Building Classification System to Predict Risk factors of Diabetic Retinopathy Using Text mining

S.Sagar Imambi
Asst.Professor, Department of Computer Science
TJPS College,Guntur,A.P, India

T.Sudha
Professor & Head , Department of Computer Science
Vikram Simhapuri University ,A.P,

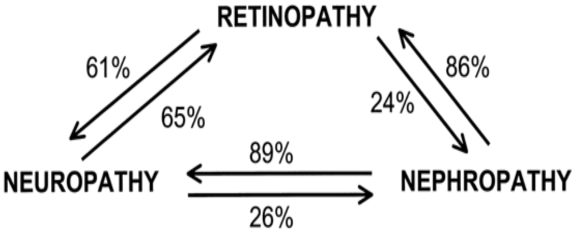
Abstract: -**This Making medical decisions such as diagnosing the diseases that cause a patient's illness is often a complex task. The Diabetic retinopathy is one of the complications of diabetes and Diabetic retinopathy is one of the most common causes of blindness. Unfortunately, in many cases the patient is not aware of any symptoms until it is too late for effective treatment. Analysis of the evoked potential response of the retina, optical nerve and optical brain centre will pave a way for early diagnosis of diabetic retinopathy and prognosis during the treatment process. The objective of this study is to identify the prevalence and severity of diabetic retinopathy and to determine the relationship between risk factors, prevalence and severity of diabetic retinopathy. We collected 3450 patients history, who are suffering with type 2 diabetes .As the available data is not in structured format, we apply text mining classification technique to predict the risk factors of the diabetic retinopathy. This study shows that a relatively short duration of case management instituted before onset of clinically identifiable retinopathy, significantly reduce the risk of developing retinopathy in patients with type 2 diabetes. A total of 1402 patients (39.8%) had evidence of retinopathy. This comprised of 32% of initial stage of DR , 20% Retinal haemorrhages, 14% patients with Mild non proliferate diabetic retinopathy, 18% with Moderate non proliferate DR, 1% with Proliferate DR ,14%with High risk**.

Keywords: Text mining ,Classification, Clinical records, Diabetic retinopathy.

I. Introduction

Making medical decisions such as diagnosing the diseases that cause a patient's illness is often a complex task. The complexity is in recognizing reliable predictive factors associated with the diseases. Though hospitals are maintaining large amount of clinical data, single human cannot process all the data available. Thus there is growing pressure for intelligent data analysis techniques to discover knowledge which supports physicians.

Diabetes is a disease which affects over 20 million Indians with almost as many being at high risk of developing the disease [9]. Early prediction of complexity is needed, to live a healthy life with diabetes. Diabetic can affect many parts of body. It can lead to complications like blindness (Retinopathy), kidney problems (Nephropathy), high BP leading to heart attack and paralysis (Neuropathy) e.t.c. Interrelationship between diabetic risk factors is given in the below fig.



.

Figure 1—Inter- relationship between diabetic complications in the follow-up
Source :Study by Pirart

Diabetic retinopathy is one of the most common causes of blindness. Unfortunately, in many cases the patient is not aware of any symptoms until it is too late for treatment.[3]. The prevalence of Diabetic Retinopathy is needed as it is growing very fast in India and other countries. We are investigating a more effective approach to predict the risk factors for prevalence of diabetic retinopathy by using text mining.

Text mining is defined as knowledge discovery in textual databases, allows us to create a technology that combines the human's linguistic capabilities with the speed and accuracy of computer. Text mining process includes information pre-processing - representation, Information Extraction, Information mining and Interpreting the result. Although most biomedical IE activities are related to literature mining and terminology extraction, (Bunescu et al., 2003), (Tveit and Saetre 2005), clinical patients record mining is not a new research.

A Statement of the problem:

We collected 3450 clinical reports from Diabetic Health care center , Andhra Pradesh. We focus on disease, symptoms, life styles, other risk factors of type-2 diabetes. Symptoms may be common for more than one disease. A physician, who has to work with this large collection of medical data should know the possibilities of complications of disease. The objective of this study is to provide a classification model which helps the physician to identify the prevalence and severity of diabetic retinopathy and to determine the relationship between risk factors, prevalence and severity of diabetic retinopathy from the narrative clinical reports.

