Section 1. Medical science

https://doi.org/10.29013/YSJ-22-7.8-3-11

Cheng Annika

PREDICTING BREAST CANCER USING ARTIFICIAL NEURAL NETWORK AND LOGISTIC REGRESSION

Abstract

Objective: This study aims to build a predictive model for breast cancer using artificial neural network and compare its performance to logistic regression model.

Methods: Wisconsin Diagnostic Breast Cancer (WDBC) data was used in this study. Features were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They described characteristics of the cell nuclei present in the image.

All the participants who were eligible were randomly assigned into 2 groups: training sample and testing sample. Two models were built using training sample: artificial neural network and logistic regression. We used these two models to predict the risk of breast cancer in the testing sample. Receiver operating characteristic (ROC) were calculated and compared for these two models for their discrimination capability and a curve using predicted probability versus observed probability were plotted to demonstrate the calibration measure for these two models.

Results: A total of 569 patients were included in this analysis, 357 (62.74%) benign, 212 (37.26%) malignant breast cancer patients.

According to the logistic regression, number of concave portions of the contour and texture (standard deviation of gray-scale values) were at important predictors for malignant breast cancer.

According to this neural network, the top 5 most important predictors were worst area, mean of severity of concave portions of the contour, worst of severity of concave portions of the contour, worst of symmetry, worst of compactness.

For training sample, the ROC was 1.0 for the Logistic regression and 1.0 for the artificial neural network. Artificial neural network performed better clearly. While in testing sample, the ROC was 0.92 for the Logistic regression and 0.99 for the artificial neural network. Artificial neural network had better performance.

As to calibration measure, predictions made by the neural network are (in general) less concentrated around the 45-degree line (a perfect alignment with the line would indicate an ideal perfect calibration) than those made by the Logistic model.

Conclusions: In this study, we identified several important predictors for breast cancer e.g., number of concave portions of the contour, worst of symmetry, worst of compactness. This provided important information for providers and patients for timely accurate diagnosis. We built a predictive model using artificial neural network as well as logistic regression to provide a tool for timely accurate diagnosis. When compared to artificial neural network model, logistic regression had a worse discriminating capability and a better calibration between predicted probability and observed probability.

Keywords: Breast Cancer, Statistics.

1. Instruction

In the United States, breast cancer is the most common cancer in women. Breast cancer is a disease in which cells in the breast grow out of control. There are different kinds of breast cancer. The kind of breast cancer depends on which cells in the breast turn into cancer. In 2014, 236,968 women and 2,141 men in the United States were diagnosed with breast cancer. A total of 41,211 women and 465 men in the United States died from breast cancer in 2014 [1].

Women who have changes in certain breast cancer genes (BRCA1 and BRCA2), or have close relatives with these changes have increased risk of breast cancer [2]. About 12 percent of women in the general population will develop breast cancer sometime during their lives (4). By contrast, according to the most recent estimates, 55 to 65 percent of women who inherit a harmful BRCA1 mutation and around 45 percent of women who inherit a harmful BRCA2 mutation will develop breast cancer by age 70 years [3]. Other risk factors include Ashkenazi Jewish heritage, treatment with radiation therapy to the breast or chest during childhood or early adulthood according to the US Centers for Disease Control and Prevention [4].

This study aims to: 1) examine the predictors of breast cancer; 2) build a predictive model for breast cancer using artificial neural network and compare its performance to logistic regression model.

2. Data and Methods:

Data:

Wisconsin Diagnostic Breast Cancer (WDBC) data was used in this study. Features were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They described characteristics of the cell nuclei present in the image [5].

It is a public data avaialble at: URL: http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29

Models:

Artificial neural netwrok consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. In more practical terms neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data. Using neural networks as a tool, data warehousing firms are harvesting information from datasets in the process known as data mining. The difference between these data warehouses and ordinary databases is that there is actual anipulation and cross-fertilization of the data helping users makes more informed decisions.

