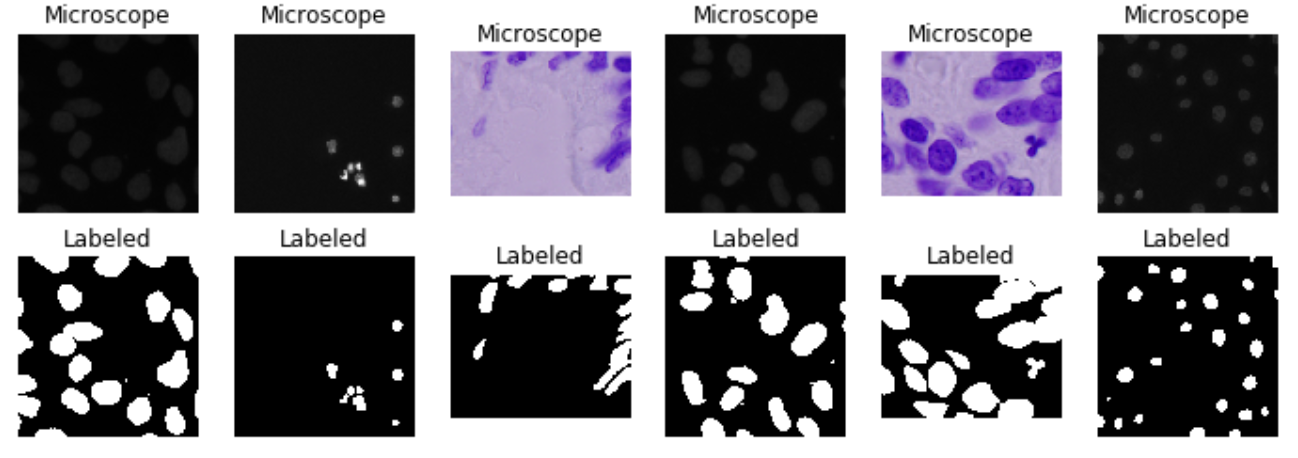
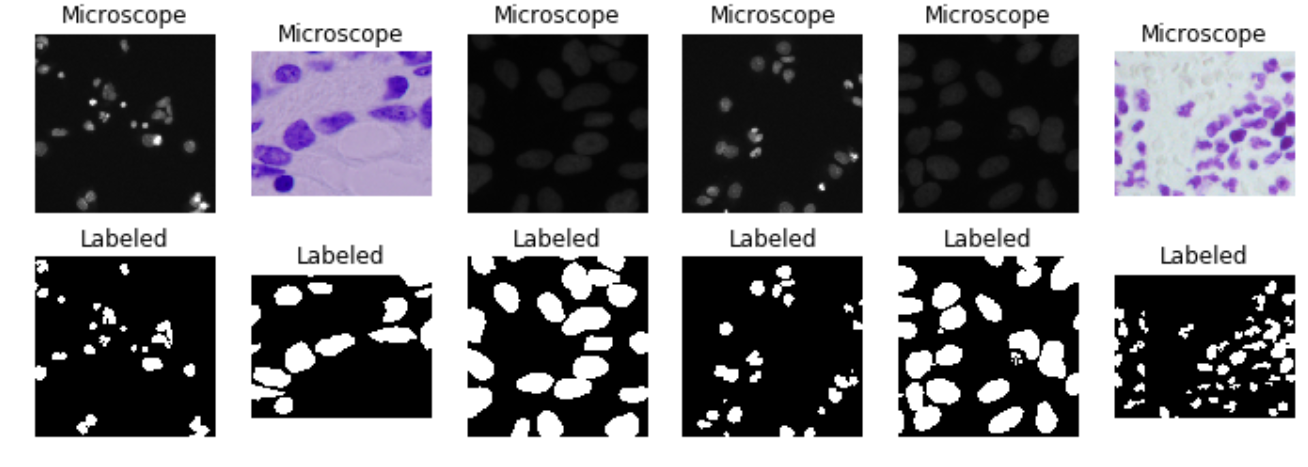
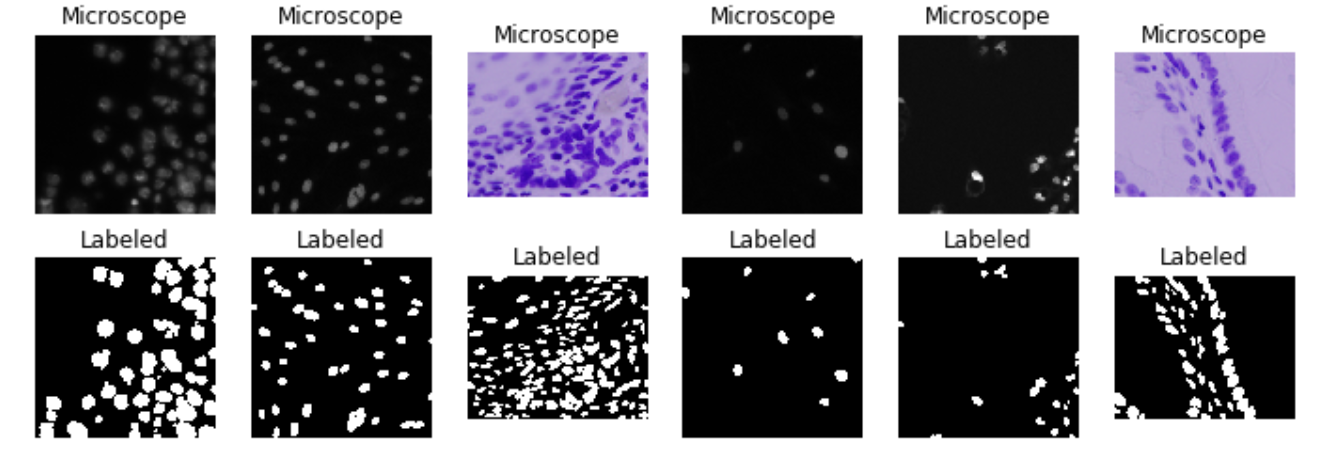
To start on the Kaggle challenge, we wanted to look at the data and understand the problem better. Thus, we wanted to sample both the training data, the masks associated with the images of the training data, and the test data that we are given. We wanted to first visualize these images to get a sense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



