Konstantin Howard
Computational Methods for Precision Health
Advised by Olga Troyanskaya

Analyzing Tumor Histologies to Diagnose Cancer and Estimate Tumorous Cell Ratio

Motivation and Goal

When tumors arise from abnormal cell growth, doctors take biopsies of the tumorous tissue to diagnose whether the growth is benign or malignant. From the biopsies, slide images of tissue stained with Hematoxylin & Eosin (to give a purple-pink definition to the cells within) are studied. With these samples pathologists are also able estimate to the ratio of tumorous to healthy cells, giving a measure of malignancy. The counting of tumorous cells is performed manually under a microscope in a laborious, expensive, and notoriously error-prone process. My goal is to use machine learning to perform this analysis given histology slide images and patient data as input. Tumor analysis not only informs cancer diagnosis but is also crucial in deciding which therapies would be most effective to treat that patient.

Problem Background and Related Work

The issues with manual analysis are well documented. One study comparing the tumor cell ratio estimates of several pathologists found that only 14% of observations were accurate, with significant deviations (Smits et al., 2014). Error resulted from both systematic bias but also from random mistakes, implying that even more-skilled pathologists would not constitute a complete solution. Recent research has successfully attempted to use machine learning to the same end as my project: a model developed by NEC Laboratories obtained an average absolute error of 6% whereas the error was 20% for human pathologists (Cosatto et al., 2021). However, such methods are not widely used and the work tends to still be done by hand.

Critically, my project will incorporate other data about the patient, particularly demographic information, to hone the effectiveness of the model. This will allow the model to be evaluated for any biases towards different groups of patients and account for any discrepancies in diagnoses. These kinds of biases plague many methods of diagnosis and are an area of healthcare that precision health methods attempt to tackle. If the model I develop does not work well for some patients, that must be captured in my results.

My model will also make use of U-Net, a CNN segmentation architecture designed specifically for biomedical imaging. U-Net is very effective at delineating individual cells within a slide image, even when borders are ambiguous and inner-cell contents are visible, outperforming sliding-window CNNs in competition (Ronneberger et al., 2015). On less-magnified images like the histologies in my data, a U-Net model was successful in Gleason-grading prostate health (Li et al., 2017). Previous research has used U-Net architecture for various segmentation tasks for tumors, but not for estimating tumor cell ratio.

Approach

The data will be drawn from cases in the National Cancer Institute's Genomic Data Commons and Tissue Image Library.[1] I plan to implement a CNN model using U-Net architecture and train it on tumor histology slide images to classify not only whether the growth is benign or malignant but also to accurately estimate the ratio of tumor cells in the sample. U-Net will help the model to segment slide images, thus providing insight on what features are critical to a diagnosis. I will use patient data to evaluate the model's effectiveness on different demographics. Given the results of this analysis, patient data will be incorporated to combine the predictive power of the U-Net CNN with an FNN using patient data as input.

Plan

1. I will download the slide images from the dataset. Slide images will be divided into smaller frames to narrow the focus of the classifier and obtain more images to train on from a smaller number of individual cases. Images may be augmented for training purposes as well. Images will need to be associated their respective case by ID tags. This will be done for several different types of cancers/organ sites by my partners as well since some types may be easier to classify than others.
2. I will implement the U-Net CNN using PyTorch (several open-source versions already exist). I will begin with just the binary classification of benign or malignant before attempting the tumor cell ratio. Given the results in this step, I will narrow my focus to one type of cancer/organ site.
3. I will implement the tumor cell ratio estimation. This step may prove to be very difficult and is contingent on my initial success in classifying just malignancy.
4. I will evaluate the effectiveness of the model based on different patient groups. If discrepancies arise, I will add an FNN to the model to account for them.

Evaluation

The success of this model will be determined by classic machine learning methods: withholding a portion of the dataset from training for testing purposes. The binary classification of malignancy is easy to evaluate. False negatives are particularly important in the case of cancer diagnoses, so this will be an important measure of accuracy. Since the tumor cell ratio estimation outputs a range from 0-100% (usually divided into 10% increments), evaluation will also consider how much the estimate deviates from the ground truth. Once again, undercounts will be more important to track. With the addition of the FNN, the model will be evaluated again to see how the incorporation of other patient data affects it. I will also be able to compare the success of my model with those of my partners, who will be using histologies from other kinds of cancer/organ sites.

[1] GDC and GTEx- publicly available at https://portal.gdc.cancer.gov/ and https://biospecimens.cancer.gov/gtexbiobank/histology_viewer.asp

Works Cited

Eric Cosatto, Kyle Gerard, Hans-Peter Graf, Maki Ogura, Tomoharu Kiyuna, Kanako C. Hatanaka, Yoshihiro Matsuno, Yutaka Hatanaka, A multi-scale conditional deep model for tumor cell ratio counting, Proc. SPIE 11603, Medical Imaging 2021: Digital Pathology, 1160308 (15 February 2021); https://doi.org/10.1117/12.2581108

Li, J., Sarma, K. V., Chung Ho, K., Gertych, A., Knudsen, B. S., & Arnold, C. W. (2018). A Multi-scale U-Net for Semantic Segmentation of Histological Images from Radical Prostatectomies. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, *2017*, 1140–1148.

Olaf Ronneberger, Philipp Fischer, Thomas Brox
Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS, Vol.9351: 234--241, 2015

Smits, A., Kummer, J., de Bruin, P. *et al.* The estimation of tumor cell percentage for molecular testing by pathologists is not accurate. *Mod Pathol* **27**, 168–174 (2014). https://doi.org/10.1038/modpathol.2013.134