The Break Through Tech AI program helps participants gain skills in machine learning, data science, and AI that prepares them for working on a team to solve industry challenges. Being able to work on real world problems allows us to gain hands-on experience and learn more about the industry. My team and I had the opportunity to work with Memorial Sloan Kettering on a project related to image processing.

Memorial Sloan Kettering Cancer Center is one of the best cancer hospitals in the U.S.. We were tasked to accurately segment nuclei and separate overlapping and touching cells in H&E- stained tissue sections using deep learning. H&E staining is often the gold standard in histopathology and provides pathologists a detailed view of the tissue. Manual separation of cells is a long and tedious process, so developing a deep learning model that results in more accurate images will take less time. Accurately segmented nuclei will provide important insights to pathologists such as physical characteristics of nuclei and spatial distribution. This is important to gaining further insight into cell features and functionality.

For preprocessing we decided to use stain normalization and tiling on the images. Then apply this data to a U-Net model for semantic augmentation with a dice score coefficient. When training the model we started with converting the images to grayscale which the best accuracy score was around 20%. Then we added a validation set which was 10% of the training data and the best accuracy score was around 40%. We also decided to apply data augmentation which increases the amount of data needed to train robust AI models. This was on the fly data augmentation which was applied to both images and masks which resulted in the best accuracy score of around 60%. Before getting the final results for the U-Net model there was a mask and image tile error, some of them were not lined up that had to be fixed. Once completed the final accuracy score from the model was around 75%.

The dataset that we used was the MoNuSeg dataset which was open source. This dataset was obtained by carefully annotating tissue images of several patients with tumors from different organs that were diagnosed from multiple hospitals. Training data contained 30 images of size 1000x1000 and the test data contained 14 images of size 1000x1000. During our data preprocessing we broke down the images to size 256x256 to infuse more information into the dataset.

Some insights we found were that training with the RGB images provided more information and improved the model. Also when deciding between the Dice Score and Jaccard Index, they are similar metrics but we found that the Dice Score provides more nuance in comparing pixel similarity. Some obstacles we ran into were ensuring the dimensions of the input images for the U-Net were matched up. Using on the fly augmentation removed that issue. We also found that using virtual machines on Google Cloud allowed us to run our U-Net model for more epochs to get a more accurate model. We also prepared a second model which was Mask RCNN that would have been used for instance segmentation. The results from this model would have provided a mask that shows each distinct nuclei cell. The mrcnn package for the Mask RCNN model was only compatible with a certain version of TensorFlow that didn't work with any of the machines we had. If we had more time and potential next steps of the project we would try to use PyTorch to create the Mask RCNN model. To further improve our current U-Net model we would train the model more to try and receive a better accuracy score and get more training data, so the model can be more versatile.

Overall I enjoyed working on this project with my team and learned a lot from it. Deep learning, computer vision, and image processing techniques were all new to us coming into this project, so there was a lot to learn in the 3 months we worked together. Thank you to MSKCC for allowing us to take on this challenge and our advisor Alex Hollingsworth for helping and supporting us along the way.