International Journal of Scientific Research in Mathematical and Statistical Sciences
Vol.6, Issue.2, pp.284-289, April (2019)
DOI: https://doi.org/10.26438/ijsrmss/v6i2.284289

Research Paper

E-ISSN: 2348-4519

A Comparison of Statistical Discriminant Analysis and Artificial Neural Network Model for the prediction of breast cancer

R Jaisankar[1*], D Victorseelan[2]

[1]Department of Statistics, Bharathiar University, Coimbatore, India
[2]Department of Mathematics, Nehru College of Management, Coimbatore, India

Corresponding Author: r_jaisankar@rediffmail.com

Available online at: www.isroset.org

Received: 21/Apr/2019, Accepted: 28/Apr/2019, Online: 30/Apr/2019

Abstract- Artificial Neural Network (ANN) is one of the widely used statistical learning techniques in machine learning and cognitive science which is inspired by biological neural networks and basically consists of several non-linear processing units, called neurons or nodes. Though Artificial Neural Network has a wide variety of applications, it can also be used for discrimination of subjects. Statistical Discriminant analysis developed by R.A. Fisher (1936) is still prevailing as a novel methodology for discrimination. This paper presents the results of an experimental comparison of Statistical Discriminant Analysis and Artificial Neural Network (ANN) for predicting the patients affected by breast cancer. Samples of 116 patient's profiles collected from various private and government hospitals in Coimbatore, India, were used. The power of the model is measured by correct prediction rate. The study reveals that higher accuracy is provided by Neural Network analysis than Discriminant analysis in terms of prediction.

Keywords- *Discriminant Analysis, Machine learning, Artificial Neural Networks, Back propagation algorithm, Training and testing.*

I. INTRODUCTION

Predicting breast cancer is critical for medical Industry because it allows them to develop strategic programs that will help to decrease the affected. The present work aims on an experimental comparison study of Statistical Discriminant Analysis and Artificial Neural Network for predicting breast cancer. Samples of 116patient's profiles were used. The power of the models is measured by correct prediction rate.

Several researchers have conducted studies related to breast cancer and some researchers attempted to development and performance of an Artificial Neural Network for predicting the affected and also applying sensitivity analysis on the ANN developed, to identify the factors. (Joana Crisostomo (2016) and Miguel Patricio (2018))[1]. Some of the factors which are significant includes Age in years, BMI, Glucose level, Insulin, HOMA - Homeostasis model assessment scores, Leptin, Adiponectin, Resistin and MCP-1[2].

The data collected from different hospitals in Coimbatore district, Tamilnadu, INDIA are used and the independent variables mentioned above were taken up for analyses Statistical Discriminant Analysis and Artificial Neural Network modeling[3][4].

The following section consists of a brief description of statistical techniques – Discriminant analysis and artificial neural networks that were used in this study. The final section contains the conclusions and possible future work.

A. Discriminant Analysis

Discriminant analysis is used to describe the differences between groups and use these differences for classifying a new member to these groups based on the observations taken from the member. Discriminant analysis is also called classification in many references. However, several sources use the word classification to mean cluster analysis. Some applications of Discriminant analysis include medical diagnosis [5], market research, classification of specimens in anthropology, predicting company failure or success, placement of students (workers) based on comparing pretest results to those of past students (workers), discrimination of natural versus man-made seismic activity, fingerprint analysis, image pattern recognition, and signal pattern classification.

Discriminant Analysis characterizes an individual or a phenomenon, by a vector of variables X_1, $X_{2,}$...,X_n that constitute a multivariate density function.

The Discriminant function maps the multidimensional characteristics of the density function of the population's variables onto a one-dimensional measure, by forming a linear combination [6]. The linear Discriminant function is as follows:

$$\boldsymbol{Z_i = XA = a_0 + a_1X_{i1} + a_2X_{i2} + \cdots + a_nX_{in}}$$

Where,
Z = Discriminant score for the patient,
X = vector of *n* independent variables.
A = vector of Discriminant coefficients

Discriminant Analysis computes the Discriminant coefficients and selects the appropriate weights that will separate the average values of each group, while minimizing the statistical distance of each observation and its own group means.

