Journal of Theoretical and Applied Information Technology
10th May 2016. Vol.87. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A COMPUTATIONAL HYBRID MODEL WITH TWO LEVEL CLASSIFICATION USING SVM AND NEURAL NETWORK FOR PREDICTING THE DIABETES DISEASE

[1]NASIB SINGH GILL, [2] POOJA MITTAL

[1] Professor, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana, India

[2] Assistant Professor, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana, India
Email: **mpoojamdu@gmail.com**

ABSTRACT

Data Mining is a collection of number of computational approaches. These approaches are used to develop Knowledge inference systems by identifying the hidden and convincing patterns from the input data. The aim of this study is to propose a computational Hybrid Prediction Model (HPM) for efficient diabetes prediction. The Pima diabetic dataset is used as the data source, obtained from the University of California, Irvine (UCI), the machine learning repository. At first stage of the proposed HPM the filtration feature selection method of MATLAB is used for selecting the most discriminatory predictors, reflecting the possibility of diabetes occurrence. At second stage, a two-layered classification is applied on the filtered data, by combining Support Vector machine (SVM) and Neural Network, to enhance the overall recognition rate of the model. The proposed hybrid model gained 96.09% of overall accuracy. The comparative study is also conducted and it is evident that the proposed model had obtained the significant classification accuracy. The accuracy rates achieved by many researchers in the past, on the same data set, ranges from 59.4% to 92% of accuracy. Further, for validating and evaluating the results, Recognition rate, Mean Absolute Error (MAE) and Receiver Operating Characteristics (ROC) performance measures are used. This research can be helpful to the physicians for predicting and detecting the diabetes at an early stage, efficiently.

Keywords: *Support Vector Machine, Neural, Diabetic, Mean Absolute Error, Accuracy Rate*

1. INTRODUCTION

Diabetes is one of the most common chronic, debilitating complex and intricate disease found in all age groups ranging from infants to old age people in all populations. Insulin is a vital and an essential component of our human body metabolism which is used for its proper growth. Diabetes may be caused due to inadequate production of insulin or inappropriate utilization of the produced insulin. In any of the situation, plenty of glucose is available in the blood stream which human body could not utilize it properly [1], resulting in the diabetes. Broadly, diabetes is categorized into two categories: Type 1 and Type 2. Due to insulin deficiency, Type 1 diabetes is caused, also known as juvenile diabetes. The only solution to this type of diabetes is to inject the required amount of insulin as supplement in the patient's body. Whereas: Type 2 diabetes is more commonly known as Adult-onset diabetes. It is the most common type of diabetes, which usually develops at the age of 40 and older.

According to Diabetes Atlas, it is estimated that about 194 million people worldwide are suffering from diabetes and it is estimated that this will increase to 333 million or 6.3% by 2025 [2]. Type 2 diabetes is spreading like an epidemic and constitutes about 85% to 95% of all diabetes in developed countries and even higher in developing countries [3]. Diabetes is taking shape of a serious global problem which is treated as a red alert to human life. Diabetes is associated with severe complications and a number of adverse effects of all kinds including loss of eye sight, poor immune system, heart complications, skin ailments, kidney disease, nerve damage and even blood vessel damage. There is no absolute solution and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

treatment for diabetes, which can cure the diabetes completely.

Thus, it becomes necessary to predict the disease at an early stage to prevent its malicious effects on the human health.

Though, it is nearly impossible to cure the diabetes completely, yet it can be controlled so that a patient can lead a healthy life. Early prediction of disease is crucial and is one of the major research area thrust. For an early and efficient diagnosis, a computational intelligent tool must be used for assisting the physicians. These reasons motivated us to pursue our present research for this cause. The objective of this paper is to propose a hybrid prediction model as an intelligent tool for achieving the better recognition rate.

