International Journal of Electrical, Electronics and Computer Engineering **6**(1): 164-172(2017)

ISSN No. (Online): 2277-2626

Prediction of Diabetes Disease Using Entropy and Gain based Data Mining Approach

Asma Aziz Khan* and Vipin Verma**

**Department of Computer Science & Engineering, ASCT, Bhopal, (Madhya Pradesh), INDIA*

***Department of Computer Science & Engineering, ASCT, Bhopal, (Madhya Pradesh), INDIA*

(Corresponding author: Asma Aziz Khan)

(Received 10 April, 2017 Accepted 28 May, 2017)

(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: The number of diabetic patient is growing day by day and various causes has been discover to diagnosis the diabetes but early prediction of the cause is very indispensable to get rid of from this disease. Data Mining is a technique which plays an imperative role by unhidden the important data related to diabetes. These data can be utilized for rapid and improved clinical decision making for protective and suggestive medicine. This paper proposes a hybrid approach SVM classifier using KNN and GA technique which can effectively discover the effectual data to diagnose the diabetes disease. The simulation and experimental analysis of the propose approach is done using MATLAB toolbox and measuring parameter specificity, accuracy and sensitivity. The simulation results of proposed approach give improved value than the existing approach.

Keywords: Diabetes, Data Mining, MATLAB, SVM, KNN –GA

I. INTRODUCTION

DIABETES mellitus is a metabolic disease, portrayed by nearness of hyperglycemia coming about because of flawed insulin discharge or its handling in the body. Basically, it comes about because of bodies that don't have enough insulin to breakdown glucose (repercussion of starch), or bodies impervious to the impacts of insulin. Glucose, as a principle wellspring of energyfor cells that makes up the muscles and different tissues, is created from the nourishment we eat and in our liver. Sugar (or glucose) is caught up in the circulation system and goes into a cell by the assistance of insulin. The liver stores glucose as glycogen so that, if glucose turns out to be low, the liver reconverts the put away glycogen into glucose to standardize the glucose level [1]. Diabetes is analysis from glycemia related with microvascular infection [2]. Data mining procedure can be to a great degree valuable for Medical specialists for removing concealed medicinal knowledge. It would some way or another be inconceivable for customary example coordinating and mapping procedures to be so viable and exact in anticipation or conclusion without utilization of information mining methods. This work goes for connecting different diabetes input parameters for effective order of Diabetes dataset and ahead to mining valuable examples. Information revelation and information mining have found various applications in business and logical space. Profitable learning can be found from utilization of information mining strategies in human services frameworks excessively [3]. Information preprocessing and change is required before one can apply information mining to clinical information. Knowledge discovery and data mining is the center stride, which brings about disclosure of concealed however helpful learning from huge databases [3]. This paper propose a hybrid approach SVM using KNN and GA data mining technique to unhidden the useful information from the massive dataset of diabetic patient and its experimental work is perform in MATLAB simulation toolbox using performance measuring parameter specificity, sensitivity and accuracy.

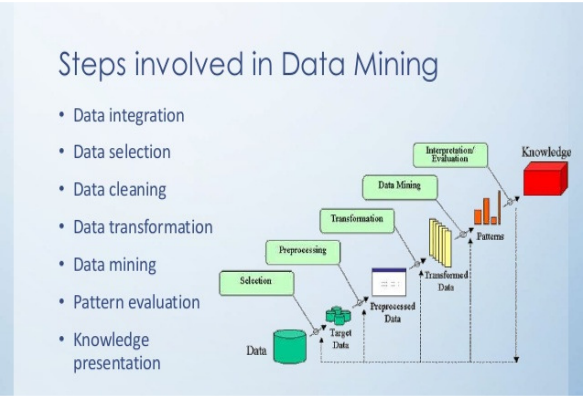

Fig. 1. Various steps involved in the process of data mining.

The data mining involves various stages to process the data which is shown in Fig.1.

