Efficient quality control of whole slide pathology images with human-in-the-loop training

Abhijeet Patil [a],[*], Harsh Diwakar [a], Jay Sawant [a], Nikhil Cherian Kurian [a], Subhash Yadav [b], Swapnil Rane [b], Tripti Bameta [b], Amit Sethi [a]

[a] *Indian Institute of Technology, Bombay, Mumbai, India*
[b] *Tata Memorial Centre - ACTREC, HBNI, Navi Mumbai, India*

ARTICLE INFO

ABSTRACT

Histopathology whole slide images (WSIs) are being widely used to develop deep learning-based diagnostic solutions, especially for precision oncology. Most of these diagnostic softwares are vulnerable to biases and impurities in the training and test data which can lead to inaccurate diagnoses. For instance, WSIs contain multiple types of tissue regions, at least some of which might not be relevant to the diagnosis. We introduce HistoROI, a robust yet lightweight deep learning-based classifier to segregate WSI into 6 broad tissue regions—epithelium, stroma, lymphocytes, adipose, artifacts, and miscellaneous. HistoROI is trained using a novel human in-the-loop and active learning paradigm that ensures variations in training data for labeling efficient generalization. HistoROI consistently performs well across multiple organs, despite being trained on only a single dataset, demonstrating strong generalization. Further, we have examined the utility of HistoROI in improving the performance of downstream deep learning-based tasks using the CAMELYON breast cancer lymph node and TCGA lung cancer datasets. For the former dataset, the area under the receiver operating characteristic curve (AUC) for metastasis versus normal tissue of a neural network trained using weakly supervised learning increased from 0.88 to 0.92 by filtering the data using HistoROI. Similarly, the AUC increased from 0.88 to 0.93 for the classification between adenocarcinoma and squamous cell carcinoma on the lung cancer dataset. We also found that the performance of the HistoROI improves upon HistoQC for artifact detection on a test dataset of 93 annotated WSIs. The limitations of the proposed model are analyzed, and potential extensions are also discussed.

Introduction

Computational pathology—the use of computational (specifically, deep learning) techniques for diagnostic analysis of digital pathology whole slide images (WSIs)—is well on its way to being incorporated for clinical diagnosis in the coming few years. However, variance in the quality of tissue and slide preparation as well as scanning of WSIs, remains a concern in ensuring that deep learning models trained on curated datasets during the development phase work well during deployment as well. Due to this reason and also due to the lack of sufficient computational power and software libraries that can handle WSIs, most of the research on computational pathology in the last few years was concentrated on the analysis of clean and carefully curated datasets of patches (sub-images of manageable size) extracted from histopathology WSIs.[1,2] With increasing computational resources, research on using WSIs for diagnosis has gained traction, where either region-level annotations or weakly supervised learning methods are used. The former is laborious and prone to human bias and errors due to the

large (gigapixel) size of WSIs. On the other hand, while weakly supervised learning can eliminate the burden of regional annotations and reduce human bias,[3,4,5] it can introduce its own biases as we lose control over the accidental association of slide-level labels to even some of the irrelevant patches in a WSI.

In order to improve patch selection for various levels of supervision, we propose a fast, semi-automated, and iterative method that combines dataset preparation and model training to classify WSI patches into various commonly used tissue segment classes. Our method has the following advantages. Firstly, it keeps a human-in-the-loop for cluster level weakly supervised annotations. Secondly, its number of iterative model refinements can be manually controlled to trade-off between the quality of its results and annotation effort. Thirdly, the trained model generalizes from only one dataset to unseen datasets and organs for patch-level classification. And lastly, we show that the results of weakly supervised learning algorithms can be improved using the patch classification based on our model as a quality-control (QC or filtering) strategy for multiple problems and organs.

We elaborate upon the problem that we have solved as follows. Many diagnostic tasks, such as identifying cancer in breast tissue or subtypes of renal carcinoma, depend on the structure of epithelial cells.[6,3] Structural features of the stroma have been shown to be related to the survival and recurrence of breast cancer.[7,8] Similarly, the interaction between epithelial cells and lymphocytes is used as an indicator of survival in breast cancer patients.[9] However, most of these tasks do not depend on the analysis of the adipose (fat) regions or the artifacts. In fact, it has been shown in multiple studies that the performance of histopathology classification algorithms degrades if artifacts are not removed from the analysis.[10,11,12] Artifacts in a WSI can get introduced during various stages, including improper tissue fixation, irregularities in thickness while cutting tissues, tissue folds introduced while putting tissue sections on glass slides, over- or understaining, the introduction of foreign objects, the introduction of air bubbles while putting slide cover, marking with a pen over the tissue area, and improper focus while scanning. Artifact removal can further improve both strongly and weakly supervised classification pipelines. While neural networks take patches as input instances, in weakly supervised learning, the labels are available only at the bag level, where bag refers to a set of patches extracted from a WSI. Introducing too many irrelevant instances in a bag, such as those containing artifacts, can confuse neural networks.

Though a variety of histopathology datasets are publicly available on platforms, such as TCGA[13] and grand challenges,[14] there is no general and reliable solutions available for pre-processing patches pathology-specific labels, such as artifacts. A few solutions proposed earlier for the detection of artifacts in WSIs[15,16,17] provide good results only for specific organs and datasets. Detection of artifacts in WSI is challenging because the range of visual features associated with artifacts is very broad and prone to subjectivity. Another challenge in developing a model for whole slide image segmentation or patch classification is the unavailability of training data containing enough variations for generalization.

In this study, we present a lightweight classifier—HistoROI—to segregate patches of WSIs into one of the following classes—epithelium, stroma, lymphocytes, adipose, artifacts, and miscellaneous. Fig. 1 shows an input–output map of HistoROI on a sample WSI. HistoROI is trained with a novel human-in-the-loop training paradigm to ensure variation in training data and robust classification and to reduce the annotation and labeling burden on human experts. Our method runs a few iterations of over-clustering of patches, label assignment to pure clusters, sampling from the impure clusters for labeling, and improving the classifier.

We have validated HistoROI on a patch-based colon cancer dataset[18] to show its ability to generalize to organs other than the breast tissue on which it was trained. We have also demonstrated the applicability of the proposed model for detecting artifacts in WSIs. Further, we have shown improvement in the performance of the weakly supervised learning models for the CAMELYON[19] and TCGA-Lung[13] datasets after using HistoROI to remove obviously unwanted patches from WSI bags.

Specifically, our contributions are the following:

- We introduce HistoROI, the classification model trained using the human-in-the-loop training paradigm. This method makes optimal use of the annotators' time by ensuring presenting them only difficult and diverse patches. [Publically available HistoROI.]
- We prepared a dataset of histopathology patches with 6 different classes. This dataset contains more than 2 million patches with enough variation for each class. It includes a variety of artifacts and can be used to prepare quality control (QC) solutions. [Publically available HistoROI-dataset.]
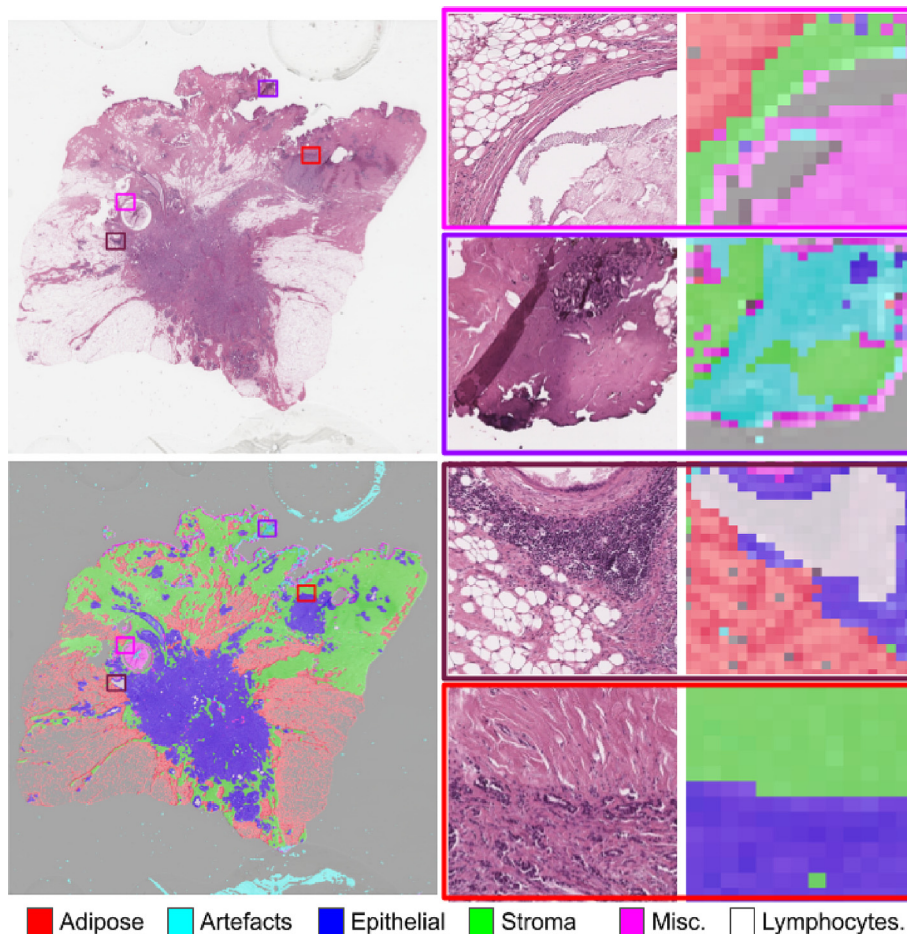

