Reprint from The Review of DIABETIC STUDIES Vol 9 No 1 2012

Computational Intelligence-Based Diagnosis Tool for the Detection of Prediabetes and Type 2 Diabetes in India

Shankaracharya[1], Devang Odedra[1], Subir Samanta[2], and Ambarish S. Vidyarthi[1]

[1] *Department of Biotechnology, Birla Institute of Technology, Mesra, Ranchi 835215, India.* [2] *Department of Pharmaceutical Sciences, Birla Institute of Technology, Mesra, Ranchi 835215, India. Address correspondence to: Shankaracharya, e-mail: shankaracharya@bitmesra.ac.in*

Manuscript submitted April 13, 2011; resubmitted April 27, 2012; accepted May 4, 2012

■ Abstract

BACKGROUND: The incidence of diabetes is increasing rapidly across the globe. India has the highest proportion of diabetic patients, earning it the doubtful distinction of the 'diabetes capital of the world'. Early detection of diabetes could help to prevent or postpone its onset by taking appropriate preventive measures, including the initiation of lifestyle changes. To date, early identification of prediabetes or type 2 diabetes has proven problematic, such that there is an urgent requirement for tools enabling easy, quick, and accurate diagnosis. **AIM**: To develop an easy, quick, and precise tool for diagnosing early diabetes based on machine learning algorithms. **METHODS**: The dataset used in this study was based on the health profiles of diabetic and non-diabetic patients from hospitals in India. A novel machine learning algorithm, termed "mixture of expert", was used for the determination of a patient's diabetic state. Out of a total of 1415 subjects, 1104 were used to train the mixture of expert system. The remaining 311 data sets were reserved for validation of the algorithm. Mixture of expert was implemented in matlab to train the data for the development of the model. The model with the minimum mean square error was selected and used for the validation of the results. **RESULTS**: Different combinations and numbers of hidden nodes and expectation maximization (EM) iterations were used to optimize the accuracy of the algorithm. The overall best accuracy of 99.36% was achieved with an iteration of 150 and 20 hidden nodes. Sensitivity, specificity, and total classification accuracy were calculated as 99.5%, 99.07%, and 99.36%, respectively. Furthermore, a graphical user interface was developed in java script such that the user can readily enter the variables and easily use the algorithm as a tool. **CONCLUSIONS**: This study describes a highly precise machine learning prediction tool for identifying prediabetic, diabetic, and non-diabetic individuals with high accuracy. The tool could be used for large scale screening in hopsitals or diabetes prevention programs.

Keywords: type 2 diabetes · diabetes diagnosis · prediabetes · computational intelligence · machine learning algorithm · mixture of expert

Introduction

The worldwide incidence of diabetes mellitus is increasing rapidly and has reached epidemic proportions [1]. The consequences for both health and economy are devastating [2]. Type 2 diabetes is traditionally considered a 'silent disease' and patients may appear to be relatively free of symptoms for many years [3]. However, chronic organ complications become more serious the longer impaired glucose metabolism remains undetected and untreated. Therefore, early detection of diabetes is very important so that appropriate action can be taken for the prevention of fatal organ complications.

Because of the subtle course of the disease, the identification of prediabetic states is very difficult. Many factors need to be analyzed to correctly diagnose diabetes, including age, gender, fasting plasma glucose, postprandial glucose, waist circumference, body mass index, family history of diabetes, lifestyle, smoking, drinking of alcohol, exercise, hypertension, diastolic blood pressure, high cholesterol, major food intake, regularity in food, and type of diet (vegetarian or non-vegetarian) [4]. Undoubtedly, the evaluation of data taken from patients, and experts' decisions, are critical for diagnosis. If physicians are inexperienced in dealing with diabetes, erroneous diagnoses could be made, which may subsequently lead to serious late complications because of untreated diabetes. There is a wealth of knowledge from randomized controlled trials showing that early lifestyle changes or medical interventions can prevent type 2 diabetes in a large number of high risk individuals [5-7].

