**Abstract:**

**Background:** Acute respiratory infections (ARI) are one of the most important global cause of morbidity and mortality of under-five children and come up with significant health complications for developing countries like Bangladesh.

**Objectives:** In this study, we compared the performance of LR and ANN models in predicting ARI of under-five children in Bangladesh.

**Methods:** The cross-sectional data were collected from the 2014 Bangladesh Demographic and Health Survey (BDHS), which designed to provide information on demographic indicators, maternal health, and children's health and nutritional status. After excluding missing values and variables that are not related with child ARI, a total of 966 children under-five years were acceptable for our study. Children with a cough coincide with chest related short, rapid breathing in the 2 weeks before the study were assumed having an ARI. Artificial Neural Network (ANN) and logistic regression (LR) model were used to predict ARI in this study. The efficiency of two models was compared by receiver operating characteristic (ROC) curve and accuracy rate.

**Results:** By using accuracy rate and Kappa statistic Among the 10 ANNs, the ANN with 30 neurons in the hidden layer was considered the best ANN model on the basis of highest accuracy rate and kappa statistic. The results showed that the accuracy rate and Area Under Receiver Operating Characteristic (AUROC) curve for ANN (30) model was 96.38% and 93.9% while for LR model it was 87.06% and 78.2%.

**Conclusion:** In comparison with conventional LR model, the ANN models in this study appeared to be more accurate in predicting ARI of under-five year children.

**Introduction:**

For classifying/predicting of the binary outcome variable (ARI), a number of methods have been developed such as linear discriminant analysis, logistic regression analysis, different data mining 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 Acute Respiratory Infection (ARI), Lung Cancer, Breast Cancer, Low Back Pain, Psychological Symptom, Hospital-acquired infections (HAI), Academic failure; various machine learning algorithm specially Artificial Neural Network (ANN) were used widely in association with some traditional statistical methods such as logistic regression analysis and discriminant analysis (Ray & Chunara, 2017; Chang et al., 2017; Brigham et al., 2018; Kononenko, 2001; Teshnizi et al., 2015; Ayer et al., 2010). Uses of ANNs as clinical prediction models has been explored in many areas of medicine, including nephrology (Gabutti et al., 2004), microbiology (Maiellaro et al., 2004), radiology (Lim et al., 2004) and neurology (Loukas et al., 2004). Logistic regression (LR), and artificial neural networks or neural networks (NN) are two of the most commonly used models for data classification in medical diagnosis. **(Oner et al., 2013), (Morteza et al., 2013), (Ayer et al., 2010), (Parsaeian et al., 2012).** The results of a meta-analysis with 28 studies showed that in 36% of them, ANN, in 14% logistic regression method, performed better and in other studies (50% of cases) both modes had a similar performance (Song et al., 2005) (**Sargent, 2001).** Many of the studies used ANN as the classification tool as the performance is usually good compared to interpretation made by experienced researcher.

ANN employ nonlinear mathematical models to mimic the human brains own problem-solving process and a machine learning tool that has turned out to be useful for complex pattern recognition problems. ANNs are complex and flexible nonlinear systems with properties not found in other modelling systems. These properties include robust performance in dealing with noisy or incomplete input patterns, high fault tolerance and the ability to generalize from the input data (Eftekhar et al., 2005; Patterson D.W., 1996). ANNs have more complex structures with many extra nodes (called hidden nodes) compared with other models, and this complexity provides ANNs with the power to classify any data with complex relationships (Ayer et al., 2010).

Comparing ANN models with standard statistical generalized linear models such as logistic regression is an important step in the development procedure. 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 ARI prevalence and some of them predicting ARI by using different classification method, but a very little number of studies that attempt to compare the performance of classification methods for predicting ARI. For this reason in this study, we applied logistic regression model and ANN to predict the Acute Respiratory Infection (ARI) based on effective factors and then compared the ability of each of these models to classify ARI among under five years children of Bangladesh.