A package called "neuralnet" in R was used to conduct neural network analysis. The package neuralnet focuses on multi-layer perceptrons (MLP, Bishop, 1995), which are well applicable when modeling functional relationships. The underlying structure of an MLP is a directed graph, i.e. it consists of vertices and directed edges, in this context called neurons and synapses. The neurons are organized in layers, which are usually fully connected by synapses. In neuralnet, a synapse can only connect to subsequent layers. The input layer consists of all covariates in separate neurons and the output layer consists of the response variables. The layers in between are referred to as hidden layers, as they are not directly observable. Input layer and hidden layers include a constant neuron relating to intercept synapses, i.e. synapses that are not directly influenced by any covariate Neural networks are fitted to the data by learning algorithms during a training process. Neuralnet focuses on supervised learning algorithms.

The backward propagation of errors or backpropagation, is a common method of training artificial neural networks and used in conjunction with an optimization method such as gradient descent. The algorithm repeats a two phase cycle, propagation and weight update. When an input vector is presented to the network, it is propagated forward through the network, layer by layer, until it reaches the output layer. The output of the network is then compared to the desired output, using a loss function, and an error value is calculated for each of the neurons in the output layer. The error values are then propagated backwards, starting from the output, until each neuron has an associated error value which roughly represents its contribution to the original output.

We also used logistic regression models to calculate the predicted risk. Logistic regression is a part of a category of statistical models called generalized linear models, and it allows one to predict a discrete outcome from a set of variables that may be continuous, discrete, dichotomous, or a combination of these. Typically, the dependent variable is dichotomous and the independent variables are either categorical or continuous.

The logistic regression model can be expressed with the formula:

$$ln(P/P-1) = \beta_0 + \beta_1{}^*X_1 + \beta_2{}^*X_2 + \ \dots\ .+ \beta_n{}^*X_n$$

Model evaluation:

The two criteria to assess the quality of a classification model are discrimination and calibration. Discrimination is a measure of how well the two classes in the data set are separated; calibration determines how accurate the model probability estimated is to the true probability. To provide an unbiased estimate of a model's discrimination and calibration, these values have to be calculated from a data set not used in the model building process. Usually, a portion of the original data set, called the test or validation set, is put aside for this purpose. In small data sets, there may not be enough data items for both training and testing. In this case, the whole data set is divided into n pieces, n_1 pieces are used for training, and the last piece is the test set. This process of n-fold cross-validation builds n models; the numbers reported are the averages over all n test sets. An alternative to cross-validation is bootstrapping, a process by which training sets are sampled with replacement from the original data sets.

The discriminatory ability – the capacity of the model to separate cases from non-cases, with 1.0 and 0.5 meaning perfect and random discrimination, respectively– was determined using receiver operating characteristic (ROC) curve analysis. ROC curves are commonly used to summarize the diagnostic accuracy of risk models and to assess the improvements made to such models that are gained from adding other risk factors. Sensitivity, specificity, and accuracy will be also calculated and compared. For all these measures, there exist statistical tests to determine whether one model exceeds another in discrimination ability.

The contingency table can derive several evaluation "metrics" (see infobox). To draw a ROC curve, only the true positive rate (TPR) and false positive rate (FPR) are needed (as functions of some classifier parameter). The TPR defines how many correct positive results occur among all positive samples available during the test. FPR, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test.

A ROC space is defined by FPR and TPR as x and y axes respectively, which depicts relative tradeoffs between true positive (benefits) and false positive (costs). Since TPR is equivalent to sensitivity and FPR is equal to 1 – specificity, the ROC graph is sometimes called the sensitivity vs (1 – specificity) plot. Each prediction result or instance of a confusion matrix represents one point in the ROC space.

The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). The (0,1) point is also called a perfect classification. A random guess would give a point along a diagonal line (the so-called line of no-discrimination) from the left bottom to the top right corners (regardless of the positive and negative base rates). An intuitive example of random guessing is a decision by flipping coins. As the size of the sample increases, a random classifier's ROC point migrates towards the diagonal line. In the case of a balanced coin, it will migrate to the point (0.5, 0.5).

The diagonal divides the ROC space. Points above the diagonal represent good classification results (better than random), points below the line represent poor results (worse than random). Note that the output of a consistently poor predictor could simply be inverted to obtain a good predictor.

