Volume 5, Issue 9, September 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Diabetes Mellitus Forecast using Kernel Discriminant Analysis (KDA) with Neural Network Classifier

Sukhjinder Singh
M.Tech Student
Department of Computer Science & Engineering
SGGSWU, Fatehgarh Sahib, Punjab, India

Kamaljit Kaur
Assistant Professor
Department of Computer Science & Engineering
SGGSWU, Fatehgarh Sahib, Punjab, India

Abstract: Diabetes Mellitus, simply called as Diabetes, is a harmful disease, in which a person is affected with high blood glucose level. The main cause of this disease is that body fails to produce insulin or not properly utilization of insulin. The Diabetes can results in Insulin Resistance, age, central obesity, Stress, Polyuria, Polydipsia disease etc. High level of Diabetes is associated with the heart disease. The Diabetes Disease is highly prevalent in world. There is system used to predict the complications with the use of clinical dataset. But few systems prediction based on risk factors. Systems based on risk factors helps not only of experts but also warn patients in advance. This paper proposes a methodology that aims to predict complications regarding diabetes disease in advance on basis of risk factors. This methodology uses the data mining technique for prediction. In this paper, two data mining techniques: Feed Forward Neural Network and Kernel Discriminant Analysis (KDA) technique is used for classifying the medical databases. The parameters used for disease identification have been designed in such a way that user can predict himself either he is affected with diabetes or not.

Keywords: Diabetes Mellitus, Risk Factors, Medical Databases, Kernel Discriminant Analysis, Feed Forward Neural Network.

I. INTRODUCTION

Diabetes is one of the most dangerous diseases that causes of death [1]. Diabetes is metabolic disorder that occurs due to failure of body due to produce insulin properly. According to W.H.O, by 2015 a total of 3 hundred millions of the world population will be affected by diabetes [2]. It has been noticed that diabetes affected a more fatal persons and also the women than men. The cause of worst affect on women is their lower survival rate and poor quality of life. A cause of diabetes is also that many of the peoples don't have knowledge this disease [3]. Human body needs energy for activation the carbohydrates are broken down to glucose. That is the important energy source for body cells. Insulin is necessary to translate the glucose into body cells. The blood glucose is supplied with insulin [4]. In the world, there are many systems that are used for the advanced complication predictions of diabetes symptoms and produce the results on the basis of these symptoms. Most of these systems predict the results based on datasets available in clinical labs. But some the systems predict the causes of diabetes based on the risk factors. Such as Insulin Resistance, age, central obesity**,** Stress, Polyuria, Polydipsia disease etc, but still the major problem of these systems are to diagnose the disease correctly and costly medical tests [1]. Many data mining techniques are used to solve these problems. Data mining techniques helps the experts and patients to calculate the diabetes risks, on the basis of risks they can know in advance either they affected with diabetes or not.

II. DATA MINING TECHNIQUES

In order to find symptoms of diabetes, data mining techniques are used with help of different algorithms. Different types of data mining data types are used by researchers for prediction of diabetes such as: Neural Network, kNN, Fuzzy Neural Network, Genetic Algorithm. **Durairaj et al [4] in 2015** proposed that Neural Networks are one of the soft computing techniques that can be used to make predictions on medical data. A detailed survey was conducted on the application of different soft computing techniques for the prediction of diabetes. This survey was aimed to identify and propose an effective technique for earlier prediction of the disease. The earlier detection using soft computing techniques help the physicians to reduce the probability of getting severe of the disease. Their data set chosen for classification and experimental simulation was based on Pima Indian Diabetic Set from (UCI) Repository of Machine Learning databases**.** **Anand et al [1] in 2013** applied Neural Network techniques successfully for diagnosis of Type II diabetes. He proposed a K-Fold cross validation method for classification of PIMA Indian diabetes data set. The classification accuracy was computed with PCA preprocessing and higher order neural network. The problem of missing data in the analysis and decision making process was handled through PCA. **Rajesh et al [5] in 2012** have discussed that Medical professionals need a reliable prediction methodology to diagnose Diabetes. Data mining is applied to find useful patterns to help in the important tasks of medical diagnosis and treatment. This project aims for mining the relationship in Diabetes data for efficient classification. The data mining methods and techniques will be explored to identify the suitable methods and techniques for efficient classification of Diabetes dataset and in mining useful patterns. **Pradhan et al [6] in 2012** have

