COMPARISON OF LOGISTIC REGRESSION AND NEURAL NETWORK MODEL WITH AND WITHOUT HIDDEN LAYER

Raghavendra B.K. [#1], Dr. S.K. Srivatsa [*2], Raghavendra S. [*3], Shivashankar S.K. [#4]

Dr. M.G.R. Educational and Research Institute, Chennai, INDIA
Senior Professor, St. Joseph's college of Engineering, Chennai, INDIA
Assistant Professor, Ghousia College of Engineering, Ramanagaram, INDIA
Lecturer, P.E.S. College of Engineering, Mandya, INDIA
[1] raghavendra_bk@rediffmail.com
[2] profsks@rediffmail.com
[3] raghav.trg@gmail.com
[4] shivashankarsk@rediffmail.com

***Abstract*— Logistic regression model has been widely used in many biomedical fields such as cancer diagnosis, survival prediction. It is a powerful and well-established method both in statistics and biomedical fields. It is also recommended that data mining techniques should be compared to logistic regression when conducting clinical data mining. It enables us to investigate the relationship between a categorical outcome and a set of explanatory variables. Artificial neural networks (ANNs) are considered as a field of artificial intelligence. ANN have been applied in many disciplines, including biology, psychology, statistics, mathematics, medical science, and computer science. It has also been applied to a variety of business areas such as accounting and auditing, finance management and decision making, marketing and production. In this an attempt has been made to evaluate logistic regression model using neural network with and without hidden layers on publicly available medical datasets. The classification accuracy is used to measure the performance of the model. From the experimental results it is confirmed that the neural network model without hidden layer gives more efficient result.**

***Keywords*— Artificial neural network, classification accuracy, hidden layers, logistic regression, medical dataset.**

I. INTRODUCTION

Medical applications of data mining include prediction of effectiveness of medical decisions. Nowadays statistical methods constitute a very powerful tool for supporting medical decisions. The size of medical data that any analysis or test of patients makes that doctors can be helped by statistical models to interpret correctly and to support their decisions. The models are a very powerful tool for doctors and these can not substitute their viewpoint. On the other hand, the characteristics of medical data and the huge number of variables to be considered as fundamental point for the development of new technique as neural network for the analysis of the data [1].

Logistic regression is a technique for analysing problems in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable. In logistic regression, the dependent variable is binary or dichotomous, i.e., it only contains data coded as 1 (TRUE, success, etc.) or 0 (FALSE, failure, etc.).

Logistic regression model has been widely used in many biomedical fields [2-5], such as cancer diagnosis, survival prediction; et al. Logistic regression is a powerful and well-established method both in statistics and biomedical fields. It is also recommended that data mining techniques should be compared to logistic regression when conducting clinical data mining [6].

Neural networks are considered as a field of artificial intelligence. The development of the models was inspired by the neural architecture of human brain. ANN have been applied in many disciplines, including biology, psychology, statistics, mathematics, medical science, and computer science. It has also been applied to a variety of business areas such as accounting and auditing, finance, management and decision making, marketing and production. Recently, artificial neural networks (ANNs) become a very popular model and have been applied to diagnose disease and predict the survival ratio of the patients. Many researchers have compared ANN versus LR. Some of them found that ANN and LR have similar classification performance. Compared to LR, neural network models are more flexible [7].

In this research work we show that it is possible to reliably improve the neural network model without hidden layer is simple and more efficient when compared with the neural network layer with hidden layer. The rest of the paper is organized as follows: Section II reviews the prior literature, Logistic Regression technique and design of neural network is discussed in Section III and IV. Experimental validation using publicly available medical dataset is given in Section V.

Section VI includes Experimental results and discussions followed by conclusion.

II. Literature Survey

There is an approach that examines the problem of efficient feature evaluation for logistic regression on very large data sets. The authors present a new forward feature selection heuristic that ranks features by their estimated effect on the resulting model's performance. An approximate optimization, based on back fitting, provides a fast and accurate estimate of each new feature's coefficient in the logistic regression model. Further, the algorithm is highly scalable by parallelizing simultaneously over both features and records, allowing us to quickly evaluate billions of potential features even for very large data sets [8].

