ISSN: 2321-7782 (Online)
Volume 2, Issue 10, October 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Hybrid Prediction Model for the Risk of Cardiovascular Disease in Type-2 Diabetic Patients

P. Radha[1]
Ph.D Scholar, Dept. of Computer Science,
Karpagam University
Asst. Prof, Dept. of Computer Science,
Vellalar College for Women, Thindal
Erode, Tamil Nadu, India

B. Srinvasan[2]
Associate Professor, Dept. of Computer Science,
Gobi Arts and Science College,
Gobichettipalayam,
Tamil Nadu, India

Abstract: ***For the general population with type 2 diabetes (T2D), the risk prediction of cardiovascular disease (CVD) based on risk prediction models cannot be carried out well. The medical researchers and practitioners necessitate an effective predictive modelling and hence cardiovascular risk model could be well applied to such patients. The development, validation and impact assessment of cardiovascular risk model in the view of primary prevention of CVD is to be focused. The effect of patient characteristics and measurements on the incident of CVD factors along with T2D is also to be investigated. This study proposes Hybrid type 2 diabetes Prediction Model based on Improved Fuzzy C Means (IFCM) clustering algorithm for validating the selected class label of given data and the Support Vector Machine (SVM) classification algorithm for applying to the result set. SVM algorithm finds its importance in building the final classifier model by using the k-fold cross-validation method. The proposed study consists of four major steps involving pre-processing and dimensionality reduction of type 2 diabetes for CVD factors using Principal Component Analysis (PCA). Attribute values are measured using entropy and information gain parameters. The performance of the Hybrid prediction model is evaluated by the measure of sensitivity and specificity as used in medical classification studies. The prediction accuracy of the proposed model for T2D patients along with CVD factors is observed to be higher rather than the existing classification accuracy of C4.5***

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

I. INTRODUCTION

In present days, the most common disease affecting the population of all age groups is diabetes. Diabetes is caused due to the excess of glucose in the blood stream as the insulin which controls the blood glucose level is not properly secreted or utilized [1]. There are basically two types of diabetes namely type-1 and type-2 diabetes. When the body's immune system does not function well, the pancreatic beta cells responsible for insulin production is been destroyed which ultimately results in failure of insulin secretion. However, type-2 diabetes (T2D) is caused by relative insulin deficiency which means that the amount of insulin secreted in patient's body may not be sufficient or effective may not be effective to control blood glucose [2]. T2D is the most common type of diabetes [3], which usually occurring at an age 40 and more. T2D is serious global health problem across many countries and 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.

Also, patients with T2D have a twofold increased risk of cardiovascular disease (CVD) [4]. Hence, for avoiding the risk of CVD incident on T2D patients, the early prediction of CVD is required. The calculation of the risk of CVD among such patients is an important guideline for the management of T2D [5-6]. Many prediction models have been developed over the past decades

to predict the risk of CVD. Despite most prediction models developed could be applied for the diabetic general population, only few such models have been specifically developed for people with type 2 diabetes [7].

The development of CVD prediction models for diabetic population has been systematically reviewed by Chamnan et al [9] and then validated. Later on, several new prediction models for the diabetic population with CVD risks have been validated on the people with diabetes. The application of such models in affecting the treatment of diabetic patients for the reduction of CVD outcome in clinical practice is still ambiguous. However, estimates of the CVD risk can be used as prognostic information and support for the choice of therapeutic strategies for individual diabetic patients. Therefore, for efficient and precise clinical decision making, accurate prediction models for validating the estimates of the risk of the targeted outcome is needed, Based on this, there are three stages of prediction modeling identified: (1) model development that includes determining the clinically relevant predictors, assigning the relative weights to these predictors and finally estimating the ideal predictive performance of the model after optimizing or over fitting with internal validation techniques; (2) Assessment of the model's predictive performance in new patients (external validation studies); (3) quantifying the application of a prediction model in daily clinical practice for improving the decision-making. Among all these stages of prediction of CVD risks in T2D patients, data collection and segregation of the useful data from the irrelevant one plays an important role to improve the prediction results.

