

IMILIA: interpretable multiple instance learning for inflammation prediction in IBD from H&E whole slide images

Thalyssa Baiocco-Rodrigues*

THALYSSA.BAIOCCO-RODRIGUES@OWKIN.COM

Antoine Olivier*

ANTOINE.OLIVIER@OWKIN.COM

Reda Belbahri

REDA.BELBAHRI@OWKIN.COM

Thomas Duboudin

THOMAS.DUBOUDIN@OWKIN.COM

Pierre-Antoine Bannier

PIERRE-ANTOINE.BANNIER@OWKIN.COM

Benjamin Adjadj

BENJAMIN.ADJADJ@OWKIN.COM

Katharina Von Loga

KATHARINA.VONLOGA@OWKIN.COM

Nathan Noiry

NATHAN.NOIRY@OWKIN.COM

Maxime Touzot

MAXIME.TOUZOT@OWKIN.COM

Hector Roux de Bezieux

HECTOR.ROUX-DEBEZIEUX@OWKIN.COM

Owkin Inc., Paris, France

Editors: Under Review for MIDL 2026

Abstract

As the therapeutic target for Inflammatory Bowel Disease (IBD) shifts toward histologic remission, the accurate assessment of microscopic inflammation has become increasingly central for evaluating disease activity and response to treatment. In this work, we introduce IMILIA (Interpretable Multiple Instance Learning for Inflammation Analysis), an end-to-end framework designed for the prediction of inflammation presence in IBD digitized slides stained with hematoxylin and eosin (H&E), followed by the automated computation of markers characterizing tissue regions driving the predictions. IMILIA is composed of an inflammation prediction module, consisting of a Multiple Instance Learning (MIL) model, and an interpretability module, divided in two blocks: HistoPLUS, for cell instance detection, segmentation and classification; and EpiSeg, for epithelium segmentation. IMILIA achieves a cross-validation ROC-AUC of 0.83 on the discovery cohort, and a ROC-AUC of 0.99 and 0.84 on two external validation cohorts. The interpretability module yields biologically consistent insights: tiles with higher predicted scores show increased densities of immune cells (lymphocytes, plasmocytes, neutrophils and eosinophils), whereas lower-scored tiles predominantly contain normal epithelial cells. Notably, these patterns were consistent across all datasets. Code and models to partially replicate the results on the public IBDColEpi dataset can be found at <https://github.com/owkin/imilia>.

Keywords: Multiple Instance Learning, Inflammatory Bowel Disease, Inflammation prediction, Interpretability, H&E, WSI, Epithelium segmentation, Cell segmentation and classification, Weakly-supervised learning, Histology.

1. Introduction

Background. Inflammatory Bowel Disease (IBD) is a lifelong, chronic inflammatory disorder of the gastrointestinal tract, manifesting as Crohn’s disease (CD) or ulcerative colitis (UC) (Graham and Xavier, 2020). IBD places significant strain on healthcare systems

* Contributed equally

(Burisch et al., 2025), affecting nearly 7 million patients worldwide (Feuerstein and Cheifetz, 2014; Molodecky et al., 2012; Ng et al., 2017), with established high burdens in Europe (2 million) and North America (1.5 million). While clinical symptoms, ranging from abdominal pain to extra-intestinal manifestations, are primary indicators, they are often non-specific as the disease course is characterized by periods of active flare-ups interspersed with remission. Consequently, diagnosis and management heavily rely on a combination of endoscopic and histological assessments (Angyal et al., 2025). Histopathological assessment of Hematoxylin and Eosin (H&E) stained slides is critical not only to confirm the diagnosis and exclude differentials (*e.g.*, granulomatosis) but also to grade severity. Interestingly, patients without lesions at endoscopy (endoscopic remission) might still harbor persistent microscopic inflammation associated with disease progression (Iacucci et al., 2021; Marchal-Bressenot et al., 2017). This has led to a shift in the therapeutic goal in IBD, beyond symptom control and towards “histological healing” (Angyal et al., 2025; Villanacci et al., 2025), as persistent microscopic inflammation is a strong predictor of relapse and colorectal cancer risk. However, the standard manual review of biopsy slides is time-consuming and dependent on the expertise of the reviewing pathologist (Feakins et al., 2023; Geboes et al., 2000). As the volume of biopsies increases, the need for automated decision-support tools becomes critical to ensure consistency and efficiency in patient care.

Related work. Modern computational pathology frameworks often rely on the combination of a feature extraction step using a foundation model (FM), and a downstream multiple instance learning (MIL) model. The FM maps small areas of tissue (or tiles) to a lower-dimensional representation space, and the MIL model further combines those representations to derive a slide-level prediction. While recent advances on FM development have yielded spectacular progress across a variety of downstream tasks (Chen et al., 2024), the majority of applications have remained limited to oncology. In the specific domain of IBD, recent research has focused on automating histological grading, with models developed to predict established severity scores such as the Nancy Histological Index (NHI), Geboes Score, or Robarts Histopathology Index (Iacucci et al., 2023; Peyrin-Biroulet et al., 2024; Rubin et al., 2024; Plattner et al., 2025; Rymarczyk et al., 2023). Beyond simple classification, efforts have also explored correlating histological features with endoscopic findings (Di Vincenzo et al., 2024; Elta et al., 1987) and quantifying specific cellular populations at sample level (Reigle et al., 2024). In (Ding et al., 2025) a graph neural network is introduced to predict WSI-level ulcer presence from H&E WSIs. Closely related to our work, (Mokhtari et al., 2024) proposed a MIL-based framework to predict disease type, macroscopic tissue appearance and endoscopic scores, while combining it with a cell detection model to derive interpretable insights. While these studies share similarities with our work, they differ primarily in the extent of their external validation, often reporting results in cross-validation only, and in their interpretability, which remains largely qualitative - relying on visual inspection of heatmaps and lacking quantitative analysis.

Contributions. In this work, we introduce IMILIA (Interpretable Multiple Instance Learning for Inflammation Analysis), an end-to-end interpretable MIL framework designed for the binary classification of IBD histology slides (inflamed vs. non-inflamed). We summarize our contributions as follows:

- We provide an extensive validation of IMILIA, using one large discovery cohort (over 3000 patients) and two external validation cohorts. To the best of our knowledge, this is the only validation of a deep learning model for IBD inflammation prediction on external datasets.
- We also provide a quantitative validation of the interpretability module, composed of HistoPLUS and EpiSeg in the context of IBD. Their combination allows to automatically derive advanced and potentially novel markers of the presence of inflammation, such as the localization of immune cells within the epithelium.

We release the code to reproduce the results on the public dataset with the Chowder and EpiSeg models (including trained Chowder) at <https://github.com/owkin/imilia>.

2. IMILIA

In this section, we introduce IMILIA and explain how a MIL model can be combined with an interpretability module to predict inflammation from H&E WSI. An overview of the method is represented in Figure 1.