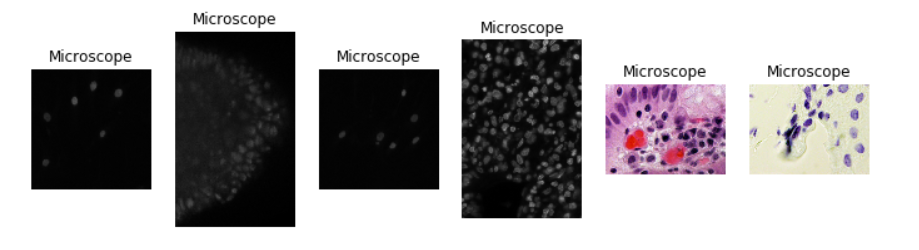


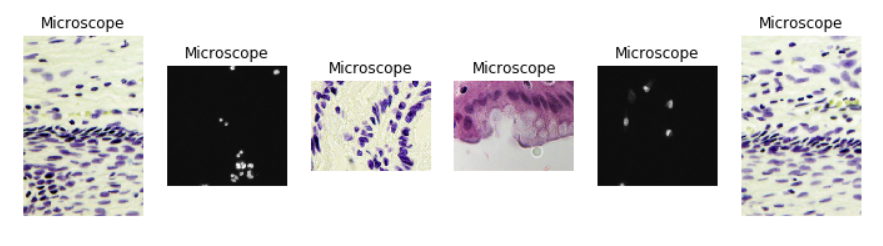
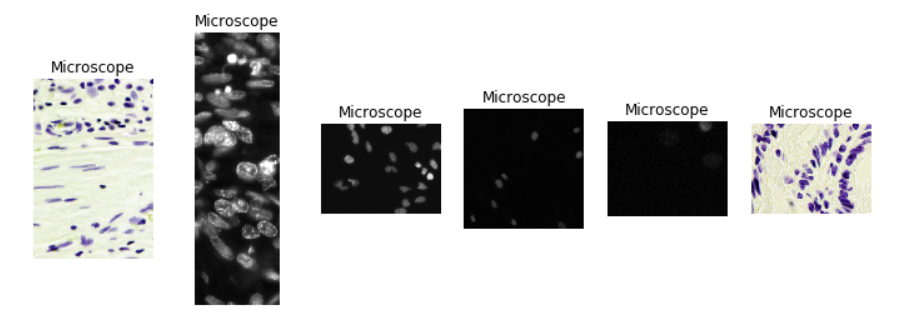


