**Science Journal of Public Health**
2013; 1(1) : 39-43
Published online March 10, 2013 (http://www.sciencepublishinggroup.com/j/sjph)
doi: 10.11648/j.sjph.20130101.16

# Application of artificial neural network and binary Logistic regression in detection of Diabetes status

**Azizur Rahman[1], Karimon Nesha[2], Mariam Akter[2], Md. Sheikh Giash Uddin[1]**

[1]Department of Statistics, Jagannath University, Dhaka-1100, Bangladesh
[2]Department of Disaster management, University of Dhaka, [2]School of Business, United International University, Dhaka, Bangladesh

**Email address:**
rahman.aziz83@gmail.com (A. Rahman)

**To cite this article:**
Azizur Rahman, Karimon Nesha, Mariam Akter, Md. Sheikh Giash Uddin. Application of Artificial Neural Network and Binary Logistic Regression in Detection of Diabetes Status, *Science Journal of Public Health,* Vol. 1, No. 1, 2013, pp. 39-43.
doi: 10.11648/j.sjph.20130101.16

**Abstract:** Various methods can be applied to build predictive models for the clinical data with binary outcome variables. This research aims to compare and explore the process of constructing common predictive models. Models based on an artificial neural network (the multilayer perceptron) and binary logistic regression were applied and compared in their ability to classifying disease-free subjects and those with diabetes mellitus(DM) diagnosed by glucose level. Demographic, enthropometric and clinical data were collected based on a total of 460 participants aged over 30 years from six villages in Bangladesh that were identified as mainly dependent on wells contaminated with arsenic. Out of 460 participants 133 (28.91%) suffered from DM, 116 (25.27%) had impaired glucose tolerance (IGT) and the remainder 211 (45.86%) were disease free. Among other factors, family history of diabetes and arsenic exposure were found as significant risk factors for developing diabetes mellitus (DM), with a higher value of odds ratio. This study shows that, binary logistic regression correctly classified 73.79% of cases with IGT or DM in the training datasets, 70.96% in testing datasets and 70.4% of all subjects. On the other hand, the sensitivities of artificial neural network architecture for training and testing datasets and for all subjects were 83.4%, 82.25% and 84.33% respectively, indicate better performance than binary logistic regression model.

**Keywords:** Artificial Neural Network (ANN), Binary Logistic (LR), Classification, Diabetes Mellitus (DM)

## 1. Introduction

Diabetes mellitus is a heterogeneous syndrome characterized by elevated blood glucose level. Most of the causes of diabetes mellitus are still unknown. However, impaired insulin secretion from the pancreas or impaired insulin action as a result of insulin resistance in the skeletal muscle, liver and adipose tissue has been noted in the diabetic patients [21]. Genetic disposition and environmental factors are important in the development of diabetes mellitus [22]. Recent studies show that environmental factors are important in the development of diabetes mellitus; among them one of the important environmental factors is arsenic contamination of well water.

Statistical methods such as discriminant analysis and logistic regression have commonly been used to develop models for clinical diagnosis and treatment [5]. But studies published in recent years have reported that the artificial neural network approach improves prediction in several situations including prognosis of breast cancer in women after surgery [17], modeling for surgical decision-making for patients with traumatic brain injury [5] and survival of alcoholic patients with severe liver disease [16]. In contrast, others have reported that artificial neural networks and statistical models yielded similar results [9, 18].

Artificial intelligence has been proposed as a reasoning tool to support clinical decision-making since the earliest days of computing [3, 4, 5, 6, 7]. Artificial neural network is a computer modeling technique based on the observed behaviors of biological neurons [8]. This is a non-parametric pattern recognition method which can recognize hidden patterns between independent and dependent variables [9]. The detailed discussion about this approach is introduced in methods and materials section.