B. Artificial Neural Network

An Artificial Neural Network (ANN) is a technique based on the neural Structure of the brain that mimics the learning capability from experiences. It means that if a neural network is trained from past data, it will be able to generate outputs based on the knowledge extracted from the data [7]. A neural network is called a mapping network, if it is able to compute some functional relationship between its input and output. Function approximation from a set of input-output pairs has numerous scientific and engineering applications.

Thus a neural network is extremely useful when you do not have any idea of the functional relationship between the dependent and independent variables. The network formed has to be trained so that it will learn an approximation. Learning in a neural network means, finding an approximate set of weights.[8][9] Thus, ANN is a universal function approximated that can approximate any function with arbitrary accuracy. It is to be noted that ANN is a nonlinear model that can be implemented for most complex real world applications.

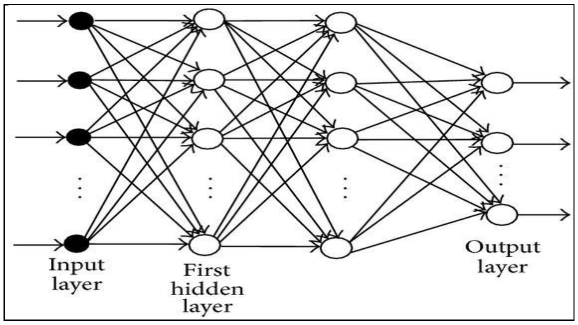


Figure.1 back- propagation neural network model

A back-propagation neural network (BPNN) is a simple and effective model of ANN. It consists of three layers which are input, hidden and output layers as shown in Figure 1. This network is also known as a feed forward back-propagation neural network.

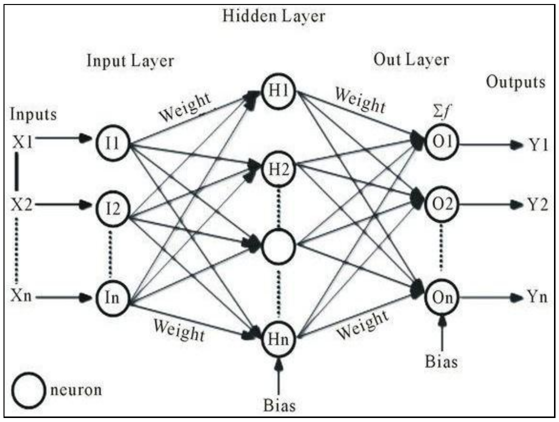


Figure 2- Learning process of the feed forward back-propagation network.

Above Figure shows the learning process of a neural network. In the training phase, the training data is fed into the input layer. It is propagated to both the hidden layer and the output layer. This process is called the forward pass [10]. In this stage, each node in the input layer, hidden layer and output layer calculates and adjusts the appropriate weight between nodes and generates an output value of the resulting sum. The actual output values will be compared with the target output values.

The error between these outputs will be calculated and propagated back to hidden layer in order to update the weight of each node again. This is called the backward pass or learning. The network will iterate over many cycles until the error is acceptable. After the training phase is complete, the trained network is ready to use for any new input data. During the testing phase there is no learning or modifying of the weight matrices. The testing input is fed into the input layer and the feed forward network will generate results based on its knowledge from the trained network [11]. A BPNN is one of the more popular ANNs that have been used for many ANN applications and is a robust neural network that can be applied easily in various problem domains.
Due to the fact that there is no conclusive superiority of any particular technique, the significance of this study is in providing insight on the contribution of various predictors in predicting the affected.

II. METHODOLOGY

The total number of records collected was from116 patients. The classification results of both Discriminant Analysis and Artificial Neural Network methods were based on the same datasets that were used for verification and calibration. This provides objectivity by comparing only the performance of each classification method. The success rate of classification is determined by the ratio of correctly classified recordings to the total number of recordings in that set. The factors taken in this study which are expected to have influence on breast cancer are as under:

A. Parameters in patient's profile

Age, BMI (Body mass index),

Glucose - a simple sugar which is an important energy source in living organisms and is a component of many carbohydrates,

Insulin - a hormone produced in the pancreas by the islets of langerhans, which regulates the amount of glucose in the blood,