The relation between several baseline predictor variables for improving the prediction results and the event of occurrence of fatal or nonfatal CVD in T2D patients is proposed to be analyzed in this study. The major important steps of the proposed work are (i) preprocessing of the collected patient's data for dimensionality and complexity reduction using principal component analysis (PCA) method, (ii) Analysis of the CVD risk factors using similarity measures like entropy and information gain. The main goal of this study is to build a Hybrid Prediction Model for performing unsupervised classification based on Improved Fuzzy C Means clustering (IFCM) for accurately classifying newly diagnosed patients into either a group with T2D or another group that is likely to develop T2D. This is followed by supervised classification using support vector machine (SVM) classification methods. The identification of all CVD prediction models for T2D patients that scores or rules and subsequently assessing their status is to be performed in this study.

II. Back Ground Study

The prevalence of heart disease in India has increased four-fold in the last four decades. Predicted to become the leading cause of death and disability by 2020, heart disease currently accounts for 29% of all deaths in the country. The worst part about the disease is that, in India, people are succumbing to heart disease and stroke in the most productive years of their lives, almost a decade earlier compared to the West. Both the government and the business community are waking up to this threat.

Managing the numerous risk factors responsible for CVD in The primary treatment approach for this complex CVD in T2D patients is dependent primarily on management of the risk factors responsible for CVD which subsequently helps the decision-making of clinical care personnel [10]. The major well-known risk factors include poor control of glycated haemoglobin (HbA1c) levels, systolic blood pressure, and lipid levels, along with age, sex, ethnicity, smoking status, and disease duration [10-11]. The demographic characteristics might be considered as fixed factors, whereas other factors are potentially changing and hence has to be addressed by continuous learning based on lifestyle choices and behavior. Numerous prediction models for other disabilities occurring along with T2D have so far been proposed and hence the background of this study excludes such models. For instance, the predictive ability of Systematic Coronary Risk Evaluation (SCORE) model [12] in T2D patients is not been considered. However, the predictive ability of SCORE in T2D patients has been assessed by three studies and was similar to other CVD prediction models reviewed for this study [13-14]. Based on the review, the predictors considered for fatal CVD over 10 years include sex, age, smoking, systolic blood pressure and either total cholesterol or ratio total/high-density lipoprotein-cholesterol.

The development and design of rat model with T2D having CVD risks were described earlier [15]. 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 likely to participate in athero protective functions such as reverse cholesterol transport.

The major early detective component of T2D is an insulin resistance and hence considered as an independent risk factor for CVD. The reversible and non-invasive features for assessing CVD would be endothelial dysfunction and increase of carotid intima-media thickness [16]. Thus, early identification of insulin resistance and impaired endothelial function helps for easy prediction of those at particular risk of CVD and hence enables targeting of aggressive risk factors to control the CVD. The initial task before studying the T2D patients for CVD risk factor analysis is to study the major important factors to analysis the results of CVD.

II-A Important risk factors in Cardio Vascular Disease (CVD) with Type 2 Diabetes (T2D)

CVD is a serious but preventable complication of type 2 diabetes (T2D) however if untreated would lead to substantial disease burden, increased use of health services and higher risk of premature mortality. Following are the specific modifiable and non-modifiable risk factors for the prediction of CVD in T2D patients.

Modifiable risk factors

- 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 CVD
- Social history including cultural identity, ethnicity, socioeconomic status and mental health Related conditions
- Diabetes
- Kidney function (microalbumin ± urine protein, eGFR)
- Familial hypercholesterolaemia
- Evidence of atrial fibrillation (history, examination, electrocardiogram

However, managing the numerous risk factors responsible for CVD in T2D becomes a present challenge for primary care clinicians and thus majorly influencing their decisions about treatment approaches for this complex disease [17].The established risk factors include poor control of glycated haemoglobin ($HbA1_c$) levels, systolic blood pressure, and lipid levels, along with age, sex, ethnicity, smoking status, and disease duration [18-19]. The efforts to reduce the average $HbA1_c$ values for diabetic patients are still been continuous and challenging over the past few years [20]. The control of CVD risk factors such as blood pressure, diet, exercise, and treatment adherence has not been still improved inspite of the wide dissemination of evidence-based

guidelines and the availability of new therapeutic agents. Moreover, T2D patients develop hypertensive which also contributes to the premature development of vascular disease.