Logistic regression is used in power distribution fault diagnosis, while neural network has been extensively used in power system reliability researches. Evaluation criteria of the goodness of the classifier includes: correct classification rate, true positive rate, true negative rate, and geometric mean [9].

The features of logistic regression and ANN have been compared and an experiment has been conducted on graft outcomes prediction using a kidney transplant dataset. The results shown reveal that ANN coupled with bagging is an effective data mining method for predicting kidney graft outcomes. This also confirms that different techniques can potentially be integrated to obtain a better prediction. Overall, the results reveal that in most cases, the ANN technique outperforms logistic regression [10].

In another research work the author compares the performance of LR, NN, and CART decision tree methodologies and to identify important factors for the small business credit scoring model on a Croatian Bank dataset. The models obtained by all three methodologies were estimated and validated on the same hold-out sample, and their performance is compared. The result shows that the NN model is better associated with data than LR and CART models [11]. The use of Artificial Neural Networks is to construct distributions to carry out possible reasoning in the field of medicine. It describes a comparison between Multivariate Logistic Regression (MLR) and the Entropy Maximization Network (EMN) in terms of explicit assessment of their predictive capabilities. The EMN and MLR have been used to determine the possibility of harboring lymph node metastases at the time of initial surgery by assessment of tumor based parameters. Both predictors were trained on a set of 84 early breast cancer patient records and evaluated on a separate set of 92 patient records. Differences in performance were evaluated accurately than MLR model with AZ=0.839, compared to the MLR model with AZ=0.809. The difference was statistically significant with two-tailed P value of less than 0.001. Accurate estimation of prognosis would provide better stratification of patients for further treatment or investigation [12].

III. LOGISTIC REGRESSION

Logistic regression is a mathematical modeling approach that is used to describe the relationship between several explanatory variables X's to a dichotomous dependent variable Y. Logistic regression can be used to predict the outcome from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. That is, logistic regression makes no assumption about the distribution of the independent variables. They do not have to be normally distributed, linearly related or of equal variance within each group. The dichotomous dependent variable can take the value of 1 with a probability of success P, or the value of 0 with probability of failure 1-P. This type of variable is called Bernoulli (or binary) variable.

The relationship between the predictor and response variables is not a linear function in logistic regression, instead, logistic regression function is used which is the logit transformation of P [13].

$$\ln\left(\frac{P}{1+P}\right) = a + b_1 x_1 + b_2 x_2 + \ldots b_j x_j$$

$$\frac{P}{1+P} = e^{a + b_1 x_1 + b_2 x_2 + \ldots b_j x_j}$$

$$P = \frac{1}{1 + e^{a + b_1 x_1 + b_2 x_2 + \ldots b_j x_j}}$$

where P is the probability of a 1, *e* is the base of the natural logarithm and *a* and *b* are the parameters of the model.

IV. ARTIFICIAL NEURAL NETWORK

Artificial neural networks are networks of units called neurons that exchange information in the form of numerical values with each other via synaptic interconnections. Neural network is a complex nonlinear modeling technique based on a model of a human neuron. A neural net is used to predict outputs (dependent variables) from a set of inputs (independent variables) by taking linear combinations of the inputs and then making nonlinear transformations of the linear combinations using activation function. It can be shown theoretically that such combinations and transformations can approximate virtually any type of response function. Thus, neural nets use large numbers of parameters to approximate any model. Neural nets are often applied to predict future outcome based on prior experience. For example, a neural net application could be used to predict who will respond to a direct mailing.

Neural networks are becoming very popular with data mining practitioners, particularly in medical research, finance and marketing. This is because they have proven their predictive power through comparison with other statistical techniques using real data sets. There are two main types of feed-forward neural networks, which are known as universal approximators, to create a nonlinear mapping between a set of input variables and the output variables; they are multi-layer perceptrons and radial basis function neural networks. Back-

propagation or Multi-Layer Perceptron is the most common ANN used in many research areas. It is composed of connected nodes (neurons) and weights as processing elements [14].

