International Journal of Emerging Engineering Research and Technology
Volume 2, Issue 7, October 2014, PP 231-239
ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

Comparison for Classification of Clinical Data Using SVM and Nuerofuzzy

Nilufar Zaman

Computer Engineering, AISSMSCOE, Pune, India
nilufar.zaman@mescoepune.org

Prof. D P Gaikwad

Aissms College of Engg, Pune

Abstract: *The soft computing techniques like SVM and nuerofuzzy is used here to find the accuracy of the clinical data. By comparing the accuracies of both the soft computing techniques it is being seen that for the given dataset SVM gives more accuracy than the nuero fuzzy techniques. More over it is being seen that pre-processing techniques also help alot in increasing the accuracy. It is being observed that SVM gives more accuracy with attribute selection than neuro fuzzy techniques.*

Keywords: *SVM, Nuero fuzzy techniques, Clinical data, pre processing techniques, attribute selection, LEM2, LERS 92.*

1. INTRODUCTION

The main aim of classification is to determine whether the particular attribute belongs to a particular class or not i.e. whether a∈C or a∉C where a is the attribute and C is class to be considered. In this paper we are classifying the person who has daibetes, who doesnot have diabetes or may have it. Diabetes is a metabolic disorder where the body either do not respond to the insulin that is produced or the body may stop producing insulin which results in high blood pressure[1]. The dataset used here is the Pima dataset which is used to determine whether the person is having diabetes or not. Initially SVM is used to classify pima dataset and is being compared with the nuero fuzzy technique [1]. The data set used here is further being pre-processed with pre processing techniques like normalization, dicretization, and attribute selection. It is found that the pre-processing techniques also contributes alot for increasing the accuracy. For the nuero fuzzy technique LEM2 algorithm is compared with Wang and Mendel algorithm [1]. The first algorithm used for nuero fuzzy technique is the Learnable Evolution Model (LEM2) [1][3][4] algorithm which is implemented in Learning from Examples based on Rough sets (LERS-92) application. The second algorithm used is the Wang and Mendel algorithm [1][5][6]. It is being observed that when only LibSVM technique is used then the accuracy rate for correctly classified objects can be shown as below:

LEM 2 > Wang and Mendel > SVM.

But when pre-processing techniques are used like Normalization, Discretization and Attribute Selection, it is being observed that slowly the accuracy level increases and when all these pre-processing techniques applied at a time SVM happens to give more accuracy than LEM 2 and Wang and Mendel Model i.e.

SVM > LEM 2 > Wang and Mendel.

2. RELATED WORK

There are various works done on the diagnosis of diabetes using Pima dataset to improve the accuracy of the classification.

In 1994 Michie, Taylor and Spiegelhalter uses various machine learning techniques to show the missing Classification Error Rate. [12] They used classifiers such as discrm, logdisc, Dipol 92, Smart, Itrule, RBF and Backprop which shows the miss classification rate from 22% to 24.5%. For more accuracy they used classifiers as Cal 5, CART, CASTLE, NaiveBay, Quadisc, C4.5, IndCART, Baytree, LVQ and Kohonen which shows the miss classification rate from 25% to 27.3% . Further they have used classifiers like AC2, CN2, New-ID, ALLOC80 and K-NN which shows the miss classification rate as 27.6% to 32.4%.

In 2004 Lena Kallin Westin discussed on his paper handles the missing data using various pre-processing methods.[13]

In 2009 for Neural Networks performance on Pima dataset Jeatrakul and Wong[14] have used various classifiers such as BPNN, GRNN, RBFNN, PNN and CMTNN which shows the miss classification rate from 23.44% to 24.74%.

In 2002 Bylander used Naive Bayes, Decision Tree and two types of Belief Networks (BN) i.e. the Belief Network itself and Belief Network with laplace on Pima dataset[15] which gives the miss classification rate ranging from 27.5% to 28.5 %.

In 2009 Robert Nowicki, Member, IEEE on his paper uses LEM2 and Wang & Mendel on various datasets of which he has used pima dataset to show the accuracy. [2]

3. PROPOSED SYSTEM

In this paper an automated approach for classification of diabetes disease is being done here with the help of PIMA datasets.

