

# A Method for Classification Using Machine Learning Technique for Diabetes

Aishwarya. R<sup>1</sup>, Gayathri. P<sup>2</sup> and N. Jaisankar<sup>3</sup>

M.Tech Student<sup>1</sup>, Assistant Professor (Senior)<sup>2</sup> and Professor<sup>3</sup>

School of Computing Science and Engineering, VIT University, Vellore – 632014, Tamil Nadu, India.

<sup>1</sup>aishwaryar.rajendran@gmail.com, <sup>2</sup>pgayathri@vit.ac.in, <sup>3</sup>njaisankar@vit.ac.in

**Abstract** - Machine learning has been one of the standard and improving techniques with strong methods for classification and reorganization based on recursive learning. Machine learning allows to train and test classification system, with Artificial Intelligence. Machine learning has provided greatest support for predicting disease with correct case of training and testing. Diabetes needs greatest support of machine learning to detect diabetes disease in early stage, since it cannot be cured and also brings great complication to our health system. One of the promising techniques in machine learning is Support Vector Machine (SVM). SVM is used for classification of system. Upshot of SVM has provided with classification of system.

**Keywords** - Machine Learning, Classification, Diabetes

## I. INTRODUCTION

Machine learning technique has been an excellent support for making prediction of a particular system by training. Machine learning is about learning structures from the data which is provided. Machine learning in recent years have been the evolving, reliable and supporting tool in medical domain. Automatic learning has fetched a greater amount of interest in medical domain due to less amount of time for detection and less interaction with patient, saving time for patients care.

Diabetes is one of the chronic (life long) disease. Diabetes is caused due to increase in blood sugar. Major cause for diabetes is little insulin and resistance to insulin. Prolong diabetes cause more health complications.

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.

Symptoms for diabetes include as blurry vision, fatigue, hungry, urinary often excess thirst and weight loss or gain. High and low Blood Pressure, Smoking and BMI can also be one of the reasons for diabetes. These are the major changes of pre-diabetic. Pre-diabetes gives heavy chance to cure the diabetes with proper food and exercise by increasing the insulin resistance. The diabetes type2 has the major amount of symptoms. Major study have proved prolong diabetes may lead to complications. Major complications are cardio vascular disease, Macro vascular, ischemic heart disease, stroke peripheral heart disease and chronic renal failure.

## II. RELATED WORK

Machine learning is one of the major methods for predicting or finding through underlying mechanism. So the major focus in machine learning technique is the mechanism or algorithm which yields an intelligent output by recognizing complex patterns. Yield patterns or predictions thought to be features of the underlying mechanism that generated the data. A learner can take advantage of examples (data) to capture characteristics of interest of their unknown underlying probability distribution. Data can be seen as instances of the possible relations between observed variables. Machine learning has made various designs for recognizing patterns to make intelligent decision for the input data. Major challenge for machine learning technique is behavior of inputs which should be trained during observed examples. Hence they are trained with all possible inputs to produce efficient and sensitive output.

Machine learning is one of the major ways for classification of disease. Variety of method has been proposed so far for easy deduction and classification of disease. Disease like diabetes needs more training with proper data sets. Since diabetes shows various signs for the presence of blood glucose and various instances.

Recent study tells that 80% of complications can be prevented by identification .intelligent data analysis method like machine learning technique are valuable in identification which can increase in early detection [2]. Automatic disease diagnosis systems have been used for many years. While these systems are constructed, the data used needs to be classified appropriately. For this purpose, a variety of methods have been proposed in the literature so far. Hybrid method for classification is also used in machine learning techniques. Hybrid technique includes Artificial Neural Networks (ANN) and Fuzzy - Neural Networks (FNN). They have verified for two real-time problems. In order to evaluate the performance of the proposed method accuracy, sensitivity and specificity performance measures that are used commonly in medical classification studies were used. The classification accuracies of these datasets were obtained by k-fold cross-validation. The proposed method achieved accuracy values 84.24% and 86.8% for Pima Indians diabetes dataset and Cleveland heart disease dataset respectively [3].