Related Work

SVM [4] is a versatile and powerful classifier. Though, SVM is having strong mathematical foundation and promising experimental results, yet the performance of SVM as classifier, is adversely affected, if applied on large sized data set. SVM is also not favorable for heterogeneous data with missing and noisy data. Core Vector Machine (CVM) was proposed by Tsang et. al.[5] as a decomposition algorithm. Several works had been proposed for the analysis of medical data for diabetic patients, focusing on exploiting the classification techniques [6]. The k-means algorithm was integrated with C 4.5 decision tree, the artificial neural network (NN) and 2D graphs by Mohamudally N. & Khan D. M. [7] to predict, classify and visualize the medical diabetic dataset. Three prediction models were compared by Meng at. El. [6] comprising of logistic regression, decision tree and artificial neural networks (NN) for diabetic patients classification. Su et. al. [8] applied four primary classifiers: the neural network, decision tree, logistic regression and rough sets to enhance the accuracy ratio of the model. The hierarchical clustering and adaptive clustering were utilized [9] to select the reduced sized data sample, later the data sample was fed into single SVM for fast training. Clustering Support Vector Machine (CSVM) has shown remarkable results for small sized data [10]. But when applied on big data, the derived results were not optimum. Later MLSVM was proposed to overcome the weaknesses of the previous models. A Multi level Support Vector Machine (MLSVM) was proposed [11] to predict the clinical charge profiles for patients diagnosed with chronic disease, by organizing the given data into tree of similar clusters for creating effective partitions of the dataset to improve the performance in terms of accuracy and speed of the training process. Authors employed multiple SVM's, each of which was applied on local data distribution. Other popular existing SVM models are Core Vector Machine (CVM) [5] and Adaptive Clustering based SVM (ACSVM) [12].

The above survey of the related work clearly indicates that intelligent techniques like SVM and Neural suffers from number of constraints that affect the overall performance of the system. Thus, the objective of this research is to propose more efficient hybrid approach for diabetes prediction. The objective of this section is to highlight the gaps present in the related literature and to discuss the earlier prediction models along with their strengths and shortcomings to frame the research problem for the current study.

2. BACKGROUND

Many researchers worked with subjugated data mining techniques to analyze the medical data by addressing varied pathologies and facets of wide range of diseases [13] by applying individually as well as in the hybrid form. A hybrid prediction model is proposed here, comprising of the following multiple stages including data preparatory, data mining techniques (to perform the actual classification) and finally the validation and verification by using the performance evaluators. In this section, the integral components of the proposed model are described.

2.1 Data Mining Tool

For the implementation of the proposed Hybrid Prediction Model (HPM), MATLAB (http://in.mathworks.com)[14] tool is used. Main reason for using MATLAB as an analysis tool in the present study is its strong mathematical foundation, which can well support the mathematically influenced SVM and NN methods, used in this model. Slow speed of SVM can also be accustomed by using fast analysis MATLAB tool. The present model is logically staged into three major phases, for optimizing the derived results from the data mining techniques. MATLAB is used at every stage of the proposed model. The filtration method is used at stage 1 to obtain the most appropriate reduced dataset. At stage 2, the SVM and Neural approach are applied

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

on the reduced classified dataset, to enhance the classification rate. Lastly, to validate and to analyze the obtained results, the graphical assets of MATLAB are deployed.

2.2 Data Mining Methods

Wide range of data mining techniques is available, fully capable of extracting meaningful and useful hidden patterns from large data bank. Among incredible applications of data mining, prediction is the most admired work area. Predictive data mining techniques are originated from different research areas and apply diverse modeling approaches depending on the type of application. Among all predictive techniques, SVM and Neural Network approach are used in the proposed model, described in this section.

2.2.1 Neural Network

An artificial neural network (ANN), also called as neural network (NN), is a mathematical or computational model based on biological neural networks. As ANN is designed on the grounds of human biological systems, it learns by an example. Learning is defined as an adjustment to the synaptic connections that exists between the neurons. It consists of an interconnected group of artificial neurons and processes information by using a linking approach to perform computations. An ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Neural networks are remarkably capable of deriving the desired information from complicated or imprecise data, for extracting the patterns and detecting the trends that would be otherwise very complex or nearly impossible for either humans or for other mining techniques. Neural networks are best example of adaptive learning which is self organized, designed for performing real time operation with high degree of fault tolerance.

