6.S897/HST.956 Machine Learning for Healthcare

Lecture 12: Machine Learning for Pathology

Instructors: David Sontag, Peter Szolovits

1 Introduction

This guest lecture was taught by Andrew H Beck MD PHD, CEO at PathAI.

2 Background to Pathology

When patients go to the doctor, it is common for their symptoms and signs of disease to be taken down. From there if there is concern that there is something more structural, patients are sent to radiology to get images from inside of the body to evaluate the signs and symptoms ideally to distinguish between diseases. This information however is often not data-rich. More often than not, Radiology simply gives impressions of what is believed to be going on and requires further testing to get a better diagnosis. The more serious results will require tissue specimens which will ultimately determine the type of treatment that will have to be applied.

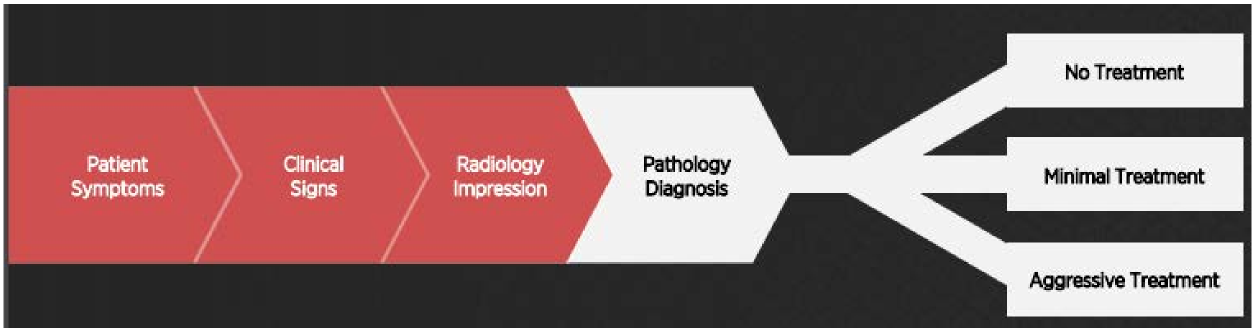


Figure 1: Infographic of pathology process

3 Pathology

Pathology is extremely visual. Pathologists will look at tissues and stains to understand how cells are being impacted. There are often patterns that arise within these images such as certain stains reacting with cells (see Figure 2.

Being able to observe these images in full detail is extremely time consuming. More often than not, pathologists have to select small areas of the image, using their intuition, that are more likely to have the disease present in order to increase the number of patients they are able to see in a day.

Another issue is that pathologists often disagree between each other and sometimes even among themselves! [SLJJGE17] showed that only about 48% of the time do pathologists give the same interpretation as other pathologists and 53% of the time with their own previous diagnosis for atypia. This lack of concordance was also observed in [BMJ17] with pathologist interpretation of melanocytic neoplasms on skin biopsies (see Figure 3).

It is only natural to see why people would be trying to use computers to automate/improve on this task.

