Global Journal of
Computer Sciences:
Theory and Research

Volume 07, Issue 1, (2017) 1-7

www.gjcs.eu

Estimation of HbA1c value using artificial neural networks

Ali Sasar, Information Systems Engineering, Mugla Sitki Kocman University, Mugla 48000, Turkey.
Osman Ozkaraca, Information Systems Engineering, Mugla Sitki Kocman University, Mugla 48000, Turkey.
Musa Peker*, Information Systems Engineering, Mugla Sitki Kocman University, Mugla 48000, Turkey.
Gurbuz Akcay, Department of Pediatri, Mediklinik Hospital, 20010, Turkey.

Suggested Citation:
Sasar, A., Ozkaraca, O., Peker, M. & Akcay, G. (2017). Estimation of HbA1c value using artificial neural networks. *Global Journal of Computer Sciences: Theory and Research. 7*(1), 1-7.

Received December 21, 2016; revised February 20, 2017; accepted April 14, 2017.
Selection and peer review under responsibility of Prof. Dr. Dogan Ibrahim, Near East University, North Cyprus.

Abstract

Diabetes is a life-long disease that occurs because of ineffectiveness or lack of the insulin hormone. Although the blood sugar, fructose and haemoglobin A1c (HbA1c) values are commonly used for diagnosis, the latter give more accurate results. The HbA1c value gives information about the blood sugar levels over the past 2 to 3 months, which is required for treating diabetes. Follow-up data of diabetic patients have been used in this study. In the classification phase, a feed-forward artificial neural network (ANN) was used to estimate the factors affecting HbA1c. The designed ANN has 26 features used as input parameters. The output layer comprises two outputs: normal (HbA1c < 6.5) and high (HbA1c ≥ 6.5). An accuracy rate of 90.33% was obtained with the proposed method. The results, which show accurate estimation of the HbA1c level parameters, will be used in future studies to investigate which parameter affects the HbA1c levels, and in what way.

Keywords: Neural network, haemoglobin A1c, diagnosis of diabetes disease.

* ADDRESS FOR CORRESPONDENCE: **Musa Peker,** Information Systems Engineering, Mugla Sitki Kocman University, Mugla 48000, Turkey. *E-mail address*: musa@mu.edu.tr / Tel.: +90-252-56-71

1. Introduction

In recent years, diabetes continues to be a major health problem both in the developing and the developed countries. The fact of high blood sugar levels in the body has increased the risk of cardiovascular and other diseases, resulting in the death of 2.2 million people annually [1]. Studies show that the prevalence of the current 382 million diabetic patients in the world today is expected to rise to 471 million by 2035 [2].

Blood sugar, fructoseamine and haemoglobin A1c (HbA1c) values are widely used for the diagnosis of diabetes mellitus. Although the role of insulin in diagnosing diabetes is great, the HbA1c value is used for more accurate results. This is because the HbA1c value gives information about the past 2 to 3 months of blood sugar, which is required in the treatment of diabetes. Thus, this value has started to be used increasingly in recent years. The American Diabetes Association approved the HbA1c test in 2010 as a diabetes diagnostic test. It has been emphasised that the normal range of this value is between 3% and 6%, while 6.5% is chosen as the diagnostic criterion [3, 4].

There are numerous studies carried out on diabetes in the literature. Soltani and Jafarian [5] emphasised the necessity of using methods with a minimum error rate, in order to better diagnose a dangerous disease such as diabetes. They have used the probabilistic artificial neural network (ANN) approach to diagnose type 2 diabetes. Their results showed training accuracy and test accuracy to be 89.56% and 81.49%, respectively.

Amato *et al.* [6] examined the abilities and limitations of ANNs by applying a medical diagnosis on cancer and diabetes data. As a result, ANNs are considered a powerful aid in diagnosis, with advantages of processing big data and rapid diagnosis. However, it has been emphasised that the decision of expert physicians is important for establishing a definite diagnosis. In the study by Leema *et al.* [7], a computer-assisted diagnostic system was developed. In this system, ANNs, particle swarm optimisation and gradient-descent-based backpropagation algorithms are proposed for classification of clinical datasets. The study used Pima Indian Diabetes, Wisconsin Breast Cancer and Cleveland Heart Disease datasets obtained from the California Irvine machine learning repository. As a method, an algorithm that is backpropagation based, differential variant and based on universal knowledge is preferred. The obtained results show that the proposed method achieves an accuracy rate of 85.71% for diabetes, 98.52% for breast cancer and 86.66% for heart disease. Rau *et al.* [8] have proven that there is a strong association between diabetes and the risk of liver cancer. In this study, data mining techniques were used to predict the risk of liver cancer within 6 years of type 2 diabetes patients. The data were obtained from the National Health Insurance Research Database, which covers approximately 22 million individuals. The risk factors were determined from the literature review and some features were obtained by chi-square calculation. Next, classification was carried out using ANN and logistic regression methods. It was reported that the logistic regression method gives better results. Choubey and Paul [9] compared various techniques used to classify the diagnosis of medical diabetes on various datasets. These techniques were analysed and compared based on their advantages, problems and successful performance.