HOMA - Homeostasis model assessment score (insulin resistance),

Leptin - Leptin is a hormone predominantly made by adipose cells that helps to regulate energy balance by inhibiting hunger,[2]

Adiponectin - is a protein hormone which is involved in regulating glucose levels as well as fatty acid breakdown. In humans it is encoded by the ADIPOQ gene and it is produced in adipose tissue,

MCP-1 (Monocyte Chemoattractant Protein-1)

Table1: Variables Descriptions Output Variables Code

Variables Description	**Output variable**	**Code**
Cancer Score	Classification	1 = Healthy controls, 2 = Patients

There are 10 predictors, all quantitative, and a binary dependent variable, indicating the presence or absence of breast cancer. The predictors are anthropometric data and parameters which can be gathered in routine blood analysis. Prediction models based on these predictors, if accurate, can potentially be used as a biomarker of breast cancer.

III. ANALYSIS

A. Discriminant Analysis Approach

First the Discriminant analysis has been conducted based on the variables with the help of IBM SPSS package, version 20. The results of Statistical Discriminant Analysis are shown below together with the interpretation.

The canonical Discriminant function coefficient table gives un-standardized coefficients which are used to create the Discriminant equation.

Table2: Regression Equation

Canonical Discriminant Function Coefficients	
	Function
	1
Age	-0.0124
BMI	-0.1142
Glucose	0.0587
Insulin	0.1742
HOMA	-0.5191
Leptin	-0.0054
Adiponectin	-0.0066
Resistin	0.0333
MCP.1	0.0004
(Constant)	-2.5003

It operates just like a regression equation. In this case the observed Discriminant function is

$$\begin{aligned}\mathbf{D} = (-\mathbf{0.0124} * \mathbf{Age}) + (-\mathbf{0.1142} * \mathbf{BMI}) \\ + (\mathbf{0.0587} * \mathbf{Glucose}) \\ + (\mathbf{0.1742} * \mathbf{Insulin}) \\ + (-\mathbf{0.0066} * \mathbf{Adiponectin}) \\ + (\mathbf{0.0333} * \mathbf{Resistin}) \\ + (\mathbf{MCP.1} * \mathbf{0.0004}) - \mathbf{2.5003}\end{aligned}$$

A Discriminant score can be calculated based on the weighted combination of the independent variables.

Table3: Functions at Group Centroids

Classification	**Function**
Healthy Control	-0.696
Patient	0.565

Centroids are the mean Discriminant scores for each group. This table is used to establish the cutting point for classifying cases. If the two groups are of equal size, the best cutting point is half way between the values of the functions at group Centroids (that is, the average)[12]. If the groups are unequal, the optimal cutting point is the weighted average of the two values.

As the groups taken are unequal the weighted mean (i.e Discriminant score) is calculated as Weighted mean = 1.86833.

88.5 % sensitivity of healthy control and 68.8% specificity of patient were observed. The classification results reveal that 88.5% of respondents were classified correctly into 'healthy' groups.

Table 4: Classification Statistics

Classification		**Predicted Group Membership**		
		Healthy Control	Patient	**Total**
Count	**Healthy Control**	46	6	52

	Patient	20	44	64
Percentage	**Healthy Control**	88.5	11.5	100
	Patient	31.3	68.8	100
	77.6% of original grouped cases correctly classified.			

This overall predictive accuracy of the Discriminant function is called the 'hit ratio'. Healthy Control classification is done with slightly better accuracy (88.5%) than Patient class (68.8%).

Table 5: Classification accuracy (Cross Validated value)

DA	Healthy Control	Patient	Sub Total
Healthy Control	43	9	52
Patient	22	42	64
Sub Total	65	51	116

The above Table shows the classification of accuracy in forecasting. The independent variables in the prediction equation predicted the classification group membership correctly with a correct prediction rate77.6%.

B. Neural Network Approach

The neural network technique is applied to the same data used for the Discriminant analysis method. A three- layer feed forward network was considered for the analysis. The input layer consists of nine nodes and the designed network model has ten hidden layers.

Determining the optimal number of hidden layers and nodes it is a crucial yet complicated one, which can be found by trial and error.

