Breast Cancer Analysis

Amey Hari Madane

1. Abstract

The dataset is collected from 699 breast cancer biopsies performed at the University of Wisconsin, using fine needle aspiration cytology. It examines nine different characteristics in a scale of one through ten for cell size and shape, which dictates the healthiness of the cells. The key objective is to see if these variables alone can correctly classify the tissue sample as benign or malignant. Assuming these women are a random subset showing symptoms of breast cancer, the project will try to study this dataset in detail. It will involve the fitting of a logistic regression model by best subset selection and the implementation of the Lasso penalty method. In addition, Linear Discriminant Analysis will be employed. The idea is to assess the dependability of these features in separating benign from malignant tissue of the breast. A successful outcome could significantly impact breast cancer diagnosis, aiding in more informed treatment decisions.

2. Data Exploration

First of all, the exploration and preparation of data was done by changing the variables from factors to numerical representations. Then class variables were changed into numerical where 'benign' was represented as 0 and 'malignant' as 1. Interestingly, there were 16 missing attributes in the 'Bare.Nuclei' column. In order to handle it, the rows with missing attributes were deleted. This resulted in the dataset being left with 444 observations for benign and 239 observations as malignant.

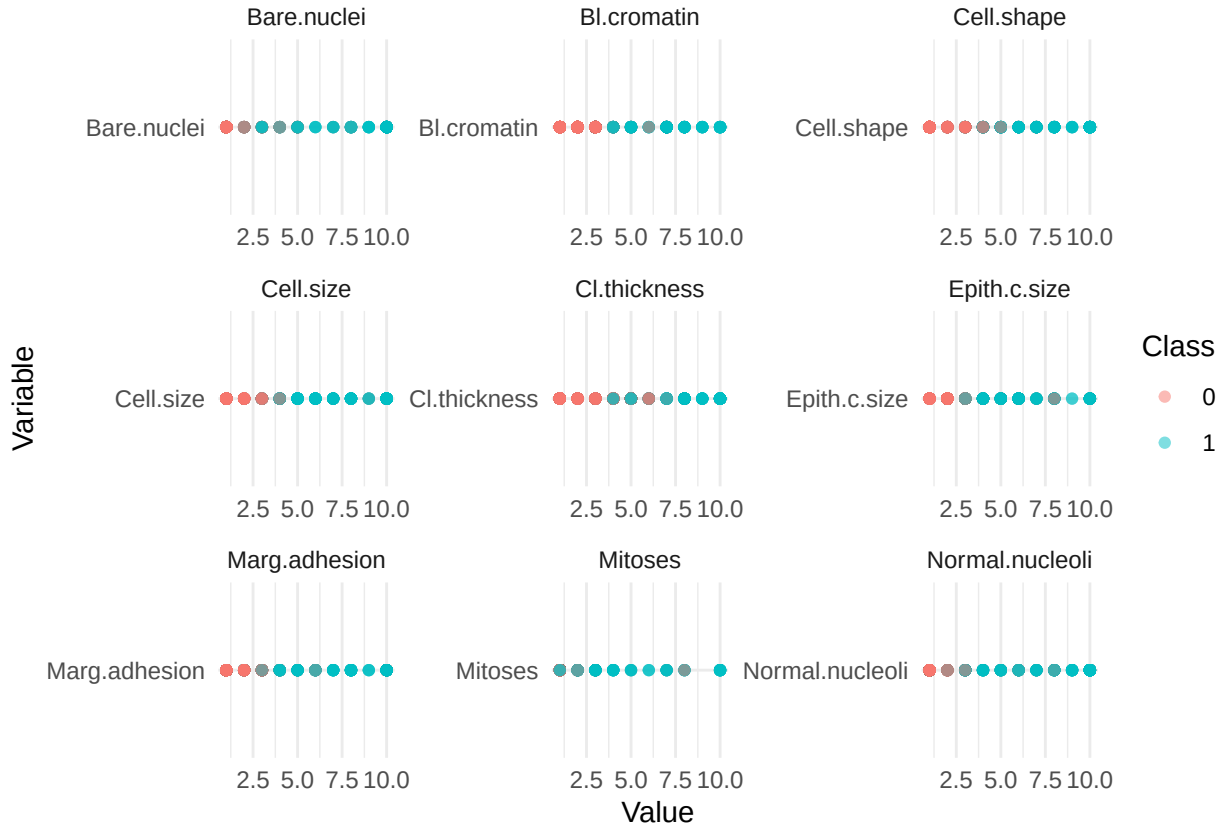
2.1 Data Summary

```
##   Cl.thickness      Cell.size        Cell.shape     Marg.adhesion
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.00
##  1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.00
##  Median : 4.000   Median : 1.000   Median : 1.000   Median : 1.00
##  Mean   : 4.442   Mean   : 3.151   Mean   : 3.215   Mean   : 2.83
##  3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000   3rd Qu.: 4.00
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.00
##   Epith.c.size     Bare.nuclei      Bl.cromatin     Normal.nucleoli
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.00
##  1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 1.00
##  Median : 2.000   Median : 1.000   Median : 3.000   Median : 1.00
##  Mean   : 3.234   Mean   : 3.545   Mean   : 3.445   Mean   : 2.87
##  3rd Qu.: 4.000   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 4.00
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.00
##     Mitoses
```

```
##  Min.   : 1.000
##  1st Qu.: 1.000
##  Median : 1.000
##  Mean   : 1.603
##  3rd Qu.: 1.000
##  Max.   :10.000
```

