Question 1: Report on Paper "Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study"

1. Overview

This paper reports on the development of a deep multiple instance learning (MIL) model that is able to predict tumor purity from H&E stained histopathology slides. In this report, the cancer domain significance and the machine learning methods covered will be further explored.

2. Cancer Domain Application

The estimation of tumor purity in histopathology slides samples is important in cancer domain application. In genomic analysis, where there is potential for development of personalized medicine, being able to accurately estimate the tumor purity is crucial and can affect the analysis quality.

Tumor content has also been found to be associated with various clinical variables. In patient prognosis, low tumor purity is associated with poor prognosis in certain types of cancer. Tumor purity has also been found to be a potential indicator of therapeutic responses in some cancer types.

The two methods of estimating tumor purity are the percent tumor nuclei estimation and genomic tumor purity inference. Percent tumor nuclei estimation can be time consuming and can be affected by inter-observer variability as a pathologist would be required to count the tumor nuclei. While genomic tumor purity inference is the current gold standard in tumor purity estimation as it infers from genomic data, it lacks the spatial information on tumor microenvironment. The MIL model developed in this paper is able to mitigate these issues.

3. Machine Learning Methods

For the prediction of tumor purity, two types of machine learning models are typically used. They are namely the patch-based models and the multiple instance learning (MIL) models.

Patched-based models learn from labeled patches cropped from sample slides. The predictions of all patches in a sample slide are used to aggregate the tumor purity for that slide. This method requires pixel-level annotation labels by pathologists and are not readily available due to the tedious annotation process.

In MIL models, each sample is represented as a bag of patches that are cropped from that sample's slide. The bags are labelled according to the tumor purity of their corresponding sample slides. Due to the method of labelling, they provide weaker aggregate information instead of pixel-level information.

In the paper, the MIL model was used due to pixel-level annotations being rare and expensive to obtain. Besides tumor purity, the MIL model was also used to classify between cancer and normal slides through the MIL model output (with tumor purity > 0.5 being classified as a cancer slide).

The MIL model was also used to generate spatial tumor purity maps by getting tumor purity for bags of patches within regions of interest (ROIs) over the entire sample. The tumor purity is then allocated to the centre of each of these ROIs.

The paper also demonstrated the ability of the model to obtain pixel level segmentation maps of cancerous and normal pixels. This was obtained through hierarchical clustering of discriminant features learned by the MIL model.

The model was trained using data from 10 TGCA cohorts and transfer learning was used to train on the Singapore cohort.

4. Conclusion

The MIL model developed in this paper was able to predict the tumor purity of H&E stained histopathology slides in both the TCGA cohorts and the Singapore cohort with close accuracy to the genomic tumor purity values, and was able to classify slides samples into cancer and normal with an AUC very close to 1.

The MIL model was also able to obtain spatial tumor purity maps and cancerous vs normal segmentation maps. These were found to be consistent with histopathology during qualitative assessment of normal tissue structures and regions invaded by cancerous cells.

These results show that the MIL model can be used to assist pathologists in their estimation of tumor purity and can reduce their workload and increase precision by reducing inter-observer variability.