In Bangladesh, a population of some 30-70 million people living in 41 districts out of the 64 are probably exposed to arsenic from drinking water containing >50mg/L level of arsenic for a long period [19]. The exposure probably started in late 1960s when drilling of tube wells began as part of a wide irrigation plan [20]. In another study, reference [19]

further examined the relation between arsenic exposure and glucosuria (taken as a proxy for diabetes mellitus) in subjects. In the Peoples Republic of Bangladesh, there are about 7 million individuals affected by Diabetes Mellitus (DM) in 2011 and with increasing urbanization, the prevalence of DM is rising rapidly. So from this point of view the present investigation has been under taken in order to identify and manage patients with arsenic exposure having DM, especially in groups at higher risk for the disease and its complications with the application of artificial neural network and binary logistic regression [19].

## 2. Methods and Materials

### *2.1. Artificial Neural Network*

Artificial Neural Network (ANN) modeling, a paradigm for computation and knowledge representation, is originally inspired by the aspect of information processing and physical structure of the brain with a web of neural connection (see figure 1). Therefore some writers classified it as a "microscopic", "whole box" system and an expert system as a "microscopic", "black-box" system [1]. Artificial neural networks are used in three main ways: (i) as models of biological nervous system and intelligence, (ii) as real-time adaptive signal processors controllers implemented in hardware for applications such as robots, (iii) as data analytic methods [2].

The main principle of neural network computing is the decomposition of the input-output relationship into a series of linearly separable steps using hidden layers [6]. There are three distinct steps in developing an ANN based solution: i) data transformation or scaling, ii) network architecture definition, when the number of hidden layers, the number of nodes in each layer and the connectivity between the nodes and set, iii) construction of learning algorithm in order to train the network [5,8]. Figure 2 shows the simple architecture of a typical network that consists of an input layer, series of hidden layers, an output layer and connection between them.

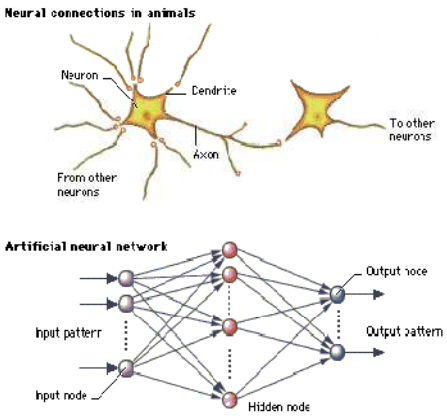


***Figure 1***. *The Neural connection in animals (biological neuron in top) and the counterpart Artificial Neural network structure (in bottom).*

Nodes in the input layer represent possible influential factors that affect the network outputs and have no computation activities, while the output layer contains one or more nodes that produce the network output. Hidden layers may contain a large number of hidden processing nodes. A feed-forward back-propagation network propagates the information from the input layer to the output layers, compares the network outputs with known targets and propagates the error term from the output layer back to the input layer, using a learning mechanism to adjust the weights and biases [5,10].

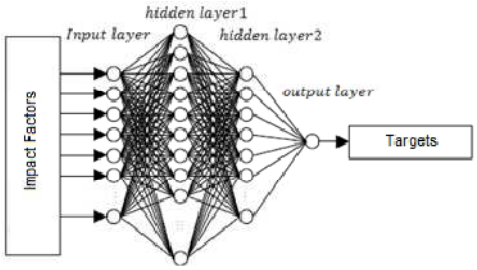


***Figure 2.*** *Simple structure of a typical neural network.*

In 1957, Rosenblatt invented the perceptron, an artificial neuron, in which dendrites are replaced by weighted inputs that are summed inside the artificial neuron and pass through a suitable threshold (activation) [10]. The activated outputs transfer from inner to output layers and produce an output to simulate a desired output (target) at the end. By a learning algorithm, the neural net achieves a form of learning by modifying weights proportional to the difference between the target and the gained output [9]. Artificial neural network have been applied to diagnosis and decision-making in various medical fields [12, 13, 14, 15, 16].