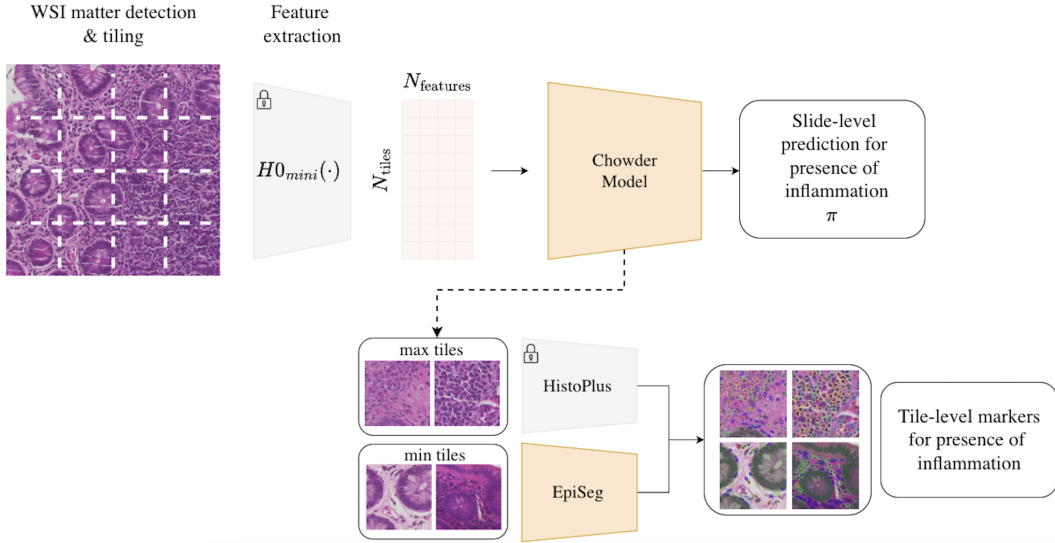


Figure 1: IMILIA: Interpretable Multiple Instance Learning for Inflammation Analysis. A multiple instance learning model is combined with two interpretability blocks to predict the presence of inflammation from H&E slides in IBD, and derive quantitative biological markers driving the predictions. Beige colors denotes models specifically trained for this paper.

WSI preprocessing. An in-house segmentation network is used to detect regions with relevant tissue, excluding background and acquisition artifacts. The tissue regions are then tessellated into non-overlapping tiles of fixed dimensions.

Feature extraction. Throughout this study, and for consistency, we use H0-mini (Filiot et al., 2025) as a feature extractor to transform tissue tiles into low-dimensional representations. We use H0-mini because it offers one of the best trade-offs between performance and computational efficiency, while exhibiting strong robustness to variations in scanning and staining conditions.

Chowder. Similar to other MIL architectures (Ilse et al., 2018), Chowder (Courtiol et al., 2020) identifies a WSI to a collection of tiles. Following the feature extraction step, each slide is therefore represented by a matrix $W \in \mathbb{R}^{n_{\text{tiles}} \times d}$, where n_{tiles} is the number of tiles in the slide and d is the dimension of the embeddings. A first 1D convolution layer maps W to a score vector $S \in \mathbb{R}^{n_{\text{tiles}} \times 1}$. Then, extreme tiles are selected by keeping only the r max and min scores in S , resulting in a vector $S_{\text{extr}} \in \mathbb{R}^{(2 \times r) \times 1}$. A final MLP is then applied to produce the model’s prediction. In this study, we use a multi-channel version of Chowder, where the first 1D convolution layer is replaced by K 1D convolution layers that operate in parallel, resulting in a vector $S_{\text{extr}} \in \mathbb{R}^{(2 \times r \times K) \times 1}$.

Selection of extreme tiles. As a by-product of its training process, Chowder learns tile score representations S , where tiles with maximal scores are associated with a positive prediction and tiles with minimal scores are associated with a negative prediction. In the rest of this paper, such tiles are referred to as max and min tiles. At inference, scores are computed for all slides, allowing to compare the content of tiles providing positive evidence of inflammation (higher scores) versus tiles providing negative evidence (lower scores).

HistoPLUS. As a first block of the interpretability module, we use HistoPLUS (Adjadj et al., 2025), a cell detection model trained on a pancancer dataset covering 6 indications, and supporting 13 different cell types, and consisting of a CellViT (Hörst et al., 2024) model. While we note that HistoPLUS was originally trained on oncology WSIs, we provide quantitative evidence in Section 5 of its transfer performance to IBD. Following the extreme tiles selection, HistoPLUS can be applied to derive tile-level cell types predictions.

EpiSeg. EpiSeg is the second block of the interpretability module, an epithelium segmentation model, trained on top of H0-mini’s representations of patch tokens. Given a tile of size $W \times H$, and a patch size P , the tile is divided into $W/P \times H/P$ small patches. Following the H0-mini inference, each patch is then represented by a d -dimensional vector, or patch embedding. A logistic regression is then trained to predict the presence of epithelium at patch level, resulting in segmentation maps of size $W/P \times H/P$. Similar to HistoPLUS, EpiSeg can then be applied on the extreme tiles of Chowder to produce coarse segmentation maps of the presence of epithelium.

3. Data

For this study, we use three datasets: the SPARC IBD (Raffals et al., 2021) dataset as our discovery cohort, the FINBB and IBDColEpi (Pettersen et al., 2022) datasets as our two external validation datasets for the inflammation prediction module. IBDColEpi was used as training for the EpiSeg model.

SPARC IBD. The Study of a Prospective Adult Research Cohort with IBD (SPARC IBD) is a multicentered longitudinal study of adult IBD patients. We included 3322 H&E

slides obtained from intestinal mucosal biopsies of patients to compose our discovery cohort, among which 50.2% correspond to patients diagnosed with CD and 24.8% with UC (missing diagnostic information for other samples). 67.8% of samples were extracted from the colon and 26.9% from the ileum (missing location information for other samples). Ground-truth labels for inflammation presence were derived from the macroscopic tissue appearance of slides, assessed by a pathologist and originally categorized into 1) normal, 2) possible inflammation and 3) severe inflammation, with presence of erosion or ulcers. We grouped categories 2 and 3 to compose the “inflamed” class (the positive class used for training). The prevalence of inflamed samples is 31% ($N = 1022$).

FINBB. This dataset has been provided by the Finnish Biobank Cooperative (FINBB) and includes comprehensive multimodal data collected from Helsinki Biobank and Finnish Clinical Biobank Tampere. We used a total of 314 H&E slides from this dataset as external validation, among which 59.2% corresponded to patients diagnosed with UC and 40.8% with CD. 76.4% samples were extracted from colon, 20% from ileum and 1.6% from the intersection ileum/colon (other samples were classified as “unclear”). As for SPARC IBD, labels were derived from the macroscopic tissue appearance of slides, categorized into 1) normal and 2) inflamed. We note the prevalence shift compared to the discovery cohort, with inflamed samples accounting for 95% of the population.

IBDColEpi. The dataset consists of 140 H&E WSIs from biopsies of colonic mucosa of active and inactive IBD with pixel-level annotation of the epithelium. “Active” is defined as the presence of intraepithelial granulocytes in one or more locations in the biopsies. Still, the changes may be focal, hence the majority of the epithelium may still lack intraepithelial granulocytes or other signs of active disease (crypt abscesses, granulation tissue, etc.). For external validation, we use 132 H&E slides (8 were excluded due to insufficient tissue area), with a prevalence of 41% for positive (active disease) samples. No information is provided regarding categorization of samples into UC vs CD. IBDColEpi is publicly available¹.

