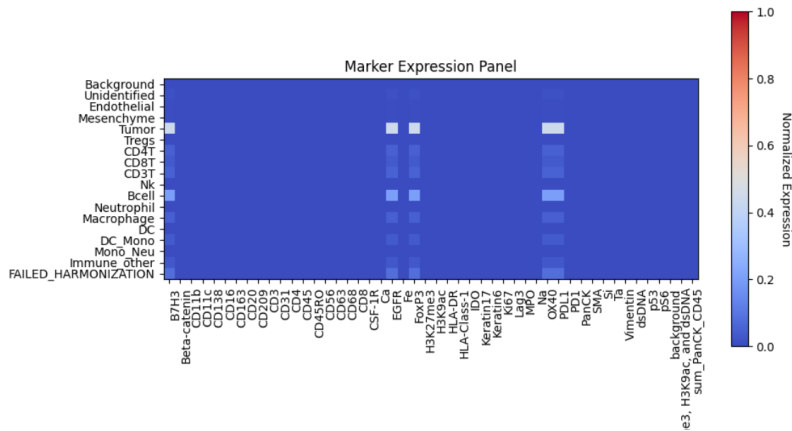


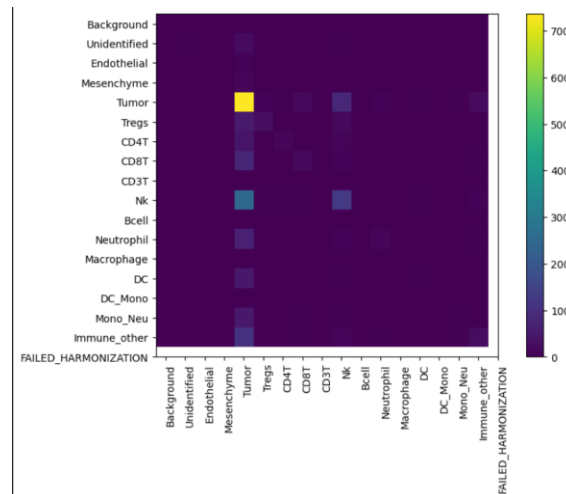
BEBI 205 First Set Report

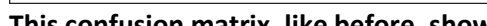
1. Marker Expression Panel



- a. This marker expression panel looks not great, but after testing it seems this is due to a data issue. I was not able to get the compute server for the class working properly and did everything locally. As such, I needed to severely limit the amount of data (~2000) images, not due to hardware constraints but due to jupyter issues on my local machine. The code is setup to do more images, and I have left a comment for the small changes needed for that.

2. Confusion Matrix



- a. 
- b. This confusion matrix, like before, shows a need for more data that I was not able to accommodate. The code is setup to handle as much data as needed, I just was not able to provide it. As we can see, the model is very good at distinguishing tumors, which is likely due to the data being skewed towards tumor cells. This is also missing some cells due to the train/test split removing them due to lack of data, again, but the code should be equipped to handle more data and that issue would go away.

3. <https://github.com/emesic23/bebi205>
4. It took a very long time to figure out how to efficiently process the data. My first issue was that normalization was taking a long time, and eventually I figured out an ok speed method. After that, I spent another very long time figuring out cropping the cells. This is mainly due to a misunderstanding of the nature of the data, where the X is a single massive image with the y

being a mask that we can use to cut out cells. I had done things with different assumptions and kept running into walls until I figured out exactly what the data meant. After figuring out the cropping method, another issue I had was that my environment would crash when I tried using more than ~2000 cells, which led to issues with my marker expression panel as certain cell types were much more prominently featured than others. This led to downstream issues in training, as I was only able to train off of 2000 images for my CNN, and given the uneven balance of cell types, this also led to the confusion matrix looking off. Although, the model did seem to recognize tumors quite well along with a few other cell types.

Most of my ML for image experience has been with object detection given a dataset of several images with a clear label or set of bounding boxes for each image, so this was a novel experience of having to create my own images from a segmentation mask.