### *2.2. Study Population*

Six villages in two districts of Bangladesh (Cumilla, Jhenidah) were selected for the study on the basis of existing survey reports of arsenic measurements in drinking water. The study was cross-sectional and was performed by the door-to –door visits to interview families with known arsenic concentration in their wells. Eligible subjects were included those who had lived in the study areas throughout their lifetimes and who had used the same well as long as it had existed. A total of 460 subjects >= 30 years of age were identified. A total of 307 individuals had histories of arsenic exposure, were further interviewed by questionnaire, and were examined for identification of diabetes mellitus, according to the American Diabetes Association (ADA) criteria. Among this population, aged >= 30 years (33.83% male and 66.17% female) who had record of DM and had complete information were the subjects of the presents study.

### *2.3. Participants' Demographic and Clinical Characteristics*

Fasting plasma glucose (FPG) level was used to classify the glucose metabolism status of each subject according to

American Diabetes Association (ADA) criteria [21]. A blood sample was drawn into vacutainer tubes between 07:00 and 09.00 hours from all study participants after a 12-14 hour overnight. Subjects were classified as: normal glucose or disease free (FPG<110 mg/dL), IGT(FPG>=110<126 mg/dL) or diabetic (FPG>=126 mg/dL). The demographic and clinical data used as predictors in the models were: patient's age, sex, body mass index (BMI), family history of diabetes, history of Arsenic exposure. Arsenic exposure was defined as any prior diagnosis of this disease by a physician. Weight and height were measured according to standard protocols. BMI was calculated by dividing the weight (in kilogram0 by the square of the height (in meters).

### *2.4. Prediction Models*

We applied two different models to the patient data. The first was a standard binary logistic regression analysis. The second was a standard feed-forward error back-propagation multilayer perceptron with a three layer topology(input, hidden and output layers) with four neurons in the hidden layer (determined by trial and error process) and no direct connection from the input to output layers[11]. The error back propagation learning algorithm is a powerful approach and, despite its slow convergence, is one of the most popular and successful algorithm for pattern recognition [24].

The two different models were compared in their ability to predict glucose metabolism status from the patients' demographic and clinical data. To do this we first merged the objects in the DM and IGT groups. Then we split the database into two groups: a training data-set containing approximately 75% of the sample and testing data-set containing 25% of the subjects. The training dataset was used to develop the logistic regression and perceptron models by introducing the disease status of the subjects into the models. The testing data set was used by the models for classifying the glucose status of subjects.

### *2.4. Software*

The neural network development software used in this study was R, version 2.5.1 package (nnet version 7.2-290). Other statistical analyses were performed by the SPSS version 13.0.

## 3. Results

Among 460 participants aged 30 years or over, 133(28.91%) suffered from DM, 116(25.27%) had IGT and the remainder 211(45.86%) were disease-free by the American Diabetes Association (ADA) criteria.

The mean age in this study was 45.3(standard deviation (SD) 13.045) years overall and 44.12(SD 12.44) years for the disease-free group (Table 1). One way ANOVA indicated that the mean age of the three groups was significantly different and Tukey post hoc multiple comparison test showed that the disease-free group was younger than the DM and IGT patients.

Those in the disease-free group had a higher mean BMI than those in the DM and IGT groups in table 1. The chi-squared test indicated that there was a significant association between glucose level status and history of arsenic exposure (P<0.001). Moreover, table 1 show that the IGT and particularly the DM group had a higher proportion of subjects with a positive family history of diabetes compared with the disease-free group (78.94%,73.27% and 67.29%) and also for history of arsenic (72.94%, 64.65% and 63.98%) for the DM, IGT and disease-free groups respectively).

It is clear that one should specify the training and test dataset before conducting any training neural network architecture. Table 2 illustrates the glucose tolerance status of the training and testing datasets of the sample.

