Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study

1.Background

1.1 Definition and Importance

A tumor consists of a complex mixture of cells, the percentage of cancer cells within the tumor is called tumor purity.

An accurate tumor purity estimation is crucial for accurate pathologic evaluation and for sample selection to minimize normal cell contamination in high throughput genomic analysis.

1.2 Current methods and models

Percent tumor nuclei estimation: The pathologist counts the percentage of tumor nuclei over a region of interest in the slide. Referred to as percent tumor nuclei. Methods are used for sample selection and interpretation of results in the molecular analysis. Widely applicable, cellular level resolution but tedious and time-consuming, exists inter-observer variability between pathologists' estimates.

Genomic tumor purity inference: Inferred from different types of genomic data. Referred to as genomic tumor purity. Methods are used in genomics analysis and in correlational studies. Accepted as the golden standard. Produce consistent values on different cancer data sets in TCGA. Do not apply to the low tumor content samples. Do not provide spatial information of the locations of the cancer cells. Lose information about the spatial organization of the tumor microenvironment.

Patch-based models: Trained on a patch cropped from a slide, patch label determined by pathologists' pixel-level annotations. Predictions are obtained from the trained model, aggregated to obtain sample-level tumor purity prediction. Limited coverage, rarely available, expensive, and tedious.

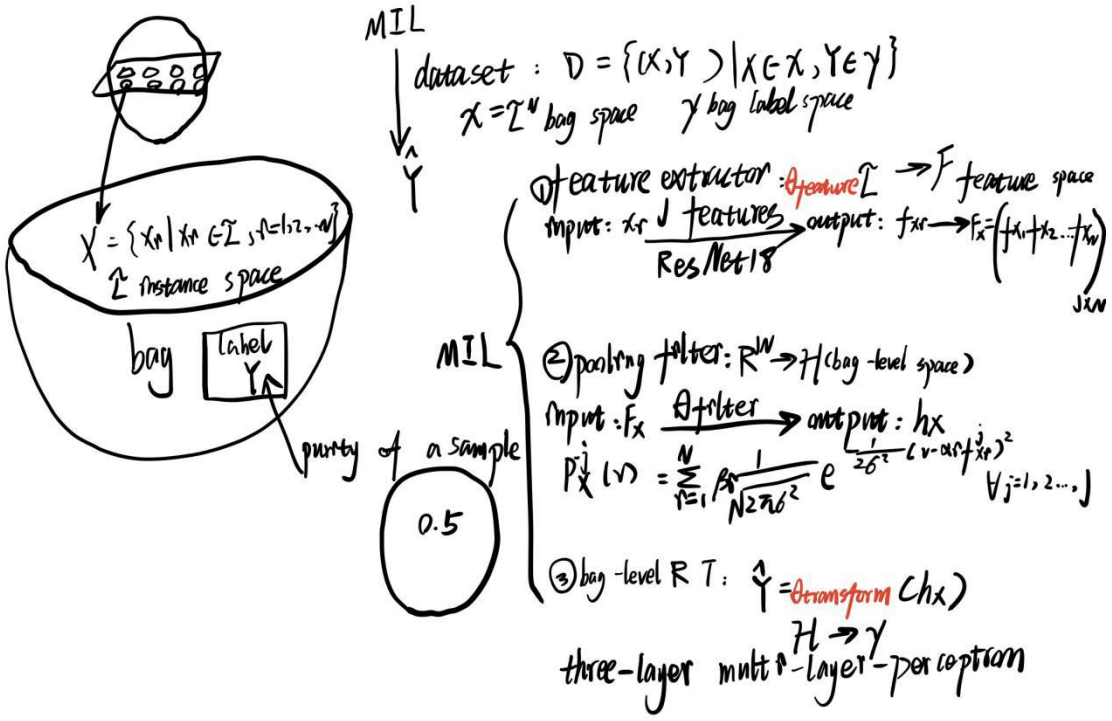
Multiple instance learning (MIL) models: Represents a sample as a bag of patches cropped from the sample's slides and uses a sample-level label as the bag label. Weak labels provide only aggregate information. Does not require pixel-level annotations. Easily be collected.

2.Method

2.1 Data Sets

H&E stained histopathology slides and corresponding genomic sequencing data for ten different cohorts in TCGA. Each patient had a tumor sample, and some patients also had matching normal samples. Digital histopathology slides in Singapore. Randomly segregated the data at the patient level into training, validation, and test sets. Tissue regions were detected by applying OTSU thresholding, image dilation, median filtering, and hole-filling, respectively. Over the detected tissue regions, non-overlapping512×512RGB images at20×zoom level were cropped.

