F	Frozen from TCGA	epochs: 200; embedding size: 512; learning rate: 0.0002; weight decay: 0.005
DSMIL-E2E	epochs: 100; learning rate: 0.001; weight decay: 0.000001; batch size: 1024; arch: ResNet18	epochs: 200; embedding size: 512; learning rate: 0.0002; weight decay: 0.005
HIPT-F	Frozen from TCGA	epochs: 20; layers: 1; heads: 3; dropout: 0.25, learning rate: 0.0003
HIPT-E2E	epochs: 30; weight decay: 0.04; learning rate: 0.0005, warmup epochs: 10	epochs: 20; layers: 1; heads: 3; dropout: 0.25, learning rate: 0.0003

Appendix B. Additional Experiments

Since DSMIL can take as input patches extracted at different magnifications, we assessed the performance of DSMIL in predicting disease diagnosis with different magnification datasets. These results are summarised in Table 4. In order to be consistent with HIPT we used 40x patches to train DSMIL-E2E for all main experiments (Table 1). However, we find that for predicting macroscopic appearance, slightly improved performance can be achieved with lower magnifications, although there is no statistically significant difference between these models trained at different magnifications and HIPT-E2E still significantly outperforms DSMIL-E2E at all magnifications.

Table 4: Comparison of DSMIL-E2E trained on different magnifications in predicting macroscopic appearance.

Dataset	**AUROC ± 1 SE**
5x	0.645±0.008
10x	0.752±0.006
20x	0.758±0.006
40x	0.750±0.006

Appendix C. Additional Attention Maps and Pathologist Feedback

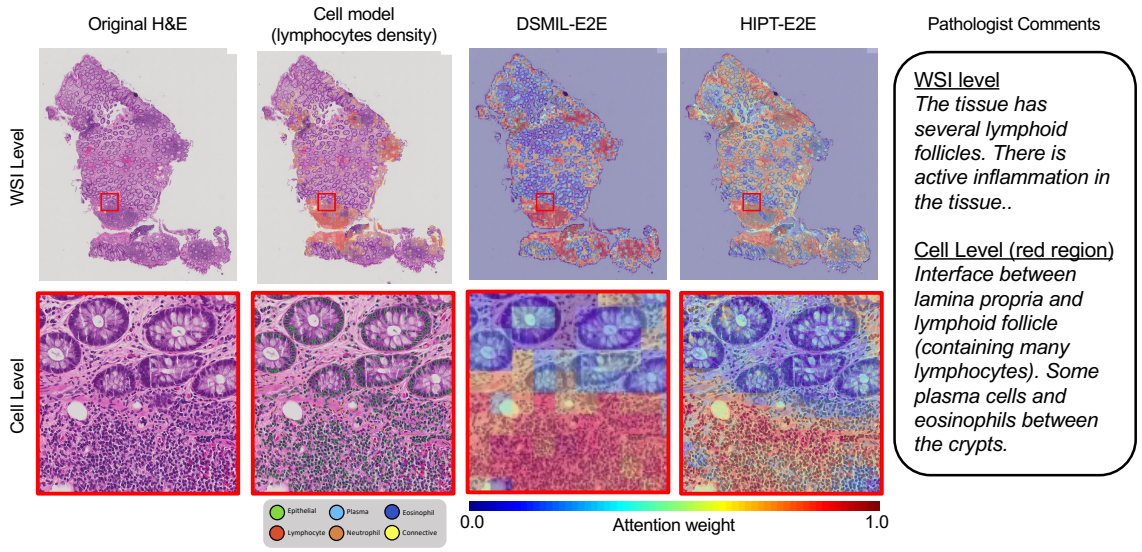

Figure 5: H&E image, Attention maps, cell level predictions and pathologist comments for UC patient with lesional macroscopic appearance and high endoscopic score.