The predictive risk analysis of CVD in T2D patients has been done by the development of hybrid prediction model in this study. It includes pre-processing method to reduce the irrelevant and missed data information using principal component analysis (PCA) and dimensionality reduction. The attribute analysis of the risk factors was then performed based on the information gain and entropy values. In Hybrid prediction model, the prediction is based on the unsupervised classification methods known as clustering methods to measure the similarity among the attributes for labelling the attributes under classes and then performing support vector machine learning based classification.

III. PROPOSED METHODOLOGY

In Cardiovascular complications are now the leading causes of diabetes-related morbidity and mortality. The public health impact of cardiovascular disease (CVD) in patients with diabetes is already enormous and is increasing. Managing the numerous risk factors responsible for CVD in T2D represents an ongoing challenge for primary care clinicians, strongly influencing their decisions about treatment approaches for this complex disease. The major objective of this proposed work is to examine the common clinical and behavioral factors that contribute to cardiovascular disease (CVD) risk (ie, 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 patients (T2D) data with CVD risk using principal component analysis (PCA) and dimensionality reduction is also performed using PCA. Then CVD risk factors are estimated based on the metrics like information gain and entropy measurements to enhance the prediction accuracy results. The estimated CVD risk factors are used for unsupervised classification using Improved Fuzzy C Means (IFCM) clustering methods, which data is used prediction of T2D for CVD risk factors. Then perform supervised classification task for prediction of type 2 diabetes patients with CVD risk are predicted using support vector machine (SVM) .The entire representation of the proposed system is illustrated in Figure 1.

A. Dataset Information

From the clinical diabetic patient's records, datasets were collected which will be included in one of the following attributes. Number of Pregnancy, Plasma glucose concentration, Diastolic blood pressure (mm Hg),Triceps skin fold thickness (mm), 2 h serum insulin (mu U/ml),Body mass index (weight in kg/(height in m)^2), Diabetes pedigree function, Age (years),Class variable (0 or 1). These datasets would fall under 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 haemoglobin) (%), Fibrinogen (mg/dl), ultrasensitive C reactive protein (us-CRP) (mg/L). The change in the value of each and every attribute will helps to analyse the risk factor of CVD for T2D patients. The management of all these complex varied attributes becomes the major challenge for improving the efficient diagnosis and decision making by primary care clinicians.