C-mean clustering mechanism for classification disease is used. The accuracy of c-mean clustering for classification of diabetic disease was 86.4 % [4]. Automatic detection of disease can be done with Linear Discrimination Analysis and Morlet Wavelet Support Vector Machine classifier (LDA-MWAVM) system. There are three stages, feature extraction and feature reduction stage by using LDA. Classification is done using Morlet Wavelet Support Vector Machine (MWSVM). Efficiency of this system is 89.74%. Efficiency is less compared to other methods [6]. A new hybrid system was adapted based on Generalized Discriminate Analysis (GDA) and Least Square Support Vector Machine (LS – SVM). The proposed system consists of two stages. GDA is used for recognizing difference with healthy and unhealthy people as preprocessing technique. LS provided 78.21% with 10 cross fold. LS - SVM is used for classification using 12-fold. The efficiency percentage obtained is 82.05%. Efficiency has been improved. The output efficiency is checked using cross-validation and confusion matrix [7].

Classification using decision tree support using Classification and Regression (CART) is common technique used for classification and finding the parents. Complications like dyslipidemia, hypertension, cardiovascular disease, retinopathy and end-stage renal disease generally occurs when HgbA1c > 9.5 only, 10 predictor factors are used for detecting [8]. ANN is one of the fast developing and promising technique used in real world problems. Accuracy can be improved using the preprocessing technique, since most medical data updates are not complete. ANN gives classification efficiency experimentally of about 99% [9].

Ant Colony Optimization (ACO) is successful in rule based classification. New algorithm is presented in this paper for extracting If-then rule for diagnosing diabetes. The proposed method is FADD. The method is based on an online fitting procedure and it is evaluated using eight biomedical datasets and five versions of the random forests algorithm (40 cases). The method decided correctly the number of trees in 90% of the test cases. Efficiency is less while comparing. Need more attributes [10]. Learning Vector Quantization (LVQ) is one such powerful method for classification. The major advantage of this system is successful disease diagnosis rate. This LVQ is further developed as Adaptive LVQ with Reinforcement Mechanism (ALVQ). The parameters of reinforcement are updated in adaptive way in network. Here comes the major disadvantage with adaptive method. The missing values are added using median values. With this algorithm they have checked breast cancer and thyroid. The accuracy is about 99.5 % [5].

SVM is one of the machines learning technique. SVM operates by finding linear plane methods for classification model data are used to implement classification by developing new model. Major aim lies in classification are reliability. Reliability increases when there is increase in efficiency and support for the system by ease for use. SVMs provide a promising tool for the prediction of diabetes ,where a comprehensible rule set have been generated, with prediction accuracy of 94%, sensitivity of 93%, and specificity of 94% [2].

### III. PROPOSED WORK

We developed a system which detects diabetes and reveals about the complications due to diabetes. In this paper a system is proposed which uses Principal Component Analysis (PCA) for preprocessing techniques. PCA is one of the most promising techniques for preprocessing. Classification of diabetes disease, with non diabetes and diabetes is done using SVM. The following fig 1 shows the architecture of system.