2.2 MIL Model

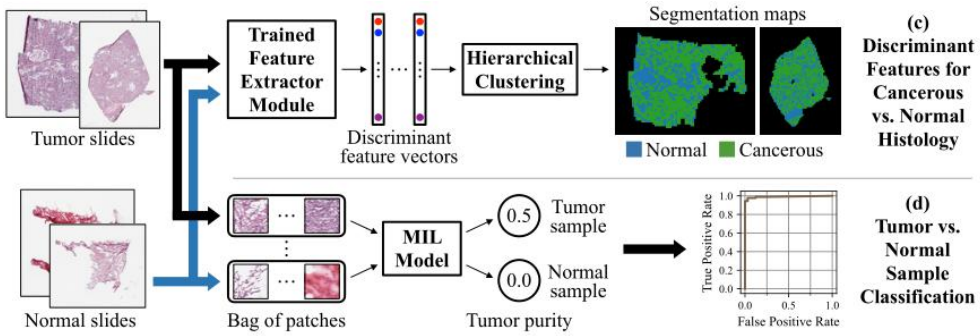


2.3 Training of MIL Model

ADAM optimizer with a learning rate of lr= 0.0001 and L2 regularization on the weights with a weight decay of weight_decay= 0.0005. The batch size was 1. Used absolute error as the loss function and employed early-stopping based on loss.

2.4 Segmentation of Histopathology Slides

For each patient with a matching normal sample, we used the trained feature extractor module to extract features of patches cropped over the slides of the tumor and normal samples of the patient. Then, we clustered the patches using hierarchical clustering over the extracted feature vectors. We calculated patient-specific distance thresholds in hierarchical clustering to capture inter-patient variations. Finally, we assigned a cancerous or normal label to each cluster based on the slide type.



2.5 Statistical Analysis

We obtained 95% confidence intervals for SRCC and area under the receiver operating characteristic curves using the percentile bootstrap method.

To compare the performance of two methods, we used Fisher's z transformation based method

on SRCC and Wilcoxon signed-rank test on absolute error values.
All statistical tests were two-sided and statistical significance was considered when P <0.05. We used scipy.stats (v1.4.1) python library for statistical tests.

3.Result

3.1 The MIL model's tumor purity predictions correlate significantly with genomic tumor purity values

SRCC is used as the performance metric. Obtained significant correlations (P < 0.05) between genomic tumor purity values and models' predictions from digital histopathology slides. The minimum correlation value obtained with MIL predictions was higher than the maximum correlation value obtained with pathologists' percent tumor nuclei estimates. This implies that MIL predictions are more consistent with genomic tumor purity values than the pathologists' percent tumor nuclei estimates.

3.2 MIL models' predictions have lower mean absolute error than percent tumor nuclei estimates

We used the Wilcoxon signed-rank test on absolute error values. Absolute error values in MIL predictions were significantly lower than ones in pathologists' percent tumor nuclei estimates in all cohorts except the LGG cohort. Two methods performed similar (Pcomp= 5.4e−02>0.05) in the test set of the LGG cohort.

3.3 Tumor purity varies spatially within a sample: top and bottom slides of a sample are different in tumor purity

The predictions of two slides are statistically compared using the Wilcoxon signed-rank test. There is a significant difference between the MIL predictions on the top and bottom slides of the same tumor sample. In all cohorts, at least 75% of samples have p-value P <1.0e−08and at least 95% of samples have p-value P <0.05.

3.4 Spatial tumor purity map analysis reveals the probable cause of pathologists' high percent tumor nuclei estimates

We obtained tumor purity maps by our trained MIL models in different TCGA cohorts and conducted error analysis over them to test hypothesis.
Pathologists may tend to select high tumor content regions to estimate percent tumor nuclei.

3.5 The MIL model learns discriminant features for cancerous vs. normal tissue histology

For each patient having both tumor and matching normal samples, features of patches cropped over the slides of the tumor and normal samples were extracted using the trained feature extractor module of the MIL model. Then, slide-level cancerous vs. normal segmentation maps were obtained by performing a clustering over the extracted feature vectors.
We observed that segmentation maps were consistent with the LUAD histopathology during the qualitative assessment of the segmentation maps. While healthy tissue components were labeled normal, regions invaded by neoplastic cells were labeled cancerous.