II TEXT MINING CLASSIFICATION PROCESS

Text mining always involves (a) getting some texts relevant to the domain of interest; (b) representing the content of the text in some medium useful for processing (natural language processing, statistical modelling, etc.); and (c) doing something with the representation (finding associations, dominant themes, etc.) (Perrin, 2001). Humphreys, Demetriou, and Gaizauskas (2000) defined information extraction as "extracting information about predefined classes of entities and relationships from natural language texts and placing this information into a structured representation called a template" to build a database of information about enzymes, metabolic pathways, and protein structure from full text biomedical research articles. The process of text-mining needs a well-organized integration of the phases of knowledge discovery. Every phase of the text-mining process can be addressed with several different methods and technologies. The text mining phases are shown in the fig2.

In our work, we tried to extract representative medical terms from text database that could be used as important keywords to characterize the report. We specifically extracted four medical concepts from the report namely; diseases, sign or symptoms, life style, other risk factors like BMI, weight, hypertension etc.

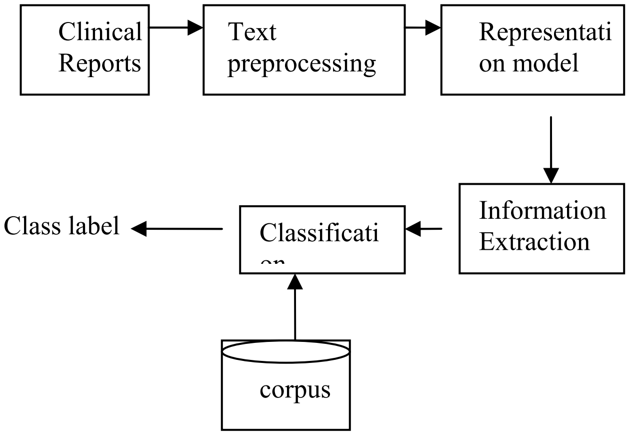


Figure 2 : Architecture for Clinical Data Classification

A. The Algorithm for predicting class label by machine learning approach:

Step 1: Prepare a set of training data. Attach topic information (class label) to the document in a target domain.

Collection of Clinical Records D= {d1,d2,d3……dn)

Collection of Classes C={c1,c2…c6}

Step2: Represent the data in vector form.

Step3: For each class i of training documents in k nearest neighbours, compute first singular vector.

Step4: Using cosine values of k nearest neighbours and frequency of documents of each class i in k nearest neighbours, compute average cosine value for each class i, Avg_Cosine(i).

Step 5: Assign (i.e., classify) the testing document a class label which has largest average cosine..

M-Classifier Function is function Φmc : D-> C that maps the documents to classes.

B. Risk factors of Diabetic retinopathy

Gender
Duration of Diabetes
Glycogenic control
Hypertension
Renal disease
Elevated Serum lipids
Pregnancy
Alcohol
Anemia
Obesity
Cataract Surgery

Figure 4: Risk factors associated with the development of diabetic retinopathy

C. . The Various levels of Diabetic retinopathy:

They are

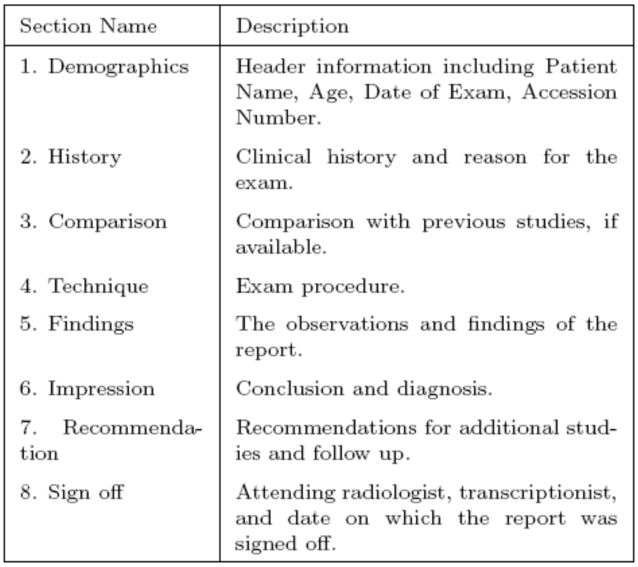
1. Initial state of DR(Diabetic Retinopathy)
2 .Retinal hemorrhages only
3. Mild non proliferative diabetic retinopathy
4. At this stage microaneurysms occur. Micoaneurysms are small areas of ballon like swelling in the retina's blood vessels.
5. Moderate non proliferative DR
6. As the disease progress , some blood vessels that nourish the retina are blocked.
7. Proliferative DR