With an estimated 50.8 million people living with diabetes, India has the world's largest diabetes population [8, 9]. Studies reported a high prevalence of undiagnosed diabetes in the Indian community [10, 11]. Overall 99% of diabetics in the South East Asia region live in India, Bangladesh, or Sri Lanka (IDF 2011). Furthermore, an additional 23.8 million people living in these regions have impaired glucose tolerance. Estimates project that this number will increase to 38.6 million by 2030. The expected increase to 10.2% regional prevalence of diabetes in 2030 is a consequence of several factors, including increasing life expectancy, a larger proportion of population over 50 years, and the rapid urbanization in India (IDF, 2011). The vast majority of individuals are ignorant of their disease status, and are thus left untreated. These individuals are prone to micro- and macrovascular complications. It is necessary that they are identified and offered early therapy.

Abbreviations:
BMI - body mass index
BP - blood pressure
CURES - Controlled Substance Utilization Review and Evaluation System
EM - expectation maximization
FPG - fasting blood glucose
HDL - high-density lipoprotein
IDF - International Diabetes Federation
IDRS - Indian Diabetes Risk Score
ME - mixture of expert
MLP - multilayer perceptron
M-step - maximization step
PID - Pima Indian diabetes
PPG - postprandial glucose
RML - Ram Manohar Lohia

Artificial intelligence strategies have been extensively explored and tested on Pima Indian diabetes dataset from the USA [12, 13]. However, it is surprising that even though India has the highest number of diabetes patients, such innovative strategies are relatively unexplored. In this study, the mixture of expert (ME) algorithm has been applied to train and test data obtained from Indian patients in Indian hospitals.

Table 1. Attribute vectors with population characteristics in the training dataset

No.	Attribute	Mean	SD	Type of variable
1	Age (yr)	47.0	13.9	Continuous
2	BMI (kg/m^2)	32.0	6.9	Continuous
3	Waist circumference (cm)	89.4	23.6	Continuous
4	Diastolic BP (mmHg)	86.5	7.4	Continuous
5	Hypertension	-	-	Binary
6	Lifestyle	-	-	Binary
7	Gender	-	-	Binary
8	High cholesterol	-	-	Binary
9	Physical exercise	-	-	Binary
10	Smoking	-	-	Binary
11	Food type (veg/non-veg)	-	-	Binary
12	Major cereal (rice/wheat)	-	-	Binary
13	Family history of diabetes	-	-	Binary
14	Drinking habit	-	-	Binary
15	Class	-	-	Binary

Legend: BMI - body mass index, BP - blood pressure, SD - standard deviation, veg - vegetarian.

Methods

Data collection

Data from more than 1500 diabetic and non-diabetic patients were collected from the Ram Manohar Lohia (RML) Hospital, New Delhi, and the local population of Ranchi during November 2009 to August 2011. The RML Hospital data comprised different Indian populations as the hospital accepts referrals from across the country. The data were collected by questionnaire from patients with various risk factors for diabetes. In this regard, a family history of diabetes was defined as the existence of a first-degree relative (mother or father) having diabetes. High cholesterol was defined as ≥240 mg/dl.

The characteristics of the population in the dataset are described in Table 1. A few datasets (7-8%) were missing in a substantial number of attributes. These datasets were excluded from the study. In some datasets, only one or two attributes were missing. Their values were estimated after normalization. Finally, 1415 samples from the original dataset were considered for further processing; 947 were diabetic and 468 were non-diabetic individuals.

Estimation of mean values for missing attributes in datasets and its normalization

The mean values of the diabetic or non-diabetic population, respectively (Table 2), were used when a single attribute was missing from a particular database. Those datasets with several missing attributes were removed completely to avoid a bias in prediction. All samples were normalized between 0 and 1 to exclude the bias caused by different means, and to account for non-linearity in the sigmoid activation functions.

Partition of the data set into training and testing data

Out of the total 1415 sample datasets, 1104 samples (train dataset) were used randomly to train the ME algorithm. The remaining 311 samples (test dataset) were reserved for testing and validation of the algorithm.

