ISSN:2320-0790

COMPUSOFT
An International Journal of Advanced Computer Technology

Feature Selection Using Particle Swarm Optimization for Predicting the Risk of Cardiovascular Disease in Type-II Diabetic Patients

P. Radha[1], Dr. B. Srinivasan[2]

[1] Ph. D Scholar, Dept. Of Computer Science, Karpagam University, Coimbatore & Asst. Prof, Dept. Of Computer Science, Vellalar College for Women, Thindal, Erode.
[2]Associate Professor, Dept. Of Computer Science, Gobi Arts and Science College, Gobichettiplayam.

Abstract: Diabetes is the most common disease nowadays in all populations and in all age groups. A wide range of computational methods and tools for data analysis are available to predict the T2D patients with CVD risk factors. Efficient predictive modelling is required for medical researchers and practitioners to improve the prediction accuracy of the classification methods .The aim of this research was to identify significant factors influencing type 2 diabetes control with CVD risk factors, by applying particle swarm optimization feature selection system to improve prediction accuracy and knowledge discovery. Proposed system consists of four major steps such as pre-processing and dimensionality reduction of type 2 diabetes with CVD factors, Attribute Value Measurement, Feature Selection, and Hybrid Prediction Model. In proposed methods the pre-processing and dimensionality reduction of the patients records is performed by using Kullback Leiber Divergence(KLD) Principal component analysis (PCA), then attribute values measurement is performed using Fast Correlation-Based Filter Solution(FCBFS), feature selection is performed by using Particle Swarm Optimization (PSO), finally hybrid prediction model which uses Improved Fuzzy C Means (IFCM) clustering algorithm aimed at validating chosen class label of given data and subsequently applying Extreme Learning Machine(ELM) classification algorithm to the result set.

Keywords: Classification, Hybrid Prediction Model, Fuzzy c means clustering(FCM), Principal Component Analysis (PCA), Kullback Leiber Divergence(KLD), Extreme Learning Machine (ELM), Fast Correlation-Based Filter Solution(FCBFS), Particle Swarm Optimization (PSO)

1. INTRODUCTION

The worldwide prevalence of type-II diabetes is growing rapidly, reaching epidemic proportions .One of the major reasons of the increased prevalence in developing countries is the adoption of the so-called western lifestyle that is, a high intake of energy dense food and a low physical activity pattern. These lifestyle changes lead to one of the key abnormalities underlying type 2 diabetes that is, insulin resistance. Insulin resistance is associated with central obesity, hyperinsulinaemia, polycystic ovary syndrome, hypertension, and dyslipidaemia. Hyperglycaemia, the established diagnostic marker of diabetes mellitus, is the result of the second key feature, progressive pancreatic b-cell failure. It is well recognised that type 2 diabetic patients have an excess risk of developing atherosclerosis, resulting in high cardiovascular disease morbidity and mortality [1] . Therefore, with the rise of the prevalence of diabetes, it may be expected that the global burden of cardiovascular disease will also increase.

Among individuals with type 2 diabetes, cardiovascular disease (CVD) is the leading cause of morbidity and mortality [2] adults with diabetes have a two- to fourfold higher risk of CVD compared with those without diabetes[3-4] . Diabetes is also accompanied by a significantly increased prevalence of hypertension and dyslipidemia [5] .It is reasonable to postulate that in many individuals, excess weight gives rise to diabetes, hypertension, and dyslipidemia, thereby leading to frank CVD [6] . This seemingly simple algorithm is undoubtedly more complex because (1) Many studies show that hyperglycemia at pre-diabetic levels is an independent risk factor for CVD [7] (2) Central obesity (i.e., intra-abdominal or visceral fat) may have a greater detrimental effect than overall weight/BMI [8] , and (3) there is a complex relationship between lipid metabolism and hyperglycemia.

Cardiovascular disease (CVD) is a serious but preventable complication of type 2 diabetes mellitus (T2DM) that results in substantial disease burden, increased health services use, and higher risk of premature mortality. People with diabetes are also at a greatly increased risk of cardiovascular, peripheral vascular, and cerebro-vascular disease [8], known as macrovascular complications. There are two main classes of diabetes, which are diagnosed ultimately by the severity of the insulin deficiency. Insulin-dependent diabetes mellitus or type 1 diabetes is an insulinopenic state, usually seen in young people, but it can occur at any age [9]. Non-insulin-dependent diabetes mellitus or type 2 diabetes is the more common metabolic disorder that usually develops in overweight, older adults, but an increasing number of cases occur in younger age groups. However, new prediction models for the diabetes population have been developed since this review, and many more prediction models exist that can be applied to people with diabetes.

Among these stages analysis of type 2 diabetes with CVD risk factors, important features in the dataset are not selected ,irrelevant data in the T2D patients records are also removed so it degrades the performance of the T2D patients prediction results .In order to overcome these issues in this work first the collection data and removal of the irrelevant data ,selection of the most important features for prediction plays major important role to improve the prediction results of the type 2 diabetes results with CVD risks. The aim of this study was to analyze CVD risk factors in type 2 diabetic patients. The major important steps of the proposed works as follows: Preprocessing of the data using dimensionality reduction Kullback Leiber Divergence(KLD) -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 CVD are analyzed using similarity measure Fast Correlation-Based Filter Solution(FCBFS) . The purpose of this study is to reduce the irrelevant or unimportant features in the type 2 diabetes patient records after the similarity measurement from the FCBFS for CVD risk factors, then 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 Extreme Learning Machine (ELM) classification methods. The proposed hybrid prediction model is different from the existing methods such as IFCM-SVM and K-C4.5 methods since it select important features in the T2D patients records using the particle swarm optimization algorithm The aim of this study was to identify all CVD 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

In modern medicine, large amounts of data are generated, but there is a widening gap between data collection and data comprehension. It is often impossible to process all of the data available and to make a rational decision on basic trends. Thus, there is a growing pressure for intelligent data analysis such as data mining to facilitate the creation of knowledge to support clinicians in making decisions.