II. CLASSIFICATION AND DIAGNOSTIC CRITERIA FOR DIABETES

The Diabetes is classified into categories and follows the various criteria to diagnosis the diabetes.[4]

A. Classification Diabetes

Earlier classifications. The first widely accepted classification of diabetes mellitus was published by WHO in 1980 (1) and, in modified form, in 1985 (3). The 1980 and 1985 classifications of diabetes mellitus and allied categories of glucose intolerance included clinical classes and two statistical risk classes. The 1980 Expert Committee proposed two major classes of diabetes mellitus and named them, IDDM or Type 1, and NIDDM or Type 2. In the 1985 Study Group Report the terms Type 1 and Type 2 were omitted, but the classes IDDM and NIDDM were retained, and a class of Malnutrition-related Diabetes Mellitus (MRDM) was introduced. In both the 1980 and 1985 reports other classes of diabetes included Other Types and Impaired Glucose Tolerance (IGT) as well as Gestational Diabetes Mellitus (GDM). These were reflected in the subsequent International Nomenclature of Diseases (IND) in 1991, and the tenth revision of the International Classification of Diseases (ICD-10) in 1992. The 1985 classification was widely accepted and is used internationally. It represented a compromise between clinical and aetiological classification and allowed classification of individual subjects and patients in a clinically useful manner even when the specific cause or aetiology was unknown. The recommended classification includes both staging of diabetes mellitus based on clinical descriptive criteria and a complementary aetiological classification.

Revised classification. The classification encompasses both clinical stages and aetiological types of diabetes mellitus and other categories of hyperglycaemia, as suggested by Kuzuya and Matsuda (15). The clinical staging reflects that diabetes, regardless of its aetiology, progresses through several clinical stages during its natural history. Moreover, individual subjects may move from stage to stage in either direction. Persons who have, or who are developing, diabetes mellitus can be categorized by stage according to the clinical characteristics, even in the absence of information concerning the underlying aetiology. The classification by aetiological type results from improved understanding of the causes of diabetes mellitus.

Application of the new classification. The new classification contains stages which reflect the various degrees of hyperglycaemia in individual subjects with any of the disease processes which may lead to diabetes mellitus.

All subjects with diabetes mellitus can be categorized according to clinical stage, and this is achievable in all circumstances. The stage of glycaemia may change over time depending on the extent of the underlying disease processes (Figure 2). The disease process may be present but may not have progressed far enough to cause hyperglycaemia. The aetiological classification reflects the fact that the defect or process which may lead to diabetes may be identifiable at any stage in the development of diabetes - even at the stage of normoglycaemia. Thus the presence of islet cell antibodies in a normoglycaemic individual makes it likely that that person has the Type 1 autoimmune process. Unfortunately there are few sensitive or highly specific indicators of the Type 2 process at present, although these are likely to be revealed as aetiology is more clearly defined. The same disease processes can cause impaired fasting glycaemia and/or impaired glucose tolerance without fulfilling the criteria for the diagnosis of diabetes mellitus. In some individuals with diabetes, adequate glycaemic control can be achieved with weight reduction, exercise and/or oral agents. These individuals, therefore, do not require insulin and may even revert to IGT or normoglycaemia. Other individuals require insulin for adequate glycaemic control but can survive without it. These individuals, by definition, have some residual insulin secretion. Individuals with extensive beta-cell destruction, and therefore no residual insulin secretion, require insulin for survival. The severity of the metabolic abnormality can either regress (e.g. with weight reduction), progress (e.g. with weight gain), or stay the same.

B. Diagnostic Criteria

The clinical diagnosis of diabetes is often prompted by symptoms such as increased thirst and urine volume, recurrent infections, unexplained weight loss and, in severe cases, drowsiness and coma; high levels of glycosuria are usually present. A single blood glucose estimation in excess of the diagnostic values indicated in black zone establishes the diagnosis in such cases. Figure 1 also defines levels of blood glucose below which a diagnosis of diabetes is unlikely in non-pregnant individuals. These criteria are as in the 1985 report (3). For clinical purposes, an OGTT to establish diagnostic status need only be considered if casual blood glucose values lie in the uncertain range (i.e. between the levels that establish or exclude diabetes) and fasting blood glucose levels are below those which establish the diagnosis of diabetes. If an OGTT is performed, it is sufficient to measure the blood glucose values while fasting and at 2 hours after a 75 g oral glucose load. For children the oral glucose load is related to body weight: 1.75 g per kg.

Table 1: Values for diagnosis of diabetes mellitus and other categories of hyperglycaemia.

