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

Abstract

The paper proposes a novel deep multiple instance learning (MIL) architecture to predict tumor purity from H&E-stained digital histopathology slides. Tumor purity is the percentage of cancer cells in the tumor tissue, which is critical for pathological evaluation and sample selection. MIL predictions are highly consistent with the purity of genomic tumor, which shows that the method is effective. Furthermore, the paper also identifies the reasons of the inaccuracy of pathologist's estimates and produces a cancer purity map to show tumor purity spatially within a sample. The study can be used for high throughput sample selection for genomic analysis and other cancer study.

Introduction

AI has made great progress in the past 50 years, and has also carried out in-depth research and preliminary application in the medical field. There are two traditional methods to estimate the purity of tumor: estimating the percentage of tumor nuclei by pathologists and genomic tumor purity inference. However, those methods are time consuming and fail to produce tumor purity spatially within the sample. There are two machine learning models that can be used to predict tumor purity from digital histopathological slides: patch-based model and multiple instance learning (MIL) model. The patch-based model is trained on a patch cropped from a slide using the corresponding pixel-level labels. The coverage of this approach is limited because they require pixel-level annotations that are rarely available, expensive and cumbersome, while the MIL paradigm only need bag-level annotations. It represents a sample as a bag of patches cropped from the sample slide and uses a sample level label as the bag label. Sample-level labels can easily be collected from pathology reports, electronic health records, or different data modalities. This study designed a novel MIL model to predict tumor purity from H&E-stained histopathology slides.

Methodology

MIL model consists of three parts. The first part is a feature extractor and the paper used resnet18 as feature extractor. It extracts features from each patch in a bag, and stack the feature vectors together as an input to the second part of the model. The second part of the model is to produce bag level features. It calculates each feature's distribution over patches in the bag. The third part is to transform bag level features into labels.

The paper tests their model on fresh-frozen sections in TCGA cohorts and ffpe sections in a local Singapore cohort. When testing on the Singapore cohort, the paper use transfer learning only adapting the first layer weights to train the network in order that the feature extractor performs better on ffpe slides. The paper calculates Spearman's rank and mean absolute error to evaluate the performance and proves the model's superiority compared to pathologists' estimates.

Results

The model can predict a sample's tumor purity accurately. The tumor purity is spatially calculated within a tumor sample. By comparing the predictions from upper slide and bottom slide using the Wilcoxon signed-rank test, the paper finds that in some cancer cohorts, the predictions vary significantly, while in some other cancer cohorts, the spatial variation is insignificant. Thus, predicting a sample's tumor purity using the top and the bottom slides is better than using only one slide.

The paper explores the reason of why pathologists tent to produce higher tumor purity of a slide. By analysing the tumor purity map, the paper discovers that the tumor purity within a slide is not homogeneous and pathologists might choose the region of interest which has higher tumor purity as the slide's tumor purity. The paper proves the assumption by analysing the

change of mean absolute errors between pathologists predicts and the average of the k highest tumor purity patches.

The model can learn discriminant features for cancerous and normal tissue histology. The paper extracts feature of patches cropped over a slide using trained MIL model and the segmentation maps are obtained by clustering over the extracted features vectors. The map is consistent with the LUAD histopathology during the qualitative assessment of the segmentation maps.

Discussion

MIL model proposed by the paper has several superiorities compared to pathologist's estimates and traditional deep learning method which predicts tumor purity on patch level.

The novel MIL model has lower costs. Pathologist's estimates need a lot of manual work and suffers from inter-observer variability. Traditional deep learning methods which produce predictions on patch level need a lot of pixel-level annotations which are hard to collect. The model, however, only needs sample-level annotations.

The MIL model produced tumor purity map in slide level, which is more accurate than using the average tumor purity in the region of interests as the slide tumor purity. The author points out pathologist's estimates often tend to produce higher tumor purity which might due to the region of interests they select are the higher in tumor purity respectively. Furthermore, the MIL model can help study the tumor distributions spatially.

The author also analyses several limitations of the model. First, the model may not predict tumor purity on low tumor purity samples as accurately as on high tumor purity samples. Second, the model can't be validated on the ffep dataset due to the different distributions of train set and validation set. Third, the model is deep learning based, so more data will enhance the model performance.

