The notebooks used for data gathering, cleaning, preprocessing, and model training can be found in the “Notebooks” folder.

## Preprocessing notes

The raw data came as a zip file which unzipped into two folders and a CSV file. The folders each contained the testing and training datasets as DICOM files, and the CSV file contained a table of ICH labels for the training data. My “create\_labels” notebook formats the CSV file into a manageable format suitable for binary classification and saves into a new CSV file.

The “cluster\_ct\_scans” and “cluster\_ct\_scans\_2” notebooks will each create a CSV file that (1) contains lists of DICOM image IDs, where each list represents DICOM images belonging to the same CT scan; (2) contains rows of DICOM image IDs and their (x,y,z) coordinates detailing which layer in the scan that they represent (this is necessary to order the images from low to high and find surrounding slices).

The “convert\_dicoms” notebook will take each 512x512 DICOM image, apply the brain, subdural, and bone window filters (these filters accentuate different parts of the scan, such as bone or soft tissue, making them more visible), and saves the image as a 512x512x3 RBG PNG scaled between 0 and 255. These PNG files are then used for training.

## Training notes

All 3 model experiments are included in the notebooks “model\_experiment\_1”, “model\_experiment\_2”, and “model\_experiment\_3”. Model Experiment 2 has no comments as it was deemed a failure partially into training, as its precision and recall remained below 10% for much of training. Comments on the performance for Model Experiments 1 & 3 can be found in their respective notebooks.

I was unable to automate and tune many models as the rubric requests, because of the high training time and computational cost that is inherent to my data. However, I believe I was able to gain some decent insight into the best model architecture to use in my final project.