Predicting Lung Cancer Onset Using Segmentation and Classification

Trivedi, Neelay

Lung cancer claims 150,000 Americans yearly, yet only 15.7% are diagnosed pre-metastasis. Current computer diagnostics focus on Bayesian networks using rule-based prediction. In this project, I developed a regular classification algorithm to identify potential lung nodules from CT scans and to classify whether they would be cancerous in a 12-month timespan. I was working with data from the Lung Image Database Consortium, and I used scans from 90 patients. I preprocessed the balanced data (45/45 positive/negative ratio) by applying provided masks indicating the location of lung tissue. To further isolate lung tissue, I used a Hounsfield filter with a 395-405 HU range and normalized slice number for each patient to 256. I created a random training/testing set of 65 scans/25 scans. Next, I used U-net segmentation, a small dataset-optimized convolutional neural network, to predict nodule masks and extract a feature set from cancerous slices during ground truth comparison. I designed six U-Net models with varied hyperparameters including learning rate, optimizer, activation function, and number of layers with a Dice coefficient objective function. Model 6 achieved the highest Dice(0.64) using a 0.001 learning rate, 9 layers, Adam optimizer, and Leaky ReLU activation. I then used Model 6's feature set to train a classifier that determined whether unknown nodules were cancerous. Using XGBoost classification, I varied learning rate, optimizer, activation function, and decay, using a logarithmic loss objective function. The most effective XGBoost classifier achieved 0.59 loss and a 76%/72% accuracy on the training/testing sets. Future work will focus on using stratified k-fold cross validation to minimize sample bias and using grid search to improve hyperparameter optimization.

Awards Won:
Fourth Award of $500