International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 2, February - 2014

# Design of Multi-Class Classifier for Prediction of Diabetes using Linear Support Vector Machine

Akshay Joshi
Department of Computer Engineering
AISSMS College of Engineering
Pune, Maharashtra, India

Anum Khan
Department of Computer Engineering
AISSMS College of Engineering
Pune, Maharashtra, India

Omkar Kulkarni
Department of Computer Engineering
AISSMS College of Engineering
Pune, Maharashtra, India

Prathamesh Palaskar
Department of Computer Engineering
AISSMS College of Engineering
Pune, Maharashtra, India

M. A. Pradhan
Department of Computer Engineering
AISSMS College of Engineering
Pune, Maharashtra, India

***Abstract* - Diabetes is a chronic disease. During the last 20 years the total number of people with diabetes worldwide has risen from 30 million to 230 million, according to the diabetes federation. It is a major health problem worldwide. Diabetes disease diagnosis with the proper interpretation of the diabetes data is an important classification problem. This project proposes a classifier for detection of diabetes using support vector machine (SVM). Linear SVM will be used in a multi-layered pattern to classify patients into diabetic, pre-diabetic or non-diabetic. This SVM based classifier can assist in the accurate decisions about Diabetes disease.**

***Keywords—Classifier, svm hyperplane, diabetes, SVM, Multi-Class.***

## I. Introduction

Diabetes is rapidly emerging as a major public health problem all over in India. In the urban areas, the prevalence is around 15 to 18%. Also, in the rural areas the prevalence is on the rise. In one of the studies conducted by India Diabetes Research Foundation centre showed an increase in prevalence of diabetes from 2.2% (1989) to 6.4% (2003) in Southern India. As of today, India is the Head quarters for diabetes in the world. Urbanization, changing lifestyle and genetic predisposition, have all contributed in the increasing prevalence of diabetes. However the fact remains that the diabetes can be prevented.

Diabetes can be classified as diabetes 1, diabetes 2 and gestation diabetes. Diabetes 1 is caused to less amount of insulin or no insulin secretion in body. This type of diabetes is caused majorly to young children, teens and young adults. They are mainly caused due to little insulin I in their body. This type of diabetes needs insulin to be injected in their body. Till now there is no exact cause for this type of diabetes. These types of patients are called Insulin Diabetes Dependent patients (IDDM). Diabetes type2 is cause due to resistance less in insulin. Type2 diabetes is majorly found in adults but now found in younger people also. This type of patients has insulin in their body which is not sufficient. Major cause for type2 diabetes is high obesity rate, majorly when BMI is greater than 25 then there exist greater percentage of risk. These types of patients are Non –Insulin Dependent patients (NIDDM). Gestations Diabetes is cause during pregnancy period. This type of diabetes can be cured after birth of child. Proper treatment has to be followed during this type of diabetes else there is heavy chance to change into type2 diabetes [1].

Automatic disease diagnosis systems have been used for many years. In this system, the task is to identify and select a useful subset of pattern- representing features from a larger set. Classification is an important data mining problem which has training set that consists of multiple records, each having multiple attributes. In order to identify the class of diabetes each record is tagged with a special class label. The main purpose of classification is to analyze the input data and to develop a model for each class using the features present in the data. This classifier will be further used to classify the unclassified records.

## II. Related Work

Numerous classification techniques have been proposed so far used for the classification of diabetes.

Nowadays the development of an effective diabetes diagnosis system by taking advantage of computational intelligence is regarded as a primary goal. Many approaches based on artificial network and machine learning algorithms have been developed and tested against diabetes datasets, which were mostly related to individuals of Pima Indian origin. Moreover a doctor has to rely on his past experiences and his knowledge to compare the results of a patients test results. Particular detection of diabetes depends on a large set of values. In order to optimize the result a need of a automatic classification arises.