4. Experimental setup.

Chowder implementation. For the MIL model, a multi-channel Chowder model with 5 channels was implemented. The number of selected extreme tiles per channel is 50, 25 with maximal scores and 25 with minimal scores, resulting in 250 extreme tiles. A final MLP with 2 hidden layers maps the tiles scores to the final prediction. The model is trained on the SPARC IBD dataset with a 5-fold cross-validation, using the Adam optimizer, and dropout is applied to the linear layers with probability 0.5. At inference, on the external validation cohorts, the 5 models are ensembled by averaging their predictions (including for the min and max tile scores). We provide in Appendix A an extensive list of the hyperparameters used in the training process.

For each cohort, two subsets of 1000 tiles, one with maximal and the other with minimal scores, are selected and processed by the two interpretability models, EpiSeg and HistoPLUS.

1. <https://www.kaggle.com/datasets/henrikpe/251-he-cd3-wsis-annotated-epithelium-ibdcolepi>

EpiSeg training. We leverage the epithelium annotations in IBDColEpi, and start from images and their corresponding pixel-level segmentation masks of epithelium of the same size. We note that the 224×224 tiles used to train the MIL model are too small to correctly predict the presence of epithelium, therefore, and to increase the context, we trained EpiSeg on tiles of size 1022×1022 , which simply corresponds to resampling the images available in the dataset to a spatial resolution of 0.5 micrometer per pixel (mpp). H0-mini operates by dividing images into small patches of constant size 14×14 . Following the forward pass in H0-mini, each patch is associated with an embedding x_p of dimension $d = 768$. To obtain lower resolution segmentation masks, we simply apply a 14×14 convolution filter with stride 14 and constant weights (and produce the equivalent of an average-pooling layer over patches). At this stage, each patch is associated with a continuous label y_p which can be interpreted as the area of epithelium within this patch. A logistic regression is finally trained on pairs (x_p, y_p) . A 3-fold cross-validation is performed to optimize the $L2$ -regularization parameter C in the model, which yields an optimal value $C = 10^{-2}$ used to train the final model. We use the train / test split provided by the authors of the dataset to train and evaluate our model. Illustrating samples can be found in Figure 2.

EpiSeg inference on extreme tiles. At inference time, for consistency in the patches representations by H0-mini, we ensure images have the same size 1022×1022 as was used during the training of EpiSeg. We thus extract expanded versions of the minimal and maximal tiles to match this size, with the tile at its center. The expanded image is used for EpiSeg inference and the final epithelium mask for the tile of interest is obtained by cropping the center of the EpiSeg’s output mask.

Interpretable features computation. For each max and min tiles, HistoPLUS and EpiSeg were used to generate cell masks (one for each cell type) and epithelium masks, respectively. They are further combined to compute the density ρ^c of several cell types c within the epithelium region, as:

$$\rho^c := \frac{\sum_{k=0}^{N^c} E(c_k)}{\sum_{j=0}^H \sum_{i=0}^W E(x_i, y_j) \times \text{mpp}_x \times \text{mpp}_y} \quad (1)$$

where, for a pixel location (x, y) , $E(x, y)$ is one if (x, y) falls within the predicted epithelium region; N^c is the total number of instances for the cell type c , H and W are the height and width of the epithelium mask, $c_k := (c_{k,x}, c_{k,y})$ is the centroid of the k -th cell instance, and $(\text{mpp}_x, \text{mpp}_y)$ is the pixel size.

5. Results

IMILIA shows strong cross-validation and transfer performance on the three cohorts. IMILIA’s classification module (Chowder model) shows a ROC-AUC of 0.83 in cross-validation in the discovery dataset SPARC IBD. It shows a great transferability in external datasets, reaching an AUC of 0.84 on FINBB. Remarkably, the model achieved near-perfect classification on the IBDColEpi dataset, with an AUC of 0.99. We attribute this to the curated nature of the dataset, which likely represents distinct inflammatory phenotypes with minimal “grey-zone” cases.

HistoPLUS demonstrates robust transfer performance in a new therapeutic area, beyond oncology. To assess the generalizability of HistoPLUS beyond oncology, and its suitability for IBD, we validated it in 90 tiles extracted from the SPARC IBD dataset, annotated with 13291 cells overall. Originally developed for oncology applications, the discrimination between cancer and healthy epithelial cells was part of HistoPLUS’ training. However, its training dataset may not adequately represent the full spectrum of epithelial morphologies, in particular in the heavily immune-infiltrated setting characteristic of IBD where epithelial cells frequently exhibit morphological irregularities that can resemble neoplastic features. Given these considerations, we elected to remap all predicted cancer cells to the epithelial class.

We use the same notation as in the HistoPLUS original paper, and denote by HistoVAL ($N = 530$) the validation set of HistoPLUS in the oncology setting. Table 1 and Table 3 summarize the model’s performances, providing a direct comparison between SPARC IBD and HistoVAL. HistoPLUS exhibited consistent validation performances on IBD, achieving a detection quality score of 0.774 (95%CI, [0.760;0.789]) on the IBD samples, and a segmentation quality score of 0.755 (95%CI, [0.749;0.761]). For cell classification, HistoPLUS demonstrated good capabilities on epithelial cells, immune cells, endothelial cells and fibroblasts with no significant performance drop between oncology and IBD.

Table 1: Cell classification performance of HistoPLUS in HistoVAL (oncology) and SPARC IBD. We report mean values of F1 scores and confidence intervals at 95% level, obtained by bootstrapping with 1000 repeats.

| Dataset | Epith. | Lymph. | Plasm. | Eosin. | Neutro. | Endoth. | Fibro. |
|----------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| HistoVAL | 0.42 (0.28;0.54) | 0.42 (0.28;0.54) | 0.48 (0.43;0.53) | 0.46 (0.37;0.52) | 0.24 (0.18;0.31) | 0.33 (0.29;0.38) | 0.38 (0.35;0.40) |
| IBD | 0.70 (0.67;0.73) | 0.53 (0.48;0.57) | 0.52 (0.46;0.57) | 0.45 (0.40;0.51) | 0.36 (0.29;0.42) | 0.44 (0.36;0.51) | 0.33 (0.27;0.39) |

EpiSeg, a logistic regression trained on top of H0-mini’s representations can efficiently predict the presence of epithelium. We use the test set of IBDColEPI to evaluate the logistic regression. It reaches an average precision score (defined as the area under the precision-recall curve) of 0.98 to predict the presence of epithelium at patch level. Qualitative examples of EpiSeg’s predictions are displayed on Figure 2, and the full precision-recall curve can be found in Appendix C.

As an additional validation of EpiSeg, we found that HistoPLUS and EpiSeg show a strong agreement regarding the prediction of epithelium. The pearson correlation between the number of epithelial cells identified by HistoPLUS in a tile and the epithelium area from EpiSeg’s predictions in the same tile is 0.85 ($p < 10^{-8}$) for SPARC IBD, 0.74 ($p < 10^{-8}$) for FINBB and 0.83 ($p < 10^{-8}$) for IBDColEpi.

We also note that EpiSeg’s performance provides strong evidence of the quality of H0-mini’s representations in the context of IBD, since a simple logistic regression model is able to learn how to discriminate patches corresponding to epithelium zones.