Although there are many approaches for estimating the risk of diabetes and CVD [10] virtually none have been validated much beyond the population from which they were constructed. There is one such tool, however (available free on the Internet at http://www.diabetes.org/diabetesphd), that has been extensively validated across many widely differing clinical trials, and it incorporates virtually all known CVD risk factors. Although it can be used to predict the risk of developing CVD/diabetes or the effects of treatment after developing diabetes/CVD, this tool and other risk-assessment algorithms are rarely used in clinical practice

Conversely, emerging evidence suggests that simply ascertaining a person’s blood glucose level, blood pressure, LDL cholesterol level, and tobacco use and noting the presence of obesity may be sufficient to initiate the appropriate interventions to prevent or identify diabetes and emerging CVD [11-12]. Even borderline abnormalities, especially if they are multiple, may well presage future problems and should be addressed.

Classification techniques such as support vector machines [13], neural networks [14]. Su et al. [15] used four data mining approaches (neural network, decision tree, logistic regression and rough sets) to select the relevant features for the diabetes diagnosis, and also evaluated their performance. For example, the features selected by the neural network were evaluated by neural network; that is to say, every method is both a feature selector and a classifier. Every technique has its specific advantages and disadvantages, and is applicable for different research problems. These methods don’t follow feature selection methods to improve classification accuracy.

FSSMC [16] which has been successfully applied in data mining applications [17], was used to investigate those important factors in the type 2 diabetes data set. ReliefF did not address the problem of multi valued attributes. At present, the similarity measurement applied in ReliefF is a numerical method, and if the two selected instances have the same categorical value, the result of difference function is 0, otherwise is 1. This definition cannot measure the contribution of multi-class (3) values to class labels. Among them feature selection methods CVD risk factors are also not analyzed accurately for T2D patients.

3. PROPOSED METHODOLOGY

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. The pathophysiology of CVD in diabetes is complex and not dependent on the effects of hyperglycaemia alone. InT2D a constellation of risk factors contribute to the development of early CVD, including hypertension and dyslipidaemia. These result in metabolic changes which, coupled with a sedentary lifestyle, obesity and smoking, enhance the deleterious effects of hyperglycaemia and accelerate atherosclerotic disease in the vasculature. People with T1DM are generally diagnosed at a young age and exposure to hyperglycaemia takes place over a prolonged time period compared with type 2 diabetes patients (T2D) data. Selection of the important features in the T2D also becomes a difficult task, in order to overcome these problem and select most important features in the CVD risk factors, in this work mainly focus on the feature selection methods it selects the CVD risk factors important features in the T2D patient data 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 patient (T2D) data with CVD risk using hybrid principal component analysis (HPCA) in which the weight values of the PCA are estimated using the kullback leiber divergence it is named as KLD-PCA and dimensionality reduction is also performed using KLD-PCA. Then CVD risk factors are estimated based on the Fast Correlation-Based Filter Solution(FCBFS) ,to reduce unimportant features in the data feature selection is performed using Particle Swarm Optimization (PSO) to enhance the prediction accuracy results. The selected feature with estimated CVD 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 Extreme Learning Machine (ELM) .The entire representation of the proposed system is illustrated in Figure 1.

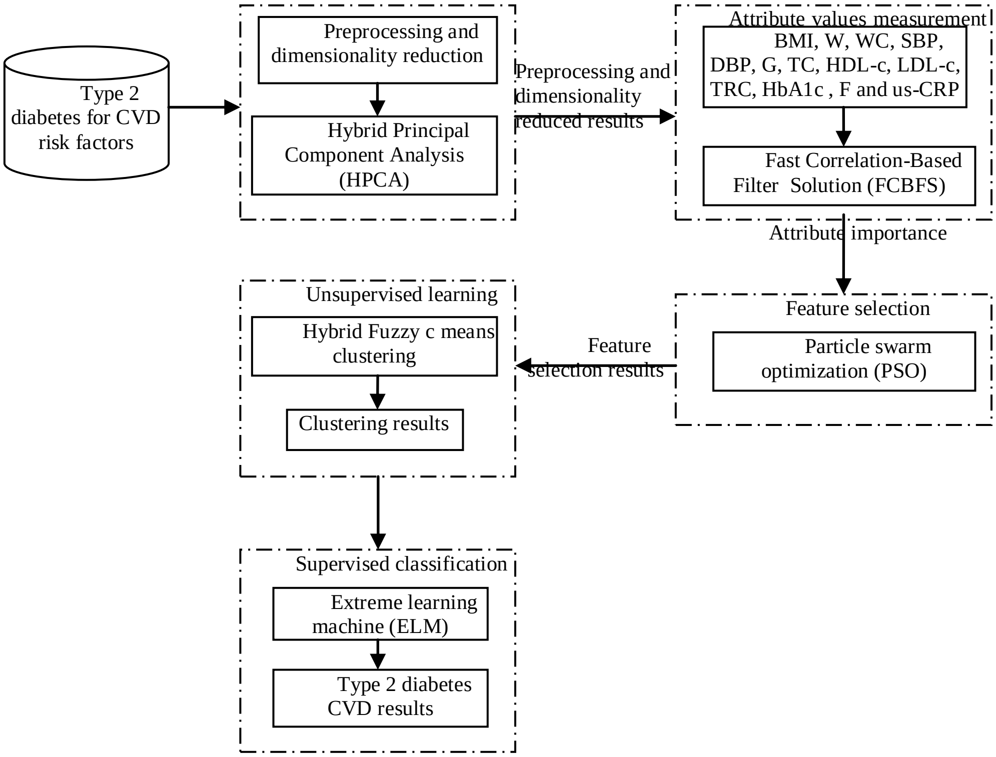
Type 2 diabetes for CVD risk factors
Preprocessing and dimensionality reduction
Hybrid Principal Component Analysis (HPCA)
Preprocessing and dimensionality reduced results
Attribute values measurement
BMI, W, WC, SBP, DBP, G, TC, HDL-c, LDL-c, TRC, HbA1c , F and us-CRP
Fast Correlation-Based Filter Solution (FCBFS)
Attribute importance
Feature selection
Particle swarm optimization (PSO)
Feature selection results
Unsupervised learning
Hybrid Fuzzy c means clustering
Clustering results
Supervised classification
Extreme learning machine (ELM)
Type 2 diabetes CVD results