2.2.2 Support Vector Machines (SVM)

In comparison to neural networks we may describe SVM as a feed-forward neural net with one hidden layer. The main building blocks of SVM's are structural risk minimization, originating from statistical learning theory which was mainly developed by Vapnik V. [4] , non-linear optimization and duality and kernel induced features spaces, underlining the technique with an exact mathematical framework. Several extensions to the basic SVM have been introduced, e.g. for multi-class classification as well as regression and clustering problems, making the technique broadly applicable in the data mining area. All SVM algorithms can be broadly classified into two classes as decomposition algorithms and second as selective sampling techniques [15]. Despite of its strong mathematical foundation and promising results, SVM is not much supportive for large scale mining tasks, when applied independently. The data set used in the present study is described in the next section.

2.3 Pima Dataset Description

The reliability and accuracy of any prediction system largely depends on the dataset. In this work, an authenticated Pima dataset [16] is used. This dataset contains the examined data for Pima Indians generated by the National Institute of Diabetes and Digestive and Kidney Disease. This institution has the name to predict various diseases at high recognition rates. The institution has defined several experiments to reduce the risk factor and to identify the diabetes over the patients. The present work used the same dataset to improve the prediction performance. The dataset description is shown in the following table:

Table 1: Description Of The Pima Dataset Extracted From UCI Repository.

Property	Description
Dataset URL	http://archive.ics.uci.edu/ml/datasets/Diabatic+Disease
Number of Attributes	9
Number of Records	768
Attribute Names	Preg, Plas, Pres, Skin, Insu, Mass, Pedi, Age, Result
Class Attribute	1
Process Attributes	3
Numerical Attributes	8

The basic statistical information regarding the dataset source is defined in the table. But this information is not enough to describe the attribute set. To have complete understanding of the dataset, it is required to have complete information about each and every attribute. The description of individual attribute is defined here. The complete dataset is defined with 9 attributes out of which 8 are considered as the information attributes and one is considered as the result attribute. The description of these dataset attributes is shown in table 2.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2: Attribute Set Description Of Pima Dataset With Specific Valid Range Respectively

Attribute	Description	Unit	Format	Range
Preg	Defines the No. of Pregnancies	No. of times	Numeric	0-17
Plas	Plasma Glucose Cont.(2 Hr.)	Mg/dl	Numeric	0-199
Pres	Blood Pressure	mmHg	Numeric	0-122
Skin	Skin Fold Thickness	Mm	Numeric	0-99
Insu	Serum Insulin (2 Hrs.)	Mu U/ml	Numeric	0-846
Mass	Body Mass	Kg/m2	Numeric	0-67.1
Pedi	Diabetes Pedigree Function	-	Numeric	0.078-2.42
Age	Age of Patient	-	Numeric	21-81
Result	Class of Disease	-	Nominal	-

For better result analysis, the data set is reduced by performing the vertical partitioning, by applying filtration method. The objective of vertical partitioning is to find out the most appropriate data subset for further analysis. The attribute selection step is described in detail in the next section.

.2.4 Feature Selection Technique

The feature selection process, also known as attributes selection or relevance analysis, is a vital step of the Knowledge Discovery from Data (KDD) process, responsible for ensuring the quality of data. Searching and evaluation are two stages of any feature selection method. In the present study, filtration method is used for achieving mRMR(minimum Redundancy Maximum Relevancy)at data level, described in the following flow statement:

Dataset→Original Feature Set→Subset Evaluation→ Value Deviation Selected Attributes

Filtration technique is one the most common and yet popular approach used for attribute selection. It can be applied on all types of data to reduce the data size significantly. Due to its statistical background, it produces high accuracy.

3. PREDICTIVE CLASSIFICATIONS ON PIMA DIABETIC DATASET IN MEDICINE: RELATED REVIEW

In the last few years, a great potential has been witnessed in the field of predictive data mining to originate clinically relevant models from the given patient's data. Main objective behind such models is to infer hidden patterns from the given set of data to provide decision support to medicine decision makers. In past, many renowned researchers contributed their valuable work in this clinical domain, by taking Pima diabetic dataset as an input data. Following table represents the different accuracy levels attained by the respective researchers while working with all types of classification techniques when applied on Pima diabetic data:

Table 3: Recognition Rate Analysis Of Various Models For Pima Diabetic Dataset

S.No	Method	Accuracy (%)	Reference
1.	k-NN	67.6	Michie
2.	CART	74.5	Michie
3.	Naïve Bayes	73.8	Michie
4.	C4.5	73	Michie
5.	ARTMAP-IC	81	Carpenter et. al.
6.	Neural Network	75.4	Bioch et. al.
7.	Bayesian Approach	79.5	Bioch et. al.
8.	Hybrid model	84.24	Humar et. al.
9.	GDA	82.05	Polat et. al.

