Non-invasive Breast Tumor Diagnosis in Machine Learning

Olivia Fan, Alicia Gong

19 December, 2022

Introduction

Breast cancer is the most common cancer worldwide and the most common cancer diagnosed in the US. Each year in the US, about 264,000 cases of breast cancer are diagnosed in women and about 2,400 in men.

Early diagnosis of the condition is crucial to improve the survival rate and relieve suffering in patients. Breast carcinoma is one of the most common cancers occurring in the female population world-wide. Mammography is an effective X-ray imaging technology that detects breast cancer early. In clinical oncology, doctors usually perform morphimetric analysis of mammographic images. Nuclear changes occurring during these transformational steps need to be assessed objectively. Variations in nuclear structure are the morphologic hallmark of cancer diagnosis. There is a gradual shift in the nuclear parameters as the disease progresses from benign to malignant. Nuclear size, shape, chromatin pattern, and nucleoli size and a number have all been reported to change in breast cancer. These nuclear morphometric features have been shown to predict the prognosis of the breast cancer patients. Classically, benign or malignant breast tumors are diagnosed by radiologists' interpretation of mammograms based on morphometric parameters. However, diagnosing cancer is challenging even for the most skilled doctors. The symptoms are often shared with diseases and conditions that are unrelated to cancer, leading doctors to improperly diagnose the disease. According to an expansive study conducted by Dartmouth College, the University of Vermont, and the Fred Hutchinson Cancer Research Center, and published in the March 2015 issue of the Journal of American Medical Association, approximately 13% of the diagnoses missed Stage 1 breast cancer. Meanwhile, 48% failed to detect atypia hyperplasia, a precursor to breast cancer. A significant number also over-diagnosed atypia hyperplasia.

There is, therefore, an urgent need to find new tools that can identify patients with breast cancer. Our study aims to build supervised machine-learning models to predict the diagnosis of breast cancer in a non-invasive framework and understand the most important variables, to assist doctors and radiologists in accurately interpreting mammography imaging.

We built 3 models in total: Lasso-penalized logistic regression, SVM, and Random Forest. First, we used Lasso-penalized logistic regression to select variables and to predict the probabilities. Then, we used the predictors selected by LASSO to build a SVM model. Finally, we build another Random Forest model as a comparison to the SVM model, also for a more straightforward interpretation.

Data

We obtain the Breast Cancer Wisconsin (Diagnosis) Data Set from Kaggle. The dataset contains diagnosis results and features of the cell nuclei computed from a digitized image of a fine needle aspirate (FNA) of a breast mass for 568 patients. The size of the nucleus is expressed by the features radius and area. The shape is expressed by the features smoothness, concavity, compactness, concave points, symmetry, and fractal dimension. The perimeter expresses both the size and shape of the nucleus. A higher value of shape features corresponds to a less regular contour and, therefore, to a higher probability of malignancy. For each of the

features the mean value, worst value (mean of the three largest values), and standard error are computed for each image, resulting in 30 features of 568 images.

We examined an extensive list of variables in our models, and a full table of them can be found in the appendix.

Data Processing

The original dataset contains a blank column `...33`, so we dropped it. We also dropped the `id` column, and rename several columns, `concave points_mean`, `concave points_worst` and `concave points_se` that contains blank space in their names.

In order to fit the SVM on the data, we processed the data to encode the `diagnosis` variable into a factor variable with level 1 associating with malignant tumor and -1 associating with benign tumor.

We partition the data into training and testing sets using a 70-30 percentage split (70% of the original data as the training set, and 30% as the testing set) in order to examine the performance of the models on future data.

Exploratory Data Analysis

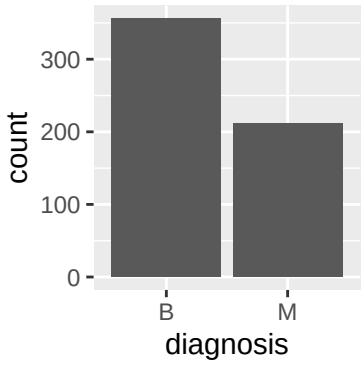


Figure 1: Distribution of Malignant and Benign Tumors

The bar plot (Figure 1) shows that there is a larger number of benign than malignant tumors.

We divide the data into 3 categories according to their features, namely, mean, standard error and worst.

Based on the correlation matrix (Figure 2), we have several major observations:

- Radius_mean, perimeter_mean, and area_mean are highly correlated.
- Compactness_mean, concavity_mean and concave_points_mean are highly correlated.

Furthermore, we observe from the pairwise scatterplot matrix (Figure 3) that the two classifications seem to be generally separable, with distinct regions in the visualization that cleanly cluster without much mingling or mixing. Overall across malignant and benign tumors, there seems to be a strong positive linear relationship between `radius_mean` and `parameter_mean`, `radius_mean` and `area_mean`, as well as `area_mean` and `parameter_mean`, which hints again at the collinearity issue which we will later tackle at through variable selection. While the two classifications together constitute a roughly linear relationship between predictors, malignant tumors (red) generally associate with higher values in both predictors accumulating in the right top corner, while benign tumors (blue) generally associate with lower ones in the left bottom corner.

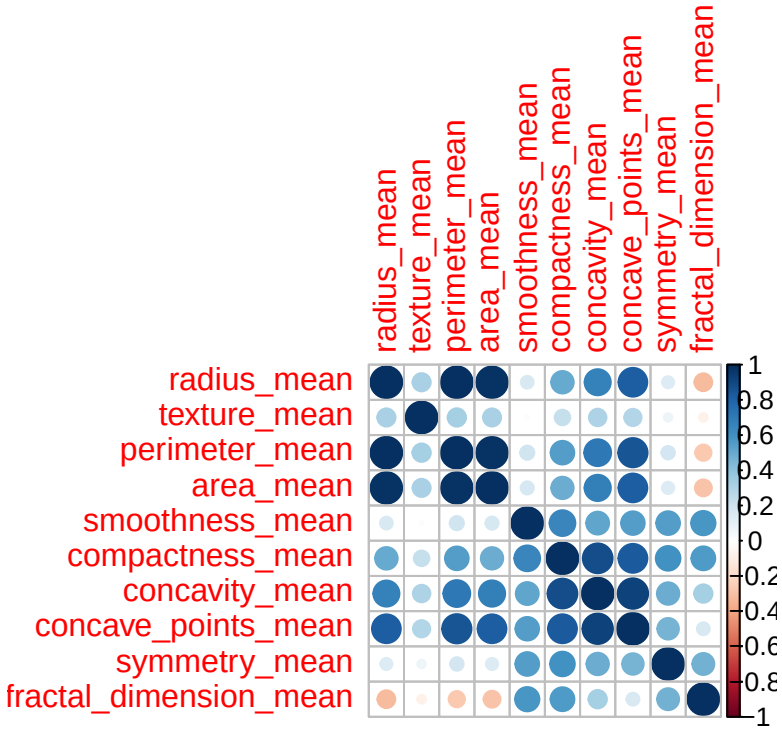


Figure 2: Correlation between mean predictors in the dataset

We observe that there does not seem to be a separating hyperplane for the two classes for predictors in the relationship between `texture_mean` and `symmetry_mean`, and between `smoothness_mean` and `fractal_dimension_mean` since the observations in two classes mingle together. This hints at the fact that these predictors might not be helpful for the two class classification problem, which we will filter out in our model through variable selection.