presented that Support vector machine (SVM) is one of the most important machine learning algorithms that has been implemented mostly in pattern recognition problem, for e.g. classifying the network traffic and also in image processing for recognition. Lots of research is going on in this technique for the improvement of Qos (quality of service) and in security perspective. The latest works in this field have proved that SVM performs better than other network traffic classifier in terms of generalization of problem. **Thirumal et al [7] in 2015** have discussed that Data mining looks through a large amount of data to extract useful information. The usage of data mining techniques in disease prediction is to reduce the test and increase the accuracy of rate of detection. One of the most common diseases among young adult is Diabetes mellitus. This develops at a middle age and more common in obese children and adolescents. In order to reduce the population with diabetes mellitus it should be detected at an earlier stage, hence a quick and efficient detection mechanism has to be discovered. The principle of this study is to apply various data mining techniques which are noteworthy to prediction of diabetes mellitus and extract hidden patterns from the PIMA Indian diabetes dataset available at UCI Machine Learning Repository. After study of various techniques, some of which discussed above, this paper proposes a new approach that combines Feed Forward Neural Network and KDA technique for diabetes prediction based on risk factors.

III. METHODOLOGY

a. The Data

There are many risk factors that become a cause of diabetes. It is very difficult to diagnose these factors easily. Most of the time the disease is diagnosed at the last stage of disease. With the help of risk factors, it is easy to diagnose disease possibilities in advance. The dataset used in this research composed of 9 risk factors are blood cholesterol, plasma glucose, diastolic blood pressure, triceps (SFT), Insulin, BMI, DPF, age class. On the basis of these risk factors the results are computed to know whether the patient has risk of Diabetes or not. The dataset contains 768 people data collected from the UCI repository. The dataset consists of 9 attributes as shown in Table 1.

Table 1: Attributes of Diabetes Dataset

S. No.	Name of the Attributes	Description
1.	Blood cholesterol	Blood cholesterol is measured in units to measure the cholesterol levels (mmol/l)
2.	Plasma glucose	Plasma glucose concentration measured using two hours oral glucose tolerance test (mm Hg)
3.	Diastolic BP	Diastolic blood pressure
4.	Insulin	2 hours serum insulin (mu U/ml)
5.	Triceps (SFT)	Triceps skin fold thickness (mm)
6.	BMI	Body Mass Index (weight kg/height in (mm)^2)
7.	DPF	Diabetes Pedigree function
8.	Age	Age of Patient
9.	class	Diabetes on set within 5 years

b. Extraction using KDA with Feed forward neural network:

The methodology used in this research for diabetes diagnosis is based on the KDA and Feed Forward Neural Network. Feed Forward is term that describes the pathway within Neural Network system, which passes the signals from sources to external environment. KDA is technique of feature extraction. In this system, KDA is used with Feed Forward Neural Network. The objective of KDA is to find a transformation maximizing between the class variance and minimizing within the class variance. The features of risk factors extracted with use of KDA are taken as an input by the Feed Forward system. On the basis of these features, Feed forward system produces signals as output. These signals are the output that show either disease of diabetes exists or not on the basis of features of risk factors taken as the input. Figure 1:Shows the process of proposed work.