Fig. 1. An example of segmentation results on a whole slide image using HistoROI into epithelium, stroma, lymphocytes, adipose, artifacts, and miscellaneous areas.

- We show the utility of the proposed model on the CAMELYON and TCGA-Lung datasets. We have used the proposed model as a pre-processor for a classification algorithm and showed an increase in AUC from 0.88 to 0.92 on the CAMELYON dataset. We also explored the effect of quality control (QC) on the automatic WSI diagnosis pipeline. We have conducted a similar study on the TCGA-Lung dataset and demonstrated promising results. Though a few studies have been done on the effects of QC on diagnosis using patches of WSIs using simulated data,[10–12] to the best of our knowledge, this is the first study to show results on WSIs using real artifacts. We also present improvement in the explainability of weakly supervised learning models with the help of HistoROI.
- To validate HistoROI for QC, we annotated the foreground tissue region for 93 WSIs with a binary mask. This dataset can also be useful for the development and validation of novel QC solutions for WSIs. [Publically available TCGA-4Org.]

We introduce the datasets used for training and evaluation in Section Datasets. We describe our human-in-the-loop active learning method in more detail in Section Methods. We share the results of using HistoROI as a pre-processing tool, its applicability to the colon cancer dataset, and using it for QC in Section Results and discussion. We conclude in Section Conclusion.

Datasets

The datasets used in this study are summarized in Table 1. BRIGHT is a dataset with 503 WSIs labeled as cancerous, pre-cancerous, or non-cancerous.[6] We have used the patches from 50 WSIs of the BRIGHT dataset for the training of HistoROI. These WSIs were selected carefully to capture variations for better generalization. The detailed process of selecting these WSIs and patch-level data preparation is described in Section 3. This dataset contains more than 2 million labeled patches. Each patch is assigned 1 of the 6 possible labels—epithelium, stroma, lymphocytes, adipose, artifacts, or miscellaneous. WSIs from all the classes in the BRIGHT dataset were considered for the preparation of patch-level datasets. The artifact class includes patches with out-of-focus areas, tissue folds, cover slips, air bubbles, pen markers, and extreme over- or understaining.

CRC dataset[18] is a colon cancer dataset of 100 000 patches of size 224x224. The patches in this dataset are segregated into 9 different classes, viz. normal (NORM), tumor (TUM), mucin (MUC), muscle (MUS), background (BACK), debris (DEB), adipose (ADI), lymphocytes (LYM), and stroma (STR). Class-wise distribution of patches in this dataset can be found in Table 1.

Classification models trained using weakly supervised learning are trained and validated on CAMELYON16 and TCGA-Lung datasets.

Table 1
Summary of the datasets used in this study.

Dataset	Composition	Purpose
BRIGHT	2 169 355 Patches from 50 WSIs 787 168 - Epithelium, 863 989 - Stroma 98 293 - Lymphocytes, 245 525 Adipose 127 393 - Artifacts, 46 987 - Miscellaneous	Training and validation of HistoROI
CRC-100k	100 000 patches of colon cancer 10 407 - ADI, 8763 - NORM, 10 566 - BACK, 11 557 - LYM, 14 317 - TUM, 13 536 - MUS, 8896 - MUC, 10 446 - STR, 11 512 - DEB	Validation of HistoROI on external dataset
TCGA-4Org	93 multi-organ WSIs annotated with annotated tissue region	Application of HistoROI for QC
CAMELYON16	399 WSIs - 270/129 in train/test set Train - 70/100 positive/negative WSIs Test - 49/80 positive/negative WSIs	Application of HistoROI as WSI pre-processing
TCGA-Lung	1034 WSIs - 634/404 in train/test set Train - 316/318 adeno/squamous WSIs Test - 211/193 adeno/squamous WSIs	Application of HistoROI as WSI pre-processing

CAMELYON16 is a dataset of 399 WSIs of sentinel lymph nodes associated with breasts, with 270 images for training and validation and 129 WSIs for testing. The TCGA-Lung dataset contains 1034 WSIs. Further details of these datasets are shown in Table 1.

To validate the effectiveness of HistoROI for identifying artifacts in WSIs, we manually annotated a dataset of 93 WSIs to delineate the foreground tissue regions (excluding artifacts). These WSIs were selected from 4 different organs and contain various artifacts and a wide range of stain intensities and colors. All the WSIs were downloaded from the TCGA data portal. This dataset contains 27 WSIs from breast tissue, 21 WSIs from lung tissue, and 21 and 24 WSIs from kidney and prostate tissues, respectively. This test dataset was selected from a different source and contains 3 additional organs to test if the model can strongly generalize after being trained solely on WSIs from the BRIGHT breast tissue dataset.[6]

Methods

HistoROI is a ResNet18-based 6-class classifier that assigns a pathology-relevant label to a patch extracted from a WSI. To reduce annotation time for a WSI, we have taken advantage of features from shallow layers of pre-trained CNNs. We have leveraged a novel human-in-loop active learning approach to ensure diversity in the annotated datasets. With the help of k-means clustering and easy-to-use user interfaces, the annotation of each WSI takes less than 15 min. The details of dataset preparation and training of HistoROI are explained in this section.

Annotation of a single WSI: We have used a clustering-based approach to annotate patches of WSIs. We extract patches of size 256x256 at 10x magnification (approx. 1 μm × 1 μm per pixel dimension) from a WSI. Patches with mostly white pixels are not considered for data annotation by removing those with more than 95% of pixels with average RGB values greater than 230. The rest of the patches are given as input to an EfficientNet-b0[20] CNN architecture pre-trained on ImageNet to extract instance (patch) features. Specifically, we extract a tensor with 40 feature channels from block-2 of EfficientNet-b0 and use global average pooling to get a 40-dimensional feature vector for a patch. We have tried extracting features from other layers of EfficientNet-b0 and different versions of EfficientNets as well as ResNets and observed that clusters created using features of block-2 of EfficientNet-b0 are semantically more homogeneous (Fig. 11). While segregating patches of WSI, we mostly care about local features, such as texture, to distinguish between different types of tissue regions, such as stroma or nucleus-dense epithelium. The diameter of nuclei at a 10x magnification level is less than a few pixels, which can be learned by such local features. Therefore, we hypothesize that the identification of classes of our interest can better be done with local features extracted by block-2 of EfficientNet-b0 than the global features we get from deeper layers of pre-trained CNNs.