According to the distribution of predictors by diagnosis types in Figure 5, the distributions of the predictors seem to be all unimodal, with no apparent outliers and generally right-skewed, with `concavity_mean` and `concave_points_mean` being particularly right-skewed, hinting at the high correlation between the two predictors. Hence we want to consider including only one of them in our model. We take a closer look at the distributions of these two predictors in Figure 4:

After deriving the histogram comparing distributions of predictors based on the two classifications, we would like to find features with little overlap between benign and malignant classes which will likely to be significant for diagnosis. We plot the distribution of the predictors separated by the benigh and malignant classes, and observe again that predictors associated with texture, smoothness, symmetry and fractal dimension are inseparable and therefore might not be helpful for the classification problem. For example, the distributions of `smoothness_se` for benign and malignant cancers almost completely overlap, same do `symmetry_se`, `fractal_dimension_se` and `texture_se`. Therefore, we will consider filter out these predictors in our final model.

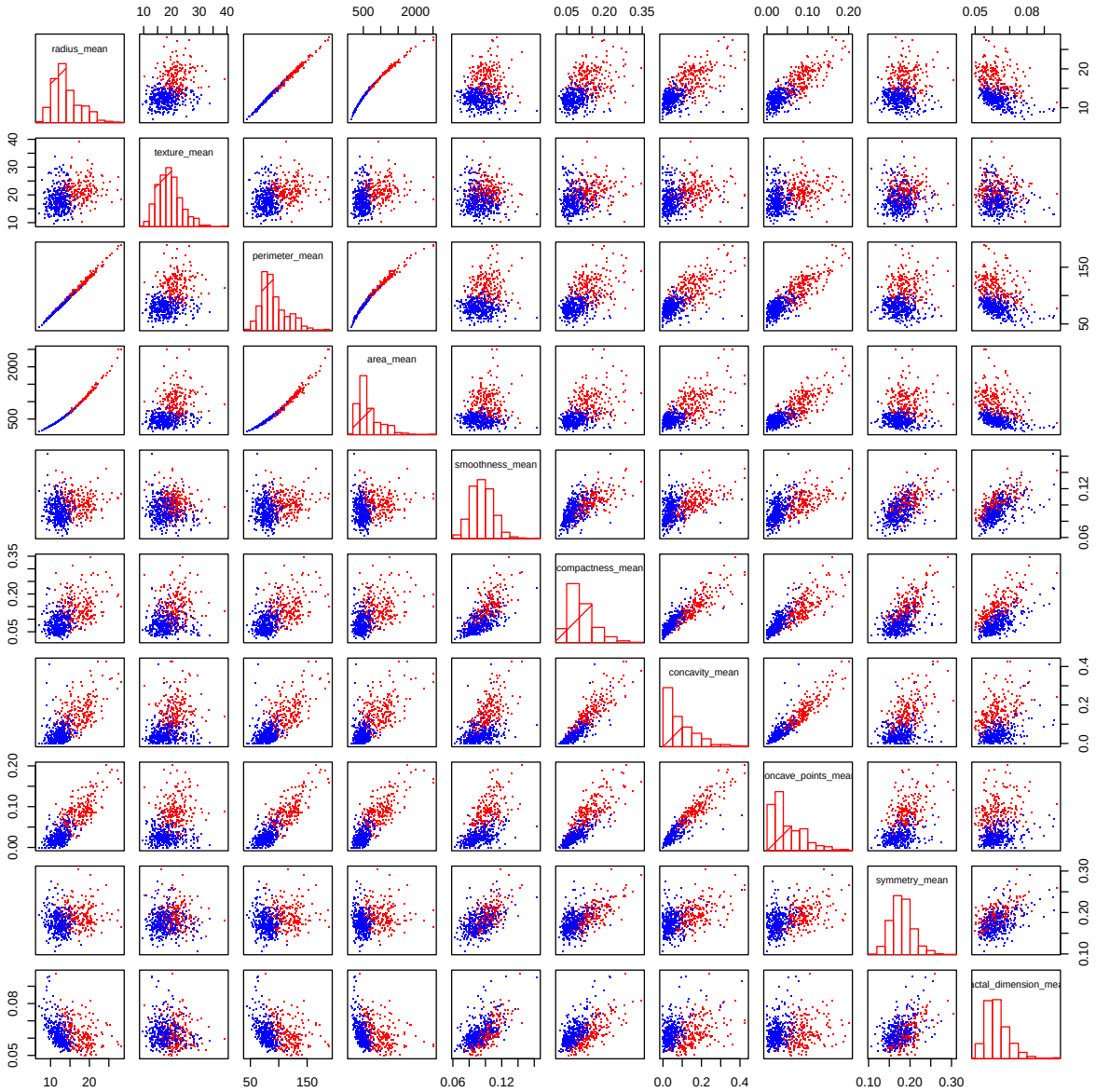


Figure 3: Scatterplot of mean predictors by malignant (red) and benign (blue) tumors