**Figure 1:** Random Sampling of the Training Data

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.

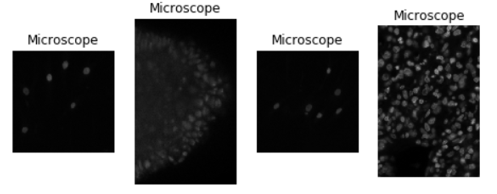
In addition to looking at the training data, I looked at the testing data. After random sampling of the testing data, I noticed that there seems to be images taken in modalities not present in the training data. Here are some random samples from the testing data.





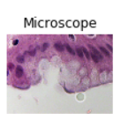
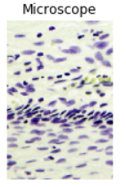
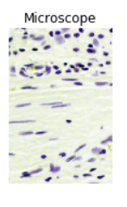
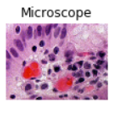
**Figure 2:** Random Sampling of the Test Data

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.



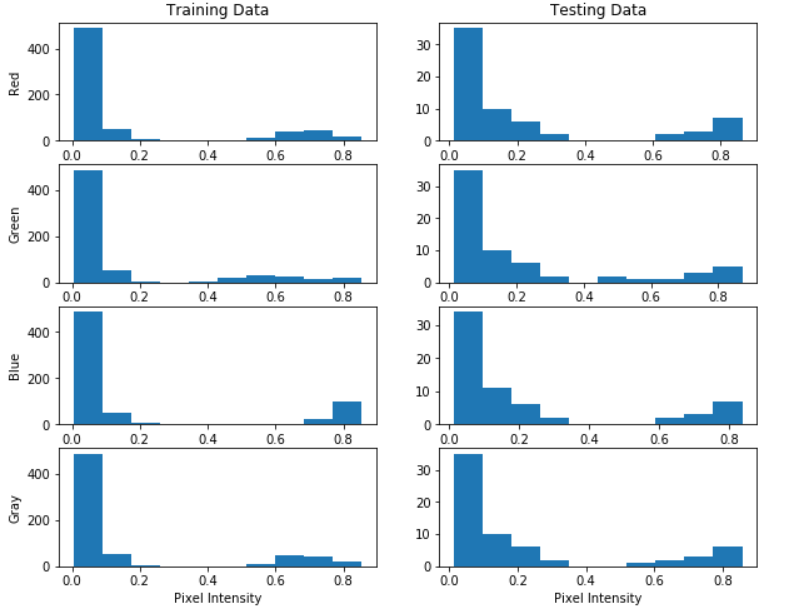
**Figure 3:** Test Data Examples that Look like Training Data

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).



