Automating Ground Truth Annotations For Gland Segmentation Through Immunohistochemistry

Tushar Kataria (tushar.kataria@utah.edu)
University of Utah
Saradha Rajamani (saradharajamani0@gmail.com)
University of Utah
Abdul Bari Ayubi (aabari888@gmail.com)
University of Utah
Mary Bronner (mary.bronner@aruplab.com)
University of Utah
Jolanta Jedrzkiewicz (jolanta.jedrzkiewicz@hsc.utah.edu)
University of Utah
Beatrice Knudsen (beatrice.knudsen@path.utah.edu)
University of Utah
Shireen Elhabian (shireen@sci.utah.edu)
University of Utah

Article

Keywords:

DOI: https://doi.org/

Additional Declarations: No competing interests reported.

Automating Ground Truth Annotations for Gland Segmentation Through Immunohistochemistry

Tushar Kataria[1,2], Saradha Rajamani[1,2], Abdul Bari Ayubi[3], Mary Bronner[3,4], Jolanta Jedrzkiewicz[3,4], Beatrice Knudsen[2,3,*], and Shireen Y. Elhabian[1,2,*]

[1]Kahlert School of Computing, University of Utah, USA
[2]Scientific Computing and Imaging Institute, University of Utah, USA
[3]Department of Pathology, University of Utah, USA
[4]ARUP Laboratories, Salt Lake City, USA
[*]Corresponding authors:shireen@sci.utah.edu, beatrice.knudsen@path.utah.edu

ABSTRACT

The microscopic evaluation of glands in the colon is of utmost importance in the diagnosis of inflammatory bowel disease (IBD) and cancer. When properly trained, deep learning pipelines can provide a systematic, reproducible, and quantitative assessment of disease-related changes in glandular tissue architecture. The training and testing of deep learning models require large amounts of manual annotations, which are difficult, time-consuming, and expensive to obtain. Here, we propose a method for the automated generation of ground truth in digital H&E slides using immunohistochemistry (IHC) labels. The image processing pipeline generates annotations of glands in H&E histopathology images from colon biopsies by transfer of gland masks from CK8/18, CDX2, or EpCAM IHC. The IHC gland outlines are transferred to co-registered H&E images for the training of deep learning models. We compare the performance of the deep learning models to manual annotations using an internal held-out set of biopsies as well as two public data sets. Our results show that EpCAM IHC provides gland outlines that closely match manual gland annotations (DICE = 0.89) and are robust to damage by inflammation. In addition, we propose a simple data sampling technique that allows models trained on data from several sources to be adapted to a new data source using just a few newly annotated samples. The best-performing models achieved average DICE scores of 0.902 and 0.89, respectively, on GLAS and CRAG colon cancer public datasets when trained with only 10% of annotated cases from either public cohort. Altogether, the performances of our models indicate that automated annotations using cell type-specific IHC markers can safely replace manual annotations. The automated IHC labels from single institution cohorts can be combined with small numbers of hand-annotated cases from multi-institutional cohorts to train models that generalize well to diverse data sources.

1 Introduction

Image analysis for digital pathology has emerged as a computer-assisted tool for histopathologic diagnosis of inflammatory diseases and cancer[1]. With an increasing number of annual cancer diagnoses[2], declining numbers of pathologists[3,4], and the qualitative nature of disease activity assessments[5–8], there exists a dire need to increase the efficiency of rendering a diagnosis and provide tools for quantitative slide evaluations by practicing pathologists.

Computer-assisted diagnosis has undergone improvements over the past decade through the development of deep-learning models[9]. However, machine learning models using highly supervised training approaches are data-hungry, requiring enormous amounts of annotated data for reliable predictions. Thus, the generation of manual annotations constitutes a major roadblock to the deployment of clinical-grade models. The generation of cell type-specific annotations for digital pathology faces three main difficulties. First, pathologists need to spend time annotating regions of interest. Second, interobserver disagreement[5,6] requires a consensus of multiple pathologists to obtain accurate ground truth. Lastly, some of the cell types that require annotations may be challenging to detect in H&E-stained tissue sections[8]. To increase the efficiency and accuracy of generating manual annotations inside H&E digital slides, we propose an automated annotation process. The workflow uses antibodies to automatically mark specific cells with a brown color by immunohistochemistry (IHC). A threshold of brown color intensity is applied to generate a binary mask (or gland outline) that is automatically transferred from IHC to H&E slides. The transferred IHC masks are used instead of manual annotations for the training of machine-learning models. Hundreds of antibodies are available for tissue staining using a fully automated staining process and can be used for cell-type-specific ground truth generation. Thus, IHC has the potential to replace manual ground truth labels for the training of algorithms. Using IHC for ground truth generation requires little to no oversight by pathologists, enabling a way to generate an unlimited amount of annotated data.

Inflammatory bowel disease (IBD) is characterized by inflammation of the gastrointestinal tract. The disease is painful and predisposes patients to cancer. Therefore afflicted individuals require intense clinical surveillance with annual biopsies[2]. IBD includes Crohn's disease and ulcerative colitis. Whereas Crohn's disease affects the bowel intermittently with skip lesions and transmucosal inflammation, ulcerative colitis affects the full length of the bowel and is centered in the colonic mucosa. In both IBD disease subtypes, the affected segments of the bowel may require surgical removal, with likely peri/postoperative complications[2]. Although endoscopic evaluation is essential for disease surveillance, recent studies[10–12] have shown that histologic analysis more accurately assesses disease activity[13,14].

Gland[15–18] and nuclear segmentation[19–21] using deep learning models has emerged as important problems in the field of histopathology image analysis. Under the microscope, ulcerative colitis shows inflammatory infiltrates in the lamina propria, leading to the separation of glands and distortion of gland architecture. Inflammatory cells, in particular neutrophils, infiltrate glands, destroy the epithelial gland lining, and form abscesses in the gland lumen. Multiple scoring systems have been developed by pathologists to report disease activity in IBD. They are based on the density and location of inflammatory cells relative to glands. However, pathologic reporting of disease activity is not standardized, is semi- or non-quantitative, and is cumbersome to use.

Towards a computer-assisted diagnostic tool for the evaluation of disease activity in IBD, we developed a gland segmentation pipeline using IHC as ground truth. A similar approach has previously been applied to prostate[22], colon, and breast cancer[23], and using immunofluorescence staining[24]. The goal of our work is to expand and improve previous methods in several ways: compared to Boulten et al.[22] our pipeline is simpler with no human intervention; compared to Brázdil et al.[23] our approach avoids adaptive thresholding and nuclear centroids for registrations. In this study, we systematically investigate IHC markers of proteins expressed in the cytoplasm (cytokeratin-8/18, CK8/18), on the cell surface (epithelial cell adhesion molecule - EpCAM), or in the nucleus (caudal type homeobox-2 - CDX2). We compare the three IHC markers for their ability to generate accurate ground truth annotations. As a whole, our method eliminates the need for manual annotations by pathologists, thereby removing a major bottleneck for the development and deployment of deep learning models in real-world scenarios.