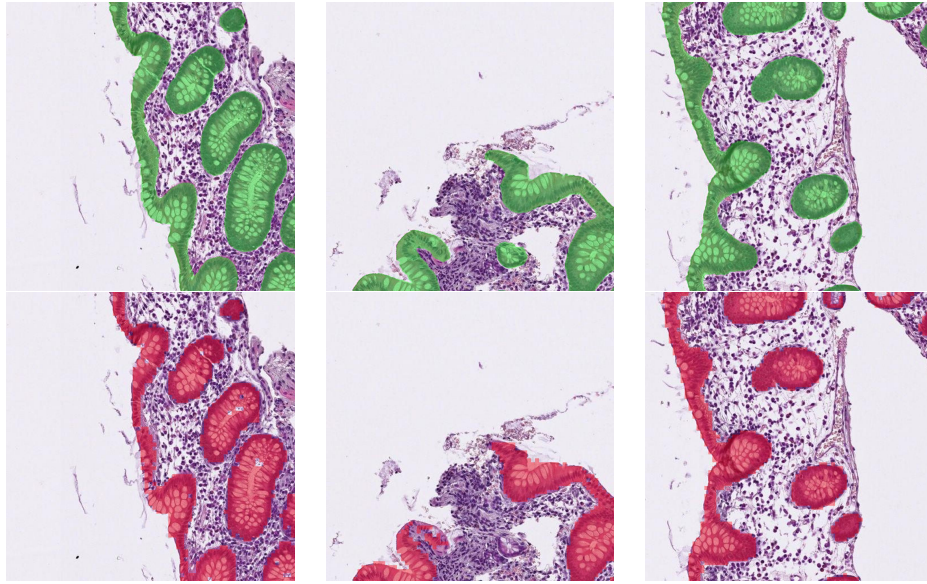


Figure 2: Visualizations of EpiSeg’s predictions for the presence of epithelium at patch level (size 14×14 at a spatial resolution of 0.5 mpp). First row (in green), is the ground truth, and second row (in red) is the model’s predictions.

The interpretability modules of IMILIA provide consistent tile-level patterns across cohorts. Tile scores were computed for all slides, allowing to extract max and min tiles for each dataset. Interpretable features were then computed for each extreme tile, including cell type counts (from HistoPLUS predictions) and density of several immune cell types in epithelium (through a combination of HistoPLUS and EpiSeg). Figure 3 illustrates extreme tiles sampled randomly within the min and max subgroups in the SPARC IBD cohort, with the predictions of HistoPLUS and EpiSeg. Additional visualization examples for the external validation cohorts FINBB and IBDColEPI can be found in Appendix D.

Figure 4 shows the average composition of minimal and maximal tiles in terms of cell types counts. Minimal tiles are richer in epithelial cells, while maximal tiles have a higher concentration of immune cells (lymphocytes, plasmocytes, eosinophils, neutrophils) and endothelial cells. These patterns were consistent across all three datasets, indicating consistent interpretability of IMILIA. Visual examination of the minimal and maximal tile samples by an expert pathologist corroborated the quantitative findings derived from HistoPLUS predictions. To avoid bias, the pathologist was blinded to the model-predicted cell counts during the assessment. Minimal tiles predominantly represented non-inflamed tissue, characterized by epithelial areas with crypts and collagen-rich stroma (primarily dense collagen with occasional myxoid areas). In contrast, maximal tiles exhibited dense lymphocytic and plasmacytic inflammation, with scattered eosinophils and neutrophils and minimal epithelial tissue.

The combination of HistoPLUS and EpiSeg allows to derive refined markers of the presence of immune cells in the epithelium, expressed as the number of cells per unit surface

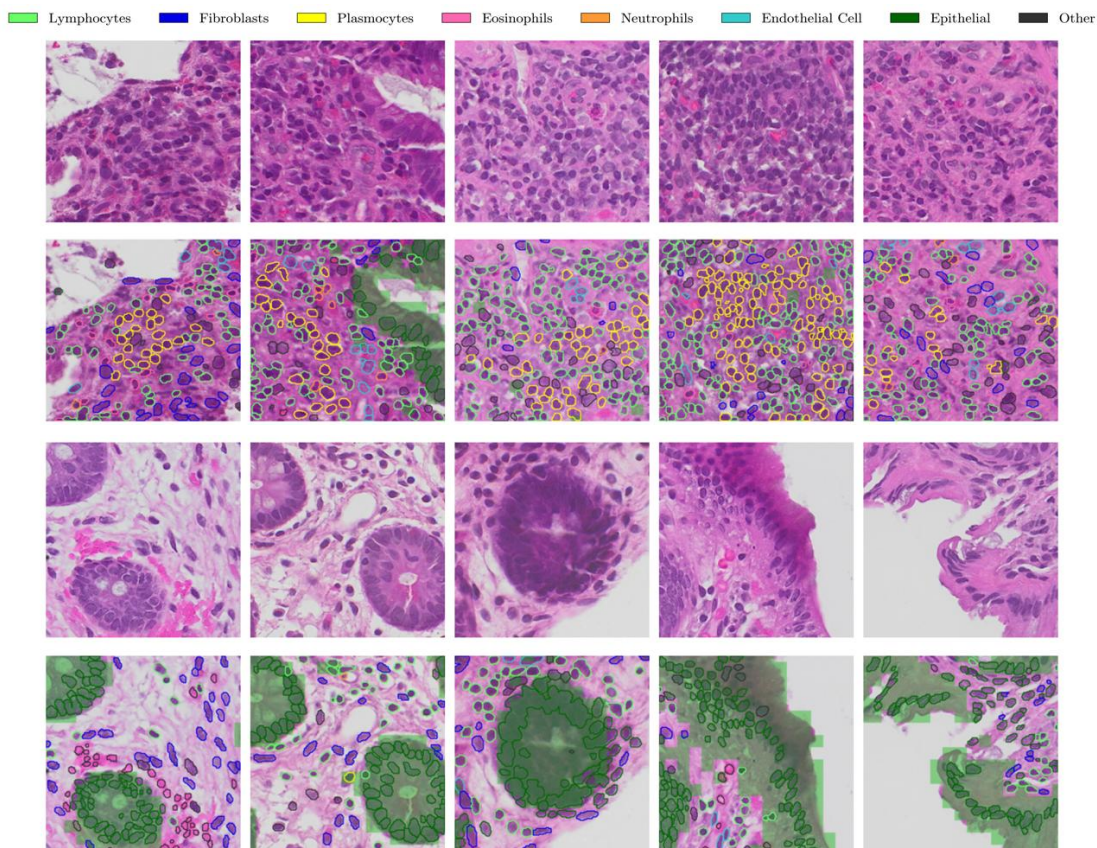


Figure 3: Examples of max (top two rows) and min (bottom two rows) tiles from the SPARC IBD cohort, with overlays of the predictions from HistoPLUS (cell contours) and EpiSeg (epithelium zones in green).

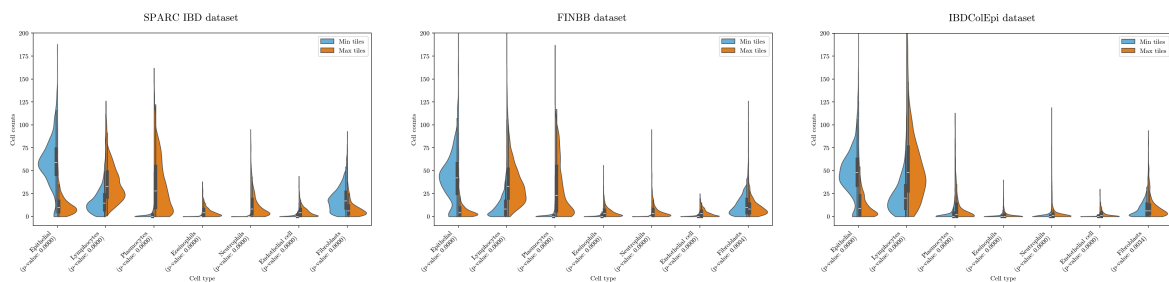


Figure 4: Violin plots of the distribution of cell types within min and max tiles, as analyzed by HistoPLUS, and by cohort.

area of epithelium, integrating a location awareness dimension. Comparison between extreme tiles shows that the difference is most significant for the density of neutrophils in the epithelium, as illustrated in Figure 5, and in line with known clinical indicators of the presence of inflammation in IBD.