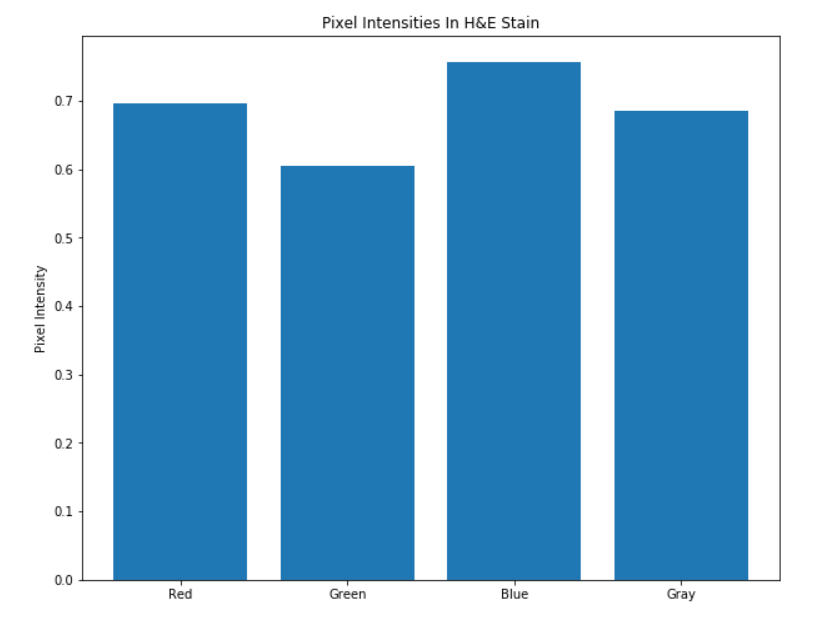
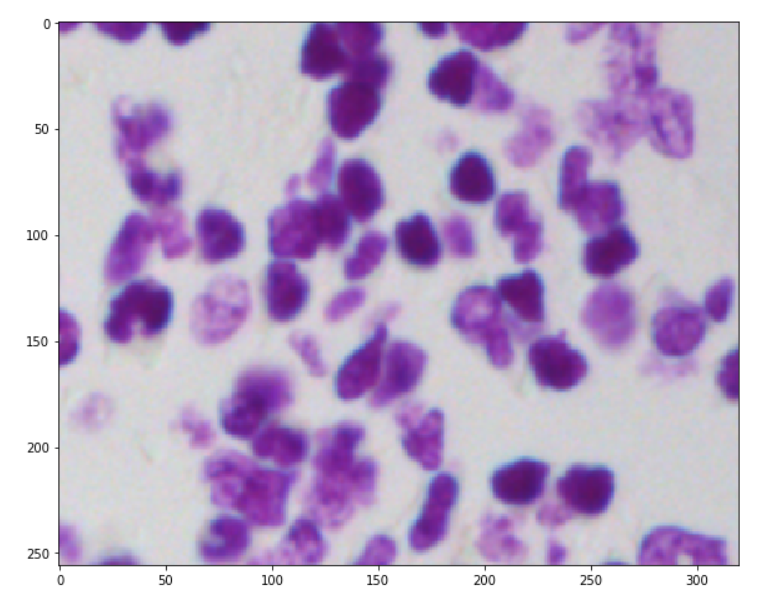
**Figure 4:** Test Data Examples That Looks Different From Training Data

