Automatic Tumor Grade Classification From MRI

Kelvin Zhang
UID: 304569586
Email: k.zhang661@gmail.com

Shalini Dangi
UID: 304465145
Email: shalinidangi@ucla.edu

***Abstract*—In this paper, we discuss methods by which prostate cancer can be detected automatically from MRI images using machine learning techniques. Specifically, we explore the results of testing various ML models on T2-weighted images to classify cancerous regions.**

I. Introduction

Prostate cancer (PCa) is one of the most common types of cancer in men, and manifests in the form of malignant growths on the prostate gland. The current standard for PCa diagnosis is transrectal ultrasound biopsy, however manual physician diagnoses is subjective and difficult to standardize and test for errors. MRIs represent a faster, non-invasive alternative and existing research has been directed at automated, software based solutions to prostate cancer detection that are faster, more scalable, and less prone to error than human processing methods. This project will focus on applying machine learning techniques to T2-Weighted MRIs, one of the main parameters in multiparametric MRI analysis, in order to detect the presence and location of PCa growths.

II. Methods

The data set used for this project consisted of T2-weighted MRI images and corresponding mask images showing cancerous, prostate, and non-prostate regions for 62 different patients, each with varying levels of cancerous tissue (Figure 1). Masks in this data set were manually contoured by a clinician and serve to mark the cancerous pixels shown in the MRI image. In order to provide our models with sufficient classification data, a feature vector was extracted from each pixel in each T2 image. Given a particular pixel, its associated feature vector consists of the pixel's immediate neighbors within an N-pixel radius (Figure 2). Note that for our most successful results, feature vectors centered around non-prostate regions were removed before training our models. Once extracted, the labeled feature vectors were then used to train supervised learning models provided by SciKit-Learn, which were then tested using leave-one-out and k-fold cross validation.

This project explored the following models: random forest classification, AdaBoost (adaptive boosting) classification, gradient boosting classification, k-nearest neighbors classification, decision tree classification, and Multilayer Perceptron (neural network) classification. Additionally, experimental adjustments that we explored in pursuit of optimal performance included feature vector size variation, feature vector pruning, selective pruning of patients, prediction methods after model fitting (binary vs. probabilistic), and threshold variation.

Feature vectors were varied by radius N (Figure 2) to determine the vector size yielding the best classification results. After careful testing of N ranging from N = 2 to N = 20, the most successful vector size resulting in the highest model evaluation scores was determined to fall within the range of N = 4 to N = 6. In addition to varying vector size, feature vectors were also pruned to create a more balanced, relevant data set as mentioned earlier. This was attempted in three different ways: 1) feature vectors associated with all 0s in the mask (represented a region completely outside of the prostate) were discarded, 2) feature vectors associated with at least one 0 in the mask (contained some region outside of the prostate) were discarded, and 3) feature vectors with a center mask value of 0 (centered around a non-prostate region) were discarded. Ultimately, the third pruning attempt was found to be the most successful and was used in the testing of our models.

Next, after viewing the T2-weighted images and corresponding mask sets for each patient, we discovered large variations in clarity and contrast within the images. As a result, we attempted to selectively prune patients from the set in order to produce a more uniform data set. However, after training models on this narrowed data set, the resulting scores showed little difference from previous runs utilizing all 62 patients. This is perhaps a result of manual pruning without a clearly defined threshold for selection. Thus, this experimental adjustment was discarded before the training and testing of our models. Additionally, another experimental adjustment we tested dealt with the predicted matrix after fitting. After fitting the model to the test set, we obtained both the predicted probability matrix and the predicted binary matrix. The former was more successful as it indicated the probabilities of a positive/negative prediction for each pixel according to the model, and used this information to plot the receiver operating characteristic (ROC) and precision-recall (PR) curves. Finally, we also experimented with increasing the threshold value of the predicted probability matrix to determine if that would

increase the accuracy of our results. This was found to be inconclusive, and the threshold was left at 0.5.

III. Evaluation

The main metrics we used to evaluate the effectiveness of our models were ROC-AUC (area under the receiver operating characteristic curve) and PR-AUC (area under the precision recall curve). We selected ROC-AUC since we have an extremely unbalanced data set where there are many more non-cancerous pixels compared to cancerous ones, and being insensitive to class balance would yield reliable measurements regardless of patient or feature vector pruning. Also, since our goal is to be able to compare different learning models, ROC-AUC is useful because it considers all operating points/thresholds instead of focusing on a particular one.

PR-AUC was used because unlike the ROC curve, it also takes into account the fraction of true positives among positive predictions. This means that it depends on the frequency of the positive class, in our case cancerous pixels, which could give us a better representation of the differences in our algorithms given that our positive class is not only much rarer, but also much more interesting than the negative case. For our application, the true negative does not offer much value and we would much rather extract more information about the recall of the models. Failing to detect a tumor would be significantly more disastrous than mistaking healthy tissue as cancerous. Additionally, the precision recall trade-off would be essential in selecting the final model for actual application as it should prioritize high recall and the ability to detect all of the cancerous regions, but also demonstrate the compromise of precision so as to not waste expensive and labor-intensive medical resources.