Application on the MNIST Dataset

Apply the MIL model in the paper on MNIST dataset. First download the MNIST dataset and select all the pictures containing 0 and 7. Shuffle the data randomly and then produce the zero percentages of each bag randomly in train set, validation set and test set. These percentages will be used as ground truths and be used to generate the data accordingly. After preparing the data, construct the model structure by defining patch-level feature extractor, distribution pooling layer and bag feature transformer. Then, train the data end to end for 200 epochs and generate loss curve for train and validation dataset (see Figure 1). Lastly, test the trained model on the test set and generate the relation curve between predictions and ground truth for each bag. Each point represents a bag in the test set (see Figure 3). The MSE

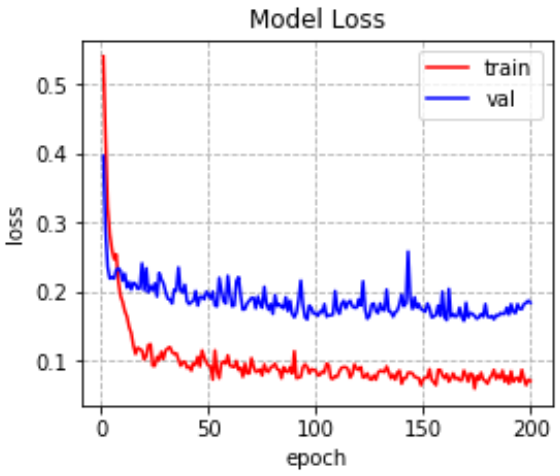

Figure 1 Loss curve for train and validation dataset

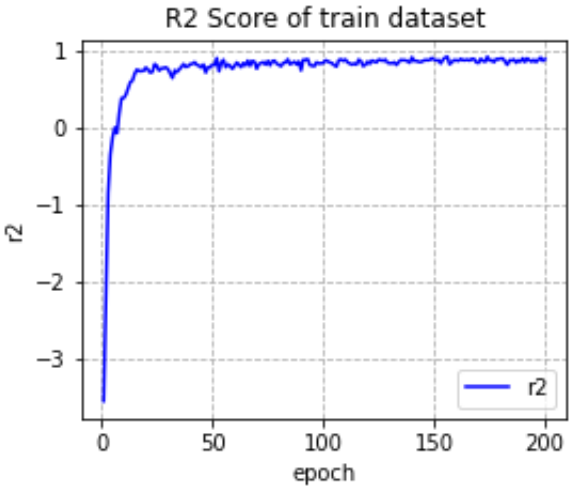

Figure 2 R2 Scores of train set

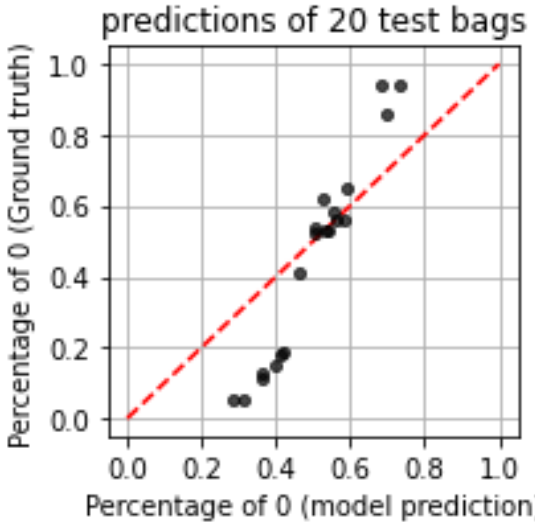

Figure 3 Comparison between model prediction vs. ground Is calculated between test set ‘s ground truth and predictions.

Table 1 Part of the predictions on test set

	Ground truth	Prediction
Bag 0	0.560	0.580
Bag 1	0.150	0.396
Bag 2	0.860	0.699
Bag 3	0.130	0.362
Bag 4	0.410	0.464
Bag 5	0.050	0.287
….	…..	…..
Mean Square Error	0.02806545	