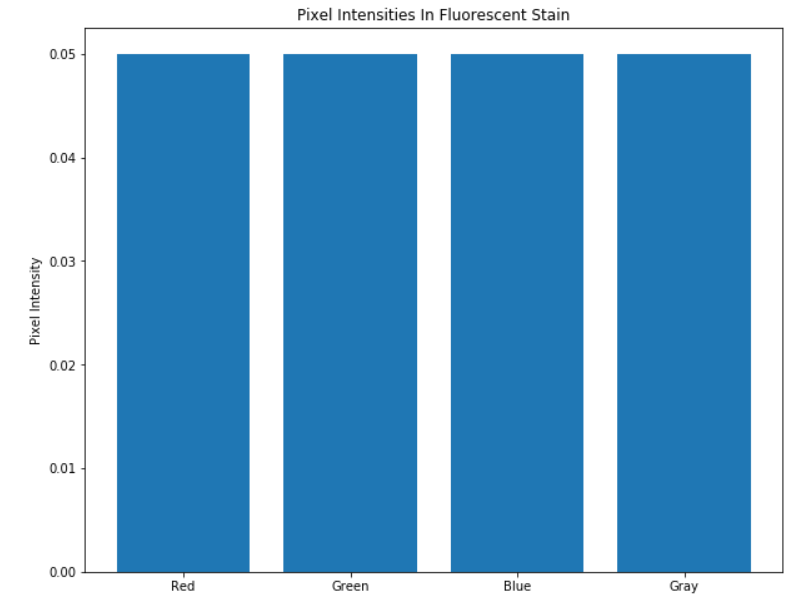
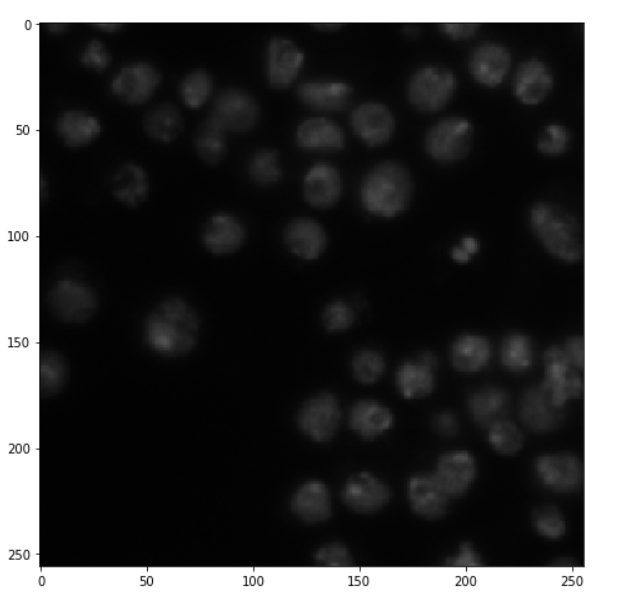
To further explore the data, we looked at the pixel intensities of the images. I wanted to get the overall pixel intensities in all 3 colors as well as grayscale overall and then in specific image cases. I think the pixel intensities look very different from image to image (depending on the imaging modality 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

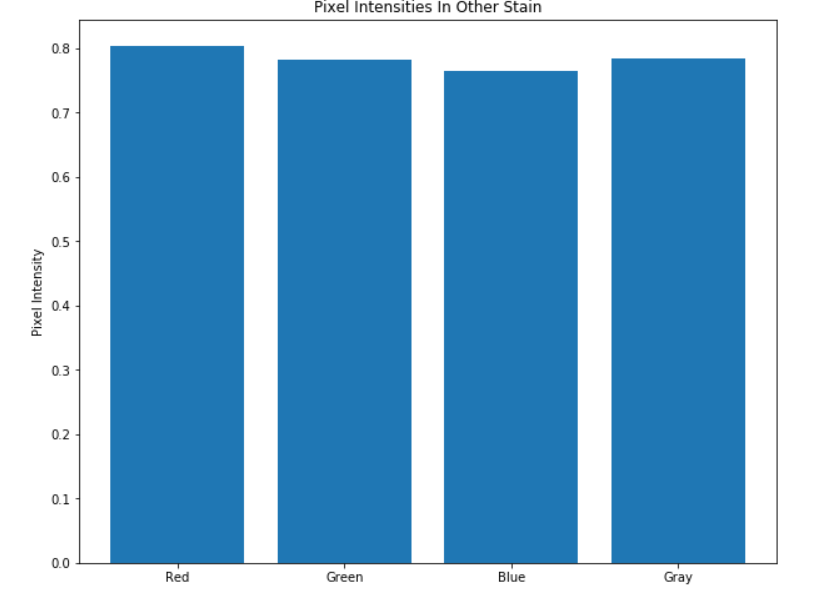
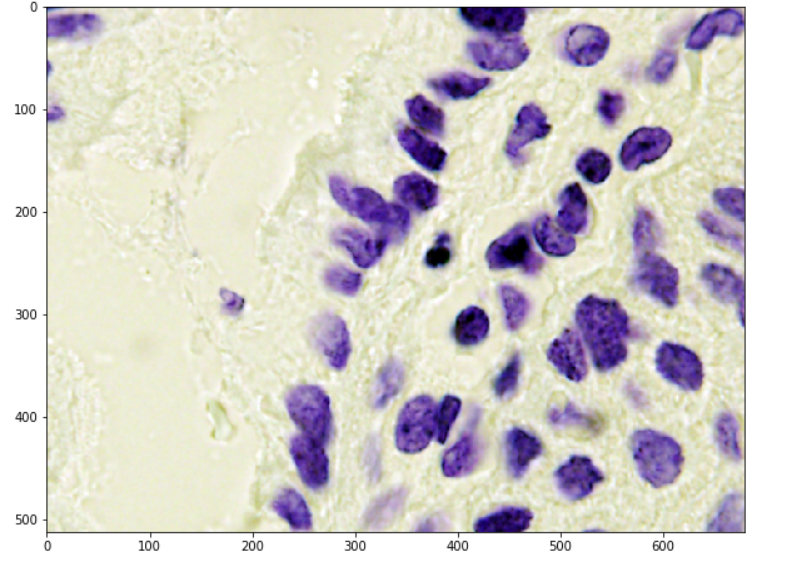
The testing data seems to have a higher proportion of images with a larger variety of colors. There are many more with pixel intensities that are greater than 0 within the testing data set than the training data. There also appears to be a higher proportion of training data with high blue colors which 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.

