Report for Q1

Introduction

Cancer is a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body[1]. The savage cells are called cancer cells and are the main object in cancer therapy. Hence, tumor purity, which marks the proportion of cancer cells in complex tumor tissue, is a very important clinical indicator tightly affecting therapeutical effect. However, nowadays most tumor purity calculation still relies on the tedious estimation of pathologists or genomic tumor purity inference which cannot be applied to low tumor content samples. Thus it is necessary to optimize these limited methods. This report focuses on the paper *Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study*. The paper presents a new model using deep multiple instance leaning principles to predict tumor purity from H&E stained digital histopathology slides. This novel model is creative and practical because it not only overcomes the disadvantages of existing methods but also reveals how spatial factors influence tumor purity predictions. This model also have functions to do segmentations and classifications. At the same time, the comprehensive experiments listed made the paper much persuaded.

Summary

In this part, I will summarize the trunk of this paper, including logic of the model, the data utilized and how experiments are organized.

1. Model

   The Multiple Instances Learning model is the most important part to predict sample level tumor purity. And the process is : First, patient level top and bottom slides are inputted to be sampled and organized as a bag of patches. Then, the patches will be extracted to be feature vectors (ResNet18) . Third, those vectors will be filtered to be bag-level representations and finally used by the representation transformation module (3-layer perceptron) to calculate the tumor purity.

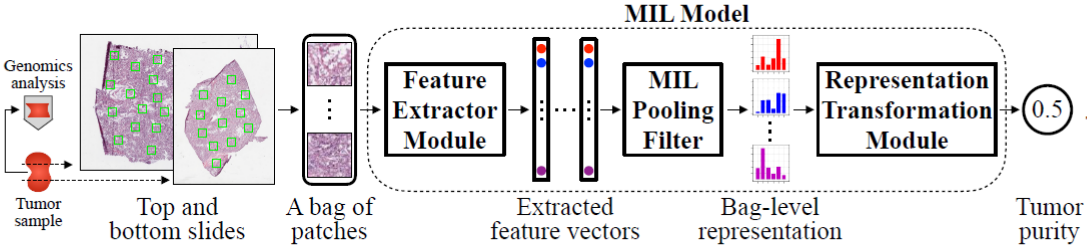


Figure 1: The structure of the MIL model[2]

   The model uses golden standard genomic sequences as labels. The most creative module of MIL Model is the Pooling Filter, which is based on a distribution formula below and proved to be superior to the point estimate-based counterparts.

$$\tilde{p}^j_X(v) = \sum_{i=1}^{N} \beta_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left(v-\alpha_i f^j_{x_i}\right)^2} \quad \forall_{j=1,2,\cdots,J}$$

   The utilization of the MIL Model is not only restricted by tumor purity prediction. The paper presents three extended usages of MIL Model that strengthen the novelty of this paper. By ROI (Region of Interest) analysis, the variation of tumor purity over a slide and the influence of spatial elements on tumor purity predictions can be obtained. And the feature extracted module of MIL model can extract discriminant features of sample level slide, which are used by hierarchical clustering module to obtain normal vs cancerous segmentation maps. What's more, depending on the tumor purity value obtained from MIL Model (threshold = 0.5), the input slides can be classified into normal samples and tumor samples.

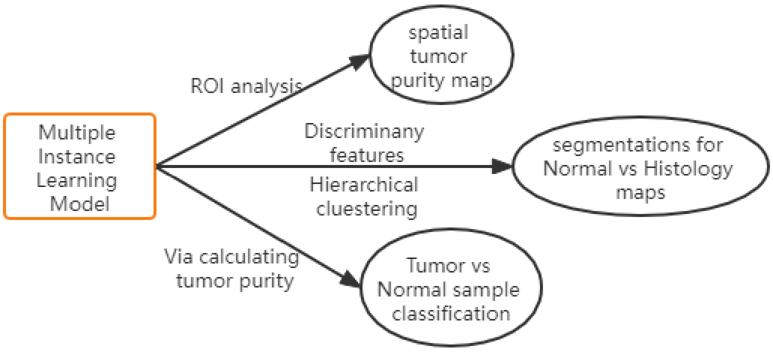
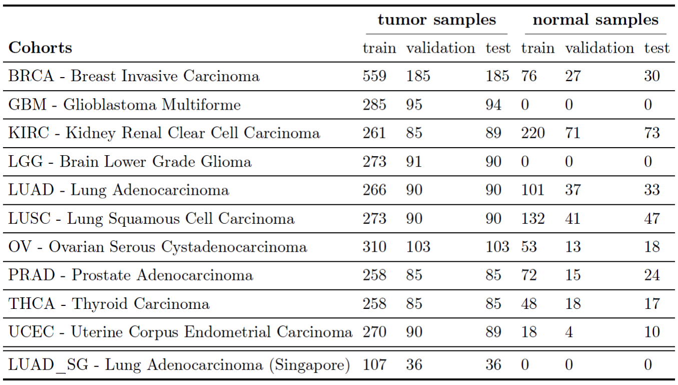

