INTRODUCTION:

Even more research are demonstrating the importance of the tumour microenvironment (TME) in predicting outcomes and treatment success.[1] The proportion of cancer cells in the tumour tissue is defined as tumour purity, which indicates the characteristics of TME. It has an influence on the appropriate evaluation of molecular and genetic characteristics. [2]From H&E stained digital histopathology slides, an unique deep multiple instance learning model predicted tumour purity. In eight separate TCGA cohorts, the model correctly predicted tumour purity from slides of fresh-frozen sections. Additionally, spatially resolved tumour purity maps demonstrate that tumour purity changes spatially within a sample.

TUMOR PURITY:

The tumour content of the samples, which is measured as tumour purity, is one of the most important parameters determining the quality of genomic analysis. Furthermore, the sample must have a high tumour content in order to identify any genetic alterations. As a result, an accurate tumour purity assessment is critical in clinical practise. For example, a sample chosen based on an inflated tumour purity might result in a false-negative test result.

There are two methods for estimating tumour purity: percent tumour nuclei estimates and genomic tumour purity inference. In essence, the pathologist counts the number of tumour nuclei in a specific area on the slide. It interprets data at the molecular level. Counting tumour nuclei, on the other hand, is difficult and time-consuming.

The second technique, which is often regarded as the gold standard, is genetic tumour purity. In The Cancer Genome Atlas(TCGA), genomic approaches generate consistent values across various cancer data sets . They do not, however, apply to samples with minimal tumour concentration. Furthermore, they do not give spatial information about the cancer cells' locations. This is a critical component of therapeutic response. As a result, the strengths and limitations of genomics methodologies and pathologists' slide reading procedures are distinct.

A machine learning model is created to predict tumor purity using H&E stained histopathology slides to be compatible with genomic tumor purity values. The model is cost-effective compared to genomics methods or pathologists’ readings since it uses readily available histopathology slides in the clinic and involves a few manual steps. It also provides information about the spatial organization of the tumor microenvironment.

Patch-based models and multiple instance learning (MIL) models are two types of machine learning algorithms that may be used to predict tumour purity from digital histopathology slides. The appropriate patch label derived based on pathologists' pixel-level annotations is used to train a patch-based model on a patch clipped from a slide. Annotations at the pixel level are not required in the MIL paradigm. A sample is represented as a bag of patches clipped from the sample's slides, with the bag label being a sample-level label. They may also be easily gathered from pathology reports, electronic health records, and other data sources.

COMPARISON BETWEEN THE EFFICIENT MODELS AND INEFFICIENT METHODOLOGY:

To evaluate the models’ performance in 10 different TCGA cohorts, correlation analyses between genomic tumor purity values and the MIL models’ predictions are conducted. 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. One of the primary reasons for this superiority is that the MIL models were trained directly on genomic tumor purity values, which enabled the MIL models to learn associated features.

Apart from Spearman’s correlation coefficients, the mean-absolute errors were checked between genomic tumor purity values and MIL models’ predictions, genomic tumor purity values, and pathologists’ percent tumor nuclei estimates. It is observed that MIL predictions had lower mean-absolute-error and higher Spearman’s correlation coefficient than pathologists’ percent tumor nuclei estimates. Hence, the MIL model performs better than the other methods.