In this research work, two layered network without hidden layer and three layered network with multiple nodes of hidden layer and single node output layer are used. Number of input nodes may be used based on the results of feature selection algorithm. The simplest feed-forward neural networks (FFNN) with and without hidden layer is shown in Fig. 1 and 2 respectively. Fig. 1 consists of two layers: an input layer and an output layer and Fig. 2 consists of three layers: an input layer, hidden layer, and an output layer. In each layer there are one or more processing elements (PEs). PEs are meant to simulate the neurons in the brain and this is why they are often referred to as neurons or nodes. A PE receives inputs from either the outside world or the previous layer. There are connections between the PEs in each layer that have a weight (parameter) associated with them. This weight is adjusted during training. Information only travels in the forward direction through the network - there are no feedback loops. And for regression problems neural network takes only one output node.

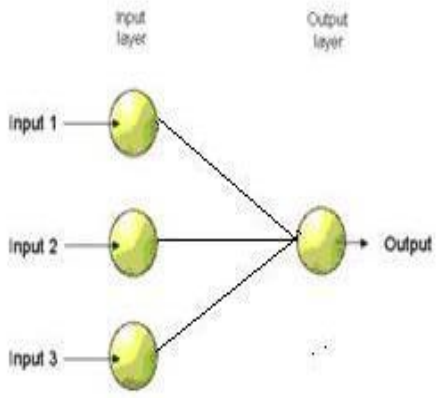

Fig. 1: Example of a simple feed forward neural network with two layers

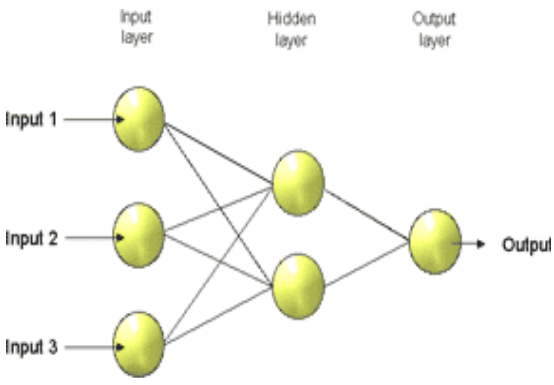

Fig. 2: Example of a simple feed forward neural network with three layers

V. EXPERIMENTAL VALIDATION

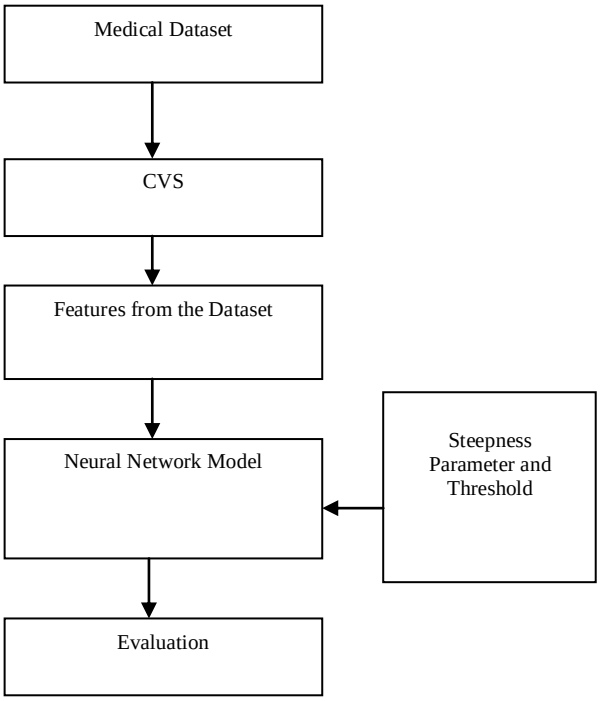

Fig. 3: Neural Network Model Framework

The framework for neural network model with and without hidden layer is shown in Fig. 3. The process of evaluation is as follows: The logistic regression and neural network model with and without hidden layer has been evaluated for the effectiveness of the classification. The logistic function with steepness parameter (σ) is calculated using the following equation.