Model development

Various models were created by varying the ME parameters. The following parameters were included:

- Number of experts
- Number of hidden units for the expert multilayer perceptrons (MLPs)
- Number of hidden units for gate MLPs
- Number of iterations
- Expert output activation function
- Type of ME

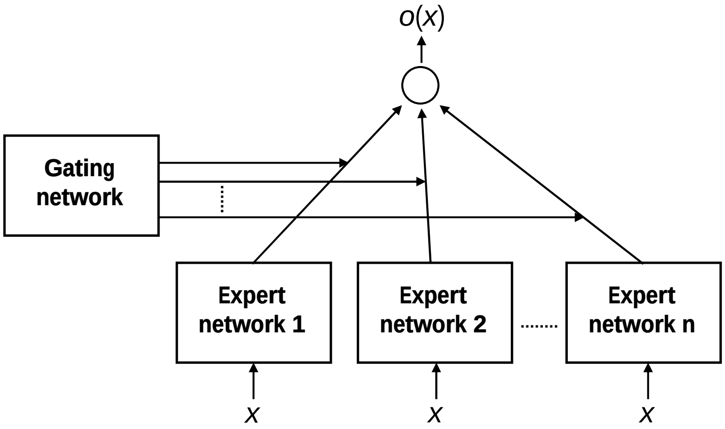


Figure 1. General architecture of the mixture of expert (ME) system. In the ME system, x is the input vector which includes the values of all variables. The value of x is passed through the expert and gating networks. Each expert network is comprised of several multilayer perceptrons (MLPs). The gating network is based on the expectation maximization (EM) algorithm. The final output of the ME system, o(x), is the sum of multiplications of the outputs from the gating and expert networks.

The developed model was optimized by studying two parameters simultaneously, while keeping other parameters constant. A total of six cases have been envisaged that influence the performance of the model. The weights of nodes were iteratively optimized with the conjugate gradient descent method. Each expert network used the sigmoid activation function, while the gate network used the softmax activation function. Thus, based on supervised learning, each expert network produced its own output from which the gate network learns how to choose the more precise final output.

The algorithm was trained and tested using Matlab. The simulation of the network was performed on a windows operating system with dual core processor using mixlab [14] and netlab (http://www.ncrg.aston.ac.uk/ netlab/). From these models, the one with minimum mean square error and maximum accuracy was selected as the prediction model.

Mixture of experts system

Mixture of experts (ME) is a supervised learning algorithm, which subdivides a learning task into appropriate subtasks, each of which can be solved by a simple expert network [15]. The global output of our ME system was derived as a convex combination of the outputs from a set of *n* experts, in which the overall predictive performance of the system was generally superior to any of the individual experts. The ME error function was based on the interpretation of an ME system as a mixture model with conditional densities as mixture components (for the experts) and gating network outputs as mixing coefficients.

The ME architecture was composed of several expert networks and a gating network (Figure 1). The gating network produced a scalar output from the vector input *x*, and worked on the generalized linear function, with the output for the *i*-th input variable, given by:

$$g(x, v_i) = \frac{e^{\xi_i}}{\sum_{k=1}^{n} e^{\xi_k}}$$

where $\xi_i = v_i^T x$, and v_i is a weight vector.

Each expert network produced an output vector for an input vector from the following generalized linear equation:

$$o_i(x) = f(W_i x)$$

where W_i is a weight matrix.

The final output of the ME system was the sum of the multiplication of the output from the gating and expert networks:

$$o(x) = \sum_{k=1}^{n} g(x, v_k) o_k(x)$$

The use of a soft-max function in the gating network and the conditional densities (θ_j) guaranteed that the distribution was normalized, with:

$$\int p(t|x)\,dt = 1$$

This distribution formed the basis for the ME error function, which could be optimized using the gradient descent or the expectation maximization (EM) algorithm.

Results

The data collected from the different sources were not consistent in all variables. Therefore, training and test datasets were prepared by normalizing the instances of the data, as per Table 2. Furthermore, all values were normalized between 0 and 1.

Table 2. Normalization of training and test data used to fill missing data

No.	Risk factor	Diabetic	Non-diabetic
1	FPG (mg/dl)	185	105
2	PPG (mg/dl)	255	155
3	BMI (kg/m^2)	34	26
4	Waist circumference (cm)	106	75
5	Diastolic BP (mmHg)	88	83
6	Hypertension	Yes	No
7	Lifestyle	Inactive	Active
8	High cholesterol	Yes	No
9	Physical exercise	No	Yes
10	Smoking	Yes	No
11	Food type (veg/non-veg)	Non-veg	Veg
12	Major cereal (rice/wheat)	Rice	Wheat
13	Family history of diabetes	Yes	No
14	Drinking habit	Yes	No

Legend: Data in the table are mean values of variables in the two groups, diabetic and non-diabetic individuals, respectively. Appreviations: BMI - body mass index, BP - blood pressure, FPG - fasting plasma glucose, PPG - postprandial glucose, SD - standard deviation, veg - vegetarian.

