International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:9, No:9, 2015

# Performance Analysis of Artificial Neural Network with Decision Tree in Prediction of Diabetes Mellitus

J. K. Alhassan, B. Attah, S. Misra

***Abstract*—**Human beings have the ability to make logical decisions. Although human decision - making is often optimal, it is insufficient when huge amount of data is to be classified. Medical dataset is a vital ingredient used in predicting patient's health condition. In other to have the best prediction, there calls for most suitable machine learning algorithms. This work compared the performance of Artificial Neural Network (ANN) and Decision Tree Algorithms (DTA) as regards to some performance metrics using diabetes data. WEKA software was used for the implementation of the algorithms. Multilayer Perceptron (MLP) and Radial Basis Function (RBF) were the two algorithms used for ANN, while RegTree and LADTree algorithms were the DTA models used. From the results obtained, DTA performed better than ANN. The Root Mean Squared Error (RMSE) of MLP is 0.3913 that of RBF is 0.3625, that of RepTree is 0.3174 and that of LADTree is 0.3206 respectively.

***Keywords*—**Artificial neural network, classification, decision tree, diabetes mellitus.

## I. Introduction

WITH the current trend of computerization in hospitals, a huge amount of data is collected. Although human decision-making is often optimal, it is insufficient when there are huge amounts of data to be classified. Medical data mining has great potential for exploring hidden patterns in the data sets of medical domain. These patterns can be used for clinical diagnosis [1]. Data Mining is a technology used to describe knowledge discovery and to search for significant relationships such as patterns, association, and changes among variables in databases. The discovery of those relationships can be examined by using statistical, mathematical, artificial intelligence and machine learning techniques to enable users to extract and identify greater information and subsequent knowledge than simple query and analysis approaches [2]. Neural network techniques have the potential to handle complex, nonlinear problems in a better way when compared to traditional techniques. However systems developed using neural network model suffer from certain drawbacks like local minima, model over fitting [3].The World Health Organization (WHO) and International Diabetes Federation (IDF) in a global report gave estimate of 347 million people as living with diabetes, with 19.8 million from Africa. As reported by IDF 2013 report, Nigeria has the main number of people with diabetes in Africa, with 3.9 million cases and 4.9 per cent national occurrence rate, [4].

Diabetes also known as Diabetes Mellitus is a disease of the body in which the body does not generate or correctly use insulin, a hormone that is required to change sugar, starch and other food into energy required for daily life. Both genetics and environmental appears to take part in role in its cause. There are two most important types, Type 1 and Type 2. The former is insulin-dependent, an autoimmune disease in which the body does not generate any insulin. Such people have to receive daily insulin injections to maintain life. This often occurs in the children and young adult. Whereas Type 2 is considered non-insulin dependent, it is metabolic disorder resulting from the body incapability to create adequate or correctly use of the insulin. This is common with people from the age of 30.

The objective of this study is to compare performances of various classification methods to diagnose the presence of albuminuria in patients with Type 2 diabetes mellitus. MLP and RBF were the two models used for ANN, while RegTree and LADTree were the DTA models used to predict albuminuria. The classification accuracy, sensitivity, specificity and Youden index of these classification methods in another independent set of data records were evaluated.

## II. Related Studies

Usually standard statistical classification techniques have been used in classification difficulties when dependent variable is dichotomous. Data mining applications with higher accuracy and efficiency any longer are used by researchers, with popular classification techniques similar to artificial neural networks (ANN), decision trees (DT) and random forests (RF) used for medical prediction [5]. These classification techniques determine the predictor associated with outcome in addition to predicting the outcome of a disease. Also, relationships hidden deep into datasets and also identify the risk groups. The performances of classification techniques (MLP, C&RT, LR, RBF and self-organizing feature maps) were compared in order to envisage the occurrence of coronary artery disease by [6]. The best models to predict the survival of breast cancer patients was presented by [7]. For this principle they used ANN, logistic regression (LR), DT and Bayesian model by comparing their performances. Maroco et al. [5] compared discriminant analysis, LR, RF, classification trees, support vector machines, ANN (multilayer perceptron (MLP) and radial basis function (RBF)) for prediction of dementia patients. Ture et al. [8]

J. K. Alhassan is with the Federal University of Technology, Department of Computer Science, Minna, Nigeria (Phone: +234 8035961620; Fax: e-mail: jkalhassan@ futminna.edu.ng).