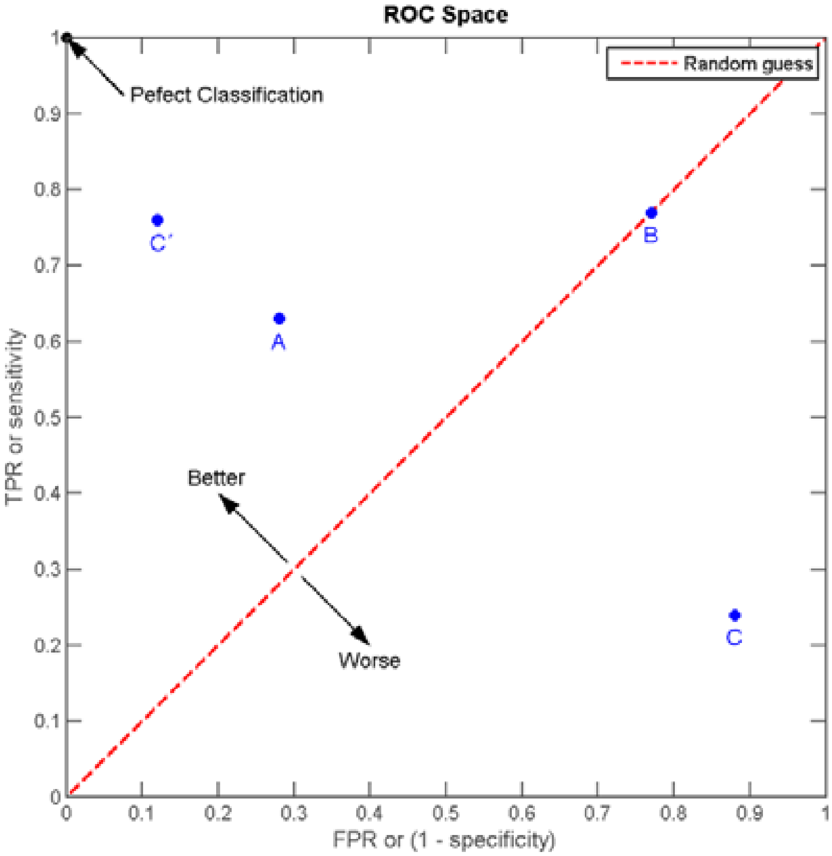

Figure 1.

Calibration is a measure of how close the predictions of a given model are to the real underlying probability. Almost always, the true underlying probability is unknown and can only be estimated retrospectively by verifying the true binary outcome of the data being studied. Calibration thus measures the similarity between two different estimates of a probability. One of the ways to assess calibration is to take the difference between the average observation and the average outcome of a given group as a measure of discalibration. A more refined way to measure calibration requires dividing the sample into smaller groups sorted by predictions, calculating the sum of predictions and sum of outcomes for each group, and determining whether there are any statistically significant differences between the expected and observed numbers by a simple method.

Variables:

The outcome variable is where or not the patients had breast cancer based on variable called Diagnosis (M = malignant, B = benign).

Ten real-valued features are computed for each cell nucleus as below.

Table 1.– Independent variables used in this study

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area – 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension («coastline approximation» – 1)

The mean, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 20 features. In this study, we did not use standard error features in our predictive modeling exercises. All feature values are recoded with four significant digits.

3. Results

A total of 569 patients were included in this analysis, 357 (62.74%) benign, 212 (37.26%) malignant breast cancer patients.