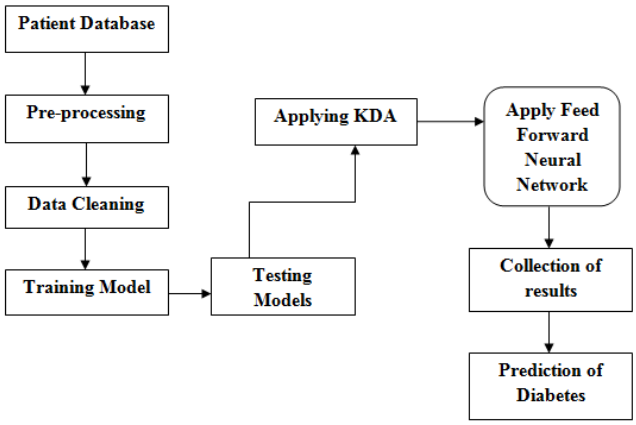

Figure 1: Framework for the proposed work

IV. PERFORMANCE MEASURES

In this approach the accuracy rates of classification for the data sets is measured. The system is made using MATLAB 2012a. In this research data of 768 people is collected based on the risk factor related to diabetes disease. The different output parameters are used to achieve the accurate results. This problem has four possibilities of outcome: True positive (TP) Rate, True Negative (TN) Rate, False positive (FP) Rate and False Negative (FN) Rate. Where TP and TN are correct classification and FP is the outcome incorrectly predicted as positive and FN is outcome incorrectly predicted as negative.

Table2. A Confusion Matrix for Prediction Outcomes

		Disease	
Prediction		Positive	Negative
	Positive	TP	FP
	Negative	FN	TN

The following set of Evaluation measures are being used to find out the results [8].

Sensitivity: A high sensitivity is clearly important where the test is used to identify a serious but treatable disease.

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

Specificity: The specificity of a clinical test refers to the ability of the test to correctly identify those patients without the disease.

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

Accuracy: Accuracy measures correctly figured out the diagnostic test by eliminating a given condition.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

Precision: is the fraction of retrieved instances that are relevant. Precision is calculated by:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall: - is the fraction of relevant instances that are retrieved. Recall is calculated by:

$$\text{Recall} = \frac{TP}{TP+FN}$$

F-measure: - A measure that compile precision and Recall is the harmonic mean of Precision and Recall, the traditional F-measure and balanced F-score. It totally depends upon value Precision and Recall. F-measure is calculated by:

$$\text{F-measure} = \frac{(2*Precision\ *Recall\ )}{(Precision\ +Recall\ )}$$

The results on the dataset will be displayed in the form of two-dimensional confusion matrix having a row and column for each class.

V. RESULTS AND DISCUSSION

The input data consisted of risk factors collected from768 people available at UCI repository. The data is encoded as 70% of the data is used for training and 30% for testing and validation. A confusion matrix is produced using MATLAB and accuracy is determined as Accuracy = (TP + TN**)** / **(**TP + FP + TN + FN); where TP, TN, FP and FN denotes true positives, true negatives, false positives and false negatives respectively. The accuracy of diagnosing the Diabetes Disease on the training data and testing data is calculated as 96%. The least mean square error (MSE) achieved is 0.21544 after 1 epochs, as shown in Figure 2. Results show Kernel Discriminant Analysis (KDA) and Neural Network approach gives better average prediction accuracy than the traditional ANN.