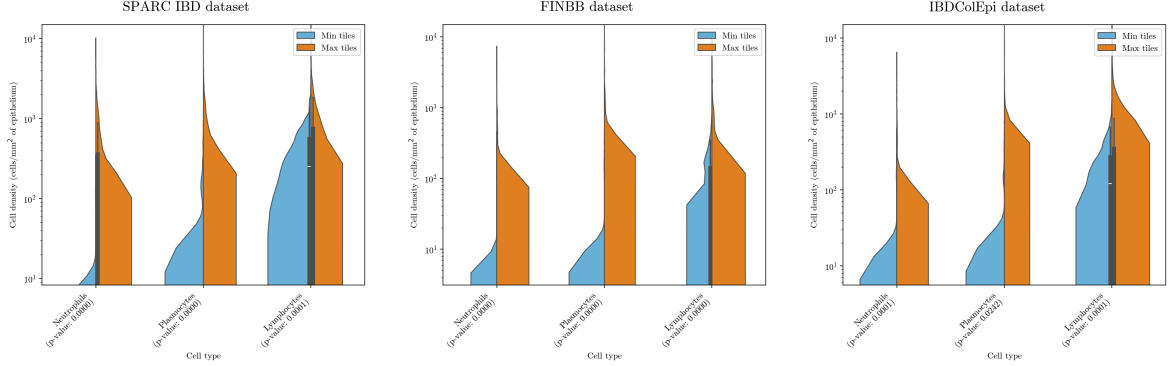


Figure 5: Violin plots of the distribution of immune cells in the epithelium.

6. Conclusion

In this study, we introduced IMILIA, an interpretable Multiple Instance Learning framework for the automated assessment of histological inflammation in IBD. Our results demonstrate that the pipeline not only achieves robust classification performance across diverse datasets but also, and crucially, provides transparent biological justifications for its predictions through quantitative cell and tissue-level analysis. One limitation of standard MIL approaches is the reliance on visual attention heatmaps for interpretability, which can lead to confirmation bias from the observer. IMILIA overcomes this by coupling the MIL attention mechanism with the HistoPLUS and EpiSeg modules, automatically deriving interpretable markers for the presence of inflammation, a first step towards automated biomarker discovery in IBD.

We also note the following limitations of our study. First, the interpretability module, fully automated, depends on the intrinsic performance of its two blocks, HistoPLUS and EpiSeg. While their performance can be estimated, inaccurate predictions can be a source of uncertainty in the full model, which would require some further validation in view of clinical usability. Second, all our models relied on a foundation model (H0-mini) developed on oncology WSIs. While it shows the strong transferability of such FMs to a new therapeutic area, future research could help assess the need for FMs tailored to IBD or inflammatory conditions in general. Finally, and as a line of future research, we note that the label “inflamed” used in this study can hide various histological patterns (*e.g.*, ulcers, architectural distortion, immune infiltration, crypt abscess, granulomas, etc.), and extending the model to these various categories is a natural perspective.

Acknowledgments

The results published here are in part from the Study of a Prospective Adult Research Cohort with IBD (SPARC IBD). SPARC IBD is a component of the Crohn’s & Colitis Foundation’s IBD Plexus data exchange platform. SPARC IBD enrolls patients with an established or new diagnosis of IBD from sites throughout the United States and links data collected from the electronic health record and study specific case report forms. Patients also provide blood, stool and biopsy samples at selected times during follow-up. The design and implementation of the SPARC IBD cohort has been previously described in (Raffals et al., 2021). It is available upon approved application to the Crohn’s & Colitis Foundation IBD Plexus.

This study utilized the national Fingenious service (www.fingenious.fi) managed by Finnish Biobank Cooperative – FINBB (www.finbb.fi). This dataset is part of a large-scale effort to advance precision medicine and research into various diseases, including Inflammatory Bowel Disease (IBD) and other conditions. The following biobanks are acknowledged for delivering samples and data to the study: Helsinki Biobank (<https://www.helsinginbiopankki.fi>) and Finnish Clinical Biobank Tampere (https://www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere).

References

- Benjamin Adjadj, Pierre-Antoine Bannier, Guillaume Horent, Sebastien Mandela, Aurore Lyon, Kathryn Schutte, Ulysse Marteau, Valentin Gaury, Laura Dumont, Thomas Mathieu, MOSAIC consortium, Reda Belbahri, Benoît Schmauch, Eric Durand, Katharina Von Loga, and Lucie Gillet. Towards comprehensive cellular characterisation of h&e slides, 2025. URL <https://arxiv.org/abs/2508.09926>.
- Dorottya Angyal, Fruzsina Balogh, Talat Bessissow, Panu Wetwittayakhleng, Akos Ilias, Lorant Gonczi, and Peter L. Lakatos. The role of histology alongside clinical and endoscopic evaluation in the management of ibd—a narrative review. *Journal of Clinical Medicine*, 14(7), 2025. ISSN 2077-0383. doi: 10.3390/jcm14072485. URL <https://www.mdpi.com/2077-0383/14/7/2485>.
- J Burisch, J Claytor, I Hernandez, JK Hou, and Kaplan GG. The cost of inflammatory bowel disease care: How to make it sustainable. *Clin Gastroenterol Hepatol.*, 23(3):386–395, 2025. doi: doi:10.1016/j.cgh.2024.06.049.
- Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, Mar 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02857-3. URL <https://doi.org/10.1038/s41591-024-02857-3>.
- Pierre Courtiol, Eric W. Tramel, Marc Sanselme, and Gilles Wainrib. Classification and disease localization in histopathology using only global labels: A weakly-supervised approach, 2020. URL <https://arxiv.org/abs/1802.02212>.