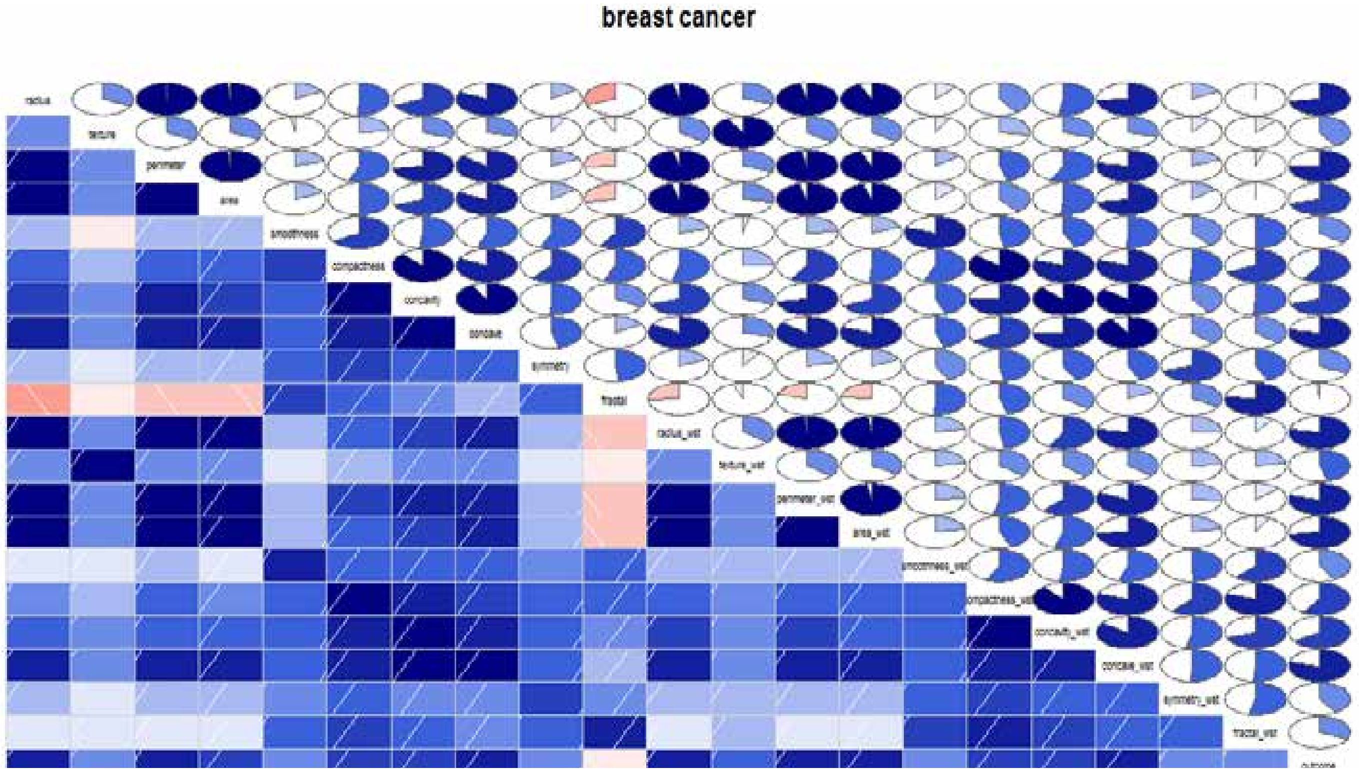

Figure 2. Matrix of correlations between variables

Basically, a corrgram is a graphical representation of the cells of a matrix of correlations. The idea is to display the pattern of correlations in terms of their signs and magnitudes using visual thinning and correlation-based variable ordering. Moreover, the cells of the matrix can be shaded or colored to show the correlation value. The positive correlations are shown in blue, while the negative correlations are shown in red; the darker the hue, the greater the magnitude of the correlation.

According to the logistic regression, number of concave portions of the contour and texture (standard deviation of gray-scale values) were at important predictors for malignant breast cancer.

Table 2.– Logistic Regression for Breast cancer

	Estimate	Std. Error	Z value	Pr(>|z|)	
radius	–4.504	8.840	–0.510	0.610	
texture	0.446	0.233	1.909	0.056	.
perimeter	0.448	1.163	0.385	0.700	
area	–0.007	0.046	–0.147	0.883	
smoothness	73.631	93.831	0.785	0.433	
compactness	–102.930	64.562	–1.594	0.111	
concavity	–3.555	29.443	–0.121	0.904	
concave	159.129	71.754	2.218	0.027	*
symmetry	11.164	27.505	0.406	0.685	
fractal	61.836	241.244	0.256	0.798	
radius_wst	0.789	2.674	0.295	0.768	
texture_wst	0.042	0.146	0.288	0.773	
perimeter_wst	–0.053	0.206	–0.255	0.799	
area_wst	0.027	0.025	1.094	0.274	
smoothness_wst	17.425	39.766	0.438	0.661	
compactness_wst	10.043	13.414	0.749	0.454	
concavity_wst	5.233	8.402	0.623	0.533	
concave_wst	6.984	24.042	0.290	0.771	
symmetry_wst	6.791	10.278	0.661	0.509	
fractal_wst	–25.984	83.068	–0.313	0.754	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