The summary gives an overview of the range, dispersion, and central tendencies of each predictor variable, hence giving a good view of how each varies within the data. Features such as 'Cl.thickness' have higher means with larger ranges, which may indicate large variability in the dataset. Mitoses has the lowest mean and variability across the dataset.

2.2 Facet Grid Scatterplot



The scatterplot matrix shows that the clear separation between the two classes in response variables proves there is a clear distinction. However, weaker separations were found within normal.nucleoli, bare.nuclei, marg.adhesion, and epith.c.size; this indicates that class values in these particular variables are overlapped. Most importantly, cell.size and cell.shape can be said to hold a very strong positive relationship; thus, as one variable increases, so does the other. These findings give valuable insight into the class separations and interrelationships among the predictor variables in this dataset.

2.3 Covariance matrix

```
##                 Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## Cl.thickness        7.956694  5.554922   5.508800      3.941776     3.283363
## Cell.size           5.554922  9.395113   8.310604      6.207468     5.134708
## Cell.shape          5.508800  8.310604   8.931615      5.872385     4.799947
## Marg.adhesion       3.941776  6.207468   5.872385      8.205717     3.786179
## Epith.c.size        3.283363  5.134708   4.799947      3.786179     4.942109
## Bare.nuclei         6.096061  7.725660   7.774099      7.000264     4.744656
## Bl.cromatin         3.826365  5.673248   5.383535      4.691541     3.366253
## Normal.nucleoli     4.598758  6.730824   6.550081      5.274024     4.268107
## Mitoses             1.715289  2.447021   2.284936      2.079140     1.851150
##                 Bare.nuclei Bl.cromatin Normal.nucleoli  Mitoses
## Cl.thickness       6.096061    3.826365        4.598758 1.715289
## Cell.size          7.725660    5.673248        6.730824 2.447021
## Cell.shape         7.774099    5.383535        6.550081 2.284936
## Marg.adhesion      7.000264    4.691541        5.274024 2.079140
## Epith.c.size       4.744656    3.366253        4.268107 1.851150
## Bare.nuclei       13.277695    6.075403        6.499229 2.141645
## Bl.cromatin        6.075403    6.001013        4.977439 1.468652
## Normal.nucleoli    6.499229    4.977439        9.318772 2.294262
## Mitoses            2.141645    1.468652        2.294262 3.002160
```

