The Kaggle 2018 Data Science Bowl presented by Booz Allen Hamilton poses the challenge of finding nuclei in various imaging modalities utilizing a single machine learning model. The challenge primarily focuses on image diversity, with many images present in the testing data sets that are not present within the training data sets. Therefore, this challenge focuses on make a generalizable model that will be able to identify nuclei across many imaging modalities, including modalities that the model has never seen before. The motivation for this challenge is to help speed up the discovery of cures for numerous diseases such as cancer and Alzheimer’s disease by identifying these nuclei. By automating nuclei identification, researchers will be able to identify cell densities, cell reactions to different treatments, and understand the underlying mechanisms behind how these diseases work. Finally, these results can be utilized to increase the speed that drugs take to develop and come to market.

To tackle the difficult challenge of creating a model to fit this dataset, we first performed exploratory data analysis. We took a look at the data and noticed that there were various imaging modalities that were used to take the images in the dataset. Some of the images in the training dataset were taken utilizing brightfield microscopy while others were taken using fluorescent microscopy. Further adding to the challenge, the testing dataset that we received in stage 1 had even more imaging modalities. Some images did not look at all like the images in the training dataset. Therefore, it is important to create a model that is able to generalize well and not overfit the data. This key concept guided our model building throughout the entire competition.

To start out the competition, we utilized k-nearest neighbors on the images to simply predict for nuclei. This approach did not utilize the training data and did not perform very well on the dataset. Although the nuclei seem to be different colors, the different backgrounds and different colored nuclei make this model generalize poorly as the color intensities can be quite similar if the background intensity and nucleus intensity are similar. In addition, the approach did not utilize the training data which we believed was important.

In addition to simple k-nearest neighbors modeling, we also performed basic computer vision techniques to attempt to separate out the nuclei. Some classic image segmentation techniques that are utilized in problems such as these include binarizing the image, performing median filtering for noise reduction, and obtaining masks utilizing thresholding. As the images should show a binomial distribution in the pixel intensities (one for the background intensities and one for the nuclei), utilizing Otsu’s threshold to separate out the nuclei seemed like a good choice. Although this worked in separating out some nuclei, just choosing the smaller pixel group seemed to give poor accuracy on where the nuclei were and therefore only scored about 0.128 on the testing data. This method fails to utilize the training data to predict on the testing data as well as will have a hard time differentiating an image that is mostly nuclei with little background.

Moving forward with other types of modeling, we looked at simple other models such as logistic regression and SVMs. We found that these models had a hard time generalizing over the different imaging modalities as the thresholding for logistic regression and SVMs would be different for each type of imaging modality. Once again, it would be hard for these types of models to generalize over the different imaging modalities due to the differing pixel concentrations contained in each type of image. The inclusion of different imaging modalities in the testing data especially proves problematic for these types of models as these models typically predict for what they have seen before. In this case, the new imaging modalities present data that has never been seen before and looks completely different from the original data and therefore generalize poorly in this competition.

After looking for a model that would generalize better, we decided to try our hand at neural networks. First, we implemented the fully connected U-Net network for biomedical image segmentation due to its high accuracy with small datasets starting from a kernel on Kaggle. (https://www.kaggle.com/keegil/keras-u-net-starter-lb-0-277?scriptVersionId=2164855) To this end, we finally started getting scores of arounds 0.28 accuracy on the testing data for stage 1. Proceeding forward, we attempted to optimize the U-Net by changing the architecture of the network but saw minimal improvements if any at all. Changing the padding and the convolution size did not seem to help increase the accuracy of the model. Changing the number of layers also did not seem to improve the model accuracy very much. One fix that did see a noticeable increase in testing accuracy was to improve the Keras metric utilized to measure if the model was performing well or not. This step increased our prediction accuracy on the stage 1 testing data to 0.311. (https://github.com/kamalkraj/DATA-SCIENCE-BOWL-2018) After we reached this score, we believed that we hit the peak for U-Net and fully connected neural networks without pre or post-processing of the data. Although fully connected neural networks can perform well on some datasets, we wanted to look into different neural networks that focused on regional differences rather than the entire image. Fully connected neural networks like U-Net may put weights on too much background especially if there are no nuclei present in the image and therefore may generalize poorly. To this end, we wanted to consider the local region around nuclei in neural networks such as convolutional neural networks.

To start working on convolutional neural networks, we chose a recent convolutional neural network called Mask RCNN to perform image segmentation. In particular, we chose the Matterport Mask RCNN to perform image segmentation on the nuclei. (https://github.com/matterport/Mask\_RCNN) Although this RCNN seems to perform poorly on fine details, we believe that it will generalize better and therefore have a better score than U-Net. After brief training on the Mask RCNN, we achieved a preliminary score of 0.28 but this was only after 5 epochs. Further training will hopefully increase the score that we get and translate better to the second dataset.

Our final approach to this problem is to utilize Mask RCNN to perform image segmentation on the problem of identifying nuclei in images. A big overarching theme in this competition is being able to apply the techniques utilized to new imaging modalities. Therefore, we see new imaging modalities whose images look extremely different in the stage 1 training and stage 1 testing set. In addition, we looked at the stage 2 testing data and this dataset includes even more images that are not included in the training data. This dataset includes false positives where there are no nuclei in the image as well as textbook scanned pictures of cells. This combined with some basic image segmentation utilizing watershed segmentation to help separate out nuclei in the training data will improve the score. Finally, we will perform ensembling (particularly bagging) on the results from the neural network to improve accuracy in our final model. Ensembling our model results with different training weights will help improve our generalizability especially on images that have never been seen before. These ensembling methods are primarily utilized to prevent overfitting of the model to the training data and therefore improving our generalizability. This improvement in generalizability should improve our score on the second testing set as avoiding overfitting to the training data while predicting accurately seems to be the largest issue in this competition.

(Insert images of examples above)

* What lessons did you learn about solving prediction problems using machine learning and large data sets?