It can be observed from table 3, an extensive work had been carried out in this specific domain by many researchers. Different accuracy levels were obtained when different techniques were applied on the same data. The hybrid approach is also proposed by many researchers that outperformed from individual techniques. In the present study, a hybrid model is proposed with an aim of improvising the predictive accuracy and minimizing the error rate, described in the next section.

Journal of Theoretical and Applied Information Technology
10th May 2016. Vol.87. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. PREDICTIVE PROCESS FOLLOWED IN THE PROPOSED STUDY

Data mining is a vital phase of KDD process, which aimed at extracting useful hidden patterns from the given dataset, capable of assisting the decision making process. To obtain the optimal results from the data mining system, number of pre-analysis phases was suggested by popular researchers. Here, we have tried to follow the guidelines framed by renowned researchers for better predictive data mining.

4.1 Problem Definition

Main objective of this present study is to investigate the impact of patient's characteristics on the occurrence of diabetes. This analysis is also focused to predict the order and degree of relevancy of each and every explanatory variable on the diabetes incidents. For this purpose, the present work is conducted on the predictors of Pima Indian women's 768 randomly selected observations, available at UCI repository as secondary data source. The dataset is gathered under few constraints like all the observations were of Pima Indian female patients, having at least 21 years of age. The dataset under consideration is described in section 3. In the present study, a hybrid model is proposed to decipher all the data constraints while achieving the enhanced recognition rate than earlier prediction models, capable of making reliable predictions and can help physicians in improving their prognosis, diagnosis and treatment planning process.

4.2 Proposed Model

After framing the research problem, a predictive prototype is defined for hybrid model comprising of the following iterative phases progressing sequentially in attaining the defined goal, as shown in the Figure 1.

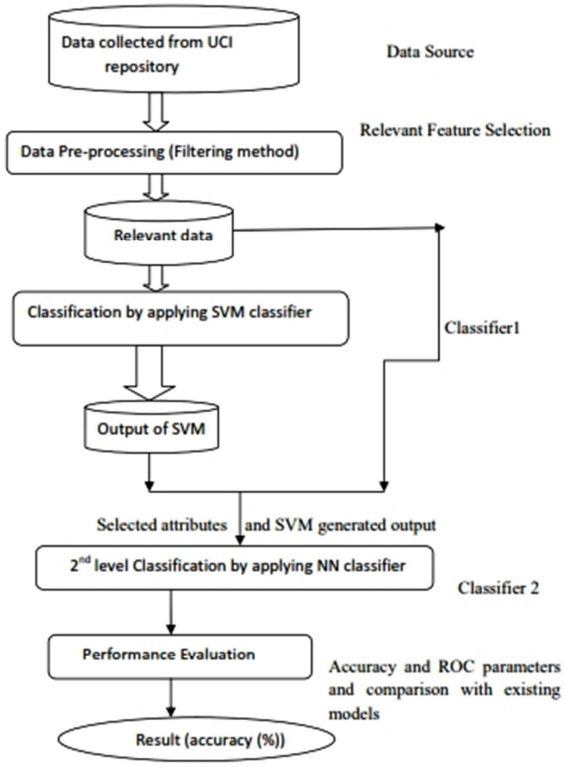

Figure 1: A Computational hybrid model with two level Classification using SVM and Neural Network

Following is the description of the HPM in terms of its integral stages:

1. At stage 1, the reliable and relevant Pima diabetic data is collected provided by the medical institutes and repositories. The relevant analyzers which may affect the Type-2 diabetes and views of respective physicians are collected.
2. Data pre-processing acted as the second stage in which data filtration method is implemented to curtail the nappropriate and inconsistent data from the source data, resulting in significant data reduction.
3. At third stage, SVM classifier is implemented on the refined dataset to predict the diabetes occurrence.
4. To improve the accuracy of prediction, at fourth stage NN method is applied on the output of SVM and the raw inputs from the dataset.
5. At fifth stage the performance of the proposed model is evaluated by measuring the recognition rate and by plotting ROC.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6. At final stage, the outcome of present model is validated and compared with the performance of existing models.