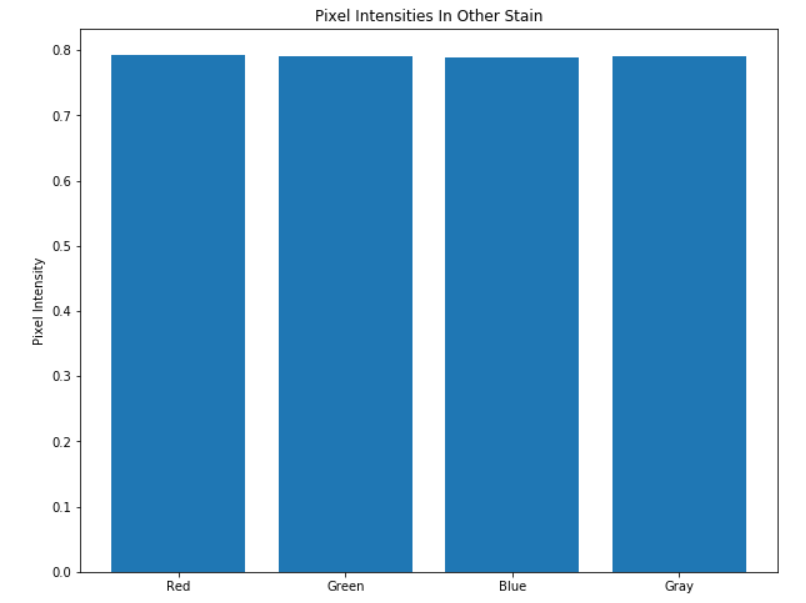
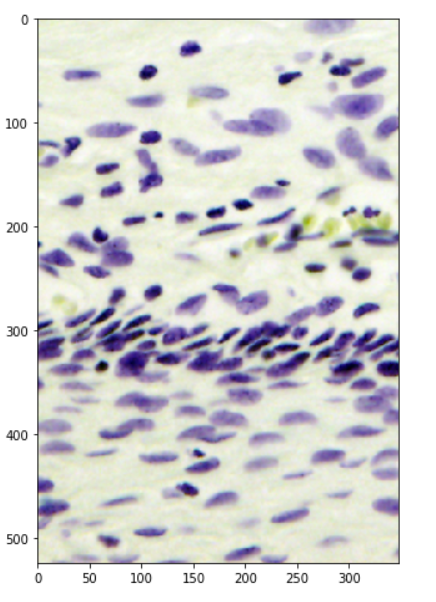




**Figure 6:** Comparing Pixel Intensities Selected Training Data Samples

The pixel intensities look different in these images due to the differing colors present in each. The fluorescent microscopy image on the bottom seems to be very gray while the top image looks more blue. In addition, the differences in background color may pose a problem with segmentation.





