Risk Level Prediction System of Diabetic Retinopathy Using Classification Algorithms

Dr.V.Ramesh[1], R.Padmini[2]

Assistant Professor[1], Research Scholar[2],
Department of CSA,SCSVMV University,
Kanchipuram, India.

***Abstract*: Diabetic retinopathy (DR) is a sight threatening complication of systemic diabetes mellitus that results from damage to the blood vessels of the retina. It is one of the most frequent causes of blindness among adults. In the early stage of disease, mostly all patients with type 1 diabetes and >60% of patients with type 2 diabetes have retinopathy. Making medical decisions like diagnosing the diabetic retinopathy is a multifaceted task. The complexity is in recognizing predictive factors associated with the diseases. Even though hospitals are maintaining the clinical data, human cannot process all the data available. Thus there is need for intelligent data analysis techniques like data mining to discover knowledge which supports physicians. The aim of the study is to determine the correlation among the various risk factors of diabetic retinopathy and design a model to predict the risk level of diabetic retinopathy. We compared risk factors between patients having diabetics with and without retinopathy. Collected data are preprocessed for classification of risk level. We found that among different classification algorithms MLP algorithm is suitable to predict the risk factor. A model is designed using .NET platform to predict the risk level of the diabetic retinopathy. The system also suggests some recommendation to take precautionary steps to avoid diabetic retinopathy disease.**

***Keywords*: Classification Algorithms, Data Mining, Diabetic Retinopathy, Multi-layer perception.**

I. INTRODUCTION

In modern medicine, large amounts of data are generated, but there is a widening gap between data collection and data comprehension. It is often impossible to process all of the data available and to make a rational decision on basic trends. Thus there is a growing pressure for intelligent data analysis such as data mining to facilitate the creation of knowledge to support clinicians in making decisions. Data mining techniques like classification model construction could be used in such databases to support other research studies. The objective of this study is to contribute to the development, validation and application of data mining methods for prediction in decision support systems in medicine. The aim of this study is to design a prediction system using classification techniques to find risk level of diabetic retinopathy and provide some recommendation to the people. In this paper, various classification algorithms were applied on the collected data and designed a tool with some recommendation for people to know their risk level of diabetic retinopathy disease.

Diabetic retinopathy is damage to the retina (retinopathy), specifically blood vessels in the retina, caused by complications of diabetes mellitus. According to Medilexicon'smedical dictionary: Diabetic retinopathy means "Retinal changes occurring in diabetes mellitus, marked by microaneurysms, exudates, and hemorrhages, sometimes by neovascularization." Diabetic retinopathy can eventually lead to blindness if left untreated. Approximately 80% of all patients who have had diabetes for at least ten years suffer from some degree of diabetic retinopathy. The retina is the light-sensitive membrane that covers the back of the eye. If diagnosed and treated early blindness is usually preventable. There are two main classes of diabetes, which are diagnosed ultimately by the severity of the insulin deficiency. Insulin-dependent diabetes mellitus or Type 1 diabetes is an insulinopenic state, usually seen in young people, but it can occur any age.Non-insulin-dependent diabetes mellitus or Type 2 diabetes is the more common metabolic disorder that usually develops in overweight, older adults, but an increasing number of cases occur in younger age groups. Diabetic retinopathy generally starts without any noticeable change in vision. However, an ophthalmologist can detect the signs. Hence, it is important for diabetes patients to have an eye examination at least once or twice annually. In this regard we developed a tool to check the risk level of diabetic retinopathy by using classification algorithms. Normal Vision and the same scene viewed by a person with diabetic retinopathyare shown below:

Fig. 1 Normal Vision Fig.2 Same scene viewed by a personwith diabetic retinopathy

Though current clinical treatments forretinopathy slow its progression but they cannot fully reversevision loss. Studies have confirmed that clinical prognosis isbetter if patients are screened and treated early. Therefore, thecurrent study aims to design a

prediction system using classification techniques and develop a tool for the diabetic patients to know their risk level to take treatment. The system was developed with help of ophthalmologist andusing more than 600 diabetes patients'records obtained from different eye hospitals. Initially the preprocessing techniques are applied on data to make it suitable for the data mining process. Once the preprocessing gets over, we applied different classification algorithms on the data to obtain patterns from the data. We found that Multi Layer Perceptron is the best prediction algorithm to find the risk level of diabetic patients. Obtained results are used in developing a tool using .NET frame work for the diabetic patients to know their risk level.