Appendix D. Misclassified Cases

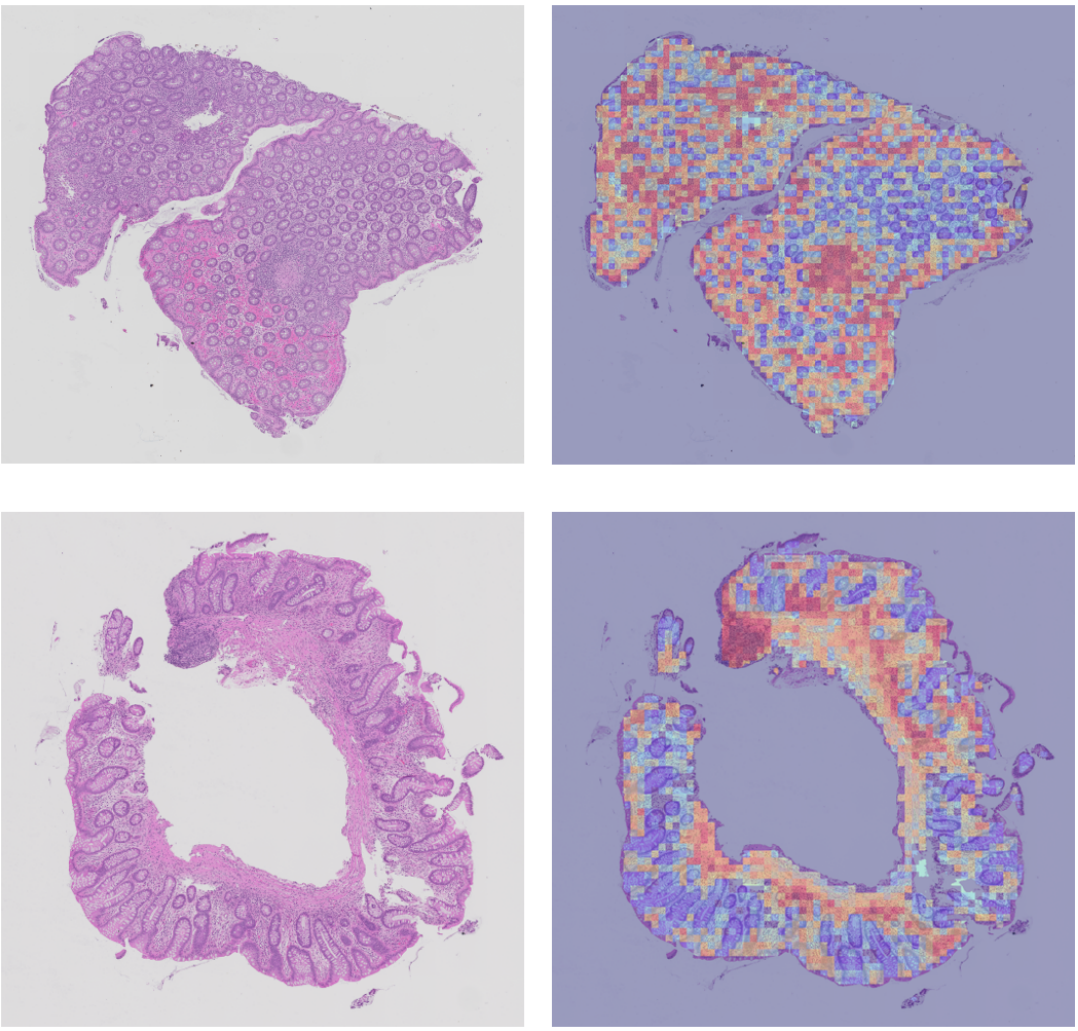

Figure 6: Two slides classified as "normal" from macroscopic appearance but predicted as "lesional" by DSMIl-E2E. Both slides were confirmed by the pathologist to contain inflammation and hence should be considered "lesional".