Kemal Polat, Salih Güneş, Ahmet [2] Arslan performed a study for diagnosis of diabetes disease, which is one of the most important diseases in medical field using Generalized Discriminant Analysis (GDA) and Least Square Support Vector Machine (LS-SVM). While LS-SVM obtained 78.21%

classification accuracy using 10-fold cross validation, the proposed system called GDA–LS-SVM obtained 82.05%. Nahla H. Barakat and Andrew P. Bradley [4] have proposed utilizing support vector machines (SVMs) for the diagnosis of diabetes with prediction accuracy of 94%, sensitivity of 93%, and specificity of 94%. Davar Giveki [8] presented a novel automatic approach to diagnose Diabetes disease based on Feature Weighted Support Vector Machines (FW-SVMs) and Modified Cuckoo Search (MCS). It also uses PCA and Mutual Information (MI). An accuracy of 93.58% is obtained by the proposed MI-MCS-FWSVM method on UCI dataset.

## III. BACKGROUND DETAIL

### A. *What is SVM?*

The Support Vector Machine (SVM) is a state-of-the-art classification method introduced in 1992 by Boser, Guyon, and Vapnik. The SVM classifier is widely used in Computational Biology due to its high accuracy, ability to deal with high-dimensional data such as gene expression, and flexibility in modeling diverse sources of data.

SVMs belong to the general category of kernel methods. A kernel method is an algorithm that depends on the data only through dot-products. The dot product can be replaced by a kernel function which computes a dot product in some possibly high dimensional feature space. The two advantages are One, the ability to generate non-linear decision boundaries using methods designed for linear classifiers. Two, the use of kernel functions allows the user to apply a classifier to data that have no obvious fixed-dimensional vector space representation.

A Support Vector Machine (SVM) is a discriminative classifier defined by a separating hyperplane. It uses, labeled training data, to output an optimal hyperplane which categorizes new examples. The basic concepts in SVM are:

1. the separating hyperplane,
2. the maximum-margin hyperplane,
3. the soft margin and
4. the kernel function.

### B. *What are hyperplanes?*

Hyperplanes with larger margin are less likely to overfit the training data. A hyperplane should separates the classes and has the largest margin.

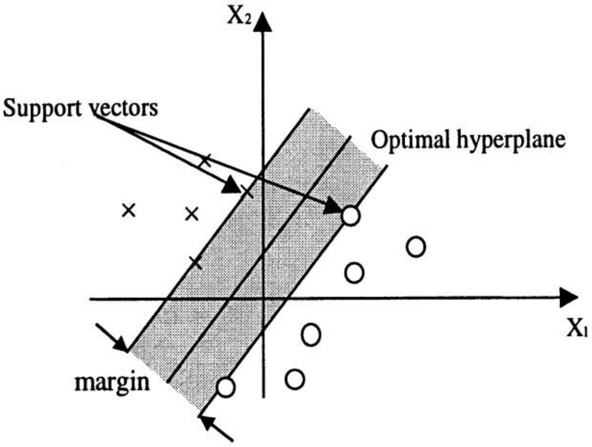


Figure 1: SVM classification with a hyperplane that maximizes the separating margin between the two classes.

A separating hyperplane can be written as

$$\mathbf{W} \bullet \mathbf{X} + b = 0$$

where $\mathbf{W}=\{w_1, w_2, \ldots, w_n\}$ is a weight vector and b a scalar (bias)

For 2-D it can be written as

$w_0 + w_1 x_1 + w_2 x_2 = 0$

The hyperplane defining the sides of the margin:

$H_1$: $w_0 + w_1 x_1 + w_2 x_2 \geq 1$ for $y_i = +1$, and

$H_2$: $w_0 + w_1 x_1 + w_2 x_2 \leq -1$ for $y_i = -1$

Any training tuples that fall on hyperplanes $H_1$ or $H_2$ (i.e., the sides defining the margin) are **support vectors**

### C. *Maximum margin*

The line that maximizes the minimum margin is a good bet. The model class of “hyper-planes with a margin of m” has a low VC dimension if m is big.