Glucose concentration, mmol l^{-1} (mg dl^{-1})			
	Whole blood	Whole blood	Plasma*
	Venous	Capillary	Venous
Diabetes Mellitus			
Fasting	>=6.1 (>=110)	>=6.1 (>=110)	>=7.0 (>=126)
or			
2-h post glucose load	>=10.0 (>=180)	>=11.1 (>=200)	>=11.1 (>=200)
or both			
Impaired Glucose Tolerance (IGT)			
Fasting (if measured)	<6.1 (<110)	<6.1 (<110)	<7.0 (<126)
and			
2-h post glucose load	>=6.7 (>=120) and	>=7.8 (>=140) and	>=7.8 (>=140) and
	<10.0 (<180)	<11.1 (<200)	<11.1 (<200)
Impaired Fasting Glycaemia (IFG)			
Fasting	>=5.6 (>=100) and	>=5.6 (>=100) and	>=6.1 (>=110) and
	<6.1 (<110)	<6.1 (<110)	<7.0 (<126)
and (if measured)			
2-h post glucose load	<6.7 (<120)	<7.8 (<140)	<7.8 (<140)

The diagnostic criteria in children are the same as for adults. Diagnostic interpretations of the fasting and 2-h post-load concentrations in non-pregnant subjects are shown in Table 1.

Change in diagnostic value for fasting plasma/blood glucose concentrations. The major change recommended in the diagnostic criteria for diabetes mellitus is the lowering of the diagnostic value of the fasting plasma glucose concentration to 7.0 mmol l^{-1} (126 mg dl^{-1}) and above (3), from the former level of 7.8 mmol l^{-1} (140 mg dl-1) and above. For whole blood the proposed new level is 6.1 mmol l^{-1} (110 mg dl^{-1}) and above, from the former 6.7 mmol l^{-1} (120 mg dl^{-1}). The new fasting criterion is chosen to represent a value which is at the upper end of the range that corresponds in diagnostic significance in many persons to that of the 2-h post-load concentration, which is not changed. This equivalence has been established from several population-based studies (6-8) and it also represents an optimal cut-off point to separate the components of bimodal frequency distributions of fasting plasma glucose concentrations seen in several populations. Furthermore, several studies have shown increased risk of microvascular disease in persons with fasting plasma glucose concentrations of 7.0 mmol l^{-1} (126 mg dl^{-1}) and over (6), and of macrovascular disease in persons with such fasting concentrations, even in those with 2-h values of < 7.8 mmol l^{-1} (140 mg dl^{-1}) (9).

Nevertheless, in less obese subjects, in some ethnic groups and in the elderly lower fasting glucose levels may be seen in persons who have 2-h post-load glucose values that are diagnostic for diabetes.

Epidemiological studies. For population studies of glucose intolerance and diabetes, individuals have been classified by their blood glucose concentration measured after an overnight fast and/or 2 h after a 75 g oral glucose load. Since, it may be difficult to be sure of the fasting state, and because of the strong correlation between fasting and 2-h values, epidemiological studies or diagnostic screening have in the past been restricted to the 2-h values only (Table 1). Whilst this remains the single best choice, if it is not possible to perform the OGTT (e.g. for logistical or economic reasons), the fasting plasma glucose alone may be used for epidemiological purposes. It has now been clearly shown, however, that some of the individuals identified by the new fasting values differ from those identified by 2-h post glucose challenge values (10,11). The latter include the elderly (12) and those with less obesity, such as many Asian populations. On the other hand, middle-aged more obese patients are more likely to have diagnostic fasting values (10). Overall population prevalence may (13) or may not (7,10,14) be found to differ when estimates using fasting and 2-h values are compared.

Individual diagnosis. The requirements for individual diagnosis differ from those of population studies. The diagnosis should not be based on a single glucose determination but requires confirmatory symptoms or blood/plasma determination. Diagnosis requires the identification of people at risk for development of complications in whom early preventive strategies are indicated. Ideally therefore both the 2-h and the fasting value should be used. These recommendations contrast with those of the ADA Expert Committee which gives primacy to the fasting plasma glucose (4).

III. DATA MINING TECHNIQUES FOR DIAGNOSIS OF DIABETES

Diabetes Mellitus has become a common health dilemma nowadays, which would distress people and lead to different complications like visual impairment, cardio vascular disease, leg amputation and renal failure if diagnosis is not done in the right time. In this discussed the two classifier techniques with principal component analysis component analysis are implemented for the forecasting of Diabetes and concluded with best forecasting techniques which has a maximum accuracy. These are given below:

A. Decision Trees

Decision tree [5] is a tree structure, which is in the form of a flowchart. It is used as a method for classification and prediction with representation using nodes and internodes. The root and internal nodes are the test cases that are used to separate the instances with different features. Internal nodes themselves are the result of attribute test cases. Leaf nodes denote the class variable. Figure1. shows a sample decision tree structure.

