Benjamin Barnett, Yaniv Bronshtein, Krishna Piratla Final Project

MSDS 534: Statistical Learning for Data Science 11/19/21

***Analysis of Breast Tumors***

**Abstract**

In this paper, we studied the features of breast cell nuclei from digitized images to diagnose early signs of breast cancer. We found that logistic regression, linear discriminant analysis, quadratic discriminant analysis, logistic GAM, random forest, generalized boosting regression, and linear and radial kernel SVM all have excellent classification performance, and are able to distinguish malignant tumors from benign tumors at different threshold values.

**Introduction**

Breast cancer is a ubiquitous disease that needs to be addressed. In the United States, there is a 1 in 8 chance that a woman is diagnosed with breast cancer in her lifetime, and is the leading cause of death among women aside from skin cancer [1]. But early detection through regular screening can help decrease mortality. Fine needle aspiration, a type of biopsy, can be used on breast mass to analyze the structure of different tumors. By studying different features such as size and shape, we hope to successfully distinguish malignant breast tumors from benign ones.

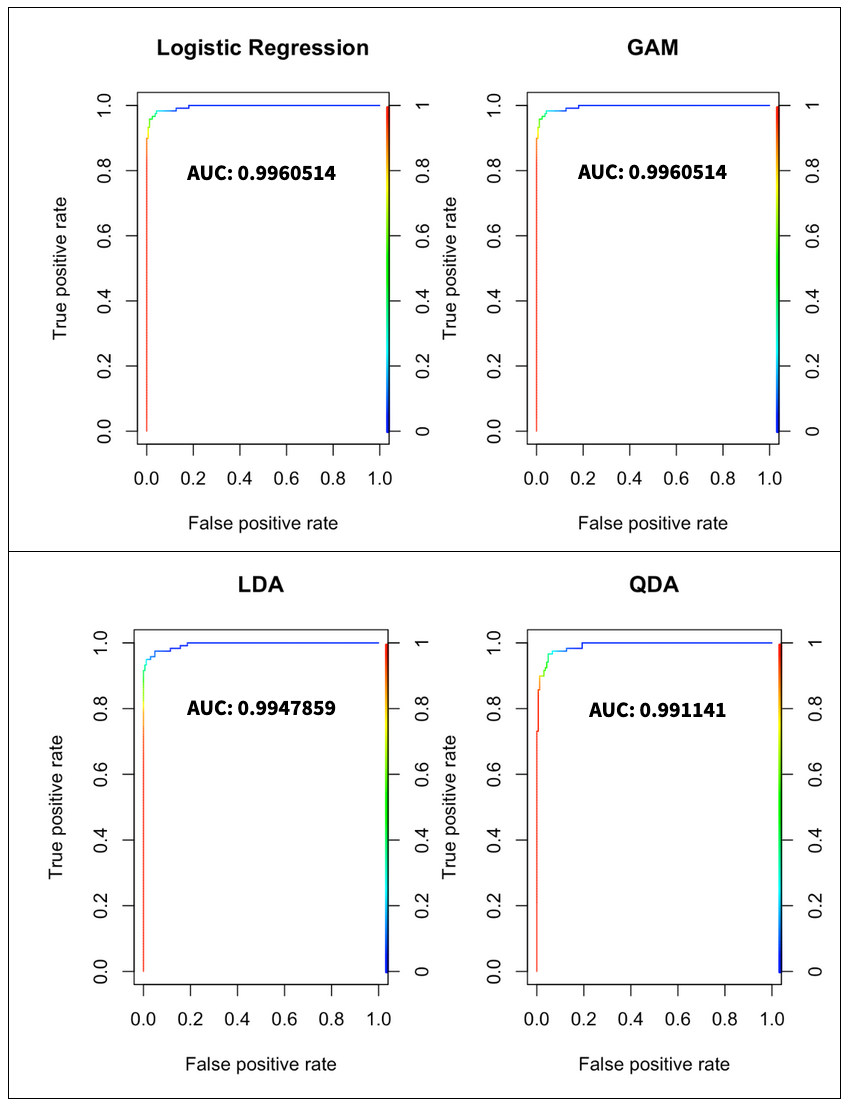
We analyzed computed features from digitized images of the FNA of breast cell nuclei including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. For each image, the mean, standard error, and mean of the largest 3 values were computed, resulting in 30 total features. Because these features were highly correlated with each other, we performed principal component analysis and used the first 7 principal components (representing 91 percent of the explained variance) in our study. Thereafter, we employed several classification methods including logistic regression, LDA, QDA, Logistic GAM, Random Forest, Support Vector Machine, and Boosting to compare and contrast the performance of each model on 50% testing data. The classification results are displayed with the ROC curves along with confusion matrices at different thresholds. Continue reading below for the methods and analysis.