- Federica Di Vincenzo, Maria A Quintero, Joao M Serigado, Tulay Koru-Sengul, Rose Marie Killian, Julio Poveda, Jonathan England, Oriana Damas, David Kerman, Amar Deshpande, and Maria T Abreu. Histologic and endoscopic findings are highly correlated in a prospective cohort of patients with inflammatory bowel diseases. *Journal of Crohn's and Colitis*, 19(6):jjae141, 12 2024. ISSN 1876-4479. doi: 10.1093/ecco-jcc/jjae141. URL <https://doi.org/10.1093/ecco-jcc/jjae141>.
- Ruiwen Ding, Lin Li, Rajath Soans, Tosha Shah, Radha Krishnan, Marc Alexander Sze, Sasha Lukyanov, Yash Deshpande, and Antong Chen. Predicting ulcer in h&e images of inflammatory bowel disease using domain-knowledge-driven graph neural network. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2025. doi: 10.1109/ISBI60581.2025.10980783.
- Grace H. Elta, Henry D. Appelman, Elizabeth M. Behler, Joanne A. P. Wilson, and Timothy J. Nostrant. A study of the correlation between endoscopic and histological diagnoses in gastroduodenitis. *The American Journal of Gastroenterology*, 82(8):749–753, 1987. URL <http://hdl.handle.net/2027.42/72433>.
- Roger Feakins, Paula Borralho Nunes, Ann Driessen, Ilyssa O Gordon, Nina Zidar, Pamela Baldin, Britt Christensen, Silvio Danese, Naomih Herlihy, Marietta Iacucci, Maurice B Loughrey, Fernando Magro, Aart Mookhoek, Magali Svrcek, and Francesca Rosini. Definitions of histological abnormalities in inflammatory bowel disease: an ecco position paper. *Journal of Crohn's and Colitis*, 18(2):175–191, 08 2023. ISSN 1873-9946. doi: 10.1093/ecco-jcc/jjad142. URL <https://doi.org/10.1093/ecco-jcc/jjad142>.
- Joseph D. Feuerstein and Adam S. Cheifetz. Ulcerative colitis: Epidemiology, diagnosis, and management. *Mayo Clinic Proceedings*, 89(11):1553–1563, Nov 2014. ISSN 0025-6196. doi: 10.1016/j.mayocp.2014.07.002. URL <https://doi.org/10.1016/j.mayocp.2014.07.002>.
- Alexandre Filiot, Nicolas Dop, Oussama Tchita, Auriane Riou, Rémy Dubois, Thomas Peeters, Daria Valter, Marin Scalbert, Charlie Saillard, Geneviève Robin, and Antoine Olivier. Distilling foundation models for robust and efficient models in digital pathology, 2025. URL <https://arxiv.org/abs/2501.16239>.
- K Geboes, R Riddell, A Öst, B Jensfelt, T Persson, and R Löfberg. A reproducible grading scale for histological assessment of inflammation in ulcerative colitis. *Gut*, 47(3):404–409, 2000. ISSN 0017-5749. doi: 10.1136/gut.47.3.404. URL <https://gut.bmj.com/content/47/3/404>.
- Daniel B. Graham and Ramnik J. Xavier. Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature*, 578(7796):527–539, Feb 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2025-2. URL <https://doi.org/10.1038/s41586-020-2025-2>.
- Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, and Jens Kleesiek. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94:103143, 2024. ISSN 1361-8415. doi: <https://doi.org/10.1016/j.media.2024.103143>. URL <https://www.sciencedirect.com/science/article/pii/S1361841524000689>.

- Marietta Iacucci, Samuel C.L. Smith, Alina Bazarova, Uday N. Shivaji, Pradeep Bhandari, Rosanna Cannatelli, Marco Daperno, Jose Ferraz, Martin Goetz, Xianyong Gui, Bu Hayee, Gert De Hertogh, Mark Lazarev, Jim Li, Olga M. Nardone, Adolfo Parra-Blanco, Luca Pastorelli, Remo Panaccione, Vincenzo Occhipinti, Timo Rath, Gian Eugenio Tontini, Michael Vieth, Vincenzo Villanacci, Davide Zardo, Raf Bisschops, Ralf Kiesslich, and Subrata Ghosh. An international multicenter real-life prospective study of electronic chromoendoscopy score picasso in ulcerative colitis. *Gastroenterology*, 160(5): 1558–1569.e8, 2021. ISSN 0016-5085. doi: <https://doi.org/10.1053/j.gastro.2020.12.024>. URL <https://www.sciencedirect.com/science/article/pii/S0016508520355633>.
- Marietta Iacucci, Tommaso Lorenzo Parigi, Rocio Del Amor, Pablo Meseguer, Giulio Mandelli, Anna Bozzola, Alina Bazarova, Pradeep Bhandari, Raf Bisschops, Silvio Danese, Gert De Hertogh, Jose G. Ferraz, Martin Goetz, Enrico Grisan, Xianyong Gui, Bu Hayee, Ralf Kiesslich, Mark Lazarev, Remo Panaccione, Adolfo Parra-Blanco, Luca Pastorelli, Timo Rath, Elin S. Røyset, Gian Eugenio Tontini, Michael Vieth, Davide Zardo, Subrata Ghosh, Valery Naranjo, and Vincenzo Villanacci. Artificial intelligence enabled histological prediction of remission or activity and clinical outcomes in ulcerative colitis. *Gastroenterology*, 164(7):1180–1188.e2, 2023. ISSN 0016-5085. doi: <https://doi.org/10.1053/j.gastro.2023.02.031>. URL <https://www.sciencedirect.com/science/article/pii/S0016508523002160>.
- Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 10–15 Jul 2018. URL <https://proceedings.mlr.press/v80/ilse18a.html>.
- Aude Marchal-Bressenot, Julia Salleron, Camille Boulagnon-Rombi, Claire Bastien, Virginie Cahn, Guillaume Cadiot, Marie-Danièle Diebold, Silvio Danese, Walter Reinisch, Stefan Schreiber, Simon Travis, and Laurent Peyrin-Biroulet. Development and validation of the nancy histological index for uc. *Gut*, 66(1):43–49, 2017. ISSN 0017-5749. doi: 10.1136/gutjnl-2015-310187. URL <https://gut.bmj.com/content/66/1/43>.
- Ricardo Mokhtari, Azam Hamidinekoo, Daniel James Sutton, Arthur Lewis, Bastian Angermann, Ulf Gehrmann, Pål Lundin, Hibret Adissu, Junmei Cairns, Jessica Neisen, Emon Khan, Daniel Marks, Nia Khachapuridze, Talha Qaiser, and Nikolay Burlutskiy. Interpretable histopathology-based prediction of disease relevant features in inflammatory bowel disease biopsies using weakly-supervised deep learning. In Ipek Oguz, Jack Noble, Xiaoxiao Li, Martin Styner, Christian Baumgartner, Mirabela Rusu, Tobias Heinmann, Despina Kontos, Bennett Landman, and Benoit Dawant, editors, *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pages 479–495. PMLR, 10–12 Jul 2024. URL <https://proceedings.mlr.press/v227/mokhtari24a.html>.
- Natalie A. Molodecky, Ing Shian Soon, Doreen M. Rabi, William A. Ghali, Mollie Ferris, Greg Chernoff, Eric I. Benchimol, Remo Panaccione, Subrata Ghosh, Herman W.