**2. Materials and Methods**

**2.1. Study Population**

This study analyzed administrative claims data obtained from Bangladesh Demographic and Health Survey (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, including ARI symptoms. A final data set of 966 observations was obtained after excluding non-eligible cases (e.g. visitors and non-surviving children), variables that are not related with child ARI and observation with missing values.

By reviewing the valid literature, the most significant factors that are associated with ARI prevalence are shown in table 1. In this study the output or response variable was Acute Respiratory Infection (ARI). If a child had a cough accompanied by chest related short, rapid breathing in the 2 weeks before the study was defined as suffering from ARI. Thirteen factors including medication for intestinal parasites, child age (in months), mothers education, sex of child, mothers age group (in years), wealth index, type of cooking fuel, place of residence, source of drinking water, toilet facility, division, body mass index (Mother), media 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 |

**2.2. 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 (figure 1). A multilayer perceptron (MLP) network is an emerging tool for designing special classes of layered feed-forward networks (Rumelhart et al., 1986).

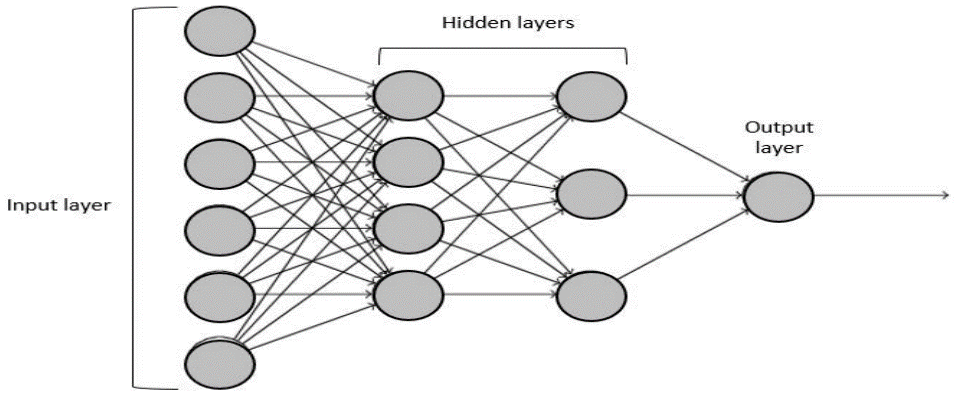


Figure 1: A Typical multilayer feed forward artificial neural network

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. Each weight is shown with directional arrows in figure 1**.** 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 back-propagation (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).

**2.3. 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. If denote n predictor variables, Y denotes the presence (Y = 1) or absence (Y = 0) of disease, and denotes the probability of disease presence (i.e. the probability that Y = 1), the following equation describes the relationship between the predictor variables and :

Where is a constant and are the regression coefficients of the predictor variables. The regression coefficients are estimated from the available data. The probability of disease presence p can be estimated with this equation. 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.

**2.4. 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 13 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 Acute Respiratory Infection (ARI) based on effective factors where the dependent variable is ARI (Yes, No) and the independent variable is previously discussed 13 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 (survey package) and SPSS (IBM SPSS 22).

**W1**

**∑**

**W2**

**W3**

**W4**

**∑**

**b (bias)**

**Results**

The ANN-based approaches used 3-layer networks and the relative weights of neurons to predict ARI. To select the most appropriate ANN, the 9 perceptron models with 13 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 compared 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) | 91.30 | 8.70 | 0.892 | 0.77 |
| ANN (10) | 94.00 | 6.00 | 0.907 | 0.85 |
| ANN (15) | 95.65 | 4.35 | 0.930 | 0.89 |
| ANN (20) | 92.96 | 7.04 | 0.897 | 0.82 |
| ANN (25) | 94.82 | 5.18 | 0.917 | 0.87 |
| ANN (30) | **96.38** | **3.62** | **0.939** | **0.91** |
| ANN (35) | 94.82 | 5.18 | 0.923 | 0.87 |
| ANN (40) | 95.65 | 4.35 | 0.934 | 0.89 |
| ANN (45) | 94.20 | 5.80 | 0.909 | 0.85 |
| ANN (50) | 94.62 | 5.38 | 0.913 | 0.86 |

