Improving Histological Image Analysis for Human Colorectal Cancer Detection

Team BIC: Boyang Wu, Jason Nguyen, Sahinde Dogruer, Shashwata Moitra

https://github.com/qln-cmu/CRC_NCT_classification

Overview

This project confronts the global challenge of colorectal cancer (CRC), a leading cause of cancer-related deaths. We aim to elevate CRC detection from histological images, capitalizing on the potential of machine learning with the "NCT-CRC-HE-100K" dataset. The end goal is not just to benchmark but also to deploy a user-centric web application for practical use. The dataset comprised 100,000 distinct non-overlapping image patches from H&E stained histological images, providing a rich source for analysis.

Objective

- Implement EfficientNet models (b0-b7) across different resolutions (224x224, 112x112, and 56x56 pixels) to find the most effective setup for CRC detection.
- Optimize image resolution to maintain a balance between accuracy and storage efficiency.
- Develop a web application that allows users to upload their histological images for instant classification results.

Stakeholders: Medical professionals: pathologists and oncologists; researchers; patients; healthcare institutions: hospitals and diagnostic labs.

KPIs: validation accuracy, sensitivity/recall, precision, F1-Score.

Model Evaluation Strategies

In our extensive evaluation of **360** model checkpoints generated from **24** combinations of different models' parameters, we implemented EfficientNet models b0-b7 across three resolutions for **15** epochs. This approach allowed us to understand the trade-offs between resolution, accuracy, and efficiency. Our focus was on three crucial aspects:

- Finding the best overall model with the highest validation accuracy,
- Enhancing the recall for the cancer-associated stroma (STR) class, and
- Ensuring accurate identification of the colorectal adenocarcinoma epithelium (TUM) class without increasing the false positive rate.

We incorporated F-1 and F-beta scores (beta = 1.2) for a precise assessment of these two classes.

Results and Improvements

A pivotal outcome of our project was the efficiency gained from resizing images. By resizing the dimensions of the input image from 224x224 pixels to 112x112 and even 56x56 pixels, we achieved a marginal improvement (**0.62%**) in validation accuracy (from 95.75% to **96.38%**) while significantly cutting down on training time (**35%** for three models) and storage demands (**75%** in data space and **10%** in model space). These findings were critical in selecting the top three models for our application.

Web Application Development

The culmination of our research is a streamlit web application that utilizes the three best models. This platform enables patients to upload their histological images and receive immediate classification results with emphasis on Tumor and Stroma identification, empowering them with rapid insights into their potential CRC status. This tool bridges the gap between advanced CRC detection technology and patient accessibility, marking a significant stride in cancer diagnostics.

Conclusion & Future Directions

Our project marks a substantial leap in histological image analysis for CRC detection, blending advanced machine learning models with practical applications. Future endeavors will focus on refining the models' efficiency and expanding the web application's capabilities to encompass a broader range of diagnostic requirements.