Diagnosing Heart Diseases for Type 2 Diabetic Patients by Cascading the Data Mining Techniques

P. Radha
Ph.D Research Scholar,
Computer Science Department,
Karpagam University, Coimbatore
&
Assistant Professor, Vellalar College for Women,
Erode, Tamil Nadu.
radhasakthivel09@gmail.com

Dr. B. Srinivasan
Computer Science Department,
Gobi Arts and Science College
Gobichettipalayam, Tamil Nadu.
srinivasan_gasc@yahoo.com

Abstract— Motivated by the world-wide increasing mortality of heart disease patients each year, researchers have been using data mining techniques to help health care professionals in the diagnosis of heart disease. Heart disease is the leading cause of death in the world over the past 10 years. Researchers have been using several data mining techniques to help health care professionals in the diagnosis of heart disease. To review the primary prevention studies that focused on the development, validation and impact assessment of a heart disease risk model, scores or rules that can be applied to patients with type 2 diabetes. Efficient predictive modeling is required for medical researchers and practitioners. Attribute values measurement using entropy and information gain parameters. This study proposes Hybrid type 2 diabetes Prediction Model which uses Improved Fuzzy C Means (IFCM) clustering algorithm aimed at validating chosen class label of given data in which incorrectly classified instances are removed and. pattern extracted from original data. Support Vector Machine (SVM) algorithm is used to build the final classifier model by using the k-fold cross-validation method. The aim of this paper is to highlight all the techniques and risk factors that are considered for diagnosis of heart disease. This paper will provide a roadmap for researchers seeking to understand existing automated diagnosis of heart disease.

Keywords- *Classification, Hybrid Prediction Model, Fuzzy c means clustering(FCM), Pima Indians diabetes cardiovascular disease (CVD), Principal Component Analysis (PCA),Support Vector Machine(SVM).*

*****

1. INTRODUCTION

Diabetes is the most common disease nowadays in all populations and in all age groups. It is a disease in which the body does not produce or properly use insulin. The cells in our body require glucose for growth for which insulin is quite essential. When someone has diabetes, little or no insulin is secreted. In this situation, plenty of glucose is available in the blood stream but the body is unable to use it [1]. Basically there are two types of diabetes, viz. Type-1 and Type-2. Type-1 diabetes occurs when the body's immune system is attacked and the beta cells (these cells produce insulin) of pancreas are destroyed. This results in insulin deficiency. The only treatment to Type-1 diabetes is insulin. On the other hand, Type-2 diabetes is caused by relative insulin deficiency. Pancreas in Type-2 diabetes still produces insulin but it may not be effective or may not produce sufficient amount of insulin to control blood glucose [2]. Type-2 diabetes is the most common type of diabetes [3], which usually develops at age 40 and older.

Type-2 diabetes is serious global health problem, which, for most countries, has evolved in association with rapid cultural and social changes, ageing populations, increasing urbanization, dietary changes, reduced physical activity and other unhealthy lifestyle and behavioral patterns . People with type 2 diabetes have a twofold increased risk of heart disease [4-5]. Guidelines for the management of type 2 diabetes advocate calculating heart disease risk to guide the initiation of appropriate treatment [5-6]. Over the past decades many prediction models (or risk scores) have been developed to predict heart disease of which only a small number have been specifically developed for people with type 2 diabetes [7].

Among these stages analysis of type 2 diabetes with heart disease risk factors, first the collection of data and removal of the irrelevant data plays major important role to improve the prediction results of the type 2 diabetes. The major important steps of the proposed works as follows: Preprocessing of the data using dimensionality reduction principal component analysis (PCA) method it is also used for dimensionality reduction to reduce the complexity of the dataset. Once the dimensionality is reduced in the data then risk factors of heart disease are analyzed using similarity measures like entropy, information gain. The purpose of this study is to build a Hybrid Prediction Model that should perform unsupervised classification methods based on Improved Fuzzy C Means clustering (IFCM) accurately classify newly diagnosed patients into either a group that is likely to develop type 2 diabetes. Then perform supervised classification using support vector machine (SVM) classification methods. The aim of this

study was to identify all heart disease prediction models (or scores or rules) that can be applied to patients with type 2 diabetes, and subsequently to assess their status.