Figure 1: Architecture of the proposed methodology

The major objective of this proposed work is to examine the common clinical and behavioral factors that

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/(height in m)^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). If the value of each and every attributes values are changed to analysis the risk factor of CVD for type 2 diabetes (T2D). 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 [25].

3.2 PREPROCESSING AND DIMENSIONALITY REDUCTION USING KULLBACK LEIBER DIVERGENCE WITH PRINCIPAL COMPONENT ANALYSIS (KLD-PCA)

The quality of the data is the most important aspect as it influences the quality of the results from the analysis of data preprocessing in order to improve the quality of the mining result and the efficiency of the mining process [18-19]. In this study, the Kullback Leiber Divergence with Principal Component Analysis (KLD-PCA) for preprocessing of type 2 diabetes (T2D) with CVD risk factors is applied as mentioned above. In PCA 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 major problem of the PCA method is that the weight value of the PCA are randomly generated in order to overcome these problem the weight values are estimated based on the kullback leiber divergence (KLD) . Proposed KLD-PCA for preprocessing of T2D with CVD risk factor .As mentioned above $\mathrm{N} = (\mathrm{X}_1, \mathrm{X}_2, \ldots . \mathrm{X}_\mathrm{n})$ be the number of type 2 diabetes patient's hospital data with the CVD risk factors and t dimension of dataset D, respectively. KLD -PCA finds a subspace of the attribute value whose basis vectors correspond to the maximum-variance direction of the original T2D data space. Let $\mathbf{W}$ represents the linear transformation that maps the original $\mathrm{t}-$dimensional T2D data space into an $\mathbf{f}$-dimensional reduced irrelevant and missing attribute data where normally $\mathrm{f} \ll \mathrm{t}$. Equation (1) shows the new reduced dimensional and reduced irrelevant data variable vectors $\mathrm{z}_\mathrm{j} \in \mathrm{R}^\mathrm{f}$

$$z_j = W^T x_j \, , j = 1, \ldots . N \tag{1}$$

$$\lambda_j e_j = Q e_j \, , j = 1, \ldots . N \, , \tag{2}$$

where $Q = XX^T \, , X = \{x_1, \ldots . x_N\}$

Here Q is the covariance matrix and λ_j the eigen value associated with the eigenvector e_j.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 [20]. PCA employs the entire T2D patient hospital record variables with CVD 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 CVD risk factor ones. In equation (1) the weight transformation matrix is calculated based on the KLD methods in the PCA. $KLD(C|a)$ appears in the information theoretic literature under various guises. For instance, it can be viewed as a special case of the cross-entropy or the discrimination, a measure which defines the information theoretic similarity between two probability distributions. In this sense, the $KLD(C|a)$ is a measure of how dissimilar our a priori and a posteriori beliefs are about C–useful class value imply a high degree of dissimilarity. It can be interpreted as a distance measure where distance corresponds to the amount of divergence between a priori distribution and a posteriori distribution. It becomes zero if and only if both a priori and a posteriori distributions are identical,

$$KLD(C|a_{kl}) = \sum_c P(c|a_{kl}) \log \left(\frac{P(c|a_{kl})}{P(c)} \right) \tag{3}$$

where a_{kl} means the l value of the k^th attribute in training data. The weight of a attributes can be defined as the weighted average of the KL measures across the attribute values. Therefore, the weight of attribute k, denoted as $w_{avg\,(k)}$, is

$$w_{avg}\,(k) = \sum_{l|k} \frac{\#a_{kl}}{N} KLD(C|a_{kl}) \qquad ((4)$$

$$= \sum_{l|k} P(a_{kl}) \, KLD(C|a_{kl})$$

where $\#a_{kl}$ represents the number of instances that have the value of a_{kl} and the N means the total number of training instances. In this formula, $P(a_{kl})$ means the probability that the attribute k has the value of a_{kl} . The final form of the weight value for attribute is denoted as,

$$w_k = \frac{\sum_{l|k} P(a_{kl}) \sum_c p(C|a_{kl}) \log\left(\frac{p(C|a_{kl})}{p(C)}\right)}{-Z\,.\sum_{l|k} P(a_{kl}) \log\left(p(a_{kl})\right)} \qquad (5)$$

The new variance of k^{th} attribute is calculated as follows:

$$\delta_{newk}\,(N-1) = \sum_{j=1}^{N} (n_{x_{kj}} - n\vec{x}_k)^2 \qquad (6)$$

$$n = \sqrt{\frac{\delta_{newk}\,(N-1)}{\sum_{j=1}^{N}(x_{jk} - \vec{x}_k)^2}} \qquad (7)$$

N is the number of samples and x_{ji} , $\vec{x}_i$are i^{th} attribute of j^{th} sample and mean of k^{th} attribute respectively. After this adjustment, PCA is employed on data.