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 (%) | 1-specificity (%) | AUROC |
| LR | 68.63% | 0.57 | 0.76 | 0.668 |
| ANN | 96.38% | 1.00 | 0.124 | 0.939 |

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 |
| Logistic Regression | No | 146 | 88 | 0.3411 | <0.001 |
| Yes | 109 | 285 |
| ANN with 30 Neuron | No | 684 | 35 | 0.91 | 0.0001 |
| Yes | 0 | 247 |

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

**Figure 2: Variable Importance of final ANN Model**

(Song et al., 2005). Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses. [Jae H. Song](javascript:void(0);), [Santosh S. Venkatesh](javascript:void(0);), [Emily A. Conant](javascript:void(0);), [Peter H. Arger](javascript:void(0);), [Chandra M. Sehgal](javascript:void(0);) ; DOI: <https://doi.org/10.1016/j.acra.2004.12.016>; [April 2005](https://www.academicradiology.org/issue/S1076-6332(00)X0063-3), Volume 12, Issue 4, Pages 487–495 ( Need download)

(Sargent, 2001). Comparison of Artiﬁcial Neural Networks with Other Statistical Approaches Results from medical data sets; Daniel J. Sargent, CANCER Supplement April 15, 2001 / Volume 91 / Number 8, 1636-1642

(Ray & Chunara, 2017). Predicting Acute Respiratory Infections from Participatory Data. Bisakha Ray, Rumi Chunara. ISDS 2016 Conference Abstracts; Online Journal of Public Health Informatics \* ISSN 1947-2579 \* http://ojphi.org \* 9(1): e71, 2017.

(Chang et al., 2017). Predicting Hospital-Acquired Infections by Scoring System with Simple Parameters; Ying-Jui Chang, Min-Li Yeh, Yu-Chuan Li, Chien-Yeh Hsu, Chao-Cheng Lin, Meng-Shiuan Hsu, Wen-Ta Chiu. PLoS ONE, August 2011 | Volume 6 | Issue 8 | e23137. doi: 10.1371/journal.pone. 0023137.t001

(Brigham et al., 2018). BrighamK, GuptaS, BrighamJC. Predicting responses to mechanical ventilation for preterm infants with acute respiratory illness using artificial neural networks. April 2018, International Journal for Numerical Methods in Biomedical Engineering, 34(1) DOI: 10.1002/cnm.3094

(Kononenko, 2001). Machine learning for medical diagnosis: history, state of the art and perspective; [Igor Kononenko](https://www.sciencedirect.com/science/article/pii/S093336570100077X#!), [Artificial Intelligence in Medicine](https://www.sciencedirect.com/science/journal/09333657), [Volume 23, Issue 1](https://www.sciencedirect.com/science/journal/09333657/23/1), August 2001, Pages 89-109, <https://doi.org/10.1016/S0933-3657(01)00077-X>. (Need download)

(Teshnizi et al., 2015). A Comparison of Logistic Regression Model and Artificial Neural Networks in Predicting of Student’s Academic Failure; Saeed Hosseini Teshnizi, Sayyed Mohhamad Taghi Ayatollahi, ACTA INFORM MED, 2015 OCT 23(5): 296-300, doi: 10.5455/aim.2015.23.296-300

(Ayer et al., 2010). Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation; Turgay Ayer, MS • Jagpreet Chhatwal, PhD • Oguzhan Alagoz, PhD Charles E. Kahn, Jr, MD, MS • Ryan W. Woods, MD, MPH • Elizabeth S. Burnside, MD, MPH, MS; RadioGraphics 2010; 30:13–22

(Eftekhar et al., 2005). Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data; Behzad Eftekhar, Kazem Mohammad, Hassan Eftekhar Ardebili, Mohammad Ghodsi and Ebrahim Ketabchi; BMC Medical Informatics and Decision Making; !5 February 2005, 5(3), PP: 1-8, doi: 10.1186/1472-6947-5-3

(Patterson D.W., 1996). Artificial Neural Networks: Theory and Applications; D. W. Patterson, Prentice Hall, Englewood Cliffs, 1996

(Tu J.V., 1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes; [Jack V.Tu](https://www.sciencedirect.com/science/article/abs/pii/S0895435696000029" \l "!), Journal of Clinical Epidemiology 1996; 49(11), PP: 1225–1231.

