Background

The most common cancer in the U.S. is female breast cancer. All cancers can be broken down into smaller categories called *cancer subtypes*. The most common subtype of breast cancer is *Invasive Ductal Carcinoma* (IDC), which is when the cancer spreads from its origination in the breast ducts and invades other breast tissue.

The purpose of this project will be to classify breast cancer images as whether they are in IDC or not, which is a common task of pathologists.

Motivations for choosing this topic include:

- image classification lends itself well to using convolutional neural networks
- I am currently creating a Cancer Data Visualizer for the GW Cancer Center, and I wanted to apply a Deep Learning method to related data

Data

The data that will be used include 277,524 images of breast cancer cells, each of which has a 50x50 resolution. 198,738 of them are classified as non-IDC and the other 78,786 are classified as IDC. They can be found on Kaggle: https://www.kaggle.com/paultimothymooney/breast-histopathology-images#10253_idx5_x1001_y1001_class0.png.

Method

First, the data is definitely large enough to warrant using a neural network. In Exam 1, we used a Multi-Layer Perceptron to classify images. Now, I would like to see how model performance will improve with

using a CNN, especially without having to resize images into vectors beforehand. I may have to create more augmented images from the IDC training images to keep a balanced training set.

Second, I came across a similar Github project that used Keras for this purpose: https://www.kaggle.com/vbookshelf/part-1-breast-cancer-analyzer-web-app/notebook.

I would like to see how differently the results will look by using PyTorch instead.

Lastly, if time permits, I would like to create a simple web app using Shiny (an R package) that allows you to upload one of the test images and get a response from my model. The above project did something similar with their Keras model using JavaScript.