Project Proposal: Lung and Colon Cancer Histopathological Image Classification

Problem Statement

The medical community is seeking an automated approach to accurately diagnose adenocarcinoma and squamous cell carcinoma in lung tissue and adenocarcinoma in colon tissue using histopathological images. An effective solution will be able to supplement the opinions of medical professionals and could also provide a cancer screening option for individuals/communities with limited access to healthcare.

Context

Currently, the gold standard approach for diagnosing cancer, and identifying cancer/tumour type, is analysis of histopathological images by a pathologist. Effective automated classification of histopathological images could help to reduce pathologist workloads, provide a “second opinion” to supplement a pathologist’s conclusion, and serve as a cancer screening option for individuals/communities with limited access to healthcare. Therefore, a model will be built in order to classify lung and colon cancer from histopathological slides with high accuracy. The model will be trained on histopathological images with five image classes: benign lung tissue, lung adenocarcinoma, lung squamous cell carcinoma, benign colon tissue, and colon adenocarcinoma. An effective model has the potential to reduce the workload on pathologists, supplement the opinions of pathologists, and provide cancer screening in situations where access to healthcare is limited. Therefore, benefits could be seen both within the medical community and on patient outcomes.

References:
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1685-x
https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images

Criteria for Success

A model will be developed that can classify lung and colon histopathological images into the following groups with a high degree of accuracy: benign lung tissue, lung adenocarcinoma, lung squamous cell carcinoma, benign colon tissue, and colon adenocarcinoma.

Scope of Solution Space

A dataset containing histopathological images with the following five image classes will be used to train the model: benign lung tissue, lung adenocarcinoma, lung squamous cell carcinoma, benign colon tissue, and colon adenocarcinoma. Both models built from scratch and models built using transfer learning techniques will be trained and tested.

References:
https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images

Constraints Within Solution Space

- The model will be limited to image classification. Image segmentation will not be performed. Therefore, the location of tumours in the images will not be identified.
- In addition to normal lung and colon tissue, the dataset contains only two types of lung cancer (squamous cell carcinoma, adenocarcinoma) and one type of colon cancer (adenocarcinoma). Therefore, the model's classification will be limited to those tissue and cancer types.
- The size of the original dataset was enhanced from 750 images to 25000 images using data augmentation techniques. Therefore, it is possible that the model will not have been trained on enough original data to generalize well to data outside of the dataset.

Stakeholders

- Medical Community (especially Pathologists)
- Patients who have lung or colon cancer, are at risk for lung or colon cancer, or are being screened for lung or colon cancer

Key Data Sources

- Kaggle - Lung and Colon Cancer Histopathological Images (https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images)
    - Original data source: https://arxiv.org/abs/1912.12142v1