Our models were tested using cross-validation. To perform the actual cross validation of our models, we took two difference approaches. We first used leave-one-out (LOOCV) using a single patient as the test subject. However, this led to large variance in our diagnostic statistics between patients and was thus, not as useful as a measure for the success of our models. For example, our random forest classifier would vary in the range of 0.4 to 0.7, even between separate iterations of the cross validation. We then decided to opt for k-fold cross validation since it mitigates this by creating rotating testing sets of size k. k = 10 was selected since we had a pronounced class imbalance and wanted to increase the number of repetitions in the sampling. This was done despite it being slightly more pessimistically biased in terms of performance results.

IV. Results

From this project, we were able to make multiple conclusions about the success of the different models that we trained and tested. ROC-AUC varied dramatically across models using LOOCV, with some underperforming near 0.5 and the gradient boosting classifier at N = 6 achieving the highest score of 0.8404. Using k-fold cross validation, the gradient boosting classifier at N = 5 achieved the highest average ROC-AUC of 0.5805. Following that, the random forest classifier and multi-layer perceptron classifier averaged a 0.5347 ROC-AUC and 0.5388 ROC-AUC across k-fold cross validation, respectively. The rest of the classifiers were mostly random in performance, and although a classifier with ROC-AUC of 0.5 is not necessarily useless, we determined that it was more probable that our model was a weak fit, rather than an underlying random process behind the feature vectors of cancerous tissue.

PR-AUC tended to fluctuate between specific images as expected since it is affected by class balance. We considered precision recall performance by taking the difference between PR-AUC and baseline precision of positive labels over total labels. In our LOOCV testing, this was maximized by the gradient boosting classfier at N = 4 with a score of 0.589 and an improvement over the baseline of 0.411. The k-fold validation was maximized by the gradient boosting classifier at N = 4 with a score of 0.413 and an improvement over the baseline of 0.241. This was easily the strongest result as the rest of the classifiers all had deltas of less than 0.2, and results between iterations were inconsistent within our computing time. This represents a significant improvement in average precision.

In addition to diagnostic statistics, we were also interested in finding the ideal threshold for this application in order to choose a particular model. In Figure 3, we can see the best performing receiver operating characteristic and precision recall curves. Purely looking at the ROC curve, averaged over N = 6, we found that the mode threshold value that maximized the difference between the true positive rate and the false positive rate was 0.2. However, as we discussed previously, this is likely not the best metric for our use case. Based on the importance of finding all of the positive cases, we should be prioritizing recall heavily, and thus, from the PR curve we can see that this can be obtained for a relatively linear trade-off in precision. Our ROC curve also encourages this as recall increases steeply with respect to the false positive rate, allowing us to achieve very high recall with small concessions.

V. Discussion

Ultimately, we observed the best ROC-AUC and PR-AUC performance from the SciKit-Learn Gradient Boosting Classifier (Figure 3). It had the overall best scores for both ROC-AUC and PR-AUC for both leave-one-out and k-fold cross validation. Additionally, it performed the best despite the feature vector size (N) that was chosen. Given more time to explore this project, extensions that might result in better ROC-AUC and PR-AUC scores might involve the following: 1) normalizing pixel values to adjust for discrepancies in quality among patients, 2) more intelligently pruning outliers with regard to patients with unclear data or feature vectors/mask outside of a relevant range of pixels, 3) incorporating different

types of MRI images including diffusion-weighted MRIs and dynamic contrast-enhanced MRIs, and 4) testing more deep learning models which are essential since they are able to learn adaptive image features and perform image classification at the same time.

The next logical step in our project would be to use our predictions to create fully segmented reconstructions of the prostate cancer in the tested patients. This would consequently allow for on-the-fly Gleason score classification and other medical diagnoses. Although we were unable to generalize strong results throughout all of the models we tested, based on the results we have gathered from this project, we hypothesize that with additional fine-tuning of the individual estimators, careful curation of the data, and further experimentation including the extensions listed previously, machine learning models have the potential to offer a very powerful approach to this medical imaging problem.

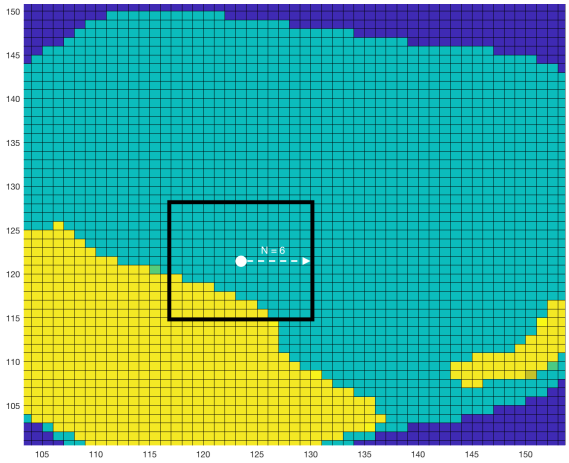

Fig. 2. Illustration of feature vector consisting of all of the pixels within the black box with the center pixel designated by a white circle. The size of the patch is determined by N. (In this example, N = 6.) Feature vectors are discarded if the center pixel corresponds to a 0 in the mask (marked by a dark blue color) as this indicates a region outside of the prostate.