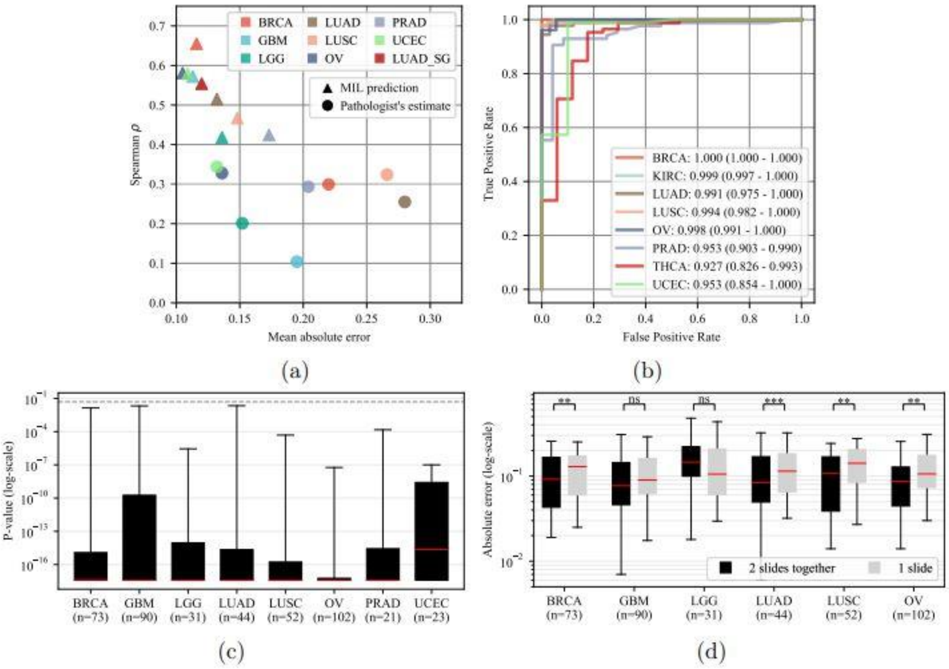

Figure1. (a,b) MIL models perform better than percent tumor nuclei estimates and successfully classify samples into tumor vs. normal. (c,d) The top and bottom slides of a tumor sample are different in tumor purity

<u>**SOURCES OF ERROR IN MIL PREDICTIONS:**</u>

TIn comparison to standard deep learning data sets, which contain millions of independent samples, the data set employed a small number of patient records. We also predict that as the number of patients grows, the performance will improve.

Histopathology slides from the top and bottom regions of the tumour part are used in the MIL model. Significant differences in tumour purity are discovered between the top and bottom regions of the tumour samples. As a result, the prediction error for samples with only one slide is likely to be larger.

The model's predictions are based on histopathology slides stained with H&E. However, genomic tumor purity values were based on DNA data, and all the effects of genetic changes may not be observable from the slides due to the particular dying characteristics of H&E staining.

METHODOLOGY:

Datasets:

H&E stained fresh-frozen section histopathology slides and corresponding genomic sequencing data for ten different cohorts are downloaded.

Following the TCGA Standard Operating Procedures, each sample was sliced into parts. Then, for genetic study, one of the parts was sequenced. From the top and bottom portions of the same part, one or two related histopathology slides (top and bottom slides) were also generated.

The data is divided into three sets: training, validation, and test, all of which had equal tumor purity distributions. OTSU thresholding, image dilation, median filtering, and hole-filling were used to discover tissue sections inside histopathology slides.

MIL Model:

The objective is to predict a bag label Y for a given bag of instances $X = \{xi\ |xi \in I, i = 1, 2, \cdot\cdot\cdot, N\}$ where I is the instance space and N is the number of instances inside the bag. Here, a bag label Y is the genomic tumor purity of a sample, and a bag X is a collection of cropped patches over the sample's slides. Let D be a MIL dataset such that $D = \{(X, Y)\ |\ X \in X \text{ and } Y \in Y\}$, where X = I N is the bag space, and Y is the bag label space. The first stage is a feature extractor module. ResNet18 model was used as the feature extractor module and a three-layer multilayer-perceptron as the bag-level representation transformation module.

The second stage is a MIL pooling filter module. It takes the feature matrix as input and aggregates the extracted feature vectors into a bag-level representation. The last stage is a bag-level representation transformation module. It transforms the bag level representation into the predicted bag label.

The neural networks are used to implement the feature and transform to parameterize the learning process fully for filter, novel 'distribution' pooling filter, which shown to be superior to point-estimate based MIL pooling filters, like max-pooling or mean-pooling. This system of neural networks is end-to-end trainable. Trained models using the 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. Absolute error as the loss function and employed early-stopping based on loss in the validation set are used to avoid overfitting. Then, evaluated the model on the unseen test set.

To compare the performance of two methods (MIL models' predictions and pathologists' percent tumor nuclei estimates), Fisher's z transformation is implemented on Spearman's rank correlation coefficients and Wilcoxon signed-rank test on absolute error values.

REFERENCE:

1. https://onlinelibrary.wiley.com/doi/full/10.1002/cam4.3505

2. https://academic.oup.com/bioinformatics/article/35/21/4433/5490858

3. https://academic.oup.com/bib/article-abstract/22/6/bbab163/6265216?redirectedFrom=fulltext