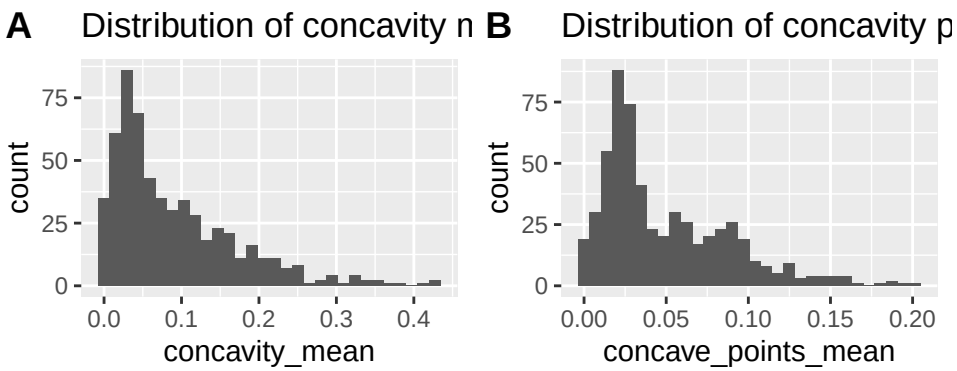


Figure 4: Distribution of concavity mean and concavity point mean

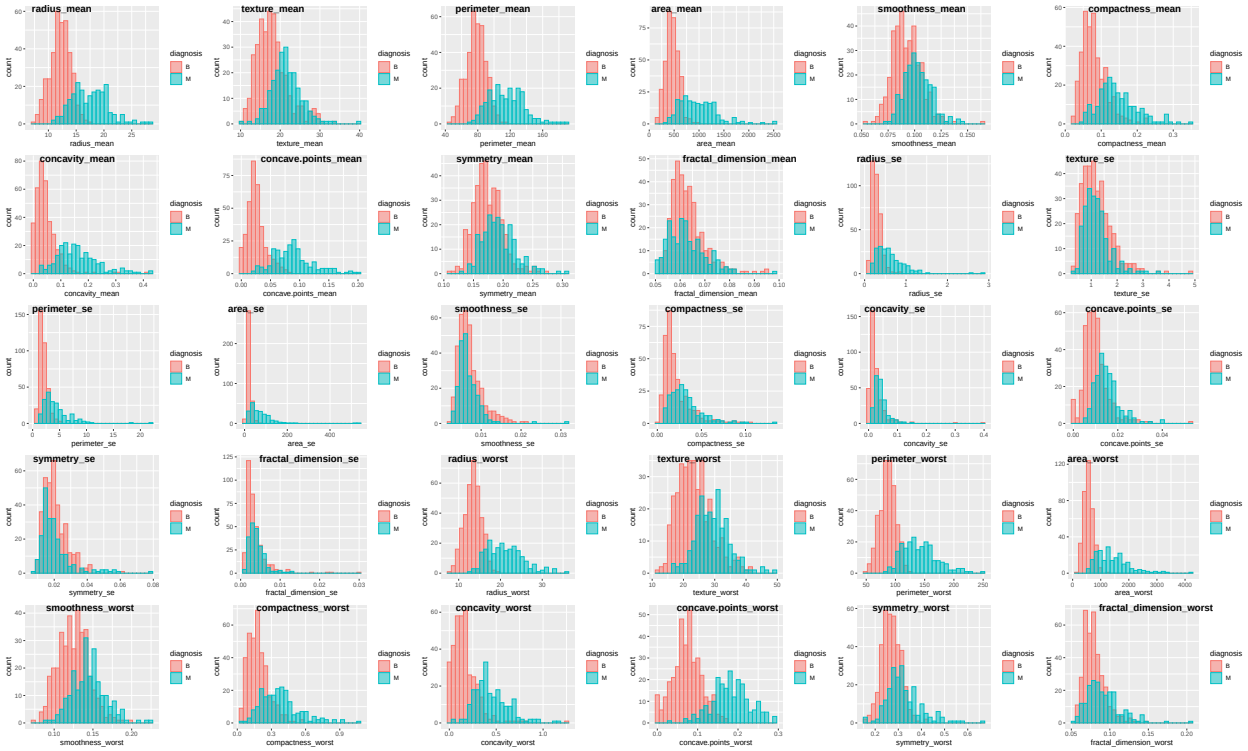


Figure 5: Distribution of predictors by diagnosis

Methodology

Model 1: LASSO-Penalized Logistic Regression

We decided to use a LASSO-penalized logistic regression model to perform variable selection by gauging insights into which predictors are the most contributive, since less significant variables are forced to be exactly zero, and the most significant variables are kept in the final model. Also because LASSO is a great tool dealing with multicollinearity issue. Since our EDA shows that many of the input variables are correlated, LASSO will drop the highly correlated features.

As explained in our exploratory data analysis above, we filter out predictors associated with texture, smoothness, fractal dimension and symmetry in our model due to the lack of separation in the values of these predictors for the benign and malignant tumor classes.

Hyperparameter Tuning

We fitted the LASSO-penalized logistic regression model using the optimal hyperparameter $\lambda = 0.001399$ via cross validation. To explore the interaction between compactness and the number of concave points, we included the interaction term `compactness_mean*concave_points_mean`.

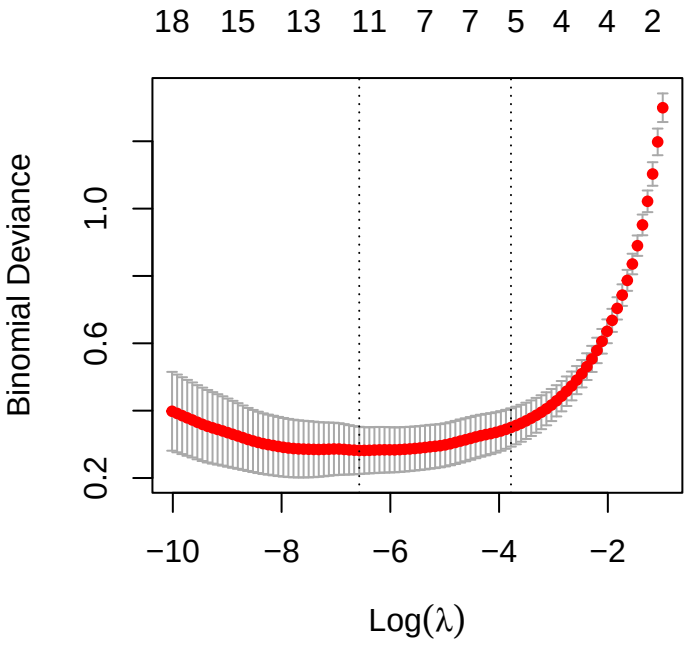


Figure 6: LASSO parameter tuning

We derived the following LASSO-penalized logistic regression model coefficients using the optimal hyperparameter $\lambda = 0.001399$:

```
## [1] 0.001398917
```

	coefficient
(Intercept)	-10.4871
radius_mean	-0.1130
compactness_mean	-33.6480
concavity_mean	14.3918
concave_points_mean	19.0997
area_se	0.0944
compactness_se	-30.3992
concavity_se	-31.2549
concave_points_se	34.8557
radius_worst	0.0112
area_worst	0.0049
compactness_worst	4.9213
concavity_worst	7.6360
concave_points_worst	26.8649
compactness_mean:concave_points_mean	74.5208

Result: Logistic Model

$$\log(\frac{P}{1-P}) = -10.4871 - 0.1130 \times radius\_mean - 33.6480 \times compactness\_mean \quad (1)$$

$$+ 14.3918 \times concavity\_mean + 19.0997 \times concave\_points\_mean \quad (2)$$

$$+ 0.0944 \times area\_se - 30.3992 \times compactness\_se - 31.2549 \times concavity\_se \quad (3)$$

$$+ 34.8557 \times concave\_points\_se + 0.0112 \times radius\_worst + 0.0049 \times area\_worst \quad (4)$$

$$+ 4.9213 \times compactness\_worst + 7.6360 \times concavity\_worst + 26.8649 \times concave\_points\_worst \quad (5)$$

$$+ 74.5208 \times compactness\_mean \times concave\_points\_mean \quad (6)$$

Model Interpretation

We found that while predictors `concavity_mean`, `concave_points_mean`,`area_se`, `radius_worst`, `concave_points_se`, `area_worst`, `compactness_worst`, `concavity_worst`, `concave_points_worst` and `concavity_se` are positively associated with the response log odds, predictors `radius_mean`, `compactness_se` and `compactness_mean` are negatively associated with the response log odds. Of all the predictors, `concave_points_se` and `compactness_mean` have the largest magnitude, and therefore are the most significant predictors. For each 0.01 additional unit increase in `concave_points_se`, the log odds of the probability that a patient is diagnosed as malignant cancer tends to decrease by 34.85% holding all else constant. On the other hand, for each 0.01 additional unit increase in `concavity_se`, the log odds of the probability that a patient is diagnosed as malignant cancer tends to decrease by 31.2549% holding all else constant. By the interaction term, for each 0.01 additional unit increase in `compactness_mean`, the coefficient of `concave_points_mean` is expected to increase by 0.745208, holding all else constant.

	compactness_mean
Min.	0.0193800
1st Qu.	0.0633000
Median	0.0926300
Mean	0.1027247
3rd Qu.	0.1303000

	compactness_mean
Max.	0.2867000

	concave_points_mean
Min.	0.0000000
1st Qu.	0.0202700
Median	0.0332300
Mean	0.0469293
3rd Qu.	0.0684700
Max.	0.2012000

To further interpret the interaction between `compactness_mean` and `concave_points_mean`, we created the following visualization with different levels of fixed `compactness_mean`, at 0.01938 (low), 0.26 (medium) and 0.28670 (high). To avoid interpolation, we first looked at the quantile summaries of these two predictors. There seems to be significant interactions between `compactness_mean` and `concave_points_mean`, since the effect of `concave_points_mean` on the probability changes drastically on the different levels of the `compactness_mean_level` values. For high and medium values of compactness mean level, as the number of concave points increases, the probability that a patient is diagnosed as malignant cancer also tends to increase, and high values of compactness mean level associates with a stronger effect by the number of concave points on this probability. On the other hand, for low values of compactness mean level, the probability tends to stay at nearly 1 and the number of concave points does not affect the probability by a significant amount. Therefore, through this interaction term, we discover that higher levels of compactness mean associates with a stronger effect that concave points mean has on the response probability.