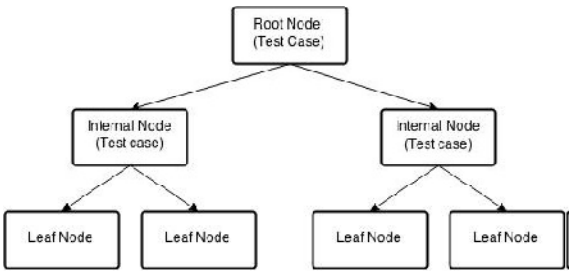

Fig. 2. Sample Decision Tree Structure.

Decision tree provides a powerful technique for classification and prediction in Diabetes diagnosis problem. Various decision tree algorithms are available to classify the data, including ID3, C4.5, C5, J48, CART and CHAID. In this paper, J48 decision tree algorithm [6] has been chosen to establish the model. Each node for the decision tree is found by calculating the highest information gain for all attributes and if a specific attribute gives an unambiguous end product (explicit classification of class attribute), the branch of this attribute is terminated and target value is assigned to it.

B. Naïve Bayes

The Naïve Bayes Algorithm is a probabilistic algorithm that is sequential in nature, following steps of execution, classification, estimation and prediction. For finding relations between the diseases, symptoms and medications, there are various data mining existing solution, but these algorithms have their own limitations; numerous iterations, binning of the continuous arguments, high computational time, etc. Naïve Bayes overcomes various limitations including omission of complex iterative estimations of the parameter and can be applied on a large dataset in real time. The algorithm works on the simple Naïve Bayes formula given below.

$$Posterior\ Probability\ P(C|X) = \frac{Likelihood\ P(X|C) \times Class\ Prior\ probability\ P(C)}{Predictor\ Prior\ Probability\ P(X)}$$

C. Principal Component Analysis

Principal component analysis (PCA) is a standard tool in modern data analysis. It is a simple non parametric method for extracting relevant information from confusing data sets.

Principal components analysis method is used for achieving the simplification and generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information. The principal components as a whole form an orthogonal basis for the space of the data. The procedure can be followed in many ways i.e. a) Using singular value decomposition method (SVD) b) using the covariance matrix method. In this work we have used MATLAB software for deriving the principal components [7].

D. Support Vector Machine

SVM is a set of related supervised learning method used in medical diagnosis for classification and regression [11,12]. SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM is called Maximum Margin Classifiers. SVM is a general algorithm based on guaranteed risk bounds of statistical learning theory i.e. the so called structural risk minimization principle. SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. The kernel trick allows constructing the classifier without explicitly knowing the feature space. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible [11, 13]. For example, given a set of points belonging to either one of the two classes, an SVM finds a hyperplane having the largest possible fraction of points of the same class on the same plane. This separating hyperplane is called the optimal separating hyperplane (OSH) that maximizes the distance between the two parallel hyper planes and can minimize the risk of misclassifying examples of the test dataset. Given labeled training data as data points of the form:

$$M = \{(x_1, y_1), (x_2, y_2), (x_3, 3) \ldots\ldots\ldots (x_n, n)$$

where y_n=1/-1, a constant that denotes the class to which that point x_n belongs. n=number of data sample. Each is a p-dimensional real vector. The SVM classifier first maps the input vectors into a decision value, and then performs the classification using an appropriate threshold value. To view the training data, we divide (or separate) the hyperplane, which can be described as:

Mapping: $W^T.x + b = 0$

where w is a p-dimensional weight vector and b is a scalar. The vector w points perpendicular to the separating hyperplane. The offset parameter b allows to increase the margin. When the training data are linearly separable, we select these hyperplanes so that there are no points between them and then try on maximizing the distance between the hyperplane. We have found out the distance between the hyperplane as 2/ |w|. To minimize |w|, we need to ensure that for all either

$$w.x_i - b \geq 1 \; or \; w.x_i - b \leq -1$$

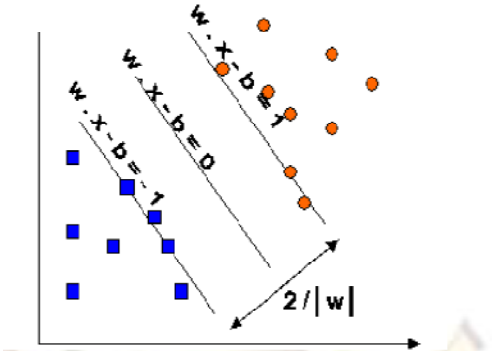

Fig. 3. Maximum margin hyperplane for SVM traned with sample from two classes.