Features of all the patches in a WSI are then clustered using a k-means clustering algorithm, for which fast implementation is available in scikit-learn.[21] We learn 32 clusters and sample 25 patches from a cluster in the form of a 5x5 grid for annotation as shown in Figs. 8, 7, and 9. Upon manual inspection, if a grid of 5x5 patches turns out to be homogeneous—that is, it contains patches of the same class—then we assign one of the following labels to all the patches corresponding to that cluster: 1. Epithelium, 2. Stroma, 3. Lymphocytes, 4. Adipose, 5. Artifact, and 6. Miscellaneous. Patches from heterogeneous clusters are pooled together and re-clustered into 32 new clusters using another round of k-means, and the labeling process is repeated, as shown by a dotted line in Fig.2. Patches belonging to heterogeneous clusters after the second stage of this human-in-the-loop labeling are discarded. Generally, less than 5% of the patches are discarded from a WSI after the second round.

We initially created a training dataset with 20 WSIs and grew this dataset as the training process progressed. The annotation of each WSI took around 15 min.

Diversity-based expansion of annotations: WSI annotation is a time-consuming task. Therefore, it is important to annotate a diverse set of WSIs instead of wasting time and effort annotating multiple WSIs with similar
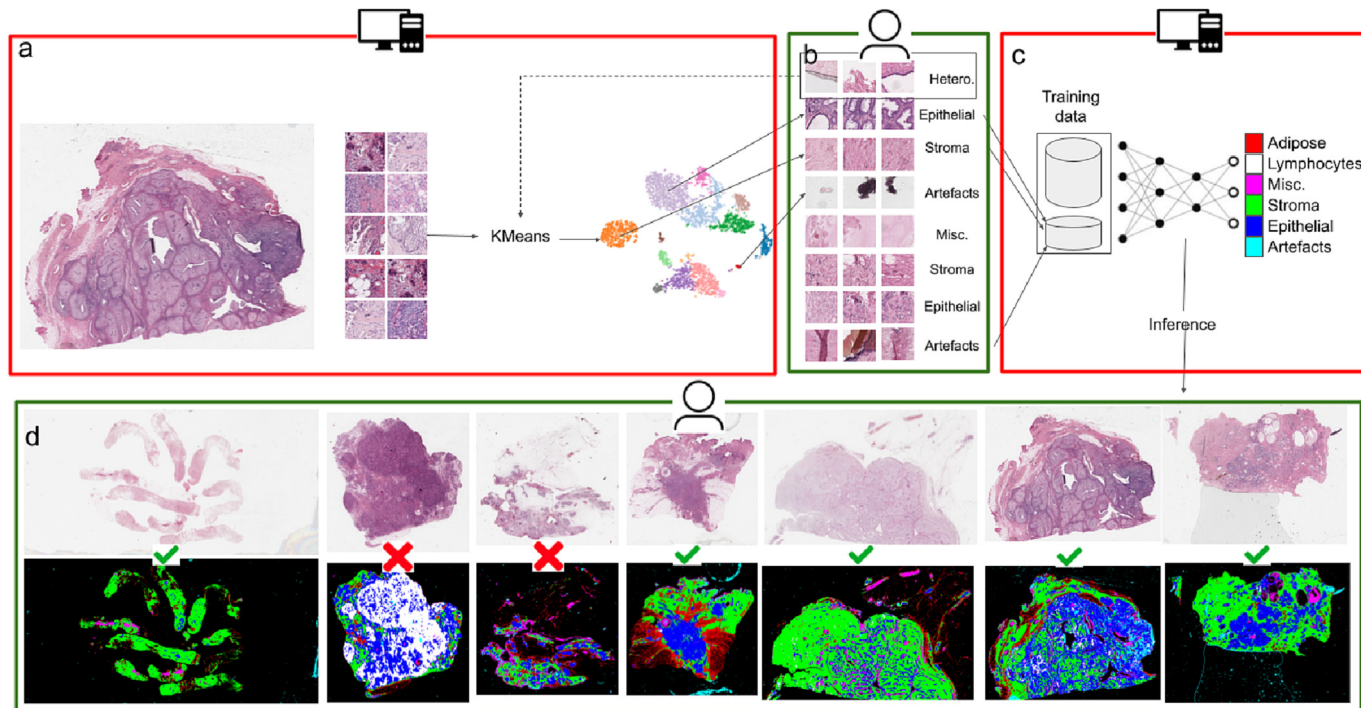
Fig. 2. Human-in-the-loop training pipeline for HistoROI: Actions in red boxes are automatic while those in green boxes are manual. (a) Embeddings of the patches of WSI are divided into clusters. (b) Clusters are manually annotated. Heterogeneous clusters are re-clustered (shown using a dotted line). (c) Newly annotated data is added to previously annotated data and HistoROI is trained with the updated data. (d) The trained model is applied (tested) on multiple WSIs. WSIs with poor performance (shown with X) are manually identified and annotated for the next iteration of training.

visual features. Diversity and label uncertainty are 2 major guiding principles of active learning approaches.[22] We trained the model and annotated additional WSIs with an approach inspired by active learning to address this problem.

We initially trained a ResNet18-based[23] 6-class classifier with the data from 20 WSIs. We used patches from 15 WSIs for training and 5 WSIs for validation. The classifier was trained on NVIDIA-3090Ti with a batch size of 128. Cross-entropy loss was optimized with Adam optimizer,[24] and the model with the least validation loss was applied on all the WSIs from the BRIGHT dataset. QuPath-compatible[25] visualizations were created for all images. WSIs and their predicted patch classifications were visually analyzed using QuPath as shown in Fig. 10. WSIs with poor performance indicate that the current WSI is out-of-

distribution with respect to the training dataset. That is, such WSIs has high label uncertainty, and it should be a good candidate for annotation in the next round to maximize information gain while minimizing annotation effort. These WSIs were manually identified, annotated, and added to the training dataset for further fine-tuning of the classifier. We repeated this cycle 3 times, adding 10 WSIs to the training dataset at a time. Thus, we created a training dataset with 50 carefully selected and annotated WSIs containing enough variation for generalization. The whole process of human-in-the-loop training is summarized in Fig. 2.

Classification models: As a representative of the downstream models that can benefit from HistoROI, we trained a widely used weakly supervised learning algorithm called CLAM for histopathology.[3] CLAM is trained with a multiple-instance learning paradigm, where the feature of a patch is used
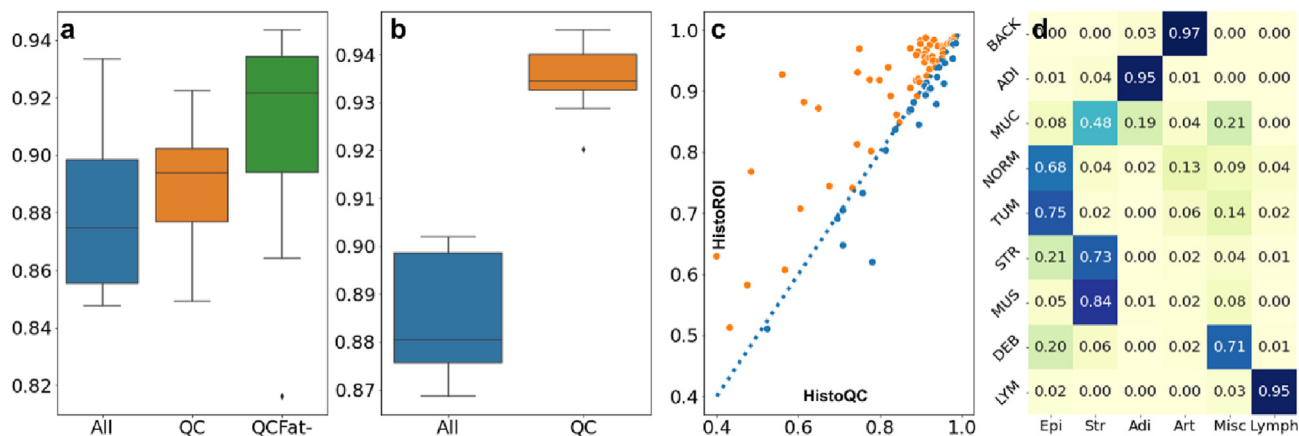
Fig. 3. HistoROI quantitative results: (a) Box plot of AUCs on the test split of the CAMELYON16 dataset shows improvement in performance when WSIs are pre-processed with HistoROI. (b) Box plot of AUCs for the TCGA-LUNSC dataset. (c) Comparison of HistoROI with HistoQC[15] for QC using TCGA-4Org dataset shows that HistoROI's Dice score is better than HistoQC for 65 WSIs out of 93 WSIs. (d) HistoROI predicts 77% of patches correctly in the CRC-100k without being trained on this dataset.