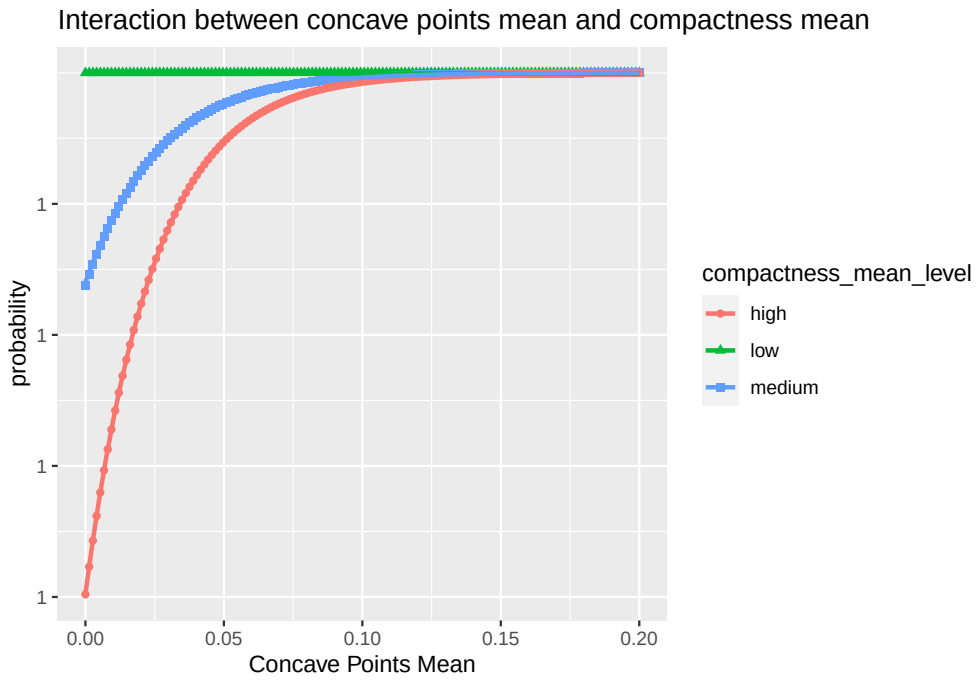


Figure 7: Interaction Term Visualization

Prediction

Using the logistic regression model, besides classification we also want to understand uncertainty - more specifically, predictive probabilities that a tumor is benign or malignant given the values of the predictors:

concave_points_mean	area_se	compactness_se	radius_worst	concavity_worst	concave_points_worst	compactness_mean	concavity_se
0.147	153.40	0.049	25.38	0.712	0.265	0.278	0.054
0.070	74.08	0.013	24.99	0.242	0.186	0.079	0.019
0.094	24.32	0.035	15.49	0.539	0.206	0.193	0.036
0.080	19.21	0.059	15.03	0.694	0.221	0.229	0.055
0.053	45.40	0.012	19.07	0.291	0.161	0.072	0.020
0.103	54.18	0.025	20.96	0.478	0.207	0.202	0.032
0.095	112.40	0.019	27.32	0.537	0.239	0.103	0.034
0.031	14.67	0.019	14.50	0.189	0.073	0.127	0.017
0.077	93.54	0.027	21.31	0.345	0.149	0.107	0.051
0.052	41.00	0.034	16.82	0.696	0.155	0.152	0.042
0.078	35.03	0.029	20.21	0.527	0.186	0.156	0.027
0.056	24.91	0.030	15.89	0.519	0.145	0.110	0.048

concavity_mean	radius_mean	concave_points_se	area_worst	compactness_worst	probabilities	predicted_class
0.300	17.99	0.016	2019.0	0.666	1.000	M
0.087	20.57	0.013	1956.0	0.187	1.000	M
0.186	13.00	0.012	739.3	0.540	0.982	M
0.213	13.73	0.016	697.7	0.772	0.976	M
0.074	14.68	0.011	1138.0	0.187	0.991	M
0.172	16.13	0.013	1315.0	0.423	1.000	M
0.148	19.81	0.015	2398.0	0.315	1.000	M
0.046	13.08	0.006	630.5	0.278	0.002	B
0.149	18.61	0.019	1403.0	0.212	1.000	M
0.122	11.84	0.010	888.7	0.578	0.994	M
0.135	16.13	0.009	1261.0	0.580	0.999	M
0.132	14.25	0.012	799.6	0.424	0.881	M

```
## [1] 0.9883041
```

We achieved a prediction accuracy of 0.9883. To interpret the predictions, we see that a patient with tumor with `concave_points_mean` of 0.147, `area_se` of 153.400, `compactness_se` of 0.049040, `radius_worst` of 25.380, `concavity_worst` of 0.711900, `concave_points_worst` of 0.26540, `compactness_mean` of 0.27760, `concavity_se` of 0.053730, `concavity_mean` of 0.300100, `radius_mean` of 17.990, `concave_points_se` of 0.712, `area_worst` of 2019.0, `compactness_worst` of 0.66560 is expected to have a 100% of being diagnosed as malignant tumor. On the other hand, a patient with tumor with `concave_points_mean` of 0.031100, `area_se` of 14.670, `compactness_se` of 0.018980, `radius_worst` of 14.500, `concavity_worst` of 0.189000, `concave_points_worst` of 0.07283, `compactness_mean` of 0.12700, `concavity_se` of 0.016980, `concavity_mean` of 0.018980, `radius_mean` of 13.080, `concave_points_se` of 0.006490, `area_worst` of 630.5, `compactness_worst` of 0.27760 is expected to have a 87.47e% of being diagnosed as malignant tumor.A patient with tumor with `concave_points_mean` of 0.055980, `area_se` of 24.910, `compactness_se` of 0.029950, `radius_worst` of 15.890, `concavity_worst` of 0.518600, `concave_points_worst` of 0.14470, `compactness_mean` of 0.10980, `concavity_se` of 0.048150, `concavity_mean` of 0.131900, `radius_mean` of 14.250, `concave_points_se` of 0.011610, `area_worst` of 799.6, `compactness_worst` of 0.42380 is expected to have a 87.47% of being diagnosed as malignant tumor.

Model 2: Support Vector Machine

We fitted a linear kernel SVM and a radial kernel SVM to compare the performance between the two kernels and select the one that provides the better formance on test data.

Linear Kernel SVM

We use the predictors selected by the LASSO penalized logistic regression as predictors for the support vector machine model in order to avoid overfitting.