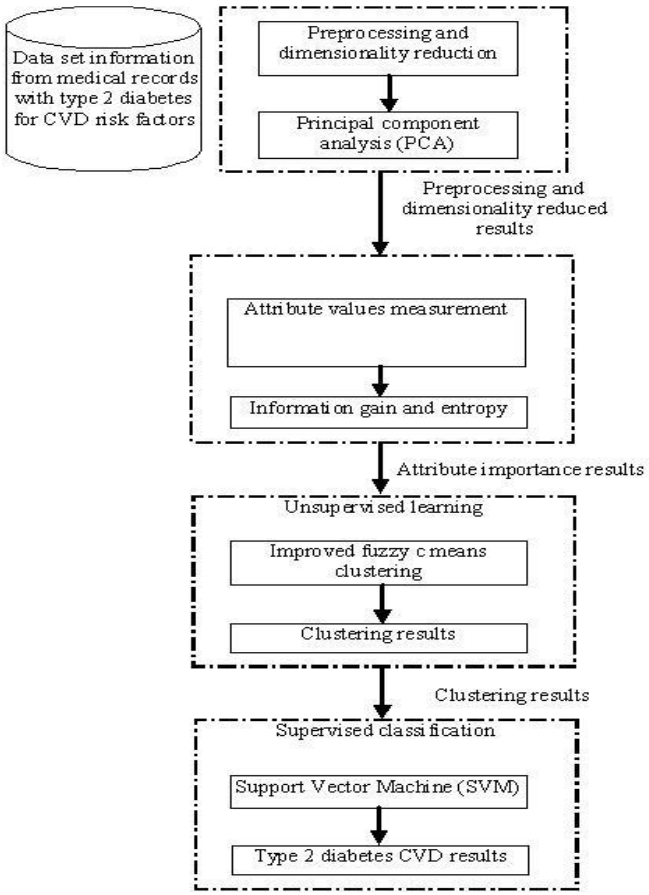

Fig.1. Architecture Diagram of Proposed Work

B. PREPROCESSING AND DIMENSIONALITY REDUCTION USING PRINCIPAL COMPONENT ANALYSIS (PCA)

The most important aspect of this model is the quality of the dataset as it directly influences the quality of the results from the analysis. Hence, the data should be carefully collected, integrated, characterized, and prepared for analysis. As mentioned earlier, principal component analysis (PCA) has been implemented for pre-processing of the datasets of T2D with CVD risk factors. Once the datasets are collected and sorted out based on the attributes as mentioned above, the datasets which does not fall on any of the attributes are removed. It is done by considering the Eigen vector of dataset that gets associated with largest Eigen value as the most important vector that reflects the greatest variance for prediction process.

In such a way all the datasets are pre-processed and removed in the pre-processing stage. A preliminary analysis of the data assumes zero for missing data. The various variables used in this study are pregnancy, plasma–glucose, diastolic BP, Body mass index, diabetes pedigree function, age, serum–insulin, triceps skin fold and class. As it does not make sense to have the value of 0 for a variable such as plasma–glucose concentration of a living human, all the observations with zero entries are removed. Also, the counts of missing values for the variables like serum–insulin and triceps skin fold are very high in this analysis. PCA for pre-processing of T2D with CVD risk factor has also been successfully applied in pattern recognition such as face classification [21]. The analysis considers $N = (X_1, X_2, \ldots . X_n)$ as the number of population of T2D patients with the CVD risk factors and t as the dimension of dataset D. The subspace of the attribute value is also found for which the basis vectors correspond to the maximum-variance direction of the original T2D data space. PCA involved here is a linear transform. Let W represents the linear transformation that maps the original $t-$dimensional T2D data space into an f-dimensional reduced irrelevant and missing attribute data represented as $f \ll t$. Equation (1) below shows the new reduced dimensional and irrelevant data variable vectors

$$z_j \in R^f$$

$$z_j = W^T x_j , j = 1, \ldots . N \quad (1)$$

where $Q = XX^T , X = \{x_1, \ldots . x_N\}$ is the covariance matrix and λ_j is the eigen value associated with the eigenvector e_j. The eigenvectors are sorted from higher to lower based on their corresponding Eigen values. The eigenvector associated with largest eigen value is the most important variable and hence considered as the data vector that reflects the greatest variance [22]. PCA thus employs the entire collected variables from the record of the T2D patients with CVD risk factors and acquires a set of projection attribute vectors to extract most important global variable and data vector from given training samples. The performance of PCA depends on the higher number of relevant data of T2D with CVD risk factor ones.

C. Attribute Values Measurement

*T*o measure the importance of the risk factor for CVD, first analysis 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 variableXwith possible values $\{x_1, \ldots, x_n\}$ be the attributes with risk factors CVD and probability mass function P(X) as:

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

Where E is the expected value operator (maximum threshold value results), and I is the information content (value of content) from patient record of X and $I(X)$ is a random variable. When a finite sample is considered, the entropy can explicitly be written as

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

Where b is the base of the logarithm used. Common value for b is 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

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

A risk equation was created for estimation of the risk of CVD, using q and the HRs for the nine predictors, after the calculation of entropy and information gain values, the CVD risk factors in the T2D patients is calculated as:

$$CVD_{risk} = (1 - \exp[-q_r \times \alpha_1^{BMI} \times \alpha_2^{W} \times \alpha_3^{WC} \times \alpha_4^{SBP} \times \alpha_5^{DBP} \times \alpha_6^{G} \times \alpha_7^{TC} \times \alpha_8^{HDL-C} \times \alpha_9^{LDL-C} \times \alpha_{10}^{TRC} \times \alpha_{11}^{HbA1C} \times \alpha_{12}^{F} \times \alpha_{13}^{CRP}]) \quad (6)$$

For each and every attribute such as 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), the highest value greater than the thresholds value is been selected.