In the present paper, three staged prediction model is proposed, to achieve better recognition rate. All the above mentioned stages are explained in the next section.

4.2.1 Data preparation using filtration method

At first stage, the Pima diabetic data is collected from the reliable UCI repository available as secondary data source with 768 observations, described in terms of nine distinct features/attributes. At first, the dataset is arranged in to two subsets with respect to class values (zero representing non diabetic and one representing diabetic patients) of class label attribute. Then, the data is horizontally partitioned by dividing the given dataset into training data (70% of dataset) and testing data (30% of data) samples by using holdout of cvpartition function available in MATLAB:

```
cvpartition(complete_data,'holdout
',230)
```

After horizontal partitioning, vertical partitioning is performed by selecting most appropriate and relevant features among all nine given attributes. Number of different possibilities is checked by varying the value of standard deviation for selecting the most relevant feature subset in data filtration stage:

```
maxdev = N; (initialzing the
standard deviation value)

opt =
statset('display','iter',...'TolFu
n',maxdev,...'TolTypeFun','abs');

inmodel =
sequentialfs(@critfun,X,Y,...'cv',
'none',...'nullmodel',true,...


'options',opt,...'direction','forw
ard');
```

Finally, the given data is reduced and the most relevant subset of attributes is extracted successfully for classification. After applying the filtration method, data size is significantly reduced that will indirectly enhance the recognition rate and will reduce the time required.

4.2.2 SVM

The current research work proposes the novelty of a hybrid scheme, which shows that the support vectors when combined with neural network offers a high level of accuracy in the field of bio genetics prediction. These vectors are produced by applying linear and radial basis kernel function on the given data. The support vector machine in the current work uses the ideal hyper plane concept. In the present study, support vectors are evaluated by applying selective feature/ attributes on the entire dataset. An Alpha Lagrange multiplier is used for creating a biased hyper plane for best classification along with selective vectors that are distinct in that hyper plane. The resultant output of this classifier is a structure network which is used for further classification.

4.2.3 Neural Networks

The neural network is fed with the attributes extracted by filtration technique from the Pima diabetic data and also with the output generated by the support vector classification as another input. The neural network used in the current work uses the feed forward architecture that best suits the predictive application:

```
network1 = feedforwardnet([20 20
30 20],'traingd');

network1.TrainParam.max_fail=100;
```

The back propagation training involved in the network reduces the error to a significant low level. The neural network training is performed on selective features, which are extracted by applying filtration method on the complete set of attributes. The imperative points which are focused in the neural network are Performance, Gradient and The maximum failure (Validation Checks). Once the neural training network is complete, it can go for any number of arbitrary tests. To assure the performance of this trained network, values of various parameters are altered, to achieve the highest degree of robustness.

4.2.4 Performance Evaluation

It is always suggested to validate and verify the constructed and trained model to assure its reliability. To ensure the authenticity of testing, the test data is considered separately from training data. To obtain the test data, cvpartition( ) method of MATLAB is used and 230 observations are holdout as testing set. The proposed model achieved 96.09% recognition rate with reasonably minimum error rate, when tested on testing

Journal of Theoretical and Applied Information Technology
10th May 2016. Vol.87. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

observations. Maximum value is also set for the system tolerance in terms of validation checks, as

```
network1.TrainParam.max_fail=100;
```

Here, network1 is the network initialized as feed forward network structure and its maximum tolerance against failure during testing is set to 100. The network will stop testing, if the validation check exceeds the tolerance value. Another performance measure used in this study is the ROC. This curve is generated and found that True positive has higher ratio than wrongly classified observations.

Accuracy or Recognition Rate:

A single prediction, results in four possible outcomes. The true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values are generated. The meaning of these four outcomes is provided in the following section: True Positive: Correctly predicted observation as Yes, True Negative: Correctly predicted observation as No, False Positive: Wrongly predicted observation as Yes when it was No, False Negative: Wrongly predicted observation as No when it was Yes. Based on these four parameters, the most popular performance evaluator, the Accuracy can be measured as:

$$\text{Accuracy} \quad [17] \quad = \frac{TP + TN}{TP + TN + FP + FN}$$

The suggested model achieved 96.09% of accuracy by correctly classifying 96 observations from the given set of 100 observations of testing set.