We tune the hyperparameter cost through cross validation considering a range of values from 0.001 to 100, and selected the optimal cost = 0.1

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost
##     5
##
## - best performance: 0.04794872
##
## - Detailed performance results:
##    cost      error dispersion
## 1 1e-03 0.10365385 0.06186851
## 2 1e-02 0.06320513 0.04380102
## 3 1e-01 0.05820513 0.03635092
## 4 1e+00 0.05044872 0.02930865
## 5 5e+00 0.04794872 0.03282235
## 6 1e+01 0.05807692 0.03211418
## 7 1e+02 0.06551282 0.03808553
```

```
##
## Call:
## best.tune(METHOD = svm, train.x = diagnosis_binary ~ concave_points_mean +
##     area_se + compactness_se + radius_worst + concavity_worst + concave_points_worst +
##     compactness_mean + compactness_se + concavity_se + concavity_mean +
##     radius_mean + concave_points_se + area_worst + compactness_worst +
##     compactness_mean * concave_points_mean, data = cancer_train,
##     ranges = list(cost = c(0.001, 0.01, 0.1, 1, 5, 10, 100)), kernel = "linear")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  5
##
## Number of Support Vectors:  48
##
##  ( 24 24 )
##
```

```
##
## Number of Classes:  2
##
## Levels:
##  -1 1
```

Then we predicted on our test set and obtained the following truth table:

predict/truth	-1	1
-1	96	3
1	2	70

```
## [1] 0.02923977
```

As previously explained, we associate level 1 with malignant tumors and level -1 with benign tumors. According to the truth table, there are 3 test samples with true category malignant predicted as benign, whereas there are 2 test samples with true category benign predicted as malignant. The misclassification rate is 0.02924.

Radial Kernel SVM

We tune the hyperparameter cost and gamma through cross validation considering a range of values of cost from 0.001 to 1000, gamma from 0.5 to 4, and selected the optimal cost = 1, gamma = 0.5

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost gamma
##     1   0.5
##
## - best performance: 0.07044872
##
## - Detailed performance results:
##     cost gamma      error dispersion
## 1  1e-01   0.5 0.09307692 0.03545438
## 2  1e+00   0.5 0.07044872 0.02835600
## 3  1e+01   0.5 0.07057692 0.04092698
## 4  1e+02   0.5 0.07307692 0.04197726
## 5  1e+03   0.5 0.07307692 0.04197726
## 6  1e-01   1.0 0.35012821 0.09475806
## 7  1e+00   1.0 0.09064103 0.02936491
## 8  1e+01   1.0 0.09820513 0.03438883
## 9  1e+02   1.0 0.10320513 0.03811214
## 10 1e+03   1.0 0.10320513 0.03811214
## 11 1e-01   2.0 0.35012821 0.09475806
## 12 1e+00   2.0 0.11076923 0.03351610
## 13 1e+01   2.0 0.11326923 0.03746702
## 14 1e+02   2.0 0.11326923 0.03746702
```

```
## 15 1e+03   2.0 0.11326923 0.03746702
## 16 1e-01   3.0 0.35012821 0.09475806
## 17 1e+00   3.0 0.18141026 0.12507978
## 18 1e+01   3.0 0.17128205 0.12520558
## 19 1e+02   3.0 0.17128205 0.12520558
## 20 1e+03   3.0 0.17128205 0.12520558
## 21 1e-01   4.0 0.35012821 0.09475806
## 22 1e+00   4.0 0.32737179 0.10661475
## 23 1e+01   4.0 0.31230769 0.11398118
## 24 1e+02   4.0 0.31230769 0.11398118
## 25 1e+03   4.0 0.31230769 0.11398118
```

```
##
## Call:
## best.tune(METHOD = svm, train.x = diagnosis_binary ~ concave_points_mean +
##     area_se + compactness_se + radius_worst + concavity_worst + concave_points_worst +
##     compactness_mean + compactness_se + concavity_se + concavity_mean +
##     radius_mean + concave_points_se + area_worst + compactness_worst +
##     compactness_mean * concave_points_mean, data = cancer_train,
##     ranges = list(cost = c(0.1, 1, 10, 100, 1000), gamma = c(0.5,
##         1, 2, 3, 4)), kernel = "radial")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##
## Number of Support Vectors:  186
##
##  ( 106 80 )
##
##
## Number of Classes:  2
##
## Levels:
##  -1 1
```

Then we predicted on our test set and obtained the following truth table:

predict/truth	-1	1
-1	96	3
1	2	70

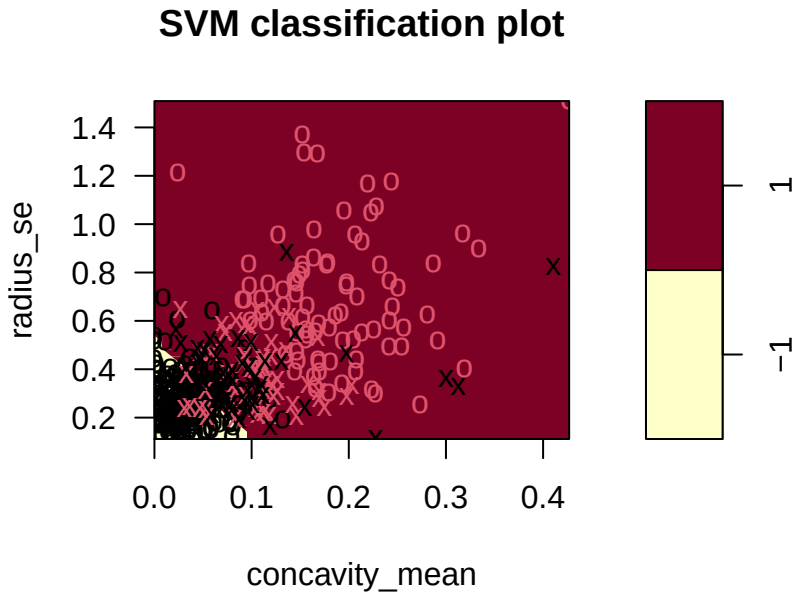
```
## [1] 0.02923977
```

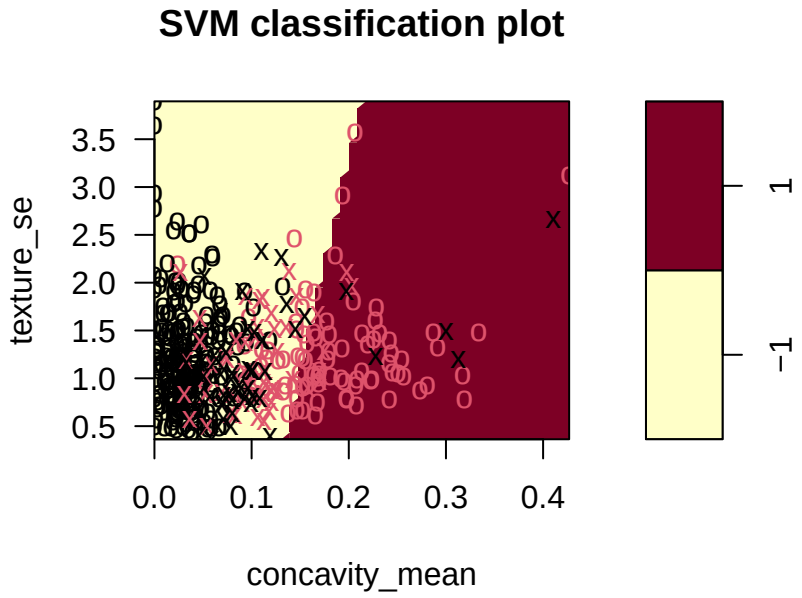
According to the truth table, there are 3 test samples with true category malignant predicted as benign, whereas there are 2 test samples with true category benign predicted as malignant. The misclassification rate is 0.02924 which is equal to that of the linear kernel. This suggests that the two classes are likely to be linearly separable so that we can find a separating hyperplane using the linear kernel, therefore the linear kernel SVM suffices for our particular dataset.