2. BACKGROUND STUDY

Managing the numerous risk factors responsible for heart disease in T2DM represents an ongoing challenge for primary care clinicians, strongly influencing their decisions about treatment approaches for this complex disease[10]. Established risk factors include poor control of glycated hemoglobin (HbA1c) levels, systolic blood pressure, and lipid levels, along with age, sex, ethnicity, smoking status, and disease duration [10-11] . This model predicts fatal heart disease over 10 years based on five predictors namely, sex, age, smoking, systolic blood pressure and either total cholesterol or ratio total/high-density lipoprotein- cholesterol. The predictive ability of SCORE in patients with type 2 diabetes, however, has been assessed by three studies and was similar to other heart disease prediction models included in our review[13-14] .

Insulin resistance is an early and major component of T2DM and an independent risk factor for heart disease. Endothelial dysfunction and increased carotid intima-media thickness [16], may be early, reversible features of heart disease and can be assessed non-invasively. Thus, early identification of insulin resistance and impaired endothelial function may identify those at particular risk of heart disease and enable targeting of aggressive risk factor control to those who will most benefit. Before studying the type 2 diabetes patient for heart disease risk factor analysis in the works, first study major important factors to analysis the results of heart disease for type 2 diabetes patients.

Important risk factors in Heart Disease with Type 2 Diabetes (T2D)

Heart disease is a serious but preventable complication of type 2 diabetes (T2D) that results in substantial disease burden, increased health services use, and higher risk of premature mortality

Modifiable risk factors are Smoking status, Blood pressure, Serum lipids, Waist circumference and body mass index, Nutrition, Physical activity level, Alcohol intake.

Non-modifiable risk factors Age and sex, Family history of premature heart disease, Social history including cultural identity, ethnicity, socioeconomic status and mental health Related conditions, Diabetes, Kidney function, Familial hypercholesterolaemia, Evidence of atrial fibrillation.

Managing the numerous risk factors responsible for heart disease in T2D represents an ongoing challenge for primary care clinicians, strongly influencing their decisions about treatment approaches for this complex disease.[17] Established risk factors include poor control of glycated hemoglobin (HbA1c) levels, systolic blood pressure, and lipid levels, along with age, sex, ethnicity, smoking status, and disease duration [18-19]. Many people with T2D are also hypertensive which contributes to the premature development of vascular disease. Diabetes is associated with a typical dyslipidaemia comprising mildly elevated levels of small dense low-density lipoprotein (LDL), reduced levels and altered composition of high-density lipoprotein (HDL) and increased triglyceride-rich lipoprotein particles. Glycated, small dense LDL is associated with increased oxidative stress within the vasculature, while reduced concentrations of altered HDL are less able to participate in athero protective functions such as reverse cholesterol transport.

The major aim of this study is propose the hybrid prediction model for the analysis of risk of cardiovascular disease in type-2 diabetic patients by applying preprocessing method to reduce the irrelevant data and missed data information which are not use for prediction of T2D for risk analysis of heart disease is removed in this stage using principal component analysis and dimensionality reduction also performed in this stage ,then perform the attribute analysis of the risk factors based on the information gain and entropy values , develop hybrid prediction models for the prediction of a Type-2 Diabetic Patient. Hybrid prediction model performs the prediction based on the unsupervised classification methods initially the unsupervised classification methods refer a clustering methods to measure the similarity among the attributes for labeling the attributes under classes ,then perform support vector machine learning based classification task for prediction .