***Table 1.*** *Characteristics of subjects in different glucose status groups.*

| **Variables** | **DF(n=211)** | | **IGT (n=116)** | | **DM(n=133)** | | **Total(n=460)** | |
|---|---|---|---|---|---|---|---|---|
| | **Mean(SD)** | | **Mean(SD)** | | **Mean(SD)** | | **Mean(SD)** | |
| Age(in years) | 44.13(12.45) | | 44.53(12.18) | | 47.82(14.39) | | 45.31(13.04) | |
| Anthropometric Measure BMI(kg/m$^3$) | 20.27(4.73) | | 19.54(3.5) | | 19.78(3.48) | | 19.94(4.11) | |
| Sex | No. | % | No. | % | No. | % | No. | % |
| Male | 63 | 29.85 | 32 | 27.58 | 45 | 33.83 | 140 | 30.43 |
| Female | 148 | 70.15 | 84 | 72.42 | 88 | 66.17 | 320 | 69.57 |
| Hist. of Arsenic | | | | | | | | |
| Yes | 135 | 63.98 | 75 | 64.65 | 97 | 72.94 | 307 | 66.73 |
| No | 76 | 36.02 | 41 | 35.35 | 36 | 27.06 | 153 | 33.27 |
| Family Hist. of Diabetes | | | | | | | | |
| Yes | 142 | 67.29 | 85 | 73.27 | 105 | 78.94 | 335 | 72.82 |
| No | 68 | 32.70 | 31 | 26.72 | 28 | 21.05 | 125 | 27.17 |

**SD=StandardSD=Standard deviation, IGT= Impaired glucose level, BMI= Body mass index, DM=Diabetic Mellitus, DF=Disease-free, History of Arsenic exposure, Family History ofDiabetes*

***Table 2***. *Distribution of glucose status of the sample in the training and testing data sets.*

| Variable | Training datasets | | Testing datasets | | Total |
|---|---|---|---|---|---|
| | No. | % | No. | % | No. |
| **Disease-free** | 158 | 75 | 53 | 25 | 211 |
| **IGT or DM** | 187 | 75 | 62 | 25 | 249 |
| **Total** | 345 | 75 | 115 | 25 | 460 |

**IGT= Impaired glucose level, DM=Diabetic Mellitus, DF=Disease-free*

### *3.1. Comparative Study*

As a common statistical method, we use binary logistic regression and it indicates that all factors were significantly associated with glucose status (Table 3). Age, sex, BMI and Arsenic exposure were significant risk factors for Diabetes Mellitus (DM). Meanwhile, those who were suffering from arsenic disease had a higher risk of DM or IGT.

***Table 3.*** *Odds ratios and coefficients of binary logistic regression analysis of factors associated with glucose status.*

| Characteristics | Coefficient | S.E. | OR | 95.0% C.I.for EXP(B) | |
|---|---|---|---|---|---|
| sex(1) | -.118 | .225 | .889 | .572 | 1.381 |
| age | .020 | .008 | 1.02 | 1.00 | 1.037 |
| bmi | .002 | .026 | 1.00 | .952 | 1.054 |
| parsc(1) | .507 | .236 | 1.66 | 1.04 | 2.636 |
| Hist.of Diabets | .678 | .314 | 1.96 | 1.45 | 2.871 |
| Constant | -2.059 | .775 | .128 | | |

**Sex(1) and parsc(1) are categorical variables*

Table 4 shows the true and predicted status of subjects in the training and testing datasets as well as for all subjects. Binary logistic regression correctly classified 73.79% of cases with IGT or DM in the training datasets, 70.96% in the testing datasets and 70.4% of all subjects. The sensitivities of the neural network architecture for the training and testing datasets and for all subjects were 83.42%, 82.25% and 84.33% respectively (Table 5)

***Table 4.*** *Number of correct diagnosis of glucose status using binary logistic regression model.*

| True Status | Predicted Status using logistic-regression | | |
|---|---|---|---|
| | Disease free No. | IGT or DM No. | Total No. |
| **Training Data** | | | |
| Disease-free | 82 | 76 | 158 |
| IGT or DM | 49 | 138 | 187 |
| Total | 131 | 214 | 345 |
| **Testing Data** | | | |
| Disease-free | 19 | 34 | 53 |
| IGT or DM | 18 | 44 | 62 |
| Total | 37 | 78 | 115 |
| **Overall** | | | |
| Disease-free | 124 | 87 | 211 |
| IGT or DM | 74 | 175 | 249 |
| Total | 198 | 262 | 460 |