E. K-Nearest Neighbour (KNN)

The idea in k-Nearest Neighbor methods [8] is to identify ‘k‘ samples in the training set whose independent variables 'm' are similar to 'n', and to use these 'k' samples to classify this new sample into a class, v. Assume that _f' is a smooth function, an idea is to look for samples in our training data that are near it and then to compute _v‘ from the values of ‘y‘ for these samples. When we talk about neighbors we are implying that there is a distance or dissimilarity measure that we can compute between samples based on the independent variables. For the moment we will concern ourselves to the most popular measure of distance is say, Euclidean distance. The initial training stage for kNN [9] consists of storing all known occurrences and their class labels. Either a tabular representation or a specialized design such as a kd-tree can be used. If we want to adjust the value of 'k', an alternate method of n-fold cross-validation on the training data set can be used. The k-NN algorithm for continuous-valued functions

- Calculate the mean value of the k nearest neighbors Distance-weighted nearest neighbor algorithm
- Weight the contribution of k neighbors according to their distance to the query point x_q
  – giving greater weight to closer neighbors

$$W = \frac{1}{d(x_q, x_i)^2}$$

  – Similarly, for real-valued target functions

F. Genetic Algorithm

Genetic algorithm [10] is a subset of evolutionary algorithm developed from Darwin's theory of gradual evolution and fundamental ideas. Process of optimization is based on a random trend in genetic algorithm. Before the genetic algorithm can be implemented, we must first find encoding system for the intended problem. The most common way to show chromosomes in the genetic algorithms is in binary form. In this case, chromosome is a bit string, the length of which is determined by some existing parameters. In other words, each parameter is related to a bit in a string. In this algorithm, for a fixed number called population, a set of target parameters is produced randomly. The genetic algorithm applies the rule of surviving the best to get the better solutions and then it assigns the number representing the fitting of that set to the member of the population. This process is repeated for every single member. With the retrieval of genetic algorithm operators such as selection, Mutation and crossover imitated from natural genetics, better approximations can be obtained from final solution and this procedure continue to get the convergence criterion. A selection operator chooses some chromosomes among the available chromosomes in a population for reproduction. The methods of selection are selection of the elite, the roulette wheel, tournament, Boltzmann, ranking, etc. The crossover operator is a random merging which some parts of chromosomes are exchanged. This issue causes that the children are not exactly like their parents and have had a combination of characteristics of their parents. After the merging, Mutation operator is applied on chromosome. This operator chooses a gene from a chromosome randomly and changes the content of that gene. There are three criteria for algorithmtermination:1-The number of generations in algorithm,2-the population does not become better,3-classification accuracy of element with the best fitness does not exceed the threshold level.

IV. PROPOSED METHODOLOGY

In this, we propose an ensemble approach SVM[11,12], KNN [9] and GA [10] algorithm for the diagnosis of diabetes in Pima Indian women dataset. Initially, apply learning algorithm (SVM) to reduce the dataset from database then later trained those selected dataset and prefer only those dataset which generated frequently. Again apply the KNN-GA algorithm on the selected dataset to improve the accuracy level of classified or misclassified dataset. After that apply crossover and mutation function for the generation of new population and calculate the local maxima of the each categories of dataset. In this we have calculated the entropy and information gain of the Pima Indian Women dataset.

A. Entropy

Entropy is a quantity which is used to describe the `business' of an image, i.e. the amount of information which must be coded for by a compression algorithm. Low entropy images, such as those containing a lot of black sky, have very little contrast and large runs of pixels with the same or similar DN values. An image that is perfectly flat will have entropy of zero. Consequently, they can be compressed to a relatively small size. On the other hand, high entropy images such as an image of heavily cratered areas on the moon have a great deal of contrast from one pixel to the next and consequently cannot be compressed as much as low entropy images.

$$\boldsymbol{Entropy = \sum_{i} -P_i log_2 P_i}$$

B. Information gain

The information gain for an attribute is defined as follows: The information gain is equal to the total entropy for an attribute if for each of the attribute values a unique classification can be made for the result attribute. In this case the relative entropies subtracted from the total entropy are 0.

$$\textbf{Entropy H(X)} = -p_1\log_2 p_1 - p_2\log_2 p_2 - p_3\log_2 p_3 - \ldots\ldots - p_n\log_2 p_n \quad \ldots\ldots\ldots\ldots eq(1)$$

$$= -\sum_{i=1}^{m} pi\ log2\ pi$$

In the equation 1, the class-wise probability has been settled then entropy has been calculated of each individual attributes.