3. PROPOSED METHODOLOGY

The major objective of this proposed work is to examine the common clinical and behavioral factors that contribute to heart disease risk (i.e. attributable risk) among those with type 2 diabetes and perform hybrid prediction model .The major steps involved in the proposed system are: preprocessing of the type 2 diabetes patient (T2D) data with heart disease risk using principal component analysis (PCA) and dimensionality reduction is also performed using PCA. Then heart disease risk factors are estimated based on the metrics like information gain and entropy measurements to enhance the prediction accuracy results. The estimated heart disease risk factors are used for unsupervised classification using Improved

Fuzzy C Means (IFCM) clustering methods, which data is used prediction of T2D for heart disease risk factors. Then perform supervised classification task for prediction of type 2 diabetes patients with heart disease risk are predicted using support vector machine (SVM) . entire representation of the proposed system is illustrated in Figure 1 .

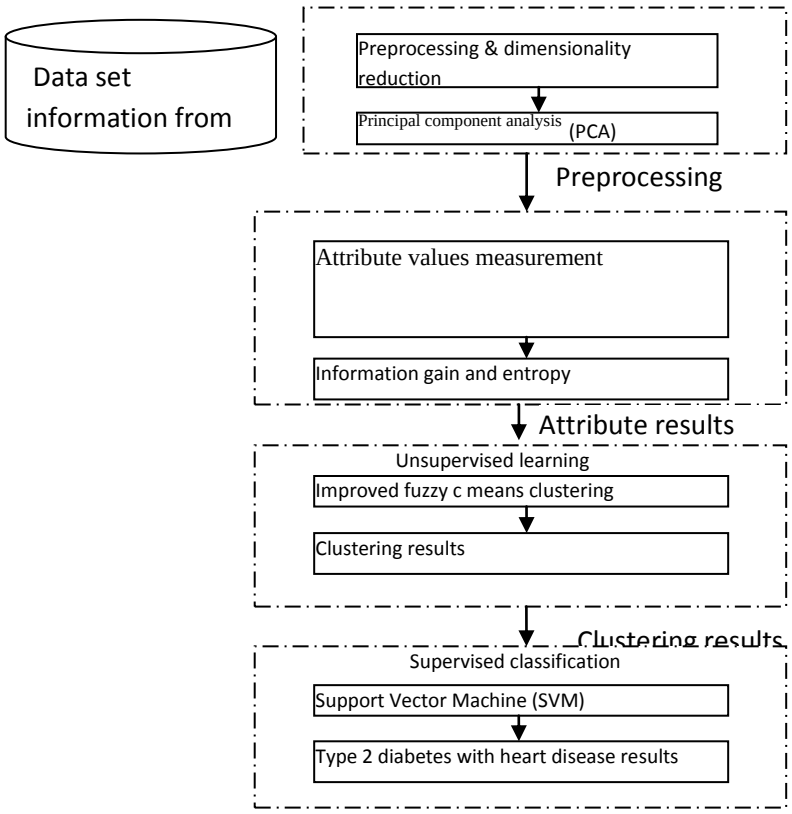

FIG 1: ARCHITECTURE DIAGRAM OF THE PROPOSED WORK

3.1 DATASET INFORMATION

The dataset collected from real patient records which includes the following attributes for diabetes patients records Number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test , Diastolic blood pressure (mm Hg) ,Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml) ,Body mass index (weight in kg/(heightinm)^2) ,Diabetes pedigree function ,Age (years) ,Class variable (0 or 1).These data are collected with the following CVD risk factors which includes BMI (Body Mass Index) , Weight (kg) ,Waist circumference (cm) , Systolic blood pressure (SBP) (mmHg) , Diastolic blood pressure (DBP) (mmHg) ,Glucose (mg/dl) ,Total cholesterol (mg/dl) , High-Density Lipoprotein cholesterol (HDL-c) (mg/dl) , Low-Density Lipoprotein cholesterol (LDL-c) (mg/dl) ,Triglycerides (mg/dl) ,HbA1c (glycosylated hemoglobin) (%) Fibrinogen (mg/dl), ultrasensitive C reactive protein (us-CRP) (mg/L).