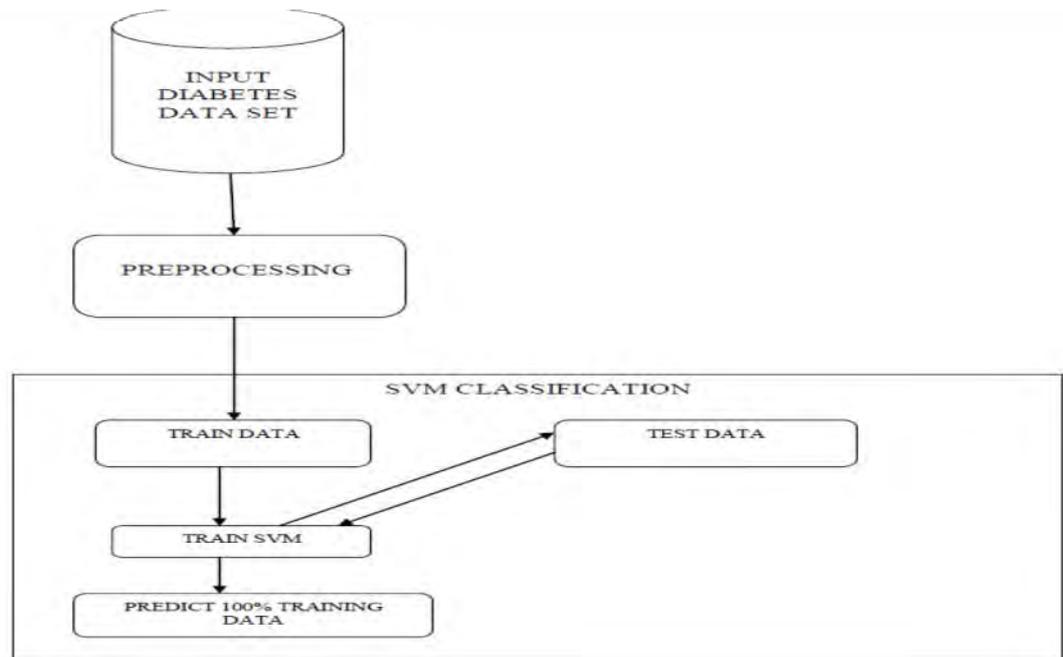


Fig. 1. System Architecture

#### A. Revealing proper data from data set

This is one of the challenging tasks for a new system to be introduced. Attribute selection needs more concern for getting exact percentage of efficiency. In this paper, a standard dataset is used for detecting the proposed system. The dataset is obtained from the Pima diabetes database. The dataset contains data for 769 patients with eight attributes. Both sick and healthy patient's data are obtained. Eight attributes are pregnancy, diastolic BP, tri fold thick, serum Ins, BMI, DP function and age along with diagnosis. The datasets are stored in excel format for preprocessing the data.

#### B. Preprocessing

Preprocessing is done using Principal Component Analysis (PCA) which is a more standard technique which is used for processing the raw data. PCA preserves dimensionality which is very important for medical data classification. The following formula gives the model of how the transformation of data can be done in PCA.

$$\begin{aligned}
 Y^T &= XW \\
 &= V\Sigma^T W^T W \\
 &= V\Sigma^T \quad [15].
 \end{aligned}$$

In the proposed system, classification of diabetes is done by PCA. Preprocessed data is further classified using SVM. PCA is one of the most promising techniques in preprocessing. PCA reduces dimensionality problem along with its exact essential data, necessary for processing. The major advantage of PCA is it balances between necessary data which has to be retrieved and data which have to be over-simplified. There is clear representation gives easy understanding. The residual data which is in A is not retained in for X. Necessity of preprocessing is to decide dimensionality. PCA has major advantage of fixing the dimensions required.

- Dimensionality reduction without losing data.
- Missing data or missing value prediction and replacing
- Database fix for classification

These are the major important tasks which are necessary to be checked for SVM classification in our system.

#### C. Classification of patients

Support Vector Machine is a supervised technique which is associated with learning algorithms. SVM technique is used for finding patterns for classification and quantitative predictions of one variable from the values of another. General format of SVM is predicting the output based on the trained data. The inputs are given the outputs are predicted within two options. New behaviors are then trained in same space such that output follows within two classes.

SVM is one the machine learning technique SVM is one the supervised learning model with learning algorithms that analysis data and recognized patterns. Training is done for classification and regression analysis. Basis data takes the input and then predicts the output based on the trained data. Only two possible classes are made possible. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. 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. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on in addition performing linear classification.

SVM can efficiently perform linear classification. Training SVM needs quadratic equations to be solved where coefficient equals the training samples. In practice problems are divided into sub problem and then limit performance iteration pairwise. Procedure for incremental and decremental is reversible and leave-one-out generation performance on training data. Training SVM incremental deals like discarding the previous data except the support vectors. Only one vector is trained at a time, Kuhn Tucker conditions retain adiabatic increments in steps which are computed analytically. Efficiency is improved by incremental and decremented orders, incremental is reversible and decrement that is unlearning makes it more advantages than SVM. Major focus on the system is

- The increment keeps the focus on the margin coefficient value changes in every step and checks for the margin are placed in equilibrium.
- Decrement is just the reverse of the incremental model; it leaves out the value and checks for new values.