Appendix E. Supervised Cell Model

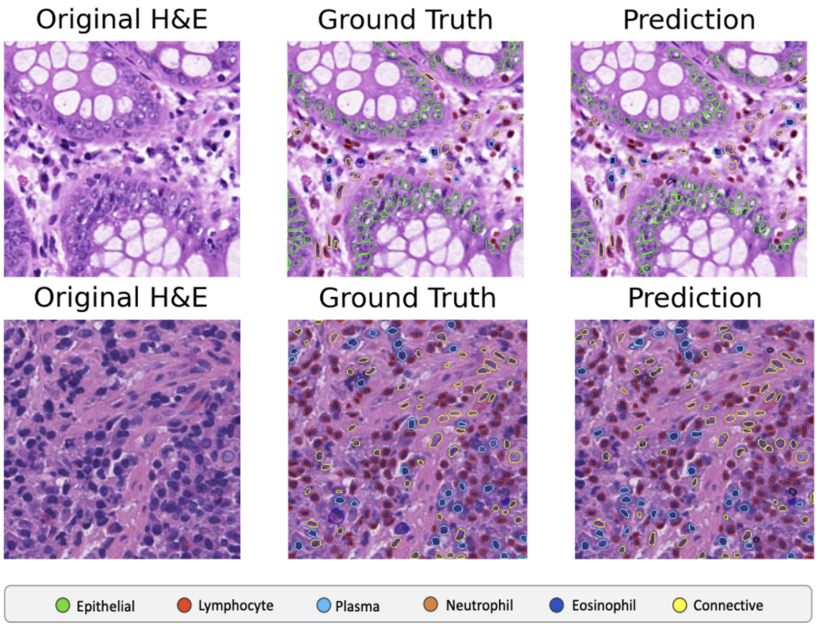

Figure 7: Six cell types predictions overlaid over two CoNIC patches.. CoNIC examples have ground truth, annotations by pathologists available in the dataset.

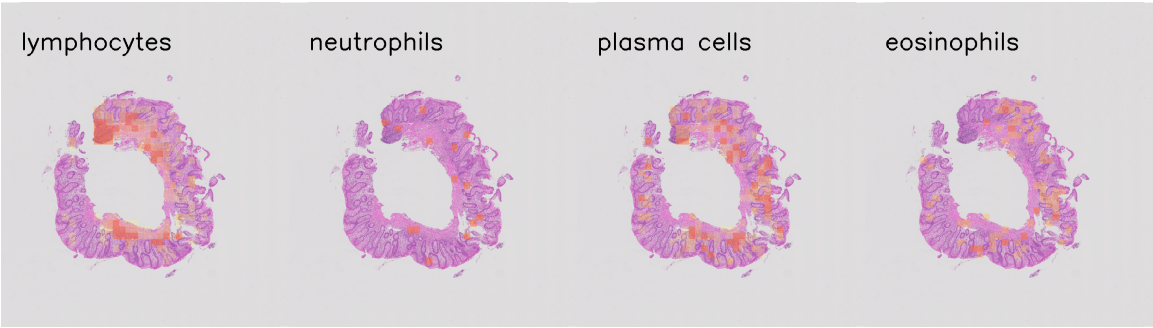

Figure 8: Four immune cell density heatmaps overlaid over an IBDplexus WSI.

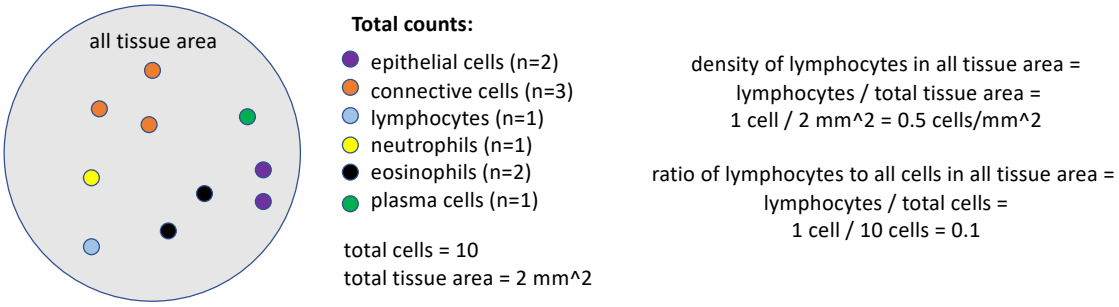

Figure 9: An example how human interpretable features features (HIFs) are calculated from six class cells predictions.

(*a*) Statistically significant HIFs for IBD Plexus UC patients.

Feature	**p value**
ratio of neutrophils to all cells in all area	0.0007
density of neutrophils in all tissue area	0.0022
density of eosinophils in all tissue area	0.0111
ratio of lymphocytes to all cells in all area	0.0125
density of lymphocytes in all tissue area	0.0150
ratio of eosinophils to all cells in all area	0.0187
ratio of epithelial cells to all cells in all area	0.0234
density of connective cells in all tissue area	0.0416

(*b*) Statistically significant HIFs for IBD Plexus CD patients.

Feature	**p value**
ratio of eosinophils to all cells in all area	0.00002
density of eosinophils in all tissue area	0.0001
ratio of neutrophils to all cells in all area	0.0018
density of neutrophils in all tissue area	0.0057
density of plasma cells in all tissue area	0.0170
ratio of lymphocytes to all cells in all area	0.0294
ratio of plasma cells to all cells in all area	0.0341

Table 5: HIFs for IBDplexus UC (a) and CD (b). The patients are split into two groups by Mayo endoscopic subscore of zero (n93) and more than zero (n84) for UC and by SES CD endoscopic subscore of zero (n193) and more than zero (n150).