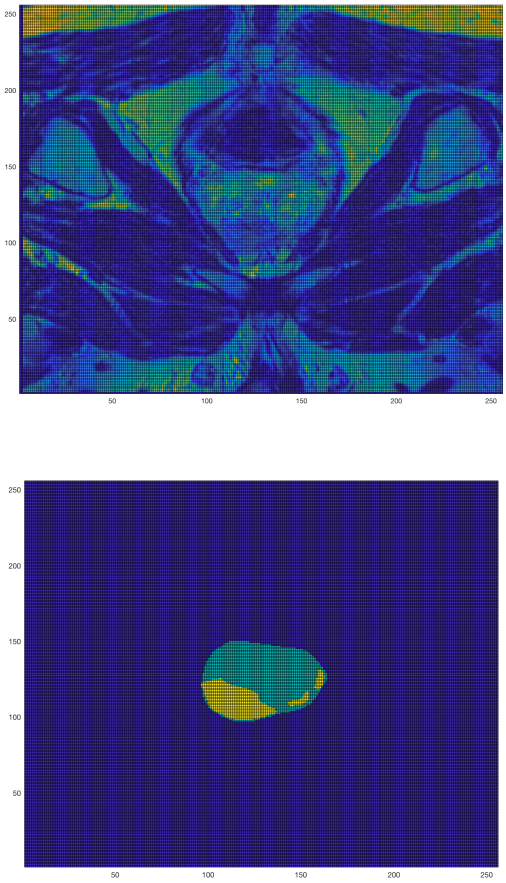

Fig. 1. Example T2-weighted image (top) and corresponding mask (bottom) for patient 1. Mask colors represent 0 - blue, 1 - green, 2 - yellow for non-prostate, prostate, and cancerous prostate regions, respectively. Note that the quality of scans and masks varies greatly among patients.

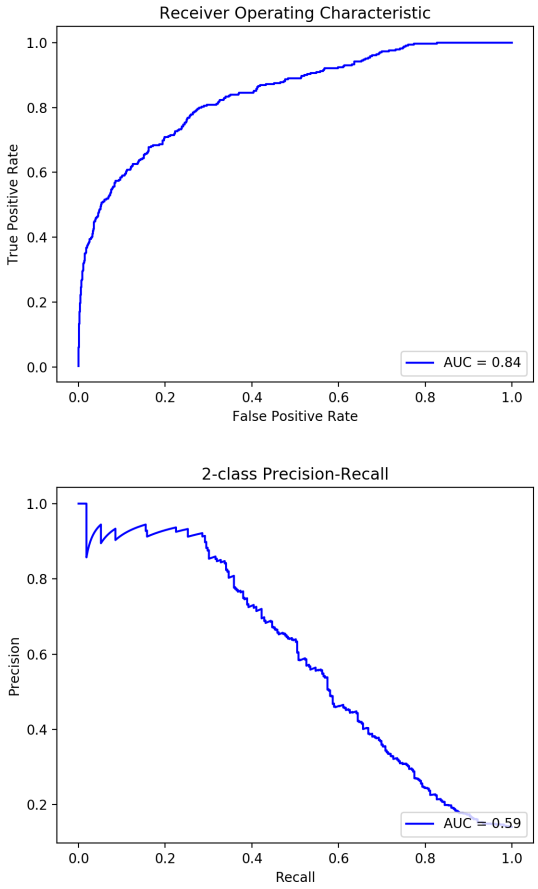

Fig. 3. Receiver Operating Characteristic curve for GBC with n=6 (top), Precision Recall curve for GBC with n=4 (bottom). These were the best performing models for their respective diagnostic statistic.

REFERENCES

[1] L. Hussain, A. Ahmed, S. Saeed, S. Rathore, I. A. Awan, S. A. Shah, A. Majid, A. Idris, and A. A. Awan, Prostate cancer detection using machine learning techniques by employing combination of features extracting strategies, Cancer Biomarkers, vol. 21, no. 2, pp. 393413, 2018.

[2] Saito, Takaya, and Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. Ed. Guy Brock. PLoS ONE 10.3 (2015): e0118432. PMC. Web. 7 June 2018.

[3] X. Wang, W. Yang, J. Weinreb, J. Han, Q. Li, X. Kong, Y. Yan, Z. Ke, B. Luo, T. Liu, and L. Wang, Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning, Scientific Reports, vol. 7, no. 1, 2017.