A Comparative Study of Machine Learning Algorithms in Classifying and Predicting Exclusive Breastfeeding Infants in Bangladesh

Mohammad Nayeem Hasan   
*Department of Statistics*  
*Shahjalal University of Science and Technology*Sylhet, Bangladesh  
nayeem5847@gmail.com

Sabrin Sultana  
*Department of Banking and Insurance*  
*University of Chittagong*Chittagong, Bangladesh  
sabrinsultana1060@gmail.com

*Abstract*— Artificial neural networks (ANNs) have recently been applied in situations where analysis based on logistic regression (LR) is a standard statistical approach; direct comparisons of results are rarely attempted. In this study we compared both the logistic regression model and the feed-forward neural network model with exclusive breastfeeding data set. Data for this study included 16 factors associated with exclusive breastfeeding (EBF), which collected from the Bangladesh Demographic and Health Survey (BDHS), 2014. LR with the forward method and feed-forward ANN were applied in 30 neurons in hidden layer to achieve best fitted model. The accuracy of the models in predicting EBF is compared by the classification accuracy and Area under Receiver Operating Curve (AUROC). Of the 10 ANNs, the hidden layer with 30 neurons presents better result compared to LR. The AUROC of the LR model and ANN with 30 neurons in hidden layers, were estimated as 0.78 and 0.94, respectively. The LR and ANN models respectively classified 87.06% and 96.38% of the exclusive breastfed infants correctly. Based on this dataset, it seems that the classification of infants in two groups with and without exclusive breastfed by ANN with 30 neurons at the hidden layer is better than the LR model.

Keywords— Logistic Regression, Artificial Neural Network, Exclusive breastfeed, BDHS.

# Introduction

Breastfeeding is a biological common diet that protects health and well-being from disease and is a safe alternative to ensure ideal growth for young children [1]. Breastfeeding has health benefits for both mother and child because it contains nutrients, antioxidants, hormones and antibodies [2]. Several national and international organizations (e.g. WHO) approve exclusive breastfeeding for the first six months (e.g. children received only breast milk) [3] and it is also recommended for two or more years as it relates to adolescence mental growth [4].

For classifying/predicting of the binary outcome variable, several methods have been developed such as linear discriminant analysis, logistic regression analysis, different machine learning methods like a decision tree, Support vector machine, and artificial neural network models. Previous studies showed that in predicting binary outcome of various domains in medical diagnosis such as lung cancer, breast cancer, low back pain, psychological symptom, hospital-acquired infections various machine learning algorithm specially ANN were used widely in association with some traditional statistical methods such as logistic regression analysis and discriminant analysis [5]–[7]. Uses of ANNs as clinical prediction models has been explored in many areas of medicine, including nephrology [8], microbiology [9], radiology [10] and neurology [11].

The two most commonly used models for data classification in medical diagnosis are LR and ANN [12]–[14]. Comparing ANN models with standard statistically generalized linear models such as logistic regression is an important step in the development process. If the results show that the gain of non-linear model, such as the ANN, is limited, one should usually go for the less complicated model (Green et al., 2006). Logistic regression always has the nice property of being fully interpretable and the assumption of independence of errors and variables is essential. Under this condition, if the relevant data is complicated, the model’s assumptions may not be true anymore (Shaﬁei et al., 2016). The advantages of NN compared with LR include the ability to detect automatically complex nonlinear relationships between predictor and outcome variables, and to implicitly discern interactions among independent variable (Tu J.V., 1996).

The main purpose of most of the studies is to find the factors that are significantly associated with EBF prevalence and some of them predicting EBF by using different classification method, but a very little number of studies that attempt to compare the performance of classification methods for predicting EBF. For this reason in this study, we applied LR model and ANN to predict the EBF based on effective factors and then compared the ability of each of these models to classify EBF among under six months children in Bangladesh.

# Methods

## Study Design

This study analyzed administrative claims data obtained from BDHS. The BDHS data set is assumed the most comprehensive and reliable data source for this study. In this study we used a nationally representative data set from the 2014 BDHS. A detailed description of the sample design and procedure is presented in the 2014 BDHS report (NIPORT et al., 2016).

## Data Preparation

Totally the mothers of 7886 children under 5 years were asked about pregnancy, demographic, economic and various health issues, and breastfeeding. A final data set of 628 observations was obtained after excluding non-eligible cases (e.g., visitors and non-surviving children), children that greater than 6 months and observation with missing values. By reviewing the valid literature, the most significant factors that are associated with EBF prevalence are shown in table 1. In this study the output or response variable was EBF. Thirteen factors including child age (in months), mothers’ education, fathers’ education, sex of child, mothers age group (in years), wealth index, disease category, place of residence, place of delivery, C-section delivery, division, body mass index (Mother), Religion, mass media, post-natal care, and child size at birth were used as possible predictor variables of ARI.