Fig. 4. Qualitative results on CRC-100k[18]: Example predictions (patch labels) on CRC-100k dataset by HistoROI for various classes (row labels) shows that misclassified patches (with red crosses) tend to have the background, staining issues, or multiple classes, which are not present in the correct classified patches (with green ticks). Here BACK stands for the background, ADI for adipose, MUC for mucus, NORM for normal, TUM for tumor, STR for stroma, MUS for muscle, DEB for debris, and LYM for lymphocytes.

as an instance, and the set of these instances from a WSI (either all patches or a subset thereof) is used to create a bag. CLAM has shown improved performance on multiple publicly available datasets.[13,19] We trained and validated CLAM in multiple patch pre-filtering settings to verify the effectiveness of the proposed classification method and model in improving the effectiveness of weakly supervised learning.

Results and discussion

HistoROI as WSI pre-processing tool

We first show the applicability of HistoROI as a pre-processing tool for a weakly supervised learning algorithm—CLAM.[3] Even though weakly
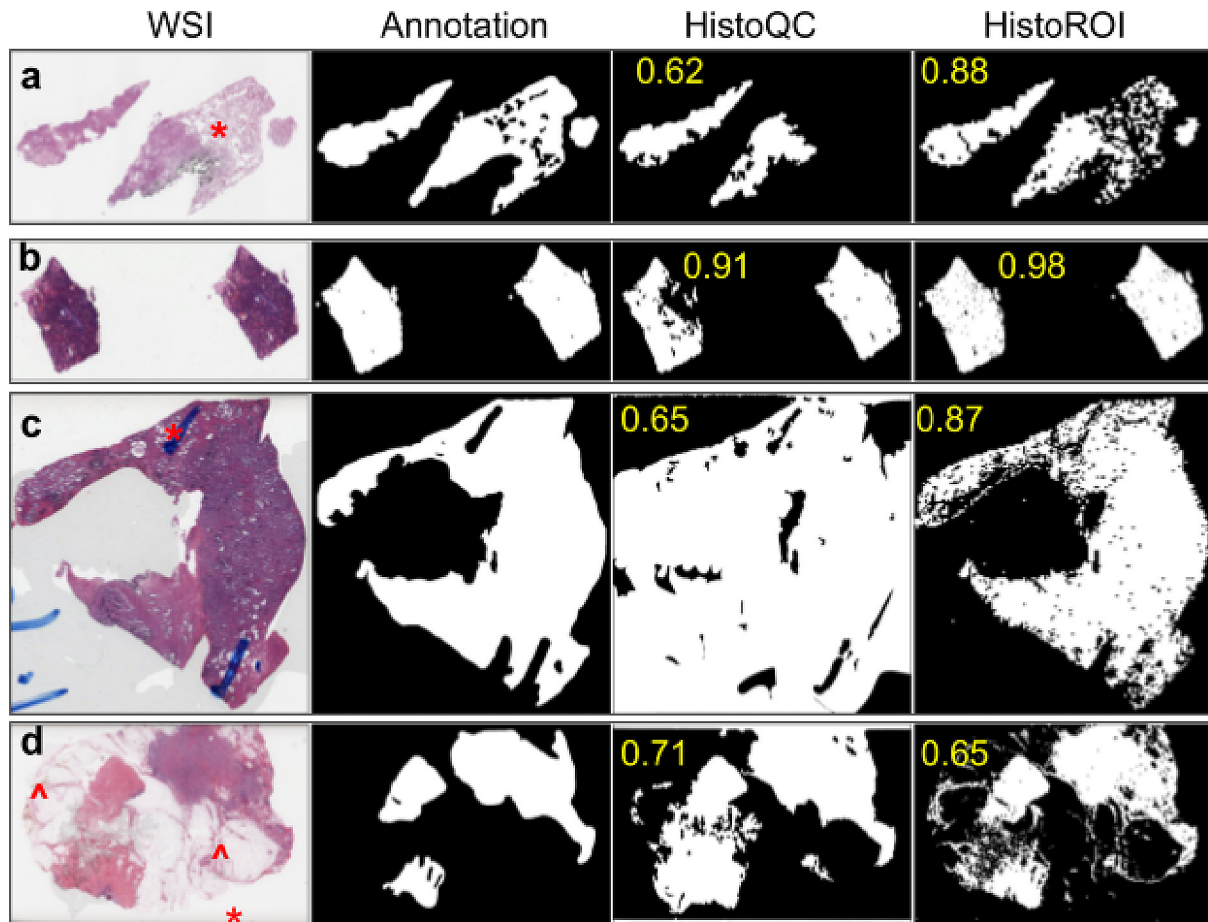
Fig. 5. Qualitative results on TCGA-4Org dataset: Foreground detection (with Dice score) for a few examples WSIs show improved foreground detected by HistoROI over HistoQC.
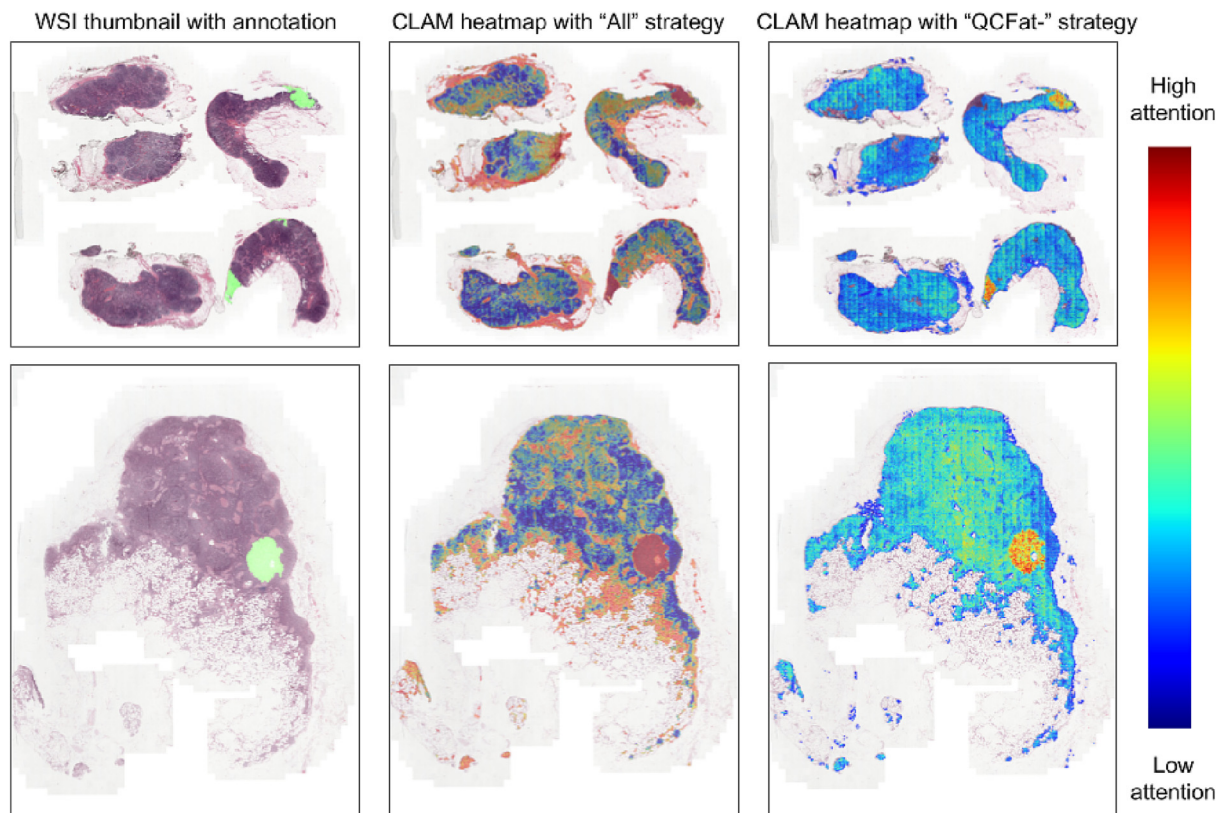