SVM Visualization

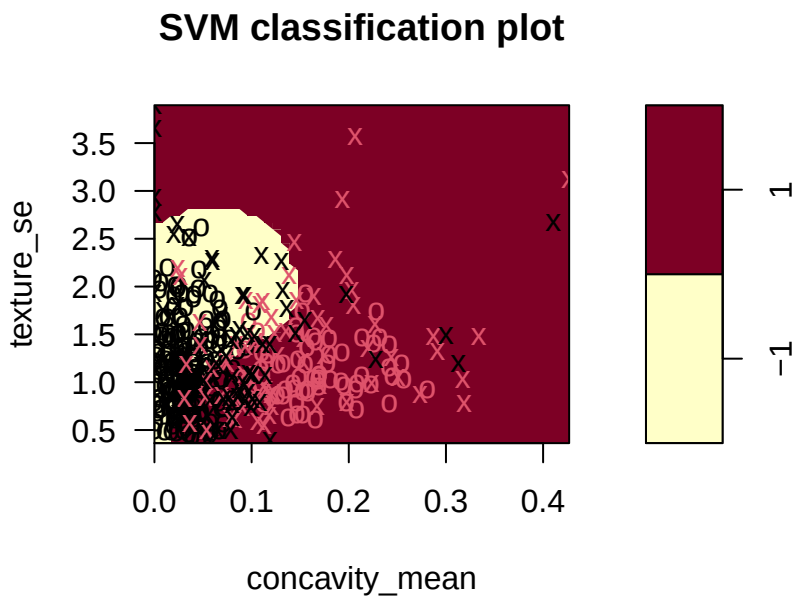
We visualize the decision boundaries of the linear and radial SVM kernels plotting pairs of predictors, taking `concavity_mean` and `texture_se` as examples:

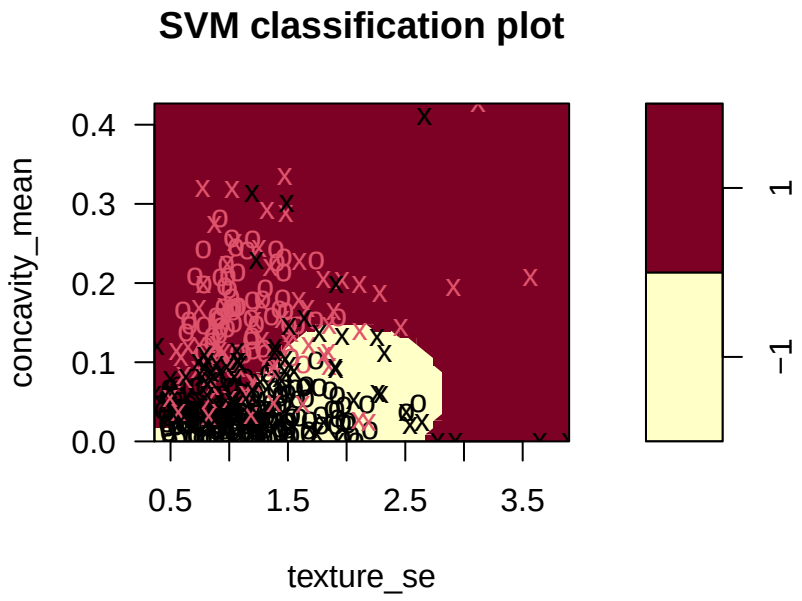
Linear





Radial

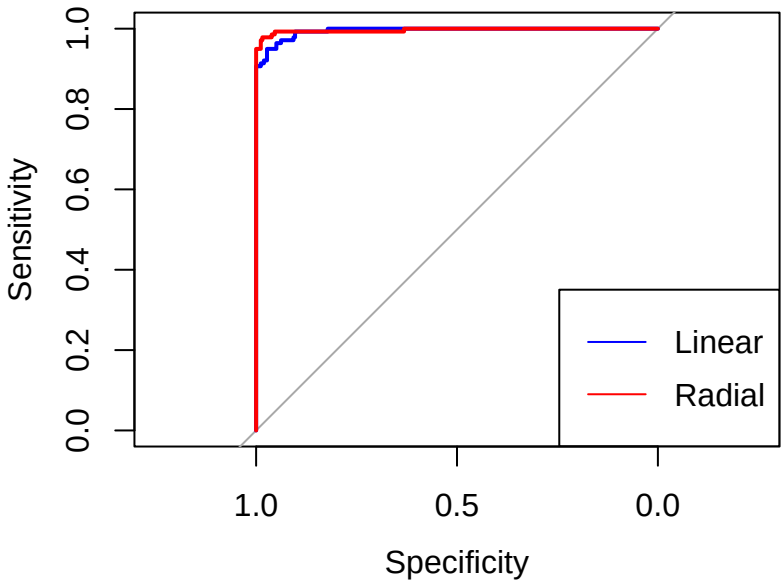

SVM classification plot
texture_se
concavity_mean
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.1
0.2
0.3
0.4
1
−1


SVM classification plot
concavity_mean
texture_se
0.4
0.3
0.2
0.1
0.0
0.5
1.5
2.5
3.5
1
−1

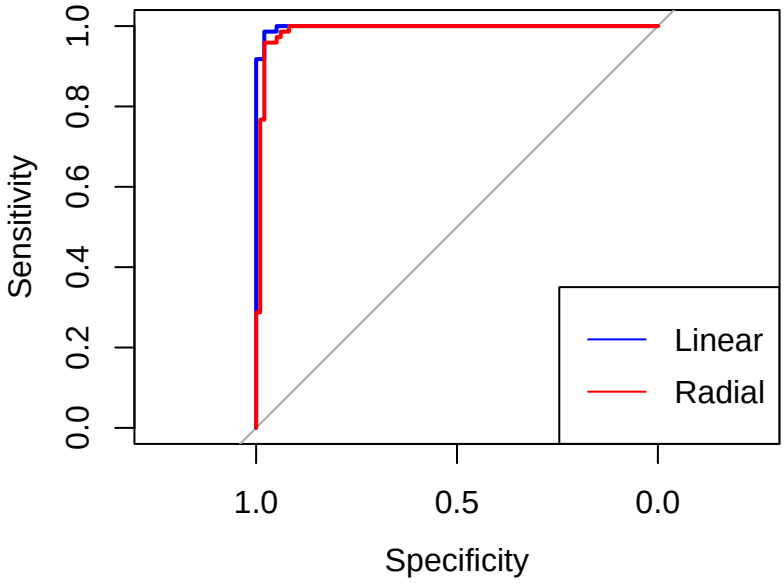
ROC

We visualize the ROC curves of linear and radial kernel SVMs on test and train data respectively.

On train data:



On test data:



Even though the radial kernel fits the training data more closely due to its higher complexity, the linear kernel performs better on the test data (since the data is likely to be linearly separable as explained above), we decided to select the model with the linear kernel, which will result in lower variance and similarly low bias according to the ROC curves above.

Model 3: Random Forest

Now, we are going to build a Random Forest model to compare with Lasso and SVM. We chose Random Forest because it uses bootstrap sampling and feature sampling, so it is not affected by multicollinearity that much since it is picking different set of features for different models and of course every model sees a different set of data points. What's more, it is easy and straightforward to interpret a tree model, so it can help us understanding the importance of the variables than the LASSO Penalized Logistic model.