2 Methods

2.1 Automated Gland Mask Generation via Immunohistochemistry

The pipeline takes advantage of the fact that the two paired digital whole slide images (WSIs), namely, the H&E and the corresponding IHC, image are derived from the exact same tissue section. The IHC labeling generates ground truth at pixel-level resolution. The IHC pipeline consists of traditional image processing algorithms for tissue edge detection, opening, and closing, as well as contour generation[25]. These algorithms are fast, easy to compute, and reproducible but they are more sensitive to noise when compared to deep learning models, and so require preprocessing steps to clean the data. This section will describe each block in the proposed annotation pipeline.

2.1.1 Tissue extraction, IHC masks and registration

Each WSI is a gigapixel image, greater than 40000 pixels in each dimension. Due to memory constraints, WSIs from H&E and IHC are separated into tissue pieces for coregistration. This approach also removes small, uninformative tissue pieces, retaining relevant information.

To segment different tissue pieces of the H&E and IHC WSI, we first converted the image to grayscale. We applied a canny edge detector[25] on the grayscale image to get the boundaries separating the foreground (tissue pieces) and background. After edge detection, using morphological operations, we removed noise and small objects, and closer objects are merged into a single object as shown in Supplementary Figure SF1-A. Lastly, we place bounding boxes around tissue pieces. Each extracted tissue piece was supplied separately to the automated gland mask generation block.

Three IHC markers, CDX2 (nuclear), CK8/18 (cytoplasmic), and EpCAM (membrane), were used to stain glandular epithelial cells in the colon. The digital IHC images were used for the automated generation of gland masks. For each tissue piece, the DAB channel[26] was separated from the hematoxylin channel, and a threshold of positive DAB pixels was calculated using Eq. 1 to generate a binary mask. The mask provides landmark locations for epithelial cells for each of the three markers, which are then converted utilizing morphological processes to encapsulate the epithelium.

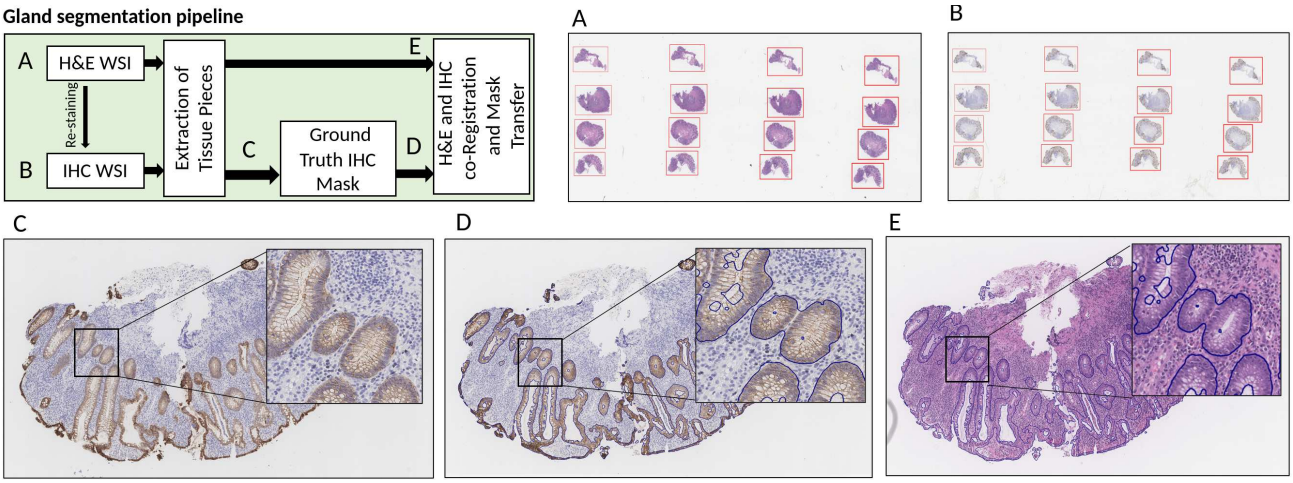

Figure 1. Gland segmentation pipeline. H&E-stained slides are digitized (H&E Whole Slide Image – WSI). The coverslips are removed, and the tissue is decolorized and re-stained by IHC with one of three antibodies: CDX2, a nuclear lineage marker of colonocyte differentiation, CK8/18, a cytoplasmic marker of epithelial differentiation, or EpCAM, an epithelial membrane protein. **A**. H&E-stained slide. **B** and **C**. IHC slides are digitized (IHC Whole Slide Image – WSI). Panel **B** shows the WSI, and panel **C** shows one of the tissue pieces stained with EpCAM in its bounding box. **D**. IHC masks. Binary IHC masks are generated based on an automated threshold in the DAB channel. **E**. Co-registration and mask transfer. A rigid body registration is applied to co-register the paired H&E and IHC tissue pieces. The binary IHC mask is transferred to the H&E images. Details of the tissue piece extraction and mask generation block are described in Supplementary Figure 1.

Several differences are identified between the three IHC markers: (1) within glandular masks generated by CDX2 IHC, holes are observed in the epithelial lining due to the nuclear localization of the CDX2 marker; (2) in EpCAM and CK8/18 IHC images, mucin-filled goblets in the apical cytoplasm of colonic epithelial cells are visually similar to the background. In both cases, the regions are filled[25] to generate the final gland outlines. Because goblets and lumens appear to be of similar intensities (see Figure 2), thresholding cannot be used to distinguish between goblets and lumens. ALtogether, we concluded that CDX2 can be used for gland outlines, but not for annotations of glandular lumens. On the other hand, EpCAM and CK8/18 markers provide both, gland and lumen annotations.

$$threshold\_value = mean(DAB\_channel) - standard\_devation(DAB\_channel) \quad (1)$$

Masks obtained from IHC images are transferred to the corresponding H&E tissue using image registration algorithms. Although the transformation between H&E and IHC is deformable at the whole slide level, at the tissue level we can simplify the assumptions. As re-staining minimally alters the size or structure of glands in corresponding tissue pieces, we can assume rigid body motion[27], i.e., the two tissues under observation are only misaligned by rotation, scaling, and translation. Consequently, using rigid body registrations[28] to coregister the corresponding tissue fragments for each marker results in accurate alignments.

2.2 Datasets

Here we describe the IBD WSI collection protocol. Tissue split for training and testing performance analysis of deep learning models and external datasets used analysis of performance generalization across different sites and disease severity type for gland segmentation.

2.2.1 Internal Dataset and manual annotations

The dataset includes H&E whole slide images from surveillance colonoscopies of 5 patients with active ulcerative colitis. It contains 15 H&E WSIs displaying 92 tissue pieces (16 – 24 tissue pieces per WSI) and approximately 3000 glands. Formalin-fixed and paraffin-embedded tissue blocks from IBD cases were retrieved from the pathology archive at the University of Utah. The pathology archive at the University of Utah followed all the informed consent guidelines for collecting samples for this study. All experimental protocols were approved by Institutional Review Board (IRB) at the University of Utah under IRB_00140202 and IRB_00057287. All the experiments done and reported in the manuscript follow the guidelines of the above stated IRB protocols. The study is approved under a waiver of consent since all HIPAA-sensitive data fields are removed prior to the use of slides. No demographic or clinical information from study participants is used for data analysis, and the link to the

medical record was destroyed before the images are processed. The waived informed consent was approved by IRB at the University of Utah.

