Automated Quality Control Analysis of Histopathology Images

MERCK

Project Coordinators: Lin Li, Rajath Soans

Berkeley UNIVERSITY OF CALIFORNIA

Students: Antoine Lavoie, Bruce Xu, Ethan Yoo, Julie Chen

Abstract

Histopathology images are widely used in clinical and research settings, but the presence of artifacts can significantly impact their quality and subsequent analysis. In this project, we aim to build an automated quality control (QC) tool for accurate detection, localization, and quantification of artifacts in whole slide images (WSIs). Our strategy is to leverage existing histopathology image QC tools to generate annotated dataset which can be used to train a deep learning model that can identify artifacts in WSIs. The project comprises three main phases: (1) Detect and segment artifacts on WSIs, (2) Use case study to evaluate the impact of the QC tool on subsequent analysis, such as deep learning models, (3) Design comprehensive QC metrics and scoring mechanism. Our goal is to improve the usability of WSI datasets by automating the artifact detection process and eliminating the need for manual curation in digital pathology workflows.

Motivation

Motivation for automatic quality control

- Histopathology is a time-consuming subject that requires technicians to go over each image by eye to assess its usability
- Algorithms yield higher efficiency
- QC of histopathology images is time-consuming.
- Automatic QC tool can help accelerate the process and improve the QC performance
- Algorithms yield higher accuracy

Motivation for using HistoQC

- HistoQC is an open-source quality control tool for digital pathology slides
- Has been used in various research/industrial practices
- Able to detect bubbles, blur regions, pen-markings

Motivation of our project

- HistoQC has room for accuracy improvements and functionality improvements

Optimization

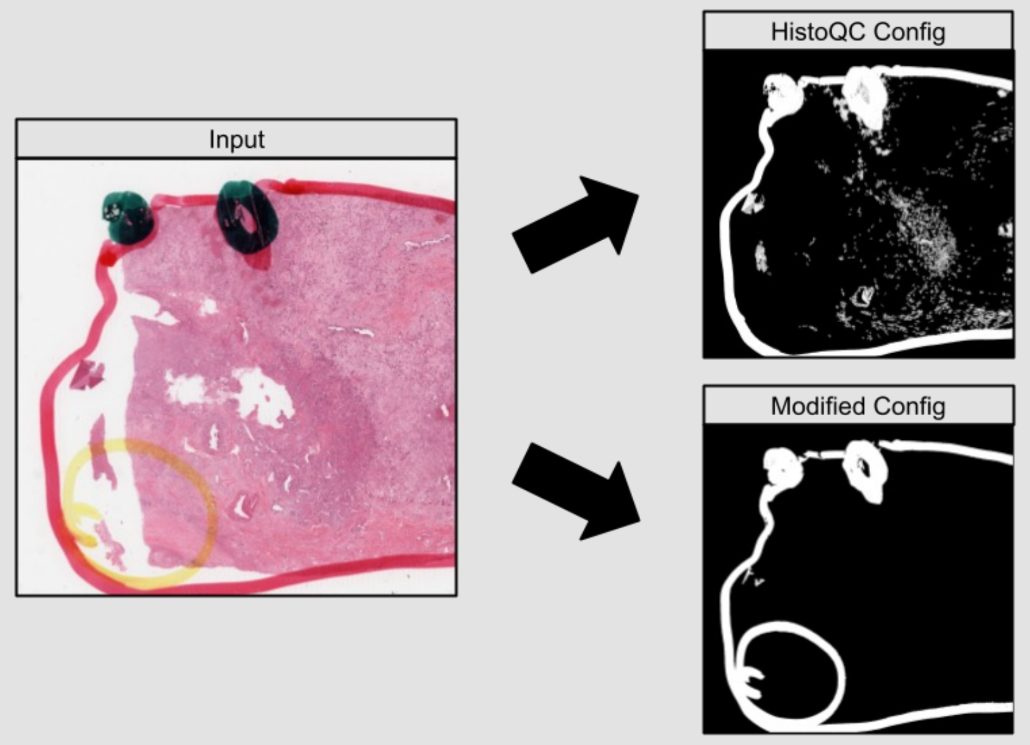
Steps:

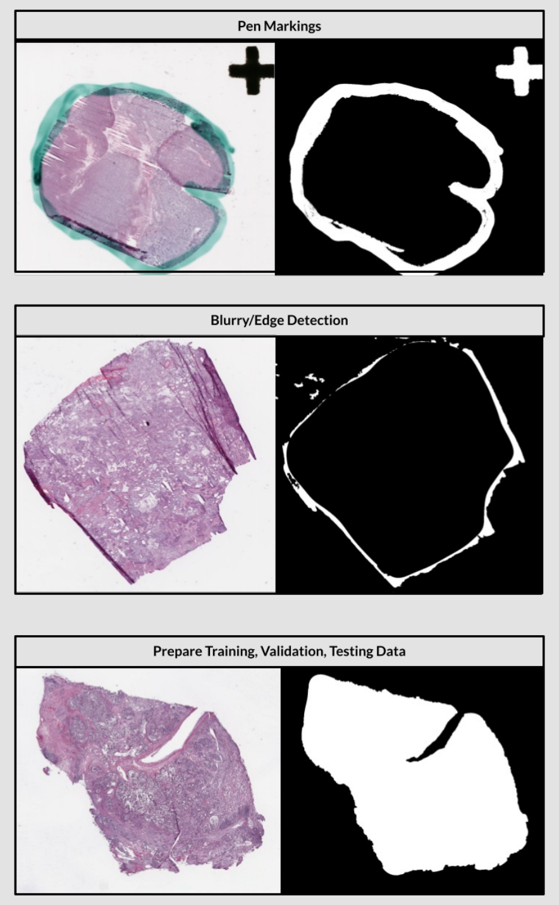
- Implemented OpenCV to create our own dataset using color thresholding to train the random forest classifier.
- Detected features were used to train a Random Forest classifier with <N> trees.
  - Random forest classifier is a machine learning algorithm that uses an ensemble of decision trees to classify input data. It randomly selects a subset of features and creates multiple decision trees, which are then combined to produce a final prediction.

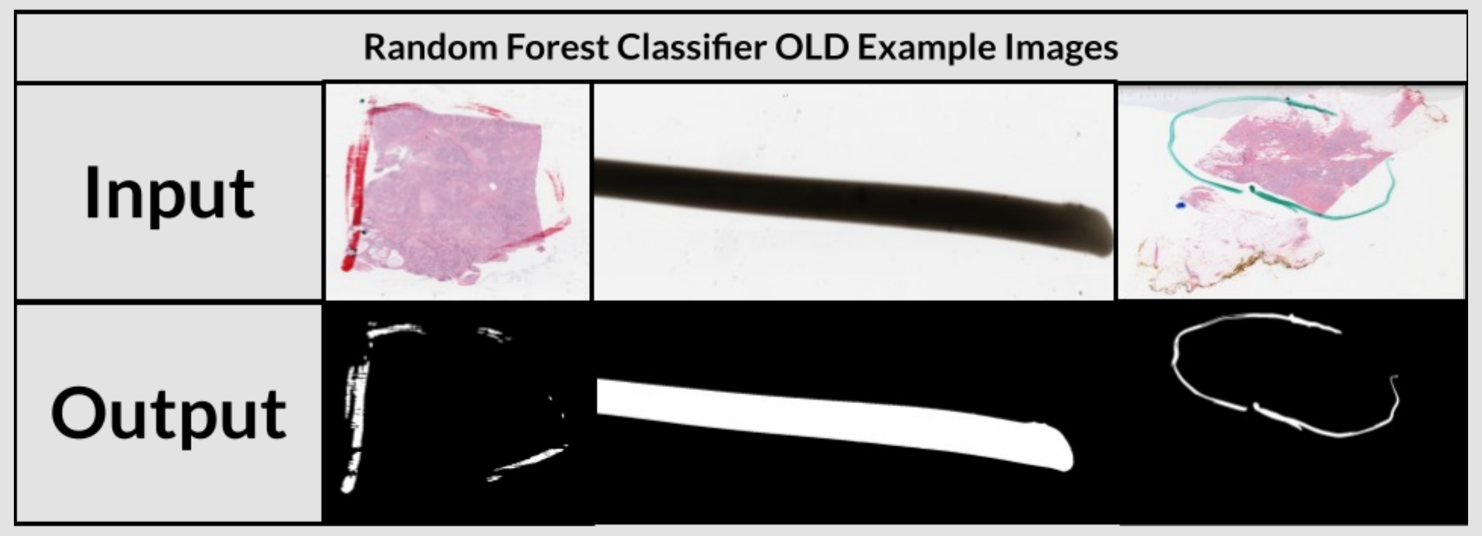
Optimized Features:

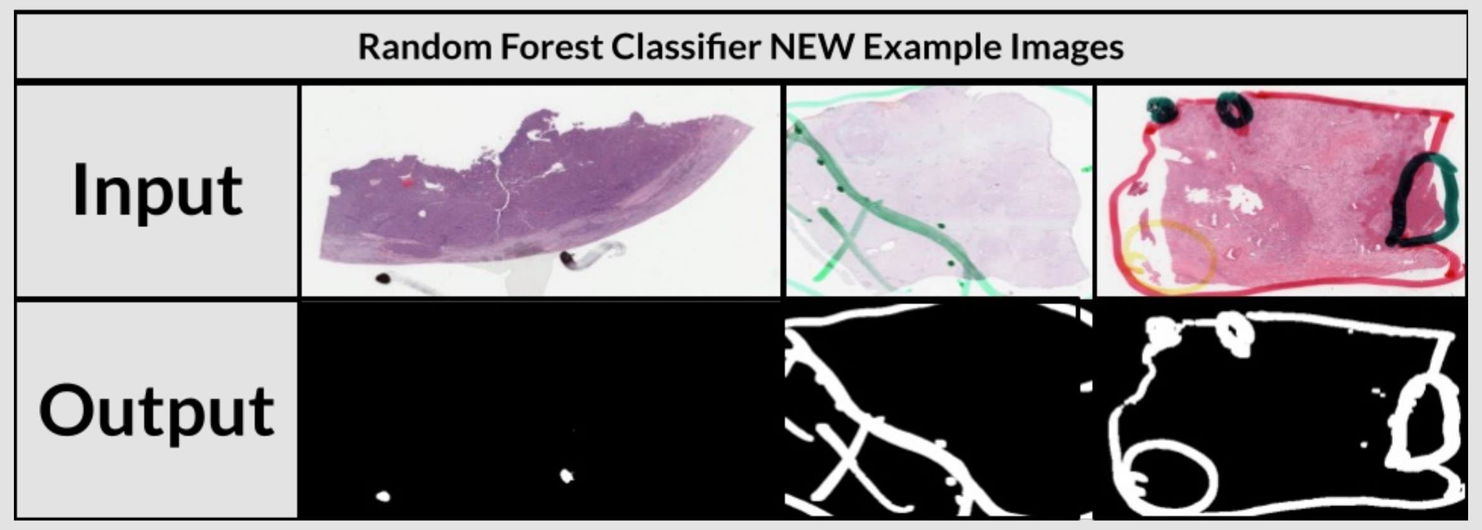
- Frangi Beta 1: 0.5 → 0.3 (weight between lines and blobs)
- Frangi Beta 2: 15 → 12 (weight between lines and background)
- Laplace: 3 → 5 (increase area the filter is applied for edge detection)
- Local Binary Pattern (LBP) Radius: 0 → 5 (larger circular area, more robust to noise and texture variation by comparing intensity to neighboring pixels)
- Median Disk Size: 0 → 5 (nonlinear filter that replaces each pixel with median of surrounding pixels)
- Gabor Sigma: 0 → (1, 3) (captures more texture using sinusoidal waves modulated by Gaussian, usually used for face recognition)
- Gaussian Sigma: 0 → 3 (weighs pixels based on the Gaussian curve, where sigma determines the size of the kernel and amount of smoothing)
- Gaussian Multichan: False → True (applied the filter to each color channel instead of just a singular channel)
- Area Threshold: 100 → 1500 (increases area threshold needed to be detected, takes out smaller noise)

Example of module improvement (Pen Marking module):
Cannot detect green pen markings —> Can detect pen markings now