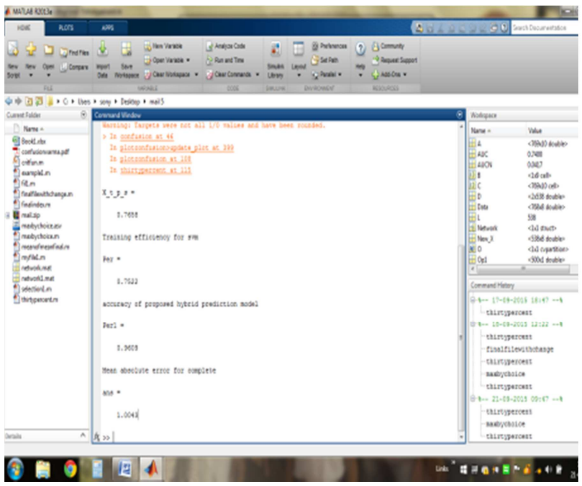

Figure 2: Result Analysis Of The Proposed Model In Terms Of Accuracy Measure

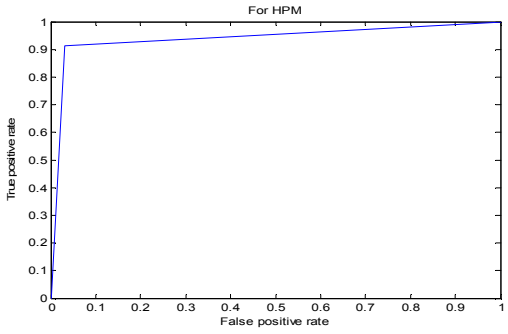

Figure 3: ROC Diagram Based On The True Positive And False Positive Values

As shown in the result window in Figure 2, the proposed hybrid model achieved the 96.09% recognition rate whereas when the same techniques were applied separately on the same Pima data set , less accuracy rate was obatined. When the given data is classified with SVM in the present study, (71% to 78%) accuracy is achieved. Nueral Network approach obtained (75% to 78%) of recognition rate. It is also concluded that feature selection stage significantly not only improve the accuracy of the proposed model but also reduces the required time significantly.

Receiver Operating Characteristics (ROC)

ROC is a useful and precise way for organizing the classifiers and visualizing their quality. It is a useful tool for interpreting medical test results and other decision making machine learning approaches [18]. The *receiver operating characteristic* is a metric used to check the quality of classifiers. For each class of a classifier, roc applies threshold values across the interval [0,1] to the outputs. For each threshold, two values are calculated, the True Positive Ratio (the number of outputs greater or equal to the threshold, divided by the number of one targets), and the False Positive Ratio (the number of outputs less than the threshold, divided by the number of zero targets). The results of this function can be visualized with plot roc command (mathworks.com),, as explained:

[tpr,fpr,thresholds] = roc(targets,outputs)

where tpr are true positive records, fpr are false positive records with respect to the value of threshold provided (http://in.mathworks.com). The ROC diagram depicts the ratio of True Positive Rate to the False positive rate obtained by the present hybrid prediction model for Pima diabetic

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

data set, as shown in Figure 3. By varying the value of standard deviation, used for the feature selection, different values for different performance measures like accuracy, ROC and Area Under Curve (AUC) are derived, as presented in the following table:

Table 4: Performance Evaluators For The Proposed Model With Different Values Of Deviation With Comparative Analysis Of The Proposed Model With The Individual Mining Techniques In Terms Of MAE, ROC And AUC

Max Deviation	NN Accuracy %	SVM Accuracy %	Hybrid Accuracy %	MAE	ROC	AUC
0.2	77.70	71.74	93.48	1.04	0.13	70.75
0.5	78.44	74.35	90.87	0.91	0.13	72.88
0.9	76.39	78.26	93.04	0.93	0.18	76.42
1.0	76.58	75.22	96.09	0.10	0.19	74.00
1.5	75.28	75.65	89.57	0.91	0.15	74.04
2	78.81	73.48	92.61	0.99	0.15	73.90
2.5	78.07	72.61	91.74	1.05	0.15	71.70

It is clearly evident from the table 4, the present model gives the maximum accuracy when the feature selection is performed with deviation value equal to 1.0. The performance of the system decreases when either the deviation value is increased or decreased. Moreover, the minimum error rate is achieved equal to 0.10. In the next section, validation of derived results is presented by comparing the present model with already existing models.

