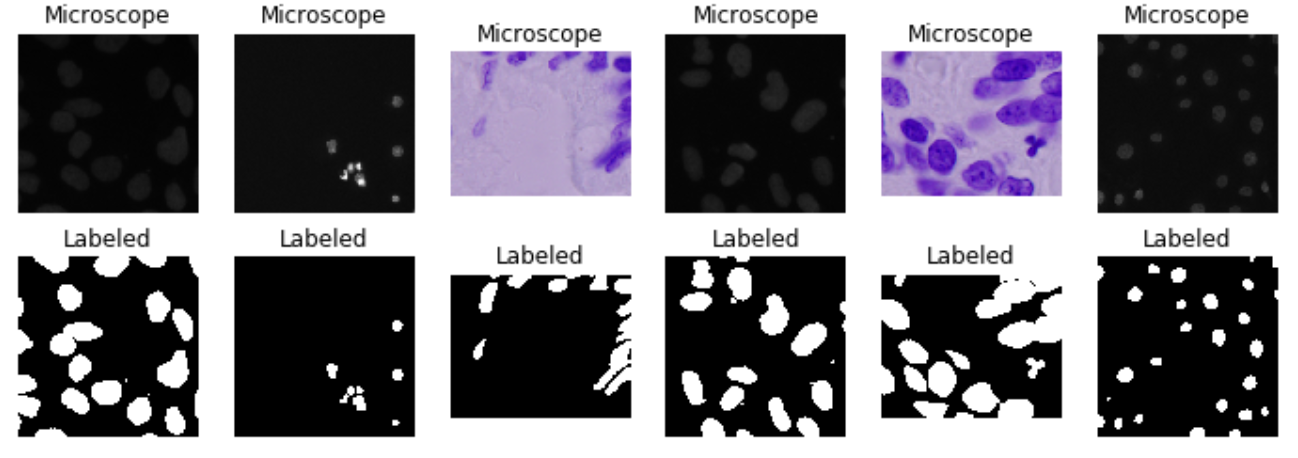
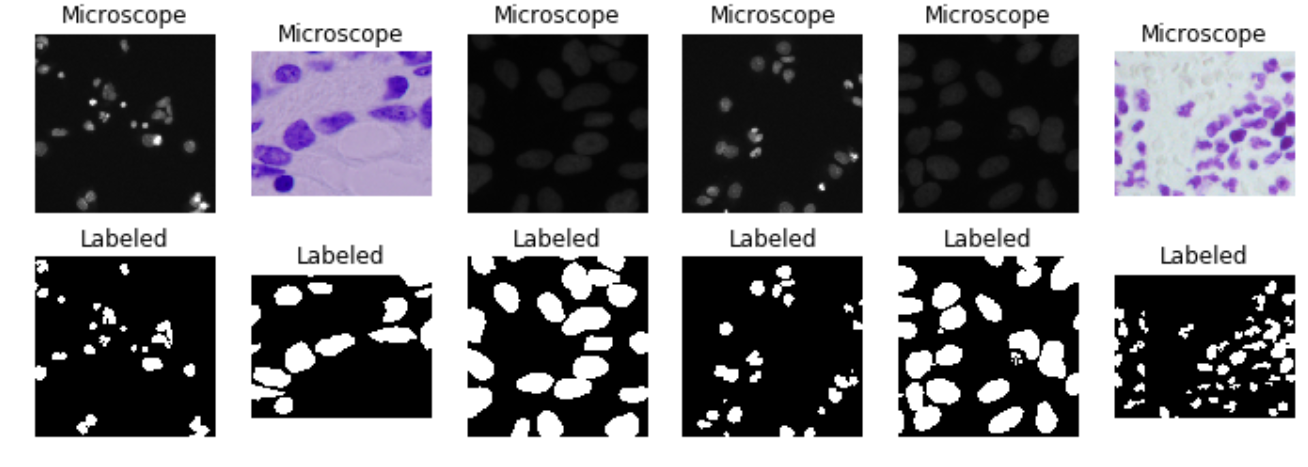
To start on the Kaggle challenge, we wanted to look at the data and understand the problem better. Thus, we wanted to sample both the training data, the masks associated with the images of the training data, and the test data that we are given. We wanted to first visualize these images to get a sense of what the images look like and then look at the pixel intensity graphs for the images, both overall and in specific image cases.





