A Computational Model for the Early Detection of Primary Breast Cancer Using Differential Expression Analysis of Microarray Data from Serum microRNA

Krishna Sudhir

Abstract: Breast cancer remains one of the leading causes of cancer death in women as the risk of developing breast cancer reaches as high as 12% in the average American woman. With limited non-invasive treatment options, and long recovery time frames, doctors face difficulty in administering appropriate treatments in accordance to the nature of the type of cancer. Different types of breast cancer call for differing treatment approaches in order to minimize risk of recurrence. Having foresight on a patient's genetic predisposal to the type of cancer will allow patients with a good prognosis to avoid unnecessary aggressive chemotherapy. This research project endeavors to develop computational models that will predict a patient's chances of having breast cancer by use of statistical modeling using microarray data measuring microRNA expression patterns. These models utilize the expression values of top 6 microRNAs to predict the chances of breast cancer being present in the patient, represented by one binary dependent variable. The main statistical methods employed were logistic regression and artificial neural networks, and the model accuracy for both methods was measured by use of an RMSE score analysis and prediction error analyses. The final models were built using the training dataset and these found to be reliably predictive of the outcome when validated with the test dataset. This promises scope for future expansive study and in turn could lead to building health care applications that could leverage such computational models. As a proof of concept, these models were exported to Python serialized objects and built a web application to demonstrate the efficacy of using computational models in cancer diagnosis.

Key words: Breast Cancer, microRNA, Microarray, Differential Expression Analysis, Logistic Regression, NCBI GEO, R, Computational Sciences, Statistical Modeling, Neural Networks

1 Introduction

Several studies have been conducted to find the genetic linkage to the occurrence of breast cancer, which is a major form of cancer in the developed world. There are groundbreaking studies which explored various computational methods for early detection of breast cancer using miRNA expression analysis [1]. Could computational methods be employed to create a simple model to detect the occurrence of breast cancer using microarray expression values of select miRNAs found in blood serum? This study focuses on creating such a model using differential expression analysis, logistic regression and neural networks.

Breast cancer, being one of the most common cancers in women, is the second leading cause of cancer deaths in women. Currently, the average risk of developing breast cancer in American women is approximately 12%, with incidence rates rising slightly in recent years. Death rates for breast cancer victims have dropped by 40% in the past 4 decades, however with an incidence probability of 1 in 8, breast cancer continues to be significantly relevant in current medical research [2].

Ductal Carcinoma In Situ is a type of breast cancer found in the lining of breast milk ducts. This type of breast cancer, however, is noninvasive and do not spread past the breast duct tissue unless left untreated. Invasive ductal carcinomas will begin growing in the milk ducts and continue to invade the surrounding tissue. Up to 80% of all breast cancer diagnoses are invasive ductal carcinomas, and the most prevalent type of breast cancer in men. IDCs are also known as infiltrative ductal carcinomas. Triple negative breast cancer is used to indicate that the most common receptors to induce breast cancer (estrogen, progesterone, and HER-2) are not found in the tumor by testing the cancer cells. Hormonal therapies that target common receptors such as estrogen, progesterone, and HER-2 are ineffective and chemotherapy is normally used to treat triple negative breast cancer. Inflammatory breast cancer (IBC) is an aggressive breast cancer that begins in the lymph vessels in the breast after penetrating the skin. Due to its rapid growth and aggressive nature, survival rates for inflammatory breast cancer are significantly low. Surgery, radiation therapy, chemotherapy and hormone treatments are all used to treat IBC. Metastatic breast cancer will spread to other parts of the body such as the lungs, liver, bones and even the brain.

Cancer cells invade nearby healthy cells in order to replicate more cancer cells. Those cancer cells then infiltrate the lymphatic or circulatory system and further, travel through blood and lymph vessels to invade other parts of the body. Once cancer cells become embedded within capillaries in various parts of the body, they will divide and repeat replication, forming smaller tumors or micrometastases.

Chemotherapy comes in four different forms. Curative chemotherapy aims to eradicate all cancer cells. Adjuvant chemotherapy aims to eradicate all surrounding tumor cells post surgery in order to prevent recurrence. Neoadjuvant chemotherapy aims to remove tumors too large prior to surgery in order to lessen the invasiveness of surgery. Palliative chemotherapy is performed when all tumor cells are no longer able to

be removed and is used to alleviate symptoms or slow the progress of the disease, however is temporary.