From each block we obtained three unstained slides. The slides were stained with H&E using the automated clinical staining process and scanned on the Aperio AT2 slide scanner with a pixel resolution of 0.23 nm at x40. After scanning, coverslips were removed and the slides were placed in the Leica Bond 3 autostainer for restaining by immunohistochemistry (IHC) with antibodies reactive with CDX2 (caudak type homeobox-2), EpCAM (epithelial cellular adhesion molecule) or CK8/18(cytokeratin-8/18). The heat retrieval prior to the antibody incubation decolorizes the H&E slides, foregoing manual removal of H&E. After IHC, slides were scanned on the Aperio AT2 at x40 and the digital IHC images was paired with the corresponding H&E image.

Board-certified pathologists were tasked to manually outline glands and lumens in 10 H&E images of tissue samples without the assistance of IHC slides. These manual annotations were never used for training, but rather to compare the performance of an automated annotation framework to that of a pathologist.

Variability within digital slides have a negative impact on the performance of deep learning models[15,22,29]. A sample with higher disease activity loses the gland-like structure, thereby confusing deep-learning models. These datasets also have inflammatory cells surrounding glands changing the color composition of both H&E and IHC images. To study the impact of gland diversity in annotations, we requested that pathologists assess the disease activity, i.e. the severity of active inflammation, in each tissue separately and assign scores of 0 (no inflammation), 1 (mild inflammation), and 2 (moderate inflammation). These annotations helped demonstrate the robustness of our proposed method in inflamed tissues.

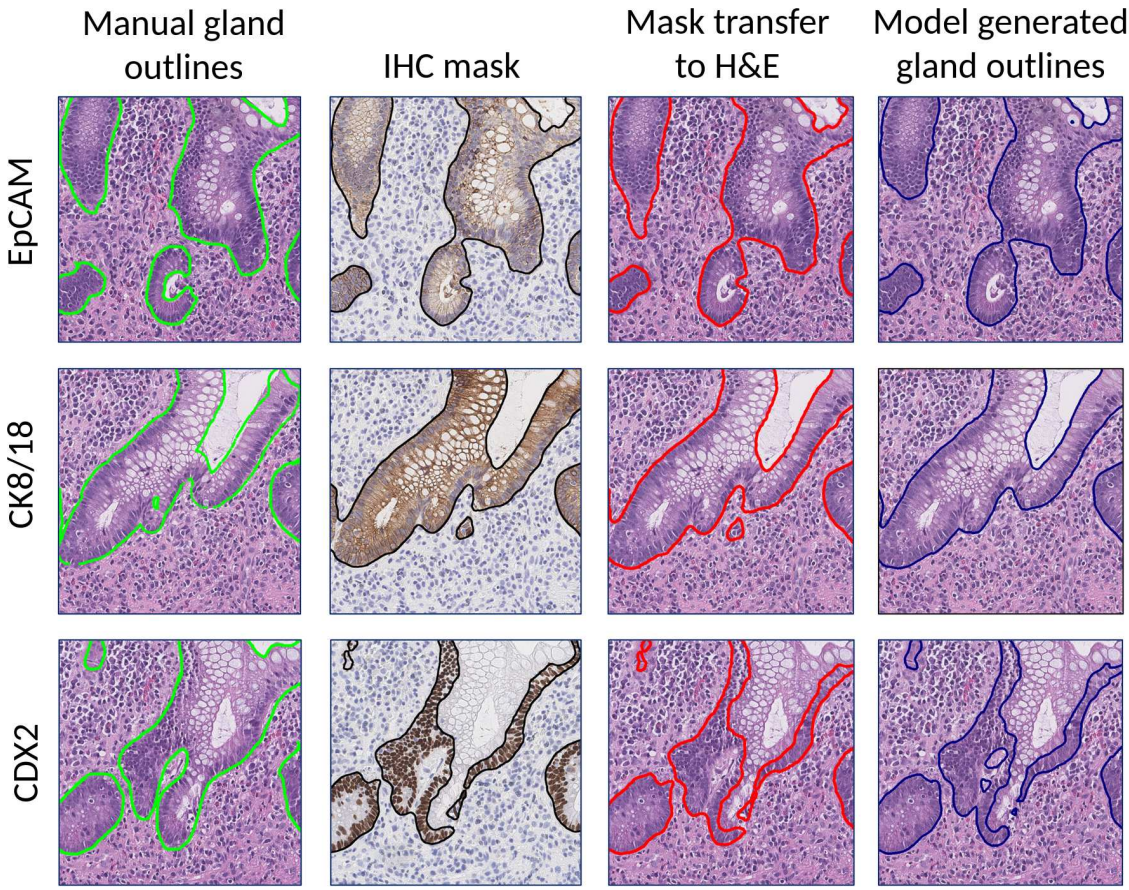

Figure 2. Gland segmentation results. Gland outlines are generated manually (column 1), based on IHC staining (column 2), or by transfer of the IHC mask to the H&E image (column 3). After the transfer of gland outlines from IHC to H&E, an algorithm is trained on gland outlines in H&E images. To generate gland outlines by IHC, slides were stained with CDX2, CK8/18 or EpCAM.

2.2.2 CRAG and GLAS external datasets

Two colon cancer datasets, GLAS and CRAG[15,16], were used for testing the generalization of trained algorithms. Both datasets possess ground truth gland segmentation annotations for normal and colon cancer glands.

The CRAG dataset was extracted from 38 H&E strained WSIs, scanned using Omnyx VL120 scanner with a pixel resolution of $0.55\mu m$/pixel. The training dataset consists of 168 images and the testing dataset consists of 40 images with a patch size of 1512x1516 pixels, with corresponding gland segmentation ground truths.[15]

The training dataset for GLAS was extracted from 16 H&E strained WSIs, scanned using Zeiss MIRAX MIDI Slide Scanner with a pixel resolution of $0.465\mu m$/pixel. The testing dataset for GLAS consists of 80 images split in Test A and Test B datasets with patch sizes up to 755x522 pixels and their corresponding gland segmentation ground truths. Test A and Test B were released at separate times during the challenge phase. The training dataset consists of 88 images.[16]

2.3 Model Training

Three models are used for training U-Net[30], Feature Pyramid networks[31], and DeepLabV3[32] to make sure that segmentation results do not have a model architecture bias. Model architectures were used from the Pytorch segmentation models library[33]. U-Net[30] is a popular medical data semantic segmentation network. Because of skip connections, U-Net can transmit minute details to the decoder branch, resulting in better segmentation results. The main advantage of U-Net is its fast convergence and stability for small datasets. Other variants of U-Net[17,34–36] further improve results due to complex model architectures, but for our experiments, we have limited our scope to U-Net for simplicity. Feature Pyramid networks (FPN) is a model architecture that uses multi-scale deep representations, achieving high accuracy for detection and semantic segmentation tasks. Because multiscale architectures have been shown to be more effective for histopathology images[17,37], we use an FPN-based segmentation model to establish another baseline. DeepLabV3 employs atrous convolution. The enlarged field of view[32,38] of the atrous convolution layers aids in increasing the performance of the semantic segmentation models. For all three models, the encoder weights were initialized to the Resnet34 model[39]. Without loss of generalization, assume all the results reported in section 3 are on x20 magnification and utilize the U-Net framework as the CNN model. Other magnifications and comparisons to other model frameworks are provided in supplementary figures.

