Literature Review on
Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning in A Pan-cancer Study

Yufei WU

Jan. 11, 2022

Contents

1 Introduction

Tumor purity is the proportion of cancer cells in the tumor issue. In this paper, a deep multiple instance learning model is developed to accurately estimate the tumor purity. The advantage of the model is that it is cost-effective compared with estimating by pathologists, and it is more objective. The model may also provide information about the spatial organization of the tumor microenvironment which is crucial to the tumor formation and therapeutic response.

2 Multiple Instance Learning Model

Overview of the MIL Model

Figure 2.1 shows the process of the multiple instance learning model to predict the tumor purity.

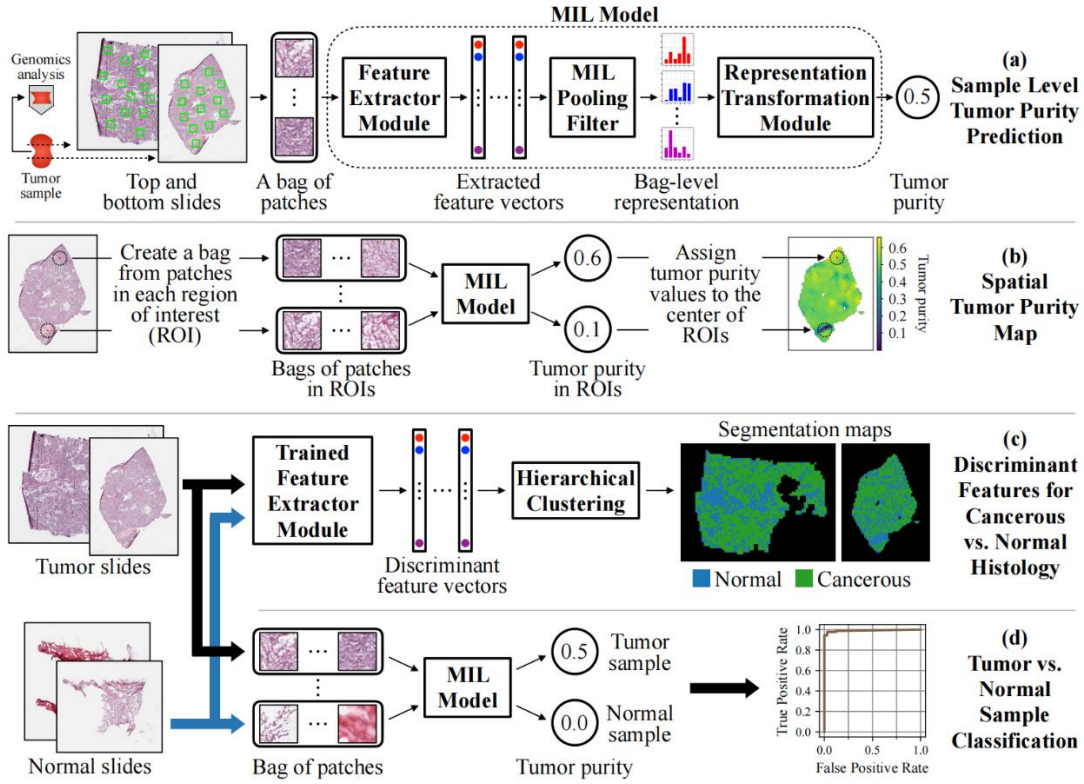

Figure 2.1: A novel MIL model (a) The model accepts a bag of patches cropped form the top and bottom slides of a sample as input. The feature extractor module extracts a

feature vector for each patch inside the bag. (b) Obtaining a spatial tumor purity map. (c) The MIL model learned discriminant features for cancerous vs. normal histology from samples and trained feature extractor module extracts feature of patches from tumor and normal slides of a patient. (d) The MIL model successfully classifies samples into tumor vs. normal.

Three-Stage Process

The multiple instance learning model performs on bag-level data. The objective of the model is to predict a bag label Y for a given bag of instance X. To obtain the predicted bag label $\hat{Y}$, the model in the paper contains three stages.

The first stage is to extract the features of the input x_i. A feature extractor module $\theta_{feature}$ transforms the instance space to the feature space. ($\theta_{feature}: I \rightarrow F$). For each input x_i, we extract J features and outputs a feature vector. Then for n inputs, we will get a feature matrix with size $J \times N$.

The second stage is a MIL pooling filter module $\theta_{filter}: F \rightarrow H,$ which transforms to the bag-level representation space. In this stage, the model takes the feature matrix and aggregates the extracted feature vectors into a bag-level representation.

The third stage of the MIL model is a bag-level representation transformation module $\theta_{transformation}: H \rightarrow y$. The bag-level representation is transformed into the predicted bag label $\hat{Y}$.