Then gain was calculated as follows:

$$\textbf{Gain = Entropy(X) - Entropy (X|Y)} \quad \ldots\ldots\ldots\ldots \textbf{eq (2)}$$

So as per the above process feature reduction has been done, where gain was higher than that attribute has been qualified for the process and less gain was reduced from dataset.

C. Algorithm Steps

Step 1: Make X1 reduced datasets from a database.

Step 2: Set a learning algorithm to individual pattern for test dataset.

Step 3: Set a learning algorithm to individual pattern training dataset.

svmStruct = svmtrain(X1(train(:,i1),:),groups(train(:,1))

Step 4: Object with unknown found to do with each of the X1 classifiers predictions.

Step 5: Select the most repeatedly predicted samples.

KNNGA steps:

Step1: Initialize population = *X1*

Step2: Apply genetic search into selected dataset

Step3: Apply KNN classifier for testing of all five data which is classified or misclassified data.

Step4: Each attribute will organize as their ranks.
Step5: Higher ranked attribute will select.
Step6: Apply KNNGA () on the each five subset of the attributes for enhance the accuracy level.

Step7: If
knnga_classifier(class_knn)>knn_classifer(class_knn)
data_class = class_knn;
else
data_class = class_knnga;

Step8: Perform the reproduction
Step9: Apply crossover operator
Step10: Perform mutation then produce new population *X'1*
Step11: Calculate the local maxima for each category.
Repeat the steps till iteration is not finished
Step12: For each test X'1, start all trained base models then prediction of result by combining of all trained models, and separate the misclassified by optimized knnGA.
Classification: goal wise classification result obtained

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we present the results from our extensive experiments to compare the performance of SVM and proposed method SVM-KNN and GA on real health care data of Pima Indian Women [14, 15]. All the experiment are conducted on the MATLAB platform, which includes three Intel 3.4 GHz machines, each running 16GB RAM.

A. Pima Indian Women Dataset

A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. We used the 532 complete records after dropping the (mainly missing) data on serum insulin. In this work we usage the Pima.tr
Pima.tr2, Pima.te. These data frames contain the following columns:
Npreg number of pregnancies.
glu plasma glucose concentration in an oral
glucose tolerance test.
Bp diastolic blood pressure (mm Hg).
Skin triceps skin fold thickness (mm).
bmi body mass index (weight in kg/(height in m)\^2).
ped diabetes pedigree function.
ageage in years.
type Yes or No, for diabetic according to WHO criteria.
The training set Pima.tr contains a randomly selected set of 200 subjects, and Pima.te contains the remaining 332 subjects. Pima.tr2 contains Pima.tr plus 100 subjects with missing values in the explanatory variables.

B. GUI Environment

This section shows the main GUI environment of the proposed methodology and entropy and information gain process of it.

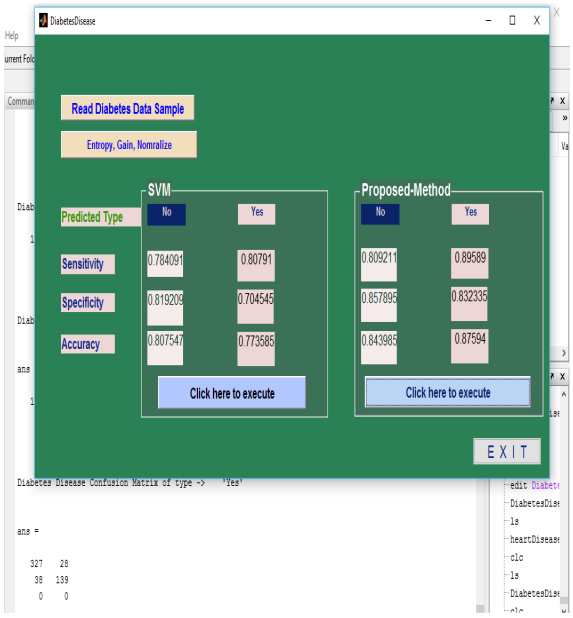

Fig. 4. Main GUI of the Proposed Methodology

C. Result Analysis

The comparative analysis of the proposed work is performed using the accuracy, sensitivity and specificity parameter and for the processed attributes confusion matrix is formed which is shown below:

Table 2: Confusion Matrix for Existing Method SVM.