***Table 5.*** *Number of correct diagnosis of glucose status using artificial neural network architecture.*

| True Status | Predicted Status using ANN architecture | | |
|---|---|---|---|
| | Disease free No. | IGT or DM No. | Total No. |
| Training Data | | | |
| Disease-free | 82 | 76 | 158 |
| IGT or DM | 31 | 156 | 187 |
| Total | 113 | 232 | 345 |
| Testing Data | | | |
| Disease-free | 19 | 34 | 53 |
| IGT or DM | 11 | 51 | 62 |
| Total | 30 | 85 | 115 |
| Overall | | | |
| Disease-free | 124 | 87 | 211 |
| IGT or DM | 39 | 210 | 249 |
| Total | 163 | 297 | 460 |

# 4. Discussion

The study by reference [25] carried out in BFD area in Taiwan is the first cross-sectional epidemiologic study indicating an association between arsenic exposure form drinking water to diabetes mellitus. The method used to diagnose diabetes mellitus is classical and followed the criteria of the World Health Organization (WHO). The dose-response relation between arsenic exposure and the prevalence of diabetes mellitus is persuasive after adjusting for possible confounders. However, the temporality of cause and effect is not well clarified by a cross-sectional study [22]. The studied carried out in Bangladesh are more recent. The first report in 1998 [26] demonstrated a higher risk for diabetes mellitus in a group of 163 subjects with keratosis (used as an indicator of arsenic exposure) while compared with 854 external controls without exposure. A significant trend for increased risk of diabetes mellitus was observed for increasing dosage of arsenic exposure. In our study, it indicates the same significant effect of increasing dosage of arsenic exposure in development of diabetes mellitus.

# 5. Conclusions

In this study, we used the primary database of the patients of diabetes mellitus with arsenic contaminated well water to develop models to try to distinguish patients with IGT or DM from disease-free patients.

The accuracy of the artificial neural network and binary logistic regression models in predicting a subject's glucose status were compared. Here, binary logistic regression correctly classified 73.79% of cases with IGT or DM in the training datasets, 70.96% in the testing datasets and 70.4% of all subjects. The sensitivities of the neural network ar-

chitecture for the training and testing datasets and for all subjects were 83.42%, 82.25% and 84.33% respectively. Thus we conclude that this study demonstrate a significant performance of artificial neural network than the binary logistic regression models in detection of IGT and DM patients from disease-free ones. Finally, we clearly state that this results only apply to a population with the same characteristics and that models which applied here cannot universally be applied to all diabetics.

## References


[1] Y. L. Eldon, "Artificial Neural Networks and Their Business Applications", Inf and Mang, 1994, 27(5): 303-313.

[2] S. S. Warren, "Neural Networks and Statistical Models", Preceedings of the 19the Annual SAS Users Group International conference, April 1994, USA.

[3] A. A. Betanzos, "Applying statistical uncertainty-based and connectionist approaches to the prediction of fetal outcome: a comparative study," Arti Intelli in Med, 1999, 17(1): 37-57.

[4] P. J. A. Lisboa,"A review of evidence of health benefit from artificial neural networks in medical intervention," Neural Networks, 2002, 15: 11-39.

[5] Y. C. Li, W. T. Chui, W. S. Jian, "Neural networks modeling for surgical decisions on traumatic brain injury patients," Int J Med Info, 2000, 57: 389-405.

[6] W. B. Schwartz, "Medicine and the computer: the promise and problems of change," New England Journal of Medicine, 1970, 283: 1257-64.

[7] E. H. Shortliffe, "The edolesence of Al in medicine: will the field come of age in the '90s?," Arti Intelli in Med, 1993, 5(2): 93-106.