The first and third stage $\theta_{feature}$ and $\theta_{transformation}$ will be implemented by neural networks. For θ_{filter}, a novel 'distribution' pooling filter will be used. This system of neural networks is end-to-end trainable/

Distribution Pooling Filter

One of the special designs of the MIL model in the paper is a novel 'distribution' pooling filter. It produces stronger bag-level representations from patches' features than standard pooling filters like max and mean pooling. The 'distribution' pooling summarizes extracted features into a bag-level representation by estimating marginal feature distributions. Let $\tilde{p}_X^j$ denote the estimated marginal distribution obtained over

j^{th} extracted feature, it can be calculated by using kernel density estimation, which employs a Gaussian kernel with standard deviation σ. The equation is given bellowed.

$$\tilde{p}_X^j(v) = \sum_{i=1}^{N} \beta_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(v-\alpha_i f_{x_i}^j)^2} \quad \forall j = 1, 2, \dots J$$

We also proved that this kind of pooling filters are superior to the point estimate-based pooling filters regarding the amount of information captured.

MIL Model Training

The data of the patients are randomly divided into training set, validation set and test set. The input is a bag of patches from slides of the tumor sample, and the ground-truth label is the tumor purity value obtained from genomic sequencing data by ABSOLUTE. A tumor purity value of 0.0 is assigned to a normal sample as the ground-truth label. The loss function is the absolute error. Based on the loss in the validation set, the model employed early-stopping in case of overfitting.

3 Applications on MNIST Datasets

In this part, we will try to implement the method on MNIST datasets.

Datasets

The MNIST data set is a widely-used data set in the area of machine learning, which contains tens of thousands of scanned images of handwritten digits, together with their correct classifications. The data set was split into 50,000 training images, 10,000 validation images and 10,000 test images.

Figure 3.1 MNIST Data

Results

We trained the model with 20 epochs and displayed the accuracy of the prediction.

Figure 3.2 gives the accuracy of the model prediction through training.

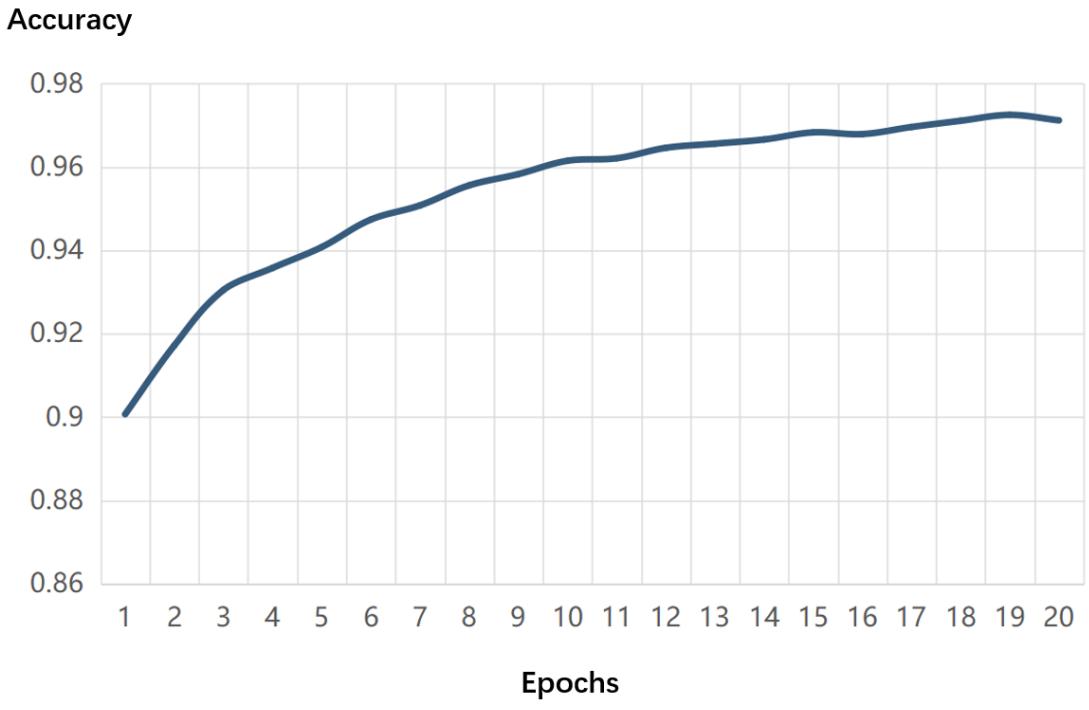

Figure 3.2 The Model Accuracy

Reference

Mustafa Umit Oner et al., Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning in A Pan-cancer Study, 2021