3.3 ATTRIBUTE VALUES MEASUREMENT

In this work measure the values of the attributes for prediction of the CVD risk factor in T2D patients based on Fast Correlation-Based Filter Solution for prediction of T2D with CVD risks factors. 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 X with possible values $\{\mathrm{a}_1, \dots, \mathrm{a}_\mathrm{n}\}$ be the attributes with risk factors CVD and probability mass function P(a) as:

$$H(a) = E[I(a)] = E[-ln(P(a))] \qquad (8)$$

Here E is the expected value operator (maximum threshold value results), and I is the information content (value of content) from patient record of X. $I(a)$ 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

$$H(a) = \sum_i P(a_i) I(a_i) = -\sum_i P(a_i) \log_\mathrm{b} P(a_i) \qquad (9)$$

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

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

$$\mathrm{Gain(S,A) = Entropy\ (S)} - \sum_{\mathrm{v \in Values\ \ (A)}} \frac{|\mathrm{S_v}|}{|\mathrm{S}|}.\mathrm{Entropy(S_v)} \qquad (10)$$

(10)

But this work sometimes the linear correlation measures may not be able to capture correlations that are not linear in nature, in order to solve this problem proposed a symmetrical uncertainty,

$$\mathrm{SU(a,b)} = 2\left[\frac{\mathrm{IG(a|b)}}{\mathrm{H(a) + H(b)}}\right] \qquad (11)$$

It compensates for information gain's bias toward attributes with more values and normalizes its values to the range [0, 1] with the value 1 indicating that knowledge of the value of either one completely predicts the value of the other and the value 0 indicating that a and b are independent. Given a data set with N number of the samples with m number of the attributes, their values and a class C, the algorithm finds a set of predominant attributes set value from the equation (11) A_{best} t for the class concept. It consists of two major parts. In the first part (line 2-7), it calculates the $\mathrm{SU(a,b)}$ value for each number of the attributes , selects relevant attributes into A'_{list} based on the predefined threshold δ, and orders them in descending order according to their $\mathrm{SU(a,b)}$ values. In the second part (line 8-20), it further processes the ordered list A'_{list} to remove redundant attribute values and only keeps predominant ones among all the selected relevant attributes. According to Heuristic 1, a attribute $A_p$that has already been determined to be a predominant attribute can always be used to filter out other attributes that are ranked lower than A_p and have A_p as one of its redundant peers. The iteration starts from the first element (Heuristic 3) in A'_{list} (line 8) and continues as follows. For all the remaining features (from the one right next to A_p to the last one in A'_{list}), if A_p happens to be a redundant peer to a attributes A_q, $A_q$will be removed from A'_{list} (Heuristic 2). After one round of filtering attributes based on A_p, the algorithm will take the currently remaining attribute values for same attribute right next to A_p, as the new reference (line 19) to repeat the filtering process. The algorithm stops until there is no more feature to be removed from A'_{list}. The first part of the above algorithm has a linear time complexity in terms of the number of attributes with best information gain value. As to the second part, in each iteration, using the predominant attributes values in the list A_p identified in the previous round, remove the less information gain value which is not used in the current iteration. The best case could be that all of the remaining attributes following A_p in the ranked list will be removed; the worst case could be none of them.

```
Input: $= \{X_1, \dots . X_N\} be\ the\ number\ of\ samples$ ,
          $A = \{a_1, \dots a_m\}\ no\ of\ attributes , \delta$
Output :$A_{best}$ information gain value
1:Begin
2:for $i = 1\ \ to\ N$ ,For j=1 to m do begin
3: Calculate $\mathrm{SU_{j,c}}$ for $\mathrm{a_i}$
```

```
if $(\mathrm{SU}_{\mathrm{j,c}} \geq \delta \& A_{best})$
Append $a_i\, to\, A'_{list}$
end
Order $A'_{list}$ in descending $\mathrm{SU}_{\mathrm{j,c}}$ value
$\mathrm{A_p} = \mathrm{getfirstelement}(A'_{list})$;
do begin
$\mathrm{A_q} = \mathrm{getnextelement}\left(A'_{list}, \mathrm{A_p}\right)$;
if $(\mathrm{A_q} <> Null)$
do begin
$\mathrm{A'_q} = \mathrm{A_q}$;
if $\left(SU_{p,q} \geq SU_{q,c}\right)$
remove $\mathrm{A_q}$ from $A'_{list}$
$\mathrm{A_q} = \mathrm{getnextelement}\left(A'_{list}, \mathrm{A'_q}\right)$;
else $\mathrm{A_q} = \mathrm{getnextelement}\left(A'_{list}, \mathrm{A_q}\right)$;
end until $\left(A_p == NULL\right)$
$\mathrm{A_p} = \mathrm{getnextelement}\left(A'_{list}, \mathrm{A_p}\right)$;
end until $\left(A_p == NULL\right)$
$A_{best} = A'_{list}$;
End
```

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 then calculate the risk factor of CVD for prediction of the Type 2 diabetes patients ,

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

For each and every attribute values select highest value which is greater A_{best} 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). In the above step we consider all the attributes are features with highest attribute value, but some of the attributes as mentioned above is not useful for prediction of the T2D patients for CVD risk factor . To improve the efficiency of classification algorithms, feature selection is used to identify and remove as much of the irrelevant and redundant information as possible. In the treatment of diabetes, hundreds of attributes are routinely collected but only a small number are used, i.e. the clinicians routinely perform ad-hoc feature selection.