4.2.5 Validation with the existing models

In this study, the proposed model is evaluated on the basis of accuracy or recognition rate. Recognition rate is defined ratio of correctly classified observations and total number of observations. A model is proposed for diabetic prediction by Humar & Novruz [19] and achieved an accuracy of 84.24% by applying 10-fold cross validation. Many researchers who worked on the same Pima data set had obtained the recognition rate from 78.04% to 81%. These results are tabulated in the table 5:

Table 5: Validation Table For Various Prediction Models For Pima Diabetic Data

Method	Accuracy (%)	References
Hybrid Prediction Model	**96.09**	**Present Study**
Hybrid Prediction Model	92.38	Patil et al. (2010)
Hybrid Model	84.5	Humar Kahramanli (2008)
Logdisc	77.7	Statlog
Gaussian fuzzy decision tree	75.8	Varma et al. (2014)[20]
Least square SVM	82.05	Kemal Polat et al. (2008)
Fuzzy based ACO(Ant Colony Optimization)	84.21	Beloufa et al. (2013)

It can be witnessed from the previous studies; a wide range of accuracy was achieved when different approaches were applied on the same data set. The 92.38% of accuracy was achieved by the HPM suggested by Patil et. al.[3] when k-means clustering and C4.5 decision tree algorithm were applied. Based on the results presented in table 5, the proposed model outperformed other models, by gaining the highest accuracy rate of 96.09% in predicting the disease.

5. RESULT ANALYSIS AND DISCUSSIONS

The accuracy performance evaluator is one of the most desired measures among all other quality measures. Based on this accuracy value, other performance measures can be derived like specificity and sensitivity. The proposed hybrid model obtained the accuracy of 96.09%. In the present model, the max deviation feature selection method is applied on the train data set, to reduce the data size significantly. Then SVM is applied on this extracted dataset as a classifier for analysis. Finally, NN approach is implemented collectively on the initially extracted attributes and output of SVM as input data. The output fetched after applying both classifiers is 96.09% of classification accuracy. ROC is another

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

performance measure, used in this study, which proved that present model outperformed the previous models with same data set. It is also proved that hybrid model attain better accuracy rate than the individual techniques when applied in the same environment. The different classification accuracies obtained by applying SVM with and without feature selection, NN with and without feature selection and the proposed hybrid model on the same Pima dataset, are presented in the following table:

Table 6: Findings Of The Present Study In Terms Of Accuracy (%) Derived With SVM, NN And HPM With And Without Feature Selection.

Classification Method	Accuracy Rate (%)
SVM without Feature Selection	74
SVM with Feature Selection	75.22
Neural Network without Feature Selection	75
Neural Network with Feature Selection	76.58
Hybrid Prediction Model(FS+SVM+NN)	96.09

where FS is feature selection, SVM is Support Vector Machine and NN is Neural Network

Another important finding from the present study is that the use of appropriate feature selection method, always improvise the results of classifier. Depending on the nature of the data, the appropriate feature selection technique must be selected to reduce the time requirements and to increase the classification recognition rate.

6. CONCLUSION & FUTURE SCOPE

In this paper, a novel computational hybrid model is developed using MATLAB, to predict the diabetic disease based on the symptom analysis. The proposed model is defined with comprehensible definition of its each intermediate stage. The proposed model has achieved 96.09% accuracy rate when filtration feature selection, two level classification techniques (SVM and the Neural Network) are implemented. The comparative analysis proved that the proposed model had outperformed the earlier models and approaches. The suggested hybrid model can be used as an expert system application, under the guidance of diabetic expert to assist the physicians for taking decisions regarding the early diagnosis of the disease.

The proposed system can largely and effectively contribute in the decision making process followed by the physicians for their patients. This model can assist the physicians in predicting the diabetes disease at an early stage which may reduce the chances of having adverse effects of diabetes on the patient's life. Further, the proposed model can be applied on any real dataset comprising of both male and female data for rich patterns. As a future scope, the missing values of the data set can be more precisely substituted before analysis. The results of this study can be enhanced by applying more precise classifiers to attain higher accuracy rate.