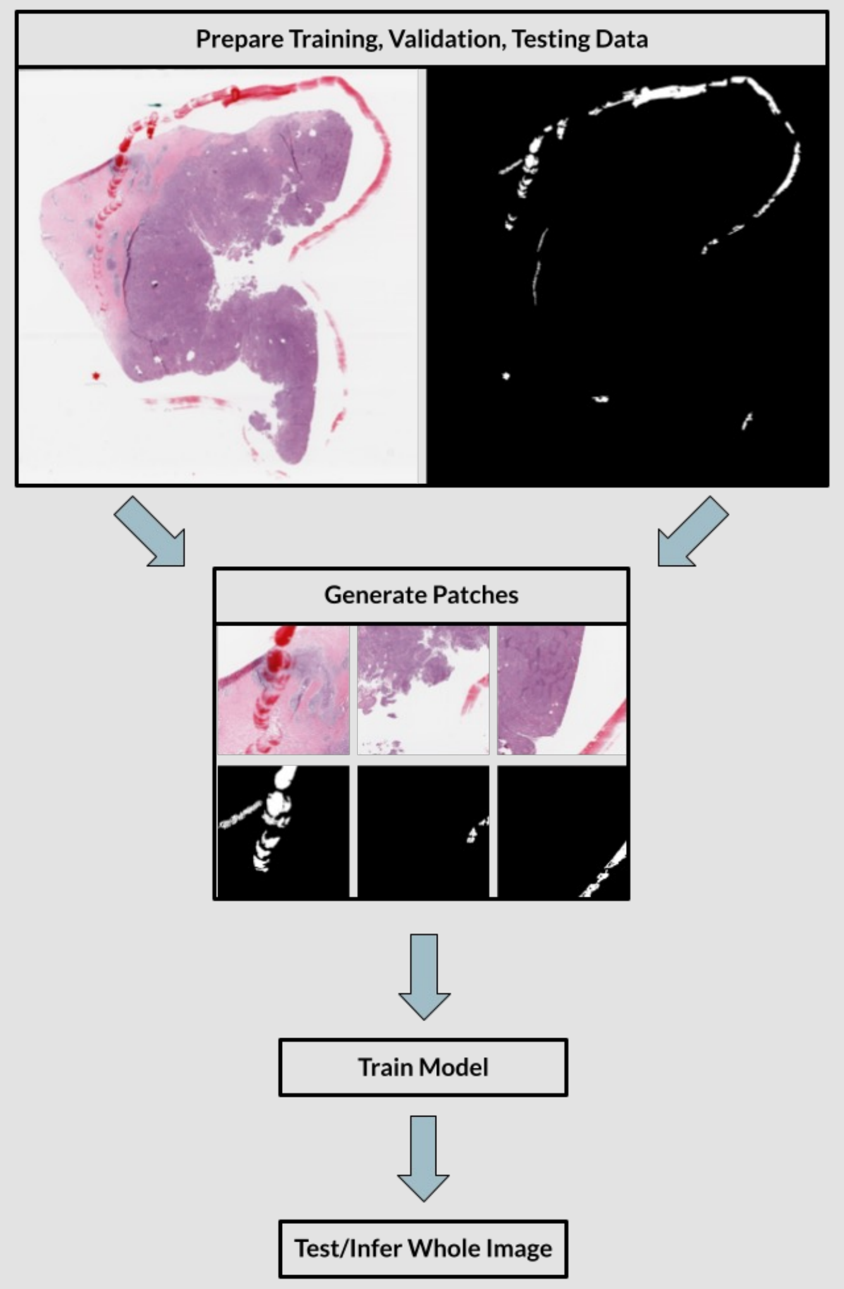
Deep Learning

Why

- Pathology images are complex and large, requiring time and resources to analyze each and every slide
- Allows for hierarchical feature representation that can outperform current image analysis method

How

- Create the dataset through HistoQC, resulting in the ideal input and output.
- Determine the hyperparameters and optimize them (learning rate, batch size, layers and neurons, regularization, etc.)
- Parse and section the dataset to feed through the deep learning model for tile extraction at a higher resolution
- Return a slide level decision heatmap and quality control score



Challenges

Challenge 1: High-Performance Computing

- Spent a large chunk of time trying to sort out Savio in order to train models that require multiple gpu
- Ultimately unsuccessful in overcoming this obstacle but better prepared for future terms

Challenge 2: Limited Flexibility with HistoQC

- Limited to 3 examples for optimizing HistoQC outputs

Challenge 3: Data Collection

- The GDC website is the place to download training data, but the interface is hard to work with
- Not having the feature to preview images, need for manual filtering, which is time-consuming

Challenge 4: Complex Algorithm

- The structure and parameters of HistoQC were difficult to interpret

Next Steps

- Complete of the current implementation of the deep learning model we are working with
- Explore academic papers to optimize the model and choices we make in the process
- Use real-world dataset from Merck Pathologist

Conclusion

- The team is able to improve the accuracy of the original algorithm, especially in pen-marking detections, building on the existing image QC tools and leveraging AI/ML-based approaches.
- For future directions, deep learning models can be implemented to further improve the accuracy of the algorithm. The pen-marking detection algorithm can also be modified to serve a larger variety of purposes, such as tissue fold, blur region, and bubble detection.
- Automated quality control for histopathology images is an important yet emerging field. We anticipate that there will be an influx of novel and innovative research and development initiatives in the foreseeable future.