The dataset used here consists of 9 attributes of which 8 attributes are inputs and the last one is the output which is the class. The 8 attributes are explained as below:

- *Pregnant:* This attribute consists of the record of number of times the women get pregnant.
- *Plasma Glucose:* It consists of the plasma glucose concentration measured using 2 hours oral glucose tolerance test. It is measured in "mmHg".
- *Diastolic BP:* It's the Diastolic Blood Pressure.
- *Skin:* It's the Triceps skin fold thickness. It is measured in terms of "mm".
- *Insulin:* It is measured by fasting with 2 hours serum insulin. It is measured in terms of "muU/ml"
- *Mass:* It's the Body mass index. It is measured in terms of "kg/height in (mm)2"
- *Pedi:* It is the Diabetes Pedigree function.
- *Age:* Age in terms of year also helps in detecting diabetes as aged persons are more prone to diabetes.
- *Class:* Diabetes on set within 5 years.

In this model the dataset is pre-processed before classifying with the pre-processing techniques like:

1. *Normalization:* It is a data transformation technique where the data or the attributes are so scaled that it will fall within a small specified range such as from -1 to 0 or 0 to 1.
2. *Discretization:* It is a data reduction pre-processing technique where raw data values for attributes are replaced by ranges or higher conceptual levels.

Attribute or Feature Selection: It is the technique that determines the splitting methods of the tuples at a given node. It produces subset from the original features. In the procedure of subset generation the new subset produced is being checked with the previous one and the procedure is continued till we get the best subset among the subsets and the stopping point comes when some stopping criteria is applied [10][11]. Then validation is done for the best subset.

The proposed method is shown below:

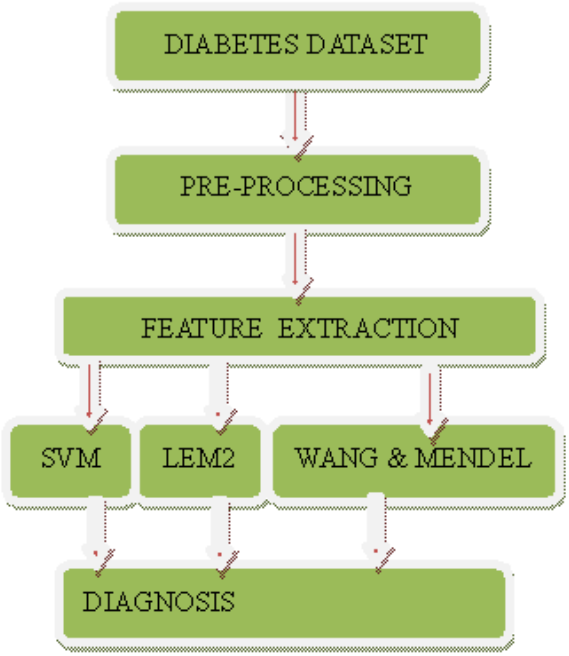

Fig1. *Block diagram for the comparison between classifiers for Diagnosis of diabetes.*

Cross Validation: Cross Validation is a technique of classifier which estimates its performance.

The Cross Validation can be "n" folded where n is the number of sets. In "n" folds 1 set is used for testing whereas "n-1" is used for training the dataset for example 3- fold Cross Validation can be shown as below:

Test Set	Training Set	Training Set
Training Set	Test Set	Training Set
Training Set	Training Set	Test Set

Fig2. *Block Diagram showing 3-Fold Cross Validation.*

4. SUPPORT VECTOR MACHINE

Data can be either linear or nonlinear. If data is linear we can use LibSVM but for nonlinear data kernel functions are used. For nonlinear data selecting Kernel function is a difficult task, so for this case we can use Selection and Adaption method.

In our case for Pima Dataset we have used LibSVM as it consists of only linear data and pre-processing techniques are also being used before classifying with SVM which helps in increasing the accuracy.

Support Vector Machine is the classifier that analyzes the training data to find an optimal way to classify correctly. It uses Vapnik's Principle [8]. SVM is the classifier that classifies the object by drawing the hyperplane. Support Vectors are the weight vectors and the parameter of the linear model which is written down in terms of a subset of the training set.

Let us consider two classes with labels +1 and -1.

Let the sample be S = { ct, at }

Where if at = +1 then ct ∈ C1

Or if at = -1 then ct ∈ C2

It means that for label -1 we will consider the class C1 and for label +1 we will consider class C2. Thus in SVM while classification of the data, it automatically distinguishes the object into two different classes and will automatically assigns different labels for both the classes.