[8] J. Park, D. E. Edington, "A sequential neural network model for diabetes prediction," Arti Intelli in Med, 2001, 23(3): 277-93.

[9] U. Ergun, "Classification of carotid artery stenosis of patients with diabetes by neural networks and logistic regression," Comp in Bio and Med, 2004, 34: 389-405.

[10] F. Rosenblatt, "The perceptron: a perceiving and recognizing automation," Cornell Aeronautical Laboratory report 85-460-I. Ithaca, New York, Cornell Aeronautical Laboratory, 1957.

[11] C. M. Bishop, "Neural networks for pattern recognition," 4th edition. Oxford, Oxford University Press, 1995.

[12] A. L. Ronco, "Use of artificial neural networks in modeling associations of discriminant factors: towards an intelligent selective breast cancer screening," Arti Intelli in Med, 1999, 16(30): 299-309.

[13] R. L. Kennedy, "An artificial neural network system for diagnosis of AMI in the accident and emergency department: evaluation and comparison with serum myoglobin measurements," Comp Meth and Prog in Bio, 1997, 52(2): 93-103.

[14] S. S. Cross, "Image analysis of low magnification images of fine needle aspirates of the breast produces useful discrimination between benign and malignant cases," Cytopathology, 1997, 8: 265-73.

[15] R. Dybowski, and V. Gant, "Artificial neural network in pathology and medical laboratories," Lancet, 1995, 346: 1203-7.

[16] P. Lapuerta, S. Rajan, and M. Bonacini, "Neural networks of outcomes in alcoholic patients with severe liver disease," Hepathology, 1997, 25: 302-306.

[17] P. J. A. Lisboa, "A Bayesian neural networks approach for modeling censored data with an application to prognosis after surgery for breast cancer," Arti Intelli in Med, 2003, 28(1): 1-25.

[18] E. Tafeit, "The determination of three subcutaneous adipose tissue compartments in non-insulin-dependent diabetes mellitus women with artificial neural networks and factor analysis," Arti Intelli in Med, 1999, 17: 181-193.

[19] M. Rahman, M. Tondel, I. A. Chowdhury, and O. Axelson, "Relations between exposure to arsenic, skin lesions, and glucosuria," Occup Env Meth, 1999, 56: 277-281.

[20] P. Bangla, and J. Kaiser, "India's spreading health crisis draws global arsenic experts," Science , 1996, 274: 174-175.

[21] R. A. DeFronzo, R. C. Bonadonna., and E. Ferrannini, Pathogenesis of NIDDM. In: Alberti, K.G.M.M., Zimmet, P., DeFronzo, R.A., Keen, H. (Eds.), "International Test book of Diabetes Mellitus," 2nd edition. Wiley, New York, 1997, pp. 635-711.

[22] C. H. Tseng, C. P. Tseng, H. Y. Chiou, Y. M. Hsueh, C. K. Chong, and C. J. Chen," Epidemiologic evidence of diabetogenic effect of arsenic," Toxicology Letters, 2002, 133(1): 69-76.

[23] M. Rahman, M. Tondel, S. A. Ahmed, and O. Azelson, "Diabetes mellitus associated with arsenic exposure in Bangladesh", Am J of Epi, 1996, 148: 196-203.

[24] A. Kazemnejad, Z. Batvandi and J. Faradmal. "Comparison of artificial neural network and binary logistics regression for determination of impaired glucose tolerance/diabetes", Eastern Mediterranean Health Journal, 2010, 16(6):615-630.

[25] M. S. Lai, Y. M. Hsueh., C. J. Chen, et al. ., "Ingested inorganic arsenic and prevalence of diabetes mellitus," Am J of Epidemiol, 1994, 139: 484-492.

[26] M. Rahman, M. Tondel, S. A. Ahmed, and O. Azelson, "Diabetes mellitus associated with arsenic exposure in Bangladesh,"Am J of Epidemiol, 1998, 148: 198-203.