300 patches were randomly sampled from each tissue in the training set. These patches and the corresponding ground truths from the IHC marker pipeline were used for training the deep learning models. With a lower number of patches (N=50 or 100 per tissue piece), the model did not converge to a satisfactory solution, and a higher number of patches from each tissue (N=600, 2000) led the model to overfit the data. The A30 NVIDIA Graphics unit was used to train all the models. The learning rate and weight decay of each model were set to 1e-4. The learning rate reduces by 10 after half the number of epochs. During inference, a window size of 128 pixels was used. Results from overlapping windows were averaged. All experiments employed binary cross entropy with sigmoid activation as the loss function.

We used two performance metrics to analyze the results Dice score and Jaccard score (also known as intersection over union). Definitions of these metrics are given below.

$$Dice = \frac{2 * P * GT}{P + GT}$$

$$Jaccard = \frac{P \cap GT}{P \cup GT}$$

where P is the predicted segmentation mask and GT is the ground truth segmentation mask. These metrics are commonly used for segmentation tasks[16,30,32,40].

2.4 Statistical Analysis

We train multiple copies of models from random initialization evaluating the mean, median, standard deviations, and interquartile ranges of their performance measures. Deep learning models are inherently stochastic because of random initialization and the path they take to minima. Analyzing the mean and standard deviation of different models trained on the same data is a better indicator of performance than a single model observation. Reporting average results also help to determine if there are statistically significant differences between the reported results for different IHC markers or domain adaption experiments. We used violin plots for these performance comparisons. As violin plots are a combination of the box plot and the kernel density estimate plot, they can be interpreted similarly to box plots.

Calculation of p-values: Null hypothesis testing in statistics relies on the independent sampling of data distributions. The generalization of machine learning or deep learning model outputs is determined using cross-validation approaches. Simple statistical tests on probabilities obtained from models that are trained on dependent (cross-validated) data have a high probability

of reporting type I error incorrectly[41], whereas paired t-test[42] and 5x2cv[41] test have a low probability for type I error. We report p-values using paired t-tests because that is applicable to the experimental setup explained above.

3 Results

3.1 Generation of Automated Gland Masks

To automate gland annotations for ground truth generation in H&E digital slides, we transferred gland outlines from corresponding IHC to the H&E images. IHC marks the expression of specific proteins in tissue sections based on the visualization of antibody binding through the deposition of brown DAB chromogen. Multiple IHC markers specifically stain cells in colonic glands and thus can potentially be used for gland masking. It is unclear whether nuclear, cytoplasmic, or membrane IHC markers are best suited to generate binary IHC masks. To address this question, we develop a pipeline that automates the generation of masks in IHC images and their transfer to H&E images. The accuracy of these generated masks is then determined by comparing automated gland outlines to manual annotations by board-certified pathologists.

Three tissue sections from the same block were used to compare masks derived from different IHC markers. First, tissue sections were stained with H&E and scanned to produce digital slides. The H&E WSI contains four rows of biopsy tissues, and each row represents a separate biopsy (Figure 1). Each biopsy tissue appears four times on the same slide in deeper sections at 4-micron intervals. After removing the coverslips, the H&E stained tissue section was placed into an autostainer for IHC. During the IHC staining process, the antigen retrieval step heats the tissue to $97^{\circ}C$ in an aqueous buffer solution that reverses chemical modifications introduced during tissue fixation. Interestingly, at the same time, the buffer formulations decolorize the H&E stain. The destained slides are restained with either a nuclear marker (CDX2), a cytoplasmic marker (CK8/18), or a membrane marker (EpCAM).

Figure 1 illustrates the main steps in the proposed pipeline. The digital IHC slide is processed first. To generate a binary mask, we applied an automated threshold that separates brown pixels within glands from background pixels. Next, tissue pieces within IHC or H&E slides are placed into bounding boxes and extracted for registration. Details of the gland mask generation block are shown in SF1. After co-registering of IHC and H&E tissue pieces, gland masks are transferred from the IHC to the H&E image. Each IHC marker generates a slightly different mask that we compare to the manual gland outlines. The results are shown in Table 1. The EpCAM-derived IHC masks demonstrate the highest accuracy in matching human gland outlines followed by CK8/18 and CDX2.

IHC	Dice	Jaccard
CDX2	0.807	0.6775
EpCAM	**0.8631**	**0.7656**
CK8/18	0.8395	0.7245

Table 1. Binary IHC masks against manual gland. Binary masks generated by IHC markers listed in the first column. The masks are overlaid on manual gland annotations to calculate Dice score and Jaccard index

3.2 Gland Segmentation Models Trained on Automated Gland Masks

The gland masks from CDX2, EpCAM, and CK8/18, and one from all three IHC markers combined in the H&E are used to train and test four types of deep-learning models for gland segmentation, The performances of the trained models are evaluated on tissue pieces not included in the training set (60 percent) that are either from the same patients (internal test set, 20 percent) or different patients not used for training (held out test set, 20 percent). Models are trained on each of the 5-fold cross-validations sets multiple times, and results are reported as explained in section 2.4. All hyperparameters, including learning rate, batch size, weight decay, number of patches from each tissue, and magnification are established by tuning the model on the validation set. The test set is not evaluated until all the parameters are optimized.

To identify the optimal model architecture and conditions for training, we performed several ablation experiments. Keeping all other experimental parameters constant, we trained models on x40, x20, and x10 images of the same tissue pieces (Figure SF5). For all three IHC datasets, models trained on x40 and x10 exhibit greater variance than models trained on x20. We concluded from these experiments that the optimal pixel size for training deep learning gland segmentation models is x20 magnification. Using x20 images, we compared the performance of three model architectures, U-Net[34], Feature Pyramid networks[31], and DeepLabV3[18]. Results on the internal test set with Jaccard and Dice scores are shown in Figure SF6. The similar mean and distribution of performance scores across models and IHC datasets suggest that the three model architectures

have comparable performances. DeepLabv3-trained models slightly outperform U-Net models, but the difference is not statistically significant. For subsequent experiments, we use U-Net and x20 magnification images.