Diabetes Disease Confusion Matrix of type -> 'No'	
64	42
24	135
0	0
Diabetes Disease Confusion Matrix of type -> 'Yes'	
150	29
27	59
0	0

Table 3: Confusion Matrix for Proposed Method.

Diabetes Disease Confusion Matrix of type -> 'No'	
133	44
34	321
0	0
Diabetes Disease Confusion Matrix of type -> 'Yes'	
309	46
44	133
0	0

Accuracy Analysis. The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

$$\text{Accuracy} = (TP+TN)/TP+TN+FP+FN$$

For this parameter the comparison between SVM and proposed method is perform in which it is found that the accuracy rate of SVM is about 78% and our method is about 86% which means our method generates better accuracy rate than the existing SVM method.

Table 4: Accuracy result analysis of the proposed 2tier method and SVM Method.

Accuracy		
Method/Type	SVM	Proposed
No	0.807547	0.843985
Yes	0.773585	0.87594

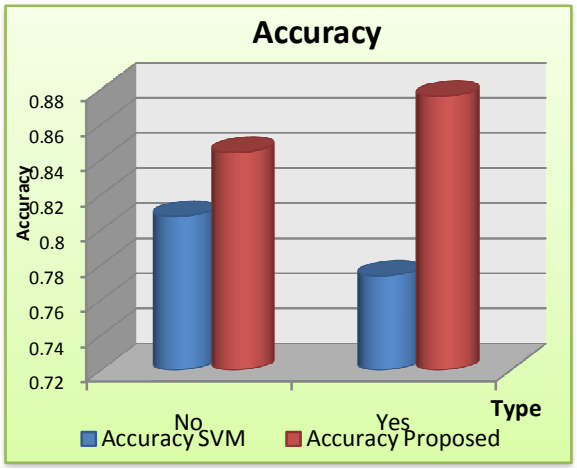

Fig. 5. Accuracy graph between SVM and proposed Method.

Specificity Analysis. The sensitivity of a test is its ability to determine the patient cases correctly. To estimate it, we should calculate the proportion of true positive in patient cases. Mathematically, this can be stated as:

$$\text{Sensitivity} = TP/TP+FN$$

For this parameter the comparison between SVM and proposed method is perform in which it is found that the accuracy rate of SVM is about 76%, M and our method is about 84% which means our method generates better specificity rate than the existing SVM method.

Table 5: Specificity result analysis of the SVM and Proposed method.

Specificity		
Method/Type	SVM	Proposed
No	0.819209	0.857895
Yes	0.704545	0.832335

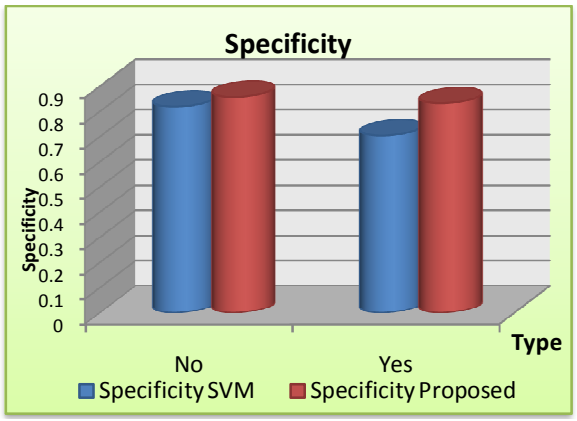

Fig. 6. Specificity graph between SVM and Proposed method.

Senstivity Analysis. Sensitivity (also called the true positive rate, the recall, or probability of detection in some fields) measures the proportion of positives that are correctly identified as such (i.e. the percentage of sick people who are correctly identified as having the condition).

$$\text{Specificity} = TN/TN+FP$$

For this parameter the comparison between SVM and proposed method is perform in which it is found that the accuracy rate of SVM is about 79.5% and our method is about 85% which means our method generates better specificity rate than the existing SVM method.

Table 6: Sensitivity result analysis of the SVM and Proposed method.