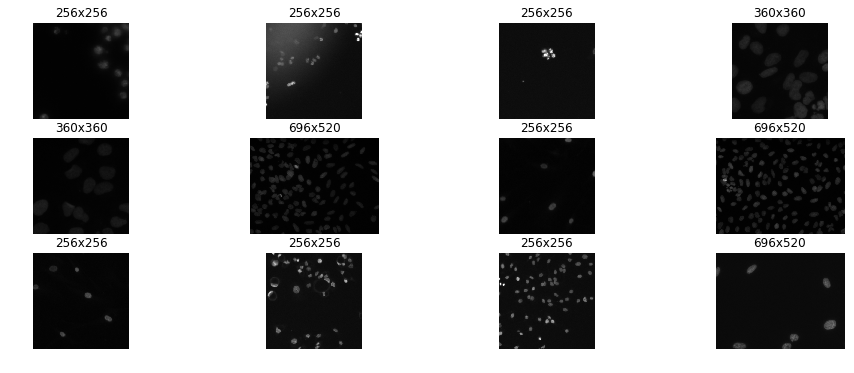
**Figure 7:** Pixel Intensities of Other Types of Staining In the Testing Data

These pictures look very different from anything in the training data and what we would expect to be the most challenging to classify. Although the bottom image also shows dark purple nuclei, it also shows yellow cells not present before. The top picture shows cells similar to the training data but displaying different pixel intensities. The mean pixel intensities of the top image look similar to pictures that look vastly different which could cause some issues when attempting to identify nuclei. Finally, perhaps attempting to make a homogenous background for these images would be helpful to prevent any mix-ups between pixel intensity and the location of a cell (although current computer vision techniques typically can pick out the background from the image utilizing thresholding).

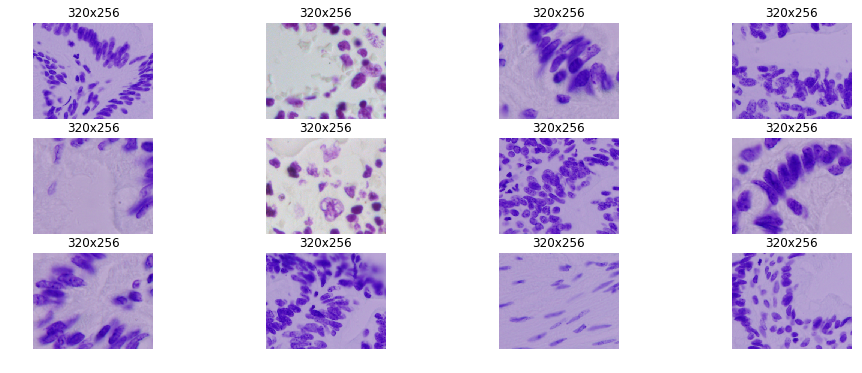
Realizing the existence of multiple image types in the dataset, we did a cluster analysis of the training set to group images based on their image modalities. Here, we used the Kmeans method described by kernel “Stage1 EDA: Microscope image types clustering” (<https://www.kaggle.com/mpware/stage1-eda-microscope-image-types-clustering)>. The clustering result shows that the training set can be broken down into three categories, as shown in the table below:

|  |  |  |
| --- | --- | --- |
| **Image Modality** | **Number of Images** | **Percentage in the Training Set** |
| Fluorescent | 546 | 81.49% |
| Histological | 108 | 16.12% |
| Bright-field | 16 | 2.39% |