In this study, it has been aimed to use ANNs for analysis of diabetes data. ANNs (a classification technique) are one of the most effective and widely used techniques in various applications, such as the medical diagnosis of diabetic patients. In the scope of this study, numerous features have been presented as input data to the model. Experiments have been carried out using numerous classification methods to evaluate the performance of the neural network during classification.

2. Material and Methods

2.1. Data

Access to the diabetes follow-up data used in this study was provided by Koycegiz, Dalaman and Ortaca state hospitals in Turkey. There are 27 properties in this dataset. The attributes and explanations of the 27 parameters are presented in the Appendix. This dataset contains observations on 963 male and 1289 female patients, between the ages of 6 and 89.

2.2. Feed-Forward Artificial Neural Networks

ANNs are mathematical systems paradigm that are composed of several processing units linked together in a weighted manner. The processing units receive signals from other neurons, combine them, transform and generate a numerical result.

Typically, the processing units roughly correspond to actual neurons and are linked in a network; these also structure neural networks. In this study, the neural network model of the feed-forward neural network is used. Feed-forward ANNs are found primarily in three different layers. These layers are, respectively, the input layer that holds data entering the ANN; the hidden layer or layers in which operations are conducted and educated according to the desired conclusion; and finally, the output layer that shows the output values [10].

3. Application and Experimental Results

The block diagram of the proposed system is presented in Figure 1. In the data pre-processing stage, in order to make the classification process more efficient, 0–1 normalisation process is applied to the data. The min–max method given in Eq. (1) is used as the normalisation method

$$x^{'} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where $x^{'}$ is the normalised data, x_i are input values, x_{min} is the smallest number in the input set and x_{max} is the largest number in the input set. After the normalisation stage, the values are assigned to the instead of missing data, using the random values replacement method in the missing data analysis module of the Orange Software. In the classification phase, feed-forward neural networks have been used.

