Solutions to Q1

1 Brief summary of literature:

Tumor purity is the percent of cancer cells present in a sample of tumor tissue, which is very important in the diagnosis of cancer because an accurate pathologic evaluation is based on it. There are 2 main types of ways to determine the tumor purity, **percent tumor nuclei estimation** and **genomic tumor purity inference**. To be more specific, percent tumor nuclei estimation is performed by pathologists via the way of counting the percentage of tumor nuclei over a region of interest, which usually costs time and energy and exists inter-observer variability. Genomic tumor purity analysis is considered as the golden standard but it is not applicable to low tumor content samples and there is no spatial information of locations of cancer cells.

The study presents multiple instance learning **(MIL)** model (presented in Figure 1.) to predict tumor purity from H&E stained histopathology slides and some notable improvements are presented as below:

(1) Use the **bag-level label** to represent a sample as a bag of patches and the weak labels can be more easily collected from pathology reports, electronic health records, or different data modalities.
(2) The MIL model has a novel **distribution pooling filter** which could produce stronger bag-level representations from patches' features.
(3) Obtain **the spatially resolved tumor purity maps** showing the variation of tumor purity over a slide.

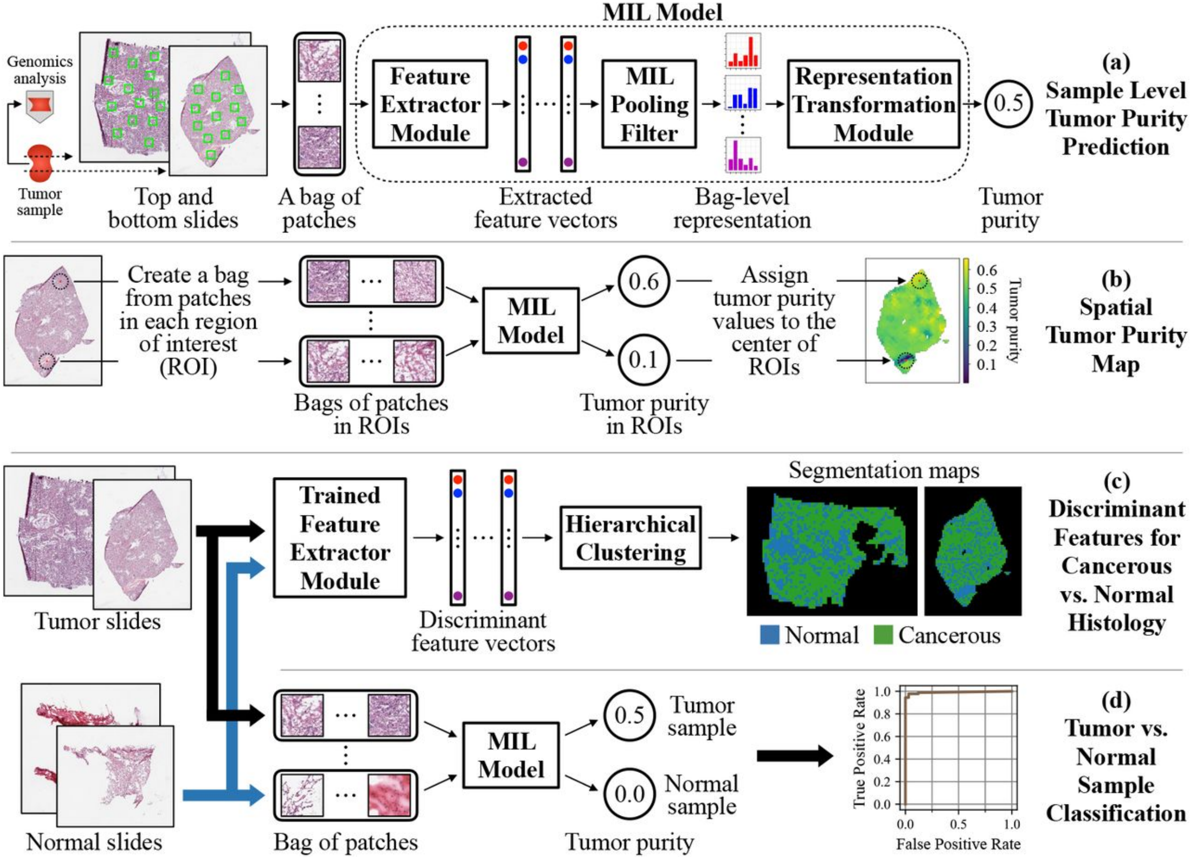

Figure 1[1]

In results, MIL model used in tumor purity prediction has shown its good performance in various aspects, such as lower mean absolute error, good qualities in classification of tumor or normal. Also, some other new discoveries in the determination of tumor purity are included:
(1) Tumor purity **varies spatially within samples** and this could be the probable reason for higher percent tumor nuclei estimation completed by pathologists in region of interest (ROI).
(2) Accounting for the spatial distribution of cancer cells, it is better to predict the tumor purity of a sample by **using both top and bottom slides**.
(3) The MIL model is based on tumor samples with a broad range of tumor purity values, thus it needs to be strengthened by samples with **low tumor content** in training.

3 MNIST data set:

MNIST data set is downloaded from http://yann.lecun.com/exdb/mnist/ with training data and testing data (including images and labels) and it is stored in the 'Original_dataset' folder.

2 Neural network architecture:

To train for the model, scripts are as below and model weights are saved in "saved_models" folder.

```
distribution_pooling_filter.py
resnet_no_bn.py
model.py
dataset.py
train.py
plot_loss.py
```