Predicting Tumor Purity from Histopathology Slides, a Deep Learning Approach

Tumor Purity and its Implications

The tumor microenvironment consists of a heterogenous mix of tumor as well as normal cells. The normal cells include stromal cells, capillary endothelial cells, immune cells and fibroblasts. The ratio of tumor to normal cells within a tumor is known as the tumor purity. In certain cancer types, low tumor purity levels indicate poor prognosis. For example, in gastric cancer, low tumor purity is an indicator of immune evasion, as well as propensity of tumor cells to metastasis [1].In colorectal cancer, low tumor purity was associated with worse survival in patients, expression of immune evasion molecules such as PDL-1, as well as an enrichment of pro-tumor microenvironment cells such as M2 macrophages and neutrophils[2].

In addition to contributing to tumor prognosis, a low tumor purity may also confound genomic techniques in diagnostics. Many targeted therapies rely on the presence of activating mutations in oncogenes for their therapeutic effect. Thus, it is crucial that these relevant mutations be detected in cancer biopsies. A low tumor purity results in low signal levels of the mutated allele, buried beneath the signal of the normal allele. This impacts the success rate of mutation calling during sequencing. Mutation calling becomes unreliable when the tumor purity falls below 60% [3].

Measuring Tumor Purity

Thus, ensuring a reasonable tumor purity in biopsy samples is essential for the diagnosis of cancer using genomic techniques. Tumor purity is estimated using three general classes of techniques. The first technique involves manual examination of biopsy slides by a trained pathologist. The morphology, particularly the nuclei, of cancer cells appear different to that of normal cells, and pathologists can use these differences to discriminate between cancer and normal cells. However this method is tedious, expensive and results in inconsistent results between different pathologists [4]. The second technique involves estimating the tumor purity using molecular techniques. The genomes, transcriptomes and proteomes of cancer cells are generally distinct from normal cells, and thus , high throughput sequencing or proteomics might uncover disparities between the signatures of tumor biopsies and normal biopsies, the disparity increasing with tumor purity. However, this requires prior knowledge of the molecular signature of the tumor, such as the presence of mutations , copy number variants and expression profiles. Furthermore, these techniques produce bulk averages of molecular signatures over the entire biopsy, and are unable to discern regions of high purity and low purity [4]. Finally, these techniques are susceptible to tumor heterogeneity, as variations in profiles across tumor populations may result in high variance of the expression levels and mutation signatures in biopsies compared to the normal controls.

The third technique is relatively novel and adopted by this paper[4]. It involves substituting the traditional pathologists' examination with machine learning algorithms trained to discriminate between the morphologies of normal and cancer cells. Typically,feature extraction is done using a deep convolutional neural network (CNN), and several different types of ML algorithms are trained over the features extracted

by the CNN,including support vector machines, gradient boosted trees and multilayer perceptrons, using the tumor purity labels as supplied by pathologists or molecular techniques.

Convolutional Neural Network and Multiple Instance Learning Problems

A convolutional neural network is a neural network that specializes in extracting information from images. It exploits the fact that pixels in the same neighbourhood are spatially correlated, to produce features such as edges and shapes. The CNN is a stack of layers known as convolution layers. Convolution detects the extent in which a particular feature, such as an edge, exists at a particular (n x n pixel) region in an image. It does this by computing the elementwise product of a set of weights corresponding to each feature (represented as a n x n matrix) over the grid of pixels of the same size in the region, and summing up the products. The weights themselves are learned using a process known as backpropagation, where the weights are updated using the derivative of the error of the model (or loss) for each training example (w.r.t to the weight). Over many iterations, the loss goes down and the model learns features that approximate the image more accurately.

The outputs of convolution layers are fed through a pooling layer which computes a summary statistic about the outputs of the convolutions over a (m x m pixel) neighbourhood. The most commonly used max pooling filter reports the maximum extent that a particular feature exists in the neighbourhood. The outputs of pooling layers are then used as inputs to the next convolution layer. Later layers learn more complex features (for example, the layers in a face detection CNN may learn edges, the next circles,then eyes, and then faces). Finally, the last layer learns global features, which are then weighted and summed up using a (few) fully connected layer(s). The output of the layer(s) gives the prediction of the model.
Oftentimes, the architectures and weights of successful models trained on a particular problem are reused for separate problems. Removing the fully connected layers produces a feature extractor, and custom fully connected layers or other ML models can be trained on the features extracted to solve the new problem. For example, the authors repurposed ResNet18 [6], originally trained to classify everyday objects including wheels, pencils and birds, into a model that is able to differentiate between normal and tumor cells by adding their own custom fully connected layers and training them on histopathology slides.