Fig. 6. Qualitative results on CAMELYON16: CLAM[3] trained on regions filtered by HistoROI generates better attention maps. Metastasized region is annotated with cyan color in the first column. CLAM attention map for the "All" strategy is scattered compared to the accurate and specific one generated using the "QCFat–" strategy.

supervised learning methods are supposed to filter out irrelevant regions, including artifacts, automatically, we show that these methods can still benefit from data pre-filtering based on HistoROI. As a baseline, we train CLAM with the bag prepared by features of all the patches from a WSI. Then, for comparison, we filter out patches from the bags based on the prediction of HistoROI and analyze changes in the performance of CLAM applied to this filtered data. More specifically, we perform 2 experiments. The first experiment shows the utility of HistoROI at train and test times. In this experiment, we use the CAMELYON dataset to demonstrate improvement in the performance of CLAM when data filtering is used at both training and test times. In the second experiment, we try to emulate a real-world scenario. Generally, training data goes through a few iterations of labeling/annotations before being fed to the deep learning algorithm. Through this process, training data is cleaned and curated, hence expected to contain lesser anomalies. On the other hand, test data cannot be manually curated. Hence, in the second experiment, we do not use data filtering on training data but filter the patches in test WSI using HistoROI at inference. We have used the TCGA-Lung dataset to distinguish between adenocarcinoma and squamous cell carcinoma for this experiment and have shown significant improvement in the performance when HistoROI is used for data filtering. This experiment also shows the possibilities of using HistoROI on algorithms that are independently developed. We have created bags for CLAM in 3 different ways in these experiments: (1) Bags with the features from all the non-white patches ("All"), (2) bags without patches identified as artifacts by HistoROI ("QC"), and (3) bags without artifacts and adipose patches ("QCFat-").

We trained CLAM 10 times with different train-validation splits on the CAMELYON16 dataset. We used the same test split created by the CAMELYON16 challenge organizers for testing. We used 80% of training WSIs for training and the remaining 20% for validation to tune hyperparameters. The mean AUC on test data was observed to increase from 0.88 to 0.90, while the median AUC increased from 0.87 to 0.92, as shown in Fig. 3 (a). Filtering out the patches with artifacts clearly improves the performance compared to using all the patches. Further removing patches with adipose tissue increased performance even further. We tried various other patch selection criteria along with different color normalization methods,[26,27] and did not notice a significant change in test accuracy. Interestingly, we noticed a significant reduction in test accuracy when we
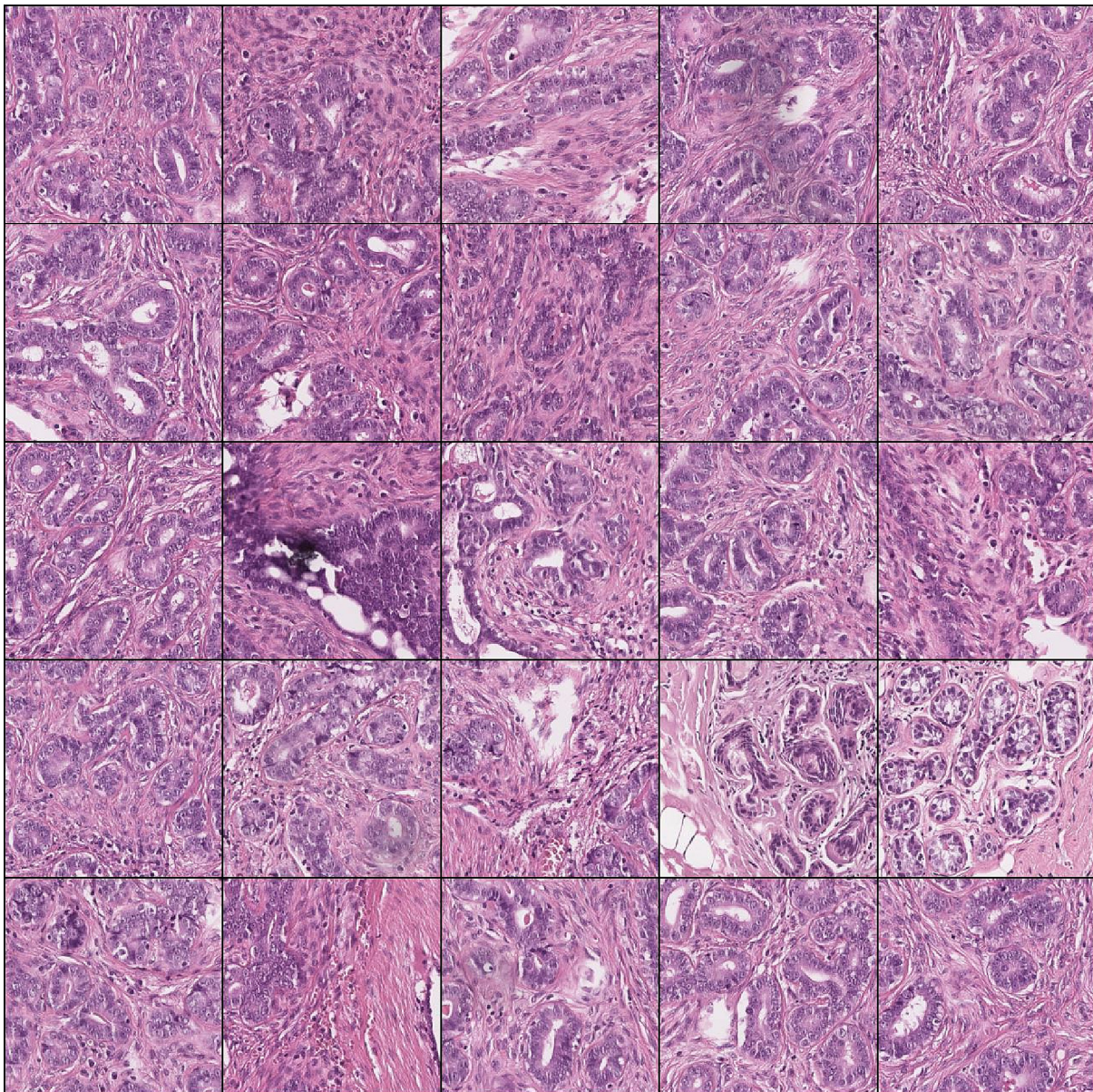

Fig. 7. Grid of 5x5 patches with predominant epithelium.

used only epithelial patches. Upon closer inspection, we noticed that for a few WSIs, most patches contain both epithelial nuclei and lymphocytes. HistoROI predicted those patches as lymphocytes, and hence informative patches were filtered out from the training algorithm. This observation highlights that assigning multiple labels for a patch can be more useful than assigning a single one.

For our next experiment, we assessed the impact of using HistoROI for a scenario in which the training data is relatively cleaner than the test data, which is to be expected in practice. We hypothesize that if real-world data is filtered using a reliable quality control pipeline, the performance of the model trained on relatively cleaner data can be improved during test time. To demonstrate this, we have conducted experiments with TCGA-Lung cancer data. For this experiment, we intentionally prepared a test split with WSIs that contain a relatively higher proportion of artifacts. The list of WSIs in the test split is listed on the webpage with the code of our method HistoROI. We trained CLAM 10 times with different train-validation splits and analyzed the change in performance at test time with

and without data filtering. Quantitative results are shown as boxplots in Fig. 3 (b). We observe that the performance of CLAM significantly improves when inferred WSI is pre-processed using HistoROI.