This maximum-margin separator is determined by a subset of the datapoints. Datapoints in this subset are called “support vectors”. It will be useful computationally if only a small fraction of the datapoints are support vectors, because we use the support vectors to decide which side of the separator a test case is on.

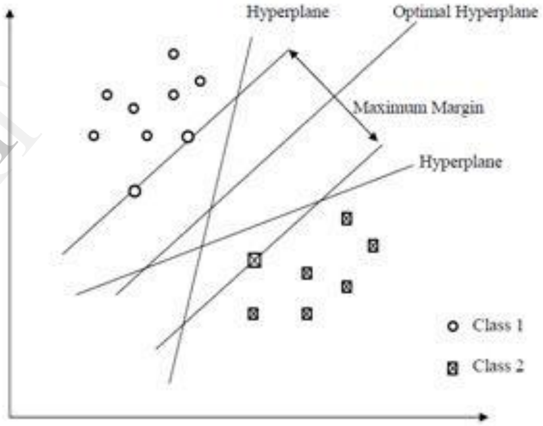


Figure 2: Maximum Margin

### D. *Kernel Functions*

Now the big bonus occurs because all the machinery we have developed will work if we map the points xi to a higher dimensional space, provided we observe certain conventions.

Let Φ(xi) be a function that does the mapping. So the new hyperplane is

$$\sum_{i=1}^{N} w_i \phi_i(\mathbf{x}) + b = 0$$

For simplicity in notation define

$$\phi(\mathbf{x}) = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_{m_1}(\mathbf{x}))$$

where $m_1$ is the new dimension size and by convention $\phi_0(\mathbf{x}) = 1.$

Then all the work we did with x works with $\phi(\mathbf{x}).$

The only issue is that instead of $\mathbf{x}_i^T\mathbf{x}_j$ we have a Kernel function, $K(\mathbf{x}_i,\mathbf{x}_j)$ where,

$$K(\mathbf{x}_i,\mathbf{x}_j)=\phi_i(\mathbf{x})^T\phi_j(\mathbf{x})$$

Examples
Polynomials

$$(\mathbf{x}_i^T\mathbf{x}_j+1)^p$$

Radial Basis Functions

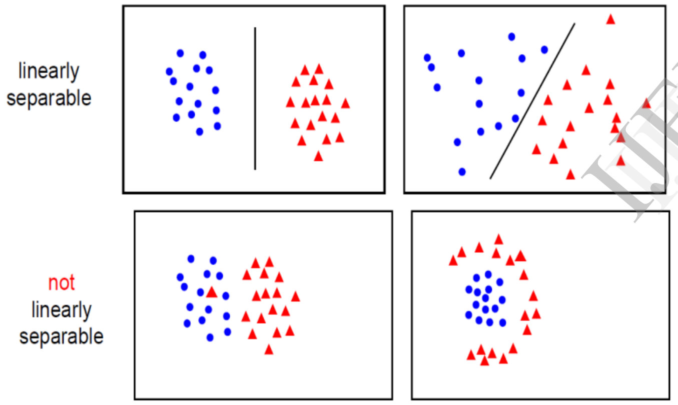
$$\exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_i\|^2}{2\sigma^2}\right)$$

### *E. Linear Support Vector Machine:*

Given training data $\mathbf{(x_i, y_i)}$ for $\mathbf{i = 1. . . N}$, with $\mathbf{x_i} \in \mathbf{R}^d$ and $\mathbf{y_i} \in \{-1, 1\}$, learn a classifier f(x), such that

$\mathbf{f(x_i) \geq 0 \quad y_i = +1}$
**or**
$\mathbf{f(x_i) < 0 \quad y_i = -1}$

Linear separability:



Linear Classifier has the form:

$$\mathbf{f(x) = w \,.\, x + b = 0}$$

where,

'**.**' denotes the dot product and $\mathbf{W}$ the normal vector to the hyperplane.
x is the vector of attribute values representing instance i in the training data.
w is the vector of weights for all attributes.
b is the real number representing the y-intercept.

