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

**Mohammad Nayeem Hasan1,** **Sabrin Sultana2**

1Department of Statistics, Shahjalal University of Science & Technology, Sylhet-3114, Bangladesh, [nayeem5847@gmail.com](mailto:nayeem5847@gmail.com)

2Department of Banking and Insurance, University of Chittagong, Chittagong, Bangladesh,

[sabrinsultana1060@gmail.com](mailto:sabrinsultana1060@gmail.com)

Corresponding Author: Mohammad Nayeem Hasan, [nayeem5847@gmail.com](mailto:nayeem5847@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.

**Materials and Methods**

***Study Population:***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). 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.

Table 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 |

***Artificial Neural Network:*** Artificial neural network is a data processing mechanism generated by the simulation of human nerve cells and nervous system in a computer environment. The ANN used in this study was a standard feed-forward, back-propagation neural network with three layers: an input layer, a hidden layer, and an output layer. A multilayer perceptron (MLP) network is an emerging tool for designing special classes of layered feed-forward networks.

The number of nodes in the input and output layers is determined based on the data structure, but ﬁnding the optimum number of hidden nodes is a crucial step in the architecture of the neural network. The most popular method of finding the optimal number of hidden nodes is cross-validation technique. A feed-forward network, the neurons in each layer only connect with the neurons in the next layer with their weights. These connections are unidirectional, which means signals or information being processed can only pass through the network in a single direction, from the input layer, through the hidden layer(s). A MLP is usually trained by a backpropagation (BP) algorithm with forward and backward phases. When the ANN is trained using the back-propagation algorithm the weights and biases are optimized. The objective function employed for optimization is the sum of the squares of the difference between a desirable output and an estimated output. In the MLP with back error propagation, transfer function in the hidden layer neurons is a nonlinear function such as hard limit, linear or sigmoid which able to associate training patterns with outputs. But for simple derivative and related derivatives with function, mostly used sigmoid function (Jiang et al., 2010).

***Logistic Regression:*** A well-known statistical method for modeling a binary response variable is logistic regression. Logistic regression examines the relationship between a binary outcome (dependent) variable such as presence or absence of disease and predictor (explanatory or independent) variables. The outcome variables can be both continuous and categories. The parameters of logistics regressions are estimated via maximizing logarithmic likelihood function. Logistic regression models generally include only the variables that are considered “important” in predicting an outcome. The stepwise method, backward and forward selection methods are generally preferred in the literature for selecting independent variables.

***Statistical Analysis:*** To fit ANN model to the dataset, first assume a standard feed-forward, back-propagation neural network with three layers: an input layer, a hidden layer and an output layer. The input layer consists of 16 neurons; in order to prevent the over-fitting of data, the hidden layer contained a different number of neurons such as 10, 15, 20, 25, 30, 35, 40, 45 and 50 and the output layer contained two neurons. We used the sigmoid function in both hidden layer and output layer for activation. Logistic regression model used to predict the EBF based on effective factors the independent variable is previously discussed 16 factors.

***Performance estimation:*** The Area under the Receiver Operating Characteristic (AUROC), the indicators of sensitivity, specificity and kappa coefficient used to compare ANNs with different neurons and also for comparing the best ANN with logistic regression curves to evaluate the predictive accuracy of two models. The higher ROC areas indicated a better performance of the models. The statistical analysis and data management for this study had been carried out using R.

**Results**

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.

Table 2: Selecting the best ANN model using Accuracy rate, Area under ROC curve and Kappa

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of Neurons | Accuracy Rate (%) | Incorrect Prediction | 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 (Table 4). The results showed that ANN (30) model was significantly more out-performed 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.

Table 4: Comparison of performance indices of the ANN and LR models for predicting ARI

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy rate (%) | Sensitivity (%) | specificity (%) | PPV | NPV | AUROC |
| LDA | 63.85% | 0.8235 | 0.5121 | 0.5357 | 0.8093 | 0.6680 |
| LR | 70.06% | 0.5843 | 0.7802 | 0.6450 | 0.7330 | 0.6822 |
| RF | 97.45% | 0.9490 | 0.9920 | 0.9878 | 0.9661 | 0.9705 |
| 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 (Table 4).

Table 3: Classification of ARI based on LR and ANN (30) models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model |  | No | Yes | Kappa Statistic | P-value |
| Linear Discriminant Analysis | No | 210 | 182 | 0.3093 | 0.012 |
| Yes | 45 | 191 |
| Logistic Regression | No | 149 | 82 | 0.3700 | <0.001 |
| Yes | 106 | 291 |
| RF | No | 242 | 3 | 0.9469 | <0.001 |
| Yes | 13 | 370 |
| ANN with 30 Neuron | No | 251 | 5 | 0.9703 | <0.001 |
| Yes | 4 | 368 |

Results of ANN (30) showed that respectively BMI, wealth index, source of drinking water, child age, and mother age were five effective factors on ARI.

|  |
| --- |
|  |

**Figure 2: Variable Importance**

**Discussion**

In this study, after changes in the number of hidden layer neurons, an MLP with 30 neurons in the hidden layer was the best structure among 10 ANN models and then it was chosen to compare with LR model. But the result showed that by increasing the number of neurons in hidden layer cannot certainly affected on the performance of the neural network.

Significantly, the AUROC for ANN (30) model (94%) was considerably higher than LR model (78%). Also, sensitivity for two models showed that the classification accuracy of EBF in ANN (96.38%) was more accurate than the LR model (87.06%). So far, many studies to compare ANN and LR models in various fields such as medicine, economics, agriculture, etc. have been conducted but there are a few studies in our field. In a study almost the same, results showed that the correct classification in predicting graduate students for the ANN (93.3%) was higher than the discriminant analysis (81.5%) (16). Also, in other studies showed that ANN model is far better than the LR model (7, 11, 17-20). Although most studies indicated ANNs are a technique alternative to conventional statistical methods for predicting, but there were a few studies which showed LR model have the same or better performance (10, 11). In addition, the results of a meta-analysis study showed that in 36% of cases ANN and 14% LR performed better and in others both models functioned well (10).

In general ANN methods as semi-parametric methods have many advantages such as, allow a large number of variables in the model, no need to assumptions such as normality and, finding the models despite missing data, detection of complex and nonlinear relationship between independent and dependent variables (21).

Logistic regression model is mainly influenced by the sample size, number of independent variables, potential multicollinearity and missing (23). The ANN model indicated respectively 10 factors, education, place of residence, wealth index, C-section, post-natal care, mass media, mother’s age, disease, child age and division had the greatest impact on EBF.

**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.

**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.