Alex Dunn, Breanna Nery
June 9, 2017

CS 188 Final Report

Executive Summary

The objective of this project was to survey machine learning methods for prostate cancer (PCa) classification. Using sets of 256x256 images (T2 MRI, DWI, ADC) from 12 patients, various machine learning algorithms including ensemble methods, deep learning, and support vector machines were used to classify individual pixels in the prostate with clinician-verified masks used as the ground truth. The Scikit-learn gradient boosting classifier ensemble model, trained and tested on conglomerate sets of T2 MRI, Gabor-filtered T2 MRI ($f = 0.3$), and ADC images, proved to be the most successful "plug-and-play" machine learning model, with an overall leave-one-out cross-validated receiver operating characteristic area under curve score of 0.7408 ± 0.12, averaged over 20 trials.

Methods

We chose to use the Scikit-learn library (sklearn) in Python 2.7.13 for our implementation of the project. We also used Keras (with Tensorflow backend) as a base for building a deep-learning neural network for classification. The available input images were T2 MRI, five different weights of diffusion weighted images (DWI: 10, 100, 400, 800, 2000), apparent diffusion constant images (ADC), and Gabor-filtered versions of all the above. The available output images were gathered from clinician-verified 'masks', images clearly delineating which pixels constitute the prostate, cancer inside the prostate, and area outside the prostate. Of the original 62 patients, we only considered the images from 12 patients with high T2 contrast marking cancerous areas. Cancer was treated as binary classification in prostate areas (cancer=1, non-cancer=0).

Since classification of each patient's image must be done on a per-pixel basis, each pixel and the surrounding *n* pixels on each side of the pixel (making $2n + 1$ by $2n + 1$ subimages) were used as input data. This data was preprocessed into flattened feature vectors and was associated with the center pixel's matching mask value (the output). Only feature vectors containing input pixels which are all inside the corresponding patient mask prostate area are used for training and testing ("pruning"). Various sklearn models and a Keras 2-layer convolutional neural network, along with several combinations of input images, were tested and scored with leve-one-out cross validation according to the receiver operating characteristic (ROC) area under curve (AUC).

Results

The initial values for the ROC AUC score for 8 models using only T2 and mask data were calculated for $n = 3$. Given the quick training times, high scores, good scaling with increasing dimensionality (dimensionality is proportional to n^2), and generalizability of ensemble methods, 5 of the 8 models were ensemble methods. The classifiers were sklearn gradient boosting ensemble (AUC of 0.6670), extra trees (0.6271), bagging (0.6252), random forest (.6189), Adaboost (0.6151), support vector machine (0.5001), multilayer perceptron neural network (0.4998) and Keras convolutional neural network (0.4995).

Given the high default score of the GradientBoostingClassifier (GBC), we chose it as our model to refine. We then independently surveyed the GBC performance when the normalized, Gabor-filtered T2 image was added as an extra feature (to resolve 'texture' attributes), the results of which are shown in the blog post "Filtering the Images" for 10 frequencies between 0.1 and 0.9. Using a Gabor frequency of 0.3 resulted in the highest AUC score of 0.6731. Only normalized and filtered images were used as extra features, as non-normalized Gabor-filtered images reduce the AUC score significantly lower for all Gabor frequencies.

Next, we included different multiparametric datasets as extra features in order to select the best images for feature extraction. For these trials, we used $n = 5$ in order to maximally stratify the AUC scores using more information from various image sets. Diffusion weighted images of any weight level reduced the cross-validation score when added as extra features to the T2 MRI alone. Using a combination of T2, ADC, DWI (2000) and Gabor-filtered T2 ($f = 0.3$) resulted in a cross-validation score of 0.716. Using T2, ADC, and DWI alone resulted in 0.711; T2, DWI, and filtered alone resulted in 0.642. The best result came from using T2, ADC, and filtered images alone, with a cross-validated AUC score of 0.724.