3.2 PREPROCESSING AND DIMENSIONALITY REDUCTION USING PRINCIPAL COMPONENT ANALYSIS (PCA)

The quality of the data is the most important aspect as it influences the quality of the results from the analysis. The data should be carefully collected, integrated, characterized, and prepared for analysis. In this study, we applied the principal component analysis (PCA) for preprocessing of type 2 diabetes (T2D) with heart disease risk factors as mentioned above. Initially the data are collected for each patients which consists of the attributes those mentioned above ,once the data are collected with heart disease risk factors then whose patient records which doesn't contain information about those attributes are removed ,and which data eigenvector associated with largest Eigen value is the most important vector that reflects the greatest variance for prediction process .From this point of the view the data are preprocessed and removed in the preprocessing stage . A preliminary

analysis of the data indicates the usage of zero for missing data. The various variables used in this study are pregnant, plasma–glucose, diastolic BP, Body mass index, diabetes pedigree function, age, serum–insulin, triceps skin fold and class. Since, it does not make sense to have the value of a variable such as plasma–glucose concentration 0 in living people; all the observations with zero entries are removed. Also, in this analysis, the count of missing values for the variables serum–insulin and triceps skin fold are very high. Principal component analysis for preprocessing of T2D with heart disease risk factors, it is successfully applied to various applications of pattern recognition such as face classification [21]. As mentioned above, N = (x1, x2, …xn)be the number of type 2 diabetes patient's hospital data with the heart disease risk factors and t dimension of dataset D, respectively. PCA finds a subspace of the attribute value whose basis vectors correspond to the maximum-variance direction of the original T2D data space. As mentioned before, PCA is a linear transform. Let W represents the linear transformation that maps the original dimensional T2D data space into an f-dimensional reduced irrelevant and missing attribute data where normally f << t Equation (1) shows the new reduced dimensional and reduced irrelevant data variable vectors xi € Rt

$$z_j = W^T x_j \, , j = 1, \dots . N \qquad (1)$$

λjej=Qej,,j=1,…N where $Q=XX^T$, X ={ x_1, ….,x_N} (2)

Here Q is the covariance matrix and λj the eigen value associated with the eigenvector ej . The eigenvectors are sorted from high to low according to their corresponding eigen values. The eigenvector associated with largest eigen value is the most important variable and data vector that reflects the greatest variance [22]. PCA employs the entire T2D patient hospital record variables with heart disease risk factors and it acquires a set of projection attribute vectors to extract most important global variable and data vector from given training samples. The performance of PCA is reduced when there are more irrelevant data ones than the relevant T2D with heart disease risk factor ones.

3. 3 ATTRIBUTE VALUES MEASUREMENT

To measure the importance of the risk factor for heart disease, first analyze the results of the attributes based on their attribute value .In this work measure the values of the attributes for prediction of the CVD risk factor in T2D patients based on their metrics like entropy and information gain ,if the attribute values results of information gain and entropy is high it shows the prediction results of T2D with CVD risks factors are high . For each and every attribute values select highest value which is greater than the thresholds value. BMI, Weight (kg) ,Waist circumference (cm), SBP (mmHg), DBP (mmHg), Glucose (mg/dl), Total cholesterol (mg/dl), HDL-c (mg/dl), LDL-c (mg/dl),Triglycerides (mg/dl), HbA1c (%), Fibrinogen (mg/dl) and us-CRP (mg/L). Shannon defined the entropy H of a discrete random variable with possible values be the attributes with risk factors CVD and probability mass function as:

$$H(X) = E[I(X)] = E[-\ln(P(X))] \qquad (3)$$

Here E is the expected value operator (maximum threshold value results), and is the information content (value of content) from patient record of is itself a random variable. If the value of the attribute results is equal to entropy value then it is selected for risk factor estimation. When taken from a finite sample, the entropy can explicitly be written as

$$H(X) = \sum_i P(x_i) I(x_i) = -\sum_i P(x_i) \log_b P(x_i) \qquad (4)$$

where is the base of the logarithm used. Common values of are 2.