B. Attah is with the Department of Computer Science, Minna, Nigeria (e-mail: blessingiganya@yahoo.com).

S. Misra is with the Computer Science Department, Covenant University, Otta, Ogun State, Nigeria (e-mail: ssopam@gmail.com).

compared a range of classification techniques to predict control and hypertension groups. They formulated models using LR, flexible discriminant analysis, multivariate adaptive regression splines, chi-squared automatic interaction detector, quick unbiased efficient statistical tree, C&RT, RBF and MLP to predict hypertension. Morteza et al. [9] predicted albuminuria in patients with type 2 diabetes mellitus by using two diverse statistical models, MLP and conditional LR. Meng et al. [10] compared the performance of LR, ANN and DT models for predicting diabetes or prediabetes using common risk factors. Ture et al. [11] investigated the consequence of some hypothetical factors on academic achievement using LR and chi-squared automatic interaction detector method in [11].

### *A. Data Collection*

The Two hundred (200) datasets used for this research was collected from Lagos State University Teaching Hospital (LASUTH), Nigeria. The obtained record has nine variables which are the age of patient, no of exercise done per week by patient, plasma glucose level (mgdl), skin fold thickness (mm), body mass index (kg/m2), Diastolic blood pressure (MgHg), smoking status, Diabetes pedigree type and diabetes probability as shown in Table I.

TABLE I
ATTRIBUTES OF MEDICAL DATASET

| S/N | Variable Name | Description |
|---|---|---|
| 1 | Age | Age of patient |
| 2 | E/W | No of exercise done by patient |
| 3 | PGL | Plasma glucose level |
| 4 | SFT | Skin fold thickness |
| 5 | BMI | Body mass index |
| 6 | DBP | Diastolic blood pressure |
| 7 | SS | Smoking status |
| 8 | DPT | Diabetes pedigree type |
| 9 | DP | Diabetes probability |

### *B. Data Preprocessing*

An important step in the data mining process is data preprocessing. One of the challenges that face the knowledge discovery process in medical database is poor data quality. For this reason data were carefully prepared to obtain accurate and correct results. Most related attributes were chosen for the mining task.

## III. DATA MINING STAGES

The data mining stage was divided into three phases. At each phase all the algorithms were used to analyze the health datasets. The testing method adopted for this research was parentage split that train on a percentage of the dataset, cross validate on it and test on the remaining percentage. Sixty six percent (66%) of the health dataset which were randomly selected was used to train the dataset using all the classifiers. The validation was carried out using ten folds of the training sets. The models were now applied to unseen or new dataset which was made up of thirty four percent (34%) of randomly selected records of the datasets. Thereafter, interesting patterns representing knowledge were identified.

## IV. EVALUATION METRICS

In selecting the appropriate algorithms and parameters that best model the diabetes forecasting variable, the following performance metrics were used:

**Time:** This referred to as the time required to complete training or modeling of a dataset. It is represented in seconds.

**Kappa Statistic:** A measure of the degree of nonrandom agreement between observers or measurements of the same categorical variable. The equation for κ is given in (1)

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \tag{1}$$

where Pr($a$) is the relative observed agreement among raters, and Pr($e$) is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then κ = 1. If there is no agreement among the raters other than what would be expected by chance (as defined by Pr($e$)), κ = 0.

**Mean Absolute Error:** Mean absolute error is the average of the difference between predicted and the actual value in all test cases; it is the average prediction error. The MAE $E_i$ of an individual program $i$ is evaluated by (2):

$$E_i = \frac{1}{n}\sum_{j=1}^{n}\left|\frac{P_{ij} - T_j}{T_j}\right| \tag{2}$$

where $P_{(ij)}$ is the value predicted by the individual program $i$ for fitness case $j$ (out of $n$ fitness cases); and $T_j$ is the target value for fitness case $j$.

For a perfect fit, $P_{(ij)} = T_j$ and $E_i = 0$. So, the MAE index ranges from 0 to infinity, with 0 corresponding to the ideal.

As it stands, $E_i$ cannot be used directly as fitness since, for fitness proportionate selection, the value of fitness must increase with efficiency.