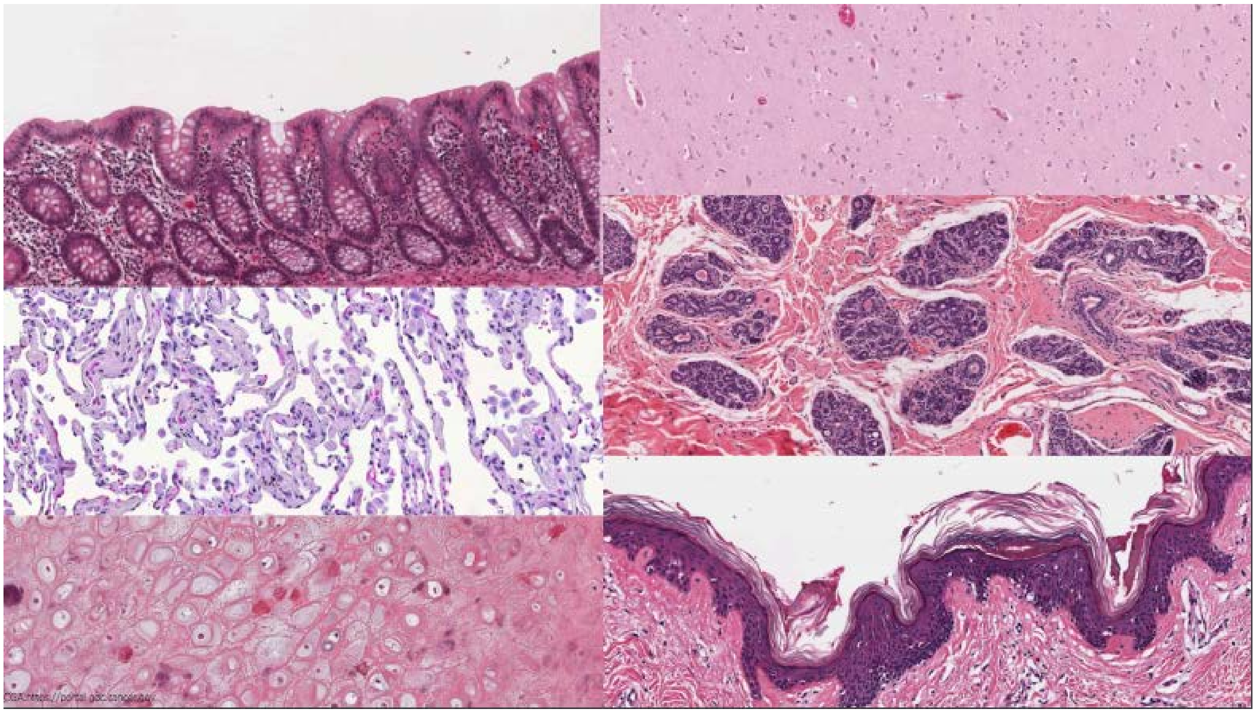
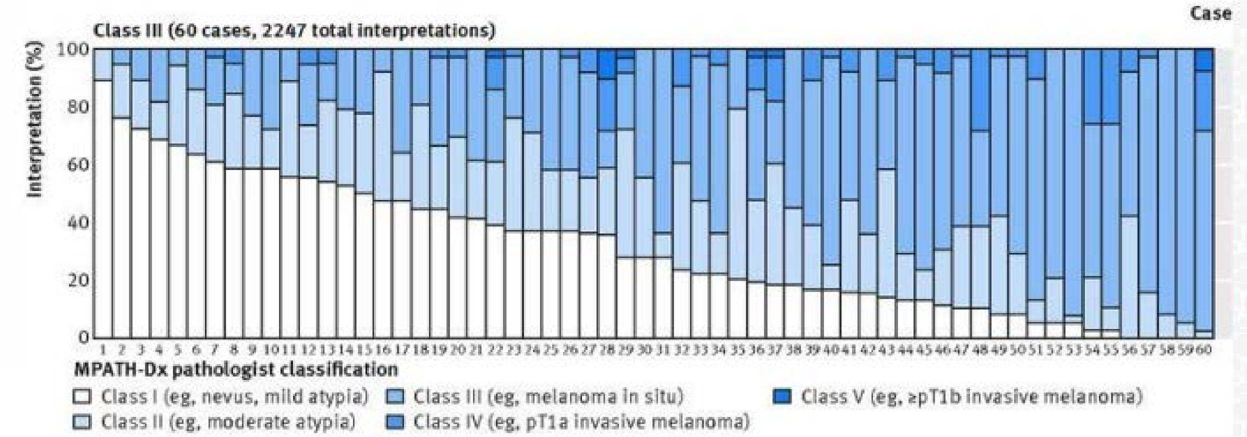


Figure 2: Example tissues that a pathologist would look at to determine if a disease is present.



Courtesy of Elmore, et al. Used under CC BY-NC.

Figure 3: Graphic describing the percent of pathologists that predicted a particular class of carcinoma by case.

[DF81] is one of the first attempts, which in 1981 used more traditional machine learning models to predict the sizes of cancer cells and predict outcomes.

In the 1990s and 2000s more attempts were made but were ultimately limited by the lack of digital health records and compute power. It wasn't until [AHBK11] when more modern techniques were used to extract quantitative features (see Figure 4) and properly analyze the image with them.