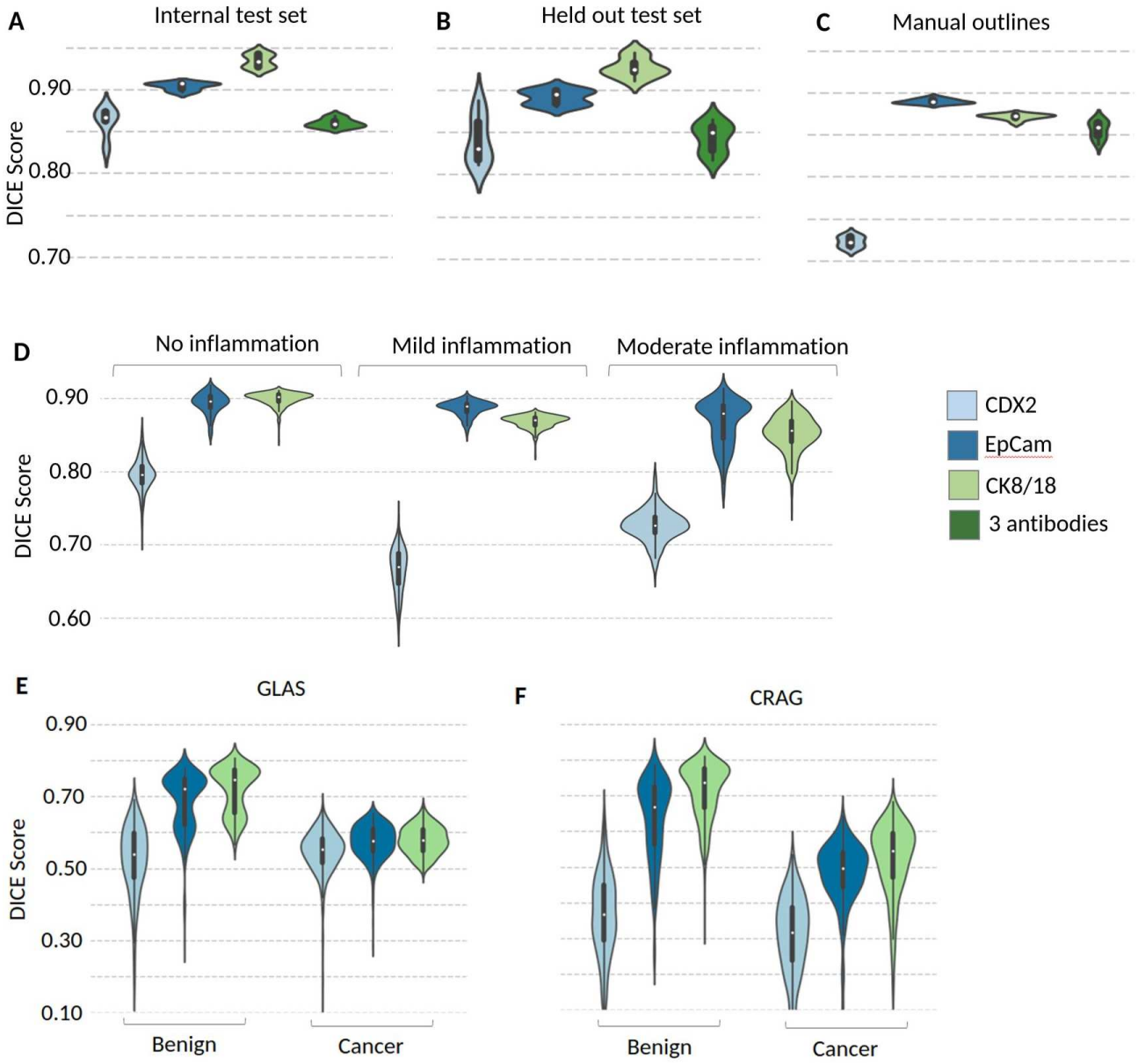

Figure 3. Performance of gland segmentation models. For each marker, we get a kernel density estimate of performance metrics, which is shown as a violin plot. **A. Internal test set**. Dice scores of models trained on ground truth masks obtained from CDX2, EpCAM, CK8/18, or on all three antibodies are shown. One tissue piece from each case is held out for training and used to evaluate the performance of the model. **B. Held out test set**. Models are applied to a held-out case, not used for training. **C. Manual annotations**. Computer-generated gland outlines are compared to outlines obtained through manual annotations. **D. Effect of disease activity on model performance.** The amount of inflammation in each tissue piece was scored by a pathologist and classified as no, mild, or moderate inflammation. **E. External GLAS dataset.** Models were tested separately on benign and cancer cases in the GLAS testing cohort. **F. External CRAG dataset.** Models were tested separately on benign and cancer cases in the CRAG testing cohort.

We train separate models on EpCAM- , CK8/18- and CDX2-derived gland masks and test their performance on held-out gland masks as well as manual annotations. The performance of models are shown by their distributions of Dice and Jaccard

scores as violin plots in Figure 3-A and -B and Figure SF3. The model trained on CDX2-derived gland masks possesses the lowest Dice score. The mean performance of models trained on CK8/18-derived gland masks is significantly higher than the performance of models trained on EpCAM masks ($p < 0.05$). Training models on all three CK8/18, EpCAM, and CDX2-derived masks reduces the performance. We compared the performance of the trained models to the manual gland outlines provided by pathologists on both internal and external test data. The respective Dice score for models trained on EpCAM-derived masks (0.89 ± 0.003) is greater than for CK8/18-derived masks ($p < 0.05$) (Figure 3-C). As a reference, we compared manual annotations to EpCAM-derived IHC masks transfered to H&E images (Table 1). The respective Dice score is lower than the Dice score obtained from the models trained on EpCAM-derived gland masks Figure 3-C. The same observation applies to CK8/18, but not to CDX2.

Next, we determine the effect of disease activity, i.e. inflammation, on the performance of the models. Inflammation can lower the staining intensities of CK8/18 and EpCAM IHC labels but does not have a significant effect on CDX2 staining intensity. Figure 3-D shows the performance of the model on tissues grouped by disease activities. Inflammation reduces the performance of all the models, most notably the model trained on CK8/18-derived gland outlines. Models trained on EpCAM-derived gland masks show the most robustness to inflammation activity.

We also conducted experiments to determine whether providing fewer training data impacts the performance of the models. The results of these experiments are shown in Figure SF7. By increasing the size of the dataset, we observe increases in mean Dice score and Jaccard index and decreases in the variation of scores. Doubling the amount of data from 25 tissues to 50 tissue pieces lowers the standard deviation of the Dice scores by 10 to 20 percent. Altogether, we observe a significant increase in performance and a decrease in the variability of Dice and Jaccard scores by increasing the amount of training data.

3.3 Testing IHC-Derived Models on Public Data

We trained models on CDX2-, CK8/18- or EpCAM-derived gland outlines and tested on public data from GLAS and CRAG gland segmentation challenges. GLAS and CRAG are multi-institutional cohorts of digital slides from normal colon and colon cancer. The cohorts are divided into training and test sets and all the glands are annotated by hand. The performances of models trained on internal IHC-derived gland outlines and applied to CRAG and GLAS cases in the test set are shown in Figure 3-E and 3-F. We divided the cohorts into images from benign colon and colon cancer. Glands in the benign colon are more similar to glands in the IBD cases used for training, whereas cancerous glands in the colon represent a domain shift.

Accordingly, we obtain Dice scores for both CRAG and GLAS greater than 0.85 for benign glands using models trained on EpCAM and CK8/18 masks, while Dice scores for cancer glands amount to 0.5 - 0.6 (Figure 3-E and 3-F). The CK8/18-derived models outperform EpCAM for both benign and cancer glands. A similar result is true for the Jaccard scores shown in Figure SF3. Altogether, models trained on a single institution cohort using an automated gland annotation pipeline perform adequately on external public data from multiple institutions with similar pathology. These results demonstrate the broad applicability of models trained on automated annotations, which can be generated by an efficient and hands-off process and without the manual labor of pathologists.