The covariance matrix shows the interaction between the predictor variables in the data. From the matrix, one can observe that the covariances between 'Cell.size', 'Cell.shape', and 'Bare.nuclei' are much higher, indicating that there is a greater positive relationship within these features. In other words, if one of these three variables increases, then the others will grow accordingly; thus, there could be some multicollinearity between them. Whereas very low covariance values, such as between 'Cl.thickness', 'Marg.adhesion', 'Epith.c.size', and other variables, are indicative of weaker associations or less linear dependence within these specific features. Finally, Mitoses has a weak positive relationship with all the variables. The elements on the diagonal of the matrix represent the variance of the variables, therefore showing the spread/variability of each predictor variable individually.

2.4 Correlation matrix

```
##                 Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## Cl.thickness       1.0000000 0.6424815  0.6534700     0.4878287    0.5235960
## Cell.size          0.6424815 1.0000000  0.9072282     0.7069770    0.7535440
## Cell.shape         0.6534700 0.9072282  1.0000000     0.6859481    0.7224624
## Marg.adhesion      0.4878287 0.7069770  0.6859481     1.0000000    0.5945478
## Epith.c.size       0.5235960 0.7535440  0.7224624     0.5945478    1.0000000
## Bare.nuclei        0.5930914 0.6917088  0.7138775     0.6706483    0.5857161
## Bl.cromatin        0.5537424 0.7555592  0.7353435     0.6685671    0.6181279
## Normal.nucleoli    0.5340659 0.7193460  0.7179634     0.6031211    0.6289264
## Mitoses            0.3509572 0.4607547  0.4412576     0.4188983    0.4805833
## Class              0.7147899 0.8208014  0.8218909     0.7062941    0.6909582
##                 Bare.nuclei Bl.cromatin Normal.nucleoli   Mitoses     Class
```

```
## Cl.thickness      0.5930914   0.5537424       0.5340659 0.3509572 0.7147899
## Cell.size         0.6917088   0.7555592       0.7193460 0.4607547 0.8208014
## Cell.shape        0.7138775   0.7353435       0.7179634 0.4412576 0.8218909
## Marg.adhesion     0.6706483   0.6685671       0.6031211 0.4188983 0.7062941
## Epith.c.size      0.5857161   0.6181279       0.6289264 0.4805833 0.6909582
## Bare.nuclei       1.0000000   0.6806149       0.5842802 0.3392104 0.8226959
## Bl.cromatin       0.6806149   1.0000000       0.6656015 0.3460109 0.7582276
## Normal.nucleoli   0.5842802   0.6656015       1.0000000 0.4337573 0.7186772
## Mitoses           0.3392104   0.3460109       0.4337573 1.0000000 0.4234479
## Class             0.8226959   0.7582276       0.7186772 0.4234479 1.0000000
```

Correlation Between Response and Predictor Variables:

The 'Class' variable correlates strongly positively with all the predictor variables: 'Cl.thickness', 'Cell.size', 'Cell.shape', 'Marg.adhesion', 'Epith.c.size', 'Bare.nuclei', and 'Bl.cromatin', whose magnitude falls between 0.71 to 0.82. This means that as these variables increase, there is a tendency to associate more with the 'Class' variable, which may indicate that these features are important to predict benign or malignant status.The 'Mitoses' variable has a weaker correlation of 0.42 with the 'Class' variable than other predictors, suggesting it has a relatively less strong relationship in predicting the class.

3. Fitting a logistic regression model.

.

```
##
## Call:
## glm(formula = y_train ~ ., family = "binomial", data = CancerTrain_data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.00202    0.36608  -2.737 0.006197 **
## Cl.thickness     1.09545    0.44071   2.486 0.012932 *
## Cell.size        0.50367    0.75026   0.671 0.502018
## Cell.shape       0.81736    0.76142   1.073 0.283064
## Marg.adhesion    0.91998    0.41029   2.242 0.024943 *
## Epith.c.size     0.09204    0.44533   0.207 0.836265
## Bare.nuclei      1.51514    0.38957   3.889 0.000101 ***
## Bl.cromatin      1.39072    0.52504   2.649 0.008078 **
## Normal.nucleoli  0.45675    0.38858   1.175 0.239817
## Mitoses          0.89052    0.61064   1.458 0.144747
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 708.985  on 546  degrees of freedom
## Residual deviance:  80.026  on 537  degrees of freedom
```

```
## AIC: 100.03
##
## Number of Fisher Scoring iterations: 8
```