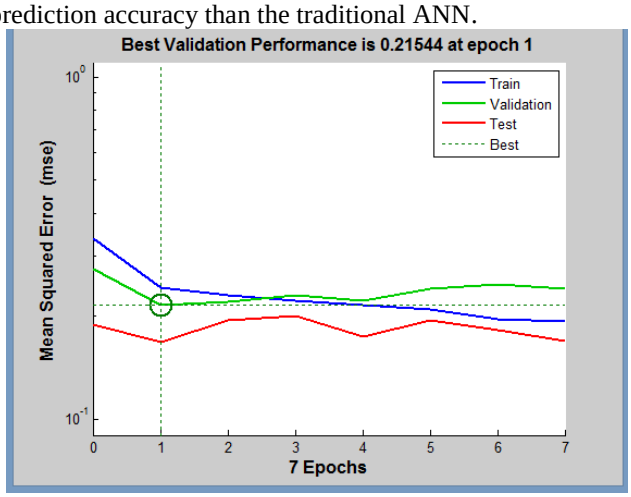

Figure 2: Performance Graph

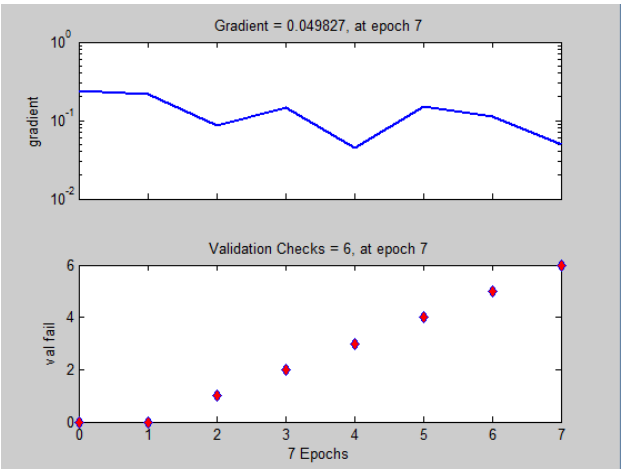

Figure 3: Training State Graph.

VI. CONCLUSION

Data mining techniques used for diabetes prediction provides best results. In this research KDA process extract the features properly and Feed Forward neural network provides the results on the basis of these features. The combination of KDA and Feed Forward Neural Network achieved the accuracy of 96%. So, the obtained results show that this approach performs better than existing techniques. For future, there is enough scope for improvement in this field and with the advent of faster and more accurate learning techniques. Results can be surely improved consider. To improve the performance of this system by using the better feature extraction technique and better classification methods to create model, this will give the efficient results.

REFERENCES

[1] Anand A. Chaudhari, Prof.S.P.Akarte" Fuzzy & Datamining based Disease Prediction Using K-NN Algorithm*" International Journal of Innovations in Engineering and Technology (IJIET) ISSN: 2319 – 1058,* Volume 3, Issue 4, April 2014.

[2] Prof.Sumathy, Prof.Mythili Thirugnanam, Dr.Praveen Kumar, Jishnujit T M, K Ranjith Kumar" Diagnosis of Diabetes Mellitus based on Risk Factors" *International Journal of Computer Applications (0975 – 8887),* Volume 10, Issue 4, November 2010.

[3] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly" Diagnosis Of Diabetes Using Classification Mining Techniques" *International Journal of Data Mining & Knowledge* Management Process (IJDKP) Volume 5, Issue 1, January 2015.

[4] M. Durairaj, G. Kalaiselvi " Prediction Of Diabetes Using Soft Computing Techniques- A Survey" International *journal of scientific & technology research, ISSN 2277-8616,* Volume 4, Issue 03, March 2015.

[5] K. Rajesh, V. Sangeetha" Application of Data Mining Methods and Techniques for Diabetes Diagnosis" *International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012International Journal of Engineering and Innovative Technology (IJEIT),* Volume 2, Issue 3, September 2012.

[6] Ashis Pradhan"Support Vector Machine- A Survey" *International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459,* Volume 2, Issue 8, August 2012.

[7] Thirumal P. C. and Nagarajan N" Utilization Of Data Mining Techniques For Diagnosis Of Diabetes Mellitus - A Case Study" *ARPN Journal of Engineering and Applied Sciences ISSN 1819-6608,* Volume 10, Issue 1, January 2015.

[8] Abdullah, A. S., and R. Rajalaxmi. "A data mining model for predicting the coronary heart disease using random forest classifier." In *International Conference in Recent Trends in Computational Methods, Communication and Controls*, pp.22-25, 2012.

[9] K. R. Lakshmi,S.Prem Kumar" Utilization of Data Mining Techniques for Prediction of Diabetes Disease Survivability*" International Journal of Scientific & Engineering Research, ISSN 2229-5518,* Volume 4, Issue 6, June-2013.

[10] Miroslav Marinov, M.S., Abu Saleh Mohammad Mosa, M.S.,1Illhoi Yoo, Suzanne Austin Boren" Data-Mining Technologies for Diabetes: A Systematic Review",*Journal of Diabetes Science and Technology*, Volume 5, Issue 6, November 2011.