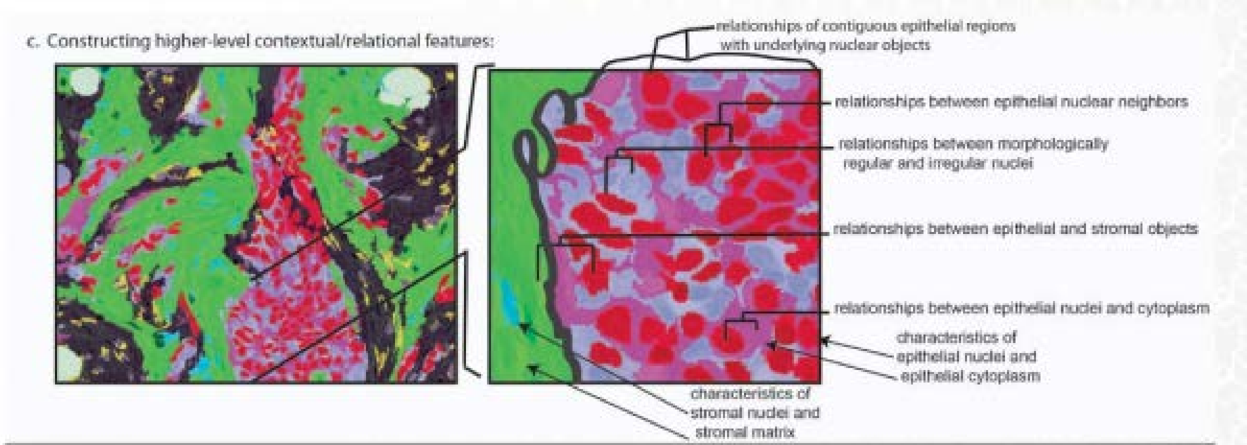

c. Constructing higher-level contextual/relational features:
relationships of contiguous epithelial regions with underlying nuclear objects
relationships between epithelial nuclear neighbors
relationships between morphologically regular and irregular nuclei
relationships between epithelial and stromal objects
relationships between epithelial nuclei and cytoplasm
characteristics of epithelial nuclei and epithelial cytoplasm
characteristics of stromal nuclei and stromal matrix

Figure 4: An explanation of features that were extracted from the image on the left

4 Computational Pathology

It is clear that the main task of pathology is image processing. The issue is how can we give computers the knowledge pathologists have to determine what cells are cancerous, the location of cancer cells, etc.? And furthermore, once we can get those features how do we use them to regress and predict proper treatment for these patients?

At PathAI there is a focus to get a computer to properly understand the image and then to use machine learning to understand the patient's health. It ultimately is a combination of modern techinques such as a Convolutional Neural Network (CNN) as well as more traditional models like logistic regression.

This gives us a few particular benefits: first, computers never get tired or distracted - this means that the entire image can be analyzed exhaustively. Secondly, the machine learning model is reproducible and able to extract more precise quantitative measures of things in the image. Thirdly, the model is efficient and can use parallelization to split the work up and get results faster. Finally, it is able to explore more than humans can and learn new relationships purely in a data driven way.

That being said, in no way can any model replace the job of a pathologist. Most (if not all) machine learning models are bad at reasoning through problems and should instead be used as a tool to assist the pathologist in identifying problem areas within some tissue sample.

5 Building the Model

The first consideration when building this model is the sheer size of images which can vary from 20,000 to 200,000 pixels per side making models that take in the full image nearly impossible to train. The resolution is so high in fact that it the image could be zoomed in and still get a good image (see Figure 5)

Instead, small patches of the full image can be extracted to be then passed into the model. This procedure allows for the full image to be explored thoroughly without having to pass in the image in its whole. These patches are initially sampled randomly throughout the entire image but can then be targeted to get more specific examples for areas in which the model is particularly weak (such as cancerous regions). At testing time, a similar approach is taken, getting patches from the test image; the results are then used to generate a heat map representing the network's prediction. (see Figure 6)

This heat map can then be used by a pathologist as a tool to focus their attention to areas of higher concern.

This model is extremely flexible and has a lot of potential applications. The following is just two examples of places where this technology can be/is being applied.

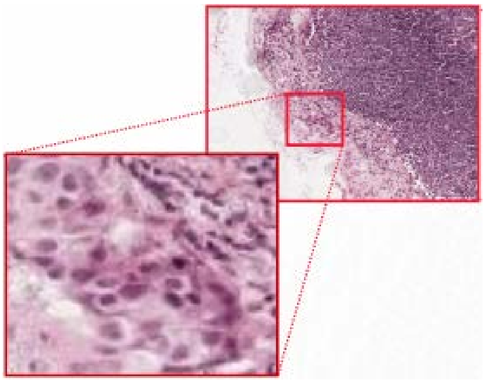


Figure 5: A small section of a tissue image that has been blown up to show the resolution.