The maximum likelihood estimates of the regression coefficients are therefore

$\hat{\beta}_0 = -1.002, \hat{\beta}_1 = 1.095, \hat{\beta}_2 = 0.503, \hat{\beta}_3 = 0.817, \hat{\beta}_4 = 0.919, \hat{\beta}_5 = 0.092, \hat{\beta}_6 = 1.515, \hat{\beta}_7 = 1.390, \hat{\beta}_8 = 0.456, \hat{\beta}_9 = 0.890$

The p-value for Cl.thickness, Marg.adhesion, Bare.nuclei and Bl.cromatin is less than 0.05 If we look at the table produced by the summary function we see that a number of the variables have very large p-values meaning that, individually, they contribute very little to a model which contains all the other predictors.Inclusion of more predictors than are necessary can inflate the variance of the parameter estimators leading to a deterioration in predictive performance.

4. Best Subset Selection in logistic regression

In the earlier model it is observed that some of the features do not have any significant effect on the model's output. Therefore to find the optimal model we apply different feature selection techniques.
We can apply best subset selection using BIC using the bestglm package.

BIC Subsets

```
##    Intercept Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## 0       TRUE        FALSE     FALSE      FALSE         FALSE        FALSE
## 1       TRUE        FALSE      TRUE      FALSE         FALSE        FALSE
## 2       TRUE        FALSE      TRUE      FALSE         FALSE        FALSE
## 3       TRUE         TRUE      TRUE      FALSE         FALSE        FALSE
## 4       TRUE         TRUE      TRUE      FALSE         FALSE        FALSE
## 5*      TRUE         TRUE      TRUE      FALSE          TRUE        FALSE
## 6       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE
## 7       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE
## 8       TRUE         TRUE      TRUE       TRUE          TRUE        FALSE
## 9       TRUE         TRUE      TRUE       TRUE          TRUE         TRUE
##    Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses logLikelihood      BIC
## 0        FALSE       FALSE           FALSE   FALSE    -354.49257 708.9851
## 1        FALSE       FALSE           FALSE   FALSE     -91.55415 189.4128
## 2         TRUE       FALSE           FALSE   FALSE     -60.82666 134.2622
## 3         TRUE       FALSE           FALSE   FALSE     -52.36195 123.6372
## 4         TRUE        TRUE           FALSE   FALSE     -46.59161 118.4010
## 5*        TRUE        TRUE           FALSE   FALSE     -43.08541 117.6931
## 6         TRUE        TRUE           FALSE    TRUE     -41.44240 120.7115
## 7         TRUE        TRUE            TRUE    TRUE     -40.31109 124.7533
## 8         TRUE        TRUE            TRUE    TRUE     -40.03435 130.5043
## 9         TRUE        TRUE            TRUE    TRUE     -40.01295 136.7659
```

BIC: Penalizes complexity more than AIC and often selects smaller models compared to AIC. .
Here BIC has selected best model with 5 predictors to be the best.