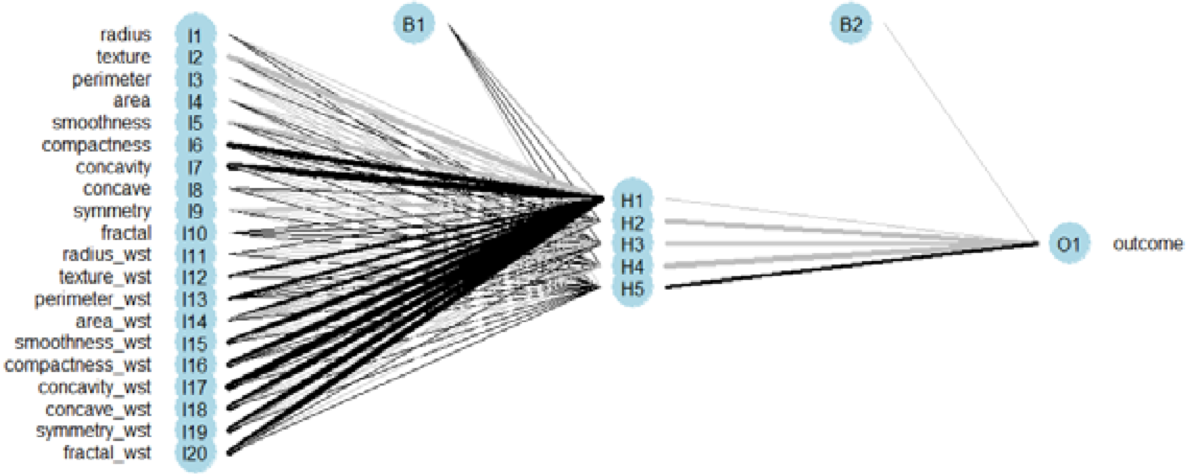

Figure 3. Artificial Neural Network in training sample

In above plot, line thickness represents weight magnitude and line color weight sign (black = positive, grey = negative). The net is essentially a black box so we cannot say that much about the fitting, the weights and the model. Suffice to say that the training algorithm has converged and therefore the model is ready to be used.