IV. IMPROVED FUZZY C MEANS CLUSTERING (IFCM)

Validation of the chosen classes using the unsupervised methods is the primary step before the application of the Classification algorithms. An Improved Fuzzy c means (IFCM) clustering is been employed in this model to validate the preprocessed dataset followed by assigning class labels to group similar cluster using the clustering algorithm. In conventional FCM clustering methods, only distance measure evaluates the difference between two individual data points thus ignoring the global view of the data distribution. Other existing fuzzy c-means based clustering algorithms also consider only hyper-spherical clusters in data space. In order to overcome these drawbacks of existing clustering algorithm, the present work follows improved fuzzy clustering in which density function measures the similarity or distance measures between the data points. Hence, the density of data points in a cluster should be distinctly different from other clusters in a data set. A regulatory factor based on cluster density is proposed to correct the distance measure in contrast with the conventional FCM. IFCM differs from other approaches as the regulator uses both the shape of the data set and the middle result of iteration operation. Further, the distance measure function is dynamically corrected by the regulatory factor until the objective criterion is achieved. Given a CVD risk factor data for type 2 diabetes dataset $X = (x_1, \ldots . x_n)$ for every data point x_i, the dot density is usually defined as:

$$z_i = \sum_{j=1, j\neq i}^{n} \frac{1}{d_{ij}} d_{ij} \leq e, 1 \leq i \leq n \quad (7)$$

Where e is the effective radius for density evaluation. As the value of e directly affects the results, it should be set appropriately according to the application scenarios. The dot density is relatively higher when e is set bigger. To avoid this uncertainty, Eq.(7) is simplified by assuming the uniform distribution of data set and there are no outliers. The new definition is thus:

$$z_i = \frac{1}{\min(\{d_{ij}\})} d_{ij} \leq e, 1 \leq i \leq n \quad (8)$$

The inverse of the minimum distance value among the neighbours can be approximated as the dot density. However, dot density considers only the local distribution of the T2D variables. Therefore, cluster density is introduced which represents the weighted linear combination of dot densities as expressed in Eq. (9).

$$\hat{z}_i = \frac{\sum_{j=1}^{n} \alpha_{ij} w_{ij} z_{ij}}{\sum_{j=1}^{n} \alpha_{ij} w_{ij}} d_{ij} \leq e, 1 \leq i \leq n \quad (9)$$

Where $\alpha_{ij}$is the category label of data point x_j and $w_{ij}$is the weight of x_j. The value of $\alpha_{ij} = 1$ when x_j most likely belongs to the cluster $i$otherwise$\alpha_{ij} = 0$. And $w_{ij}$is a positive constant which can be adjusted according to the users. Cluster density $\hat{z}_i$considers the global shape of the dataset in a cluster and uses the dynamical membership degrees in the iteration process. Furthermore, using cluster density instead of dot density can reduce the computation consumption during clustering. Using the cluster density, the distance measure is corrected as Eq. (10)

$$\hat{d}_{ij}^2 = \frac{\left|\left|x_j - v_i\right|\right|^2}{\hat{z}_i} \quad 1 \leq i \leq c, 1 \leq j \leq n, \quad (10)$$

Thus, the optimization expression can be written based on Eqs. (9) as follows

$$J_{FCM-CD}(U,V,X) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} \left||x_j - v_i\right||^2 \frac{\sum_{k=1}^{n} \alpha_{ik} w_{ik}}{\sum_{k=1}^{n} \alpha_{ik} w_{ik}\, z_k} \quad (11)$$

By applying Lagrange Multiplying Method to Eq. (10), two updated equations can be obtained as given in Eqs. (12) and (13).