- Barkema, and Gilaad G. Kaplan. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology*, 142(1):46–54.e42, Jan 2012. ISSN 0016-5085. doi: 10.1053/j.gastro.2011.10.001. URL <https://doi.org/10.1053/j.gastro.2011.10.001>.
- Siew C. Ng, Hai Yun Shi, Nima Hamidi, Fox E. Underwood, Whitney Tang, Eric I. Benchimol, Remo Panaccione, Subrata Ghosh, Justin C. Y. Wu, Francis K. L. Chan, Joseph J. Y. Sung, and Gilaad G. Kaplan. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *The Lancet*, 390(10114):2769–2778, Dec 2017. ISSN 0140-6736. doi: 10.1016/S0140-6736(17)32448-0. URL [https://doi.org/10.1016/S0140-6736\(17\)32448-0](https://doi.org/10.1016/S0140-6736(17)32448-0).
- Henrik Sahlin Pettersen, Ilya Belevich, Elin Synnøve Røyset, Erik Smistad, Melanie Rae Simpson, Eija Jokitalo, Ingerid Reinertsen, Ingunn Bakke, and André Pedersen. Code-free development and deployment of deep segmentation models for digital pathology. *Frontiers in Medicine*, 9, 2022. ISSN 2296-858X. doi: 10.3389/fmed.2021.816281. URL <https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2021.816281>.
- Laurent Peyrin-Biroulet, Shashi Adsul, Andrea Stancati, Jamshid Dehmeshki, and Olga Kubassova. An artificial intelligence-driven scoring system to measure histological disease activity in ulcerative colitis. *United European Gastroenterology Journal*, 12(8):1028–1033, 2024. doi: <https://doi.org/10.1002/ueg2.12562>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/ueg2.12562>.
- Christina Plattner, Gregor Sturm, Anja A. Kühl, Raja Atreya, Sandro Carollo, Raphael Gronauer, Dietmar Rieder, Michael Günther, Steffen Ormanns, Claudia Manzl, Stefan Wirtz, Asier Rabasco Meneghetti, Ahmed N. Hegazy, Jay V. Patankar, Zunamys I. Carrero, TRR241 IBDome Consortium, Markus F. Neurath, Jakob Nikolas Kather, Christoph Becker, Britta Siegmund, and Zlatko Trajanoski. Ibdome: An integrated molecular, histopathological, and clinical atlas of inflammatory bowel diseases, 2025. URL <https://www.biorxiv.org/content/early/2025/04/10/2025.03.26.645544>.
- Laura E Raffals, Sumona Saha, Meenakshi Bewtra, Cecile Norris, Angela Dobes, Caren Heller, Sirimon O’Charoen, Tara Fehlmann, Sara Sweeney, Alandra Weaver, Shrinivas Bishu, Raymond Cross, Themistocles Dassopoulos, Monika Fischer, Andres Yarur, David Hudesman, Deepak Parakkal, Richard Duerr, Freddy Caldera, Joshua Korzenik, Joel Pekow, Katerina Wells, Matthew Bohm, Lilani Perera, Manreet Kaur, Matthew Ciorba, Scott Snapper, Elizabeth A Scoville, Sushila Dalal, Uni Wong, and James D Lewis. The development and initial findings of a study of a prospective adult research cohort with inflammatory bowel disease (sparc ibd). *Inflammatory Bowel Diseases*, 28(2):192–199, 08 2021. ISSN 1078-0998. doi: 10.1093/ibd/izab071. URL <https://doi.org/10.1093/ibd/izab071>.
- James Reigle, Oscar Lopez-Nunez, Erik Drysdale, Dua Abuquteish, Xiaoxuan Liu, Juan Putra, Lauren Erdman, Anne M. Griffiths, Surya Prasath, Iram Siddiqui, and Jasbir Dhaliwal. Using deep learning to automate eosinophil counting in pediatric ulcerative colitis histopathological images, 2024. URL <https://www.medrxiv.org/content/early/2024/04/05/2024.04.03.24305251>.

David T Rubin, Olga Kubassova, Christopher R Weber, Shashi Adsul, Marcelo Freire, Luc Biedermann, Viktor H Koelzer, Brian Bressler, Wei Xiong, Jan H Niess, Matthias S Matter, Uri Kopylov, Iris Barshack, Chen Mayer, Fernando Magro, Fatima Carneiro, Nitsan Maharshak, Ariel Greenberg, Simon Hart, Jamshid Dehmeshki, and Laurent Peyrin-Biroulet. Deployment of an artificial intelligence histology tool to aid qualitative assessment of histopathology using the nancy histopathology index in ulcerative colitis. *Inflammatory Bowel Diseases*, 31(6):1630–1636, 09 2024. ISSN 1536-4844. doi: 10.1093/ibd/izae204. URL <https://doi.org/10.1093/ibd/izae204>.

Dawid Rymarczyk, Weiwei Schultz, Adriana Borowa, Joshua R Friedman, Tomasz Danel, Patrick Branigan, Michał Chałupczak, Anna Bracha, Tomasz Krawiec, Michał Warchoł, Katherine Li, Gert De Hertogh, Bartosz Zieliński, Louis R Ghanem, and Aleksandar Stojmirovic. Deep learning models capture histological disease activity in crohn’s disease and ulcerative colitis with high fidelity. *Journal of Crohn’s and Colitis*, 18(4):604–614, 10 2023. ISSN 1873-9946. doi: 10.1093/ecco-jcc/jjad171. URL <https://doi.org/10.1093/ecco-jcc/jjad171>.

Vincenzo Villanacci, Rachele Del Sordo, Sara Mino, Giorgia Locci, and Gabrio Bassotti. Histological healing in ibd: Ready for prime time? *Digestive and Liver Disease*, 57(5): 954–960, 2025. ISSN 1590-8658. doi: <https://doi.org/10.1016/j.dld.2025.01.039>. URL <https://www.sciencedirect.com/science/article/pii/S1590865825000404>.

Appendix A. Hyperparameters

We report in Table 2 a more complete list of hyperparameters used to train the various models.

Table 2: Hyperparameters used to train the various models.

| Chowder model | | EpiSeg | |
|----------------------------|------------------|-----------------|--------------------|
| Tile size | 224×224 | Tile size | 1022×1022 |
| Batch size | 256 | Patch size | 14 |
| Number of folds | 5 | Number of folds | 3 |
| Number of channels (K) | 5 | C | 10^{-2} |
| r | 25 | | |
| Learning rate | 0.01 | | |
| Max number of tiles | 1000 | | |
| MLP hidden layers’ neurons | [128, 64] | | |
| MLP hidden layers’ dropout | [0.5, 0.5] | | |

Appendix B. Detailed HistoPLUS transfer performance

We report in Table 3 detailed segmentation metrics for HistoPLUS, comparing its performance on the oncology validation dataset and on the SPARC IBD dataset. We refer to the original HistoPLUS paper for a precise definition of the metrics.

Table 3: Performance of HistoPLUS in cell detection and segmentation for HistoVAL (oncology) and SPARC IBD. We report mean values and confidence intervals at 95% level, obtained by bootstrapping with 1000 repeats.

| Dataset | Panoptic Quality | Detection Quality | Segmentation Quality |
|-----------|----------------------|----------------------|----------------------|
| HistoVAL | 0.605 (0.595; 0.613) | 0.753 (0.742; 0.763) | 0.801 (0.799; 0.804) |
| SPARC IBD | 0.586 (0.572; 0.599) | 0.774 (0.760; 0.789) | 0.755 (0.749; 0.761) |

Appendix C. EpiSeg precision-recall curve

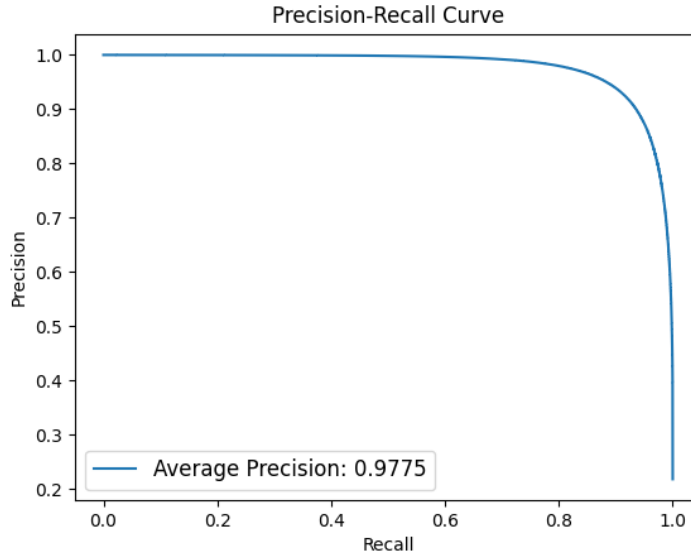


Figure 6: Precision-recall curve for the patch-level prediction of presence of epithelium by EpiSeg. The average precision metric is defined as the area under the precision-recall curve.