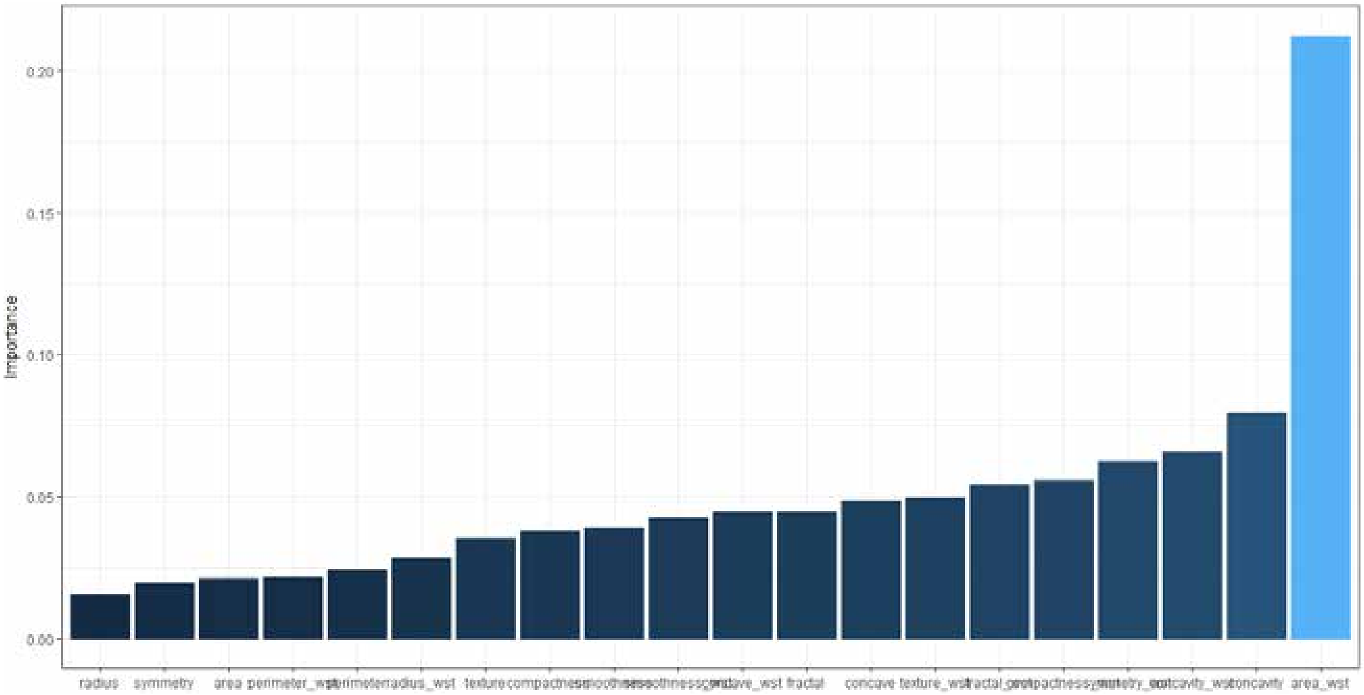

Figure 4. Variable Importance in Artificial Neural Network

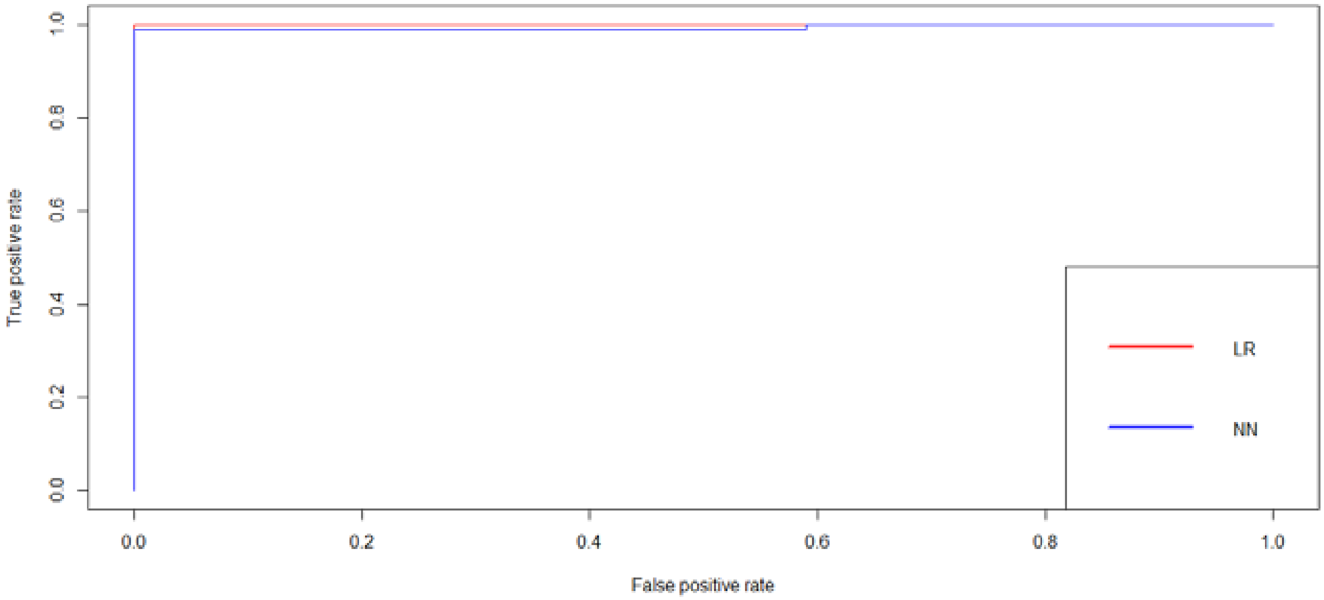

Figure 5. ROC in training sample for Logistic Regression (Red) vs Neural Network (Blue)

According to this neural network, the top 5 most important predictors were worst area, mean of severity of concave portions of the contour, worst of severity of concave portions of the contour, worst of symmetry, worst of compactness.