Information gain is a measure of this change in entropy .Suppose S is a set of instances, A is an attribute, S_v is the subset of S with A=v, and Values(A) is the set of all possible values of A, then

$$Gain\,(S, A) = Entrophy\,(S) - \sum_{v \in Values\,(A)} \frac{|S_v|}{|S|} . Entropy\,(S_v) \qquad (5)$$

A risk equation was created for estimation of the risk of heart disease, using q and the HRs for the nine predictors, After the calculation of entropy and information gain values then calculate the risk factor of heart disease for prediction of the Type 2 diabetes patients ,

$$CVD_{risk} = (1 - \exp[-q_r \times \alpha_1^{BMI} \times \alpha_2^{W} \times \alpha_3^{WC} \times \alpha_4^{SBP} \times \alpha_5^{DBP} \times \alpha_6^{G} \times \alpha_7^{TC} \times \alpha_8^{HDL-C} \times \alpha_9^{LDL-C} \times \alpha_{10}^{TRC} \times \alpha_{11}^{HbA1C} \times \alpha_{12}^{F} \times \alpha_{12}^{CRP}]) \quad (6)$$

For each and every attribute values select highest value which is greater than the thresholds value. BMI, Weight (kg) ,Waist circumference (cm), SBP (mmHg), DBP (mmHg), Glucose (mg/dl), Total cholesterol (mg/dl), HDL-c (mg/dl), LDL-c (mg/dl),Triglycerides(TRC) (mg/dl), HbA1c (%), Fibrinogen (mg/dl) and us-CRP (mg/L).

3.4 IMPROVED FUZZY C MEANS CLUSTERING (IFCM)

As the first step, before the application of the Classification algorithms, aiming at validating the chosen classes using the unsupervised methods .In this work uses an Improved Fuzzy c means (IFCM) clustering to validate the preprocessed dataset, then assign class labels to similar cluster, the clustering algorithm. The process of stepwise regression involves of these steps as follows:

(i) Choose the number of clusters c, fuzziness index m, iteration error ", maximum iterations T, and initialize the

membership degree matrix $U^{(0)}$.

1) Get the initial centroids.
2) Calculate the dot density of every data point.
3) And when the iteration index is $t(t = 1,2, \dots, T)$, updating the membership degree matrix $U^{(t)}$ and cluster centroids $V^{(t)}$).
4) Calculate the value of the objective function $J^{(t)}$.
5) If $|U^{(t)} - U^{(t+1)}| < \varepsilon$ or $t = T$, then stop the iteration and get the membership degree matrix U and the cluster centroids V , otherwise set and return to step (4).

3.5 SUPPORT VECTOR MACHINE CLASSIFICATION FOR PREDICTION OF TYPE 2 DIABETES WITH HEART DISEASE RISKS

From this results finally the cluster are formed either class label yes or class label no for classification of type 2 diabetes patients reduced dimensionality data in the PCA. Finally perform classification task for unsupervised class labels results from Improved Fuzzy c means (IFCM) clustering. The clustered results are taken as input to support vector machine (SVM) classification task for prediction of type 2 diabetes patient with heart disease risks. The Support Vector Machine (SVM) is a classification technique based on statistical learning theory [23], [24] that was applied with great success in many challenging non-linear classification problems and was successfully applied to large data sets. The SVM algorithm finds a hyper plane that optimally splits the type 2 diabetes clustered data results from IFCM training set. The optimal hyper plane can be distinguished by the maximum margin of separation between all clustered input data as training points and the hyper plane. Looking at a two-dimensional prediction problem we actually want to find a line that "best" separates type 2 diabetes with heart disease risk factor clustered data points in the positive class from points in the negative class. The hyper plane is characterized by a decision function like,

$$f(x) = sgn((w, \phi(x))) + b \quad (7)$$

where w is the weight vector for clustered data , orthogonal to the hyper plane, b is a scalar that represents the margin of the hyper plane, x is the current clustered sample tested, $\Phi(x)$ is a kernel function that transforms the input data into a higher dimensional feature space and '.',' representing the dot product. S_{gn} is the signum function.