Various n were then tested to select the highest scoring subimage size. The graphical results of the analysis are shown in Figure 1, where each trial was conducted 20 times for statistical accuracy. The highest mean score occurs at $n = 6$, corresponding to a 13x13 pixel subimage size.

The final model was chosen to be a sklearn GBC with 10 sub-estimators, a logistic-regression based loss function for optimization, and the Friedman mean squared error used as a tree split criterion.The inputs used were T2 MRI, ADC, and Gabor-filtered T2 MRI ($f = 0.3$) 13x13 subimages. The final 12-patient cross-validated ROC AUC score was 0.74 ± 0.12. An example evaluation of the model is shown below in Figure 2.

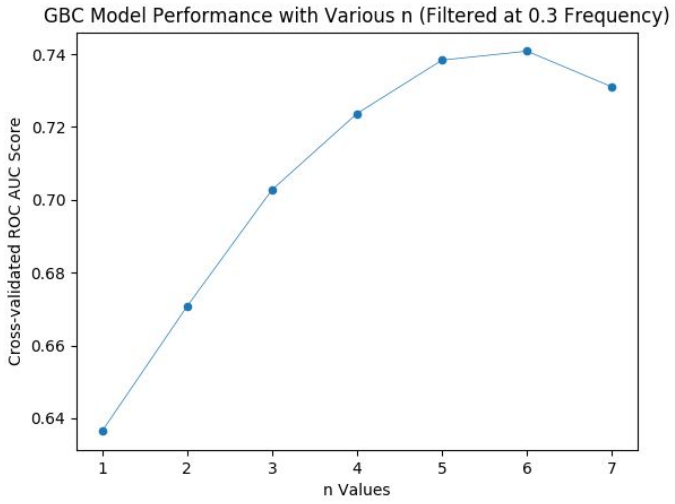

Figure 1: ROC AUC score vs. subimage dimension n.

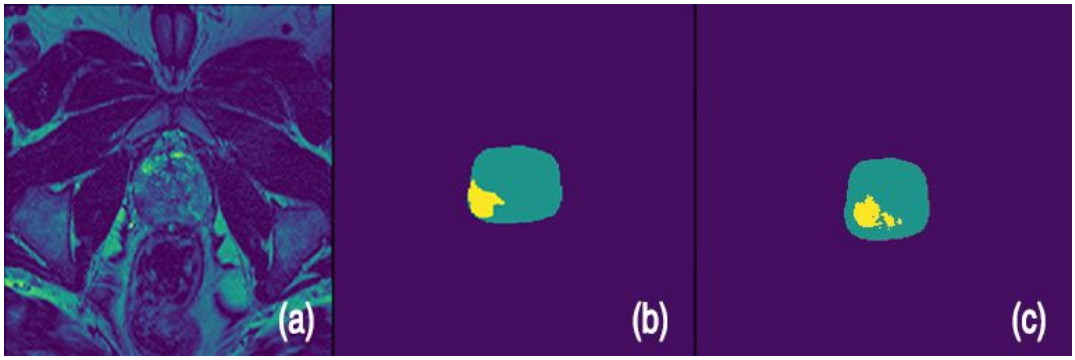

Figure 2: Patient 18 data. (a) T2 MRI image (b) Ground truth mask image. (c) GBC-predicted classification, with a score of 0.9202.

Discussion

This project investigated machine learning methods for off-the-shelf binary classification of PCa, and we observed fair ROC AUC performance from the sklearn gradient boosting classifier ensemble method. Unfortunately, while increasing subimage size increased the classification accuracy, it also decreased the available area for classification on each image. However, the end objective of machine learning models for PCa classification is often merely to *aid* clinicians in diagnosis, where clinician-reviewed masks may not always be available. Although the model described in this project has disadvantages, it may still prove useful in clinical settings since it has satisfactory predictive accuracy across the majority of the prostate.