REFERENCES:

[1] Mohamed E. L., Linderm R., Perriello G., Di Daniele N., Poppl S. J. & De Lorenzo A., " Predicting type 2 diabetes using an electronic nose-base artificial neural network analysis", *Diabetes Nutrition and Metabolism* , Vol : 15 Issue 04, 2002, 215-221.

[2] Gan, D. "Diabetes Atlas", Brussels: International Diabetes Federation (2nd ed.), http://www.eatlas.idf.org/webdata/docs/Atlas%202003-Summary.pdf.

[3] Patil B.M., Joshi R.C., Toshniwal Durga "Hybrid Prediction Model for Type-2 diabetic patients", *Expert Systems with Applications* (vol: 37), 2010, 8102-8108.

[4] Vapnik V., "Statistical learning theory", New York: John Wiley & Sons, Inc.

[5] Tsang W., Kwok J. T., Cheung P., "Core Vector Machines: Fast SVM training on very large data sets", *Journal of Machine Learning Research*, (vol: 06), 2005, 363-392.

[6] Meng X. H., Huang Y. X., Rao D. P., Zhang Q. Liu Q., " Comparison of three data mining models for predicting diabetes or pre diabetes by risk factors", The Kaohsiung Journal of Medical Sciences, 2012, 93-99.

[7] Mohamudally N. & Khan D. M., "Application of a Unified medical data miner (UMDM) for prediction, classification, interpretation and visualization on medical datasets: The diabetes dataset case", International conference on advances in data mining: Applications and theoretical aspects (ICDM), Berlin, Heidelberg: Springer-Verlag, 2011, pp. 78-95.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[8] Su C T, Yang C H, Hsu K H, Chiu W K, "Data mining for the diagnosis for type II diabetes from three-dimensional body surface anthropometrical scanning data", Computing Math Appl,(vol:51), 2006,1075-92.

[9] Huang Yue, McCullagh Paul, Black Norman & Harper Roy, "Feature Selection and classification model construction on Type 2 diabetic patients' data", Artificial Intelligence in Medicine, 2007, 41, 251-262.

[10] Award M., Khan L., Bastani F. & Yen I., "An effective support vector machines (SVMs) performance using hierarchical clustering", proceedings of the 16th IEEE international conference on tools with artificial intelligence, 2004,pp. 663-667.

[11] Zhong Wei, Chow Rick & He Jieyue, "Clinical charge profiles prediction for patients diagnosed with chronic diseases using Multi-level Support Vector Machine", Expert Systems with Applications, (vol: 39), 2012, 1474-1483.

[12] Daniael B. & Cao D., "Training support vector machines using adaptive clustering", in proceedings of the SIAM international conference on data mining, 2004, pp. 126-137.

[13] Antonelli Dario, Baralis Elena, Bruno Giulia, Cerquitelli Tania, Chiusano Silvia & Mahoto Naeem, "Analysis of diabetic patients through their examination history", Expert Systems with Applications, (vol:40), 2013, 4672-4678.

[14] http://in.mathworks.com/help/nnet/ref/roc.html.

[15] Zhong Wei, Chow Rick & He Jieyue (2012) Clinical charge profiles prediction for patients diagnosed with chronic diseases using Multi-level Support Vector Machine. Expert Systems with Applications, 39: 1474-1483

[16] http://archive.ics.uci.edu/ml/datasets/Diabatic+Disease

[17] Han Jiawei and Micheline kamber, "Data Mining: Concepts and Techniques", Second edition, Morgan Kauffman Publishers

[18] Antonin Slaby, "Proceedings of the ITI 2007 29th International Conference on Information Technology Interfaces" June 25-28, 2007, Cavtat, Croatia

[19] Humar K., & Novruz A., "Design of a hybrid system for the diabetes and heart diseases", Expert Systems with Applications, 2008, 35, pp: 82-89.

[20] Varma K.V.S.R.P, Rao A.A., Lakshmi T.S.M. & Rao P.V.N(2014) A Computational Intelligence approach for a better diagnosis of diabetic patients. Computers and Electrical Engineering, 40 ,pp: 1758-1765.