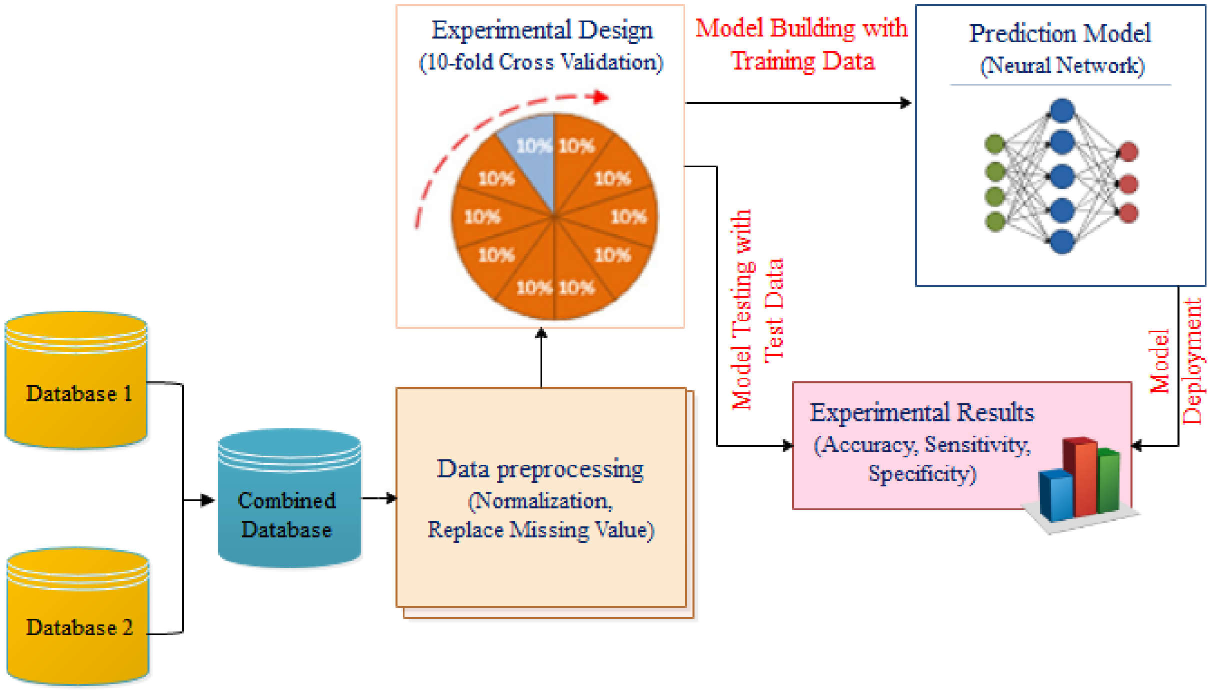

Figure 1. A graphical depiction of the methodology followed in this study

After the data selection and pre-processing stages, the training-test data distribution phase is passed. In this phase, tenfold cross-validation methods are used. Classification procedures are performed in the next stage. In the classification phase, feed-forward neural networks have been used. In this network, the learning coefficient was 0.5, momentum coefficient 0.25 and the number of iterations was set to 1,000. The sigmoid function was used as the hidden layer activation function. The experiments were carried out with four different classification algorithms to evaluate the performance of the feed-forward neural networks in this stage. These algorithms are random forest, support vector machines (SVM), k-nearest neighbour (kNN) and decision trees.

In the last stage, the criterion of statistical evaluation was used to test the effectiveness of the proposed model. In this stage, a number of evaluation methods such as accuracy rate, sensitivity, specificity and area under the ROC curve (AUC) and calibration chart were used. The confusion matrix, which provides a comparison of different classification methods, is given in Table 1. The table shows that the highest accuracy value is obtained with neural network. After this method, the highest success rate is obtained by the kNN algorithm. The lowest success rate was obtained by the random forest algorithm. It is also seen that neural network gives better results in other statistical parameters besides the accuracy rate. It is noteworthy that the high specificity value obtained from random forest algorithm is remarkable. The random forest algorithm predicts one of the two outputs with high accuracy, while the other class almost completely misjudges. This shows that the random forest algorithm fails in the class differentiation.

The success of the proposed method is also assessed with different evaluation criteria. At this stage, ROC curves and calibration graphs were used. The ROC curve is often used to allow the diagnostic test to define its own correctness and to make a reliable comparison between the tests. The AUC for a diagnostic test can range between 0.50 and 1.00. The larger this area, the more likely it is that the diagnostic test will have such a distinction.

Table 1. Confusion matrix

	Accuracy	Sensitivity	Specificity	AUC
kNN	0.8994	0.8289	0.9298	0.8793
SVM	0.7937	0.5308	0.9071	0.8298
Random forest	0.7163	0.0657	0.9970	0.6555
Neural network	0.9033	0.8582	0.9322	0.9492
Decision tree	0.7922	0.5234	0.9082	0.7965

The obtained ROC curves are presented in Figure 2. It is seen from the figure that the area value under ROC is the largest with neural network method. Calibration graphs are presented in Figure 3. In the graphs, the calibration value closest to the target line value was obtained by neural network method. According to different evaluation criteria, it is observed that the neural network method gives successful results in determining the HbA1c value every time.