For training sample, the ROC was 1.0 for the Logistic regression and 1.0 for the artificial neural network. Artificial neural network performed better clearly. However in testing sample, the ROC was 0.92 for the Logistic regression and 0.99 for the artificial neural network. Artificial neural network had better performance.

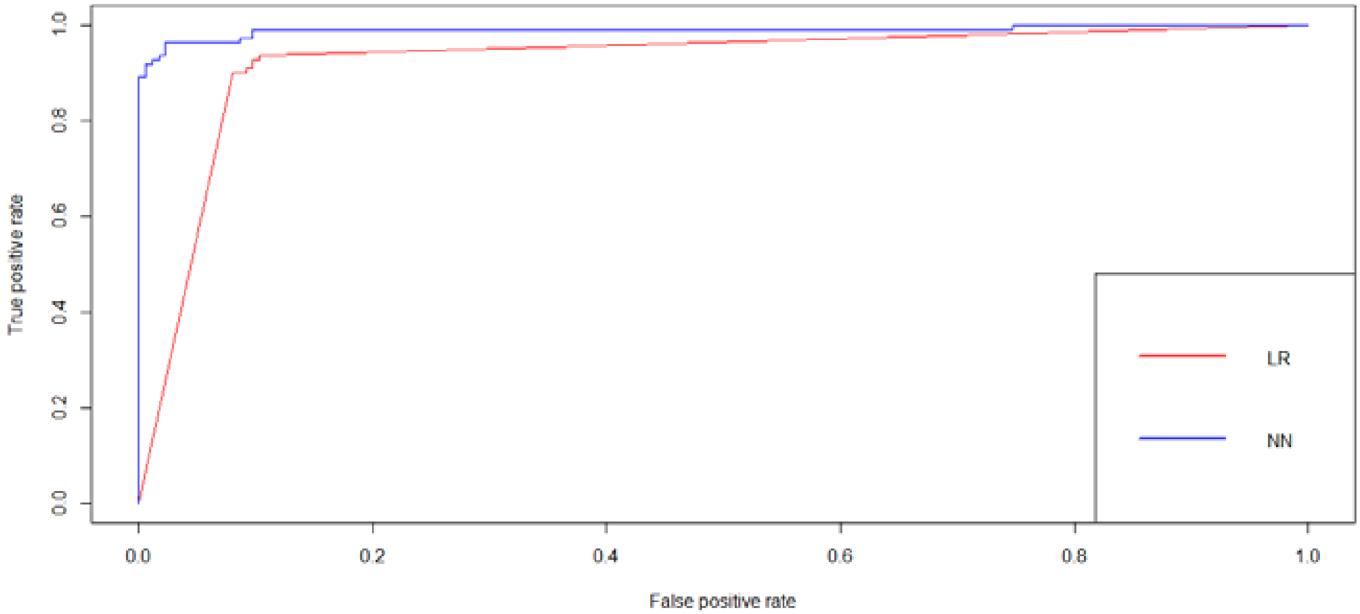

Figure 6. ROC in testing sample for Logistic Regression (Red) vs Neural Network (Blue)