Development of models and their optimization

Various configurations of different models were tested to optimize the accuracy of the algorithm by using a trial and error approach. The number of hidden nodes is associated with the mapping ability of the network. Generally, the larger the number of hidden nodes, the more powerful is the network. However, if the number of hidden nodes is too large then overtraining may occur. Consequently, the generalization of the network may get worse, which may result in poor performance of the network on the test data.

Various models were tested to optimize the model for maximum accuracy and minimum mean square error. The selected model defined the network parameters as follows:

- Number of inputs = 17
- Number of outputs = 1
- Number of experts = 2
- Number of iterations in M-steps = 5

- Number of hidden units for gate MLPs = 20
- Number of iterations = 150 (changed to reduce the training error)
- Expert output activation function = logistic
- Type of ME = standard
- Number of hidden units for the expert MLPs = 8

The overall best accuracy of 99.36% was achieved with an iteration of 150 and 20 hidden nodes. The prediction error during the training process approached almost zero. This optimized trained network was tested on 311 samples from the test dataset. The network output is shown in Figure 2.

Sensitivity, specificity, and total classification accuracy were used as measures to evaluate and validate the performance of the mixture of expert classifier. These three terms can be defined as follows:

1. Sensitivity = (number of true positive outputs/number of actual positive cases)
2. Specificity = (number of true negative outputs/number of actual negative cases)
3. Total classification accuracy = (number of correct decisions/total number of cases)

The actual and predicted classifications are presented in a confusion matrix (Table 3). Sensitivity, specificity, and total classification accuracy were calculated as 99.5%, 99.07%, and 99.36%, respectively.

A graphical user interface was developed where users may easily enter the variables (Figure 3). The input variables are age, gender, fasting plasma glucose, postprandial glucose, waist circumference, height, weight, family history of diabetes, lifestyle, smoking, drinking, exercise, hypertension, diastolic blood pressure, high cholesterol, major food intake, regularity in food, and type of diet (vegetarian or non-vegetarian). When the submit button is pressed, the diagnosis is displayed in another frame. The back button can be used to navigate to previous page for reuse.

Discussion

Automated diagnosis of diseases has always been of interest as an inter-disciplinary study amongst computer and medical science researchers. Brause (2001) showed that human diagnostic capabilities are significantly worse than that of neural diagnostic systems [16]. Although several algorithms for predicting risk and diagnosis of type 2 diabetes have been developed, there is no widely accepted diabetes prediction tool in routine clinical practice [13]. The algorithm designed in this study differed from others in several ways. First, it used more input variables. Second, the algorithm used the mixture of experts approach that has shown significant improvements in the prediction problem. This study demonstrated that the algorithm reached and accuracy of 99.36% in the Indian dataset compared with 97% in an earlier study (using a Pima Indian diabetes dataset) [12].

Diabetes risk factors in India differ from those in western countries [9]. Therefore, strategies for diabetes detection and prevention in western populations are less likely to be effective in India.