(Gabutti et al., 2004). Usefulness of artificial neural networks to predict follow-up dietary protein intake in hemodialysis patients. Gabutti L, Burnier M, Mombelli G, Male F, Pellegrini L, Marone C; Kidney Int, 2004, 66(1), PP: 399-407, doi: [10.1111/j.1523-1755.2004.00744.x](https://doi.org/10.1111/j.1523-1755.2004.00744.x)

(Maiellaro et al., 2004). Artificial neural networks for the prediction of response to interferon plus ribavirin treatment in patients with chronic hepatitis C; Maiellaro PA, Cozzolongo R, Marino P; Curr Pharm Des, 2004, 10(17), PP: 2101-2109, doi: [10.2174/1381612043384240](https://doi.org/10.2174/1381612043384240)

(Lim et al., 2004). Classification of mammographic masses using generalized dynamic fuzzy neural networks; Lim WK., Er MJ.; Med Phys 2004, 31(5), PP: 1288-2109, doi: [10.1118/1.1708643](https://doi.org/10.1118/1.1708643)

(Loukas et al., 2004). Online prediction of self-paced hand-movements from subthalamic activity using neural networks in Parkinsons disease. Loukas C, Brown p; Journal of Neuroscience Methods, 2004, 137(2), PP: 193-205, doi:  [10.1016/j.jneumeth.2004.02.017](https://doi.org/10.1016/j.jneumeth.2004.02.017)

(Shaﬁei et al., 2016). Comparison of Artiﬁcial Neural Network and Logistic Regression Models for Prediction of Psychological Symptom Six Months after Mild Traumatic Brain Injury. Elham Shaﬁei, Esmaeil Fakharian, Abdollah Omidi, Hossein Akbari, Ali Delpisheh, and Arash Nademi; Iran J Psychiatry Behav Sci. 2017;11(3): e5849

(Green et al., 2006). Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room; Michael Green, Jonas Bjork, Jakob Forberg, Ulf Ekelund, Lars Edenbrandt, Mattias Ohlsson; Artificial Intelligence in Medicine 2006, 38, PP: 305-318, doi: 10.1016/j.artmed.2006.07.006

(NIPORT et al., 2016). National Institute of Population Research and Training (NIPORT). Mitra and associates; and ICF International. Bangladesh Demographic and Health Survey 2014. Dhaka, Bangladesh and Rockville, Maryland, U.S.A; NIPORT, Mitra and Associates, & ICF International, 2016. 137 pp.

(Rumelhart et al., 1986). Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. In: Rumelhart DE and McCleland JL, editors, Cambridge, MA: MIT Press; 1986, pp 318–362.

(Jiang et al., 2010). Jiang J, Trundle P, Ren J. Medical image analysis with artificial neural networks. Computerized Medical Imaging and Graphics. 2010; 34(8): 617-631.

**(Oner et al., 2013).** Comparisons of Logistic Regression and Artificial Neural Networks in Lung Cancer Data. Yuksel Oner1, Taner Tunc1, Erol Egrioglu1, \*, Yildiz Atasoy2; American Journal of Intelligent Systems 2013, 3(2): 71-74 DOI: 10.5923/j.ajis.20130302.03.

**(Morteza et al., 2013).** Morteza A., Nakhjavani M., Asgarani F., CarvalhoFilipe L.F., Karimi R., Esteghamati A., Inconsistency in albuminuria predictors in type 2 diabetes: a comparison between neural network and conditional logistic regression, Translational Research, 161(5),397-405,2013.

**(Parsaeian et al., 2012).** Comparison of Logistic Regression and Artificial Neural Network in Low Back Pain Prediction: Second National Health Survey. M Parsaeian 1, K Mohammad 1, M Mahmoudi 1, \*H Zeraati 1; Iranian J Publ Health, Vol. 41, No.6, Jun 2012, pp.86-92