$$P=1/ (1+e^{-logit (p) * \sigma}) \quad (5.1)$$

where σ=2, 3

The response Y is then calculated as follows by using threshold (τ). Then the probability is calculated to develop a predictive model for classification using neural network. The classification accuracy is used to measure the performance of both the models. A tenfold cross validation has been used for evaluation on all publicly available medical dataset [15].

$$Y = \begin{cases} 1 & \text{if } P \geq \tau \\ 0 & \text{otherwise} \end{cases} \quad (5.2)$$

where τ=0.2, 0,4, . . .

VI. RESULTS AND DISCUSSION

In this research work we have used publicly available medical datasets for our experiments whose technical specifications are as shown in Table 1. All the chosen datasets had at least one or more attributes that were continuous. The classification accuracy is used to measure the performance of

logistic regression and neural network model on publicly available medical datasets. The results of the evaluation are given in Table 2 and 3 respectively. Fig. 4 and 5 gives classification accuracy details after evaluation process. From the experimental results it can be observed that the classification accuracy of the neural network model without hidden layer gives more efficient result.

TABLE 1

SPECIFICATION FOR THE MEDICAL DATA SET

Sl. No	Medical Dataset	No of instances	Total no. of attributes	No of classes
1	Asthma	2464	5	2
2	Blood-transfusion	748	5	2
3	Flushot	159	4	2
4	Haberman	306	4	2
5	Liver-disorders	345	7	2
6	Spect test	187	23	2
7	Echocardiagram	132	11	2

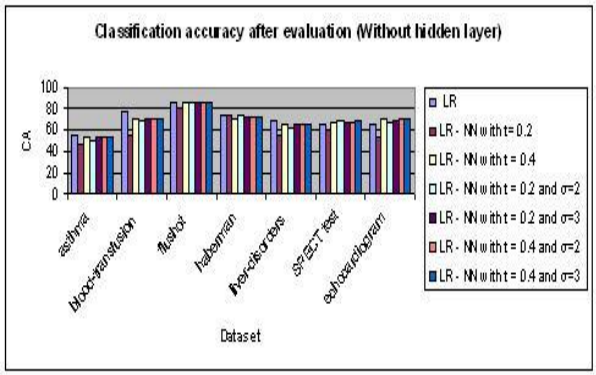

Fig. 4: Classification accuracy after evaluation (without hidden layer)

TABLE 2

LOGISTIC REGRESSION AND NEURAL NETWORK SPECIFICATION WITHOUT HIDDEN LAYER FOR THE MEDICAL DATA SET

Name of the Dataset	LR	LR - NN with t= 0.2	LR - NN with t = 0.4	LR - NN with t = 0.2 and σ=2	LR - NN with t = 0.2 and σ=3	LR - NN with t = 0.4 and σ=2	LR - NN with t = 0.4 and σ=3
asthma	56.16	46.91	52.84	50.48	52.67	53.65	53.97
blood-transfusion	77.13	56.14	71.12	68.71	70.85	70.85	71.12
flushot	86.16	80	86.26	86.25	86.25	86.25	86.25
haberman	74.18	74.5	71.89	74.5	73.2	72.54	73.2
liver-disorders	68.11	56.39	66.27	62.79	65.11	65.11	66.27
SPECT test	65.24	60.21	67.74	68.81	67.74	67.74	68.81
echocardiogram	65.15	54.54	69.69	66.66	68.18	71.21	71.21