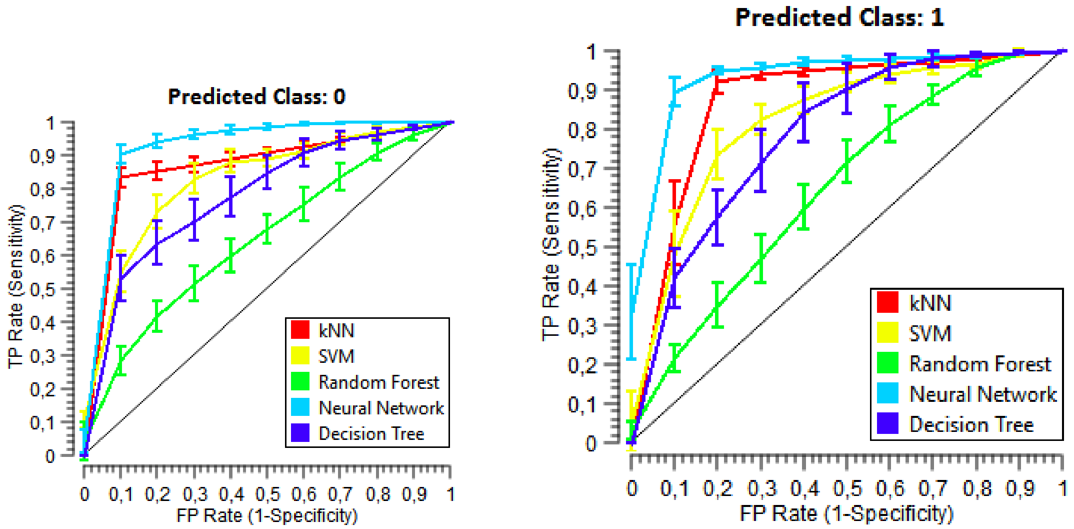

Figure 2. ROC curves

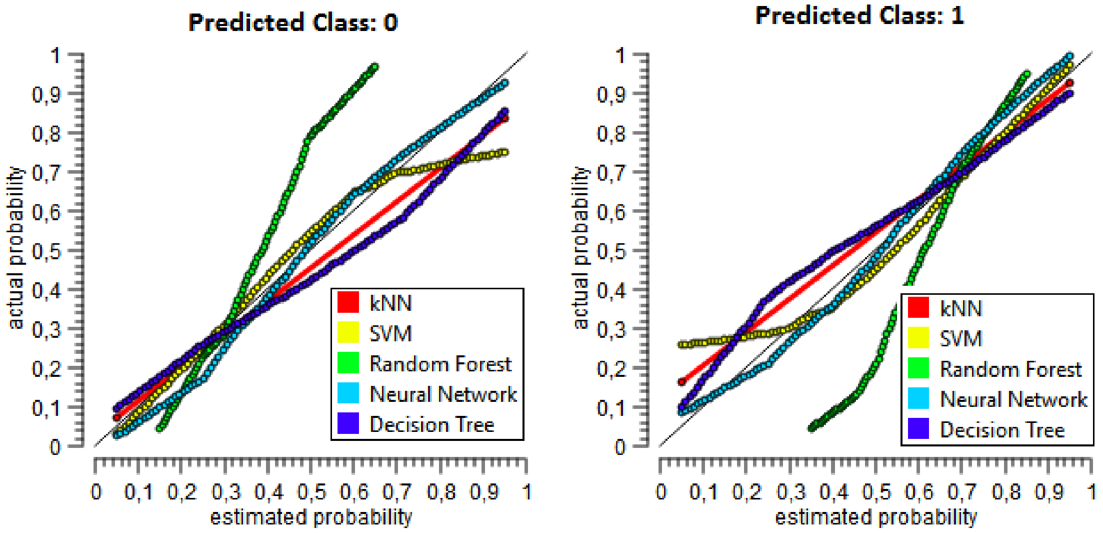

Figure 3. Calibration graph

4. Conclusion

This study suggests a data-mining-based method for detection of high accuracy values of HbA1c, which is an important parameter for diagnosing diabetes mellitus. The proposed method includes data pre-processing, data distribution, classification and performance evaluation steps, which are important steps in data mining. The novelty of this study is the detection of the HbA1c value with effective algorithms on a unique dataset by combining different datasets. Performing missing data analysis on the dataset and experimenting with different methods at the classification stage are remarkable aspects of the study. With the method proposed in this study, HbA1c value is approximately determined with 91% accuracy. In this regard, the proposed method in this study may be a good decision support system for specialists in this area.

Acknowledgements

We are grateful to the General Secretary of the Mugla Public Health Association for their permission. The study was conducted in the Köycegiz, Ortaca, Dalaman State Hospital in Turkey. Ethical clearance was obtained from the ethics committee of Muga Sitki Kocman University (Protocol no = 148).

References

[1] American Diabetes Association, “Diagnosis and classification of diabetes mellitus,” *Diabetes Care*, vol. 33, pp. 62–69, 2010.

[2] M. Brownlee and I. B. Hirsch, “Glycemic variability: a hemoglobin A1c–independent risk factor for diabetic complications,” *JAMA*, vol. 295, issue 14, pp. 1707–1708, 2001.