Along with better quantitative performance, the model trained with the assistance of HistoROI produces better explainability. We have visualized attention maps for a few WSIs trained with the "All" strategy and "QCFat–" strategies for the CAMLEYON dataset. As shown in Fig. 6, the attention map created by CLAM when trained with filtered patches is in concordance with the annotations, whereas CLAM trained with all the patches gives high attention to arbitrary locations, including regions containing adipose. According to our domain knowledge, features of the adipose region do not correlate with the presence of metastasis in lymph nodes. But high attention on the adipose region indicates the potential data bias with the correlation between the adipose region and the presence of metastasis. We have removed this potential bias by filtering out the adipose region from analysis and observed better results with the "QCFat–" strategy.
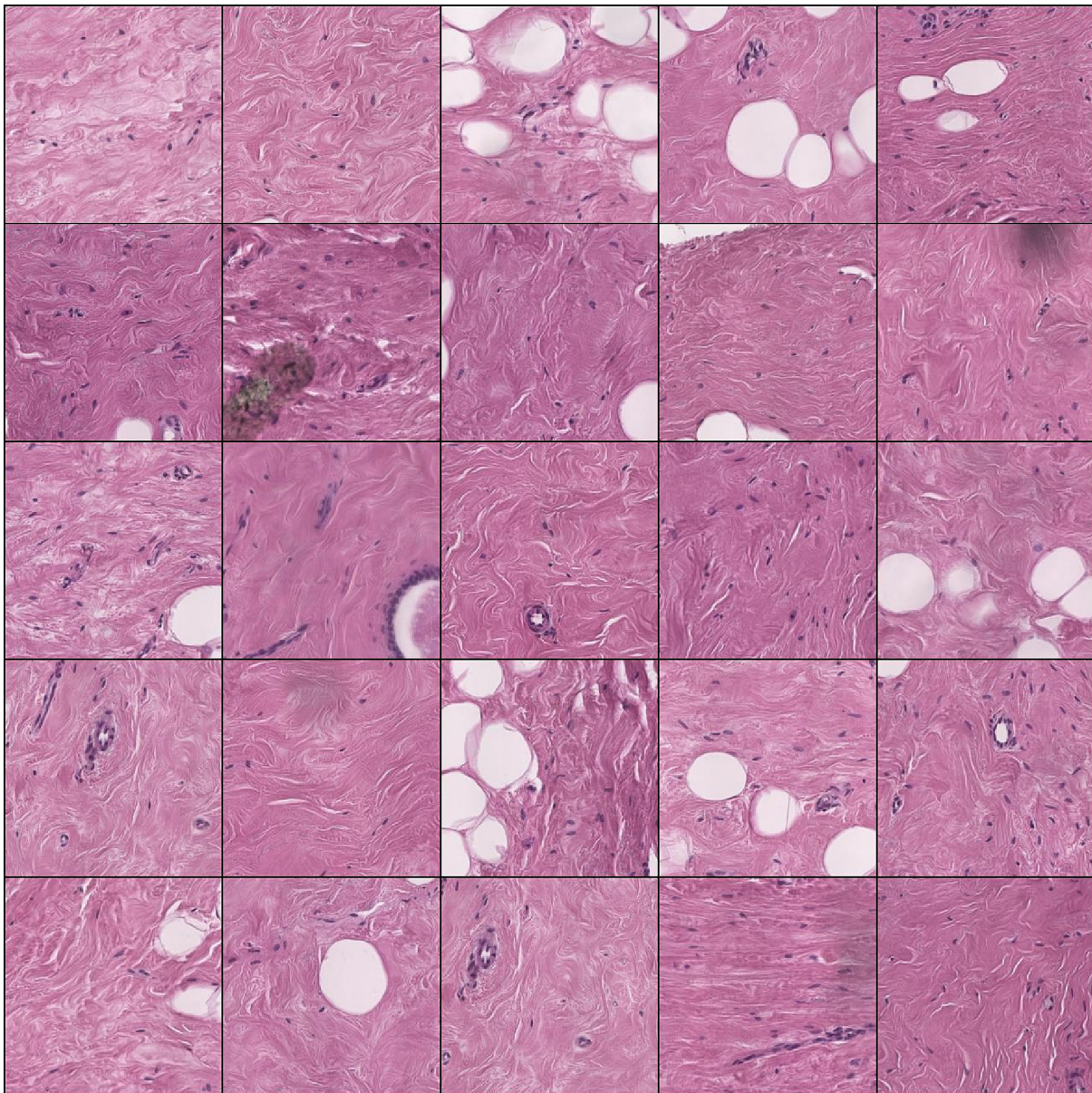

Fig. 8. Grid of 5x5 patches with predominant stroma.

Validation of HistoROI on CRC-100k dataset

We applied HistoROI to the CRC dataset to validate its generalization on the colon cancer dataset even though it was trained on the BRIGHT breast cancer dataset. The CRC dataset contains a few classes exclusive to the prediction classes of HistoROI. We expect the adipose (ADI) class of the CRC dataset to be predicted as adipose, normal (NORM) as epithelium, background (BACK) as artifacts, lymphocytes (LYM) as lymphocytes, tumor (TUM) as epithelium, mucus (MUC) and debris (DEB) as miscellaneous, and stroma (STR) and muscle (MUS) as stroma. HistoROI predicted 77% of images in the CRC-100k dataset correctly. HistoROI predicted more than 90% of patches in LYM, ADI, and BACK classes correctly. More than 70% of patches from the DEB class were correctly classified as miscellaneous. Twenty percent of DEB patches were predicted as epithelium. Most of these wrongly predicted patches contain a few epithelial nuclei in debris. For normal class (NORM), 13% of patches were predicted as artifacts. Our analysis indicates that a few of these wrongly predicted patches

contain blurry regions, and the remaining patches are overstained and thus contain some artifacts. Similarly, tumor patches misclassified as miscellaneous were observed to contain debris. Most classification errors were observed in the mucus (MUC) class. Most of the mucus patches were predicted as stroma by HistoROI as the 2 classes share visual features. Additionally, HistoROI has not seen patches with mucus while training on the BRIGHT dataset. Finally, it can be hard to distinguish between stroma and mucus in a small patch, to begin with. The dataset contains patches of size 224x244 at 20x, corresponding to 112x112 when downscaled to 10x for HistoROI.

