Breast Cancer Diagnosis Using L2 Regularized Logistic Regression
DATA 1030 Final Report- Cameron Webster

December 2, 2020

1 Introduction

While cancer diagnostic accuracy has improved over the past several decades, there is room for improvement. Much of the work is done by pathologists that analyze histology slides by looking for aberrancies in the cell compositions of breast samples. As any task performed by humans is prone to error, so is this process when it comes to misdiagnosis. If a pathologist misclassifies a patient's histology, they may undergo unnecessary and damaging treatments such as chemotherapy or radiation therapy when there isn't a need. In the case of misdiagnosis as benign, the patient may go untreated for breast cancer and suffer fatal consequences. It would benefit pathologists to consult the support of any tools that may help them more accurately predict a patient's condition.

This project attempts to create a diagnostic tool that will leverage machine learning to classify a patient's tumors as benign or malignant. The dataset used for this project came from the UCI Machine Learning repository and was published by the University of Wisconsin. It contains an id, a label of malignant or benign, and 30 characteristics of the cell nuclei of breast cell masts of 569 patients. A tool created by the University of Wisconsin researchers that automatically detects the boundaries of individual cells derived the characteristics found in this dataset. Once the tool finds the cell boundaries, it applies a real-valued number for radius, texture, perimeter, area, smoothness, compactness, concavity, number of concave points, symmetry, and fractal dimension for each cell. For the cells in a given slide, the mean, standard errors, and worst (mean of the three largest values) for each aforementioned characteristic are derived. The combination of these three descriptive statistics with the ten characteristics constitutes the 30 features of this dataset [1].

Several authors have published studies that use this dataset to answer the question of whether or not a patient has breast cancer based on the features of their histology samples' features. In Mangasarian et. al., the authors used a technique known as Multi-Surface-Method Decision Tree with 10-fold cross validation to create a classifier that predicts whether or not the patient's cancer is malignant based on three features: mean texture, worst area, and worst smoothness. Their resulting model performed well for accuracy, with a score of 97% [1]. Other papers have used this dataset for testing novel proofs-of-concept designs for machine learning methodologies. In Campbell et. al., the authors set out to demonstrate the performance of a novel Support Vector Machine training algorithm on the dataset by attempting to answer the question of whether or not a patient has cancer. In demonstrating their model, they found a peak accuracy score of 98.5% [2]. In each of these instances, the accuracy scores are very high. However, when dealing with problems that address patients' health issues, accuracy must be as close to perfect as possible. It is for this reason that the goal of this project is to improve on previous performances like those previously mentioned.

2 Exploratory Data Analysis

The following section contains several of the graphics created during exploratory data analysis.