The prediction of tumor purities from histopathology slides represents a unique challenge. Traditional CNNs are designed to produce one output label per image. In the case of estimating tumor purity from biopsy slides, the resolution of the entire image (which may be above 100k x 100k pixels) is too large for most neural networks to operate on [5]. Thus, algorithms must resort to sampling multiple patches (of typically a few hundred px) from the original image. These patches may vary widely in terms of tumor purity. During training, the label for each patch is unknown. Only the tumor purity of the entire biopsy (i.e the image) is supplied as a label. The global purity thus represents an aggregate label over all the patches in the image, which the ML algorithm has to learn from the features extracted from each image. More generally, the problem is known as Multiple Instance Learning (MIL), where the model is trained on labels given to sets of images(otherwise known as bags) and the objective is to predict the labels of the bags, without access to the true labels of each instance.

There are two general ways in which this is achieved. The first is to learn a representation of each instance, and then aggregate them together using a pooling layer/model to give the bag label. The second is to ignore the fact that each bag is made of instances , extract embedded features corresponding to the bag, and to predict the bag label in the same fashion as an instance is predicted[9] .

Study Design and Results

In the MIL example described in the paper, the author's used the first approach. First, a set of 128 features are extracted for each patch in the image (using ResNet18) and the features are pooled together to obtain a marginal distribution. To compute the marginal distribution, the feature values are binned , and the number of instances in the bag falling into each bin is computed. The authors then used a gaussian kernel to smooth out the bin edges.For each pooled feature, a fully connected neural network is trained to compute the global purity for the aggregated features.

The resultant model was then trained on histopathology slides from The Cancer Genome Atlas and SIngaporean patients , and evaluated on a separate hold-out test set from the same sources, with the labels coming from tumor purity measurements using ABSOLUTE, a genome based technique. The Evaluation set predictions showed reasonable (but not excellent) correlation with the ABSOLUTE purity (correlation coefficients around 0.2-0.5) , and also showed robust performance on formalin fixed slides prepared differently from the fresh frozen slides the model was trained on. They were also able to spatially resolve areas of different tumor purity, and use these measurements to infer that tumor purity scores tend to be inflated by pathologists because they selectively analysed areas with high purity. Finally, they found that their model was able to discriminate between tumor and normal cells, and correctly identify 90% of tumor slides as cancerous while keeping the proportion of falsely classified normal cells below 10%. The authors concluded that , by using MIL approaches, their model is superior to currently existing tumor purity prediction models, because it does not need each patch to be manually annotated with the local purity to produce the training labels, and instead uses global labels which are easily obtainable from clinical records or molecular tumor purity measurements.

References

[1] Gong, Z., Zhang, J., & Guo, W. (2020). Tumor purity as a prognosis and immunotherapy relevant feature in gastric cancer. *Cancer medicine*, *9*(23), 9052–9063. https://doi.org/10.1002/cam4.3505

[2] Mao, Y., Feng, Q., Zheng, P., Yang, L., Liu, T., Xu, Y., Zhu, D., Chang, W., Ji, M., Ren, L., Wei, Y., He, G., & Xu, J. (2018). Low tumor purity is associated with poor prognosis, heavy mutation burden, and intense immune phenotype in colon cancer. *Cancer management and research*, *10*, 3569–3577. https://doi.org/10.2147/CMAR.S171855

[3] Cheng, J., He, J., Wang, S., Zhao, Z., Yan, H., Guan, Q., Li, J., Guo, Z., & Ao, L. (2020). Biased Influences of Low Tumor Purity on Mutation Detection in Cancer. *Frontiers in molecular biosciences*, *7*, 533196. https://doi.org/10.3389/fmolb.2020.533196

[4]Oner, U., M., Chen, J., Revkov, E., James, A., Heng, S. Y., Kaya, A. N., Alvarez, J. J. S., Takano, A., Cheng, X. M., Lim, T. K. H., Tan, D. S. W., Zhai, W., Skanderup, A. J., Sung, W. K., Lee, H. K.,
Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study, Patterns,2021,100399,ISSN 2666-3899, https://doi.org/10.1016/j.patter.2021.100399.

[5].Xu, Y., Jia, Z., Wang, LB. *et al.* Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinformatics* 18, 281 (2017). https://doi.org/10.1186/s12859-017-1685-x

[6] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.