II. REVIEW OF LITERATURE

Chan [1] explored the relationshipbetween physiological data and retinopathy, nephropathy andneuropathy in Taiwan using two data mining methods, namelyC5.0 and neural network. In the C5.0 method, data withdiabetes duration more than seven years were used to generate22 rules needed for prediction whilst for the neural networkmethod, retinopathy predictions were made based on a hiddenlayer with 52 neurons. The sensitivity and specificity forretinopathy prediction were found to be 58.62 and 74.73,respectively using C5.0 whereas the values were 59.48 and99.86 for neural network, indicating the latter method has ahigher prediction power.

Conway et al. [2] investigated the role of hemoglobin level in predicting proliferative retinopathy among 426 Type 1 diabetes patients.They used stereo fundus photography to determine the presence of proliferative retinopathy, followed by Cox proportional hazards modeling with stepwise regression to determine the association of hemoglobin level with proliferative retinopathy. They found that higher hemoglobin level predicts the incidence of proliferative retinopathy, though the association varies by gender, being linear and positive in men and quadratic in women.

A model Intelligent Heart Disease Prediction System (IHDPS) built using data miningtechniques like Decision Trees, Naïve Bayes and Neural Network was proposed by SellappanPalaniappanet al. [3]. The results revealed the peculiar strength of each of the methodologies in comprehending the objectives of the specified mining objectives. IHDPS was capable of answering queries that the conventional decision support systems were not able to. It facilitated the establishment of vital knowledge, e.g. patterns, relationships amid medical factors connected with heart disease. IHDPS subsists well being web-based, user-friendly, scalable, reliable and expandable.

The prediction of Heart disease, Blood Pressure and Sugar with the aid of neural networks waspproposed by Niti Guru et al. [4]. The experiments were carried out using sample database of patients' records. The ANN is tested and trained with 13 input variables such as Age, Blood Pressure, Angiography's report and the like. The supervised network has been recommended for diagnosis of heart diseases. Training was carried out with the aid of back-propagation algorithm.

III. METHODOLOGY

The approach applied is based on data mining methods and involves four major steps : Preprocessing the data, classification for model construction, association rule extraction and development of tool using .NET platform. In this study, knowledge base was developed by collecting data from the diabetic patients by questionnaire method, literature reviews and interview with leading medical experts. A total of 641 patients (410 men and 231 women) with type 2 diabetes collected for our study. Their ages were from 25 to 70 and duration of diabetes was between 4 and 20. A total of 14 variables were selected namely, age, gender, BMI, occupation, level of blood pressure, cardiac complication, eye vision, sugar_fasting, sugar_post_parandial, type of diabetic medicine, hereditary details, food habits like veg, nonveg, smoking and drinking habits, details of physical exercises. Data are collected from the patients at different eye hospitals in around Kanchipuram and Chennai districts of Tamil Nadu. The detail of collected data is shown Table 1. Collected data were preprocessed in order to make it suitable for mining process. After the preprocessing, the classification algorithms like ID3, J48, BayesNet and MultiLayerPerceptron. We found that Artificial Neural Networks is the best prediction algorithm to find the risk level of diabetic patients. Obtained results are used in developing a tool using .NET frame work for the diabetic patients to know their risk level.

3.1 Data Preprocessing

Cleaning and filtering of the data might be necessarily carried out with respect to the data and datamining algorithm employed so as to avoid the creation of deceptive or inappropriate rules or patterns [6].The actions in the pre-processing are the removal of redundant records,normalization of values, accounting for missing datapoints and removing unnecessary data fields[8]. In order for improve the quality of data needs to be transformed. The raw data is changed into data sets with a few appropriatecharacteristics. Moreover it is essential to combine the data to reduce the number of datasets besides minimizing the memory and processing resources required by the data mining algorithm[