Algorithm checks out the all the nearest point in opposite class. If it finds out a new data or case which is presented in the opposite class then a new margin is added. This continues till all the data are perfectly checked. Classification depends completely on this margin and data which it predicts as new nearest neighbor. The following steps give out the clear description of the SVM algorithm for classification.

Step 1:

Get the data from the processed datasets. Check for the datasets are adaptable for the application. Make the data set adaptable for the application.

Step 2:

Train the data from the data sets which is available for classification. More we train the classification patter better the prediction.

Step 3:

Test the data, with the better system for training

Step 4:

Check for new updates present f there is new update and if the data is adaptable for the application then include for classification.

Step 5:

Check for the proper data set classification.

#### IV. IMPLEMENTATION DETAILS

Matlab 10 is used for implementation of SVM classification. Preprocessing and classification used bioinformatics tool available in Matlab. Bioinformatics tool was used to check prevalence of diabetes in patients. Data (benchmark) which is obtained from Pima database is preprocessed, since preprocessing always gives better results. Data is first normalized before preprocessing, normalizing data is very important before preprocessing, using raw data without normalization PCA or any algorithm will result in different results. PCA is used for reducing the dimensions by maintaining much variance as much as possible. Output from preprocessed data is now achieved as the input for classification, SVM classification technique is used for classification of diabetes. SVM classifier fix out the margin at initial level and classifies, latter margin changes at each increment operation opting for new margin. This leads the data (preprocessed) to classify as presence of diabetes or not, all the data in database is checked and falls in either of the classification group.

#### V. RESULTS AND DISCUSSION

The diagnosing of diabetes using SVM with preprocessing PCA shows better accuracy that the existing system. Sample classification done using SVM is given in fig2. In fig2, classified output can be observed, two patches can be checked in which blue color gives diabetes patients and red color gives no diabetes patients. Patient number is given along for identifying after classification. Accuracy of existing system has been 89.07 percentages and proposed system has achieved output of 95 %. Benchmark data is obtained from Pima database.

Performance analysis is done using ROC curve fig 3. The analysis is performed with true positives (TP), true negatives (TN), false positives (FP), false negatives (FN) obtained and put in a confusion matrix. With these values, sensitivity, specificity and accuracy are calculated as follows:

$$\text{Sensitivity} = (\text{TP} / (\text{TP} + \text{FN})).$$

$$\text{Specificity} = (\text{TN} / (\text{FP} + \text{TN})).$$

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

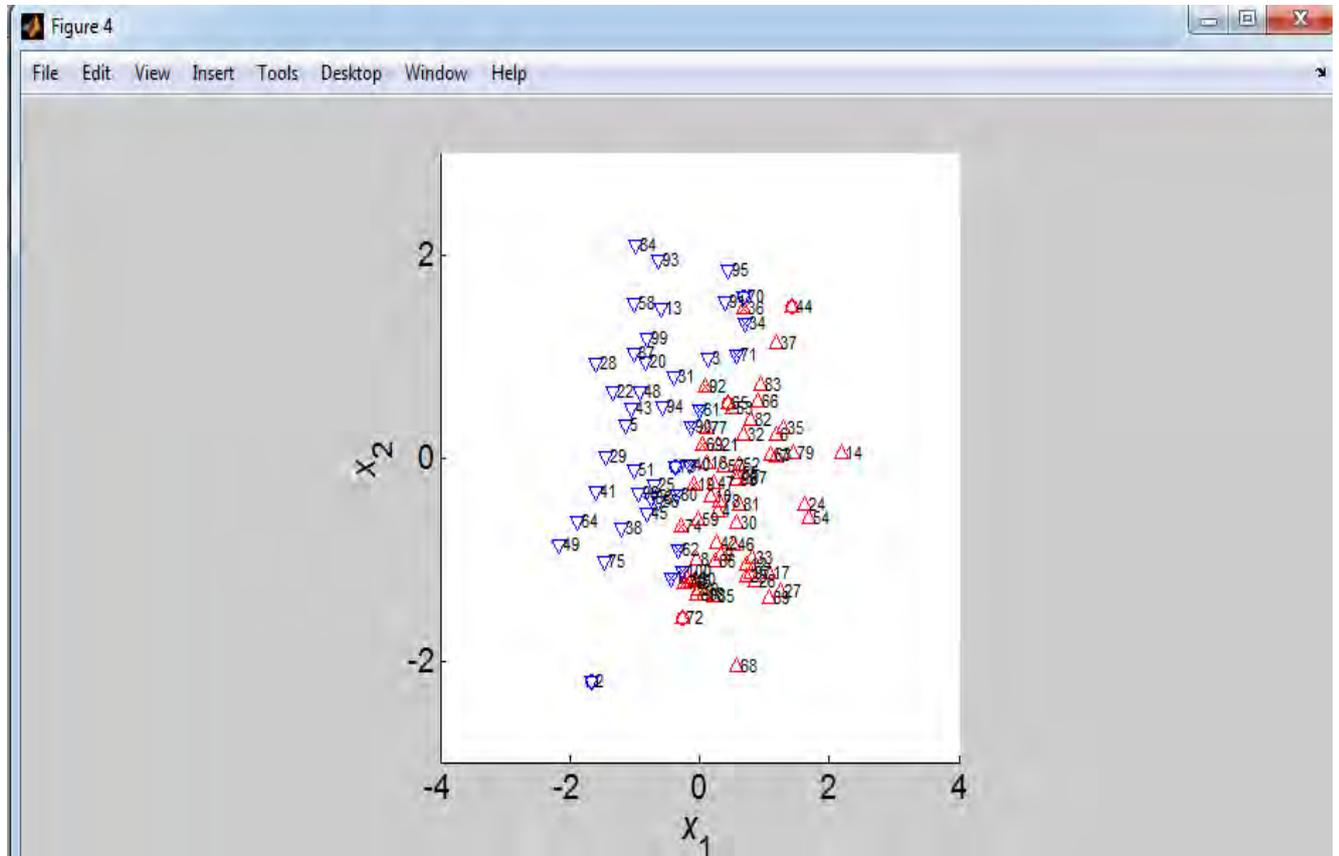


Fig. 2. CLASSIFICATION

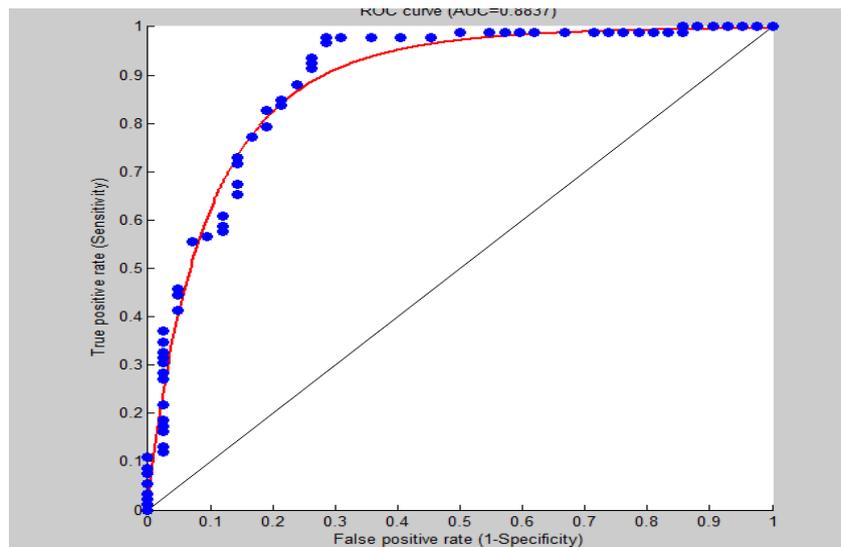


Fig. 3. ROC CURVE

## VI. CONCLUSION AND FUTURE ENHANCEMENT

The accuracy of the proposed system is good while preprocessing, when compare with previous work which has been done without preprocessing the system. Preprocessing has played a key role in classification of diabetes. Future work can be done by further predicting the complications of the patients.