3.4 PARTICLE SWARM OPTIMIZATION FOR FEATURE SELECTION

PSO is a population based optimization tool, which was originally introduced as an optimization technique for real-number spaces. In PSO, each particle is analogous to an individual "fish" in a school of fish. In this work to select most important features in the T2D for CVD risk factor analysis and prediction of the T2D patient CVD risk factors .Particle swarm optimization consists of n number of samples N moving around a D-dimensional search space. The process of PSO is initialized with a population of random number of samples with m number of features for each particles and the algorithm then searches for best optimal type-II diabetes with CVD risk factors. The optimal solutions are found by continuously updating generations. Each T2D features samples (particles) make use of its own memory and knowledge gained by the swarm as a whole to find the best solution. The position of the i^th^ T2D patients records sample features particle can be represented by $\mathrm{T2D} = (\mathrm{T2D_1},, .. \mathrm{T2D_i})$. The velocity for the i^th^ patient record features can be written as $\mathrm{v_i} = (\mathrm{v_{i1}}, \mathrm{v_{i2}}, \ldots, \mathrm{v_{iD}})$. The positions and velocities of the features are confined within $[\mathrm{T2D_{min}}, \mathrm{T2D_{max}}]^{\mathrm{D}}$ and $[\mathrm{V_{min}}, \mathrm{V_{max}}]^{\mathrm{D}}$, respectively. The best previously visited position of the i^th^ features is denoted its individual best position $\mathrm{fp_i} = (\mathrm{fp_{i1}}, \mathrm{fp_{i2}}, \ldots, \mathrm{fp_{iD}})$, a value called $\mathrm{fpbest_i}$. The best value of the all individual $\mathrm{fpbest_i}$ values is denoted the global best position $\mathrm{g} = (\mathrm{g_1}, \mathrm{g_2}, \ldots, \mathrm{g_D})$ and called gbest. At each generation, the position and velocity of the i^th^ each one of the features for T2D with CVD risk factors are updated by $\mathrm{fpbest_i}$ and gbest in the swarm. It occurs the space featuring discrete problem in order to solve this problem Kennedy and Eberhart introduced binary PSO (BPSO), which can be applied to discrete binary variables. In a binary space, a particle features may move to near corners of a hypercube by flipping various numbers of bits; thus, the overall particle velocity may be described by the number of bits changed per iteration. In BPSO, each particle features for T2D patients are updated can be based on the following equations:

$$\begin{aligned} v_{id}^{new} = w \times v_{id}^{new} + c_1 \times r_1 \\ \times \left(\mathrm{fpbest_{id}} - \mathrm{t2d_{id}^{old}}\right) + \mathrm{c_2} \\ \times \mathrm{r_2} \times (\mathrm{fgbest_{id}} - \mathrm{t2d_{id}^{old}}) \end{aligned} \tag{13}$$

If $v_{id}^{new} \notin (V_{min}, V_{max})$ then

$$v_{id}^{new} = \max(\min(V_{max}), v_{id}^{new}), V_{min}) \tag{14}$$

$$S(v_{id}^{new}) = \frac{1}{1 + e^{-v_{id}^{new}}} \tag{15}$$

If $(r_3 < S(v_{id}^{new}))$ then

$$t2d_{id}^{new} = 1 \quad \text{else } t2d_{id}^{new} = 0 \tag{16}$$

In these equations, w is the inertia weight that controls the impact of the previous velocity of a T2D patient records feature particle on its current one, r_1, r_2 and r_3 are random numbers between (0, 1), and c_1 and c_2 are acceleration constants, which control the results of the particles . Velocities v_{id}^{new} and v_{id}^{old} denote the velocities of the new and old feature particle, respectively. v_{id}^{old} is the current particle position, and v_{id}^{new} is the new, updated feature position. In Equation (14) feature (particle) velocities of each dimension are tried to a maximum velocity V_{max} . If the sum of accelerations causes the velocity of that dimension to exceed V_{max}, then the velocity of that dimension is limited to V_{max}. V_{max} and V_{min} are userspecified parameters ($in\ our\ case\ V_{max} = 6, V_{min} = -6$). If $S(v_{id}^{new})$is larger than r_3, then its position value is represented by $\{1\}$ (meaning this position is selected for the next update). If $S(v_{id}^{new})$ is smaller than r_3, then its position value is represented by $\{0\}$.