Fig. 3 (d) shows quantitative results, and Fig. 4 highlights a few correctly and wrongly predicted patches for the CRC dataset. Analysis of these results can lead to better design of models and datasets. For example, a few patches in the NORM class that were incorrectly predicted as artifacts contain artifacts in the form of blur or dark stains. These observations highlight subjectivity in labeling histopathology patches. We propose that rather than assigning hard labels to a patch, a better strategy would be to
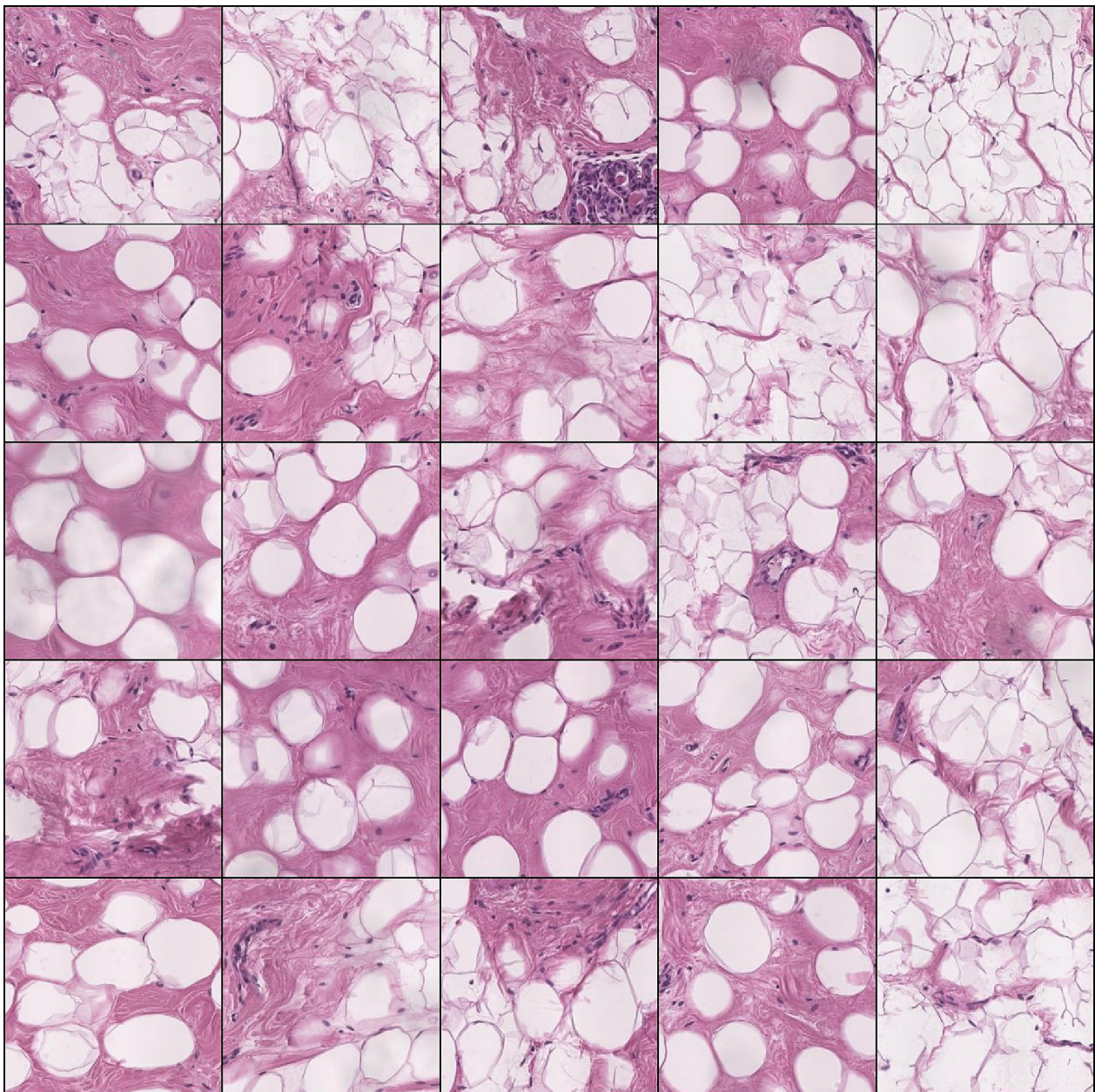

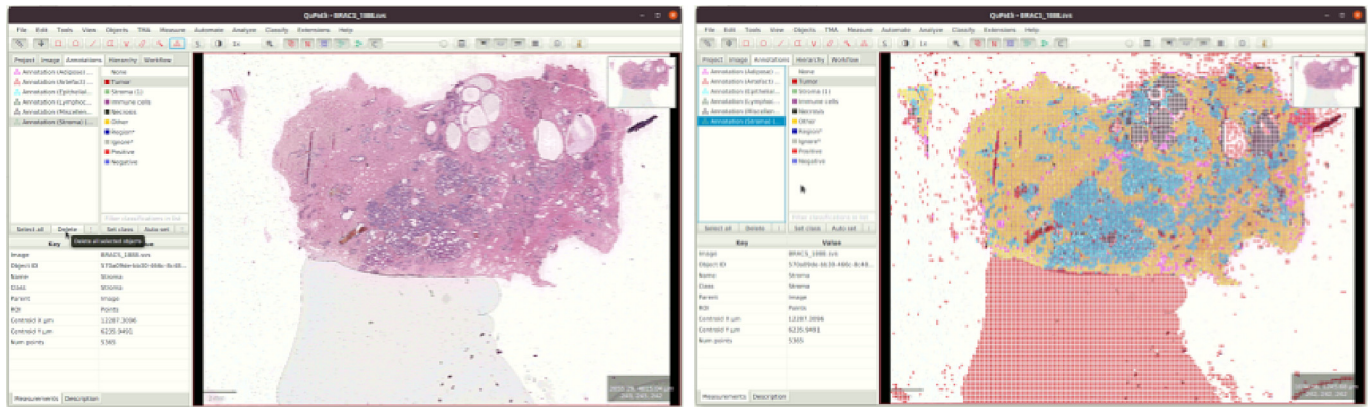**Fig. 9.** Grid of 5x5 patches with predominant adipose.

Fig. 10. QuPath point annotations for manual selection of WSIs.

assign multiple labels or assign a level of a particular class along with the label. Since we have trained a HistoROI with patches from the breast cancer dataset, the low accuracy for the MUC class shows that patches from multiple diverse organs should be used to train a robust model. The human-in-the-loop paradigm presented here provides an easy way to expand the dataset and the model generalization.

HistoROI as QC tool

The background (BACK) class in CRC dataset contains mostly white patches, which does not effectively validate predictions for the artifact class of HistoROI. To address this issue, we annotated the foreground region in 93 WSIs and compared predictions made by HistoROI with it. Annotated dataset contains WSIs from 4 different organs (27 from the breast, 21 from the lung, 21 from the kidney, and 27 from the prostate). We also compared predictions made by HistoQC[15]—a popular histopathology quality control tool—with HistoROI. The mean Dice score over WSIs between HistoROI and hand annotations was 0.87, whereas, for HistoQC, it was 0.83. A few qualitative results are shown in Fig. 5. The performance of HistoROI was better on 65 WSIs out of 93 WSIs, as shown in Fig. 3 (c).

According to our observations, HistoQC tends to identify the region with relatively less dense tissue as fat (denoted by * in Fig. 5 (a)). Also, HistoQC fails to distinguish between foreground and background when

Fig. 11. Visual comparison of Imagenet pre-trained features of ResNet-34 and EfficientNet-b0. (a) Shows a reference image. Each row contains 25 patches with minimum distance from a reference patch. Rows of patches are sorted in order of semantically most similar (row b) to least similar (row l). Row (b) corresponds to patches extracted from EfficientNetb0-block-2. Similarly, row (c) to (l) corresponds to features extracted from ResNet34-block-1, EfficientNet-b0-block-1, ResNet34-block-2, EfficientNet-b0-block-4, EfficientNet-b0-block-3, ResNet34-block-3, ResNet34-block-4, and block 5 to 7 of EfficientNet-b0, respectively.

background pixels are grayish. Because of this, the performance of HistoQC degrades in the presence of air bubbles (Fig. 5 (c)) and coverslip-related artifacts (denoted by * in Fig. 5 (d)). HistoROI performs better compared to HistoQC for the scenarios mentioned above.

On the other hand, HistoQC detects pen marks better than the proposed HistoROI. Further analysis of training data for HistoROI showed that it contains only 1 WSI with pen marks. These pen marks were also observed to be outside the tissue region. Hence, in Fig. 5 (c) (denoted by *), pen marks outside the tissue region are correctly identified as background by HistoROI. In contrast, the pen marker inside the tissue region is not delineated properly by HistoROI. Also, for a few WSIs, the Dice score for HistoQC was observed to be greater than that for HistoROI, even if visually, HistoROI performance is better. This uncovers difficulties in annotating WSIs. One such example is shown in Fig. 5 (d) (denoted by ^). Detailed analysis of the prediction of HistoROI can lead to guidelines for building the next generation of a QC solution, which we think will simply involve sampling from more diverse training data, such as one that contains more organs and slides with marker pen artifacts. Though the predictions of HistoROI are pixelated (because its model is based on patches), the overall Dice score is better than HistoQC. This indicates that the patch-based dataset contains enough variation to recognize most artifacts. A dataset for semantic segmentation with current data can be helpful for designing the next generation of QC solutions for histopathology images.