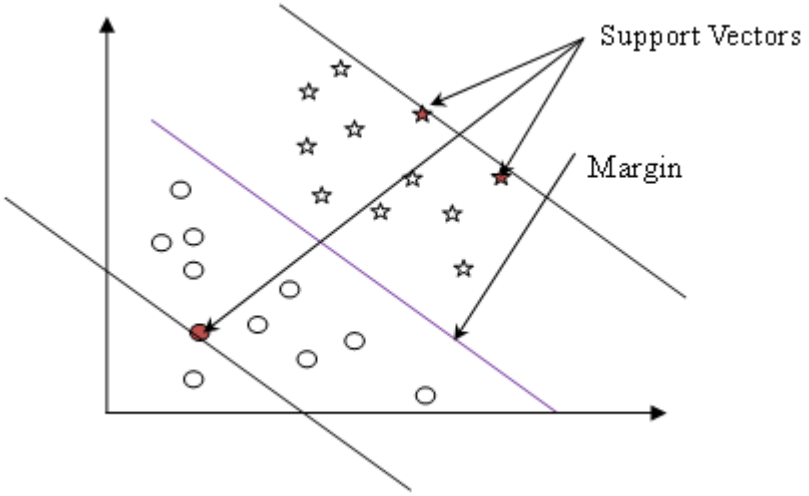

Fig3. *Support Vector Machine.*

The cases that are close to the boundary are the ones that provides knowledge extraction but the cases that are in the neighborhood of the boundary are the erroneous one.

5. NEURO FUZZY

Neuro fuzzy is an embedded hybrid technology which comprises of the benefits of both neural network and fuzzy logic. Hybrid Technology is the technology which is employed with various technologies to solve the problems efficiently.[9]

Basically there are 3 types of hybrid technologies:

- *Sequential Hybrid Systems:* Sequential Hybrid Systems(SHS) is the system where one by one technologies are added to increase the efficiency of the system
- *Auxiliary Hybrid Systems:* Auxiliary Hybrid Systems(AHS) is a system where the second technology processes the data provided by the first and is being provided for further use.
- *Embedded Hybrid Systems:* Embedded Hybrid System(EHS) is a system which provides fusion of various technologies i.e. various technologies are embedded together to increase the performance of the system.

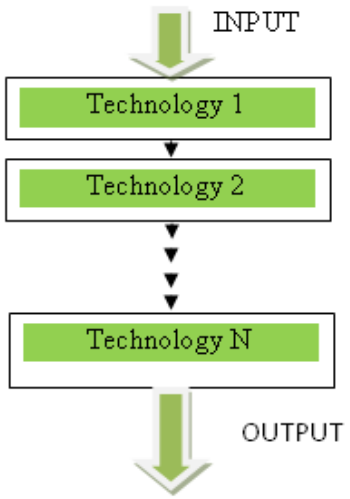

Fig4. *SHS*

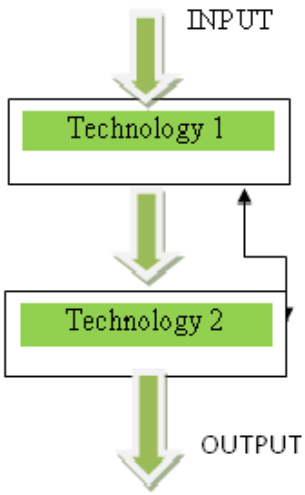

Fig5. *AHS*

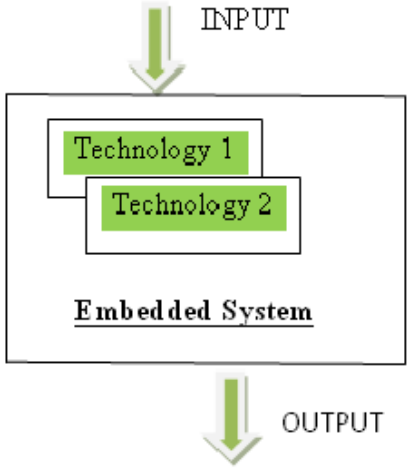

Fig6. *EHS*

Neural Network and Fuzzy Logic are two distinct methodologies but both of them have their own advantages and disadvantages.[7]

Neural Network (NN) consists of a group neuron which are interconnected to each other whereas Artificial Neural Network (ANN) comprised of artificial neurons i.e. the programs that shows the

properties of neurons. Thus NN comprises of various simple processing elements which communicates in an environment where there are huge amount of interconnections available and it works with various weights.On the other hand ANN which is subsequently referred as NN is sometimes called artificial as it helps in solving artificial intelligence problems but without creating a real biological model.