Hormone therapy is used to treat breast cancers that are hormone sensitive. These types of cancers are referred to as estrogen receptor positive (ER positive) and progesterone receptor positive (PR positive) cancers. Hormone therapy includes medications that block selected hormones from attaching to cancer cells, medications to cease estrogen production, and surgery/medication to stop hormone production in the ovaries. Hormone therapy may be administered prior to, or post-surgery in order to decrease chances of the cancer returning, or to slow the progression of the cancer. [3]

1.1 Medical Diagnosis

Test and procedures used to diagnose breast cancer may include breast examinations, mammograms, breast ultrasounds, biopsies, and breast magnetic resonance imaging. During a breast examination, the breasts and lymph nodes of the patient are manually checked by the physician in order to feel for lumps or abnormalities. This may be done in a physician's office but is most commonly performed as an at home examination. Mammograms take an X-ray of the breast and screen the breast for abnormalities. Ultrasounds may be used to determine the nature of the abnormal growth in the breast, with regards to whether it is a solid mass or a fluid-filled cyst.

A biopsy is the only definitive procedure used to diagnose breast cancer. During a biopsy, a core of tissue from the area of interest will be extracted by the help of an X-ray or other imaging technologies. The cells are then sent to a laboratory for further analysis in order to identify the nature of the cancer to decide the treatment options thereafter.

1.2 Predictive Diagnosis

As of now, active areas of research include predictive studies for breast cancer. Studies have been able to correlate the effect of exercise, abnormal weight gain or weight loss, as well as an individual's diet on risk. Additionally, environmental impacts on breast cancer risk factors has been an area of active research in recent years. Predictive studies have included the effect of common gene mutations in victims, the impact of gene variants on incidence rates.

Recent research has been able to correlate mutations in BRCA2, a gene that is responsible for suppressing cell growth, to the onset of breast cancer. The gene is responsible for producing tumor suppressing proteins as well as proteins that repair damaged DNA in order to prevent instability of the genetic material in cells. Inherited mutations in BRCA1 and this gene, is correlated with an increased lifetime risk of developing breast or ovarian cancer. Women with harmful mutations to both BRCA1 and BRCA2 genes have an increased

chance of developing breast cancer. Mutations in BRCA1 and BRCA2 have been linked with ovarian cancer, Fallopian tube cancer, prostate cancer, and peritoneal cancer in addition to breast cancer [5].

1.3 Micro RNA

In recent bioinformatics research, MicroRNA, otherwise known as miRNA expression levels have been identified as potential predictors of breast cancer. MicroRNA are small noncoding RNA that, after transcription, will regulate gene expression. Most miRNA have not shown to have any specific biological functions, however, research has reported miRNA involvement in cell division, cell differentiation, cell dedifferentiation inside mammary glands in adult females. [4]

miRNA plays a significant role in gene silencing in cells during cell development and specialization. MicroRNAs turn off genes by inactivating messenger RNAs in order to prevent genetic translation into proteins. It is through activity by microRNAs following cells until their death that allows for cells to perform their respective functions. Inactivation of microRNAs may result in a range of diseases such as cancer and heart disease.

Genes coding for microRNAs are located in genes in the DNA that produce regulatory mRNA. The transcripts produced from these genes are known as primary microRNAs, which will eventually become the final regulatory microRNAs after various methods of processing. Primary MicroRNAs are processed and cut by enzymes prior to their release into the cytoplasm through nuclear pores. Precursor miRNAs are double stranded until reacting in the cytoplasm with other enzymes to produce a single stranded RNA Induced Silencing Complex (RISC) which are then to be guided to their respective target to silence genes. The RISC may inactive the mRNA by cutting the mRNA or through inhibition of translation by preventing ribosome subunits from binding to the mRNA. The two types of inhibition will regulate protein synthesis, thus silencing the gene. [6]

1.4 MicroArray Applications in Breast Cancer

DNA microarray technology permits a researcher to investigate which genes are expressed or repressed in a cell or tissue. When a gene is expressed, it is first transcribed into mRNA which is then isolated and converted into complementary strands of DNA known as cDNA by the enzyme reverse transcriptase. The cDNA derived from the tumor tissue is separated from the cDNA derived from the normal tissue. DNA Microarrays are used to compare the two cDNA samples, each DNA microarray containing 6000 or more DNA sequences. Both cDNA samples are mixed prior to their insertion into the microArray and cDNA complementary to the nucleotides found in the microArray will hybridize and unbound cDNA are to then

be washed off. Patterns of hybridization are detected through various scans and expression is analyzed. Expression levels of the microRNA is indicated by different fluorescent lightings. Through this, researchers are then able to identify, and isolate the microRNA genes expressed.

For breast cancer, scientists have identified multiple genes whose expression levels vary at significant levels between genes of patients with breast cancer recurrence and patients with no breast cancer recurrence. This data has proven to be useful in clinical decision-making for treating patients depending on their prognosis. [7]

1.5 GEO Database

NCBI's GEO (Gene Expression Omnibus) is an international public repository containing various microarray datasets accessible for the research community. GEO aims to provide microarray, next-generation sequencing, and other forms of functional genomics data submitted and utilized by the research community [8]. GEO aims to provide a versatile database in order to store genomic data. Geo offers a location for researchers to deposit genomic data to allow other users to locate, review and download studies and gene expression profiles for independent research endeavors. The data available in GEO includes mRNA, genomic DNA, and protein abundance, as well as non-array techniques such as serial analysis of gene expression (SAGE), and mass spectrometry proteomic data. In GEO, the following are three entity types that may be supplied by users: Platforms, Samples, and Series.