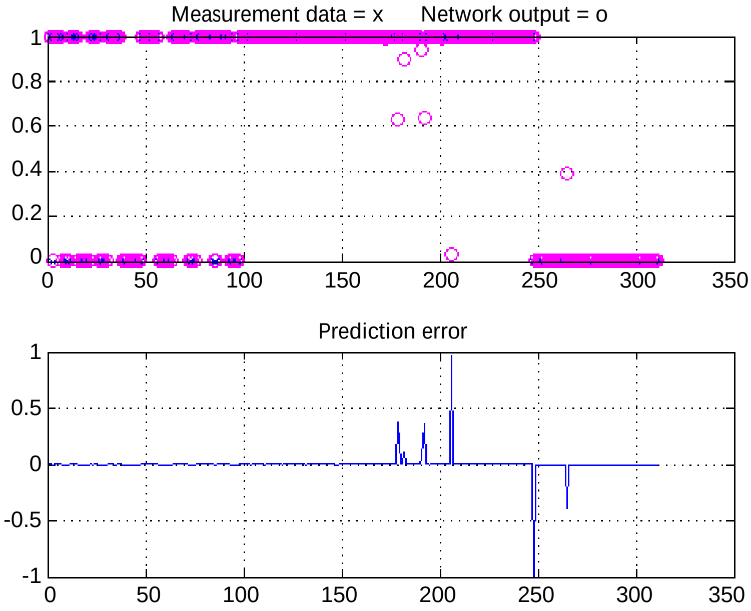


Figure 2. Network output of the testing and prediction error in testing. **A**: The outcome values are depicted as circles. Circles located at 0 and 1 clearly separate the diabetic from the non-diabetic population in the dataset. Circles located in the range]0,1[are pointing to prediabetic cases, depending on the location. **B**: Diagram showing the prediction error result. Two wrong predictions (one false positive and one false negative) are shown.

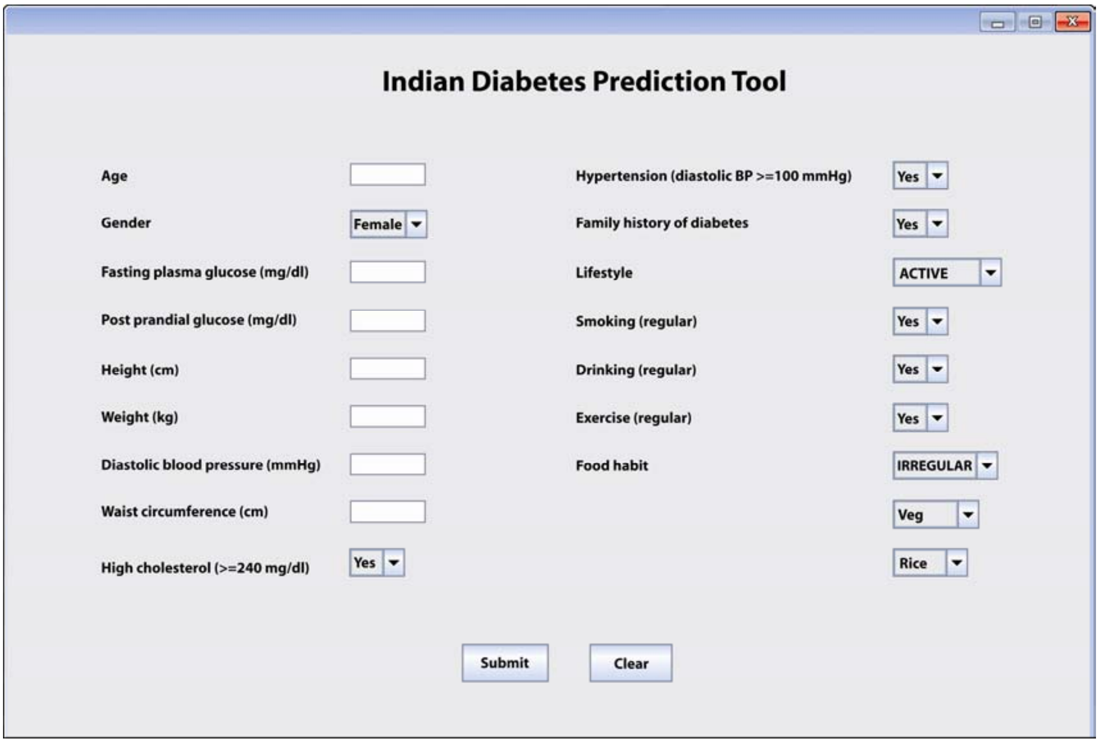
Table 3. Confusion matrix

		Predicted	
		Non-diabetic	Diabetic
Actual	Non-diabetic	107	1
	Diabetic	1	202

The current study was designed from the perspective of the Indian population. Until now, only two diabetes risk scores with relevance to the Indian ethnicity have been developed. Among them was the Indian Diabetes Risk Score (IDRS) derived from the Controlled Substance Utilization Review and Evaluation System (CURES) dataset which had a sensitivity of 72.5% and a specificity of 60.1% [17]. In another study, IDRS was validated on the Karnataka population and reached a sensitivity of 62.2% and a specificity of 73.7% [18]. A second risk score was developed from an urban Asian Indian population with 72.4% sensitivity and 59% specificity [19]. Such models failed to perform satisfactorily when applied to subjects from Indian regions that were not used to derive the model [20].