First, We select variables by running a preliminary random forest model with all the variables, to rank their importance.

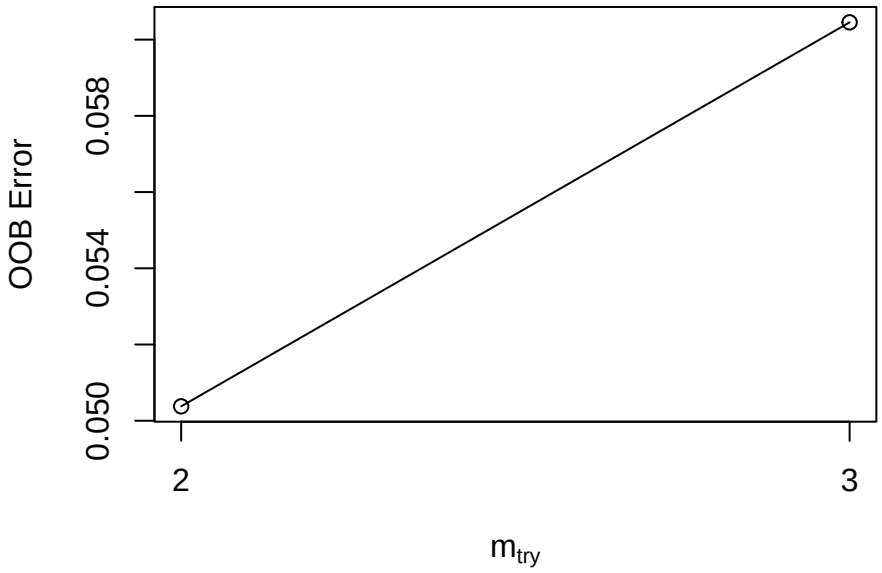
	MeanDecreaseGini
radius_mean	5.0070158
texture_mean	3.1072100
perimeter_mean	7.5925204
area_mean	9.2068964
smoothness_mean	1.3670226
compactness_mean	1.4568072
concavity_mean	12.6313700
concave_points_mean	18.2770412
symmetry_mean	0.7370356
fractal_dimension_mean	0.9026707
radius_se	2.4673699
texture_se	0.7408013
perimeter_se	3.0568472
area_se	7.1559654
smoothness_se	1.0192210
compactness_se	1.0399584
concavity_se	1.1397170
concave_points_se	1.0101496
symmetry_se	0.9229467
fractal_dimension_se	1.1310419
radius_worst	21.7267522
texture_worst	4.0267581
perimeter_worst	18.8197812
area_worst	15.9260537
smoothness_worst	2.6418060
compactness_worst	3.1437736
concavity_worst	7.1555395
concave_points_worst	23.6956861
symmetry_worst	2.0017696
fractal_dimension_worst	1.1758773

After testing different variables based on their important, and taking into accounts the collinearity issue we discussed in the EDA section, we decided to select the following variables as the predictors: concave_points_worst, area_worst, perimeter_worst, radius_worst, concave_points_mean, perimeter_mean, concavity_worst, area_se.

We chose mtry=2, because after tuning mtry, we found that mtry=2 has the lowest OOB error.

```
## mtry = 2  OOB error = 5.04%
## Searching left ...
## Searching right ...
```

```
## mtry = 3     OOB error = 6.05%
## -0.2 0.01
```



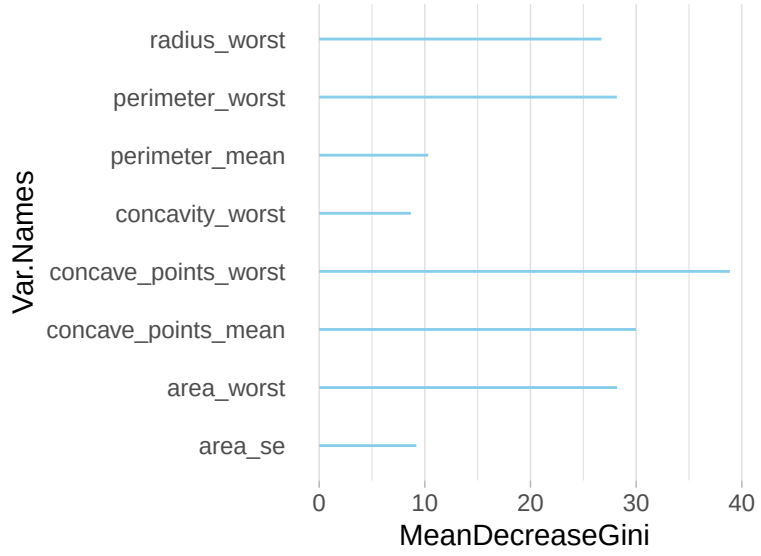
We chose the number of tree to be 500. The number of trees should be chosen carefully, since a high performance of the individual models might lead to overfitting when the number of trees is very high. However, taking 50 trees caused a lower accuracy than taking 500 trees. Therefore, this higher number of trees is chosen.

```
##
## Call:
##  randomForest(formula = diagnosis_binary ~ concave_points_worst +      area_worst + perimeter_worst +
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 5.04%
## Confusion matrix:
##     -1   1 class.error
## -1 251   7  0.02713178
## 1   13 126  0.09352518

## Confusion Matrix and Statistics
##
##           Reference
## Prediction -1  1
##         -1 96  5
##         1   2 68
##
##                Accuracy : 0.9591
```

```
##                  95% CI : (0.9175, 0.9834)
##     No Information Rate : 0.5731
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9159
##
##  Mcnemar's Test P-Value : 0.4497
##
##             Sensitivity : 0.9796
##             Specificity : 0.9315
##          Pos Pred Value : 0.9505
##          Neg Pred Value : 0.9714
##              Prevalence : 0.5731
##          Detection Rate : 0.5614
##    Detection Prevalence : 0.5906
##       Balanced Accuracy : 0.9555
##
##        'Positive' Class : -1
##
```

The RF model yields a 96% accuracy for the testing set.



From the plot we can see that the most important predictors are: concave_points_worst, concave_points_mean, area_worst, perimeter_worst, and radius_worst. Interestingly, perimeter_worst has high gini coefficient, but perimeter_mean ranks second from the last.

Conclusion & Future Work

The result of the LASSO Penalized Logistic model shows that radius, compactness, concavity, and concave points are the most important variables. The result of the RF model suggest that radius, perimeter, area, and concave points are the most important variables. The difference between two results is not a concern here due to the collinearity among predictors. We can conclude that concave points, nuclear perimeter and compactness was highly significant in differentiating hyperplasia from carcinoma.

Both of the models highlights concave points as the most crucial variable. Concavity is the severity of concave portions of the contour. A high concavity means that the boundary of the cell nucleus has indentations, and thus is rather rough than smooth. Concave points counted the number of concave portions of the contour of the cell nucleus. If the contour contains one real cell, the added concave point separates one cell into two parts. Therefore, the higher number the concave points, the more irregular the shape is. Moreover, the odds of malignant diagonosis increases as perimeter, area, radius increases.

What's more, it is worth-noticing that our LASSO Penalized Logistic model has a strong interaction term `compactness_mean:concave_points_mean`, which suggests that there is a strong positive relationship between compactness and concave points. This means that as the number of concave points grows, the value of perimeterˆ2/area also increases, which means that the cell nucleus is deformed, swelled and expanded.