A Platform record describes the list of elements that are detected and quantified in its respective experiment. Each Platform record is assigned a unique GEO accession number written as "GPLxxx". A Sample record describes the conditions under which an individual Sample was handled, the way in which it was manipulated, and the measurement of each element derived from it. Each Sample record is assigned a unique GEO accession number, written as "GSMxxx". A Series record defines a set of related Samples that are all a part of a group, the way in which the samples are related, and how they are ordered. A Series provides a center of interest and a description of the experiment as a whole. Series records may also contain tables describing extracted data, summary conclusions, or analyses. Each Series record is assigned a unique number (GSExxx).

With the advancement of computational technology and abundant availability of microarray expression data of miRNAs, could a computational model be devised that will help physicians for detecting early the occurrence of breast cancer?

1.6 Data

The data that was used for this study was freely provided by National Center for Biotechnology Information (NCBI). Named as GSE73002, this dataset contains 3,966 observations of cancer and non-cancer patients. Each observation contains miRNA expression values of 2,576 micro RNAs. [10]

1.7 Logistic Regression

Logistic regressions are a type of machine learning algorithms that are used as a statistical method to model the probability of a certain event existing. This variable is represented as a binary variable. However, this methodology may be extended to modeling the presence of several events and may collaboratively predict a nonbinary variable outcome. Logistic regressions use logistic functions to model the binary dependent variable, and when used in predictive modeling, will produce only two types of outcomes. Some examples of such outcomes are "male/female", "pass/fail", and "alive/dead". The two possible values of the dependent variable are labeled as "0" or "1".

The logarithm of odds for a given value, either a 1 or 0, is a linear combination of multiple independent variables used to predict the outcome ("predictors"). While the dependent variable must only contain discrete values (only two values), the independent variable may contain discrete or continuous variables (any real value). The model will predict the probability of the outcome variable, and this will be a number between 0 and 1. This prediction of the outcome is by use of a logistic function, which will convert the logarithm of odds into a probability.

1.8 Artificial Neural Networks

Neural networks are a set of algorithms, loosely modeled after neural networks found in the human brain, that are designed to recognize patterns and interpret data through labeling or clustering raw input. The numerical patterns they recognize are contained in vectors, into which all data must be translated. Neural networks help to cluster and classify data that is stored and managed by the programmer. Neural networks classify data when upon which a labeled dataset to train on. Neural networks use algorithms for reinforcement learning, classification and regression.

Artificial Neural Networks (ANN) are typically organized in layers that contain interconnected 'nodes' each of which holding an "activation function" within it. The "input layer" receives inputs / observations into the neural network, which will then process the data in the "hidden layers" of the neural network. Processing is done using a system of "connections" or weights assigned to each independent variable that is applied to

the data. The hidden layers will then display the answer as an output, known as the "output layer".

During the training process, the output or the predicted outcome is compared with the actual outcome. Any resulting error is fed back into the network in order to tune and correct the weights applied at each individual node. This process is known as back-propagation. By tuning the weights, the model may produce outcomes at lower error rates, and make the model more reliable. This training process is expected to continue until the error of predicted outcomes is no longer able to be decreased.

2 Methods

2.1 Data Distribution

The expression values of individual miRNAs have been visually analyzed using box plots.