Table 1 : Training Data Tuples

AGE	SEX	BP	EYEVSN	DIAYRS	DIAFAST	DIAPOST	MED	GNDIABT	DIARET
>60	Male	No	Short sight	10-15	120-160	>300	Tablet	Father	Yes
30-40	Male	No	Long sight	5-10	160-200	>300	Insulin	No	Yes
>60	Male	yes	Long sight	10-15	160-200	>300	Tablet	No	Yes
<30	Female	No	Normal	<5	120-160	180-240	Insulin	No	No
40-50	Female	No	Short sight	5-10	160-200	240-300	Tablet	No	Yes
>60	Female	yes	Normal	>15	>200	>300	Both	Mother	Yes
50-60	Female	yes	Normal	>15	160-200	240-300	Tablet	No	Yes
50-60	Female	No	Normal	5-10	>200	240-300	Tablet	No	Yes
<30	Male	No	Normal	<5	160-200	>300	Tablet	Mother	Yes
40-50	Male	No	Normal	<5	120-160	180-240	Tablet	No	No
50-60	Female	No	Long sight	10-15	>200	>300	Both	Mother	Yes
…	…	…	…	…	…	…	…	…	…

An attribute may be redundant if it can be derived from another attribute or set of attributes. The redundancies can be detected by using correlation analysis. Given two attributes, such analysis can measure how strongly one attribute implies the other. For numerical attributes, we can evaluate the correlation between two attributes, X and Y, by computing the correlation coefficient.There exist broadly two approaches to measure the correlation between two random variables. One is based on classical linear correlation and the other is based on information theory. Under the first approach, the most well known measure is linear correlation coefficient given by the formula

$$Correlation(r) = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - \sum X^2 \; N\sum Y^2 - \sum Y^2}}$$

Where X and Y are the two features/attributes.

To find out the correlation between the attributes some normalization is done.Fitness function is a simple function, which assigns a rank to individual attribute on the basis of correlation coefficients. Since strongly correlated attributes cannot be the part of DW together, only those attributes shall be fit to take part in the crossover operations that are having lower correlation coefficients. In other words we can say lower the correlation is higher the fitness value will be. The fitness function is taken over here is;

$$f(X) = 1 - \min(r_X)$$

Table 2 :Correlation Matrix

FACTORS	AGE	SEX	BP	EYEVSN	DIAYRS	DIAFAST	DIAPOST	MED	GNDIABT
AGE	1	0.1	0.55	0.22	0.8	0.34	0.37	0.06	-0.3
SEX	0.1	1	0.14	-0.28	0.31	0.49	-0.27	O.40	0.2
BP	0.55	0.14	1	-0.04	0.74	0.27	0.24	0.09	0.11
EYEVSN	0.22	-0.28	-0.04	1	0.2	0.14	0.55	0.22	-0.05
DIAYRS	0.8	0.31	0.74	0.2	1	0.46	0.5	0.3	-0.28
DIAFAST	0.34	0.49	0.27	0.14	0.46	1	0.4	0.47	0
DIAPOST	0.37	-0.27	0.24	0.55	0.5	0.4	1	0.23	-0.57
MED	0.06	0.40	0.09	0.22	0.3	0.47	0.23	1	-0.22
GNDIABT	-0.3	0.2	0.11	-0.05	-0.28	0	-0.57	-0.22	1

Table 3 : Comparison of accuracy

IF	AND	IMPLIES	CONFIDENCE
SEX = MALE	DIAPOST>= 300	DR = YES	148/150 = 98.6
DIAPOST>= 300	MED=TABLET	DR = YES	176/179 = 98.3
DIAPOST>= 300	GNDIABT=NO	DR = YES	164/167 = 98.2
BP=NO	DIAPOST>= 300	DR = YES	173/174 = 97.6
DIAPOST>= 300		DR = YES	244/250 = 97.6
DIAFAST>=200		DR = YES	147/152 = 96.7
EYEVSN=NORMAL	DIAPOST>= 300	DR = YES	169/175 = 96.5
AGE >=60	MED=TABLET	DR = YES	134/145 = 92.4
DIAYRS=10-15		DR = YES	169/185 = 91.3
Example SEX=MALE and DIAPOST>=300 IMPLIES DR = YES Confidence=98.6%			

where, min (rx) is minimum value of correlation coefficient corresponding to any attribute X. The fitness values of individual attributes computed as discussed earlier are shown in Table 2.After getting the pure data, various classification algorithms are applied on the data sets to find out the most suitable algorithm to develop a model.