TABLE 3

LOGISTIC REGRESSION AND NEURAL NETWORK SPECIFICATION WITH HIDDEN LAYER FOR THE MEDICAL DATA SET

Name of the Dataset	LR	LR - NN with t= 0.2	LR - NN with t = 0.4	LR - NN with t = 0.2 and σ=2	LR - NN with t = 0.2 and σ=3	LR - NN with t = 0.4 and σ=2	LR - NN with t = 0.4 and σ=3
asthma	56.16	48.25	52.35	49.91	52.15	54.13	54.34
blood-transfusion	77.13	68.58	77.13	76.2	76.6	78.2	78.34
flushot	86.16	73.12	85	85	85	85	85
haberman	74.18	73.52	73.85	73.2	73.2	73.52	72.54
liver-disorders	68.11	48.69	59.13	58.84	59.13	60	60.57
SPECT test	65.24	55.61	71.12	66.84	71.12	67.91	67.91
echocardiogram	65.15	63.63	71.21	70.45	70.45	72.72	72.72

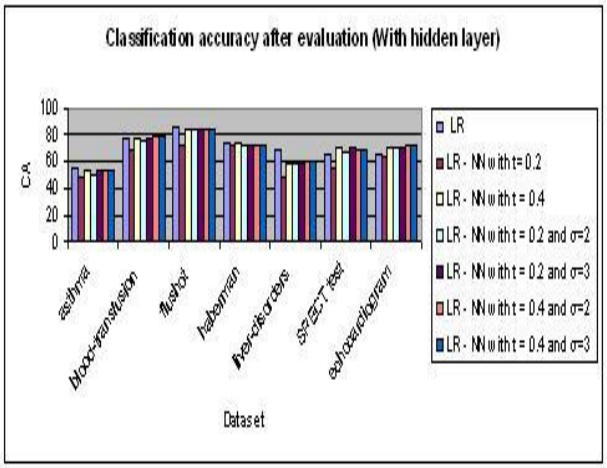

Fig. 5: Classification accuracy after evaluation (with hidden layer)

VII. CONCLUSIONS

In this research work an attempt has been made to evaluate logistic regression and neural network model with and without hidden layer on publicly available medical data sets and is used to develop a predictive model for classification using neural network with different threshold and steepness parameter. The classification accuracy is used to measure the performance of neural network model with and without hidden layer. From the experimental results it is confirmed that the neural network model without hidden layer gives more efficient result.

This document is a template. An electronic copy can be downloaded from the conference website. For questions on paper guidelines, please contact the conference publications committee as indicated on the conference website. Information about final paper submission is available from the conference website.

REFERENCES

[1] Luis Mariano Esteban Escaño, Gerardo Sanz Saiz, Francisco Javier López Lorente, Ángel Borque Fernando and José Moría Vergara Ugarriza, "Logistic Regression Versus Neural Networks for Medical Data", Monografías del Seminario Matemático García de Galdeano 33, 245-252 (2006).

[2] J. Chhatwal, O. Alagoz, M. J. Lindstrom, C. E. Kahn, K. A. Shaffer, and E. S. Burnside, "A Logistic Regression Model Based on the National Mammography Database Format to Aid Breast Cancer Diagnosis," *AJR Am J Roentgenol*, vol. 192, pp. 1117-1127, 2009.

[3] B. Eftekhar, K. Mohammad, H. E. Ardebili, M. Ghodsi, and E.Ketabchi, "Comparison of Artificial Neural Network and Logistic Regression Models for Prediction of Mortality in Head Trauma based on Initial Clinical Data," vol. 5, p. 3, 2005.

[4] E. Kupek, "Beyond Logistic Regression: Structural Equations Modelling for Binary Variables and its Application to Investigating Unobserved Confounders," *BMC Med Res Methodol*, vol. 6, p. 13, 2006.

[5] H. Khedmat, G. R. Karami, V. Pourfarziani, S. Assari, M. Rezailashkajani, and M. M. Naghizadeh, "A Logistic Regression Model for Predicting Health-related Quality of Life in Kidney Transplant Recipients," *Transplantation Proceedings*, vol. 39, pp. 917-922, May 2007.

[6] R. Bellazzi and B. Zupan, "Predictive Data mining in Clinical Medicine: Current Issues and Guidelines," *Int J Med Inform*, vol. 77, pp. 81-97, Feb 2008.

[7] Bahar Tasdelen, Sema Helvaci, Hakan Kaleagasi, Aynur Ozge, "Artificial Neural Network Analysis for Prediction of Headache Prognosis in Elderly Patients", Turk J Med Sci 2009; 39(1); pp 5-12.