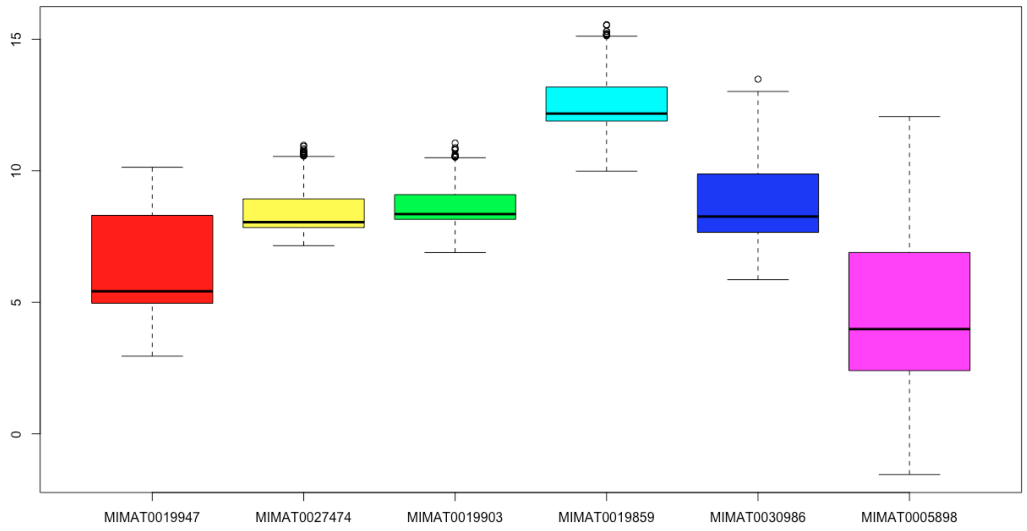


Figure 1: Box plot of miRNA Distribution

2.2 Computational Approach

1. The GSE data for GSE73002 from NCBI GEO site using GEOQuery library was imported into the R Studio session

2. Import the GSE data for GSE73002 from NCBI GEO site using GEOQuery library. This will return GEO Expression Set data structure.

3. The dataset was separated into a “Control” (subjects having no cancer) and “Test” (breast cancer subjects) group.

4. A log2 transformation was done as required on the miRNA expression values

5. Based on the above annotation, a model matrix was created using the 'model.matrix' method of the 'stat' library in R.

6. Using lmFit method of limma library, a linear model was created for each gene in the given series of arrays.

7. A contrast matrix was created based on the expression values using makeContrasts method of limma library.

8. The model statistics were found using the empirical Bayes from the microarray linear model fit using the ‘eBays’ method of the ‘limma’ library.

9. The top 250 genes from the linear model fit was identified and only the required fields were extracted. This provided a list of miRNAs sorted with respect to the differential expression values.

10. The dataframe from the Expression Dataset was created and returned by the ‘getGEO’ method from step # 1. All the miRNA IDs were eliminated from the data frame with the exception of the top 100 miRNAs.

11. The dependent variable “Y” was derived based on the diagnosis description (has cancer or no cancer).

12. The data frame was split into a training dataset (65%) and test dataset (35%).

13. A model using top 20 miRNA names was created and a logistic regression was run on the training dataset. Based on the significance of the coefficients generated, the predictor list was iteratively refined to a final list of six miRNAs. A final regression model was built using the six miRNAs as predictors and their significance was measured. The accuracy of the trained model was validated using the test dataset.

14. Using the six significant attributes, a neural network model was created using a multilayer perceptron classifier (MLP Classifier). The model was trained using the training set and validated using the test dataset.

15. Various statistical analysis methods were run on the two models created based on the predicted results to assess the accuracy of the model.

16. The logistic regression model as well as the MLP classifier model were imported as Python objects. These objects were used in a web application that will predict the chances of breast cancer presence based on a user input of expression values of the aforementioned miRNAs.

2.3 Differential Expression Analysis

Differential Expression Analysis is used to select two or more groups, say, "control" and "test" groups, and compare the average difference in expression values between the groups by statistical means. An open source suite of libraries called "Bio-Conductor" [9] was employed to conduct a differential expression analysis for this study. The miRNAs were ordered by decreasing differential expression values and the top 20 candidates were used to build the model. The expression data structure returned by Bioconductor::getGEO was converted to a data frame in order to perform a logistic regression analysis on the data frame using the candidate miRNA expression values as predictor variables.

2.4 Training and Test Data

The dataset was divided into a test set and training set. 65% of the records were used to train the model and 35% of the records was used for testing the model for prediction accuracy.

2.5 The Models

Since the dependent variable was a boolean variable (diagnosis is either 1 or 0, representing whether the individual has breast cancer or not, respectively), a logistic regression approach was considered for modeling. A binary logistic regression is a method used to model data with a dichotomous dependent variable, or a binary outcome. Logistic regression can be mathematically expressed as in the equation 1:

$$\ln(\frac{p}{(1-p)}) = a + \beta * X + e \tag{1}$$

There were 20 candidate microRNAs chosen to be examined to determine which of them were potentially significant. The expression values were visually validated using Box plots. No missing values or outliers were observed.

In the next step, a multivariate logistic regression model with all the significant variables was tried out for accuracy. The 'stat' library of R provides out of the box functions to run logistic regression as well as predict outcomes based on the model created. The function 'glm' (Generalized Linear Model) with an option to use binomial method, would conveniently return the logistic model. Using this, a multivariate logistic regression was performed with "diagnosis" as the dependent variable and the candidate miRNAs as the predictors.

The 'predict' method of the stat library returns the predicted values. By comparing the predicted outcomes with the actual outcomes in the test dataset, one can determine the accuracy of the model. Various statistical

methods were utilized to determine the predictive accuracy of the model.

