A novel approach to classification and segmentation of colon cancer imaging towards personalized medicine


*Persistent Systems Limited

Abstract—Recent advances in the field of pathology coupled with the rapid evolution of machine learning based techniques have revolutionized healthcare practices. Colorectal cancer accounts for one of the top 5 cancers with high incidence (126,240 in 2020) with a high mortality worldwide [1] [2]. Tissue biopsy remains to be the gold standard procedure for accurate diagnosis, treatment planning and prognosis prediction [3]. As an image based modality, pathology has attracted a lot of attention for development of AI algorithms and there has been a steady increase in the number of filings for FDA authorized use of AI algorithms in clinical practice [4]. The SemiCOL Challenge aims to develop computational pathology methods for automatic segmentation and classification of tumor and other tissue classes using H&E stained images. In this paper, we present a novel machine learning framework addressing the SemiCOL Challenge, focusing on semantic segmentation, segmentation-based whole-slide image classification, and effective use of limited annotated data. Our approach leverages deep learning techniques and incorporates data augmentation to improve the accuracy and efficiency of tumor tissue detection and classification in CRC. The proposed method achieves an average Dice score of 0.2785 for segmentation and an AUC score of 0.71 for classification across 20 whole-slide images. This framework has the potential to revolutionize the field of computational pathology, contributing to more efficient and accurate diagnostic tools for colorectal cancer.

Index Terms—Colorectal cancer, Computational pathology, Semantic segmentation, Whole-slide image classification, Deep learning, U-Net, EfficientNet, Binary Cross-Entropy Loss, Annotated data

I. INTRODUCTION

Artificial Intelligence has transformed the field of imaging based modalities at an unprecedented speed with more than 80 FDA filings for diagnostic use (Software as a Medical Device – SaMD) in 2021 [4]. Adoption of AI based algorithms for routine pathology analysis such as biomarker quantitation and tumor grading (Prostate cancer) has led to promising development of AI applications for complex tasks for discoveries, diagnosis, prediction and prognosis of cancer. Colorectal cancer (CRC) is one of the most prevalent malignant epithelial tumors worldwide, and early and accurate diagnosis of CRC is essential for effective treatment planning and improved patient outcomes. Traditional methods of pathology are time consuming and could lead to significant intra and interobserver variability. In this study, we explore the application of machine learning to develop custom models for slide level classification as well as multi class segmentation of colon cancer images. Our methods involve building models from the scratch in lieu of pretrained models. Although, this process is computationally intensive, it allowed us to tailor these models specifically towards the task and enabled the models to learn, adapt to the unique features that were present in the different pathology images. By automating and augmenting the current workflows using AI, we can improve the speed, accuracy and consistency of the reports thereby reducing the pathologists burden. Furthermore, the use of multi-class segmentation allows the application of these models to diverse cancer types (ex: Tumor stroma, necrosis, lymph are all present across multiple cancer types) thereby broadening the clinical utility. In this paper, we describe our approach in developing, training and testing these models highlighting the transformative potential of machine learning in computational pathology and personalized medicine.

II. METHODS AND TECHNIQUES

A. Data Preparation

We found that 50%-80% of pixels provided in the manually annotated data set were unclassified. A further 3% to 18% of pixels belonged to the background class. To ensure a good balance of tissue classes and background classes careful selection of the training images would be required.

Table 1 shows the pixel distribution among the various classes in the annotated images provided in 3 example cases. Here class 0 denotes unannotated pixels and class 10 denotes background pixels

Each 3000x3000 image was split into 8x8 tiles and resized to 256x256 pixels. Tiles that had more than 70% of pixels belonging to the background or unannotated classes were removed.

The train and test splits were create by dividing the images on the basis of case.

B. Training

Early attempts at training a single model to segment all 10 classes yielded models that would have high dice scores in a few classes and very poor dice scores in others. Improving the models performance on a particular class by increasing
Finally, the WSI was categorized as belonging to the tumor class if the total area encompassed by the tumor within the mask exceeded 10%.

### III. RESULTS

Table II below shows the average of the dice scores calculated on the segmentation task for 6 images from given challenge validation data.

<table>
<thead>
<tr>
<th>Class</th>
<th>Dice Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5605</td>
</tr>
<tr>
<td>2</td>
<td>0.2785</td>
</tr>
<tr>
<td>3</td>
<td>0.0896</td>
</tr>
<tr>
<td>4</td>
<td>0.5987</td>
</tr>
<tr>
<td>5</td>
<td>0.0805</td>
</tr>
<tr>
<td>6</td>
<td>0.5222</td>
</tr>
<tr>
<td>7</td>
<td>0.0411</td>
</tr>
<tr>
<td>8</td>
<td>0.0519</td>
</tr>
<tr>
<td>9</td>
<td>0.3404</td>
</tr>
</tbody>
</table>

**TABLE II: Class wise dice score**

An AUC score of 0.71 was achieved across 20 Whole Slide Images.

### IV. CONCLUSION

In conclusion, our study demonstrates the efficacy of denovo machine learning models at slide level classification and multi class segmentation of colon cancer images. Our approach of not using the pretrained weights, allowed us to develop new models that adapt to new histological features thereby possibly enhancing the diagnostic accuracy and reduce the inter observer variability. As we continue to refine our models to improve the accuracy, we believe our novel approach contributes to the field of computational pathology where one can envision a future for successful adoption of AI algorithms in routine clinical practice to revolutionize diagnosis, treatment and patient care.

### REFERENCES