Check-In #2 Reflec

Introduction

Around 2% to 5% of cancer cases are first diagnosed as cancer of unknown primary origin (CUP), a scenario where no obvious primary tumor is found at initial presentation. As many modern therapeutics are tailored to tumor origin, the prognosis for patients whose cancers' origins remain unknown is poor (median overall survival of 2.7-16 months). Current diagnostic work-ups for these patients are elaborate and expensive, involving several rounds of pathology, radiology, and laboratory examinations. We aim to explore and build off of a new diagnostic approach that exploits deep learning to predict tumor features from histopathological whole slide images (WSIs), with a goal of improving patient outcomes and reducing health system strain.

We base our implementation on Lu et al., 2021, who developed a multi-class classification algorithm termed TOAD (Tumour Origin Assessment via Deep Learning) that determines tumor origin and metastatic status from haematoxylin and eosin stained WSIs. We were drawn to this paper as we are broadly interested in applications of deep learning to the analysis of biological image data, and this study's use-case stood out as a promising example in the field. Hopefully our work will let us fully explore a growing world of models that will be critical to the practice of medicine in the future.

Challenges:

Our primary challenge is scaling down the original paper to fit our available resources. For example, the WSI dataset in Lu et al., 2021 is 28 terabytes, which is far more than OSCAR can support. Thus we must select a far smaller dataset and aim to predict fewer origins than the original paper. This being said, the goals of our project have shifted slightly: rather than trying to achieve the same/higher level of accuracy as the original paper by modifying our feature extraction strategy, we are more focused on comparing how well a lightweight model can perform. Specifically, we want to know if pairing EfficientNet feature extraction with JPEG images (as opposed to ResNet feature extraction with TIFF files) produces any significant results.

Insights:

There are no concrete results to display at this time, as the model is not yet functional.

Plan: Are you on track with your project?

Honestly not really. We will definitely have to dedicate a lot of time in the next few days to whip this into shape. Moving forward, we need to dedicate most of our time to the data preprocessing step. We have mostly converted the model itself from PyTorch to Tensorflow, and an initial dataset to test the skeleton of our model. The data preprocessing (standardizing the magnifications, properly storing metadata, and patching the raw images) is the main step left and probably the hardest overall. Once that is complete, the training and experimentation components will be the large hurdles left.