To improve the segmentation of glands in public datasets, we evaluated three strategies. Results are shown in Figure 4.(a) First, we trained a baseline U-Net model on the GLAS and CRAG training data and applied the model to the held-out test datasets from each cohort. The resulting gland outlines by the baseline model are shown in an example patch from the test data for both benign and cancerous glands (Figure 4-A and 4-B, column 1). We also trained the baseline model on increasing amounts of training data from GLAS and CRAG. The results are shown in Figure 4-C and 4-D (red line) and demonstrate, as expected that, the performance of the model increases as more data are used for training. (b) Next, we evaluated the performance of the model that was trained on EpCAM-derived gland outlines (out-of-box model) (Figure 4-A and -B). Gland outlines generated by the out-of-box model in GLAS or CRAG test images were more accurate for normal compared to cancer glands. While the model identified pixels within glands well, it failed to detect the boundaries of the glands. The out-of-box model performs as well as the baseline model when the baseline model is trained on only 20 percent of GLAS or CRAG training cases. However, when using more training data, the baseline model outperforms the out-of-box model.(c) To determine whether fine-tuning (Figure 4) improves the out-of-box model we used GLAS and CRAG data to retrain the model. The fine-tuned models show a slight gain in performance but are making mistakes similar to the out-of-box models, even with increasing amounts of training data (Figure 4-C and -D(blue lines)). (d) Finally, we added incremental amounts of data from GLAS and CRAG to the entire internal dataset while training the network de novo. Results of these mixed data sampling models show superior performance even when only 10-20% of GLAS or CRAG data are added to our internal data. The gland outlines generated by the mixed model possess accurate boundaries for benign and cancerous glands. Figures 4-C and 4-D (black lines)

demonstrate that including samples from the target cohort, in this case CRAG or GLAS cases, and de novo training of the model, result in superior performance.

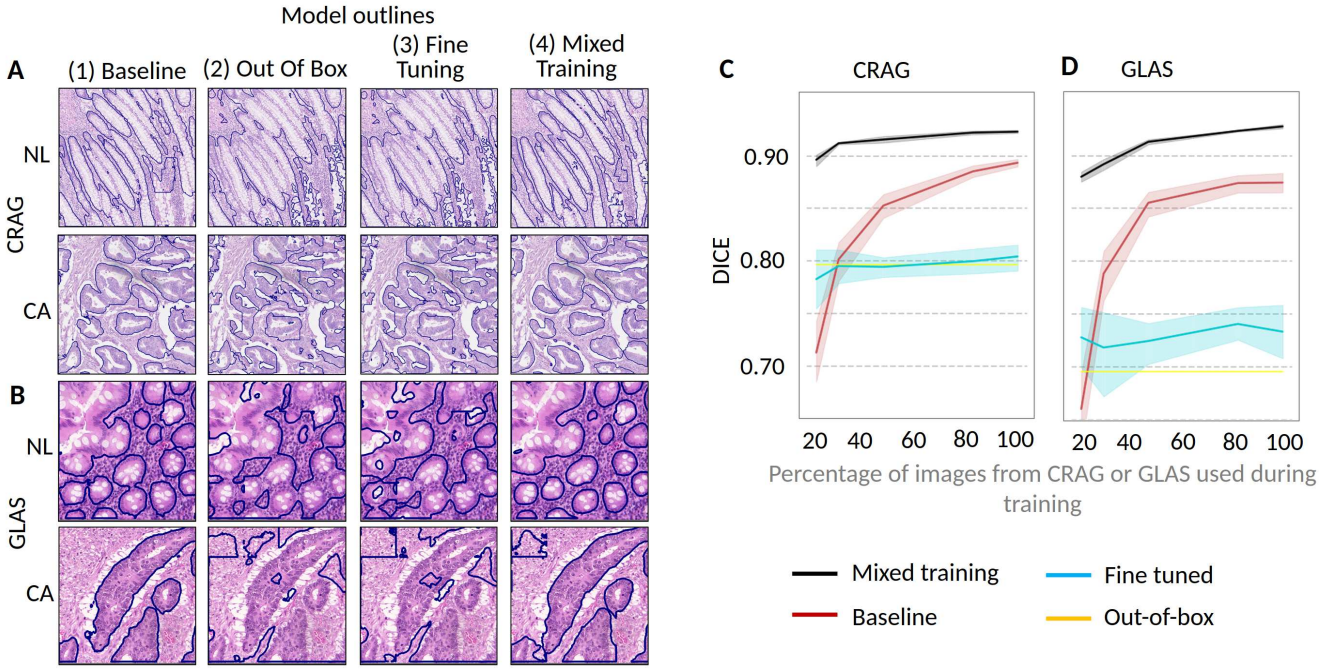

Figure 4. Domain adaptation of segmentation model. A. CRAG and B. GLAS dataset. Gland outlines are generated by four models in the test cohort: (1) A baseline model is trained on manual gland outlines in the training cohort of the CRAG or GLAS datasets and applied to the respective test cohorts. The baseline model does not use internal IBD data; (2) The out-of-box model is trained on EpCAM-derived gland outlines; (3) The fine-tuned model consists of the out-of-box model fine-tuned using CRAG and GLAS cases; (4) The mixed model is trained on data from CK8/18-derived gland outlines and in addition, training cases from the CRAG and GLAS training set. **C. CRAG and D. GLAS, Mixed model testing**. The percentage of CRAG and GLAS training cases added to the whole dataset of out-of-box training tiles is indicated on the X-axis. The Dice scores of the baseline, out-of-box, and fine-tuned models are compared to models trained on internal data with increasing proportions of CRAG or GLAS training data.

Next, we determined whether using either GLAS or CRAG data is sufficient to improve the out-of-box model. Models trained on the additional cases from a single dataset did not perform as well compared to models trained on additional cases from both datasets. However, adding images from CRAG versus training data demonstrates a greater improvement in performance (Figure SF9). This result indicates that CRAG data is more diverse than GLAS data, which helps the model achieve better performance.

As a final comparison, we trained U-Net models separately on GLAS or CRAG training data and tested the models on GLAS or CRAG test cohorts as well as on CDX2-, EpCAM- or CK8/18-derived gland masks. The models trained on or both GLAS and CRAG performed well on their own test cohorts but their performance on IHC-derived gland masks declined (SF9). This result further supports the need for multi-source sampling that we applied in Figure 4-C,D.

3.4 Lumen Segmentation

Lumen segmentation is critical for the identification of luminal neutrophils (crypt abscesses), which are a major component of inflammation in IBD cases. The color and texture of lumens resemble those of background regions and cytoplasmic mucin-filled goblets, which confuses deep-learning models. Additionally, lumens are small objects which generate a big imbalance of pixels versus background regions. This is detrimental to deep learning models, which in general perform better on large objects. Furthermore, lumen outlines extracted from the automated IHC-derived pipeline are noisier compared to gland outlines. Altogether, the lumen segmentation models we trained did not generalize well to unseen data. We used manually annotated lumens as ground truths for comparison to lumens predicted by a U-Net model. Mdels trained on images at x5 magnification performed better than those trained on x10 or x20 magnification. Based on the data imbalance, we propose that

model performance could be improved by using focal loss[43], which is effective when imbalanced data are used for training. The results of the lumen segmentation experiment are shown in SF4. The EpCAM IHC-derived models perform better when compared to CK8/18-based models, both in terms of segmentation metrics and visual results.