EXPERIMENTATION RESULTS

The data were not specifically collected for a research study. As part of routine patient management, UCHT collected diabetic patients' information from 2000 to 2004 in a clinical information system (Diamond, Hicom Technology). The data contained physiological and laboratory information for 3857 patients, described by 410 features. The patients included not only type 2 diabetic patients, but also type 1 and other types of diabetes such as gestational diabetes .Some measure of evaluating performance have to be introduced. One common measure in the literature [26] is accuracy defined as correct classified instances divided by the total number of instances. The true positives (TP) and true negatives (TN) are correct classifications. A false positive (FP) occurs when the outcome is incorrectly predicted as yes (or positive) when it is actually no (negative). A false negative

(FN) occurs when the outcome is incorrectly predicted as no when it is actually yes. In this study we use following equation to measure the accuracy Eq. (9), specificity Eq. (10), sensitivity Eq. (11)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (10)$$

$$Specificity = \frac{TN}{TN + FP} \quad (11)$$

These parameters can be used to measure accuracy, sensitivity and specificity, respectively. Sensitivity is also referred to as the true positive rate that is, the proportion of positive tuples that are correctly identified, while specificity is the true negative rate that is, the proportion of negative tuples that are correctly identified.

Table 1: Prediction methods results

Parameters	IFCM-SVM
Accuracy	93.8
Sensitivity	90.49
Specificity	54.7

5. CONCLUSION

Type 2 diabetes confers a high degree of heart disease risk brought about by multiplicative risk factors. This work presents an efficient prediction model to analysis the risk of heart disease factors in T2D patient records, to analyze the patient records. Initially the records are preprocessed and reduced dimension of the features using Principal component analysis. Risk factors are calculated using information gain and entropy measure, then unsupervised learning using Improved Fuzzy C Means (IFCM) clustering algorithm, finally support vector machine is discussed to perform prediction of T2D with heart disease risk factors. New studies investigating prediction of heart disease among patients with type 2 diabetes should, in our view, focus on further validating the performance of existing K means unsupervised learning with other techniques for assessing their impact on treatment and prevention of cardiovascular events instead of developing new prediction models.

REFERENCES

[1] Mohamed, E. L., Linderm, R., Perriello, G., Di Daniele, N., Poppl, S. J., & De Lorenzo, A. (2002). Predicting type 2 diabetes using an electronic nose-base artificial neural network analysis. Diabetes Nutrition and Metabolism, 15(4), 215–221.

[2] Guthrie, R. A., & Guthrie, D. W. (2002). Nursing management of diabetes mellitus (5th ed.). New York: Springer Publishing.

[3] Acharya, U. R., Tan, P. H., Subramaniam, T., Tamura, T., Chua, K. C., Goh, S. C., et al. (2008). Automated identification of diabetic type 2 subjects with and without neuropathy using wavelet transform on pedobarograph. Journal of Medical Systems, 32(1), 21–29.

[4] Sarwar N, Gao P, Seshasai SR, et al; Emerging Risk Factors Collaboration. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Lancet 2010, 2215-22.

[5] Woodward M, Zhang X, Barzi F, et al; Asia Pacific Cohort Studies Collaboration. The effects of diabetes on the risks of major cardiovascular diseases and death in the Asia-Pacific region. Diabetes Care 2003, 360-6.

[6] Ryden L, Standl E, Bartnik M, et al; Task Force on Diabetes and Cardiovascular Diseases of the European Society of Cardiology (ESC); European Association for the Study of Diabetes (EASD). Guidelines on diabetes, pre-diabetes, and cardiovascular diseases: executive summary. The Task Force on Diabetes and Cardiovascular Diseases of the European Society of Cardiology (ESC) and of the European Association for the Study of Diabetes (EASD). Eur Heart J 2007;28: 88-136.

[7] Chamnan P, Simmons RK, Sharp SJ, et al. Cardiovascular risk assessment scores for people with diabetes: a systematic review. Diabetologia 2009;52:2001-14

[8] Pellegrini E, Maurantonio M, Giannico IM, et al. Risk for cardiovascular events in an Italian population of patients with type 2 diabetes. Nutr Metab Cardiovasc Dis 2011;21:885-92.