After the predictor variables (miRNA expressions) were refined to a most significant set of 6 miRNAs, a neural network with 3 layers of 13 neurons each was built. The model was fit using training dataset and was validated with the test dataset. The prediction accuracy was calculated by analyzing the quality of the predicted outcomes.

2.6 Model Benchmarks

There were a number of methods to compare the accuracy of the models with one another. Using Accuracy, RMSE (Root Mean Square Error), Concordance vs Discordance ratio, misclassfication error, confusion matrix, Specificity vs Sensitivity, ROC curve and KS Plot the models were compared and classified based on their performance. Each of these methods helped in determination of the better model.

2.6.1 'p' values

The 'p' values were analyzed for each of the predictor variables to evaluate their significance. The 'p' values were reported to have low values as expected for many of the predictor variables.

2.6.2 Overfitting and Underfitting

The model prediction accuracy was assessed on the training data set and testing data set independently and found that there was no significant overfitting nor underfitting.

2.6.3 Accuracy

Prediction error was calculated by taking the mean of all the differences between predicted outcomes and actual outcomes. Accuracy was measured using Equation 2.

$$PredictionAccuracy = 1 - PredictionError \tag{2}$$

2.6.4 RMSE

RMSE is a measure of how accurate the model is by squaring the differences of predictions and actuals and then taking the square root of their average. Lower the RMSE the better the model is. The equation 3

shows how to calculate RMSE.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(\frac{d_i - f_i}{\sigma_i}\Big)^2} \tag{3}$$

2.6.5 Confusion Matrix

Confusion matrix gives a simple but useful interpretation of the model's accuracy. It gives a grid of true positives, false positives, true negatives and false negatives. This gives an idea where the model performs well and where it doesn't. From the library *InformationValue* the function *confusionMatrix()* was used to calculate the Confusion Matrix of the models that were studied.

2.6.6 Kolmogorov-Smirnov Plot

Kolmogorov-Smirnov test can be used for determining the performance of classification models. It tests the degree of separation between the positive and negative distributions in the models predicted output. From the library *InformationValue* the function *ks_plot()* was used to plot the results of K-S chart.

3 Results and Discussion

In this study, the GSE data was imported and a differential expression analysis was conducted using Bioconductor libraries. Of the top 250 miRNAs which found to be of statistical significance, 20 miRNAs were picked and used in a logistic regression model and a neural network model.

3.1 Differential Expression Analysis

Using GEOQuery and limma libraries, a contrast matrix was formed and the top 250 miRNAs were selected, that have the most differential in their expression values in the microarray data. A sample of the top miRNAs and corresponding differential values are given in figure 2.

3.2 Model Performance

Various statistical methods were employed to determine the quality and accuracy of the models. These models demonstrated impressive results on all of these measured metrics. The summary of the results from

Figure 2: Top Table from Differential Expression Analysis

adj.P.Val	P.Value	t	B	logFC	miRNA_ID	miRNA_ID_LIST
0	0	147.57584	3701.574	3.7808550	MIMAT0005951	hsa-miR-1307-3p
0	0	145.54898	3654.756	3.4297593	MIMAT0019947	hsa-miR-4783-3p
0	0	120.18101	3035.896	2.3483367	MIMAT0031000	hsa-miR-8073
0	0	113.84262	2869.605	2.7235053	MIMAT0019071	hsa-miR-4532
0	0	111.83491	2815.771	1.3633946	MIMAT0027474	hsa-miR-6787-5p
0	0	110.28653	2773.910	1.4689439	MIMAT0027623	hsa-miR-6861-5p
0	0	106.16139	2660.921	1.9841140	MIMAT0022943	hsa-miR-1233-5p
0	0	104.63013	2618.437	2.0039772	MIMAT0019757	hsa-miR-4675
0	0	-104.27247	2608.180	-2.4043564	MIMAT0004508	hsa-miR-92a-2-5p
0	0	103.17269	2577.444	1.1765288	MIMAT0019903	hsa-miR-4758-5p
0	0	100.27441	2495.454	1.7380264	MIMAT0027432	hsa-miR-6766-5p
0	0	99.14499	2463.869	1.0132662	MIMAT0027412	hsa-miR-6756-5p
0	0	-96.57634	2390.216	-0.9146768	MIMAT0005582	hsa-miR-1228-5p
0	0	94.62012	2333.591	1.9212353	MIMAT0027438	hsa-miR-6769a-5p
0	0	94.40170	2327.240	1.6834710	MIMAT0019034	hsa-miR-4419b
0	0	93.92582	2313.384	0.9794007	MIMAT0027504	hsa-miR-6802-5p
0	0	93.23987	2293.364	2.5746492	MIMAT0018943	hsa-miR-4428
0	0	92.66468	2276.534	3.1228492	MIMAT0004694	hsa-miR-342-5p
0	0	-91.80680	2251.361	-1.2215663	MIMAT0027468	hsa-miR-6784-5p
0	0	-90.69103	2218.271	-0.7875179	MIMAT0019229	hsa-miR-3940-5p
0	0	90.56840	2214.428	1.5815708	MIMAT0019859	hsa-miR-4734
0	0	90.50995	2212.702	1.8262104	MIMAT0022947	hsa-miR-1238-5p
0	0	90.06170	2199.674	2.2566057	MIMAT0004948	hsa-miR-885-3p
0	0	89.41319	2180.681	2.5997150	MIMAT0030986	hsa-miR-8059
0	0	88.87689	2164.541	2.5253591	MIMAT0015030	hsa-miR-3156-5p