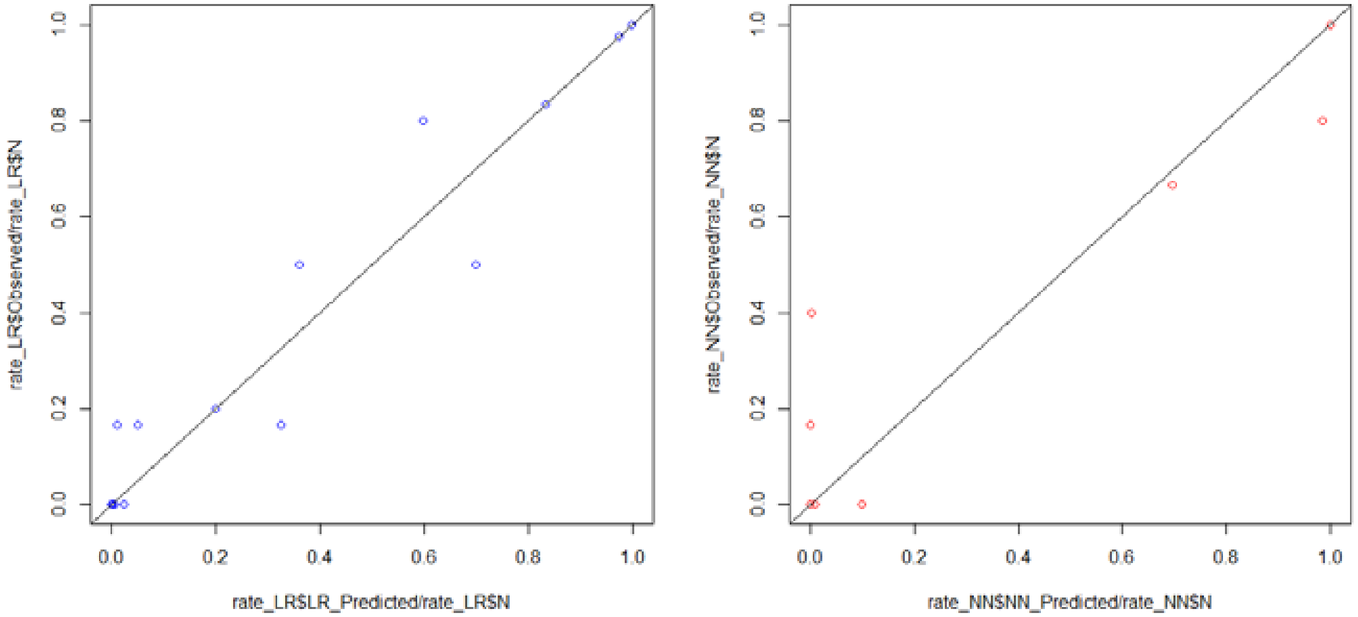

Figure 7. Predicted Probability vs. Observed Probability in testing sample for Logistic Regression (Red) vs Neural Network (Blue), sorted by predicted probability

By visually inspecting the plot we can see that the predictions made by the neural network are (in general) less concentrated around the line (a perfect alignment with the line would indicate an ideal perfect calibration) than those made by the Logistic model.

4. Discussions

In this study, we built a predictive model using artificial neural network as well as logistic regression to provide a tool for timely accurate diagnosis.

We identified several important predictors for breast cancer e.g., number of concave portions of the contour, worst of symmetry, worst of compactness. This provided important information for providers and patients for timely accurate diagnosis. Meanwhile we noticed that artificial neural network identified different predictors of great importance from predictors identified by logistic regression. It highlights the importance to explore new data mining techniques when building a predictive model.

Some known factors which might predict of breast cancer were not available in this study, like family history of breast cancer. According to the literature, women were at increased risk of breast cancer if they had close relatives (parents, siblings, or children) who were diagnosed with breast or ovarian cancer when they were younger than 45, especially if more than one relative was diagnosed or if a male relative had breast cancer.

We did not test the external validity neither for logistic regression nor for the ANN. However, we did a comprehensive split-sample validation with both strategies. Future studies could use outside data and test the performance of the outputs from these two models in this study.

A predictive model would be an extremely useful tool to timely diagnose breast cancer. When the variables included in our tool are available, the diagnosis could be acutely made. When compared to artificial neural network model, logistic regression had a worse discriminating capability and a better calibration between predicted probability and observed probability.

References:

1. U. S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2014. Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2017.
2. Antoniou A., Pharoah P. D., Narod S., et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: A combined analysis of 22 studies. American Journal of Human Genetics,– 72(5): 2003.– P. 1117–1130.
3. Chen S., Parmigiani G. Meta-analysis of BRCA1 and BRCA2 penetrance. Journal of Clinical Oncology,– 25(11): 2007.– P. 1329–1333.
4. CDC. Risk Factors for Breast Cancer in Young Women. URL: https://www.cdc.gov/cancer/breast/young_women/risk_factors.htm
5. Bennett K. P. and Mangasarian O. L. "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets". Optimization Methods and Software – 1, 1992.– P. 23–34.