In general, networks with too many hidden neurons tend to memorize the input. In this study the back- propagation learning algorithm was used for training the network. The sigmoid activation function was used at the hidden layers as well as for output layer. 70 percent of the data were used for training and 30 percent for testing the model. This resulted in a mean correct classification rate for the test data. The network was allowed to run for 1,000 epochs for each run.

Performance measures:

There can be many performance measures for a predictor and the most important measure of performance is the prediction accuracy that can be achieved with the training data.

The percent of incorrect prediction shows how well your network Architecture was created. The detail of network structure constructed is given below,

Number of input nodes= 9
Number of Output nodes=1
Number of hidden layers= Single
Number of Hidden Nodes = 10
Hidden layer activation function - Tangent Hyperbolic

Using different training and testing percentages, the model was selected as the one with the highest accuracy (94.17%) with data size 116.

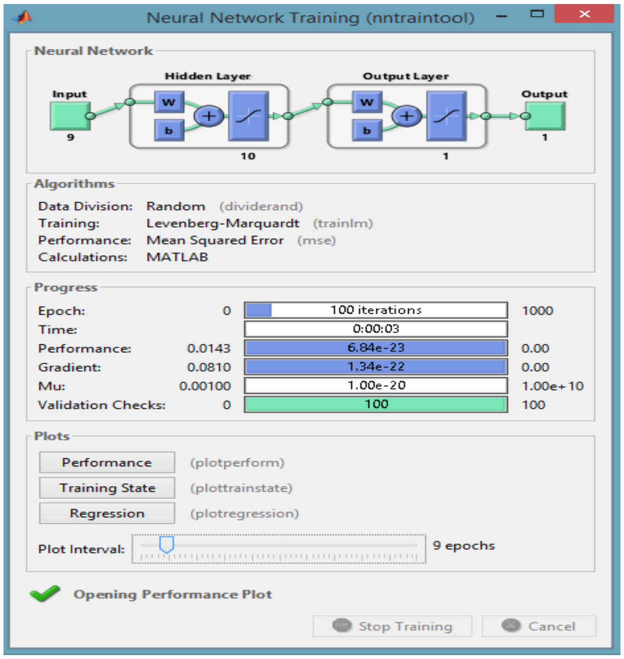
Table 6: Weight Matrix

	input nodes								
hidden nodes	0.05	-1.40	-1.79	-1.71	0.48	1.68	4.29	-2.0366;	-0.24
	-0.72	-0.60	-0.94	-0.85	-0.88	-1.00	0.31	-0.04	-0.10
	0.68	0.50	-4.85	-1.28	-0.92	0.76	1.18	0.13	0.08
	-2.49	-0.02	-1.35	-0.12	-2.10	1.14	0.85	1.31	0.33
	-1.00	0.77	-0.54	-0.67	-0.78	-0.27	0.49	0.35	-0.49
	-0.40	-1.16	-0.04	0.96	0.68	0.65	0.38	-0.48	-0.13
	-0.47	2.77	1.15	-0.07	0.89	0.08	2.64	-4.22	-0.58
	-1.98	1.51	-0.17	0.06	0.95	0.81	-1.33	0.95	0.97
	0.45	-0.73	0.12	0.27	0.22	-0.22	0.69	1.08	-0.45
	1.43	2.83	1.49	0.88	-0.62	1.84	0.45	-1.33	1.16

Training performance of the ANN

The training algorithm used Levenberg-Marquardt and performance metric used is Mean Squared Error.

With the above setting, the training performance, training states and regression results are analyzed and provided below



Data Division: Random (dividerand)
Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Calculations: MATLAB

```
Epoch:                  0   100 iterations   1000
Time:                       0:00:03
Performance:       0.0143   6.84e-23         0.00
Gradient:          0.0810   1.34e-22         0.00
Mu:               0.00100   1.00e-20         1.00e+10
Validation Checks:      0   100              100
```

(plotperform)
(plottrainstate)
(plotregression)