[3] D. M. Nathan *et al.*, “Medical management of hyperglycemia in type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy,” *Clin. Diabetes*, vol. 27, issue 1, pp. 4–16, 2009.

[4] K. Sikaris, K., “The correlation of hemoglobin A1c to blood glucose,” *J. Diabetes Sci. Technol.*, vol. 3, issue 3, pp. 429–438, 2009.

[5] Z. Soltani and A. Jafarian, “A new artificial neural networks approach for diagnosing diabetes disease type II,” *Int. J. Advanced Comp. Sci. Appl.*, vol. 7, issue 6, pp. 89–94, 2016.

[6] F. Amato *et al.*, “Artificial neural networks in medical diagnosis,” *J. Appl. Biomed.*, vol. 11, issue 2, pp. 47–58, 2013.

[7] N. Leema *et al.*, “Neural network classifier optimization using differential evolution with global information and back propagation algorithm for clinical datasets,” *Appl. Soft Comput.*, vol. 49, pp. 834–844, 2016.

[8] H. H. Rau *et al.*, “Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network,” *Comput Methods Programs Biomed.*, vol. 125, pp. 58–65, 2016.

[9] D. K. Choubey and S. Paul, “Classification techniques for diagnosis of diabetes: a review,” *Int. J. Biomed. Eng. Technol.*, vol. 21, issue 1, pp. 15–39, 2016.

[10] B. Sen *et al.*, “A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms,” *J. Med. Syst.*, vol. 38, issue 3, p. 18, 2014.

Appendix Dataset

LN	Attribute	Values	LN	Attribute	Values
1	Sex	Male: 963 and Female: 1289	15	Cerebrovasculer	(0) Yes and (1) No
2	Age	Min value: 6 and Max value: 89	16	Diabetes education	(1) Yes and (2) No
3	Body mass index	Min value: 15 and Max value: 57	17	Glukometer	(0) Yes and (1) No
4	Waist circumference	Min value: 50 and Max value: 160	18	Fasting blood glucose	Min value: 38 and Max value: 673
5	Exercise	(1) Easy; (2) medium; (3) hard; (4) three times a week or more, at least 20 min; (5) five times a week or more, at least 30 min and (6) more than the first two options.	19	Postprandial blood glucose	Min value: 83 and Max value: 688
6	Medical nutrition	(1) Not confirm; (2) confirm; (3) change list and (4) carbohydrate count	20	Trigliserit: The main component of vegetable and animal oils	Min value: 25 and Max value: 885
7	Sistolic blood pressure	Min value: 60 and Max value: 180	21	HDL: High-density lipoprotein	Min value: 14 and Max value: 109
8	Diastolic blood pressure	Min value: 60 and Max value: 130	22	LDL: Low density lipoprotein	Min value: 12 and Max value:467
9	Thyroid inspection	(1) Troid diffuse is felt by hand; (2) troid is not felt by hand and (3) troid nodular is feeled by hand.	23	ALT: An alanine aminotransferase	Min value: 1 and Max value:116
10	Diagnosis	(1) Type 1; (2) type 2 and (3) others	24	Hip environment	Min value: 60 and Max value:200
11	Medical nutrition treatment	(1) The patient obeys the medical nutrition; (2) the patient sometimes obeys the medical nutrition; (3) the patient doesn't obey the medical nutrition; (4) medical nutrition recommended but we don't know whether the patient obey the medical nutrition or not; (5) medical nutrition is not recommended and (6) nothing is known about the patient.	25	Reason of application	(1) Asymptomatic, (2) others, (3) diabetic ketoacidos, (4) kadiabetic ketosis, (5) hyperglycaemia and (6) hypoglycaemia coma
12	Exercise	(1) Exercise is recommended, patient conforms to; (2) exercise is recommended, patient sometimes conforms to; (3) exercise recommended, patient doesn't conform to; (4) not a suitable patient for exercise; (5) exercise is recommended but we don't know whether the patient obeys the exercise or not; (6) exercise isn't recommended and (7) nothing is known about the patient.	26	Exercise_proposition	(1) Not recommended, (2) contraindicated and (3) recommended
13	Reason of application	(1) General control; (2) routine examination; (3) acute metabolic complication related to the disease and (4) acute chronic complication related to the disease	27	HBA1C: Haemoglobin A_{1C} value	(0) HBA1C value less than 6.5 and (1) HBA1C value more than 6.5
14	Coronary heart disease	(0) Yes and (1) No	28		