3.5 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. In normal FCM clustering methods distance measure only evaluates the difference between two individual data points. It ignores the global view of the data distribution. Existing fuzzy c-means based clustering algorithms either considers only hyperspherical clusters in data space. In order to overcome these problems in this work presents a density function measure the similarity or distance measures between the data points. However the density of data points in a cluster could 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 the conventional FCM. It differs from other approaches in that the regulator uses both the shape of the data set and the middle result of iteration operation. And 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 $X = (x_1, \ldots\ldots x_n)$ for every data point x_i, the dot density is usually defined as:

$$z_i = \sum_{j=1, j\neq i}^{n} \frac{1}{d_{ij}} d_{ij} \leq e, 1 \leq i \leq n \tag{17}$$

Where e is the effective radius for density evaluating. Using the cluster density, the distance measure is corrected as Eq. (17)

$$\hat{d}_{ij}^2 = \frac{\left\|x_j - v_i\right\|^2}{\hat{z}_i} \; 1 \leq i \leq c, 1 \leq j \leq n, \tag{18}$$

Thus, the optimization expression can be written as follows base on Eqs. (17):

$$J_{FCM-CD}(U, V, X) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m \left\|x_j - v_i\right\|^2 \frac{\sum_{k=1}^{n} \alpha_{ik} w_{ik}}{\sum_{k=1}^{n} \alpha_{ik} w_{ik}\, z_k} \tag{19}$$

Applying Lagrange Multiplying Method to Eq. (18), can obtain the two update equations given in Eqs. (20) and (21).

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m} \; 1 \leq i \leq c \tag{20}$$

$$u_{ij} = \frac{\hat{d}_{ij}^{-2/(m-1)}}{\sum_{j=1}^{n} \hat{d}_{kj}^{-2/(m-1)}} \tag{21}$$

3.6 EXTREME LEARNING MACHINE FOR CLASSIFICATION

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 KLD-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 extreme learning machine for prediction task of type 2 diabetes patient with CVD risks. In work of [21], it was found that ELM can provide a unified solution for a generalized Single Hidden Layer Neural Network(SLFN). In order to perform the prediction for the T2D with CVD risk factors the relationship between the unsupervised feature selected learning data $UFSd_i$ and its corresponding relevance degree prediction of T2D patients $UFSOd_i$ that is approximated by the hypothesis $f\ (UFSd_i\ )$. The aim of pair wise ELM is obviously to search for a hypothesis f () such that $f\ (UFSd_{\boldsymbol{u}}^{\boldsymbol{i}}) > f\ (UFSd_{\boldsymbol{v}}^{\boldsymbol{i}})$ if .Within the framework and traditional least square loss function provided by ELM, the problem of minimizing the training error $\xi_{u,v}^i$ and the norm of output weight can be mathematically written as

$$minimize : \tfrac{1}{2}\left\|\beta\right\|^2 + C\tfrac{1}{2}\textstyle\sum_{i=1}^{n}\sum \|\, \xi_{u,v}^i \|^2$$
(22)

$$subject\ to : \left(h(UFSd_{\boldsymbol{u}}^{\boldsymbol{i}}) - h(UFSd_{\boldsymbol{v}}^{\boldsymbol{i}})\right)\beta = (UFSOd_{\boldsymbol{u}}^{\boldsymbol{i}} - UFSOd_{\boldsymbol{v}}^{\boldsymbol{i}}) - \xi_{u,v}^i, i = 1, .. n \tag{23}$$

Inspired by the work of [22], graph theory helps to store prediction classification data in an $N \times N$ symmetric matrix $\boldsymbol{W}$

$$W = \sum_{i=1}^{n} m(i) \tag{24}$$

W is one type of adjacency graph, where $W_{ij} = 1$ if the two different unsupervised feature selected learning data $UFSd_i$and $UFSd_j$ are connected to the same class and $W_{ij} = 0$ otherwise. All the diagonal entries are set to 1 in W. The corresponding Laplacian matrix **L** is then defined as $\mathrm{L} = \mathrm{D} - \mathrm{W}$, where

$$D = diag\left\{\sum_{j=1}^{N} W_{1,j} \ldots \sum_{j=1}^{N} W_{N,j}\right\} \qquad (25)$$

Therefore, with the incorporation of graph W, (10) can be rewritten in a way that is more similar to traditional regression/classification form. Hence, only the pair wise relationship between samples associated with the same class maintains

$$Min: \frac{1}{2}\|\beta\|^2 + C\frac{1}{2}W\sum_{i,j=1}^{N}\left\|\left(UFSOd_i - UFSOd_j\right) - \left(f(UFSd_i) - f\left(UFSd_j\right)\right)\right\|^2 \qquad ($$

(26)

The solution of (26) can be obtained by setting the derivative to zero

$$\beta = \left(\frac{I}{C} + H^T LH\right)^{-1} H^T LT \qquad (27)$$

Similar to [23], the kernel version for ranking problems is not difficult to be derived as

$$f(UFSd)_{kernel} = \begin{bmatrix} K(UFSd, UFSd_1) \\ . \\ . \\ K(UFSd, UFSd_N) \end{bmatrix}\left(\frac{I}{C} + L\Omega_{elm}\right)^{-1} LT \qquad ($$

(28)

4. 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. 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 [24] 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. (29), specificity Eq. (30), sensitivity Eq. (31)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (29)$$

$$Sensitivity = \frac{TP}{TP + FN} \qquad (30)$$

$$Specificity = \frac{TN}{TN + FP} \qquad (31)$$

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. 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 1: Prediction methods results

Parameters	K-C4.5	IFCM-SVM	PSO-IFCM-ELM
Accuracy	92.3	93.8	94.5
Sensitivity	89.4	90.49	92.5
Specificity	60.8	54.7	52.1