4 Discussion

We propose an automated pipeline for the annotation of single cells in H&E stained tissue sections. The method overcomes the need for manual annotations by pathologists and provides ground truth labels at pixel-level accuracy. As a test case, we used biopsies from individuals diagnosed with IBD, a condition that is characterized by acute and chronic inflammation in the colon. Over 30 biopsies are obtained for surveillance on an annual basis from individuals afflicted by IBD. The inflammation in these biopsies can distort the gland architecture by damaging the glandular epithelial lining. The inflammation and destruction of glands also increase the risk of colon cancer. Microscopic evaluation of more than 30 biopsies per individual is time-consuming, and reporting the extent of inflammation in each biopsy is cumbersome and not standardized. Therefore, a computer-assisted approach for the evaluation of IBD biopsies would improve patient management.

Gland segmentation by deep learning models can assist with the evaluation of colon biopsies. To offset the need for manual annotations of glands that are required for training and testing of deep learning models, we developed an automated pipeline. Glands are labeled by epithelial IHC markers in a tissue piece, and the IHC-derived mask is transferred to the H&E image of the exact same tissue. The gland masks in the H&E images are used for training and testing deep-learning models for gland segmentation. Our optimized pipeline provides consistent results for this gland segmentation task and demonstrates improved performance when the IHC markers stain the membrane or the cytoplasm of colonocytes. The outlines generated by our trained models also agree with manual annotations, validating the accuracy of the approach. Finally, we propose a method to test the generalization of the model across different domain shifts, such as tissue collection sites and normal versus cancer glands, by using two public datasets. We report average Dice scores on both public CRAG (Dice = 0.927) and GLAS (Dice = 0.922) datasets using an optimized data sampling method. The Dice scores we obtain are greater than previously reported Dice scores of 0.902 and 0.909[44,15,45–48]. Other publications[49–51] report "Object Dice" scores, the weighted mean of Dice scores, which does not allow a direct comparison to the Dice scores from our models.

The methods we developed can easily be deployed. To increase the reproducibility of IHC-stained slides, our IHC tissue staining was performed in a CLIA/CAP-certified laboratory. If other laboratories use different staining protocols, our thresholding method to generate IHC masks can easily be applied at other sites. Choosing the correct IHC marker for an application becomes an important task. Antibodies with high specificity for a given cell type and a large signal-to-noise ratio are best suited for reliable ground truth generation in automated annotation pipelines. To assure the best performance, we carefully tested all the hyperparameters in the model. This allows us to recommend an optimal model architecture, magnification, and IHC markers for glands in the colon. We observed that the optimal magnification for training depends on the size of the cell structure and tissue compartment of interest. For example, x20 magnification works best for glands, while x5 works best for lumen segmentation.

Histopathology protocols and tissue structures can vary, creating a domain shift even within the same disease type. Models trained on data from one site experience a significant decrease in performance when tested on data from multiple sites, which is attributed to the texture bias of deep learning models. In our experiments, models trained on normal glands performed significantly worse on glands from colon cancer (Figure 3). However, a small amount of annotated external data from tissues in the target cohort (5 percent <10 samples) resulted in large performance increases and improved generalization. We trained models on mixed data from our internal cohort and a few images from the target cohort, which outperformed models trained on 100 percent of training data from the target cohort (Figure 4). These results demonstrate that for low-data resource applications, combining data from multiple cohorts during training will result in models that generalize effectively. We also observed that adding data from multiple sources reduces the bias in the model compared to a single outside source (Figure SF7).

Limitations The proposed method for colonic gland annotations assumes that IHC markers are available. However, specific IHC markers are not available for many cell types, and the workflow to apply them to the exact same tissue section that was used for the digital H&E slide requires a tissue staining laboratory that is fully automated. This is a costly setup that does not exist in all pathology laboratories. Furthermore, pathologists have to be intimately involved in the development of algorithms, necessitating a transdisciplinary collaboration.

In our staining protocol, slides were first stained with H&E, and then after decolorization re-stained with IHC. This ensures that co-registration can occur with single-cell accuracy since only minor deformations exist across the H&E and IHC images. A

scenario where H&E - and IHC–stained images are derived from adjacent slides normally encompasses the clinical workflow in the real world. The proposed method of gland segmentation may still work with adjacent sections but may fail if there is more separation between H&E and IHC slides.

Our pipeline works well for gland segmentation in the normal colon but will need to be further optimized for lumen segmentation. The color and texture characteristics of pixels in the lumen are similar to background pixels and also to pixels in goblet cells. This problem is unique to differentiated epithelial cells in the colon, which contain goblets inside the cytoplasm. Because of the way the tissue is sectioned, there may not be a distinct separation between goblets and lumen, which limits the successful application of a semantic segmentation approach to outline the lumen. On visual examination, lumen segmentation of colon cancer glands works well, suggesting that lumen segmentation using our pipeline may work better in other glandular tissues, such as prostate or breast tissues, which do not contain goblet cells.

The approach using an admixture of images from different sources and cohorts for training may also have a few disadvantages. It requires that both training data and images from other target data sets are available for training. In addition, because we are training models de novo and not through transfer learning, the training may take longer and requires more computational resources.

5 Conclusion and Future Work

Large quantities of deeply annotated pathology images are required to train deep-learning algorithms and to optimize their performance. The proposed method permits the collection of an unlimited quantity of automatically annotated data, hence eliminating the need for manual annotation by pathologists, one of the key obstacles to the training of algorithms for computer-assisted diagnosis. In our IBD use case, EpCAM- and CK8/18-derived ground truth annotations of glands in the colon achieve the highest level of concordance with manual annotations. Moreover, the proposed data mixing strategy for domain adaptation consistently outperforms other highly supervised and fine-tuning approaches for gland segmentation. We are extending our methodology to multiple immune cell-type and nuclear annotations. The ultimate goal is to automate the assessment of disease activity in biopsies obtained for diagnosis or surveillance of IBD. By automating the assessment of inflammation in IBD biopsies, data can be used with various classification schemes to communicate the severity of the disease to the gastroenterologist. Furthermore, the quantification of gland distortion, and the enumeration and spatial quantification of immune cells in IBD, can be used in the future as a starting point for prognostic and treatment-related biomarker development.

Data Availability

Data from the University of Utah is available after the execution of a data-sharing agreement. Contact one of the corresponding authors of the paper for further details.