1. Description of variables

| Variable | Values |
| --- | --- |
| EBF | Yes, No |
| Mothers age group (in years) | 15-24, 25-34, 35-44, 45+ |
| Division | Dhaka, Rangpur, Rajshahi, Sylhet, Chittagong, Khulna, Barisal |
| Place of residence | Rural, Urban |
| Mothers’ education | No Education, Primary, Secondary, Higher |
| Husbands’ education | No Education, Primary, Secondary, Higher |
| Religion | Islam, Hinduism, Other |
| Wealth index | Poorest, Poorer, Middle, Richer, Richest |
| Body mass index (Mother) | Under weight, Normal weight, Overweight |
| Sex of Child | Male, Female |
| Delivery by caesarean section | Yes, No |
| Size of child at birth | Very large, Larger than average, Average, Smaller than average, very small |
| Childs age group (in months) | <3, 3+ |
| Mass media access | Yes, No |

## Prediction Models

1. Artificial Neural Network: ANN provides an alternative method of interpreting and recognizing complex patterns in data sets. ANNs should be considered a form of converges that iterative itself to solve many classification problems.
2. Logistics Regression: LR is used for analysing binary response data. The outcome variable Y is denoted by 1 ("success") or 0 ("failure"). The logistic model of any variable, Y, is defined as

1/(e^(-Y)+1) (1)

## Data Mining Approach

The predictive accuracy for both models was estimated by the area under the receiver operating characteristics (AUROC) curves Table II. Higher AUROC indicated a better-predicted accuracy of the models. The total classification rate for both models was determined by comparing the predicted events with the actual events Table III. The data management and data analysis for this study has been carried out using R.

# Results and Discussion

The ANN-based approaches used 3-layer networks and the relative weights of neurons to predict EBF. To select the most appropriate ANN, the 10 perceptron models with 16 neurons in the input layer, one neuron in and kappa statistics showed that the ANN with 30 neurons in the hidden layer. The activation functions of logistic sigmoid and hyperbolic tangent are used in each neuron of the hidden layer and output layer, respectively, to compare with other neural networks, had a better performance. Therefore, the desirable neural network that should be compared with the logistic regression model was a neural network with 30 neurons in the hidden layer.

1. Selecting the best ANN model using acuracy rate, AUROC and kappa statistics

| Number of Neurons | Accuracy Rate (%) | AUROC | Kappa Statistic |
| --- | --- | --- | --- |
| ANN (5) | 81.21 | 0.7872 | 0.5959 |
| ANN (10) | 84.08 | 0.8337 | 0.6691 |
| ANN (15) | 94.75 | 0.9440 | 0.8907 |
| ANN (20) | 96.66 | 0.9644 | 0.9305 |
| ANN (25) | 97.93 | 0.9776 | 0.9570 |
| ANN (30) | 98.57 | 0.9854 | 0.9703 |
| ANN (35) | 97.61 | 0.9731 | 0.9503 |
| ANN (40) | 98.41 | 0.9823 | 0.9669 |
| ANN (45) | 98.25 | 0.9815 | 0.9637 |
| ANN (50) | 97.93 | 0.9758 | 0.9568 |

The AUROC for ANN (30) and logistic models separately were compared with the reference AUROC. The results showed that ANN (30) model was significantly more outperformed than the logistic regression in terms of discrimination calibration, and accuracy (cutoff point 0.5). Compared to the MLR model, the ANN model had a superior accuracy rate. The MLR and ANN (30) models classified 87.06% and 96.38% of babies respectively, with and without ARI correctly. The MLR have sensitivity and specificity were 99.00% and 42.91%, respectively, and the ANN had a sensitivity and specificity of 100% and 12.40%, respectively. One of the diagnosis criteria for comparing the models is the area under the ROC curve that for which values 0 to 0.5 show a random classification, and values 0.5 to 1 indicate the model’s total diagnosis capacity. According to table 4, the AUROC curve in the experimental set for logistic regression and neural network models were obtained as 78.2% and 93.9%, respectively.

1. Selected Feture

| Model | Accuracy rate (%) | Sensitivity (%) | specificity (%) | PPV | NPV | AUROC |
| --- | --- | --- | --- | --- | --- | --- |
| LR | 70.06% | 0.5843 | 0.7802 | 0.6450 | 0.7330 | 0.6822 |
| ANN | 98.57% | 0.9843 | 0.9866 | 0.9805 | 0.9892 | 0.9390 |

Also, the kappa statistics for LR was 0.452, showing that the emerged classification may be due to chance and this statistic for the ANN was 0.60 which was significant; showing that the emerged classification was not due to chance.

# Discussion

In this study, we developed an individual classification method and GA model for classifying school attrition data. The results of the study show that the genetic algorithm got a good result for a small range of initial parameters. In most cases, the difference between the accuracy of the classification reported with and without GA is very small. Overall, the GA feature selectors created better classification accuracy without imply to this method. The main advantage of this approach is that it belongs to the field of controllable because GA can be secured for better results all the time by changing the fitness functions.