Logistic Regression are given in figure 3.

3.3 Model Summary (Logistic Regression)

The model summary below shows that there are many miRNAs, which play significant roles in predicting the chance of breast cancer.

3.4 Confusion Matrix (Logistic Regression)

The confusion matrix (figure 4) shows that the model predicts the outcome with impressive accuracy. It mis-predicted 7 benign cases as breast cancer and 21 breast cancer cases as benign. It predicted 1326 cases accurately.

Figure 3: Logistic Regression Model Summary

```
Call:
glm(formula = fmla2, family = binomial(), data = df_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8143  -0.0479  -0.0163   0.0009   2.7463

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -49.0030     8.2800  -5.918 3.25e-09 ***
MIMAT0019947   3.0146     0.3966   7.601 2.94e-14 ***
MIMAT0027474   5.0072     0.8302   6.032 1.62e-09 ***
MIMAT0019903   5.1323     0.9918   5.175 2.28e-07 ***
MIMAT0019859  -3.4087     0.6143  -5.549 2.88e-08 ***
MIMAT0030986  -2.3968     0.4177  -5.739 9.54e-09 ***
MIMAT0005898   1.3460     0.1737   7.749 9.23e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3039.39  on 2417  degrees of freedom
Residual deviance:  162.82  on 2411  degrees of freedom
AIC: 176.82

Number of Fisher Scoring iterations: 10
```

3.5 AUROC and Kolmogorov-Smirnov Test

The selected model was studied further for more benchmarking methods such as plotting ROC curve as well as Kolmogorov–Smirnov plots (KS Plot). ROC curve shows 99.24% of area under the curve. The ROC Plot (Figure 5) shows that the model has very good predictive abilities. Additionally the KS Plot (Figure 6) shows that the model can predict much better than the random prediction at the mean level.

3.6 Confusion Matrix and Classification Report (Neural Network)

The confusion matrix and the classification report for the neural network show (figure 7 and figure 8 respectively) very high recall and support for the neural network model, which means that the model has a very high predictive accuracy of nearly 99%

Figure 4: Confusion Matrix (Logistic Regression)

Confusion Matrix (Logistic Regression)		
	No cancer	Has Cancer
No cancer	897	21
Has cancer	7	431

4 Web Application Prototype

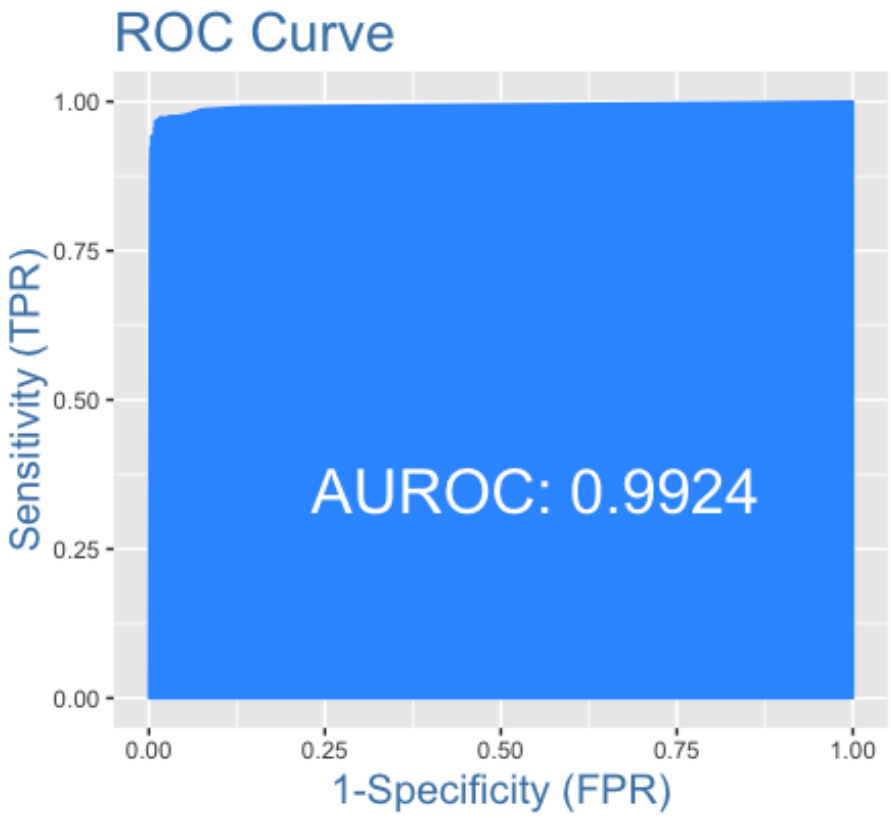
The logistic regression and neural network models were trained and exported to Python objects so that those could be used in a web application prototype. The prototype application enables the healthcare user to enter the microarray expression values for the selected miRNAs and the models would predict if the patient may have breast cancer or not. The user also have the option to "dry-run" the tool by selecting rows from a sample set of data. The same data points can be passed through either Logistic Regression or Neural Network model. Figure 9 shows a snapshot of the web application in action.