References

1. Liu, S. *et al.* Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1815–1824 (2022).
2. Dahlhamer, J. M., Zammitti, E. P., Ward, B. W., Wheaton, A. G. & Croft, J. B. Prevalence of inflammatory bowel disease among adults age 18 years—united states, 2015. *Morb. mortality weekly report* **65**, 1166–1169 (2016).
3. Metter, D. M., Colgan, T. J., Leung, S. T., Timmons, C. F. & Park, J. Y. Trends in the us and canadian pathologist workforces from 2007 to 2017. *JAMA network open* **2**, e194337–e194337 (2019).
4. Jajosky, R. P., Jajosky, A. N., Kleven, D. T. & Singh, G. Fewer seniors from united states allopathic medical schools are filling pathology residency positions in the main residency match, 2008-2017. *Hum. Pathol.* **73**, 26–32 (2018).
5. Arvaniti, E. *et al.* Automated gleason grading of prostate cancer tissue microarrays via deep learning. *Sci. reports* **8**, 1–11 (2018).
6. Eaden, J., Abrams, K., McKay, H., Denley, H. & Mayberry, J. Inter-observer variation between general and specialist gastrointestinal pathologists when grading dysplasia in ulcerative colitis. *The J. Pathol. A J. Pathol. Soc. Gt. Br. Irel.* **194**, 152–157 (2001).
7. Farmer, M., Petras, R. E., Hunt, L. E., Janosky, J. E. & Galandiuk, S. The importance of diagnostic accuracy in colonic inflammatory bowel disease. *The Am. journal gastroenterology* **95**, 3184–3188 (2000).
8. Tellez, D. *et al.* Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. image analysis* **58**, 101544 (2019).

9. Cui, M. & Zhang, D. Y. Artificial intelligence and computational pathology. *Lab. Investig.* **101**, 412–422 (2021).

10. Bryant, R. V. *et al.* Beyond endoscopic mucosal healing in uc: histological remission better predicts corticosteroid use and hospitalisation over 6 years of follow-up. *Gut* **65**, 408–414 (2016).

11. Park, S., Abdi, T., Gentry, M. & Laine, L. Histological disease activity as a predictor of clinical relapse among patients with ulcerative colitis: systematic review and meta-analysis. *Off. journal Am. Coll. Gastroenterol. ACG* **111**, 1692–1701 (2016).

12. Narang, V. *et al.* Association of endoscopic and histological remission with clinical course in patients of ulcerative colitis. *Intestinal research* **16**, 55 (2018).

13. Brandtzaeg, P., Haraldsen, G. & Rugtveit, J. Immunopathology of human inflammatory bowel disease. In *Springer seminars in immunopathology*, vol. 18, 555–589 (Springer, 1997).

14. Pai, R. K., Lauwers, G. Y. & Pai, R. K. Measuring histologic activity in inflammatory bowel disease: Why and how. *Adv. anatomic pathology* **29**, 37–47 (2022).

15. Graham, S. *et al.* Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images. *Med. image analysis* **52**, 199–211 (2019).

16. Sirinukunwattana, K. *et al.* Gland segmentation in colon histology images: The glas challenge contest. *Med. image analysis* **35**, 489–502 (2017).

17. Ibtehaz, N. & Rahman, M. S. Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural networks* **121**, 74–87 (2020).

18. Chen, H., Qi, X., Yu, L. & Heng, P.-A. Dcan: deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2487–2496 (2016).

19. Graham, S. *et al.* Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Analysis* **58**, 101563 (2019).

20. He, H. *et al.* Cdnet: Centripetal direction network for nuclear instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4026–4035 (2021).

21. Edlund, C. *et al.* Livecell—a large-scale dataset for label-free live cell segmentation. *Nat. methods* **18**, 1038–1045 (2021).

22. Bulten, W. *et al.* Epithelium segmentation using deep learning in h&e-stained prostate specimens with immunohistochemistry as reference standard. *Sci. reports* **9**, 1–10 (2019).

23. Brázdil, T. *et al.* Automated annotations of epithelial cells and stroma in hematoxylin–eosin-stained whole-slide images using cytokeratin re-staining. *The J. Pathol. Clin. Res.* **8**, 129–142 (2022).

24. Komura, D. *et al.* Restaining-based annotation for cancer histology segmentation to overcome annotation-related limitations among pathologists. *Patterns* **4** (2023).

25. OpenCV. Open source computer vision library (2015).

26. Macenko, M. *et al.* A method for normalizing histology slides for quantitative analysis. In *2009 IEEE international symposium on biomedical imaging: from nano to macro*, 1107–1110 (IEEE, 2009).

27. Szeliski, R. *Computer vision: algorithms and applications* (Springer Nature, 2022).

28. Avants, B. B., Tustison, N., Song, G. *et al.* Advanced normalization tools (ants). *Insight j* **2**, 1–35 (2009).

29. Malinin, A. *et al.* Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455* (2021).

30. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, 2015).

31. Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125 (2017).

32. Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).

33. Iakubovskii, P. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch (2019).

34. Huang, H. *et al.* Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1055–1059 (IEEE, 2020).

35. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, 3–11 (Springer, 2018).

36. Cao, H. *et al.* Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537* (2021).

37. Yang, Z., Ran, L., Zhang, S., Xia, Y. & Zhang, Y. Ems-net: Ensemble of multiscale convolutional neural networks for classification of breast cancer histology images. *Neurocomputing* **366**, 46–53 (2019).

38. Ding, X., Zhang, X., Han, J. & Ding, G. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11963–11975 (2022).

39. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

40. Kirillov, A., He, K., Girshick, R. & Dollár, P. A unified architecture for instance and semantic segmentation.

41. Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* **10**, 1895–1923 (1998).

42. Xu, M. *et al.* The differences and similarities between two-sample t-test and paired t-test. *Shanghai archives psychiatry* **29**, 184 (2017).

43. Jadon, S. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 1–7 (IEEE, 2020).

44. Graham, S. *et al.* One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification. *Med. Image Analysis* **83**, 102685 (2023).

45. Weiler, M., Hamprecht, F. A. & Storath, M. Learning steerable filters for rotation equivariant cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 849–858 (2018).

46. Ding, S., Wang, H., Lu, H., Nappi, M. & Wan, S. Two path gland segmentation algorithm of colon pathological image based on local semantic guidance. *IEEE J. Biomed. Heal. Informatics* (2022).

47. Wang, X. *et al.* Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Analysis* **81**, 102559 (2022).

48. Zheng, J., Liu, H., Feng, Y., Xu, J. & Zhao, L. Casf-net: Cross-attention and cross-scale fusion network for medical image segmentation. *Comput. Methods Programs Biomed.* **229**, 107307 (2023).

49. Wang, H., Xian, M. & Vakanski, A. Ta-net: Topology-aware network for gland segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1556–1564 (2022).

50. Wen, Y., Chen, L., Deng, Y., Zhang, Z. & Zhou, C. Pixel-wise triplet learning for enhancing boundary discrimination in medical image segmentation. *Knowledge-Based Syst.* **243**, 108424 (2022).

51. Dabass, M., Dabass, J., Vashisth, S. & Vig, R. A hybrid u-net model with attention and advanced convolutional learning modules for simultaneous gland segmentation and cancer grade prediction in colorectal histopathological images. *Intell. Medicine* 100094 (2023).

Acknowledgements

We thank the Department of Pathology and the Kahlert School of Computing at the University of Utah for their support of this project. We also thank Benjamin J. Brintz Ph.D., for statistical consultations.

Author contributions statement

Training of algorithms and code (TK, SR), conceptual framework (TK, BK, SE, SR), and manual annotations (MB, JJ, BK, ABA). (TK, BK,MB, JJ,SE) wrote the manuscript text and figures. All authors reviewed the manuscript.

Additional information

The authors have no competing interests. The study was conducted under IRB numbers: IRB_00140202 and IRB_00057287.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- IBDModernPathologysupplementary.pdf
- IBDModernPathologysupplementary.pdf
- IBDModernPathologysupplementary.pdf