# Conclusion

In general, the results of this study showed that among 10 ANNs, an ANN with 30 neurons in the hidden layers had better performance. In comparison with the conventional LR model, the ANN model in the study was more accurate in predicting academic failure and had higher overall performance indices. Therefore, based on the results of other and academic failure data, it seems that for classification of a dichotomous dependent variable, artificial neural network methods are appropriate to be used.

Finally, the application of the model and the likelihood that the required assumptions will be true will determine the best model selection methodology. In conclusion, while choosing which approach to utilize, we must consider the presumptions made for each.

##### References

[1] M. J. Sankar *et al.*, “Optimal breastfeeding practices and infant and child mortality: A systematic review and meta-analysis,” *Acta Paediatrica, International Journal of Paediatrics*, vol. 104. pp. 3–13, Dec. 2015, doi: 10.1111/apa.13147.

[2] W. H. Oddy, “Breastfeeding, Childhood Asthma, and Allergic Disease,” *Ann. Nutr. Metab.*, vol. 70, no. 2, pp. 26–36, 2017, doi: 10.1159/000457920.

[3] F. S. Santos, L. H. dos Santos, P. C. Saldan, F. C. S. Santos, A. M. Leite, and D. F. Demello, “Breastfeeding and acute diarrhea among children enrolled in the family health strategy,” *Texto e Context. Enferm.*, vol. 25, no. 1, 2016, doi: 10.1590/0104-070720160000220015.

[4] C. G. Victora *et al.*, “Association between breastfeeding and intelligence, educational attainment, and income at 30 years of age: a prospective birth cohort study from Brazil,” *Lancet Glob. Heal.*, vol. 3, no. 4, pp. e199--e205, Apr. 2015, doi: 10.1016/S2214-109X(15)70002-1.

[5] K. Brigham, S. Gupta, and J. C. Brigham, “Predicting responses to mechanical ventilation for preterm infants with acute respiratory illness using artificial neural networks,” *Int. j. numer. method. biomed. eng.*, vol. 34, no. 8, p. e3094, Aug. 2018, doi: 10.1002/cnm.3094.

[6] Y. J. Chang *et al.*, “Predicting hospital-acquired infections by scoring system with simple parameters,” *PLoS One*, vol. 6, no. 8, Aug. 2011, doi: 10.1371/journal.pone.0023137.

[7] B. Ray and R. Chunara, “Predicting Acute Respiratory Infections from Participatory Data,” *Online J. Public Health Inform.*, vol. 9, no. 1, May 2017, doi: 10.5210/ojphi.v9i1.7650.

[8] L. Gabutti, M. Burnier, G. Mombelli, F. Malé, L. Pellegrini, and C. Marone, “Usefulness of artificial neural networks to predict follow-up dietary protein intake in hemodialysis patients,” *Kidney Int.*, vol. 66, no. 1, pp. 399–407, Jul. 2004, doi: 10.1111/j.1523-1755.2004.00744.x.

[9] P. Maiellaro, R. Cozzolongo, and P. Marino, “Artificial Neural Networks for the Prediction of Response to Interferon Plus Ribavirin Treatment in Patients with Chronic Hepatitis C,” *Curr. Pharm. Des.*, vol. 10, no. 17, pp. 2101–2109, Mar. 2005, doi: 10.2174/1381612043384240.

[10] W. K. Lim and M. J. Er, “Classification of mammographic masses using generalized dynamic fuzzy neural networks,” in *Medical Imaging 2003: Image Processing*, May 2003, vol. 5032, no. 15, p. 1392, doi: 10.1117/12.481104.

[11] C. Loukas and P. Brown, “Online prediction of self-paced hand-movements from subthalamic activity using neural networks in Parkinson’s disease,” *J. Neurosci. Methods*, vol. 137, no. 2, pp. 193–205, Aug. 2004, doi: 10.1016/j.jneumeth.2004.02.017.

[12] Y. Oner, T. Tunc, E. Egrioglu, and Y. Atasoy, “Comparisons of Logistic Regression and Artificial Neural Networks in Lung Cancer Data,” *Am. J. Intell. Syst.*, vol. 3, no. 2, pp. 71–74, 2013, Accessed: Apr. 01, 2021. [Online]. Available: http://article.sapub.org/10.5923.j.ajis.20130302.03.html.

[13] A. Morteza, M. Nakhjavani, F. Asgarani, F. L. F. Carvalho, R. Karimi, and A. Esteghamati, “Inconsistency in albuminuria predictors in type 2 diabetes: A comparison between neural network and conditional logistic regression,” *Transl. Res.*, vol. 161, no. 5, pp. 397–405, 2013, doi: 10.1016/j.trsl.2012.12.013.

[14] M. Parsaeian, K. Mohammad, M. Mahmoudi, and H. Zeraati, “Comparison of logistic regression and artificial neural network in low back pain prediction: Second national health survey,” *Iran. J. Public Health*, vol. 41, no. 6, pp. 86–92, 2012, Accessed: Apr. 01, 2021. [Online]. Available: http://ijph.tums.ac.ir.