Multiple Instance Learning Model
ROI analysis
spatial tumor purity map
Discriminany features
Hierarchical cluestering
segmentations for Normal vs Histology maps
Via calculating tumor purity
Tumor vs Normal sample classification

Figure 2: Further usages of the MIL model

2. Data

   H&E stained digital histopathology slides, which are the most widely used stain in medical diagnosis, are used here to predict tumor purity. And the data sets, as being listed below, are selected from the TCGA and Singapore cohorts with more than 400 patients each. Thus standardization and reliability of data sources enhance persuasion of the paper.



	tumor samples			normal samples		
Cohorts	train	validation	test	train	validation	test
BRCA - Breast Invasive Carcinoma	559	185	185	76	27	30
GBM - Glioblastoma Multiforme	285	95	94	0	0	0
KIRC - Kidney Renal Clear Cell Carcinoma	261	85	89	220	71	73
LGG - Brain Lower Grade Glioma	273	91	90	0	0	0
LUAD - Lung Adenocarcinoma	266	90	90	101	37	33
LUSC - Lung Squamous Cell Carcinoma	273	90	90	132	41	47
OV - Ovarian Serous Cystadenocarcinoma	310	103	103	53	13	18
PRAD - Prostate Adenocarcinoma	258	85	85	72	15	24
THCA - Thyroid Carcinoma	258	85	85	48	18	17
UCEC - Uterine Corpus Endometrial Carcinoma	270	90	89	18	4	10
LUAD_SG - Lung Adenocarcinoma (Singapore)	107	36	36	0	0	0

3. Experiments

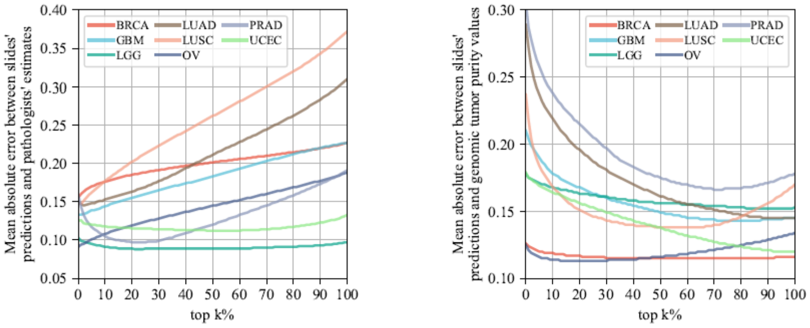
   According to the paper, 100 bags are extracted from each sample and the completed training and testing process make the model performance convinced. The accuracy of the MIL Model is compared with pathologist' estimation to represent the superiority of the MIL Model. The analysis of comparison results is written in the next part.

Results Analysis

This paper presents very detailed analysis on results, especially in statistics. The results analysis mainly focus on five parts:

1. Accuracy of tumor purity estimation (comparison)

   Taken golden standard genomic tumor purity as the evaluation reference, the predictions on the MIL model shows a better accuracy.



2. Spatial variation and spatial influence relevant to tumor purity

   From training and testing top and bottom slides separately and analyzing of p-values, we could find clear difference between MIL predictions of the same slides. And the absolute error analysis shows that using both top and bottom slides for predictions would performance better than just using single slide.

3. Causes of the bias of pathologist estimation

   From mean-absolute-error analysis, we could find that pathologists are easily to select high tumor content regions for estimation.

4. Learning discriminant features

   For patients having mixed kinds of cells, the features of of normal and tumor slides could be extracted and then the segmentation map for each kind of cells would be made.

5. Classification of normal vs tumor features

   Classification could be realized according to the tumor purity value of different tissues.

Response

I think the novel MIL model is creative and practical for three reasons.

1. Make up for the large time cost and inaccuracy of pathologist estimation. And it could applied to low tumor content samples.
2. Detailed statistics analysis make the results reliable.
3. Solving the problem of tumor purity estimation is meaningful to cancer treatment.

However, the MIL model still has disadvantages, such as the limitation of data size since insufficient data sets may lead to over-fitting and lacking robustness. What's more, I think it would be more convictive if there are comparison experiments between other ML models and the MIL model.

Reference

[1] :https://www.cancer.gov/about-cancer/understanding/what-is-cancer

[2] : Oner M U, Chen J, Revkov E, et al. Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study[J]. Patterns, 2021: 100399.