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}} \quad 1 \le i \le c \quad (12)$$

$$u_{ij} = \frac{\hat{d}_{ij}^{-2/(m-1)}}{\sum_{j=1}^{n} \hat{d}_{kj}^{-2/(m-1)}} \quad (13)$$

The process of stepwise regression involves of the following steps:

1. Choose the number of clusters c, fuzziness index m, iteration error ", maximum iterations T, and initialize the membership degree matrix $U^{(0)}$.
2. Get the initial centroids using Eq. (13).
3. Calculate the dot density of every data point using Eq. (7).
4. Update the membership degree matrix $U^{(t)}$ and cluster centroids $V^{(t)}$using Eqs. (12) and (13) when the iteration index is $t(t = 1,2,\dots,T)$,
5. Calculate the value of the objective function $J^{(t)}$using Eq. (11).
6. Stop the iteration and get the membership degree matrix U and the cluster centroids V, if $|U^{(t)} - U^{(t+1)}| < \varepsilon$ or $t = T$, otherwise set and return to step (4).

V. Support Vector Machine Classification for Prediction of Type 2 Diabetes with CVD Risks

From these results, clusters are formed as either class label 'yes' or class label 'no' for classification of T2D patients in order to reduce dimensionality data as in the PCA. Finally classification task for unsupervised class labels are performed from the results of Improved Fuzzy c means (IFCM) clustering. The clustered results are taken as input to support vector machine (SVM) classification task for the prediction of T2D patients with CVD risks. The SVM is a classification technique based on statistical learning theory [23, 24] which has been used successfully to large datasets in many challenging non-linear classification problems. The SVM algorithm finds a hyperplane that optimally splits the T2D clustered data results from IFCM training set. The optimal hyperplane can be distinguished by the maximum margin of separation between all clustered input data as training points and the hyperplane. Thus the two-dimensional prediction problem has to be solved to find a line that "best" separates clustered data points in the positive class from points in the negative class for predicting T2D with CVD risk factors [27]. The hyperplane is characterized by a decision function f(x) as:

$$f(x) = sgn(\langle w, \phi(x)\rangle) + b \quad (14)$$

where $\boldsymbol{w}$ is the weight vector for clustered data orthogonal to the hyperplane, "b" is a scalar that represents the margin of the hyperplane, "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 $\cdot,\cdot$ representing the dot product. $Sgn$is the signum function. If **w** has unit length, then $\langle w, \phi(x)\rangle$is the length of $\phi(x)$along the direction of $\boldsymbol{w}$. Generally $\boldsymbol{w}$will be scaled by $||\boldsymbol{w}||$.The algorithm needs to find the normal vector "$\boldsymbol{w}$" that leads to the largest "b" of the hyperplane. Kernel function is defined as a Mercer Kernel according to Mercer theorem [25]. This gives the mapping to clustered data as,

$$\phi(x) = \left(\sqrt{\lambda_1}\psi_1(x)\sqrt{\lambda_2}\psi_2(y), \dots\right)^T \quad (15)$$

VI. Experimental Results

The data were not collected specifically but as a part of routine patient management, UCHT collected diabetic patients' information from 2000 to 2004 from a clinical information system (Diamond, Hicom Technology). The data contained both physiological and laboratory information for about 3857 patients, described by 410 features. The patients are not only of T2D but also of other types of diabetes such as T1D and gestational diabetes. Some measure of evaluation has to be introduced for improving the performance. One common measure is the accuracy [24] 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. 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. In this study we use the following equation to determine the measure of accuracy Eq. (16), specificity Eq. (17), sensitivity Eq. (18)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$	(16)
$$Sensitivity = \frac{TP}{TP + FN}$$	(17)
$$Specificity = \frac{TN}{TN + FP}$$	(18)

These parameters can be used to measure accuracy, sensitivity and specificity respectively. The results are shown in Table 1 and are found to be better than the accuracies of other classifiers in the related studies for Pima Indian diabetes dataset.

TABLE I Prediction methods results

Parameters	K-C4.5	IFCM-SVM
Accuracy	92.3	93.8
Sensitivity	89.4	90.49
Specificity	60.8	54.7

Table 2 illustrates the cardio vascular risk components of the T2D patients, their highest chance of the risk factor values and their reference values to measure the risk factor value for prediction of CVD risk in T2D patients. The reference prediction values of the good control group exhibits younger age, shorter diagnosis duration, lower body-mass index (BMI), lower cholesterol and lower blood pressure levels compared to those in the bad control represented in bracket as standard deviation. Younger adults tend to manage their blood glucose level better; also higher cholesterol and triglycerides level are associated with bad blood glucose control and hence a well-controlled BMI may help an individual manage his/her diabetes better.