Sensitivity		
Method/Type	SVM	Proposed
No	0.784091	0.809211
Yes	0.80791	0.89589

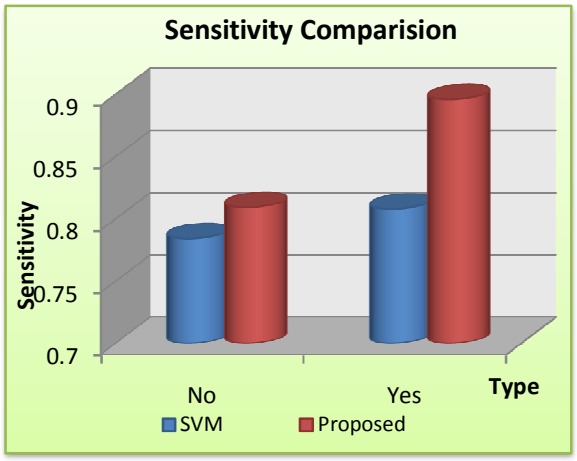

Fig. 7. Senstivity graph between SVM and Proposed method.

VI. CONCLUSION

The growth of diabetic patient is increasing rapidly so early prediction of essential data to lessen and cure it very significant. Data Mining is widely area in health care data. In this, we ensemple a mancine learning approach SVM and KNN-GA algorithm to predict the diabetic patient using Pima Indian women dataset which are collected from US National Institute of Diabetes and Digestive and Kidney Diseases. In this we uses the Pima.tr, Pima.tr2, Pima.te to predict diabetic patients using real health care data sets. The experimenta outcomes shows that the proposed data mining approach could assist health care providers to make healthier clinical decisions in identifying diabetic patients. Moreover, the approach could be further developed for patient fortification. In the future, the outcomes can be utilized to fashion a control plan for diabetes because diabetic patients are ordinarily not identified till a later step of the disease or the improvement of complications.

REFERENCE

[1]. Ojugo, A., Eboka, A., Okonta, E., Yoro, R and Aghware, F., "GA rule-based intrusion detection system", *Journal of Computing and Information Systems,* **3**(8), pp 1182 - 1194.

[2]. Goldenberg, R and Punthakee,, Z "Definition, classification and diagnosis, prediabetes and metabolic syndrome", **37**(1), S8-S11.

[3]. Harleen Kaur and Siri Krishan Wasan, "Empirical Study on applications of Data Mining Techniques in Healthcare", *Journal of Computer Science*, 2006.

[4]. "Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications", © World Health Organization 1999.

[5]. Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Morgan Kauffman Publishers, 2001.

[6]. Neeraj Bhargava, Girja Sharma, Ritu Bhargava and Manish Mathuria, "Decision Tree Analysis on J48 Algorithm for Data Mining" *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, Volume **3**, Issue 6, June 2013.

[7]. Rakesh Motka, Viral Parmar, Balbindra Kumar, A. R. Verma, "Diabetes Mellitus Forecast Using Different Data Mining Techniques", International conference on computer and Communication Technology .

[8]. Keller, J.M., Gray, M.R., Givens, J.A., A fuzzy K-nearest neighbor algorithm, Systems, Man and Cybernetics, IEEE Transactions on (Volume:**SMC-15** , Issue: 4). July-Aug. 1985, ISSN: 0018-9472.

[9]. G. Visalatchi et al, A Survey on Data Mining Methods and Techniques for Diabetes Mellitus, *International Journal of Computer Science and Mobile Applications*, Vol. **2** Issue. 2, February-2014, pg. 100-105 ISSN: 2321-8363.

[10]. S. Bahramian and A. Nikravanshalmani "Hybrid algorithm based on K-nearest-neighbor algorithm and Adaboost with selection of feature by genetic algorithms for the diagnosis of diabetes", *IJMEC*, Vol. **6**(21), Jul. 2016, PP. 2977-2986.

[11]. Cortes, C., Vapnik, V., "Support-vector networks", Machine Learning, 20(2),pp. 273-297, 1995.

[12]. V. Vapnik, "The Nature of Statistical Learning Theory." NY: Springer- Verlag. 1995.

[13]. Christopher J.C. Burges. "A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery", Springer,* **2**(2), pp.121-167, 1998.

[14]. Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C. and Johannes, R. S. (1988) Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications in Medical Care (Washington, 1988), ed. R. A. Greenes, pp. 261–265. Los Alamitos, CA: IEEE Computer Society Press.

[15]. Ripley, B.D. (1996) Pattern Recognition and Neural Networks. Cambridge: Cambridge University Press.