The parameter $\frac{b}{\|\mathbf{w}\|}$ determines the offset of the hyperplane from the origin along the normal vector **w**.

If the training data are linearly separable, we can select two hyperplanes in a way that they separate the data and there are no points between them, and then try to maximize their distance. The region bounded by them is called "the margin". These hyperplanes can be described by the equations

$\mathbf{w \,.\, x + b = 1}$

and

$\mathbf{w \,.\, x + b = -1}$

By using geometry, we find the distance between these two hyperplanes is $\frac{2}{\|\mathbf{w}\|}$, so we want to minimize $\|\mathbf{w}\|$. As we also have to prevent data points from falling into the margin, we add the following constraint: for each $i$ either

$\mathbf{w \,.\, x_i >= 1}$ for $\mathbf{x_i}$ of the first class

or

$\mathbf{w \,.\, x_i <= -1}$ for $\mathbf{x_i}$ of the second.

This can be rewritten as:

$\mathbf{y_i(w.x_i + b) >= 1}$ **for all 1<= i <= n (1)**

We can put this together to get the optimization problem:

Minimize (in $\mathbf{w}, b$)

$\|\mathbf{w}\|$

subject to (for any $i = 1, \ldots, n$)

$\mathbf{y_i(w.x_i + b) >= 1}$

## IV. PROPOSED SYSTEM

The proposed system i.e. Diabetes Prediction System is based on linear SVM(Support Vector Machine). The following diagram depicts the proposed model.

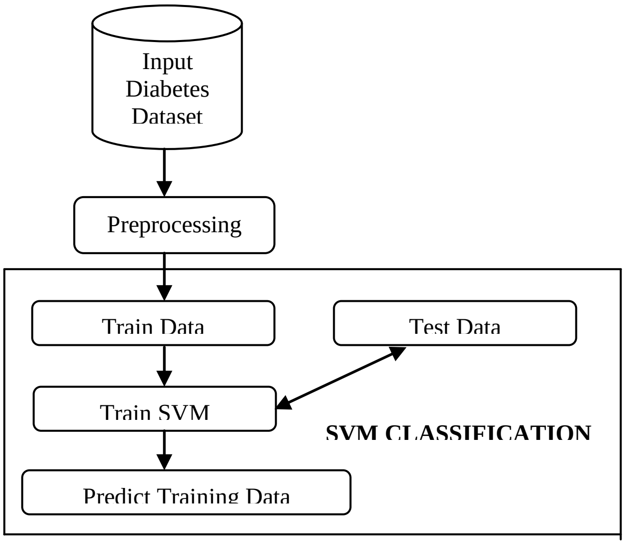


Figure 3: Proposed Model

The proposed SVM approach uses Dr. Rao's certified dataset. The dataset is trained to classify into two subsets

1. Diabetic patients
2. Non-diabetic patient
3. Pre-diabetic patient

The various blocks of the above diagram are explained as follows:

### *A. Dataset Overview*

The dataset used is obtained from Dr. Rao which contains the patient's entries which contain pre-diabetes, diabetes and non-diabetes patients. It has following attributes as:

*1) Attributes and their Dependencies.*

We have created a network structure which shows the factors leading to diabetes (and what all are caused by diabetes.)

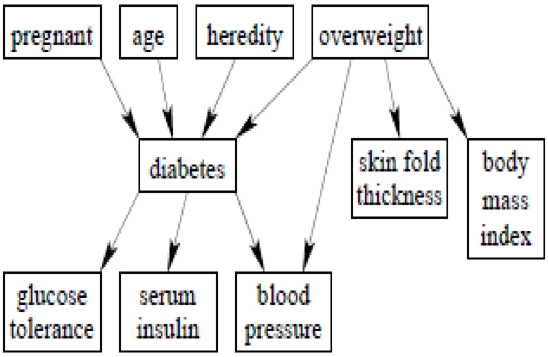


Figure 4 : Attribute Dependency

### *B. Data Preprocessing*

In order to make the data appropriate for the mining process it needs to be transformed. The raw data is changed into data sets containing a few appropriate characteristics using data processing techniques. Problems such as the removal of duplicate records, normalizing the values used to represent information in the database, accounting for missing data points and removing unneeded data fields will be resolved at this stage. The learning algorithm will heavily rely on the product of this stage, which is the final training set.