For finding the association between the different factors and diabetic retinopathy, we used SPSS software package. The results shows that there is close association between diabetic years and diabetic retinopathy (p value = 0.001), diabetic reading in fasting and diabetic retinopathy (p value = 0.001), diabetic reading past and diabetic retinopathy (p value = 0.001) and medicine intake and diabetic retinopathy (p value = 0.001). The results reveals that there is no association between age and diabetic retinopathy (p value = 0.612), sex and diabetic retinopathy (p value = 0.150), Blood pressure and diabetic retinopathy (p value = 0.717) and hereditary factors and diabetic retinopathy (p value = 0.146). For finding multiple associations between different factors and diabetic retinopathy we used WEKA tool. The results are given in Table 3.

IV. RESULTS AND DISCUSSION

All the patient data are clustered into three groups according to their degree of symptoms. Group 1: patients in Low Risk, Group 2 : patients at Medium Risk and Group 3 : patients at High Risk. Of all these 641 patients 285 patients were classified as No Risk group, 248 patients at Medium Risk Group and 108 patients at High Risk Group. Since, the objective of our work is to investigate the performance of different classification methods, various classification algorithms are analysed using WEKA Tool.From our analysis we found that MLP algorithm is suitable to predict the risk factor. The details of comparative analysis are given in Table 4.

Table 4 : Comparison of accuracy

Algorithm	CorrectClassified Instances%(Value)	IncorrectClassified Instances%(Value)	Time taken (Seconds)
MLPNN	82.17	17.83	2.73
RandomForest	81.23	18.77	1.08
Bayers net	81.29	18.71	1.02
Decision stump	80.05	19.95	0.97
NaiveBayes	78.20	21.80	0.89

The refined Retinopathy disease data set, resultant from preprocessing, is then classified using neural network algorithm.The values corresponding to each attribute in the significant patterns are as follows: blood sugar range is greater than 120, duration of Diabetic years and family history of diabetic. In this work, in addition to these significant parameters, we have used some more parameters significant to Diabetic Retinopathy with their weightage and the priority levels are advised by the medical experts. The neural network is trained with the selected significant patterns.The designed prediction system employed MLPNN with Back-propagation as training algorithm. With the help of the designed prediction system we can predict the different risk levels of Diabetic Retinopathy. The retinopathy risk evaluator receives its input when the medical expert enters the patient's data (age, duration of diabetics, sex, etc.). These inputs will be analysed by the risk evaluator and the probability of retinopathy occurrence will be displayed by using this classification rules.

Figure 1 shows a sample output for a negative retinopathy risk level prediction. In this example, the input for 10 variables was entered for a 62 years old male. Because the similarity of cases is relatively higher than the prediction of neural network, the system votes on a final negative prediction (i.e. monitoring status). The reasoning behind the final prediction is also provided together with the three most similar cases.