**Mean Squared Error:** Mean-squared error is one of the most commonly used measures of success for numeric prediction. This value is computed by taking the average of the squared differences between each computed value and its corresponding correct value. The mean-squared error is simply the square root of the mean-squared-error. The mean-squared error gives the error value the same dimensionality as the actual and predicted values. The **mean squared error** $E_i$ of an individual program $i$ is evaluated by (3).

$$E_i = \frac{1}{n}\sum_{j=1}^{n}(P_{(ij)} - T_j)^2 \tag{3}$$

where $P_{(ij)}$ is the value predicted by the individual program $i$ for sample case $j$ (out of $n$ sample cases); and $T_j$ is the target value for sample case $j$.

For a perfect fit, $P_{(ij)} = T_j$ and $E_i = 0$. So, the $E_i$ index ranges from 0 to infinity, with 0 corresponding to the ideal.

**Root Relative Squared Error:** Relative squared error is the total squared error made relative to what the error would

have been if the prediction had been the average of the absolute value. As with the root mean-squared error, the square root of the relative squared error is taken to give it the same dimensions as the predicted value. Mathematically, the **root relative squared error** $E_i$ of an individual program $i$ is evaluated by (4):

$$E_i = \sqrt{\frac{\sum_{j=1}^{n}(P_{(ij)} - T_j)^2}{\sum_{j=1}^{n}(T_j - \bar{T})^2}} \quad (4)$$

where $P_{(ij)}$ is the value predicted by the individual program $i$ for sample case $j$ (out of $n$ sample cases); $T_j$ is the target value for sample case $j$; and $\bar{T}$ is given by (5):

$$\bar{T} = \frac{1}{n}\sum_{j=1}^{n} T_j \quad (5)$$

For a perfect fit, the numerator is equal to 0 and $E_i = 0$. So, the $E_i$ index ranges from 0 to infinity, with 0 corresponding to the ideal.

**Relative Absolute Error:** Relative Absolute Error is the total absolute error made relative to what the error would have been if the prediction simply had been the average of the actual values. Mathematically, the **relative absolute error** $E_i$ of an individual program $i$ is evaluated by (6).

$$E_i = \frac{\sum_{j=1}^{n}\left|P_{(ij)} - T_j\right|}{\sum_{j=1}^{n}\left|T_j - \bar{T}\right|} \quad (6)$$

where $P_{(ij)}$ is the value predicted by the individual program $i$ for sample case $j$ (out of $n$ sample cases); $T_j$ is the target value for sample case $j$; and $\bar{T}$ is given by (7).

$$\bar{T} = \frac{1}{n}\sum_{j=1}^{n} T_j \quad (7)$$

For a perfect fit, the numerator is equal to 0 and $E_i = 0$. So, the $E_i$ index ranges from 0 to infinity, with 0 corresponding to the ideal.

## V. Experimental Design

The Artificial Neural Networks and Decision Tree algorithms were used to analyze the health data. The ANN algorithms used were Multilayer Perceptron (MLP) and Radial Basis Function (RBF), and the Decision Tree Algorithms used are RegTree and LadTree.

The ANN models were trained with 500 epochs to minimize the root mean square and mean absolute error. Different numbers of hidden neurons were experimented with and the models with highest classification accuracy for the correctly classified instances were recorded. For the Decision Tree models, each class was trained with entropy of fit measure, the prior class probabilities parameter was set to equal, the stopping option for pruning was misclassification error, the minimum n per node was set to 5, the fraction of objects was 0.05, surrogates was 5, 10 fold cross-validation was used, and generated comprehensive results.

## VI. Results and Discussion

As shown in Table II, two types of Algorithms were used for both Artificial Neural Networks and Decision Tree Model. Multilayer Perceptron (MLP) and Radial Basis Function (RBF) were the two algorithms used for ANN while RegTree and LADTree algorithms were the Decision Trees model used.