The ME classifier, configured with 2 expert networks, was implemented to compare the different models. In both hidden and output layers, the activation function was the sigmoid function, which had a range between zero and one. This function introduced two important properties. First, it was nonlinear. This allowed the network to perform complex mappings of input to output vector spaces. Second, it was continuous and differentiable. This allowed the gradient of the error to be used in updating the weights. In a similar study on a Pima Indian diabetes (PID) dataset, Ubeyli (2009) used 1100 iterations and 20 hidden layers (both in expert and gate network), and reached an accuracy of 97.93%, with a specificity of 98.01% and a sensitivity of 97.73% [21]. In contrast, our previous study with the same PID data showed that an accuracy of 70.7% was achieved with 900 EM iterations and 20 hidden nodes. With the PID data, the following accuracy results were achieved with the different combinations of EM iterations and hidden nodes:

1. 1100 EM iteration and 30 hidden nodes reached an accuracy of 85.7%
2. 1000 EM iterations and 20 hidden nodes



Indian Diabetes Prediction Tool

Age
Gender: Female
Fasting plasma glucose (mg/dl)
Post prandial glucose (mg/dl)
Height (cm)
Weight (kg)
Diastolic blood pressure (mmHg)
Waist circumference (cm)
High cholesterol (>=240 mg/dl): Yes

Hypertension (diastolic BP >=100 mmHg): Yes
Family history of diabetes: Yes
Lifestyle: ACTIVE
Smoking (regular): Yes
Drinking (regular): Yes
Exercise (regular): Yes
Food habit: IRREGULAR
Veg
Rice

Figure 3. Graphical user interface for prediction and diagnosis of diabetes type 2 in India.

reached an accuracy of 89.28%
3. 1100 EM iterations and 20 hidden nodes reached the overall best accuracy of 96.9% [12].

The tool developed in this study is superior to previously developed risk prediction tools on several accounts. Given its high specificity of 99.07% and sensitivity of 99.5%, it is more precise in differentiating diabetic from non-diabetic individuals. Moreover, it is very simple to use. As it is designed using java, it is portable across a range of operating systems. Although training and validation of the algorithm were performed on data from an Indian population, it will be interesting to broaden its application by testing it in other populations from different geographical locations.

Further trained models of diabetes risk factors may be incorporated into easy-to-use software solutions. For this purpose, graphical user interface-based tools have been developed enabling medical practitioners to simply enter the health profiles of their patients and to receive an instant diabetes prediction with an acceptable degree of confidence. Successful implementation of such prediction tools will be a step towards improved diagnosis and healthcare.

Conclusions

The present study describes a precise machine learning approach for discrimination between prediabetic, diabetic and non-diabetic patients. In India, undiagnosed diabetes is a major factor leading to an increased number of diabetic patients and fatal organ complications. The tool developed in this study provides a reliable alternative approach for detecting prediabetes and diabetes because it is built on modifiable lifestyle factors and is easy to use. It can be highly useful in large scale screening and hospitals. Moreover, it is portable and very simple to use, the user merely needs to enter the values in the computer interface and instant diagnosis is available. Such methods of early diabetes detection may be beneficial in reducing the number of diabetes patients in the future.

Acknowledgments: We are very grateful to the Department of Biotechnology, Birla Institute of Technology, Mesra, India, where this work was performed. We also acknowledge the Department of Agriculture, Government of Jharkhand (grant5/BKV/Misc/12/2001) and BITSNetSubDIC (grant BT/BI/04/065/04) from the Department of Biotechnology, Government of India for providing infrastructural facilities.

Disclosure: The authors report no conflict of interests.