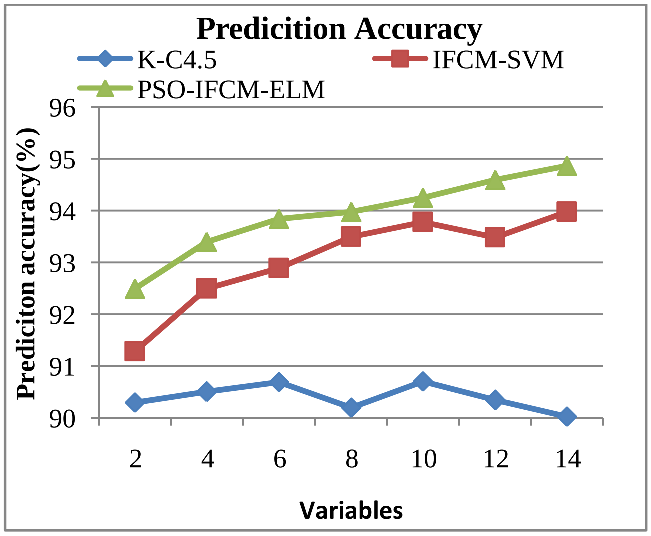

Figure 2: Prediction accuracy of the prediction methods

Prediction accuracy of the proposed PSO-IFCM-KLM based prediction methods achieves higher classification accuracy than the existing classification methods IFCM-SVM, K-C4.5 prediction accuracy is illustrated in Figure 2 , since the proposed methods selects the important features in the T2D with CVD risk factors after the completion of the preprocessing and dimensionality reduction KLD-PCA. In proposed work perform Fast Correlation-Based Filter

Solution (FCBFS) similarity measurement when compare to existing similarity measurement method.

Sensitivity accuracy of the proposed PSO-IFCM-KLM based prediction methods achieves higher Sensitivity than the existing classification methods IFCM-SVM and K-C4.5 .Sensitivity is illustrated in Figure 3, Sensitivity result of the proposed PSO-IFCM-KLM system are high because of the feature selection (PSO) is performed after the completion of the preprocessing and dimensionality reduction methods and Fast Correlation-Based Filter Solution (FCBFS) similarity measurement is performed to improve prediction accuracy.

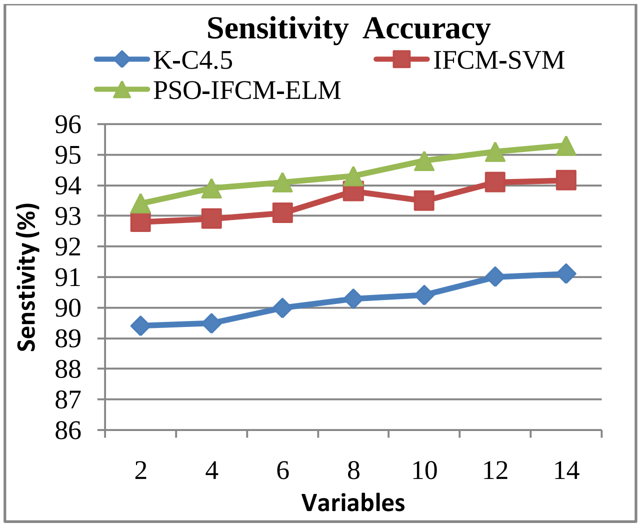

Figure 3: Sensitivity accuracy of the prediction methods

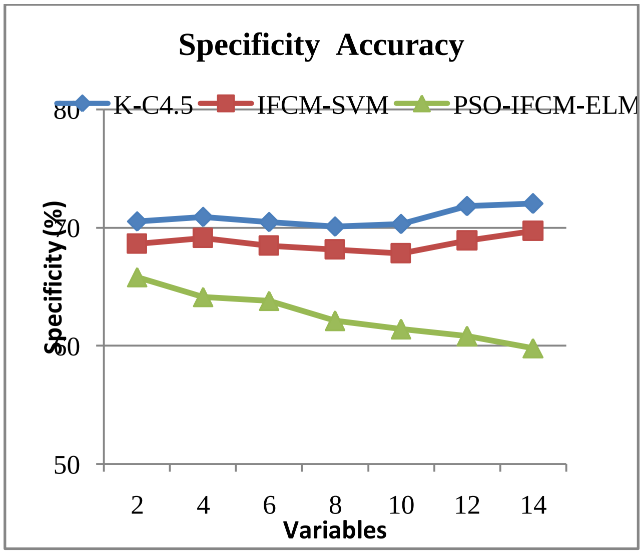

Figure 4: Specificity accuracy of the prediction methods

Specificity accuracy of the proposed PSO-IFCM-KLM based prediction methods achieves lesser specificity than the existing classification methods IFCM-SVM and K-C4.5 prediction specificity is illustrated in Figure 4 , the specificity of the proposed system are less feature selection (PSO) is performed after the completion of the preprocessing and dimensionality reduction methods and Fast Correlation-Based Filter Solution (FCBFS) similarity measurement is performed to improve prediction accuracy.

5. CONCLUSION

Type 2 diabetes confers a high degree of cardiovascular risk brought about by multiplicative risk factors. Analysis of CVD risk factors plays major importance to predict T2D patient's results. In type diabetes with CVD risk factors finding the most important features becomes major difficult task and thus reduces the prediction accuracy .In order to overcome these problem in this work propose a particle swarm optimization based feature selection. In the proposed method classifiers were able to achieve their best performance when the most important features were selected. Assessment of the impact on diabetes treatment and complications has been made for only one prediction model. In this work presents an efficient prediction model to analysis the risk of CVD factors in T2D patients records ,to analysis the patients records initially the records are preprocessed and reduced dimension of the features using KLD-PCA, risk factors are calculated using FCBFS filtering methods, features were selected using PSO ,then Improved Fuzzy C Means (IFCM) clustering algorithm is proposed for unsupervised learning, finally Extreme learning machine (ELM) is discussed to perform prediction of T2D with CVD risk factors. The experimental results verify PSO for ELM, SVM and C4,5 classifiers in the diabetic domain. PSO preserved the accuracy of the full set classifier, while using the selected available variables and significantly improved the computational efficiency of three important classification algorithms.