Plot Interval: 9 epochs

✓ Opening Performance Plot

Figure 3: Neural Network Architecture

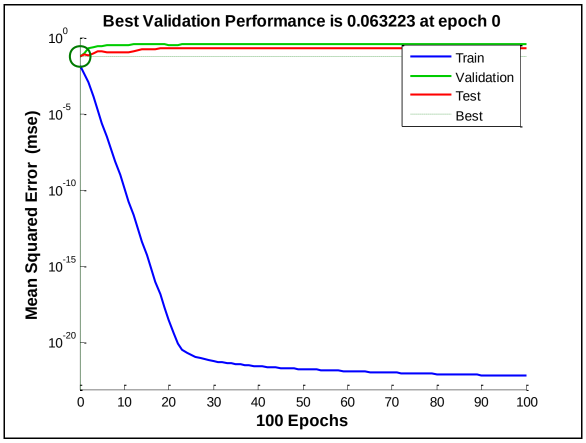


Figure 4: Training Performance of ANN

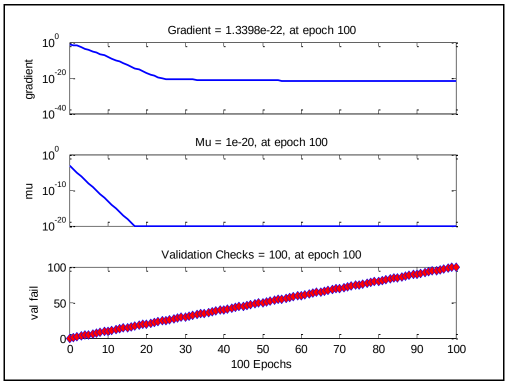


Figure 5: Training State and Validation Checks of ANN

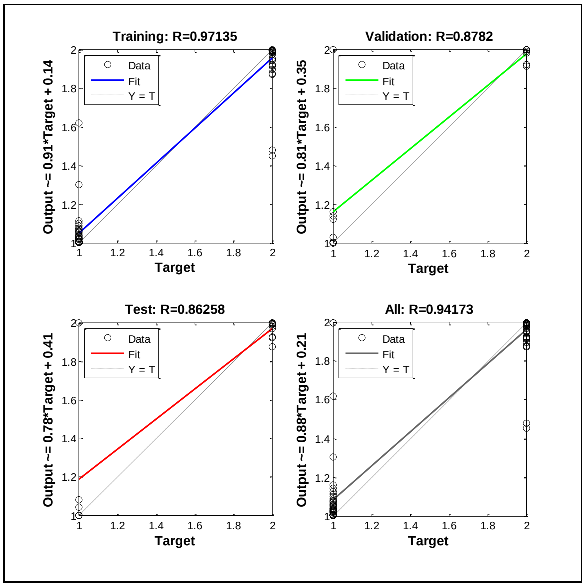


Figure 6: Regression Chart

F. Regression Equation

$$\mathbf{Y = p1 * z^{8} + p2 * z^{7} + p3 * z^{6} + p4 * z^{5} + p5 * z^{4} + p6 * z^{3} + p7 * z^{2} + p8 * z + p9}$$

Where z is centered and scaled:
z = (x-mu)/sigma
mu = 1.5517
sigma = 0.49947

Coefficients:
p1 = -2.9921e+57,
p2 = 1.3074e+57,
p3 = -2.1476e+57,
p4 = 6.2417e+56,
p5 = 1.9812e+57,
p6 = -2.4167e+57,
p7 = 1.3335e+57,
p8 = -2.7753e+57,
p9 = 3.2836e+57

Norm of residuals =
7.5058e+42

Table7: Classification accuracy of the ANN model

ANN	Healthy Control	Patient	Total
Healthy Control	50	2	52
Patient	9	55	64
Total	59	57	116
94.17% of original grouped cases correctly classified.			

The above Table shows the percentage of accuracy in forecasting. The prediction equation predicted classification groups with a correct prediction rate94.17%.

IV. COMPARISON OF DISCRIMINANT ANALYSIS AND ANN

To compare the performance of the neural network approach with the Discriminant analysis approach, the correct prediction rate of predictive accuracies in neural network and Discriminant analysis models are shown in table below. Clearly, the neural network method demonstrates a superior ability to predict the patients affected by breast cancer, shown in the following table.