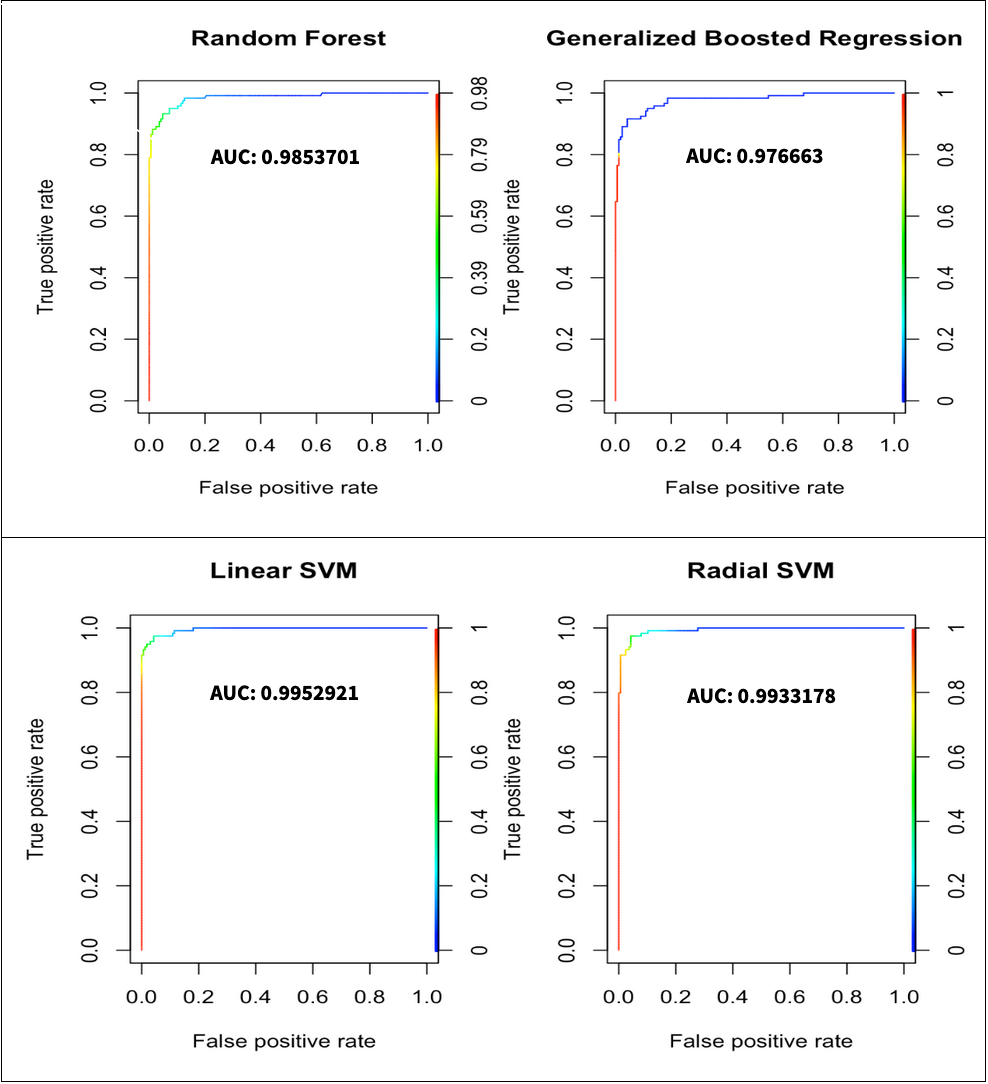
**Methodology**

In our study, we used R for Logistic Regression, LDA, QDA, logistic GAM, Random Forest, SVM, and boosting on seven principal components. To grow our Random Forest, we used 5000 trees and randomly sampled 2 variables as candidates for each split. Because other *mtry* values had similar test performance, we opted for the recommended floor(sqrt(p)) for each variable split. Regarding our boosting model, we used 5000 trees with a depth of 4 to enable variable interactions. Finally, tuning was utilized for the linear and radial support vector machine kernels.

**Results**

To help compare model performance, we initialized a global seed at 1. Below are the receiver operating characteristic curves for logistic regression, LDA, QDA, logistic GAM, random forest, SVM, and generalized boosted regression.





From the ROC curves, we can see that all of the models distinguished malignant tumors from benign tumors quite well, as the AUC values are greater than 0.9. From the confusion matrices found in the appendix, we observe that logistic regression and GAM at the threshold value of 0.5 have the best classification accuracy at over 97%, followed by linear SVM with a classification accuracy at over 96%. While we emphasize classification performance at 0.5, we also include the matrices at 0.2 and 0.8 for the reader to evaluate tradeoff at different threshold values.

Looking at the threshold of 0.2, the classification accuracy in decreasing order is LDA, Logistic Regression and GAM, QDA, Linear SVM, Random Forest, Radial SVM, and Boosting. Interestingly, the lowest performance at the threshold of 0.2 is still above 0.9, which is quite impressive. At the threshold of 0.8, the classification accuracy in decreasing order is radial SVM, logistic regression and GAM, QDA, linear SVM, LDA, Boosting, and Random Forest. The threshold value at 0.2 has better performance than at 0.8, but still has strong performance overall.