[9] Vijan S, Hayward RA. Pharmacologic lipid-lowering therapy in type 2 diabetes mellitus: background paper for the American College of Physicians. Ann Intern Med 2004;140(8):650-8.

[10] Buse JB, Ginsberg HN, Bakris GL, Clark NG, Costa F, Eckel R, et al. Primary prevention of cardiovascular diseases in people with diabetes mellitus: a scientific statement from the American Heart Association and the American Diabetes Association. Diabetes Care 2007;30(1):162-72.

[11] Stevens RJ, Kothari V, Adler AI, Stratton IM. The UKPDS risk engine: a model for the risk of coronary heart disease in type II diabetes (UKPDS 56). Clin Sci (Lond) 2001;101(6):671-9.

[12] Conroy RM, Pyo¨ra¨la¨ K, Fitzgerald AP, et al; SCORE project group. Estimation of tenyear risk of fatal cardiovascular disease in Europe: The SCORE project. Eur Heart J 2003;24:987e1003.

[13] Vander Heijden AA, Ortegon MM, Niessen LW, et al. Prediction of coronary heart disease risk in a general, pre-diabetic, and diabetic population during 10 years of follow-up: accuracy of the Framingham, SCORE, and UKPDS risk functions: The Hoorn Study. Diabetes Care 2009;32:2094e8.

[14] Chen L, Tonkin AM, Moon L, et al. Recalibration and validation of the SCORE risk chart in the Australian population: the AusSCORE chart. Eur J Cardiovasc Prev Rehabil 2009;16:562e70.

[15] Lim YK, Jenner A, Ali AB, Wang Y, Hsu SI, Chong SM, "Haptoglobin reduces renal oxidative DNA and tissue damage during phenylhydrazine-induced hemolysis," Kidney Int, vol. 58(3), pp. 1033-1044, 2000.

[16] Nathan DM, Lachin J, Cleary P et al; Diabetes Control and Complications Trial; Epidemiology of Diabetes Interventions and Complications Research Group. Intensive diabetes therapy and carotid intima-media thickness in type 1 diabetes mellitus. N Engl J Med 2003;348:2294–303.

[17] Vijan S, Hayward RA. Pharmacologic lipid-lowering therapy in type 2 diabetes mellitus: background paper for the American College of Physicians. Ann Intern Med 2004;140(8):650-8.

[18] Buse JB, Ginsberg HN, Bakris GL, Clark NG, Costa F, Eckel R, et al. Primary prevention of cardiovascular diseases in people with diabetes mellitus: a scientific statement from the American Heart Association and the American Diabetes Association. Diabetes Care 2007;30(1):162-72.

[19] Stevens RJ, Kothari V, Adler AI, Stratton IM. The UKPDS risk engine: a model for the risk of coronary heart disease in type II diabetes (UKPDS 56). Clin Sci (Lond) 2001;101(6):671-9

[20] Hoerger TJ, Segel JE, Gregg EW, Saaddine JB. Is glycemic control improving in U.S. adults? Diabetes Care 2008;31(1):81-6. Epub 2007 Oct 12.

[21] S. Chen and Y. Zhu, "Subpattern-based principle component analysis," Pattern Recognition, vol. 37, no. 5, pp. 1081–1083, 2004.

[22] A. M. Martinez and A. C. Kak, "PCA versus LDA," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 228–233, 2001.

[23] C. Nello, J. Swawe-Taylor, "An introduction to Support Vector Machines", Cambridge University Press, 2000.

[24] J. Platt, "Fast training of support vector machines using sequential minimal optimization". In B. Scholkopf, C. J. C. Burges, and A. J.

a. Smola, editors, Advances in Kernel Methods – Support Vector Learning, pages 185-208, Cambridge, MA, 1999, MIT Press.

[25] XIA Guo-en and SHAO Pei-ji."Factor Analysis Algorithm with Mercer Kernel", IEEE Second International Symposium on Intelligent Information Technology and Security Informatics, 2009.

[26] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research (JAIR), 16, 321‑357.