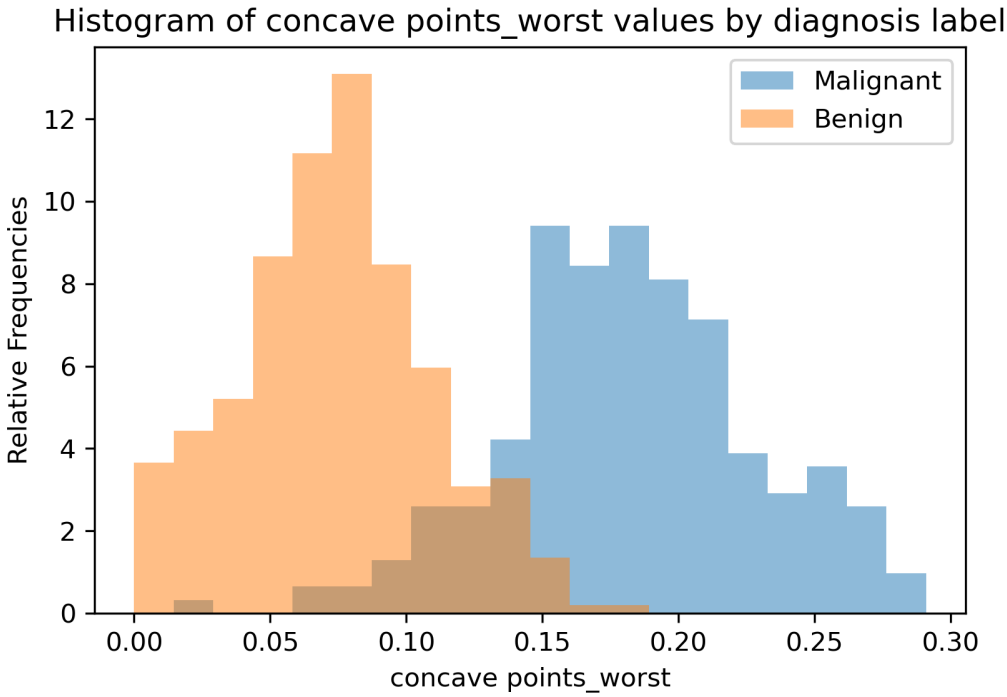

Figure 1 This figure displays the histogram distribution of the average of the number of concave portions for the three cells in a slide with the largest value for this feature, grouped by label outcome: malignant or benign. Both distributions roughly follow a Gaussian distribution. However, the mean of the distribution for the malignant patients is farther to the right of the benign distribution, leaving very little overlap between the two groups. Additionally, the variance of the malignant group appears to be higher than the benign group, as the left tail of the malignant group extends far into the left side of the benign distribution. Given the stark difference between the distributions, this feature has the potential to play an important role in the machine learning model.

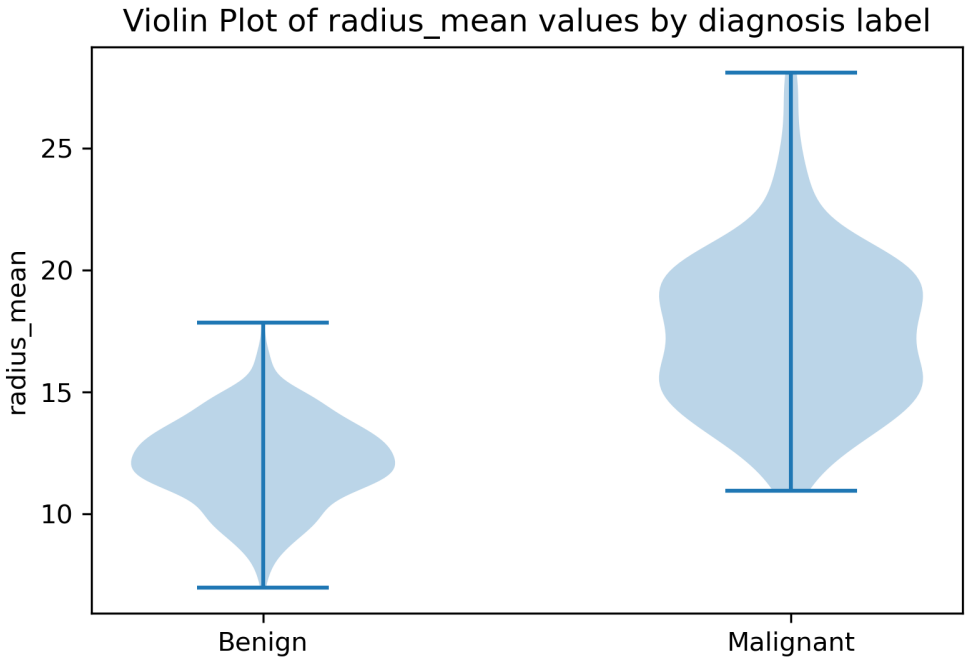

Figure 2 This figure displays the kernel density plots for the average radius of all the cells of a given patient's histology, segregated by the outcome labels. From the graph, it is evident that the feature has both a higher mean and variance for the malignant group than the benign group. Additionally, the malignant group's distribution appears to be bimodal whereas the benign group appears unimodal. While there is a modest overlap between the distributions, there exists a noticeable difference between them.