In crisp logic we normally have 2-valued truth values i.e. either true or false which can be represented by 1 or 0 but fuzzy logic have multi-valued truth values i.e. fully true, partly true , very true and so on which can be numerically represented by the range of 0 to 1. A fuzzy proposition can be represented as follows:

Let, F be the fuzzy proposition

T(F) be the truth values i.e. 0-1

Thus for the fuzzy set "F" the fuzzy membership value associated with fuzzy set P is treated as fuzzy truth value T(F) i.e.

$T(F)=P(x)$ where $0 \leq P(x) \leq 1$.

For example:

F: Rahul is innocent

If T(F)=0.7 means that F is partly innocent

If T(F)=1means that F is absolutely innocent.

Thus the integration of NN and Fuzzy Logic provides the capabilities that is beyond their individual capabilities. To adapt to an uncertain environment it increases the networks flexibility, expressiveness and adaptiveness.

6. RESULTS

The results of SVM is being compared with LEM2 [1] and Wang & Mendel Algorithms [3]. The accuracy of SVM is calculated using Weka. As we know Weka doesn't have SVM in it, so it is being integrated by creating classpaths with Weka. Diabetes dataset is already present in Weka.

It is being observed that SVM without pre-processing techniques givesless accuracy than LEM2 and Wang & Mendel Algorithms but after pre-processing the result drastically changes and increases alot. The results using SVM are shown below:

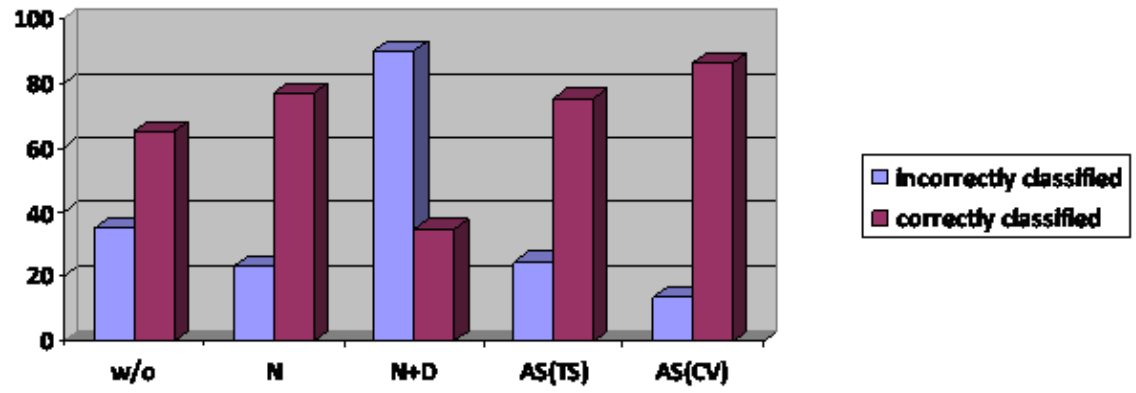

Fig7. *Graphical comparison of SVM with and without Pre-processing for PIMA dataset*

Where

w/o = without pre-processing.

N = With Normalisation.

N+D = with Normalisation and Discretization.

AS(TS) = Attribute Selection with training set.

AS(CV) = Attribute Selection with cross validation.

The results are clearly being shown below in tabular form below:

Table1. *Comparison of accuracy of SVM with and without pre-processing*

CLASSIFIER	Correctly Classified Instances	Incorrectly Classified Instances	ROC(Area under ROC)
Without Pre-processing	65.1042 % (500)	34.8958 % (268)	0.497
Normalisation	76.9531 % (591)	23.0469 % (177)	0.7044
Discretization	75.5208 % (580)	24.4792 % (188)	0.7133
After attribute selection (using cross validation)	75.5208 % (580)	24.4792 % (188)	0.7133
After attribute selection (use training set)	86.4583 % (664)	13.5417 % (104)	0.8319