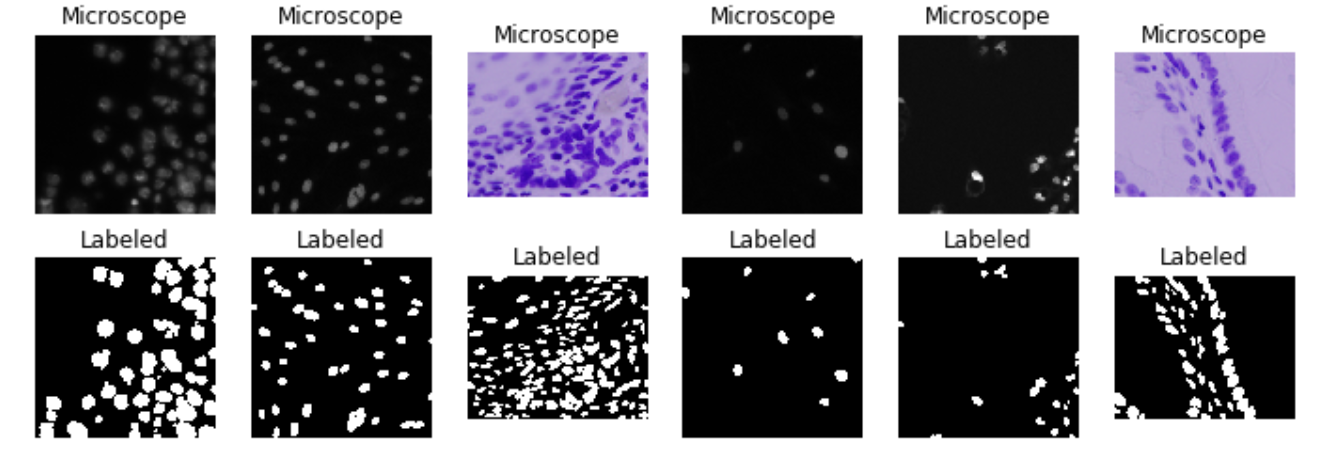
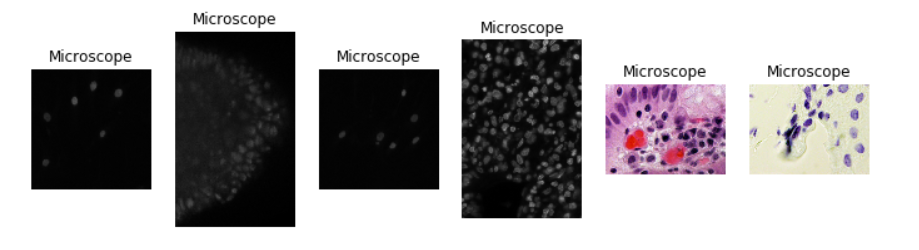


Figure 1: Random Sampling of the Training Data

The training data seems to be separated into two types of images: Purple images from H&E staining, as well as faint gray nuclei and a number of brighter white nuclei. The shapes are typically circular, but can be elongated and more ovular. In addition, some of these cell nuclei look extremely close together than the masks blend in to one another. I am a bit curious to see the pixel distribution over the two types of images as well as overall. Another thing to notice is that the pixel numbers of these images are not the same (they can be different sizes) and that might pose a challenge for comparisons.

In addition to looking at the training data, I looked at the testing data. After random sampling of the testing data, I noticed that there seems to be images taken in modalities not present in the training data. Here are some random samples from the testing data.



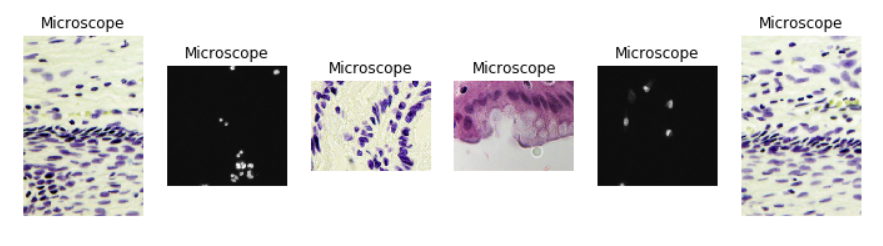
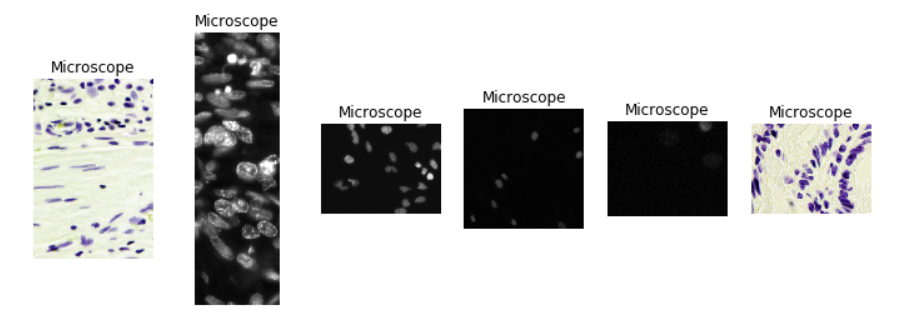