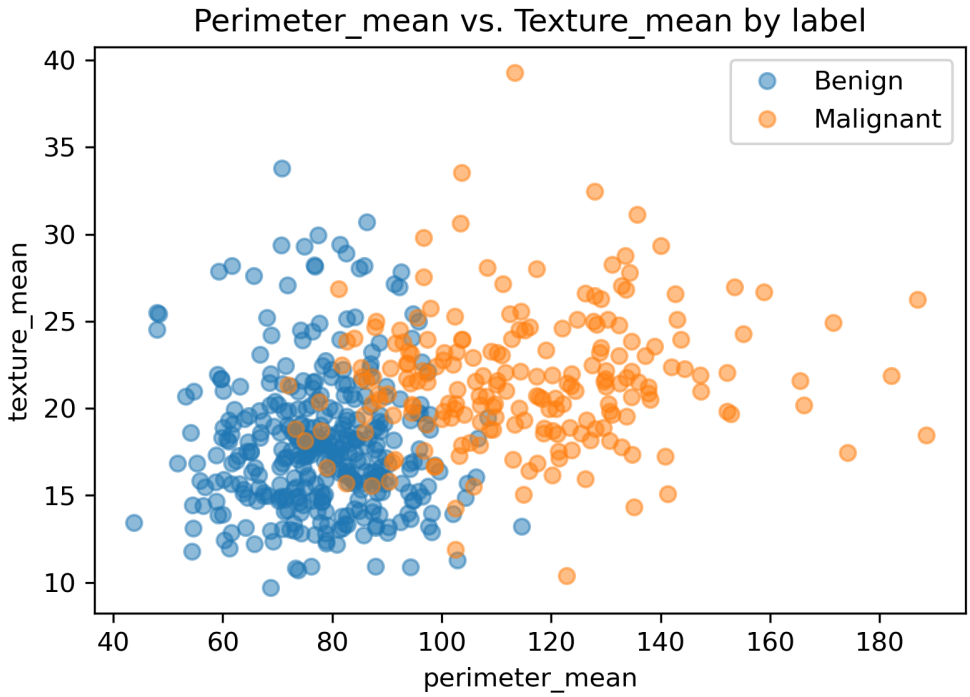

Figure 3 This figure displays the mean perimeter length for cells in an individual slide on the x-axis and mean texture (the standard deviation of the grayscale values of the pixels within a given cell) for the cells in an individual slide on the y-axis. From the plot, there appears to be a slight correlation with the mean perimeter length of cells and mean texture. Additionally, the points associated with benign and malignant labels are delineated. It appears that the center of mass in this plot for malignant observations is higher for the perimeter mean. The trend for mean texture is less obvious but appears to be that the malignant observations have a slightly higher mean for texture mean.

3 Methods

3.1 Data Splitting and Preprocessing

In the data preprocessing step, the splitting step allocated 20% of the observations to testing, the other 80% to 5 fold cross-validation. In each instance of cross-validation, the preprocessor fit and transformed the four training folds before the transforming validation and testing sets. The model trained using a cross-validation split as a way to account for the small number of observations in the set and the variability between random splits that may occur. The train-validation-testing split is 64-18-20 for the same reason– that the number of observations is small and a more substantial proportion of cases is required to properly validate and test the models. Given each observation represents one individual patent, and each patient only has one observation in the set, the data is assumed to be independent and identically distributed with no group structure or time-series property. The preprocessor applied the StandardScaler to each feature because each observation is continuous and, after examining the histogram of all the features, none of them appear to be reasonably bounded and suitable for the MinMaxEncoder. As a result, the final preprocessed

dataset has 30 features. Additionally, the preprocessor label encoded the target variable since it is a 2-category variable.

3.2 Model Selection

Using the splitting and preprocessing methodology, eight different machine learning models were trained and compared: a logistic regression model without regularization, a logistic regression with L1 regularization, a logistic regression with L2 regularization, logistic regression with ElasticNet regularization, a random forest classifier, a support vector machine classifier, a K-nearest neighbors classifier, and an XGBoost classifier. Except for the logistic regression model without regularization, all models were hyperparameter tuned using a brute-force grid search method to find the optimal parameter combination for each model. This process was repeated on 10 different random states for 10 different splits. Below are the parameters tuned and values tried for each model:

Model	Parameters
L1	**C**: 1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3, 1e4
L2	**C**: 1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3, 1e4
ElasticNet	**C**: 1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3, 1e4; **l1_ratio**: 0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99
RF	**max_features**: 1, 3, 5, 10, 20, None; **max_depth**: 1, 3, 5, 10, 20, None; **min_samples_split**: 2, 5, 10
SVC	**gamma**: 1e-2, 1e-1, 1e0, 1e1, 1e2, auto, scale; **C**: 0.1, 0.31, 1, 3.1, 10
KNN	**n_neighbors**: 1, 2, 3, 5, 10, 30, 100, 200; **weights**: uniform, distance
XGBoost	**min_child_weight**: 1, 3, 5, 7; **gamma**: 0, 0.1, 0.2 , 0.3, 0.4; **subsample**: 0.5, 0.66, 0.75, 1; colsample_by_subtree: 0.3, 0.4, 0.5, 0.7, 1

Figure 4 Parameters used for tuning of each model

After tuning, each grid search's best model parameters were extracted and used for comparison on accuracy score on a holdout set. Below are the average accuracy scores for the best of each model across the ten random states:

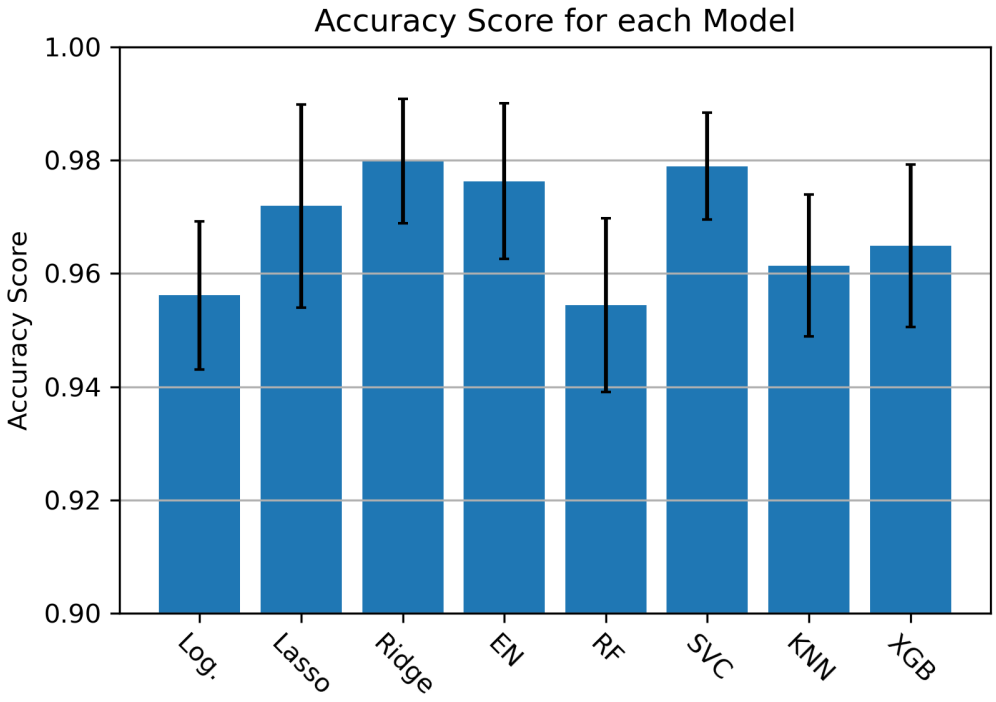

Figure 5 Average accuracy scores for the best model over ten random data splits

The logistic regression model with L2 regularization had the highest test set performance and was chosen as the model of choice. After examining the best value of the inverse regularization parameter for each random state, 1.0 was chosen for the final model because it achieved the highest score on the holdout set for 8 of the ten random states.

3.3 Final Model Formulation

After choosing the best model and hyperparameter choice, the model was retrained on new splits over new random states. Since L2 regularized logistic regression models train quickly, the model was trained on 100 different random states. For each split, 80% of the data was allocated to training and 20% was allocated to testing. For each random state, the preprocessed test set, model, and baseline accuracy score were recorded.

4 Results

4.1 Evaluation of Models

Over the 100 random states, the baseline models returned an average accuracy score of 0.62 with a standard deviation of 0.04. In comparison, the trained models returned an average accuracy score of 0.98 with a standard deviation of 0.01. The trained models achieved an accuracy that is 9 standard deviations above baseline. Similarly, the baseline model's accuracy was 36 standard deviations below the average of the trained models.