   At this advanced stage the signals sent by the retina for nourishment trigger the growth of new blood vessels. These new blood vessels are abnormal and fragile
8. High risk

   Excess fluid and lipids leak from the blood vessels into the retina causing the retina to become thickened or swollen. The selling of center part of the retina can lead to Vision loss.

We assigned the class labels as L1, L2, L3, L4, L5 ,L6.The sample Clinical record is show in the fig5.

.



Section Name	Description
1. Demographics	Header information including Patient Name, Age, Date of Exam, Accession Number.
2. History	Clinical history and reason for the exam.
3. Comparison	Comparison with previous studies, if available.
4. Technique	Exam procedure.
5. Findings	The observations and findings of the report.
6. Impression	Conclusion and diagnosis.
7. Recommendation	Recommendations for additional studies and follow up.
8. Sign off	Attending radiologist, transcriptionist, and date on which the report was signed off.

Figure5: Sample narrated Clinical record

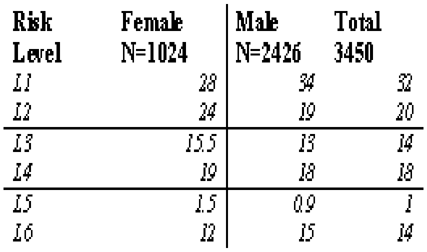
III.Experimental Results:

A total of 1402 patients (39.8%) had evidence of retinopathy. This comprised of 32% of initial stage of DR , 20% Retinal haemorrhages, 14% patients with Mild non proliferative diabetic retinopathy, 18% with Moderate non proliferative DR, 1% with Proliferative DR ,14%with High risk

Table 2 shows the details of retinopathy according to the systolic BP and it shows increasing prevalence of retinopathy with increase in hypertension.

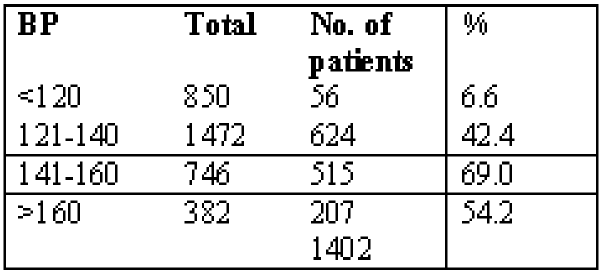
Table 3 shows the details of retinopathy according to the duration of diabetes and it shows increasing prevalence of retinopathy with increase in duration.

Table4 shows the Significant associations, which were found between the presence of retinopathy and age, high systolic BP, BMI, high pulse rate, duration of diabetes.



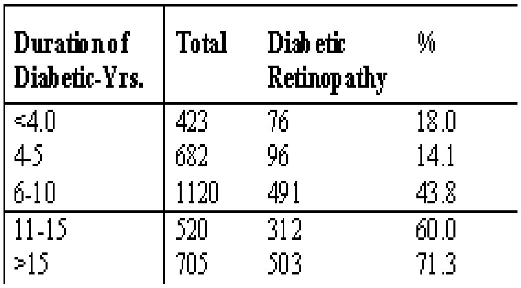
Risk Level	Female N=1024	Male N=2426	Total 3450
L1	28	34	32
L2	24	19	20
L3	15.5	13	14
L4	19	18	18
L5	1.5	0.9	1
L6	12	15	14

Table1: The % of Diabetic persons with various Retinopathy levels.