■ References

1. **Bjork S, Kapur A, King H, Nair J, Ramachandran A.** Global policy: aspects of diabetes in India. *Health Policy* 2003. 66:61-72.
2. **Wild S, Roglic G, Green A, Sicree R, King H.** Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care* 2004. 27:1047-1053.
3. **Hossain P, Kawar B, Nahas M.** Obesity and diabetes in the developing world - a growing challenge. *N Engl JMed* 2007. 356:213-215.
4. **Gao W.** Early detection of type 2 diabetes mellitus in Chinese and Indian adult population., Thesis Dissertation, Department of Public Health, University of Helsinki, Finland, 2010. pp. 15-34.
5. **Tuomilehto J, Lindstrom J, Eriksson JG, Valle TT, Hamalainen H, Ilanne-Parikka P, Keinänen-Kiukaanniemi S, Laakso M, Louheranta A, Rastas M, Salminen V et al.** Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *N Engl J Med* 2001. 344(18):1343-1350.
6. **Chiasson JL, Josse RG, Gomis R, Hanefeld M, Karasik A, Laakso M.** Acarbose for prevention of type 2 diabetes mellitus: the STOP-NIDDM randomised trial. *Lancet* 2002. 359(9323):2072-2077.
7. **Ramachandran A, Snehlatha C, Mary S, Mukesh B, Bhaskar A, Vijay V.** The Indian diabetes prevention program shows that lifestyle modification and metformin prevent type 2 diabetes in Asian Indian subjects with impaired glucose tolerance (IDPP-1). *Diabetologia* 2006. 49:289-297.
8. **Mohan V, Sandeep S, Deepa R, Shah B, Varghese C.** Epidemiology of type 2 diabetes: Indian scenario. *Indian J Med Res* 2007. 125:217-230.
9. **Diamond J.** Medicine: Diabetes in India. *Nature* 2011. 469:478-479.
10. **Ramachandran A, Snehalatha C, Kapur A, Vijay V, Mohan V, Das AK, Rao PV, Yajnik CS, Prasanna Kumar KM, Nair JD.** Diabetes Epidemiology Study Group in India (DESI). High prevalence of diabetes and impaired glucose tolerance in India: National Urban Diabetes Survey. *Diabetologia* 2001. 44(9):1094-101.
11. **Menon VU, Kumar KV, Gilchrist A, Sugathan TN, Sundaram KR, Nair V, Kumar H.** Prevalence of known and undetected diabetes and associated risk factors in central Kerala - ADEPS. *Diabetes Res Clin Pract* 2006. 74(3):289-294.
12. **Shankaracharya, Odedra D, Mallick M, Shukla P, Samanta S, Vidyarthi AS.** Java-based diabetes type 2 prediction tool for better diagnosis. *Diabetes Tech Ther* 2012. 14:251-256.
13. **Shankaracharya, Odedra D, Vidyarthi AS.** Computational Intelligence in Early Diabetes Diagnosis: A Review. *Rev Diabet Stud* 2010. 7:252-262.
14. **Moerland P.** Some methods for training mixtures of experts. IDIAP communication. *IDIAP- Com* 1997, pp. 97-105.
15. **Jordan MI, Jacobs RA.** Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 1994. 6:181-214.
16. **Brause RW.** Medical analysis and diagnosis by neural networks. *Lecture Notes in Computer Science* 2001. 99:1-13.

17. **Mohan V, Deepa R, Deepa M, Somannavar S, Datta M.** A simplified Indian diabetes risk score for screening for undiagnosed diabetic subjects. *J Assoc Physicians India* 2005. 53:759-763.

18. **Adhikari P, Pathak R, Kotian S.** Validation of the MDRF Indian diabetes risk score (IDRS) in another south Indian population through the Boloor Diabetes Study (BDS). *J Assoc Physicians India* 2010. 58:434-436.

19. **Ramachandran A, Snehalatha C, Vijay V, Wareham NJ, Colagiuri S.** Derivation and validation of diabetes risk score for urban Asian Indians. *Diabetes Res Clin Pract* 2005. 70:63-70.

20. **Babyak MA.** What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med* 2004. 66:411-421.

21. **Ubeyli ED.** Modified mixture of experts for diabetes diagnosis. *J Med Syst* 2009. 33**:**299-305.