Table8: Classification results from DA and ANN

Model	**Correct prediction Rate %**
Discriminant Analysis	77.6%
Artificial Neural Networks	94.17%

V. CONCLUSION

The target of this research work is to study the effectiveness of artificial neural networks and Statistical Discriminant Analysis in forecasting in prediction of breast cancer. A three-layer supervised neural network has been taken-up based on the back-propagation learning algorithm for ANN. It is found that the predictive accuracy of Artificial Neural Network is higher than that of the Discriminant Analysis. Extension of this work may be made by incorporating more number of medical predictors with different types of neural networks.

REFERENCES

[1] Joana Criso´stomo, Paulo Matafome, Daniela Santos-Silva.“*Hyper-resistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer*”, Endocrine, Springer Science+Business Media, New York 53, pp. 433–442, 2016.

[2] Miguel Patrício, José Pereira, and Joana Crisóstomo,“*Using Resistin, glucose, age and BMI to predict the presence of breast cancer, Patrício et al.*”, BMC Cancer, DOI 10.1186/s12885-017-3877-1, 2018.

[3] Rahul Rajawat “*On interface between artificial neural network and statistical techniques*”, Thesis submitted to the University of Rajasthan, Doctor of Philosophy 2015.

[4] Berna Yazici, Memmedaga Memmedli, Atilla Aslanargun and Senay Asma, “*Analysis of international debt problem using artificial neural networks and statistical methods*”, Neural Computing and Applications, Volume 19, Issue 8, pp 1207–1216, 2010.

[5] Chung,K.C., S. S. Tan,S.S. and Holdsworth,, D.K. “*Insolvency Prediction Model Using Multivariate Discriminant Analysis and Artificial Neural Network for the Finance Industry in New Zealand*”.International Journal of Business and Management, vol. 3, no. 1, pp. 19-28, 2008.

[6] Jeatrakul.P and Wong, K.W.“*Comparing the performance of different neural networks for binary classification problems*,” Eighth International Symposium on Natural Language Processing, Bangkok, Thailand, pp. 111-115, 2009.

[7] C. Punitha Devi, T. Vigneswari. “*A Survey on Machine Learning and Statistical Methods for Bankruptcy* Prediction”, International Journal of Computer Sciences and Engineering, Vol.-7, Issue-3, 104-110, March 2019.

[8] Silverman, D and Dracup, J.A“*Artificial Neural network and long range precipitation prediction in California*”, Journal of applied Meteorology, 39,pp.57-66, 2000.

[9] Michaelides,S.C., Neocleous, C.C. and Schizas, C.N. “*Artificial Neural Networks and Multiple linear regressions in estimating missing rainfall data. Proceedings of the DSP95*”,International Conference on Digital Signal Processings, Limassol, Cyprus, pp. 668-673, 1995.

[10] Gardner, M.W. and Dorling, S.R. “*Artificial Neural Network (Multilayer Perception)- A review of applications in atmospheric sciences*”, Atmospheric Environment, 32, pp.2627-2636, 1998.

[11] Hsieh, W.W, and Tang.T.“Applying Neural Network Models to prediction and Data Analysis in Meteorology and Oceanography”, Bulletin of the American Meteorological Society,79, pp.1855-1869,1998.

[12] Nukala V V Pravallika1and P Suresh Varma, “*Prediction of Heart Disease Using Machine Learning Algorithms”,* International Journal of Computer Sciences and Engineering, Vol.-7, Issue-3, March 2019.

AUTHORS PROFILE

Prof R Jaisankar, Ph.D. He is currently working as Professor in Department of Statistics, Bharathiar University, Coimbatore, India. He has published more than 50 research papers in reputed international and national journals. His main research work focuses on stochastic modeling in Bio-Statistics; Applied Statistics which include Design of Experiments, Regression models, Multivariate data analysis and Survival models etc. He has 29 years of teaching and Research experience.

Mr D Victorseelan, pursuing Ph.D in Bharathiar university and also working as an assistant professor in Nehru College of Management, Coimbatore. He has published a research papers in reputed international journal. His main research work focuses on Biostatistics, Machine Learning Algorithms and Computational Statistics. He has 6 years of teaching and Research experience.