Team:
Rahul Bazaz
Howard Fan
Maulik Shah
Yuqing Zhang

Project: Data Science Bowl 2017 - Lung Cancer Detection Using CT Scan Images

Description:
Home page: https://www.kaggle.com/c/data-science-bowl-2017
Lung cancer affects millions of people every year. Early detection is critical to give patients the best chance at recovery and survival. In this project, we will use thousands of CT lung scan images from patients at high cancer risk, to develop a lung cancer detection algorithm. The algorithm is expected to determine from features regarding the lesions in the lungs, whether or not the patient will be diagnosed with lung cancer within one year of the date the scan was taken. This will reduce the false positive rate that plagues the current detection technology, get patients earlier access to life-saving interventions, and give radiologists more time to spend with their patients.

Data:
Images are in DICOM format. Each image contains a series with multiple axial slices of the chest cavity. Each image has a variable number of 2D slices, which can vary based on the machine taking the scan and patient. The DICOM files have a header that contains the necessary information about the patient id, as well as scan parameters such as the slice thickness. The images in this dataset come from many sources and will vary in quality. The algorithm is expected to perform well across a range of image quality.

General aims and approaches:

- Preprocess the data (following available tutorials: https://www.kaggle.com/gzuidhof/data-science-bowl-2017/full-preprocessing-tutorial/notebook) and clean out the noises
- (Optional) Feature selection/integration or dimensionality reduction to select candidate features
    - Current ideas: Lasso, PCA, http://content.iospress.com/articles/intelligent-data-analysis/ida795
- Build a classifier using convolutional neural network (and the selected features)
- Evaluate and report performance on the test set and validation set once the solution is available.

If time allows, also:

- Search for and implement frontier classification algorithms in medical image processing or deep learning classification methods, to improve our model
- Make use of publicly available datasets (https://www.kaggle.com/c/data-science-bowl-2017/discussion/27666) to improve model performance