4.1 Best subset selection with BIC

```
##
## Call:
## glm(formula = y_train ~ ., family = "binomial", data = Cancer_data_red_BIC)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.0505     0.3242  -3.241 0.001193 **
## Cl.thickness    1.4338     0.4116   3.483 0.000495 ***
## Cell.size       1.5398     0.5093   3.023 0.002502 **
## Marg.adhesion   1.0085     0.3907   2.581 0.009847 **
## Bare.nuclei     1.6313     0.3701   4.408 1.04e-05 ***
## Bl.cromatin     1.4802     0.4845   3.055 0.002250 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 708.985  on 546  degrees of freedom
## Residual deviance:  86.171  on 541  degrees of freedom
## AIC: 98.171
##
## Number of Fisher Scoring iterations: 8
```

The maximum likelihood estimates of the regression coefficients are

$\hat{\beta}_0 = -1.05, \hat{\beta}_1 = 1.433, \hat{\beta}_2 = 1.539, \hat{\beta}_3 = 1.008, \hat{\beta}_4 = 1.631, \hat{\beta}_5 = 1.480$

The model summary clearly indicates a robust association between the predictor and response variables. Each variable exhibits positive coefficients, signifying a positive relationship. Additionally, all variables demonstrate p-values below 0.05, indicating a strong statistical significance and reinforcing the presence of a compelling positive correlation among the variables.

This model has selected Cl.thickness, Cell.size, Marg.adhesion, Bare.nuclei and Bl.cromatin variables and rest all are dropped from the model. These variables showed strong positive correlation with Class variable in the earlier correlation matrix. 4 of the variables except Cell.size had p-values less than 0.05 in earlier simple logistic regression model.

4.2 Test error

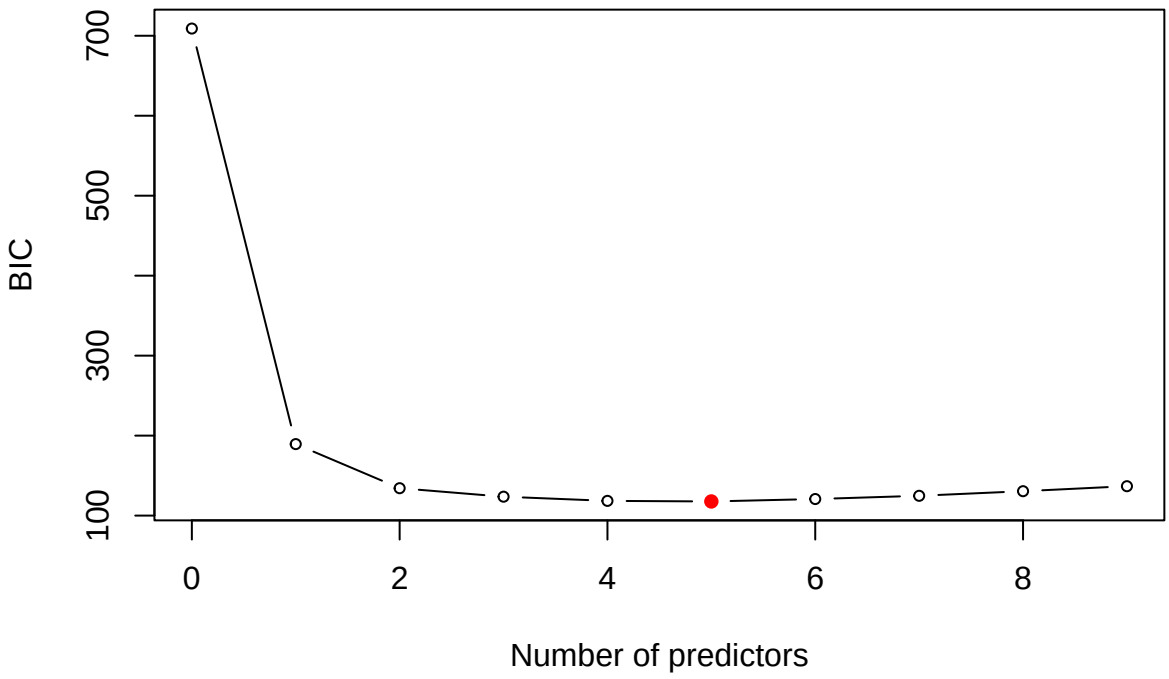
```
## [1] "Confusion matrix of subset selection with BIC"
```

```
##          Predicted
## Observed  0  1
##        0 87  2
##        1  3 44
```

```
## [1] "Test error for best subset selection with BIC is: "
```

```
## [1] 0.03676471
```

The test error for best subset selection with BIC is 3.67%. This error is less compared to the first regression model.