In attribute selection for 3 attributes i.e. Plas, mass, age we have the following results:

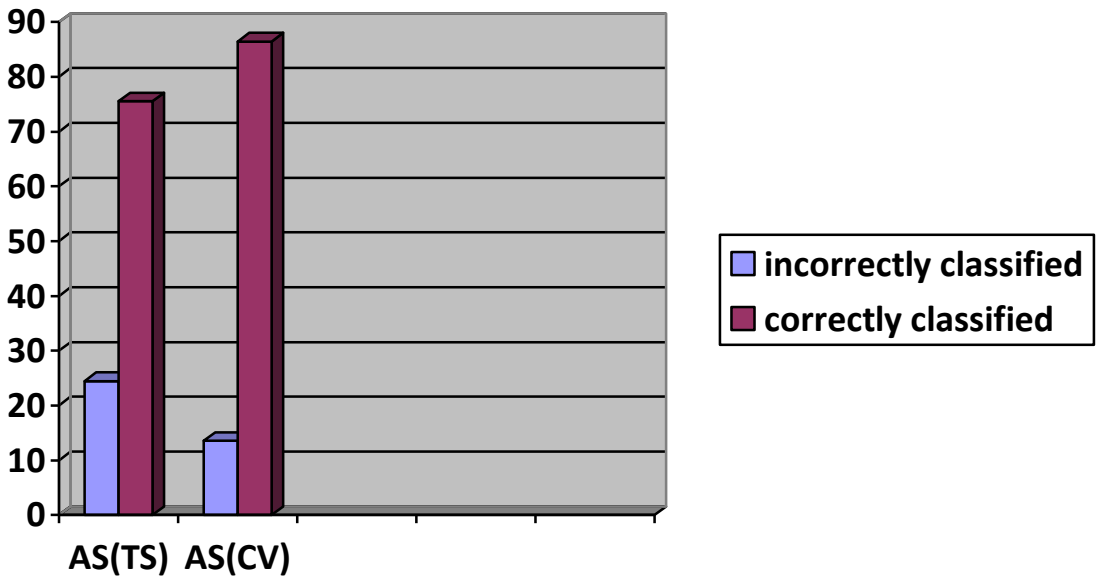

Fig8. *Graphical comparison of SVM for attribute selection with 3 attributes.*

The results are clearly being shown below in tabular form below:

Table2. *Comparison of accuracy of SVM with attribute selection for three attributes.*

CLASSIFIER	Correctly Classified Instances	Incorrectly Classified Instances	ROC(Area under ROC)
After attribute selection (using cross validation)	75.9115 %	24.0885 %	0.7120
After attribute selection (use training set)	79.0365 %	20.9635 %	0.7446

The ROC Curves for the following results are shown below:

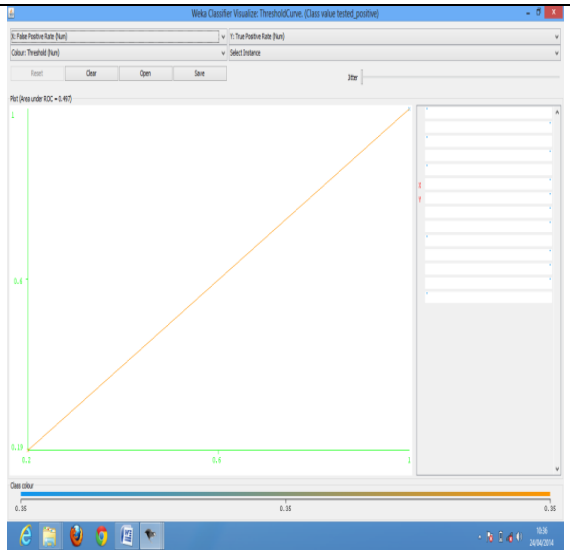

Fig9. *ROC curve of PIMA dataset without pre-processing for SVM*

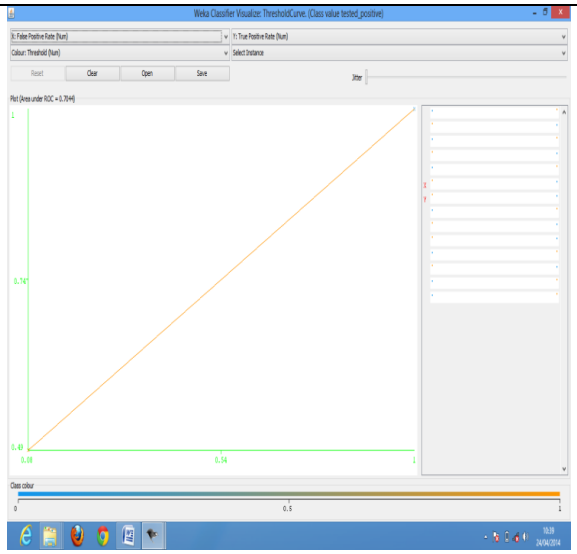

Fig10. *ROC curve of PIMA dataset with normalization for SVM.*

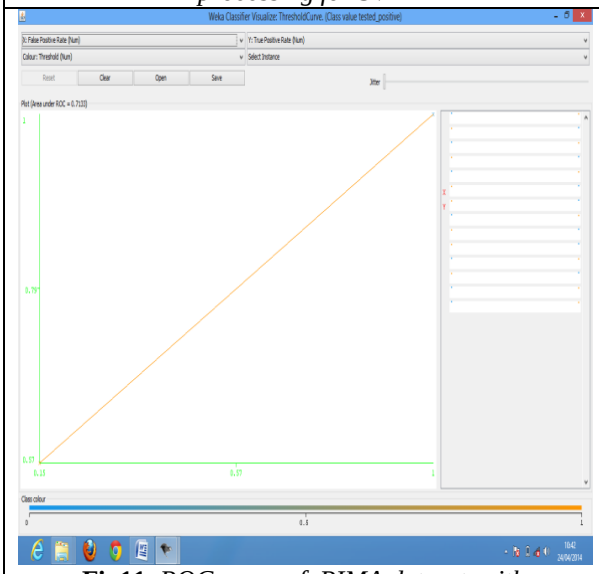

Fig11. *ROC curve of PIMA dataset with discretization for SVM.*

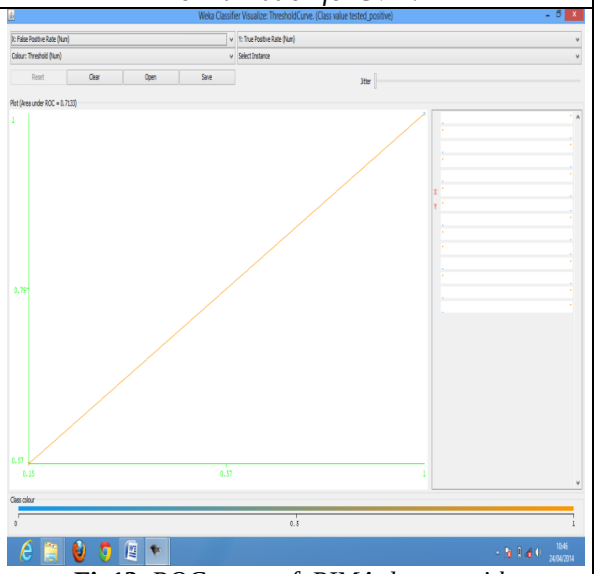

Fig12. *ROC curve of PIMA dataset with normalization and discretization for SVM.*

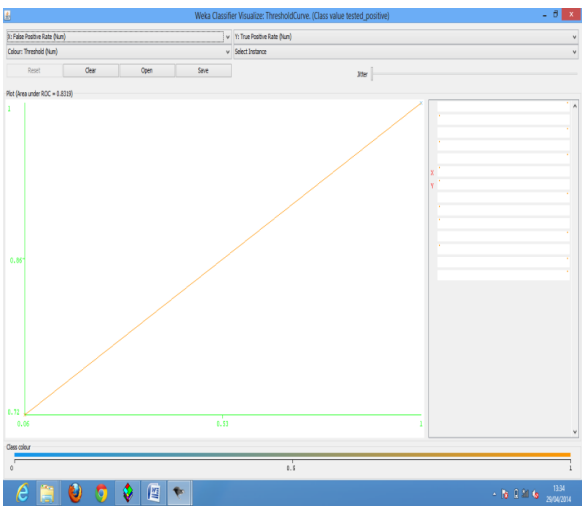

Fig13. *ROC curve of pre-processed PIMA dataset with attribute selection (using training set) for SVM.*