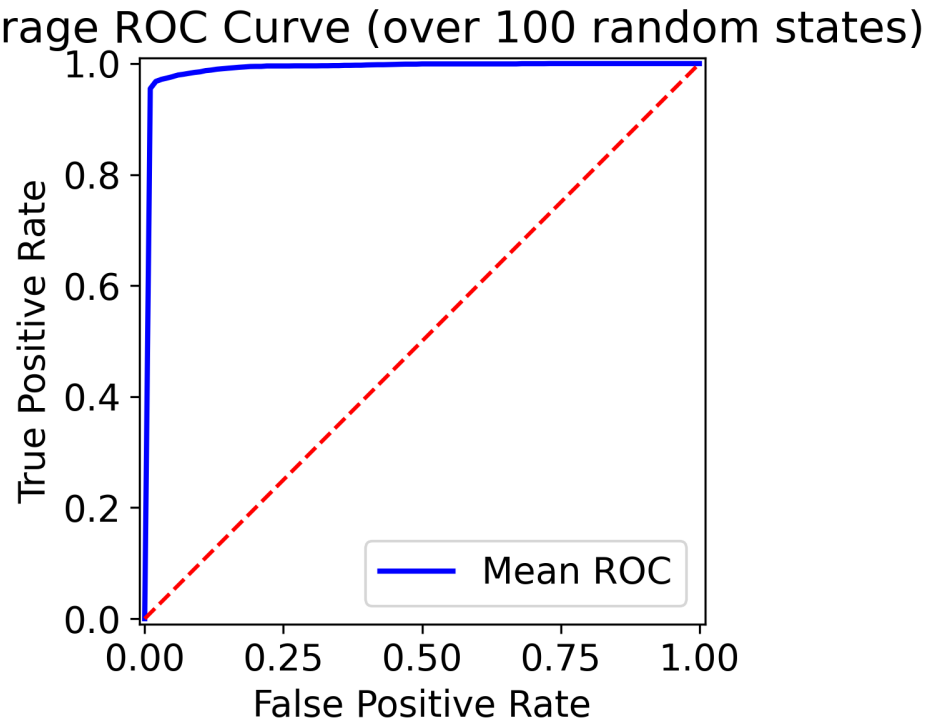

Figure 6 The Averaged ROC Curve over 100 random states

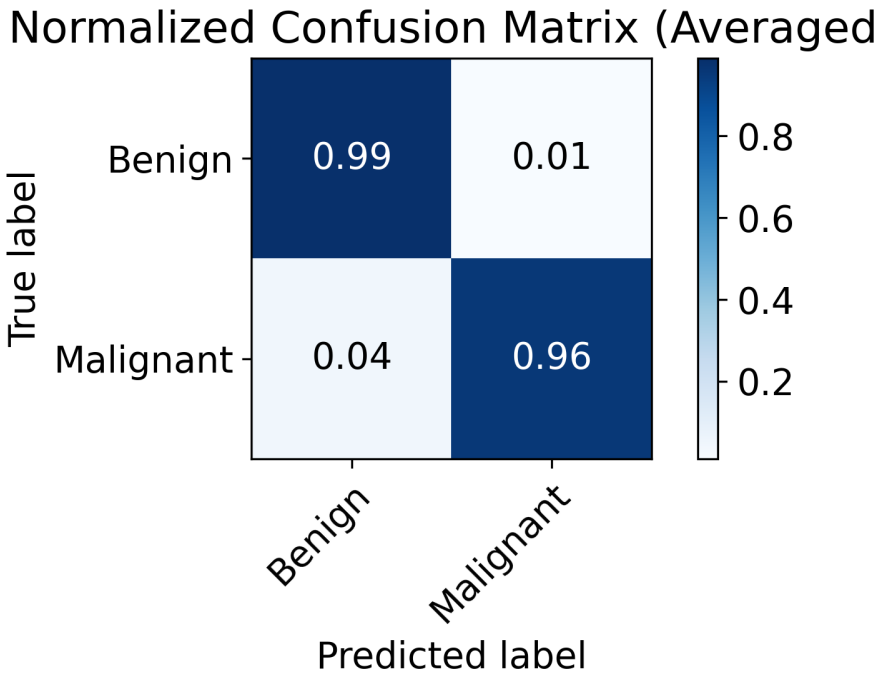

Figure 7 The Normalized Confusion Matrix over 100 random states

Above are the average ROC and confusion matrix for the 100 models, which show that while the model is very accurate, it performs slightly better at classifying negative cases than positive cases.