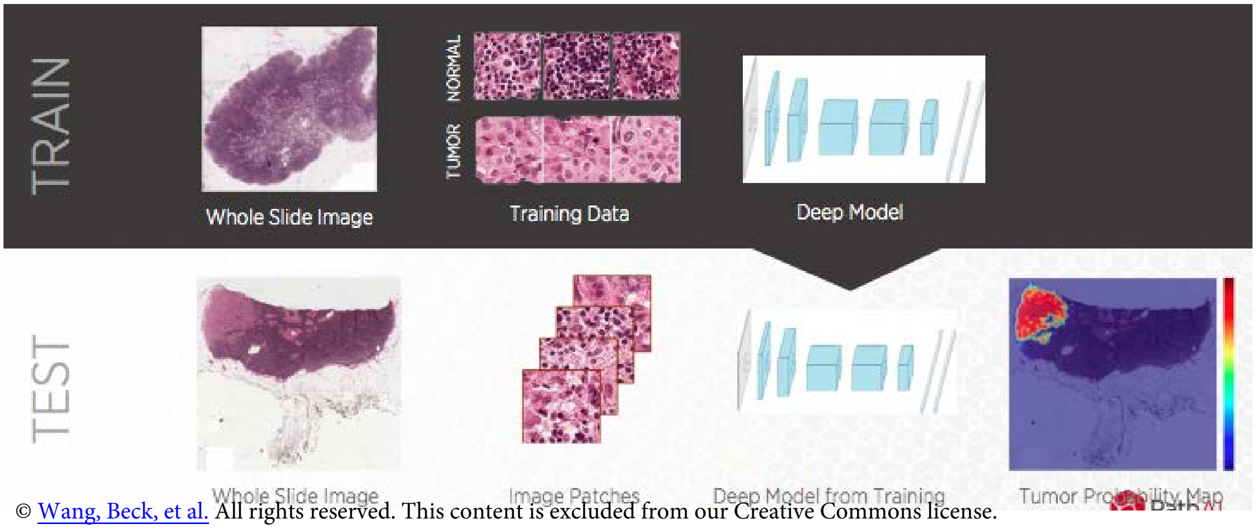


Figure 6: Example tissues that a pathologist would look at to determine if a disease is present.

5.1 Precision Immunotherapy

One specific example where this technology would be particularly useful is detecting PD-L1 in tissues. The presence of this ligand implies that a very specific drug can be used to battle cancerous cells in the immune system. Currently, humans are pretty bad at identifying when PD-L1 is present but with Computer Vision this problem is much easier to solve. (see Figure 7)

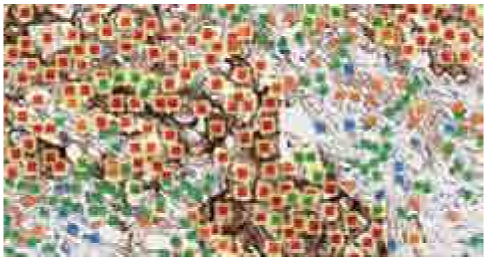


Figure 7: A tissue with PD-L1 patches highlighted.

5.2 Identifying Genetic Characteristics

Using a model with the same set up, we could identify specific cells, reactions, etc. that we are searching for - this gives us a tool to study the effects of certain drugs on the body. This helps then in identifying particular phenotypes associated with these drugs. We could then take these phenotypes and find which are most correlated with survival which would provide a way to start searching for genes that are relevant to survival.

6 Conclusion

- In the real world, 75% of the challenge in Machine Learning is building the right dataset.
- Modern Machine Learning is tons of engineeringand a little of empirical science.
- AI is great with large scale computer resources, digital data, and efficient algorithms. Without these factors, AI fails.

All images are taken from PathAI

References

[AHBK11] Samuel Leung Robert J. Marinelli Torsten O. Nielsen Marc J. van de Vijver Robert B. West1 Matt van de Rijn1 Andrew H. Beck, Ankur R. Sangoi and Daphne Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine*, 3(108):108ra113, 2011.

[BMJ17] BMJ. Pathologists diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. (357:j2813), 2017.

[BSOM$^+$14] David W Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7):1123–1131, 2014.

[DF81] Pennock JM et al. Doyle FH, Gore JC. Imaging of the brain by nuclear magnetic resonance. *Lancet*, 1(981):53–57, 1981.

[SLJJGE17] MPH Paul D. Frederick MPH MBA Margaret S. Pepe PhD Heidi D. Nelson MD MPH Donald L. Weaver MD Kimberly H. Allison MD Patricia A. Carney PhD Berta M. Geller EdD Anna N. A. Tosteson ScD Tracy Onega PhD Sara L. Jackson, MD and MPH Joann G. Elmore, MD. Diagnostic reproducibility: What happens when the same pathologist interprets the same breast biopsy specimen at two points in time? *Ann Surg Oncol.*, (5):1234 – 1241, 2017.

[WSW14] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94. ACM, 2014.