References

[1]. Dr Alan Rees, "Excess cardiovascular risk in patients with type 2 diabetes: do we need to look beyond LDL cholesterol?", Br J Diabetes Vasc Dis 2014;14:10-20.

[2]. Engelgau MM, Geiss LS, Saaddine JB, Boyle JP, Benjamin SM, Gregg EW, Tierney EF, Rios-Burrows N, Mokdad AH, Ford ES, Imperatore G, Narayan KM: The evolving diabetes burden in the United States. Ann Intern Med 140: 945–950, 2004.

[3]. Hu FB, Stampfer MJ, Solomon CG, Liu S, Willett WC, Speizer FE, Nathan DM, Manson JE: The impact of diabetes mellitus on mortality from all causes and coronary heart disease in women: 20 years of follow-up. Arch Intern Med 161: 1717–1723, 2001

[4]. Fox CS, Coady S, Sorlie PD, Levy D, Meigs JB, D'Agostino RB Sr, Wilson PW, Savage PJ: Trends in cardiovascular complications of diabetes. JAMA 292: 2495–2499, 2004.

[5]. Mokdad AH, Ford ES, Bowman BA, Dietz WH, Vinicor F, Bales VS, Marks JS: Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. JAMA 289:76–79, 2003

[6]. Thomas F, Bean K, Pannier B, Oppert JM, Guize L, Benetos A: Cardiovascular mortality in overweight subjects: the key role of associated risk factors. Hypertension 46:654–659, 2005.

[7]. Brunner EJ, Shipley MJ, Witte DR, Fuller JH, Marmot MG: Relation between blood glucose and coronary mortality over 33 years in the Whitehall Study. Diabetes Care 29:26–31, 2006.

[8]. Yusuf S, Hawken S, Ounpuu S, Bautista L, Franzosi MG, Commerford P, Lang CC, Rumboldt Z, Onen CL, Lisheng L, Tanomsup S, Wangai P Jr, Razak F, Sharma AM, Anand SS, the INTERHEART Study Investigators: Obesity and the risk of myocardial infarction in 27,000 participants from 52 countries: a case control study. Lancet 366:1640–1649, 2005.

[9]. Guthrie RA, Guthrie DW, editors. Nursing management of diabetes mellitus. 5th ed., New York: Springer Publishing; 2002.

[10]. Assmann G, Cullen P, Schulte H: Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study. Circulation 105:310–315, 2002.

[11]. Eberly LE, Prineas R, Cohen JD, Vazquez G, Zhi X, Neaton JD, Kuller LH, the Multiple Risk Factor Intervention Trial Research Group: Metabolic syndrome: risk factor distribution and 18-year mortality in the Multiple Risk Factor Intervention Trial. Diabetes Care 29:123–130, 2006

[12]. Wilson PW, D'Agostino RB, Parise H, Sullivan L, Meigs JB: Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus. Circulation 112:3066–3072, 2005.

[13]. Crone S, Lessmann S, Stahlbock R. Empirical comparison and evaluation of classifier performance for data mining in customer relationship management. In: Wunsch D, et al., editors. Proceedings of the international joint conference on neural networks, IJCNN'04. 2004. p. 443—8.

[14]. Kantardzic M, editor. Data mining: concepts, models, methods, and algorithms. New Jersey: Wiley-IEEE Press; 2002.

[15]. Su CT, Yang CH, Hsu KH, Chiu WK. Data mining for the diagnosis for type II diabetes from three-dimensional body surface anthropometrical scanning data. Comput Math Appl 2006;51:1075—92.

[16]. Huang Y, McCullagh PJ, Black ND. Feature selection via supervised model construction. In: Bramer M, editor. Proceedings of the 4th IEEE international conference on data mining. 2004. p. 411—4.

[17]. Robnik M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Mach Learn 2003;53:23—69.

[18]. Myatt, G. J. (2007). Making sense of data a practical guide to exploratory data analysis and data mining. New Jersey: John Wiley & Sons.

[19]. Han, J., & Kamber, M. (2006). Data mining: Concepts and techniques (2nd ed.). Morgan Kaufmann Publishers.

[20]. 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

[21]. Huang GB, Zhou H, Ding X, Zhang R (2012) Extreme learning machine for regression and multi-class classification. IEEE Trans Syst Man Cybern 42(2):513–529

[22]. K. Crammer, and Y. Singer. On the learn ability and design of output codes for multiclass problems. *Machine Learning*. 2002, 47 (2-3): 201-233

[23]. Pahikkala T, Tsivtsivadze E, Airola A, Boberg J, Salakoski T (2007) Learning to rank with pairwise regularized least-squares. In: Joachims T, Li H, Liu TY, Zhai C (eds) Proceedings of the SIGIR 2007workshop on learning to rank for information retrieval. ACM, Amsterdam, Netherlands, pp 27–33.

[24]. 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.

[25]. P. Radha, Dr. B. Srinivasan : Diagnosing

a. Heart Diseases for Type 2 Diabetic Patients by
b. cascading the Data Mining Techniques.

[26]. International Journal on Recent and Innovation

[27]. Trends in Computing and Communication

[28]. (IJRITCC) vol 2, issue 8, Aug 2014, 2503-2509.