Figure 2: Random Sampling of the Test Data

These random samples show some of the same imaging modalities as the training data. For example, the same dark gray images are present. An example of this type of these types of images is shown below.

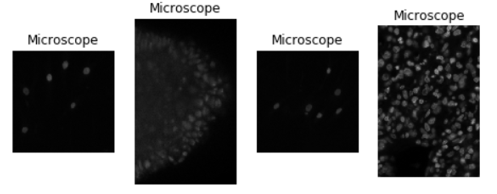


Figure 3: Test Data Examples that Look like Training Data

There does seem to be a large variety of different images that are present in the dataset not present in the training data. In particular, these images seem much more colorful than the original training data. This could pose challenges for techniques that just look at grayscale to identify nuclei especially if we have to filter out the nuclei by color (from my understanding, the nuclei are solely the dark purple portions of the images).

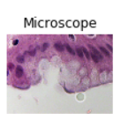
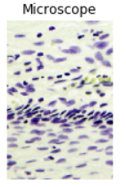
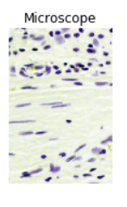
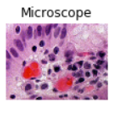


Figure 4: Test Data Examples That Looks Different From Training Data

To further explore the data, we looked at the pixel intensities of the images. I wanted to get the overall pixel intensities in all 3 colors as well as grayscale overall and then in specific image cases. I think the pixel intensities look very different from image to image (depending on the imaging modality utilized to take the image) and might be interesting to look at for analysis and determining what kind of model we want to use.

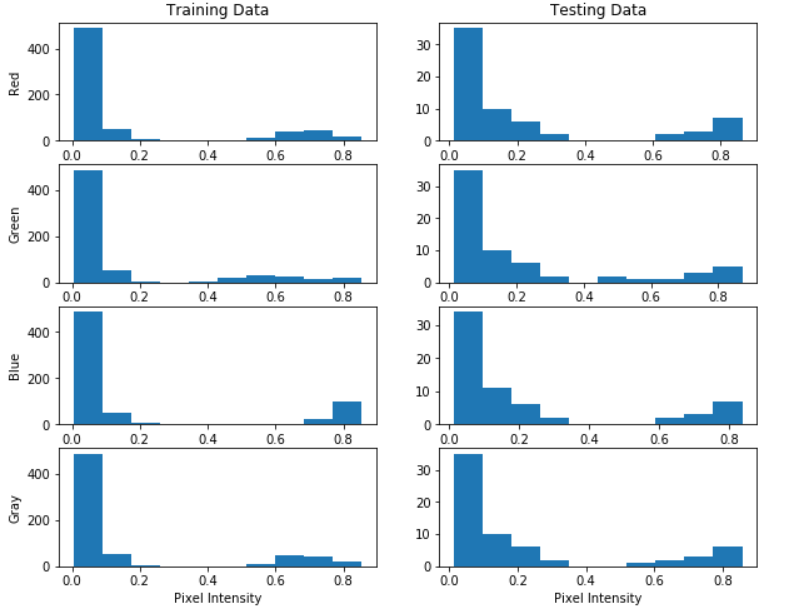


Figure 5: Overall Color Intensity of Both the Training and Testing Data

The testing data seems to have a higher proportion of images with a larger variety of colors. There are many more with pixel intensities that are greater than 0 within the testing data set than the training data. There also appears to be a higher proportion of training data with high blue colors which I believe corresponds to the purple coloring of the nuclei in the H&E stains. We will look at specific images and color intensities between the purple nuclei and the dark gray nuclei next.