*1) Data Normalization:*

Most of the available features in our application have continuous values, in which each feature is measured in a different scale and has a different range of possible values. We scale all features to a common range because it is beneficial to do so, e.g. by standardizing the data (for each feature, subtracting its mean and dividing by its standard deviation) i.e. the attribute data is scaled to fit in a specific range because the accuracy of an SVM can severely degrade if the data is not normalized as large margin classifiers such as SVM are sensitive to the way features are scaled. Therefore, it is essential to normalize the data.

### *C. SVM Training*

After gathering data and preprocessing it the next step is to determine the SVM decision function i.e.

$$f(\mathbf{x}) = \sum_{j=1}^{l_s} \alpha_j^* y_j \Phi^T(\mathbf{s}_j)\Phi(\mathbf{x}) + b = \sum_{j=1}^{l_s} \alpha_j^* y_j K(\mathbf{s}_j, \mathbf{x}) + b.$$

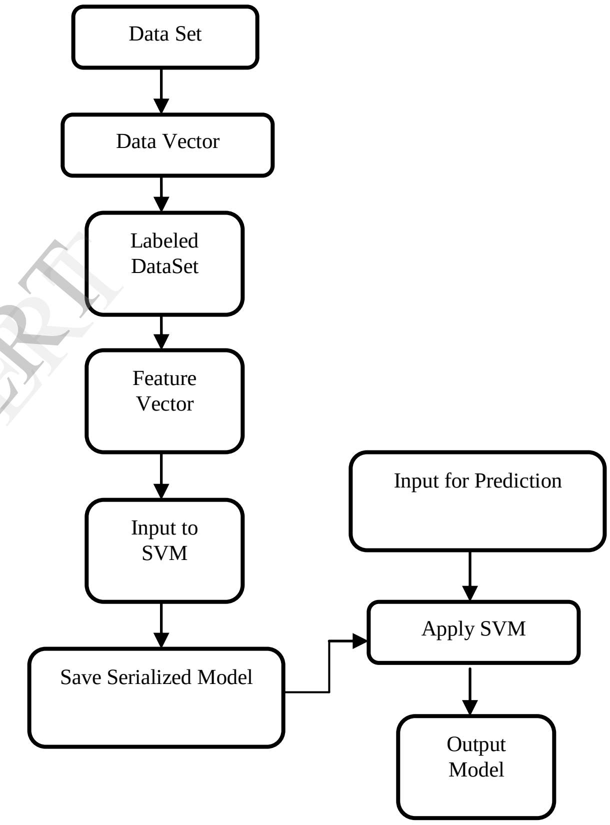
In this process, we must decide the following variables:  its associated parameter, and the regularization parameter in the structural risk function. To optimize these parameters, we applied type of cross validation to the training. This procedure consists of the following steps.

1) First, divide randomly all the available training examples into equal-sized subsets.

2) Second, for each model-parameter setting, train the SVM classifier m times; during each time one of the subsets is held out in turn while all the rest of the subsets are used to train the SVM. The trained SVM classifier is then tested using the held-out subset, and its classification error is recorded.

3) Third, the classification errors are averaged to obtain an estimate of the generalization error of the SVM classifier. In the end, the model with the smallest generalization error will be adopted.

## V. PROPOSED PLAN



The implementation of our project takes place in two stages:
1 Training.
2 Prediction.

The modules in the training part are as follows

1) *Data set*

This module is responsible for the task of importing the data set files which may contain data separated by comma means csv file (comma separated value).

2) *Data vector*

A data vector consists of a single tuple of the data set. In this, this data vector is applied serialization technique so as to maximize accuracy of the system. Serialization is the process in which the attributes of the data vector if they are in a specific range are rounded off to a predetermined value.

