The tumor microenvironment is an environment around the tumor regulating tumor survival and promotion function. These non-cancerous cells such as blood vessels and fibroblasts interact with tumor closely. Besides that, tumor purity is the proportion of cancer cells in the tumor tissue. It is a crucial factor not just affecting the quality of genomic analysis but also affecting high throughout data acquisition and analysis. [1] So, an accurate tumor purity estimation is crucial for accurate pathologic assessment and for sample selection to minimize normal cell contamination in the genomic analysis.

Normally, tumor purity can be estimated by two main methods: percent tumor nuclei estimation and genomic tumor purity inference. The former method is widely applicable, involving pathologists to estimate tumor purity by reading H&E stained histology slides. In essence, the percentage of nuclear nuclei over the region of interest in the slide is calculated manually. However, counting tumor nuclei individually is high labour-intensive and tedious. The latter method makes estimates in tumor purity using different types of genomic information, such as gene expression, somatic mutation. [2] The genomic method produces consistent values on different cancer datasets in The Cancer Genome Atlas (TCGA). However, a low tumor content sample does not include in the dataset. Between these two approaches, there are different strengths and limits to deal with. It would be best to have a hybrid solution integrating both methods and the machine learning model might be the silver bullet.

Harnessing machine learning technology in every domain has been becoming a trend in the past ten years as it offers promising benefits to the aspect of our society and the world. Machine learning applications in healthcare and biomedical science are making better decisions and providing better guidance to medical professionals and biomedical scientists. These developments in machine learning promise to make it an important tool in biomedical research. Indeed, the number of ML-related papers and patents in biomedical research has grown exponentially. [3]

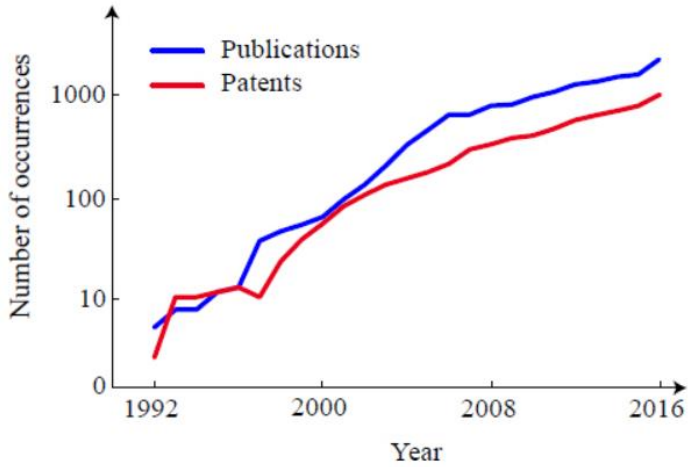

Figure 1: Trend in use of machine learning in biomedical science between 1992 to 2016. [3]

In genomics analysis, several deep learning models were proposed to classify the tumor and normal cells before this study. For example, the two current papers are Coundray et al. which deep learning model was trained on tumor vs normal classification and Fu et al. which was fine-tuned on pathologists' percent tumor nuclei estimates in a transfer learning setup. Despite the high AUC of their deep learning models, the severe data leakage problem might be a concern to pay attention to. So, this study developed a more comprehensive approach: multiple instance learning (MIL) model. It predicts the tumor purity from H&E stained histopathology slides and uses tumor purity values as a ground truth. This makes the predictions are consistent with the genomic tumor purity values inferred from the genomic sequencing data. To fully understand the capability of the MIL model, discriminant features for cancerous vs. normal histology and tumor vs. normal sample classification were studied and predicted as well. [1]

The MIL model successfully predicted tumor purity from slides in eight different TCGA cohorts and formalin-fixed paraffin-embedded sections in a local Singapore cohort. [1] The outcome showed highly consistent with genomic tumor purity values, which were inferred from genomic data. The tumor purity maps derived from the model also reveal the probable reasons for high percentage tumor nuclei estimates happened on pathologist's estimation. Inappropriate size and selection of region-of-interest by pathologists might be the probable cause. Moreover, the proposed model showed significant predictions not even on different TCGA cohorts and also classified the discriminated tumor features from the normal sample with AUC values greater or equal to 0.927. [1] In short, this model can be utilized for high throughput sample selection for genomic analysis, which will help reduce pathologists' workload and decrease inter-observer variability.

