***Breast Cancer Detection and Prediction with Machine and Deep Learning***

**David Kinney  
Spring 2021  
https://github.com/dkinneyBU/breast-cancer**

**Any surprises from your domain from these data?**

I was surprised—or maybe shocked is a better word—regarding statistics around false negatives. According to the National Cancer Institute, screening mammograms miss about 20% of breast cancers that are present at the time of screening. I really find this alarming because my (perhaps naïve) interpretation of this is, one in five women who have been screened do in fact have breast cancer and do not know it when they leave the screening process.

On a more positive note, I was *pleasantly* surprised to discover the amazing work Dr. Constance Lehman, a Professor of Radiology at Harvard Medical School, the Director of Breast Imaging, and Co-Director of the Avon Comprehensive Breast Evaluation Center at the Massachusetts General Hospital, is doing in the field of Artificial Intelligence as applied to breast cancer prediction.

**The dataset is what you thought it was?**

I am using two datasets for this project; one is tabular, one is images. Both datasets were more or less what I expected. The tabular dataset (Breast Cancer Wisconsin (Diagnostic) dataset) consists of 33 measurements; there are no categorical variables, so I will not need to deal with any encoding. The images dataset consists of 162 whole mount slide images of breast cancer specimens scanned at 40x. From that, 277,524 patches of size 50 x 50 were extracted (198,738 negative and 78,786 positive). I had attended a conference session at an AI conference here in Atlanta last year, where the presenter walked through a Deep Learning example using a dataset similar to this. So, I generally knew what to expect.

**Have you had to adjust your approach or research questions?**

My approach has not changed and remains pretty solid. I have discovered a treasure trove of research on my chosen subject, I have secured to robust and reliable datasets, and model training is going smoothly.

I must admit, when I first present my research questions I really have not done a whole lot of digging into my topic (in all three cases my topics were not areas in which I am overly familiar). Consequently, they are mostly “best guesses”. After re-reading my questions just now, I am relieved to see they are actually quite relevant, even after becoming more knowledgeable in my chosen subject.

**Is your method working?**

Yes! I am performing prediction via Machine Learning, leveraging PyCaret, and detection via Deep Learning, leveraging the fastai library. Both models are showing quite promising results, with the detection model performing at a false negative rate of 9%, less than half of the statistic provided by the National Cancer institute.

**What challenges are you having?**

The only real challenge is the trade-off between time and “beefier” Deep Learning models. A ResNet34 model can take upwards of three hours to train on my home computer. I do have two GPUs, and fastai claims to provide functionality to utilize multiple GPUs. Unfortunately, I have not been successful in getting that to work (it consistently throws a memory location error). In light of that, I have taken the approach of training a fairly simple convolutional neural network (CNN) to establish a baseline. I then gradually add more layers and tweak the hyperparameters, resulting in longer training times, but better results (one advantage of working from home is I can let the models train while I am working. I use my home computer for both, and with three monitors, can work effectively while keeping an eye on the model’s progress).

Other than that, the project is progressing nicely with no major challenges.