3) *Labeled data set*

In classification, often, an attribute may be specified as the class label attribute, whose values explicitly represent the classes. The task of this module is to label the data vectors in the training data set.

4) *Feature Vector*

In pattern recognition and machine learning, a feature vector is an n-dimensional vector of numerical features that represent some object. Many algorithms in machine learning require a numerical representation of objects, since such representation facilitate processing and statistical analysis.

Feature vectors are often combined with weights using a dot product to construct a linear predictor function that is used to determine a score for making a prediction. Thus the data set can now be applied SVM.

5) *Input to SVM*

SVM trains the data set using the feature vectors extracted earlier.

6) *Save serialized model*

The task of training the data set is now complete. This data set is known as the training data set. This data set is now stored and is to be used in the next step that is prediction.

The training part is completed. Following this the prediction part takes over. Prediction is to make inferences based on the given data set. Its modules are as follows:

7) *Input for prediction*

This module performs the task of accepting input for prediction. The input is unlabeled and it may be a single entry or a large data set.

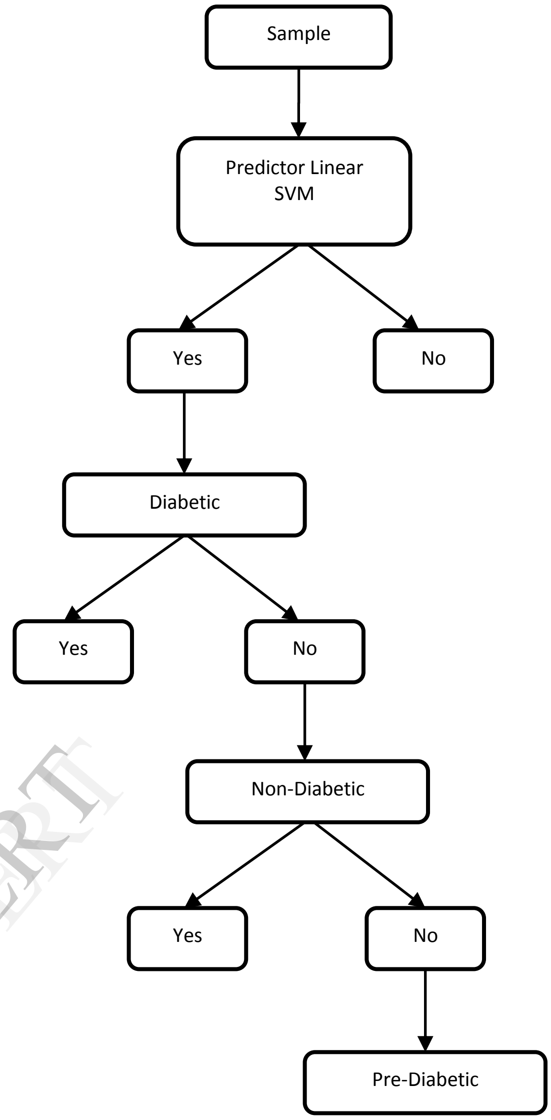
8) *Apply SVM*

SVM is applied to this data set. SVM loads the training data set. It extracts feature vector from the test data set and generates a prediction model.

9) *Output model*

The output model declares whether the person is diabetic, non-diabetic or pre-diabetic. On this basis, graphs are generated.

## A. MULTI-LAYER APPROACH



1) *Multi class classification*

In machine learning, multi class classification is the problem of classifying instances into more than two classes. While some classification algorithms naturally predict the use of more than two classes, others are by nature binary algorithms, these can be however be turned into multi class classifiers by a variety of strategies.

2) *Multi class classification using SVM*

Multi class SVM aims to assign labels to instances by using SVM, where the labels are drawn from a finite set of elements. The dominant approach for doing so is to reduce the single multi class problem into multiple binary classification problems.

### *3) Working*

In the diabetes prediction system, the linear SVM model works on the basic idea as follows

a) Decompose multi class tasks into a set of binary tasks.

b) Create ensemble of binary classifiers for binary tasks.

c) Combine outputs of ensemble as multi-class classifier.