[8] Singh, S., Kubica J., Larsen S., Sorokina D, "Parallel Large Scale Feature Selection for Logistic Regression", SIAM International Conference on Data Mining (SDM), 2009.

[9] LeXu, Mo-Yuen Chow, and Xiao-Zhi Gao, "Comparisons of Logistic Regression and Artificial Neural Network on Power Distribution Systems Fault Cause Identification", Proceedings of 2005 IEEE Mid-Summer Workshop on Soft Computing in Industrial Applications (SMCia/05), Helsinki, Finland, June 28-30, 2005.

[10] Fariba Shadabi and Dharmendra Sharma, "Comparison of Artificial Neural Networks with Logistic Regression in Prediction of Kidney Transplant Outcomes", Proceedings of the 2009 International Conference of Future Computer and Communication (ICFCC), pp 543-547, 2009.

[11] Marijana Zekic-Susac, Natasa Sarlija, Mirta Bensic, "Small Business Credit Scoring: A Comparison of Logistic Regression, Neural Network, and Decision Tree Models", 26th International Conference on Information Technology Interfaces (ITI 2004), Cavtat, Croatia, pp 265-270.

[12] Poh Lian Choong, and Christopher J.S. DeSilva (1996), "A Comparison of Maximum Entropy Estimation and Multivariate Logistic Regression in the Prediction of Axillary Lymph Node Metastasis in Early Breast Cancer Patients", The 1996 IEEE International Conference on Neural Networks, pp 1468-1473.

[13]. J. Miles and M. Shevlin,. "*Applying Regression and Correlation. A guide for Students andResearchers*". SAGE Publication Ltd., 2001.

[14] N.A. Setiawan, P.A. Venkatachalam, A.F.M. Hani, "Missing Data Estimation on Heart Disease Using Artificial Neural Network and Rough Set Theory", International Conference on Intelligent and Advanced Systems, 2007, pp 129-133.

[15] C.L. Blake, C.J Merz., "UCI repository of machine learning databases". [http://www.ics.uci.edu/~mlearn/MLRepository.html, Department of Information and Computer Science, University of California, Irvine.

Biographies and Photographs

DR S.K.SRIVATSA was born at Bangalore on 21st July 1945.He received his Bachelor of Electronics and Telecommunication Engineering Degree from Jadavpur University and Masters Degree and PhD from Indian Institute of Science, Bangalore. He is a senior Professor at St. Joseph's College of Engineering since 1st Aug 2005. He has taught 28 courses at the UG level and 40 courses at the PG level during the last 34 years. He is the author of over 450 publications in reputed journals and conference proceedings. He has produced 34 PhDs. He is the recipient of about a dozen awards. He is a life Fellow/Member in about two dozen registered professional societies. His research interests pertain to Electronics and Computer Science.

RAGHAVENDRA B. K. is working as an Associate Professor & Head, Department of Computer Science and Engineering at Ghousia College of Engineering, Ramanagaram, Karnataka, India. He has received his BE degree in Computer Science & Engineering from Bangalore University (1994) and M.Tech degree in Computer Science & Engineering from Visveswaraya Technological University, Belgaum (2004), and pursuing his PhD at Dr. MGR Educational & Research Institute, Chennai, India. His research area is Data Mining.

RAGHAVENDRA S. is working as an Assistant Professor, Department of Computer Science and Engineering at Ghousia College of Engineering, Ramanagaram, Karnataka, India. He has received his BE and MTech degrees in Computer Science & Engineering from Visveswaraya Technological University, Belgaum (2002, 2007). His research area is Cloud Computing and Data Mining.

SHIVASHANKAR S. K. is working as a Lecturer, Department of Computer Science and Engineering at P.E.S. College of Engineering, Mandya, Karnataka, India. He has received his BE degree from Mangalore University and MTech degree in Computer Science & Engineering from Visveswaraya Technological University, Belgaum (2001, 2005). His research area is Data Mining.