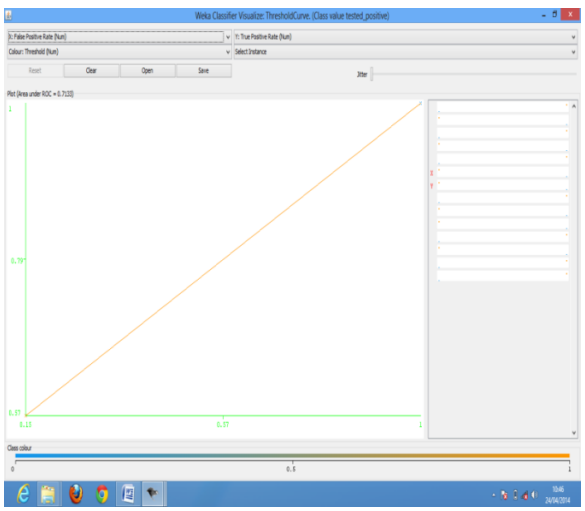

Fig14. *ROC curve of pre-processed PIMA dataset with attribute selection (using cross validation) for SVM.*

Initially Wang and Mendel algorithm is compared with LEM2 algorithm[1] which is implemented in LERS-92 .

The results of the above comparison shows that the accuracy of LERS with 8 attribute selection is 82.3% whereas for Wang & Mendel algorithm it is 79.9%. Though LERS is more efficient than Wang & Mendel Algorithm but in our paper we have shown that with SVM we can increase the accuracy further if the dataset is properly pre-processed.

7. FUTURE WORK

Till now what we have shown is the detection of the diabetes disease with as much accuracy as possible. But the accuracy can further be increased by the following ways:

- If we consider the rough fuzzy set concepts with SVM.
- After checking with PIMA dataset we can further go for checking the after effects of diabetes by Diabetes Retinopathy check which can highlight the severity of the disease more.
- The images used for Diabetes Retinopathy check can be pre-processed further for more accuracy which can easily be done using Matlab.

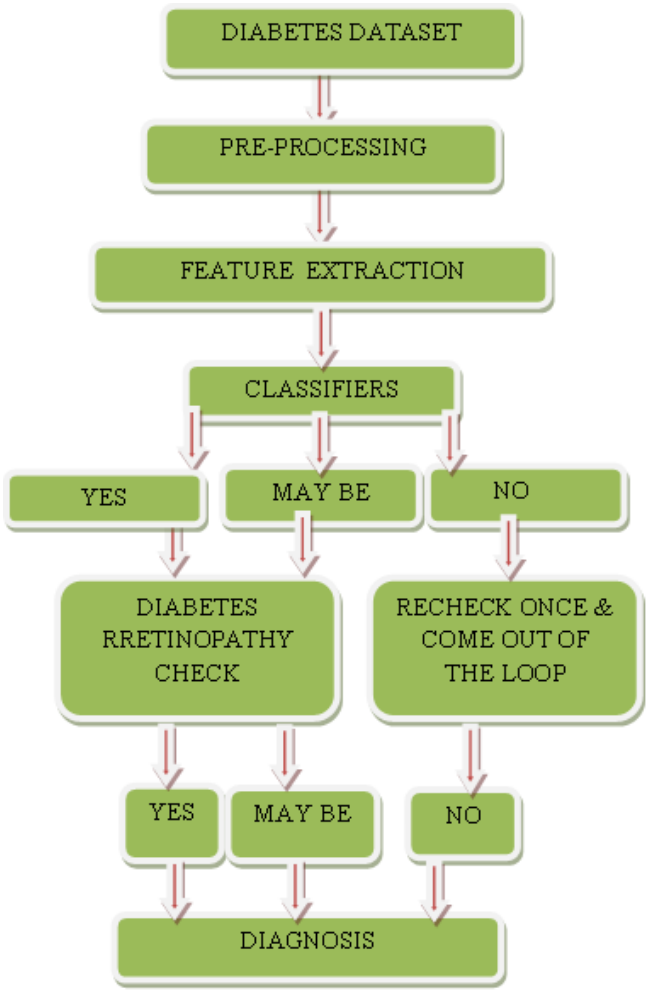

Fig15. *Block diagram showing the future scope for more accurate diagnosis of diabetes in consideration of its after effects.*

The system for the above requirement is shown below:

Step1. The diabetes dataset is pre-processed first.