In the first step, sample which may be an unlabeled data set or a single entry is imported by the system. Then, the linear SVM predictor model is loaded. The task is now to classify the patients into diabetic or non-diabetic or pre-diabetic classes.

We implement the algorithmic strategies such as divide and conquer incremental prototyping and unit integration.

The task of the classifier begins now. Each unlabeled instance is applied linear SVM so as to predict the output. In the first case, the classifier checks for the condition of diabetes. If yes, then the patient is a diabetic patient. If no, the patient may then be classified into two classes non-diabetic or pre-diabetic.

Similarly to the first case, we now undertake classification of the patient for the non-diabetic class. If yes, then the patient is a non-diabetic patient. If no, the patient may be classified into pre-diabetic.

## VI. Conclusion

Diabetes is a major health concern. The early detection and prediction of the diabetes can gave a warning at a stage where some medications and preventive measures will help the patient to increase the span of his healthy life. In this paper, we have presented a classifier for diabetes diagnosis. We have employed SVM for the diagnosis. We will be using Linear SVM in multi-layered pattern to recognize patients as diabetic, pre-diabetic or non-diabetic. The multi-layered approach will enable us to trace the result fast as it will eliminate the possibilities in stages.

The SVM and the rules extracted from it are intended to work as a second opinion for diagnosis and as a tool to predict diabetes through identifying people at high risk. The significance of our approach lies in its simplicity.

## VII. References

[1] Aishwarya. R, Gayathri. P and N. Jaisankar, authors of *“A Method for Classification Using Machine Learning Technique for Diabetes”,* International Journal of Engineering and Technology (IJET). ISSN : 0975-4024 Vol 5 No 3 Jun-Jul 2013.

[2] Polat, K., & Gunes, S. authors of, *“An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease.”* Digital Signal Processing, 17(4), 702–710.2007.

[3] Howard Robin., John. S. Eberhardt III., W. D. Muller., R. Clark.,J.Kam., authors of *"Classification of Pathology Data Using a Probabilistic (Bayesian) Model"*, in proceedings of 18th International Conference on Systems Engineering,2005.,pp. 286- 291, 16-18 Aug.2005. doi: 10.1109/ICSENG.2005.22

[4] Nahla H. Barakat, Andrew P. Bradley, Mohamed Nabil H. Barakat, authors of *“Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus”*, IEEE transaction on Information technology in bioinformatics, VOL. 14, NO. 4, JULY 2010.

[5] M.W. Aslam, A.K. Nandi, authors of *“Detection of diabetes using genetic programming”.* 18th European Signal Processing Conference. 2010.

[6] B.M. Patil , R.C. Joshi, Durga Toshniwal Department of Electronics and Computer Engineering, Indian Institute of Technology, Roorkee 247667, India, authors of *“Hybrid prediction model for Type-2 diabetic patients”*

[7] Mohamed Amine Chikh, Meryem Saidi, Nesma Settouti, authors of *“Diagnosis of Diabetes Diseases Using an Artificial Immune recognition System2 (AIRS2) with Fuzzy K-nearest Neighbor”.* Springer Journal of Medical Systems. 2011.

[8] Davar Giveki, Hamid Salimi, GholamReza Bahmanyar, Younes Khademian, authors of *“Automatic Detection of Diabetes Diagnosis using Feature Weighted Support Vector Machines based on Mutual Information and Modified Cuckoo Search”.* 2012

[9] Nesma Settouti, Meryem Saidi, and Mohamed Amine Chikh, authors of *“Interpretable Classifier of Diabetes Disease.”* International Journal of Computer Theory and Engineering vol. 4, no. 3, pp. 438-442. 2012.

[10] Shankaracharya, Devang Odedra, Subir Samanta and Ambarish S. Vidyarthi, authors of *“Computational Intelligence in Early Diabetes Diagnosis:A Review”* The Review of DIABETIC STUDIES Vol 7 No 4 2010.