TABLE II Cvd Risk Factors And their Mean, Standard Deviation Values

Variable. No	Cardiovascular risk components	Values in the sample mean ± SD [95% CI]	Reference values
1	BMI	34.8 ± 5.8 [32.9-36.7]	18.5-24.9
2	Weight (kg)	85.5 ± 14.9 [80.6-90.4]	

3	Waist circumference (cm)	106.7 ± 12.2 [102.7-110.7]	< 88 females/102 males
4	SBP (mmHg)	133.6 ± 13.8 [128.4-138.9]	< 140[1]
5	DBP (mmHg)	79.3 ± 5.9 [7781.–6]	< 90[1]
6	Glucose (mg/dl)	141.3 ± 40.7 [127.5-155]	< 100[3]
7	Total cholesterol (mg/dl)	171.9 ± 23.4 [164180]–	< 200[1]
8	HDL-c (mg/dl)	56.8 ± 11 [5360.–5]	> 50[1]
9	LDL-c (mg/dl)	92.1 ± 20.7 [85.2-99.1]	< 130[1]
10	Triglycerides (mg/dl)	123.1 ± 76.7 [97.2-149]	< 200[1]
11	HbA_{1c} (%)	6.8 ± 1.3 [6.3-7.2]	< 7%[3]
12	Fibrinogen (mg/dl)	339.3 ± 85.8 [309.8-368.8]	200-400[4]
13	us-CRP (mg/L)	6.3 ± 7.4 [3.8-8.8]	< 3[1]

Prediction accuracy of the proposed IFCM-SVM based prediction methods is marked by achieving higher classification accuracy than the existing classification methods with the prediction accuracy of K-C4.5 and is compared as illustrated in Figure 2. It is inferred that the prediction accuracy of the proposed system are high because of the performance of pre-processing and dimensionality reduction before the prediction is applied.

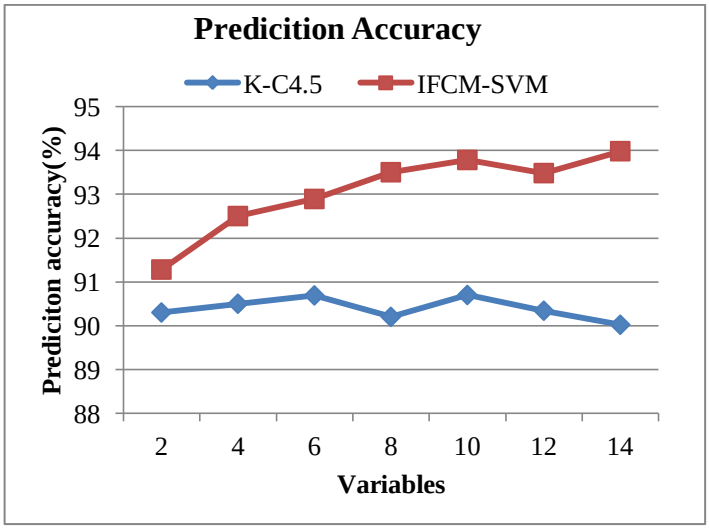

Fig.2. Prediction accuracy of the prediction methods

The proposed IFCM-SVM based prediction methods also achieve higher sensitivity than the existing classification methods K-C4.5. Figure 3 shows that the sensitivity of the proposed system is high because of the preprocessing and dimensionality reduction performed initially before the prediction is applied.