Step2. The pre-processed dataset undergoes the feature extraction procedure.

Step3. The dataset is then classified and according to the accuracy the patient is detected of either having diabetes, doesn't have diabetes or may have diabetes.

Step4. If Step 3 gives "NO" means the patient is not detected of diabetes and is asked to go through the procedure once again and come out of the system.But if it gives "YES" or "MAY BE" we need to check the after effects of diabetes which is being checked by diabetes .

Step5. After checking for diabetes retinopathy we come to the conclusion of the severity of the disease and accordingly asked them for further treatment.

The procedure for the same are:

8. CONCLUSION

This paper helps in efficient diagnosis of Diabetes. For this SVM is used as a classifier and various pre-processing techniques are used which increases the accuracy level to a great extent. The various ROC curves are also being shown here which supports the above statement. SVM is compared with LEM 2 and Wang & Mendel Algorithm which gives more accuracy. We can further extend our work by checking the after effects of diabetes which is the future scope of our paper.

ACKNOWLEDGMENT

The author would like to thank Prof. D. P. Gaikwad for his help, Prof. N F Shaikh for helping with the concepts, and the reviewers for their valuable suggestions and comments.

REFERENCES

[1] Diabetes mellitus: http://en.wikipedia.org/wiki/Diabetes_mellitus
[2] Robert Nowicki, "Rough Neuro-Fuzzy Structures for Classification With Missing Data," Member, IEEE., PART B: CYBERNETICS, VOL. 39, NO. 6, DECEMBER 2009
[3] J. W. Grzymala-Busse, "An overview of the LERS1 learning systems," in Proc. 2nd Int. Conf. Ind. Eng. Appl. Artif. Intell. Expert Syst., 1989, pp. 838–844.
[4] J. W. Grzymala-Busse, "LERS—A system for learning from examples based on rough sets," in Intelligent Decision Support: Handbook of Ap- plications and Advances of the Rough Sets Theory, R. Slowinski, Ed. Dordrecht, The Netherlands: Kluwer, 1992, pp. 3–18.
[5] L. X. Wang, Adaptive Fuzzy Systems and Control. Englewood Cliffs, NJ: Prentice–Hall, 1994.
[6] L. X. Wang and J. M. Mendel, "Generating fuzzy rules by learning from examples," IEEE Trans. Syst., Man, Cybern., vol. 22, no. 6, pp. 1414– 1427, Nov./Dec. 1992.
[7] By Obi J.C., Imianvan A.A. University of Benin, Benin City. Nigeria, "Interactive Neuro-Fuzzy Expert System for Diagnosis of Leukemia" Online ISSN: 0975-4172 & Print ISSN: 0975-4350, Volume 11 Issue 12Version 1.0 July 2011
[8] Vapnik. V, ―Statistical learning theory Wiley‖ , New York, 1998
[9] Neural Networks, Fuzzy Logic and Genetic Algorithms Synthesis and Applications by S Rajasekaran and G A VijayalakshmiPai.
[10] H.Liu and H Motoda. Feature Selection for knowledge discovery and data mining: Boston: Kluweir Academic Publishers 1998
[11] Rajiv Tiwari "Correlation based Attribute Selection using Genetic Algorithm" Volume-4-No.8, August 2010.
[12] Michei D, Speigelhalter D. J and Taylor CC."Machine Learning Neural and Statistical Classification"1994 .Editors: England: Ellis Horwood Limited, pp 1-5.
[13] Lena Kelin Westin "Pre-processing Perceptrons" Department of Computing Science Umea University. 2004, Print and Media, Umea University, ISSN-0348-0542 ISBN 91-7305-645-6.
[14] P Jeatrakul and K W Wong School of Inf. Tech., Murdoch University, Murdoch Australia "Comparing the performance of different neural networks for binary classification problems" 2009:National Conference: Natural Language Processing.
[15] "http://www.cs.utsa.edu/~bylander/cs6243/bayes-example.pdf"by Bylander.

AUTHOR'S BIOGRAPHY

Miss. Nilufar Zaman Graduated in Computer Science and Engineering. from ICFAI University Tripura, India and Pursuing M E in Computer Engineering from AISSMS College of Engineering Pune, India. Currently working as Assistant Professor in Department of Computer Engineering in Modern Education Society's College of Engineering Pune.