Our results is correspondent with the clinical evidence. The ductal carcinoma cells showed higher values for nuclear area, perimeter, diameter, compactness, and concave points when compared to fibroadenomas, fibrocystic disease, and hyperplasia. In the present study, the size related parameters (area, perimeter, diameter, concave points and compactness) of the nucleus were appropriate parameters to differentiate between benign lesions and infiltrative ductal carcinoma of the breast. These parameters showed significant differences between the benign breast lesions and carcinoma.

The following two pictures shows the cytological features for benigh Fibroadenoma and Fibrocystic disease. When the tumor is benign, the cells shows mild variation in size and shape.

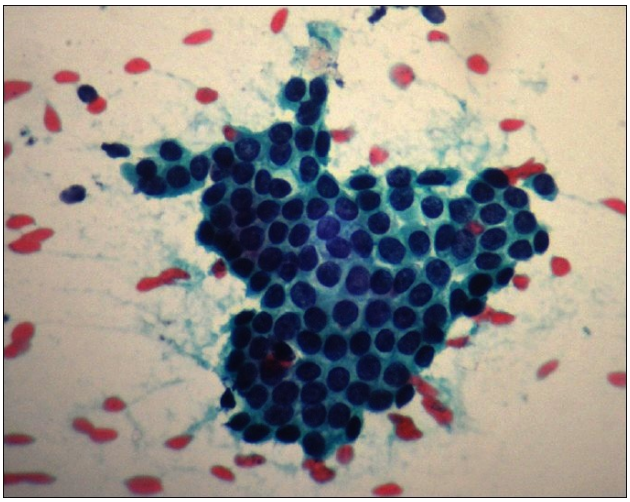


Figure 8: Benign (1)

However, when the cancer is malignant carcinoma, we can observe loosely arranged clusters of ductal epithelial cells showing nuclear pleomorphism, increased nuclear cytoplasmic ratio, nuclear indentations, and hyperchromatic nucleus. We can see from the picture that the cell nucleus is clearly deformed and almost bursting, which confirmed our interpretation to the interaction term.

When applying these 3 models to the test set, the LASSO-Penalized Logistic model yields an accuracy of 98.8%. Both linear and Radial SVM model yields an accuracy around 97%, and the Random Forest model has an accuracy of 96%. Since breast cancer is a vital disease, false negative is much more dangerous than false positive. The false negative rate in the both the linear and radial SVM is: 0.041, and in the RF is: 0.068. Therefore, we decided that SVM is a better classification model than RF in this case.

One limitation of our study is that although there are 30 variables, there are only 3 categories (mean, se, worst), and many of them are highly correlated. In future works, it would be interesting to see if other exogenous variables such as the patients' age, sex, and health conditions have potential effects in the accuracy of breast cancer diagnosis.

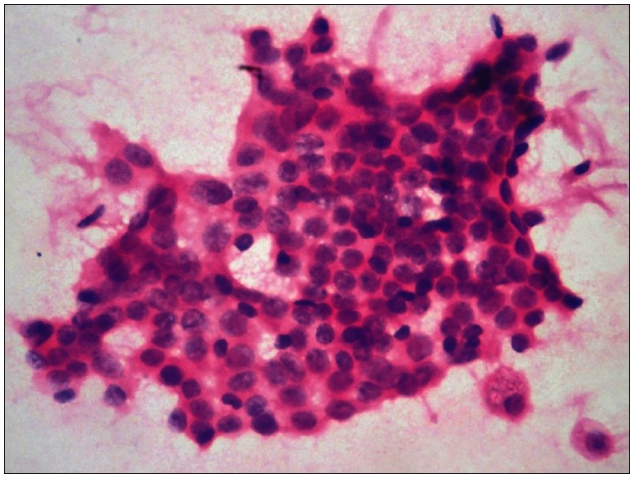


Figure 9: Benign (2)

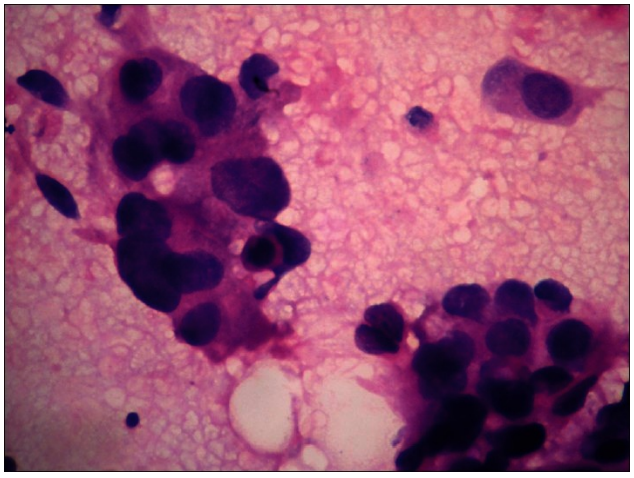


Figure 10: Malignant

Citations

Centers for Disease Control and Prevention. (2022, September 26). Breast cancer. Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/breast/index.htm

Mayo Foundation for Medical Education and Research. (2022, December 14). Breast cancer. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470

Narasimha A, Vasavi B, Kumar HM. Significance of nuclear morphometry in benign and malignant breast aspirates. Int J Appl Basic Med Res. 2013 Jan;3(1):22-6. doi: 10.4103/2229-516X.112237. PMID: 23776836; PMCID: PMC3678677.

Wolberg WH, Street WN, Mangasarian OL. Importance of nuclear morphology in breast cancer prognosis. Clin Cancer Res. 1999;5:3542–8.

Appendix

Variable Name	Description
diagnosis	The diagnosis of breast tissues (M = malignant, B = benign)
radius_mean	mean of distances from center to points on the perimeter
texture_mean	standard deviation of gray-scale values
perimeter_mean	mean size of the core tumor
area_mean	mean value of area
smoothness_mean	mean of local variation in radius lengths
compactness_mean	mean of perimeter^2 / area - 1.0
concavity_mean	mean of severity of concave portions of the contour
concave points_mean	mean for number of concave portions of the contour
symmetry_mean	mean value of correspondence in size
fractal_dimension_mean	mean for "coastline approximation" - 1
radius_se	standard error for the mean of distances from center to points on the perimeter
radius_se	standard error for the mean of distances from center to points on the perimeter
texture_se	standard error for standard deviation of gray-scale values
perimeter_se	standard error for mean size of the core tumor
area_se	standard error of area
smoothness_se	standard error for local variation in radius lengths
compactness_se	standard error for perimeter^2 / area - 1.0
concavity_se	standard error for severity of concave portions of the contour
concave_points_se	standard error for number of concave portions of the contour
symmetry_se	standard error of the symmetry measure
fractal_dimension_se	standard error for "coastline approximation" - 1
radius_worst	"worst" or largest mean value for mean of distances from center to points on the perimeter
texture_worst	"worst" or largest mean value for standard deviation of gray-scale values
perimeter_worst	"worst" or largest mean value for size of the core tumor
area_worst	"worst" or largest mean value for area
smoothness_worst	"worst" or largest mean value for local variation in radius lengths
compactness_worst	"worst" or largest mean value for perimeter^2 / area - 1.0
concavity_worst	"worst" or largest mean value for severity of concave portions of the contour
concave_points_worst	"worst" or largest mean value for number of concave portions of the contour
symmetry_worst	"worst" or largest mean value for correspondence in size
fractal_dimension_worst	"worst" or largest mean value for "coastline approximation" - 1