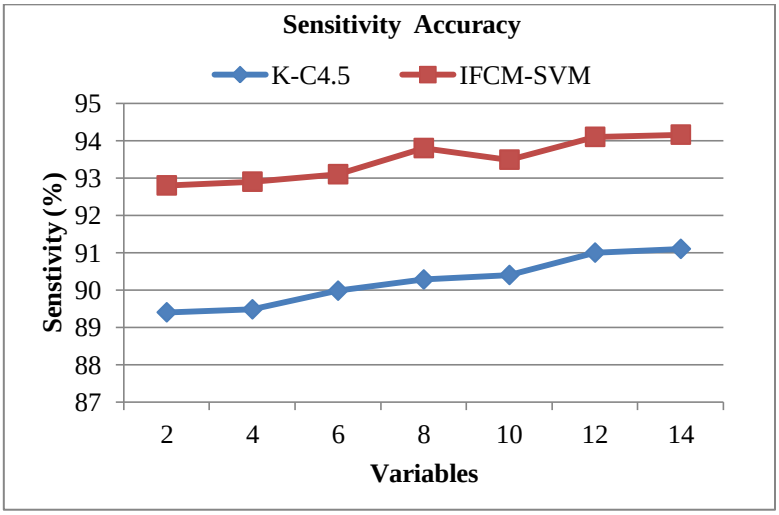

Fig.3. Sensitivity accuracy of the prediction methods

Specificity accuracy of the proposed IFCM-SVM based prediction methods is however lesser than the existing classification methods of K-C4.5 prediction. As shown in Figure 4, the specificity of the proposed system is lesser thanK-C4.5 prediction sensitivity and the reason is attributed to the preprocessing and dimensionality reduction performed before the prediction is applied.

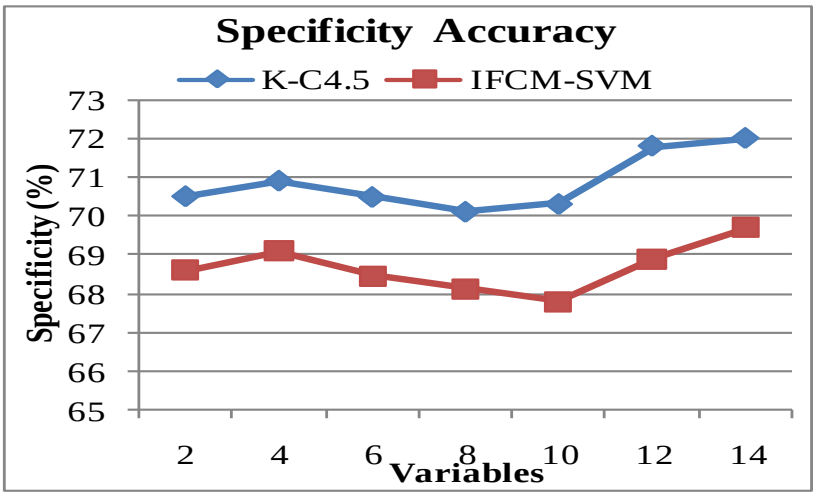

Fig.4. Specificity accuracy of the prediction methods

CONCLUSION

T2D confers a high degree of cardiovascular risk brought about by multiplicative risk factors. Earlier it was assumed that aggressive lowering of glucose level would result in cardiovascular risk reduction in T2D but this effect may not become apparent for many years. However, multiple risk factor modulation including lifestyle interventions, blood pressure lowering and lipid management along with glucose control would result in significant and early diagnosis of CVD in T2D patients throughout their lifespan. Analysis of CVD risk factors plays major role in predicting T2D patient's with CVD risks. Many studies existing present a CVD prediction model that can be applied to the diabetes population still only a minority of studies has externally validated this large number of clinical prediction models in a diabetes population. Assessment of the impact on diabetes treatment and complications has been made for only one prediction model. An improved efficient prediction model is validated in this study to analyse the risk of CVD factors in T2D patients' records. The patients records considered as the datasets are initially pre-processed followed by dimensionality reduction of the features using PCA. The information gain and entropy measure are used for calculation of the risk factors and then unsupervised learning is performed using Improved Fuzzy C Means (IFCM) clustering algorithm. Finally support vector machine (SVM) is employed to perform prediction of T2D with

CVD risk factors. New studies investigating the prediction of CVD among T2D patients should focus on further validating the performance of existing K means unsupervised learning and supervised learning with the prediction accuracy of C45. Moreover, assessing the impact of existing prediction models on treatment and prevention of cardiovascular events is the immediate need rather than 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. Xiangjin Zhou, Feng Li, Lingyun Xiang, "Blind Detection of Synthesized JPEG Image via Quantization Noise Analysis", International Review on Computers and Software (IRECOS),Vol. 7. n. 5, pp. 2016-2021.
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.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.