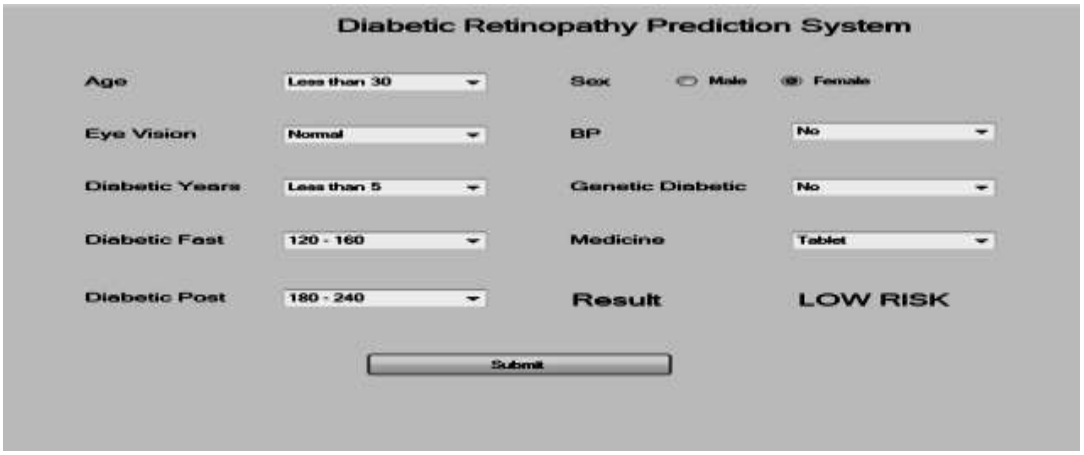

Figure 3 Application screen shots for low risk level

On the other hand, Fig 6.3 shows a positive prediction of medium risk level for a 39 years old Female. Slightly Contrary to the previous example, Figure 6.4 shows a prediction of High level of risk factors.