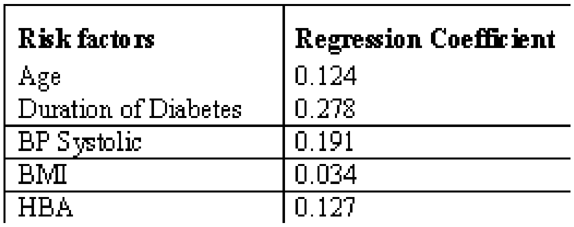
BP	Total	No. of patients	%
<120	850	56	6.6
121-140	1472	624	42.4
141-160	746	515	69.0
>160	382	207 1402	54.2

Table2:BP and Prevalence of Retinopathy



Duration of Diabetic-Yrs.	Total	Diabetic Retinopathy	%
<4.0	423	76	18.0
4-5	682	96	14.1
6-10	1120	491	43.8
11-15	520	312	60.0
>15	705	503	71.3

Table 3: Duration of Diabetes and prevalence of DR



Risk factors	Regression Coefficient
Age	0.124
Duration of Diabetes	0.278
BP Systolic	0.191
BMI	0.034
HBA	0.127

Table4:Results of multiple regression analysis showing association of various risk factors with DR

IV Conclusions

As the available clinical data is not in structured format, we apply text mining classification technique to predict the risk factors of the diabetic retinopathy. This study shows that a relatively short duration of case management instituted before onset of clinically identifiable retinopathy, significantly reduce the risk of developing retinopathy in patients with type 2 diabetes. Our goal is to develop a scalable and robust clinical report classification system that could be applied in large hospital settings to help the physicians, so that they can guide the patients easily and reduce the vision loss.

References

[1] S. Rosenbloom, R. Miller, K. Johnson, P. Elkin, S. Brown, Interface Terminologies: Facilitating Direct Entry of Clinical Data into

Electronic Health Record Systems, Journal of the American Medical Informatics Association 13 (3) (2006) 277{288.}

[2] Srinivasan,P. Text mining: generating hypotheses from MEDLINE. J. Am. Soc. Inf. Sci. Technol., 55,396–413. ,2004

[3] Maberley, D.A., King, W., Cruess, A.F., Koushik, A.Risk factors for diabetic retinopathy in the Cree of James Bay, Ophthalmic Epidemiol., 9(3), 153-167,2002

[4] Kemple AM, Zlot AI, Leman RF: Perceived likelihood of developing diabetes among high-risk Oregonians [article online]. Available from www.cdc.gov/pcd/ ssues/2005/nov/05_0067.htm. Accessed 9 March 2007

[5] International Diabetes Federation: What are the warning signs of diabetes [content online]? Available from www.idf.org/ home/index.cfm?node_11. Accessed 9 March 2007

[6] Hersh W, Cohen AM, Roberts P, Rekapalli HK. TREC 2007 genomics track overview. In Proceedings of the Sixteenth Text, REtrieval Conference, 2007.

[7] Ella Bingham, Advance in Independent Component Analysis with Application toData Mining, Dissertation for degree of Doctoral of Science in Technology at Helsinki University of Technology (Espoo Finland), 2003.

[8] Thomas Kolenda, Adaptive Tools in Virtual Enviroment, Independent Component Analysis for Multimedia, Doctoral Thesis at Technical University of Denmark, ISSN 0909-3192 2002.

[9] Sridhar, G. R. and Yarabati Venkat,Indian J. Endocrinol. Metab., 2000, 4,70–80.

[10] Sherita Hill Golden, MD,MHS et al ,Examining a Bidirectional Association Between Depression Symptoms and Diabetes, JAMA, 2008;(23):2751-2751-2759.

Author Biographies

T.Sudha
She is professor and head of the Computer Sceince Department of Vikram Simhapuri University.. She is member of CSI. She organized several UGC, AICTE Sponsored workshops, and conferences. She published several International papers and guided many Research Scholars.

S.Sagar Imambi
She is post graduate from the Madras University, Chennai. She is presently working as Asst . Professor in TJPS COLLEGE(P.G). Her research interests involve innovative methods of data mining. She is an author and coauthor of many publications on information processing and machine learning, including Predictive Data Mining