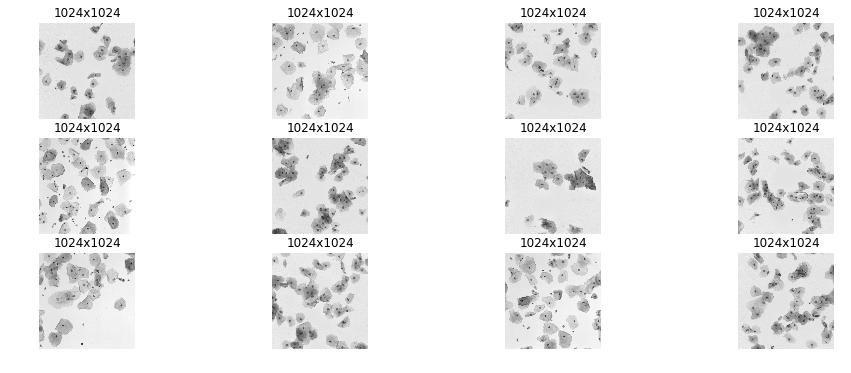
Some examples of the three types of images are shown below.



**Figure 8:** Examples of Fluorescent Images in the Training Set

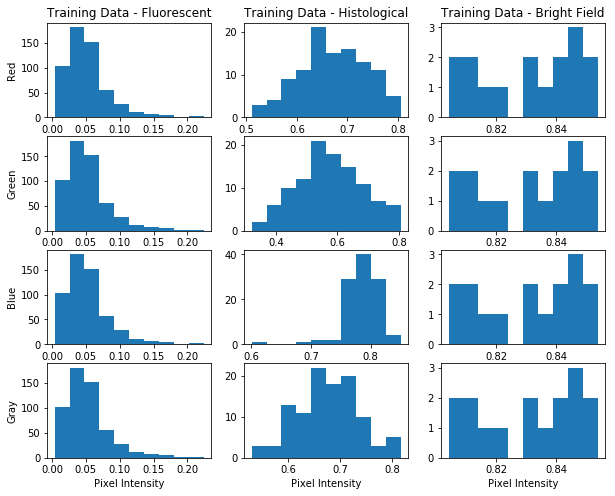


**Figure 9:** Examples of Histological Images in the Training Set



**Figure 10:** Examples of Bright-field Images in the Training Set

By intuition, different types of images have very different features, which will it challenging to develop a generalized model to identify nuclei. We were interested in accessing how different these three types of images are from each other in terms of some image metrics. If we can find some common features among all images, then these features might be useful for us to design a generalized predictive model. However, if the three types of images are significantly different from each other in terms of all metrics, then this evidence may suggest that a neural network is the best solution to solve this problem. Alternatively, in such scenario, we may build a model which first determines the most possible type of an image and then applies specialized strategy to make the prediction. Motivated by this idea, we checked the RGB value and intensity distribution of the three types of images.



**Figure 11:** Overall Color Intensity of the Three Types of Images in the Training Set

The result shows that the three types of images are significantly different in terms of the four features we accessed (red, green, blue, and grayscale intensity). From Figure 11, we can see that histological images have different pixel intensity distributions for the four features, while the fluorescent images and bright-field images have very similar distributions for the features. This result is consistent with the intuition because in the training set, only histological images are colorful. However, by our previous inspection, we found that some bright-field images in the test set are colorful too. This finding supports our previous argument that the image modalities in the training set and the test set are not exactly the same.

Moreover, the three types of images have distinct patterns of distributions, which may be used as their “signatures”. The fluorescent images have relatively low average intensity. The intensity distribution is skewed to the left with a peak at around 0.05. The histological images have normal distributions for all four types of intensities. The means of red, green, and grayscale intensities are in the range of 0.6 to 0.7, but the mean of blue intensity distribution is approximately 0.8. The bright-field images have relatively even frequency at all pixel intensities, except for intensity values around 0.83.

In conclusion, the current findings do not show any common features among the three image types. This may suggest that a neural network may be a better option than a feature-based machine learning model for nuclei detection in general microscopy images. Or we may use these features to identify the type of an image before predicting the location of nuclei.

For next steps, we will explore more image features for the three types of images in the training set and see if we can find any similarities among them. Meanwhile, we will also work more on neural net models for nuclei detection. Moreover, to better understand the difference between the training set images and the test set images, we will do cluster analysis on the test set and explore the same features of these images.