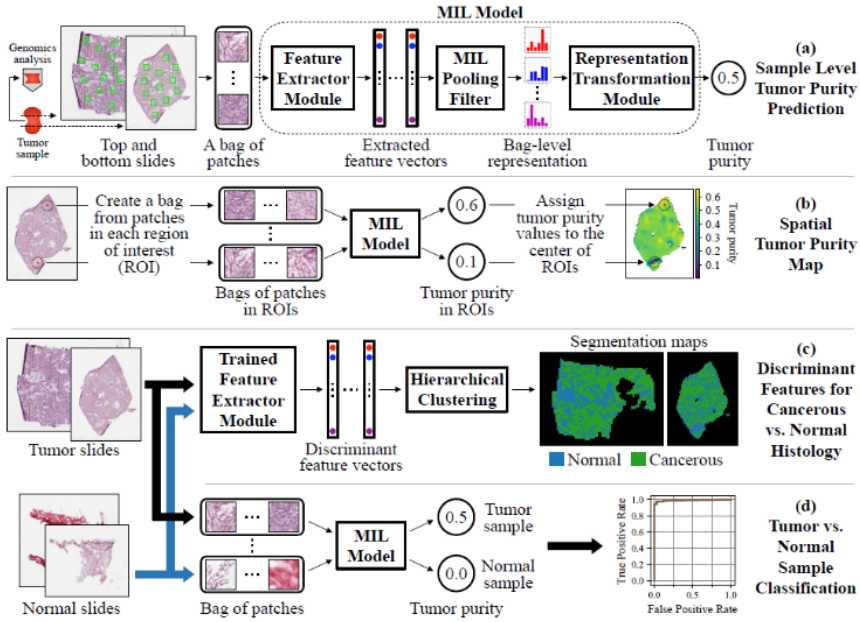

Figure 2: A novel MIL model predicts sample-level tumor purity from H&E stained digital histopathology slides. [2]

The proposed MIL workflow on the MNIST dataset is shown in Figure 2. My initial idea is to construct the MNIST data set by filtering out digit 0 and 7 data and grouping them into bags. Each bag consists of 100 images with a fraction x of digit 0 and (1-x) of digit 7. I will assign bag labels for each bag with the conditions: bag label as "1" if one instance has a label "0"; bag label as "0" if all the instance labels are "7". As shown in the Figure 2, the bag with orange color has a label of "1" while the bag with blue color has a label of 0. The proposed MIL mode consists of two modules: feature extractor module and bag level transformation module. Both modules are used ResNet modal for extracting features and predicting the outcome.

Although my code is not working and doesn't show any results, at last, I have tried my best to implement any segment of the MIL workflow within a week. It is quite challenging to develop a MIL model starting from scratch. But it has motivated me to brush up on my python skill in machine learning model development and biomedical applications after this test.

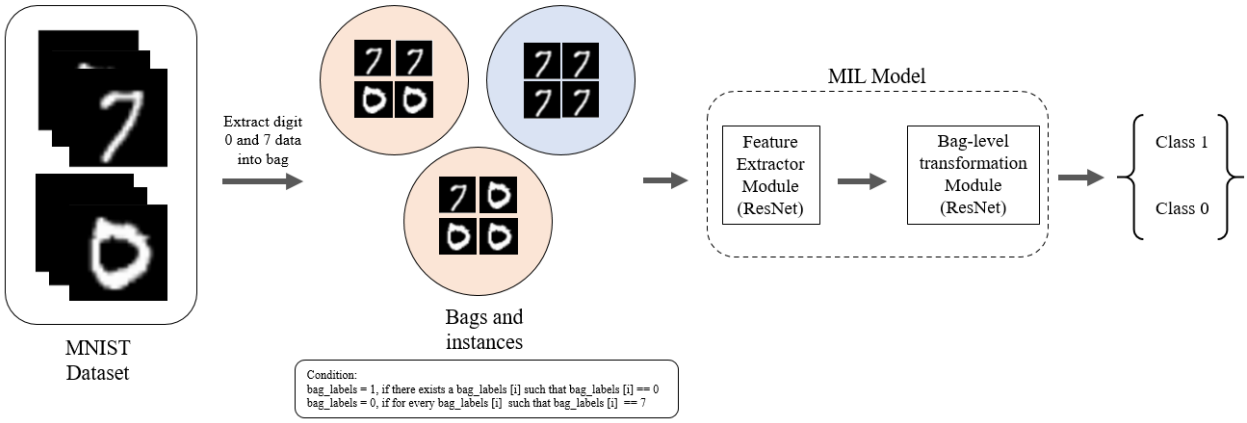

Figure 2: The proposed MIL workflow on the MNIST dataset.

Reference

1. M. U. Oner, J. Chen, E. Revkov, A. James, S. Y. Heng, A. N. Kaya, J. J. Alvarez, A. Takano, X. M. Cheng, T. K. Lim, D. S. Tan, W. Zhai, A. J. Skanderup, W.-K. Sung, and H. K. Lee, "Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study," 2021.
2. D. Aran, M. Sirota, and A. J. Butte, "Systematic pan-cancer analysis of tumour purity," *Nature Communications*, vol. 6, no. 1, 2015.
3. Kording Kp, Benjamin As, Farhoodi R, et al. The Roles of Machine Learning in Biomedical Science. In: National Academy of Engineering. Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2017 Symposium. Washington (DC): National Academies Press (US); 2018 Jan 22. Available from: https://www.ncbi.nlm.nih.gov/books/NBK481619/