**Discussion**

We employed a number of classification methods to analyze the features of breast cell nuclei. For the final report, we would like to explore the data with visualizations to see if there are notable characteristics between benign and malignant tumors. We would also like to incorporate deep learning models such as the Multilayer Perceptron Network and Recurrent Neural Network and evaluate all of our classification methods with F1 score as a new metric. By using more than just accuracy, we can hopefully distinguish which is the best method to diagnose breast cancer.

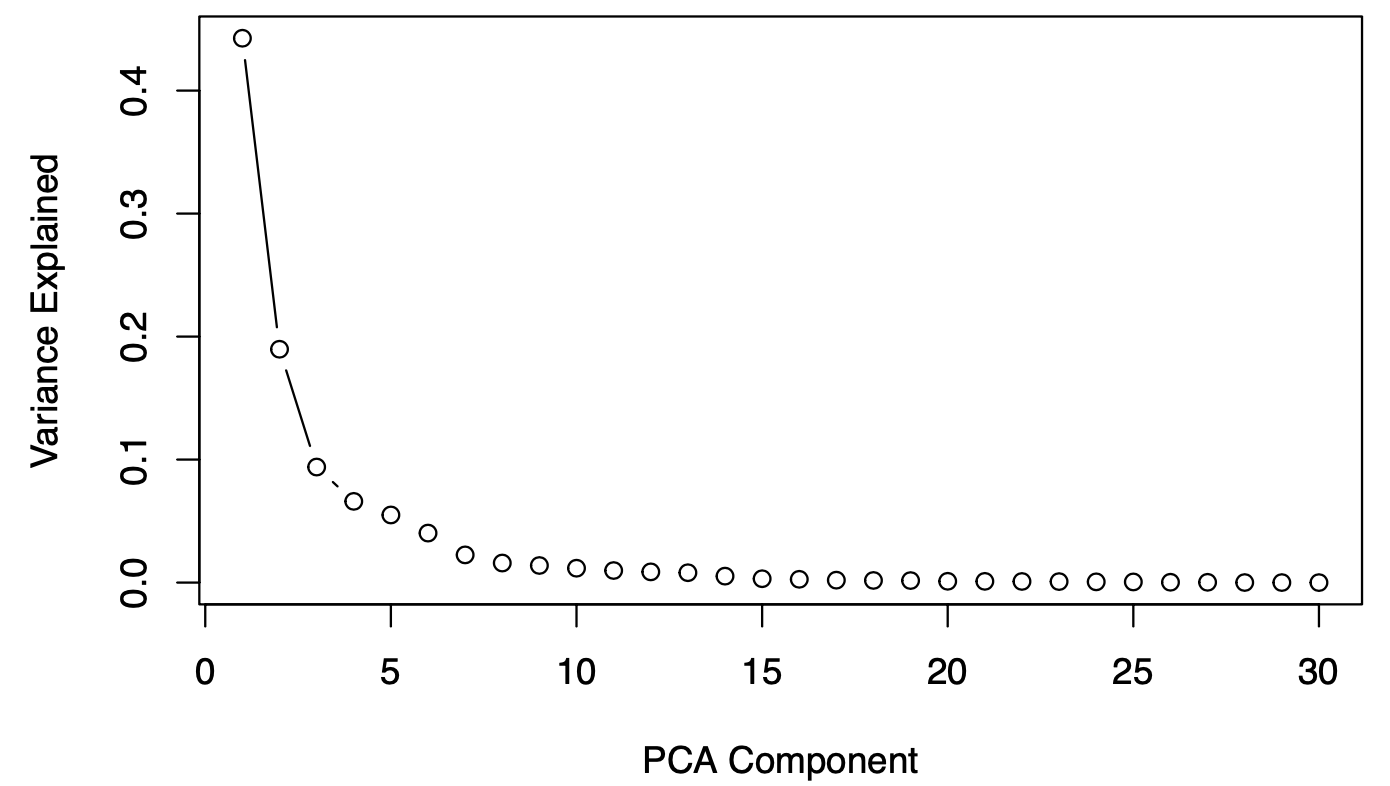
**References**

[1] “How Common Is Breast Cancer? Breast Cancer Statistics.” *American Cancer Society*, https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html.

[2] Machine Learning for Cancer Diagnosis and Prognosis*,* pages.cs.wisc.edu/~olvi/uwmp/cancer.html.

[3] Learning, UCI Machine. “Breast Cancer Wisconsin (Diagnostic) Data Set.” *Kaggle*, 25 Sept. 2016, https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.

**Appendix**



**Figure 1.** Proportion of variance explained for each principal component. By extracting the first seven components, we have 91 percent of the data variance.

| **Logistic Regression** | **GAM** |
| --- | --- |
| **LDA** | **QDA** |
| **Random Forest** | **Generalized Boosted Regression** |
| **Linear SVM** | **Radial SVM** |

**Figure 2.** Confusion matrices and accuracy scores at three different threshold values.