4.2 Interpretation of Findings

Global Feature importance for the model was calculated using both the coefficients of the models as well as a permutation test over 10 shuffles for each random state. Below are the results of these evaluations:

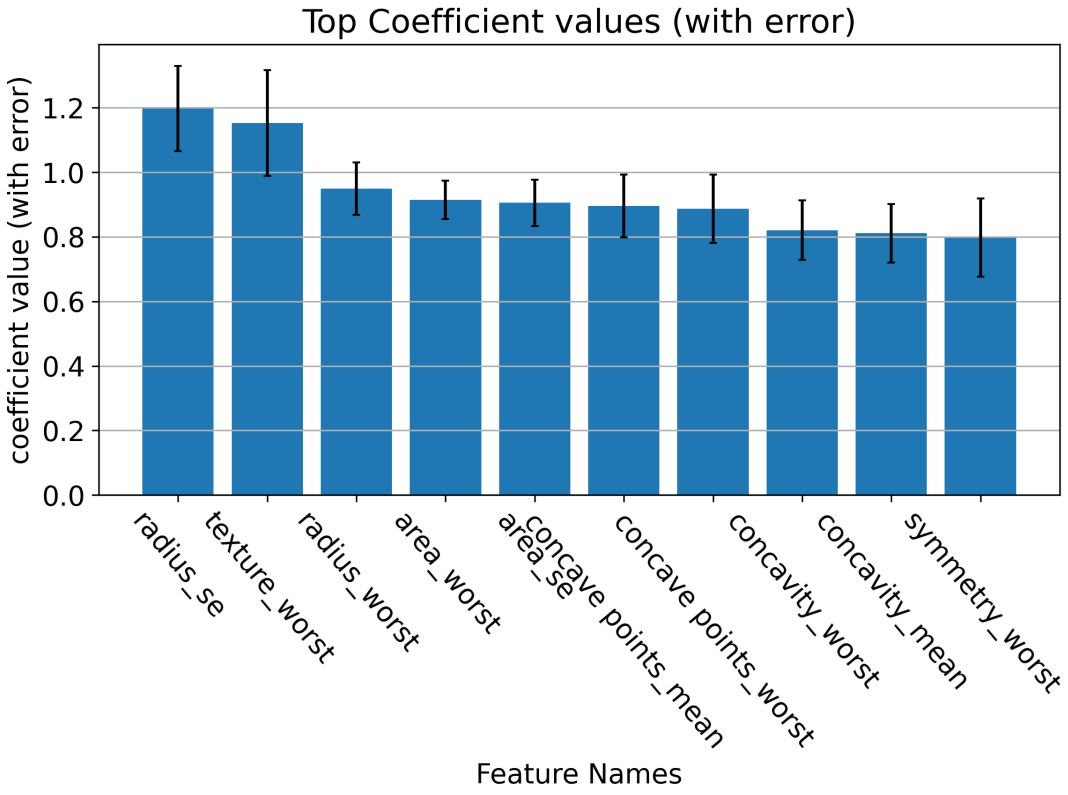

Figure 8 The top average coefficient values for the set of features

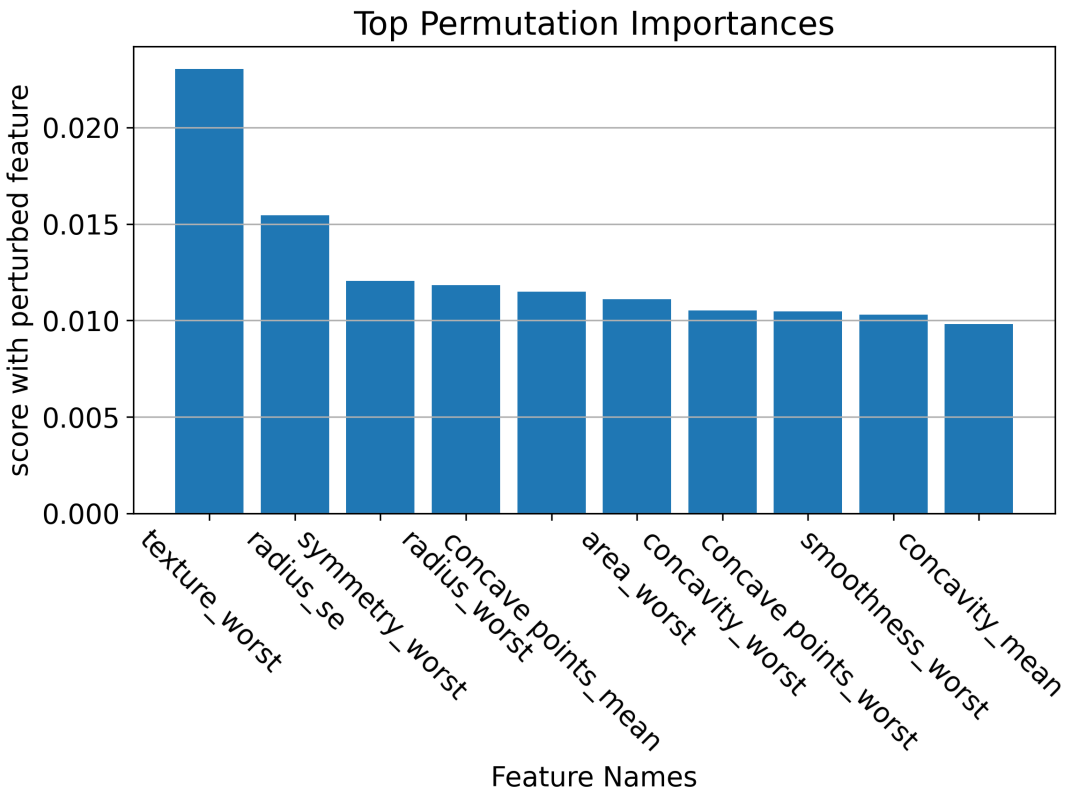

Figure 9 The top average permuation test score for the set of features

From the results, the model deemed standard error of cell mast radius and largest texture values within a given slide as the most prominent features in determining whether or not a patient has breast cancer. This contrasts with previous literature, which did not use either of these features when formulating their models. The previous work that used this dataset selected worst smoothness, mean texture, and worst area as the sole features to train their model. Only one of those features, worst area, was deemed a top feature of importance from the above tests. Given the similarity between the two tests, the feature importances are fairly robust.

5 Outlook

Given the model is a logistic regression, the results are very interpretable in comparison to past models that used support vector machines. In light of this, the radius standard error and worst texture features should be given more weight than they have previously in diagnosing patients. One issue that this model presents is the slight rate of false negatives over false positives. While this ensures that patients do not undergo unnecessary treatments, it also implies that a larger proportion of patients will go undiagnosed for breast cancer, running the risk of catching cancer after the critical window of diagnosis and potentially leading to unrecognized growing cancer masts. An improvement that would benefit this model would be to change the evaluation metric from accuracy to an f score with a beta value greater than 1. This would ensure that the model increases in sensitivity and alerts clinicians more frequently so that they may further evaluate the patient's condition. The reason the accuracy score was used for this model was the precedent set

by previous literature. This probably has to do with the unnatural balance of the dataset. In the clinical setting, a much larger proportion of patients would be benign than in the current case. Additional data that accurately reflects the prevalence of this disease would greatly benefit this study.

6 References

[1]: Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. Biomedical Image Processing and Biomedical Visualization. doi:10.1117/12.148698

[2]: Campbell, C., & Cristianini, N. (1998). Simple Learning Algorithms for Training Support Vector Machines. University of Bristol, Dept. of Engineering Mathematics.

[3]: Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Github repository: https://github.com/camweb36/breast-cancer-diagnosis