5. Regularized Logistic regression with Lasso penalty

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                          s1
## (Intercept)     -1.0040075
## Cl.thickness     0.8698421
## Cell.size        0.5183630
## Cell.shape       0.6819046
## Marg.adhesion    0.4613698
## Epith.c.size     0.1023056
## Bare.nuclei      1.2511103
## Bl.cromatin      0.8237321
## Normal.nucleoli  0.3699991
## Mitoses          0.2181063
```

At the optimal solution none of the variables drop out of the model

5.1 Test error

```
##          Predicted
## Observed  0  1
##        0 87  2
##        1  5 42
```

```
## [1] "Test error for logistic regression with Lasso is: "
```

```
## [1] 0.05147059
```

The test error (5.1%) is slightly higher for the model fitted with the LASSO penalty. Therefore of the two models, it seems that the model fitted without penalty performs better, based on this particular partition of the data into training and validation sets.

6. Bayes classifier for Linear Disciminant Analysis

All the variables have been used in the LDA model.

```
## Call:
## lda(y_train ~ ., data = data.frame(X_train))
##
## Prior probabilities of groups:
##         0         1
## 0.6489945 0.3510055
##
## Group means:
##   Cl.thickness  Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei
## 0   -0.5127239 -0.6107099 -0.6119085    -0.5198588   -0.5104219  -0.6057326
## 1    0.9480052  1.1291771  1.1313933     0.9611973    0.9437489   1.1199743
##   Bl.cromatin Normal.nucleoli    Mitoses
## 0  -0.5598119      -0.5316167 -0.3250101
## 1   1.0350689       0.9829372  0.6009302
##
## Coefficients of linear discriminants:
##                         LD1
## Cl.thickness     0.44530128
## Cell.size        0.45703871
## Cell.shape       0.28054047
## Marg.adhesion    0.12117563
## Epith.c.size     0.10777290
## Bare.nuclei      0.97393852
## Bl.cromatin      0.29214794
## Normal.nucleoli  0.32313497
## Mitoses         -0.03166195
```

Above model shows, Prior probabilities of groups:
64.89% belongs to benign cancer and 35.10% belongs to malignant cancer.
Group means
It shows the class wise average (standardised) values for each predictor variables. This helps in comparing how the average values of variables varies between two class. A large difference in average values suggests good seperation between the classes.

6.1 Test error

```
##          Predicted
## Observed  0  1
##        0 87  2
##        1  7 40
```

```
## [1] "Test error for logistic regression with LDA is: "
```

```
## [1] 0.06617647
```

The test error for the linear discriminant analysis model is 6.6% which is highest among all the methods implemented on the Breast Cancer dataset.

7. Conclusion

Among these five variants of logistic regression for the Breast Cancer data set to predict the type of cancer, namely benign or malignant, the model using best subset selection method using BIC had turned out to be the best. This model showed an error rate of 3.6%, reflecting its accuracy of prediction.

This selected logistic regression model comprises five predictor variables: Cl.thickness, Cell.size, Marg.adhesion, Bare.nuclei, and Bl.cromatin. These variables showcase a notably strong positive correlation with the target class variable. Moreover, they exhibit statistical significance with p-values less than 0.05, further affirming their relevance in the prediction process

Including more than five variables in the logistic regression model-in particular, utilizing all variables in methods like Lasso or LDA-results in higher errors. This indicates that the extra variables outside of the best subset or the full set of variables are not adding much value to enhancing the predictive power of the model.

These additional variables add nothing valuable to the model in terms of predicting type, whether it be benign or malignant. Because of this, their inclusion tends to cause noise and irrelevant information, leading to an increase in error rates without corresponding improvement in predictive accuracy. The best performance of the model is, therefore, achieved in considering a limited set of five predictor variables that are strongly associated with the target class variable, while statistical significance and error rate are low.