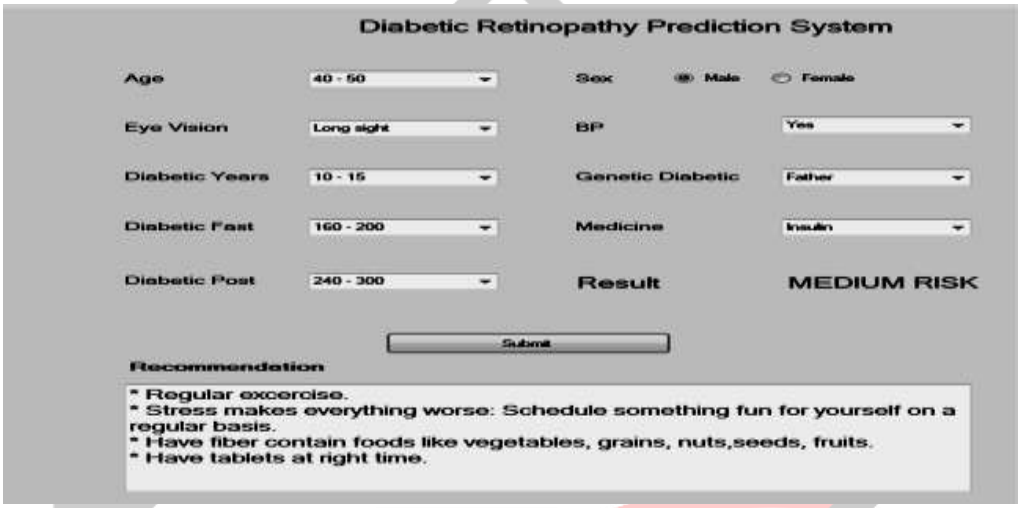

Figure 4 Application screen shots for Medium Risk Level

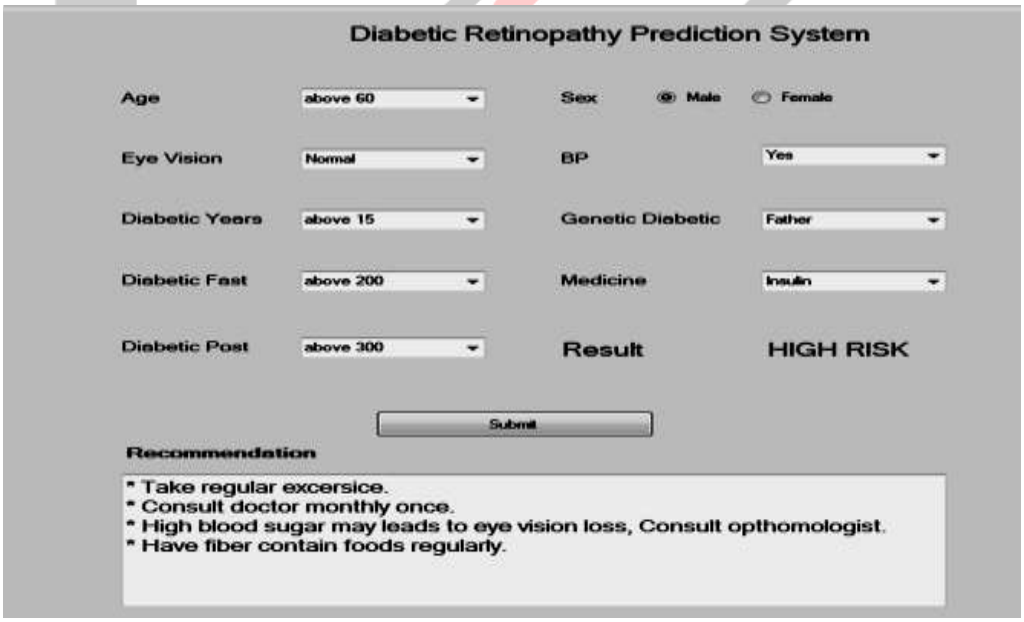

Figure 5Application screen shots for High Risk Level

V. CONCLUSION

Majority of the patients have non-insulin-dependent diabetes mellitus(NIDDM) or Type 2 diabetes. The prevalence of insulin- dependent diabetes mellitus (IDDM) or Type 1 diabetes is 10-15% of the diabetic population.The approach applied

proved successful for the classification rule based prediction of risk factors. The results obtained demonstrate the high potential of the approach and the methods developed validated to support decision making in diabetic retinopathy and other field of medicine by individual risk prediction. The system demonstrates the data mining based approaches that can be used to assess risk of factors of diabetic retinopathy of diabetic patients and recommends steps to be taken to avoid diabetic retinopathy disease. This study suggests that the patients who experience vision loss from diabetes should be encouraged to pursue visual rehabilitation with an ophthalmologist or optometrist who is trained or experienced in low-vision care. The patients with Type 1 diabetes should have an initial dilated and comprehensive eye examination by an ophthalmologist or optometrist within 3–5 years after the onset of diabetes. In general, evaluation for diabetic eye disease is not necessary before 10 years of age. However, some evidence suggests that the prepubertal duration of diabetes may be important in the development of microvascular complications; therefore, clinical judgment should be used when applying these recommendations to individual patients.

REFERENCES

[1] ChanC.L., LiuY.C., and LuoS.H., (2008) Investigation of Diabetic Microvascular complications using Data Mining Techniques," in Neural Networks, 2008. IEEE International Joint Conference on Neural Networks,2008, pp. 830-834.

[2] ConwayB.N., Miller R.G., and Klein R., Orchard T.J., (2009), Prediction of Proliferative Diabetic Retinopathy WithHemoglobin Level, Arch Ophthalmol, Vol. 127, pp. 1494-1499.

[3] SellappanPalaniappan, RafiahAwang, (2008), Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8.

[4] Niti Guru, Anil Dahiya, NavinRajpal, (2007), "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1

[5] Gerhard Münz, Sa Li, and Georg Carle, (2007) Traffic anomaly detection using k-means clustering, In Proc. of Leistungs-, Zuverlässigkeits-und Verlässlichkeitsbewertung von Kommunikationsnetzen und VerteiltenSystemen, 4. GI/ITG-Workshop MMBnet 2007, Hamburg, Germany.

[6] Ramandeep Singh, MS, Kim Ramasamy, DNB, Chandran Abraham, DO, Vishali Gupta, MS, and Amod Gupta, MS (2008), Diabetic retinopathy: An update, Indian J Ophthalmol, Vol.56(3).

[7] S.SagarImambi, T.Sudha, (2010), Building Classification System to Predict Risk factors of Diabetic Retinopathy Using Text mining, International Journal on Computer Science and Engineering Vol. 02, No. 07, 2309-2312.

[8] ShantakumarB.Patil, Y.S.Kumaraswamy (2009),Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction, International Journal of Computer Science and Network Security, VOL.9 No.2,228-236.

[9] Rajdev Tiwari, Manu Pratap Singh, (2010), Correlation-based Attribute Selection using Genetic Algorithm International Journal of Computer Applications, Volume 4– No.8, 28-34.

[10] Donald S.Fong, (2003), Diabetic Retinopathy, Diabetes Care, Vol. 26, Supplement-1.

[11] http://www.slideshare.net/Tommy96/intelligent-and-effective-heartattack-prediction-system

[12] http://www.medicalnewstoday.com/articles/183417.php