Conclusion

In this study, we have developed a human-in-the-loop active learning paradigm to prepare a patch-level dataset for classifying pathology patches. With the prepared data, we have trained a ResNet18-based 6-class classifier called HistoROI. To validate the generalization of HistoROI, we tested it for classifying colon tissue patches using the CRC dataset. We also explored the potential of HistoROI as a quality control (QC) tool for WSIs. Further, we investigated the use case for HistoROI for pre-processing WSIs in classification pipelines based on the widely used weakly supervised learning classification algorithm—CLAM. Through our experiments, we have shown improvement in the performance of CLAM when HistoROI is used for pre-filtering instances (patches) for each bag (WSI). We have also shown improvements over the widely used quality control tool, HistoQC, for the artifact detection. HistoROI was trained using the BRIGHT dataset, which is a breast cancer dataset. We conducted validation experiments utilizing datasets from multiple data centers and organs, such as TCGA, CAMELYON, and CRC-100k. The results demonstrate that HistoROI can generalize effectively on high resolution whole slide images. Nonetheless, it remains to be determined whether HistoROI can perform equally well on manually acquired microscope images, which may exhibit significant domain shifts.

Our analysis of HistoROI has brought to light certain limitations that impede its ability to accurately classify patches of histopathology images containing multiple tissue segments. These limitations stem from the single label assignment strategy to annotate the dataset, which restricts HistoROI to assigning a single label to each patch of a WSI. To address this issue, we propose several solutions, including optimizing patch size for annotation, employing segmentation masks instead of patch-wise labeling, and providing pathologists with easy-to-use interactive segmentation and active learning tools for faster annotation. Moreover, we suggest asking annotators to identify the presence of a certain class in a given patch rather than selecting from a set of labels, as well as utilizing multiple annotators to improve accuracy of annotations. By implementing these measures, we believe that HistoROI can be improved and better utilized in the field of histopathology analysis. Also, we propose that the performance of HistoROI can be improved by training it on multiple organs to develop more robust and generalizable models. For the artifact detection task, we have noticed that the performance of HistoROI does not detect pen marks well due to the lack thereof in its training data. Further analysis should be carried out to investigate cases where the proposed model cannot predict artifacts correctly.

This can lead to the preparation of better datasets using which the next version of HistoROI can be trained and released.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

References

1. Aresta G, Araújo T, Kwok S, Chennamsetty S, Safwan M, Alex V, et al. BACH: grand challenge on breast cancer histology images. Med Image Anal Aug. 2019;56:122–139. https://doi.org/10.1016/j.media.2019.05.010.
2. Spanhol FA, Oliveira LS, Petitjean C, Heutte L. Breast cancer histopathological image classification using Convolutional Neural Networks. 2016 International Joint Conference on Neural Networks (IJCNN); Jul. 2016. p. 2560–2567. https://doi.org/10.1109/IJCNN.2016.7727519.
3. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng Jun. 2021;5(6). https://doi.org/10.1038/s41551-020-00682-w. Art. no. 6.
4. Chen C, Lu MY, Williamson DFK, Chen TY, Schaumberg AJ, Mahmood F. Fast and scalable search of whole-slide images via self-supervised deep learning. Nat Biomed Eng Dec. 2022;6(12). https://doi.org/10.1038/s41551-022-00929-8. Art. no. 12.
5. Anand D, Yashashwi K, Kumar N, Rane S, Gann PH, Sethi A. Weakly supervised learning on unannotated H&E-stained slides predicts BRAF mutation in thyroid cancer with high accuracy. J Pathol Nov. 2021;255(3):232–242. https://doi.org/10.1002/path.5773.
6. Brancati N, De Pietro G, Riccio D, Frucci M. Gigapixel histopathological image analysis using attention-based neural networks. IEEE Access 2021;9:87552–87562. https://doi.org/10.1109/ACCESS.2021.3086892.
7. Xu S-J, Zhang S-Y, Dong L-Y, Lin G-S, Zhou Y-J. Dynamic survival analysis of gastrointestinal stromal tumors (GISTs): a 10-year follow-up based on conditional survival. BMC Cancer Nov. 2021;21(1):1170. https://doi.org/10.1186/s12885-021-08828-y.
8. Millar EK, Browne LH, Beretov J, Lee K, Lynch J, Swarbrick A, et al. Tumour stroma ratio assessment using digital image analysis predicts survival in triple negative and luminal breast cancer. Cancers Dec. 2020;12(12):3749. https://doi.org/10.3390/cancers12123749.
9. Zgura A, Galesa L, Bratila E, Anghel R. Relationship between tumor infiltrating lymphocytes and progression in breast cancer. Mædica Dec. 2018;13(4):317–320. https://doi.org/10.26574/maedica.2018.13.4.317.
10. Wright AI, Dunn CM, Hale M, Hutchins GGA, Treanor DE. The effect of quality control on accuracy of digital pathology image analysis. IEEE J Biomed Health Inform Feb. 2021;25(2):307–314. https://doi.org/10.1109/JBHI.2020.3046094.
11. Schömig-Markiefka B, Pryalukhin A, Hulla W, Bychkov A, Fukuoka J. Quality control stress test for deep learning-based diagnostic model in digital pathology. Mod Pathol Dec. 2021;34(12), 12. https://doi.org/10.1038/s41379-021-00859-x.
12. Kanse AS, Kurian NC, Aswani H, Khan Z, Gann PH, Rane S. Cautious artificial intelligence improves outcomes and trust by flagging outlier cases. JCO Clin Cancer Inform Oct. 2022;6:e2200067. https://doi.org/10.1200/CCI.22.00067.
13. Weinstein JN, Collisson WA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet Oct. 2013;45(10), 10. https://doi.org/10.1038/ng.2764.
14. Grand Challenge. *grand-challenge.org*. https://grand-challenge.org/. (accessed Feb. 03, 2023).
15. Janowczyk A, Zuo R, Gilmore H, Feldman M, Madabhushi A. HistoQC: an open-source quality control tool for digital pathology slides. JCO Clin Cancer Inform Apr. 2019;3:1–7. https://doi.org/10.1200/CCI.18.00157.
16. Chen Y, Zee J, Smith A, Jayapandian C, Hodgin J. Howell D.,et al. Assessment of a computerized quantitative quality control tool for whole slide images of kidney biopsies. J Pathol Mar 2021;253(3):268–278. https://doi.org/10.1002/path.5590.
17. Haghighat M, Browning L, Sirinukunwattana K, Malacrino S, Alham NK, Colling R. Automated quality assessment of large digitised histology cohorts by artificial intelligence. Sci Rep Mar. 2022;12(1), 1. https://doi.org/10.1038/s41598-022-08351-5.
18. Kather JN, Halama N, Marx A. 100,000 histological images of human colorectal cancer and healthy tissue. Zenodo Apr. 7, 2018. https://doi.org/10.5281/zenodo.1214456.
19. Bandi P, Geessink O, Manson Q, Dijk MV, Balkenhol M. Hermsen M.,et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 Challenge. IEEE Trans Med Imaging Feb 2019;38(2):550–560. https://doi.org/10.1109/TMI.2018.2867350.
20. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. arXiv Sep. 11, 2020. https://doi.org/10.48550/arXiv.1905.11946.

21. Buitinck L, Louppe G, Blondel M, Pedregosa F, Muller AC, Grisel O. API design for machine learning software: experiences from the scikit-learn project. arXiv Sep. 01, 2013. https://doi.org/10.48550/arXiv.1309.0238.

22. Sener O, Savarese S. Active learning for convolutional neural networks: a core-set approach. arXiv Jun. 01, 2018. https://doi.org/10.48550/arXiv.1708.00489.

23. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv Dec. 10, 2015. https://doi.org/10.48550/arXiv.1512.03385.

24. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv Jan. 29, 2017. https://doi.org/10.48550/arXiv.1412.6980.

25. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD. QuPath: open source software for digital pathology image analysis. Sci Rep Dec. 2017;7(1), 1. https://doi.org/10.1038/s41598-017-17204-5.

26. Vahadane A, Peng T, Sethi A, Albarqouni S, Wang L, Baust M, et al. Structure-preserving color normalization and sparse stain separation for histological images. IEEE Trans Med Imaging Aug 2016;35(8):1962–1971. https://doi.org/10.1109/TMI.2016.2529665.

27. Patil A, Talha M, Bhatia A, Kurian NC, Mangale S. Patel S.,et al. Fast, self supervised, fully convolutional color normalization of H&E stained images. arXiv Nov. 30, 2020;30. https://doi.org/10.48550/arXiv.2011.15000.