## REFERENCES

- [1] A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks- I.T.Jayalakshmi Computer Science Department CMS College of Science and Commerce Coimbatore, India , Dr .A. Santhakumaran Statistics Department Salem Sowdeswari College Salem, India 2010 International Conference on Data Storage and Data Engineering.
- [2] Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus Nahla H.Barakat, Andrew P. Bradley, Senior Member, IEEE, and Mohamed Nabil H. Barakat IEEE Transactions on Information Technology in biomedicine, vol. 14, no. 4, July 2010.
- [3] Design of a hybrid system for the diabetes and heart diseases Humar Kahramanli \*, Novruz Allahverdi Department of Electronic and Computer Education, Selcuk University, Konya, Turkey.
- [4] Nonparametric criteria for supervised classification of fuzzy data Ana Colubi a, Gil González-Rodríguez a, M. Angeles Gil a, Wolfgang Trutschnig b a Department of Statistics, University of Oviedo, 33007 Oviedo, Spain b Research Unit on Intelligent Data Analysis and Graphical Models, European Centre for Soft Computing, 33600 Mieres, Spain
- [5] A fast and adaptive automated disease diagnosis method with an innovative neural network model Erdem Alkım, Emre Gürbüz, Erdal Kılıç Department of Computer Engineering, Faculty of Engineering, Ondokuzmayıs Universities, 55139 Kurupelit, Samsun, Turkey, Pg, Neural Networks 33 (2012) 88–96
- [6] An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier Duygu Calisir a, Esin Dog̃antekin ba Istanbul University, Cerrahpas a Medical Faculty, \_Istanbul, Turkey b Firat University, Firat Medicine Center, Department of Microbiology and Clinical Microbiology, 23119 Elazi , Turkey, pg, Expert Systems with Applications 38 (2011) 8311–8315.
- [7] A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine Kemal Polat a, Salih Gu`nes a, Ahmet Arslan b a Selcuk University, Electrical and Electronics Engineering, 42075 Konya, Turkey b Selcuk University, Computer Science, 42075 Konya, Turkey, Expert Systems with Applications 34 (2008) 482–487.
- [8] Data mining a diabetic data warehouse Joseph L. Breaulta,b,\*, Colin R. Goodallc,d, Peter J. Fose, b, pg, Artificial Intelligence in Medicine 26 (2002) 37–54.
- [9] Revision of the ADA-classification of diabetes mellitus type 2 (DMT2): The importance of maturity onset diabetes (MOD), and senile diabetes (DS) Marco Vacante, Michele Malaguarnera, Massimo Motta \*, pg, Archives of Gerontology and Geriatrics 53 (2011) 113–119.
- [10] Using fuzzy Ant Colony Optimization for Diagnosis of Diabetes Disease Mostafa Fathi Ganji Faculty of Electrical and Computer Engineering University of Tarbiat Modares Tehran, Iran, Mohammad Saniee Abadeh Faculty of Electrical and Computer Engineering University of Tarbiat Modares Tehran, Iran
- [11] Data mining a diabetic data warehouse Joseph L. Breaulta, b Colin R. Goodallc, d Peter J. Fose, pg, Artificial Intelligence in Medicine 26 (2002) 37–54.
- [12] The challenge of undiagnosed pre-diabetes, diabetes and associated cardiovascular disease Ved V. Gossain , Saleh Aldasouqi 1 Division of Endocrinology, Department of Medicine, Michigan State University, East Lansing, MI, USA
- [13] Medical data mining by fuzzy modeling with selected features Sean N. Ghazavi, Thunshun W. Liao \* Industrial Engineering Department, 3128 Patrick F. Taylor Hall, Louisiana State University, Baton Rouge, LA 70803, United States.pg, Artificial Intelligence in Medicine (2008) 43, 195—206.
- [14] Link: <http://www.machinelearningtechniques.com>.
- [15] Link: [http://en.wikipedia.org/wiki/Principal\\_component\\_analysis](http://en.wikipedia.org/wiki/Principal_component_analysis).
- [16] Link: <http://www.mathworks.in/products>.