5 Conclusions

There were a number of findings from this study that were of significance. A model with 20 predictor variables was found to be having very good predictive ability.

- MicroRNA expression values have notable predictive capability.
- In practical terms, by using microRNA expression patterns in a patient, physicians can adequately identify a patient as at risk for developing breast cancer through the extent of expression in select microRNAs, and thereafter plan precautionary medical intervention.
- The model accuracy could be improved by having more observations in the underlying data for the model to train and evaluate. Collecting more data from physicians would potentially improve the statistical significance of the findings of this study as well.
- Furthermore, the data should be adjusted according to the slight multicollinearity between independent variables as was observed in this study. This would eliminate any data skewness that might limit the model's predictive capabilities.

Figure 5: AUROC Plot (Logistic Regression)



- Logistic regression and neural network are some of the most widely used methods of modeling and prediction. Newer algorithms and more cleaner data could assist physicians and other medical professionals to diagnose patients even more reliably.
- Differential expression values of the following micro RNAs seem to have significant correlation with the occurrence of breast cancer: 1. MIMAT0019947 (hsa-miR-4783-3p), 2. MIMAT0027474 (hsa-miR-6787-5p), 3. MIMAT0019903 (hsa-miR-4758-5p), 4. MIMAT0019859 (hsa-miR-4734), 5. MIMAT0030986 (hsa-miR-8059) and 6. MIMAT0005898 (hsa-miR-1246).

This analysis only scratched the surface of the untapped potential of using computational methods to detect cancer. The power and promise of computational sciences with larger and better quality data could be best utilized for improving healthcare and saving human lives.

Figure 6: KS Plot (Logistic Regression)