TABLE II
Analysis of Performances of ANN and DTA

| Performance Metrics | Artificial Neural Network | | Decision Tree | |
|---|---|---|---|---|
| | MLP | RBF | RepTree | LADTree |
| Time | 5.54 | 0.62 | 0.11 | 0.45 |
| Kappa Statistics | 0.5726 | 0.603 | 0.722 | 0.7157 |
| MAE | 1.883 | 1.2219 | 0.1822 | 0.1606 |
| RMSE | 0.3913 | 0.3625 | 0.3174 | 0.3206 |
| RAE (%) | 44.4585 | 52.3938 | 43.0115 | 37.9196 |
| RRSE (%) | 85.0425 | 78.7989 | 68.9978 | 69.6893 |

For the diabetes probability prediction, MLP algorithm used 5.54 secs to model, with kappa statistic of 0.5726, mean absolute error of 1.883 and root mean square error of 0.3913 while RBF algorithm was modeled within 0.62 secs, with kappa statistic of 0.603, mean absolute error of 1.2219 and root mean square error of 0.3625. In the case of Decision Tree Performance analysis, RepTree algorithm used 0.04 sec to modeled, with kappa statistic of 0.722, mean absolute error of 0.1822 and root mean square error of 0.3174. While LADTree algorithm was modeled within 0.45secs, with kappa statistic of 0.7157 mean absolute error of 0.1606 and root mean square error of 0.3206.

From the result analysis by comparing the techniques, Decision Tree performs better than the Neural Networks based on the error reports, number of correctly classified instances and accuracy rate generated.

## VII. Conclusion

This research work presents two examples of both Decision Tree and Artificial Neural Network building process, of most common data mining techniques. The work revealed that, Decision Tree techniques outperformed Artificial Neural Networks with a lower error metrics and higher correlation coefficient. It also, highlighted the way the stored data about diabetes record could be used in the predicting of the new patient. Prediction can be made, with certain accuracy, the diabetes probability of any patient, with the data regarding some important aspects of the patient health record.

## References

[1] *G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Application of Genetic Algorith Optimized Neural Network Connection Weighs For*

*Medical Diagnosis of Pima Indians Diabetes," International Journal on Soft Computing (IJSC), Vol. 2 No. 2. 2011, pp. 10-15.*

[2] P. Saurabh, "Mining Educational Data to Reduce Dropout Rates of Engineering Student", *International Journal of Information Engineering and Electronic Business*, 2012. Downloaded from http://www.mecs-press.org on Sept., 2014.

[3] Y. Radhika, and M. Shashi, "Atmoshere Temperature Prediction using Support Vector Machines," *International Journal of Computer Theory and Engineering*, Vol. 1 No.1, 2009, pp. 55 – 57.

[4] Z. Bobby, World Health Organization Report on Nigerian Diabetes, Downloaded from http://sunnewsonline.com/new/3-9m-nigerians-diabetic-says-report/ on 24th July, 2015

[5] J. Maroco, D. Silva, M. Guerreiro, A. de Mendonça, I. Santana. "Prediction of dementia patients: A comparative approach using parametric vs. non parametric classifiers," in *Proc.* XIX Congresso Anual da Sociedade Portuguesa de Estatistica, Portuguese, 2011.

[6] Kurt, M. Ture, A.T. Kurum. "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease". *Expert Syst Appl*, vol. 34, pp. 366-374, 2008.

[7] Endo, T. Shibata, H. Tanaka. "Comparison of seven algorithms to predict breast cancer survival". *Biomedical Soft Computing and Human Sciences*, vol. 13, pp. 11-16, 2008.

[8] M. Ture, I Kurt, A.T. Kurum, K. Ozdamar. "Comparing classification techniques for predicting essential hypertension". *Expert Syst Appl*, vol. 29, pp. 583-588, 2005.

[9] Morteza, M. Nakhjavani, F. Asgarani, F.L.F Carvalho, R. Karimi, A. Esteghamati. "Inconsistency in albuminuria predictors in type 2 diabetes: A comparison between neural network and conditional logistic regression". *Translational Research*, vol. 161, pp. 397-405, 2013.

[10] X. Meng, Y. Huang, D. Rao, Q. Zhang, Q. Liu. "Comparison of three data mining models for predicting diabetes or preetes by risk factors". *Kaohsiung J Med Sci*, vol. 29, pp. 93-99, 2013.

[11] M. Ture, Z. Akturk, I. Kurt, N. Dagdeviren. "The effect of health status, nutrition, and some other factors on low school performance using induction technique". *Trakya Univ Tip Fak Derg*, vol. 23, pp. 28-38, 2006.