Appendix D. Min and Max tiles visualization on the external validation cohorts

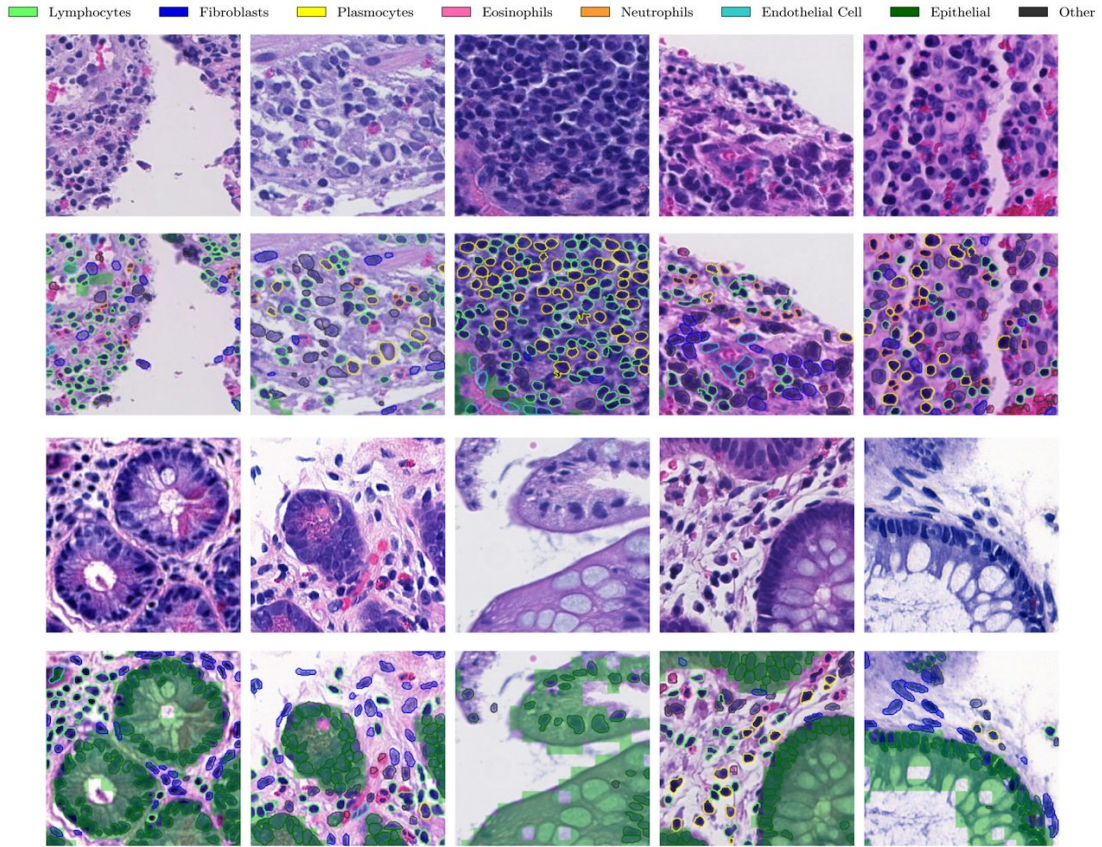


Figure 7: Max (top two rows) and min (bottom two rows) tiles within the FINBB cohort, with overlays of the predictions from HistoPLUS and EpiSeg. Interestingly, one can notice the staining distribution shift compared to discovery cohort.

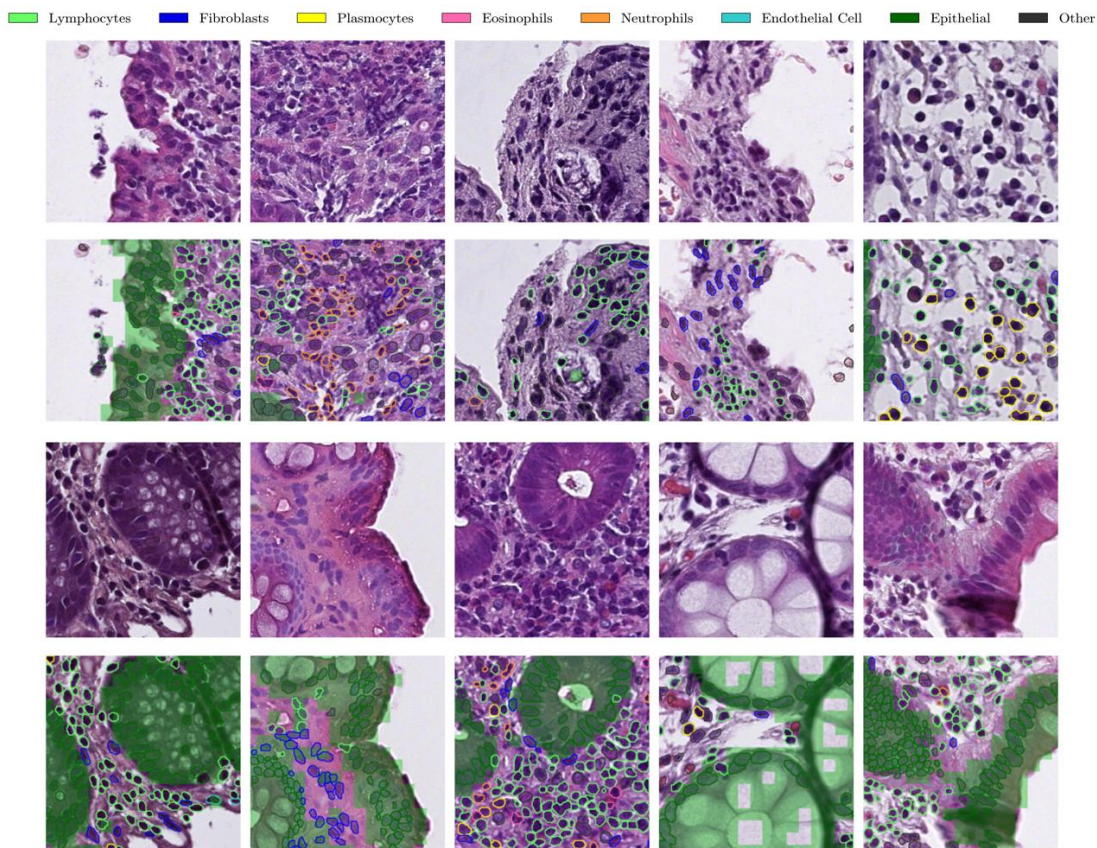


Figure 8: Max (top two rows) and min (bottom two rows) tiles within the IBDColEpi cohort, with overlays of the predictions from HistoPLUS and EpiSeg.