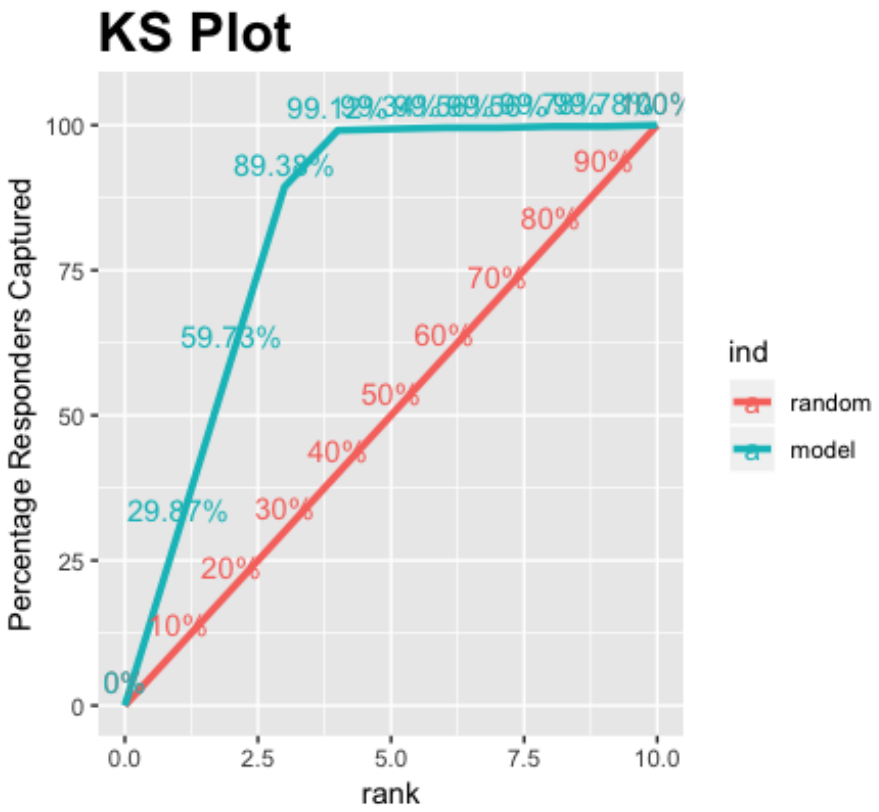


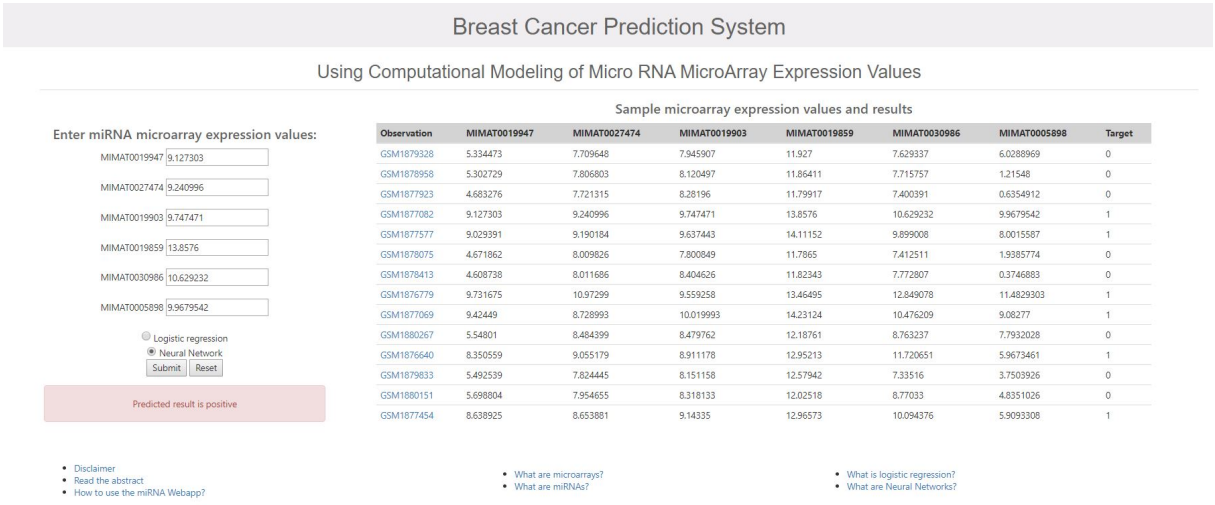
Figure 7: Confusion Matrix (Neural Network)

Confusion Matrix (Neural Networks)		
	No cancer	**Has Cancer**
No cancer	649	2
Has cancer	3	277

Figure 8: Classification Report (Neural Network)

Classification Report (Neural Network)				
	Precision	**Recall**	**f1-score**	**Support**
0	1.00	1.00	1.00	651
1	0.99	0.99	0.99	280
Accuracy			0.99	931
Macro avg	0.99	0.99	0.99	931
Weighted avg	0.99	0.99	0.99	931

Figure 9: Web Application Snapshot [11]



References

[1] Shimomura, A., Shiino, S., Kawauchi, J., Takizawa, S., Sakamoto, H., Matsuzaki, J., ... Ochiya, T. (2016). Novel combination of serum microRNA for detecting breast cancer in the early stage. Cancer science, 107(3), 326–334. doi:10.1111/cas.12880

[2] How Common Is Breast Cancer? Retrieved from https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

[3] Hormone therapy for breast cancer. (2019, February 05). Retrieved from https://www.mayoclinic.org/tests-procedures/hormone-therapy-for-breast-cancer/about/pac-20384943

[4] MicroRNA involvement in mammary gland development and breast cancer Licia Silveri, Gaëlle Tilly, Jean-Luc Vilotte, Fabienne Le Provost Reprod. Nutr. Dev. 46 (5) 549-556 (2006) DOI: 10.1051/rnd:2006026 https://www.ncbi.nlm.nih.gov/pubmed/17107644/

[5] BRCA Mutations: Cancer Risk and Genetic Testing Retrieved from https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet

[6] Bartel DP (January 2004). "MicroRNAs: genomics, biogenesis, mechanism, and function".

Cell. 116 (2): 281–97. doi:10.1016/S0092-8674(04)00045-5. PMID 14744438.: https://www.ncbi.nlm.nih.gov/pubmed/14744438

[7] DNA Microarray Technology Fact Sheet. (n.d.). Retrieved from https://www.genome.gov/about-genomics/fact-sheets/DNA-Microarray-Technology

[8] GEO Overview Retrieved from https://www.ncbi.nlm.nih.gov/geo/info/overview.html

[9] GEOquery. (n.d.). Retrieved from http://bioconductor.org/packages/release/bioc/html/GEOquery.html

[10] Geo Accession viewer. Retrieved from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73002

[11] Breast Cancer Prediction System https://mirna1.herokuapp.com/