Q1 Report

1. Machine learning in predicting tumor purity

In the past, ESTIMATE provided tumour purity predictions in individual samples with a 85% confidence interval of the validity of the prediction. However, this method has some limitations such as the inability to accurately infer tumour cellularity of hematopoietic or stromal tumours (Kosuke, 2013). Therefore, Machine learning method is introduced to predict tumor purity. According to the article, patch-based models and multiple instance learning (MIL) models are common for tumor purity.

The neural network framework used in this paper mainly consists of three parts. The first part is to extract the features of all the patches in a bag. The second part is to filter the features extracted from the first part, filter out unimportant features according to strict mathematical derivation, and filter all features of all patches in the entire bag. The third part is to implement regression prediction for the filtered features according to the multilayer perceptron. The most prominent feature of the paper is dataset processing, which is different from the general network for image segmentation based on the pixel level of medical images, and then conduct regression prediction about tumor purity based through medical image. There are two main advantages for this method. Firstly, it reduces the requirements of the dataset and increases the universality of the model. Secondly, some relatively mature image feature extraction frameworks can be used as the base. part of the frame.

In general, the application of machine learning in the field of cancer has the following benefits. Firstly, it can largely replace labour and avoid unnecessary waste of human resources. Secondly, the judgment of machine learning and neural network is completely based on the learned feature data, and will not be affected by emotional factors such as personal feelings, personal experience, etc. And thirdly, machine learning-based method could learn some difficult-to-observe cancer characteristics based on medical images, further increase the accuracy of recognition.

2. Implementation in MINST dataset

It is obviously that there is no way to achieve a perfect fit to use such simple dataset like MNIST to model the cancer purity regression predictions of the medical images in the paper. Even though we can randomly select 100 images to form a bag composed of numbers 0 and 7 and express the number of 7 as the purity of the tumor. And as mentioned in the paper, the first step is to use resnet to extract features from all images, and then filter the features, and the last step is regression. Since the images in this dataset are too simple (only 28x28 in size, and the pixel values are only 0 and 1), it is easy to overfit in such a large model. However, by simplifying the medical images in the paper with such a simple dataset and running through the model, the feasibility of this model can be verified.

For the implementation result, Figure 1 shows a graph of the loss drop when applying the MNIST dataset to the training of the model framework in the paper.

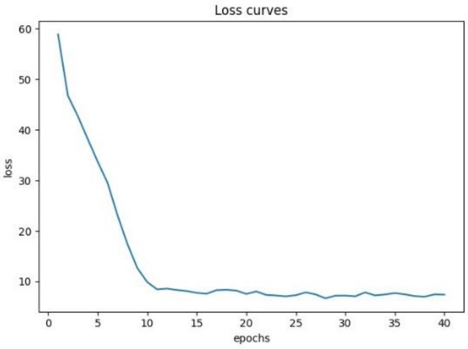

Figure 1 Training loss

Figure 2 shows prediction result on the test dataset when using the trained model. Based on the result, we can see that the model can basically fit the real data.

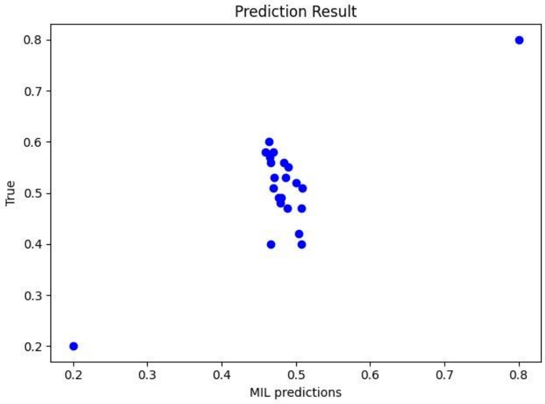

Figure 2 Prediction result

3. Biological and cancer domain knowledge

Cancer is a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body, which seriously affects people's health. According to statistics, there were an estimated 18.1 million cancer cases around the world in 2020(National Cancer Institute, 2021). Therefore, cancer is a key research object in biology and medicine, among which tumor purity is the field of recent research. Tumor purity is defined as the proportion of cancer cells in the tumor tissue, which consist of a complex mixture of cells, such as cancer cells, normal epithelial cells, and stromal cells. According to Figure3, low purity CC conferred worse survival, and tumor purity was identified as an independent prognostic factor (Mao, 2018). The benefit of tumor purity is that it can reduce the workload of pathologic evaluation, but also crucial to genomic analysis. In addition, tumor purity could help understand the microenvironment of the tumor. Therefore, it is very important to have a fast and accurate method to predict tumor purity (Mustafa, 2021).

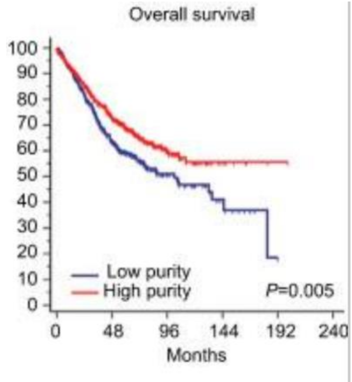

Figure 3 survival rate on low and high purity

4. Reference

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

Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2021). Cancer Statistics, 2021. *CA: a cancer journal for clinicians*, *71*(1), 7–33. https://doi.org/10.3322/caac.21654

Oner, M. U., 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. (2021). Obtaining Spatially Resolved Tumor Purity Maps Using Deep Multiple Instance Learning In A Pan-cancer Study. bioRxiv, 2021.2007.2008.451443. https://doi.org/10.1101/2021.07.08.451443