International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 12, December-2015

Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM

Parul Sinha
Department of Information Technology,
Barkatullah University Institute of Technology,
Bhopal, India

Poonam Sinha
Department of Information Technology,
Barkatullah University Institute of Technology,
Bhopal, India

***Abstract*— Chronic kidney disease (CKD), also known as chronic renal disease. Chronic kidney disease involves conditions that damage your kidneys and decrease their ability to keep you healthy. You may develop complications like high blood pressure, anemia (low blood count), weak bones, poor nutritional health and nerve damage. . Early detection and treatment can often keep chronic kidney disease from getting worse. Data Mining is the term used for knowledge discovery from large databases. The task of data mining is to make use of historical data, to discover regular patterns and improve future decisions, follows from the convergence of several recent trends: the lessening cost of large data storage devices and the ever-increasing ease of collecting data over networks; the expansion of robust and efficient machine learning algorithms to process this data; and the lessening cost of computational power, enabling use of computationally intensive methods for data analysis. Machine learning, has already created practical applications in such areas as analyzing medical science outcomes, detecting fraud, detecting fake users etc. Various data mining classification approaches and machine learning algorithms are applied for prediction of chronic diseases. The objective of this research work is to introduce a new decision support system to predict chronic kidney disease. The aim of this work is to compare the performance of Support vector machine (SVM) and K-Nearest Neighbour (KNN) classifier on the basis of its accuracy, precision and execution time for CKD prediction. From the experimental results it is observed that the performance of KNN classifier is better than SVM.**

Keywords—Data Mining, Machine learning, Chronic kidney disease, Classification, K-Nearest Neighbour, Support vector machine.

I. INTRODUCTION

Data mining deals with extraction of useful information from huge amounts of data. Many other terms are being used to understand data mining, such as mining of knowledge from databases, knowledge extraction, data analysis, and data archaeology. Basically, data mining is a crucial step in the process of knowledge discovery in databases, or KDD. The data mining techniques of classification, clustering and association helps in extracting knowledge from large amount of data.

Machine Learning is a rising field concerned with the study of huge and multiple variable data. It is evolved from the study of pattern recognition and computational learning theory in artificial intelligence and involves computational methods, algorithms and techniques for analysis. In Medical Science's perspective, Machine Learning promises to aid physicians make near-perfect diagnoses, opt the best medications for their patients, spot patients at high-risk for pitiable outcomes, and specifically improving patients' physical condition while minimizing costs.

Machine learning and data mining techniques together have proved success in prediction and diagnosis of various critical diseases. There are various applications for Machine Learning, the most vital of which is data mining. Machine learning along with data mining can often be effectively applied to such problems, as they improve the efficiency of the systems and their designs. Same set of features are used for the representation of every instance, in any dataset used by Machine learning algorithms. These features can be continuous, categorical or binary. If the instances are given with class labels or known labels i.e. with the corresponding correct outputs ,then the learning is called supervised learning, on the other hand comes unsupervised learning, where instances or class are unlabeled. Researchers hope to discover a lot of information using these supervised and unsupervised learning.

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. [Oracle Data Mining Concepts 11g Release 1 (11.1)]

Various data mining classification approaches and machine learning algorithms are applied for prediction of chronic diseases. Here we are concerned about Chronic kidney disease (CKD), also known as chronic renal disease, is an abnormal function of kidney or a progressive failure of renal function over a period of months or years. Often, chronic kidney disease is diagnosed as a result of screening of people known to be at risk of kidney problems, such as those with high blood pressure or diabetes and those with a blood relative with CKD. It is differentiated from acute kidney disease in that the reduction in kidney function must be present for over 3 months.This work predominantly focused on, prediction of chronic kidney disease. Chronic Kidney disease is predicted using classification techniques of data mining. The classifiers used here are, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) classifier. Their performance is then evaluated based on accuracy, precision and F-measure.

II. LITERATURE SURVEY

In 2015, Konstantina Kourou et.al [1] proposed a study of Machine learning applications in cancer prognosis and prediction. In this paper, they have presented a review of various recent ML approaches that are applied for the prediction of cancer detection. Here they have presented review of newly published content for the work done so far in cancer detection.

In 2015 P.Swathi Baby et. al [2] proposed a project to diagnosis and prediction system based on predictive mining. Here kidney disease data set is used and analysed using Weka and Orange software. Here the Machine learning algorithms such as AD Trees, J48, K star , Naïve Bayes, Random forest are used for the performance study of each algorithm which gives the Statistical analysis and predicting kidney diseases using the algorithms. Their observation shows that the best algorithms K-Star and Random Forest for the used Dataset ,where Build the models are less time(0 sec and 0.6 sec) and the ROC values are 1.

In 2014 K.R.Lakshmi et.al [3] proposed performance evaluation of three data mining techniques for predicting kidney dialysis survivability. In this research, various data mining techniques (Artificial Neural Networks, Decision tree and Logical Regression) are used to extract knowledge about the interaction between these variables and patient survival. A performance comparison of three data mining techniques is applied for extracting knowledge. The concepts introduced in this research have been engaged and tested using a data collected at different dialysis sites. The outcomes are reported. Finally, ANN is suggested for Kidney dialysis to get better results with accuracy and performance.

Shital Shah et.al [4] projected a research on predicting survival of kidney dialysis patients using data mining techniques. In this research, a data mining approach is used to extract knowledge about the interaction between these variables and patient survival. Two different data mining algorithms are employed for extracting knowledge in the form of decision rules. Data mining is performed on the individual visits of the "most invariant" patients as they form "signatures" for their decision categories. It concludes that the overall classification accuracy for all data mining algorithms was significantly higher using the individual visit data set over the aggregate data set. The prediction accuracy of individual visit based rule sets increased over the aggregate based rule sets.

In 2015,Mr. S Dayanand [5]proposed the research work to predict kidney diseases by using Support Vector Machine (SVM) and Artificial Neural Network (ANN). The aim of this work is to compare the performance of these two algorithms on the basis of its accuracy and execution time. From the experimental results it is observed that the performance of the ANN is better than the other algorithm.

LIMITATIONS

The existing prediction system for chronic kidney disease is fine with some limitations. Below is the table shown, describing the worked done for prediction and detection of various kidney diseases. A new CKD prediction system is still the need. A decision support system for chronic kidney disease is still the need for early prediction, as not much work is done for the same.

III. METHODOLOGY

The work proposed here uses three classification techniques to predict the presence of chronic kidney disease in humans. The classifiers used are Support vector machine and KNN classifier. The data set for chronic kidney disease was gathered and applied on each classifier to predict the disease and the performance of the classifier is evaluated based on accuracy, precision and F measure.

Architecture of Predictive Data Mining: Proposed Approach

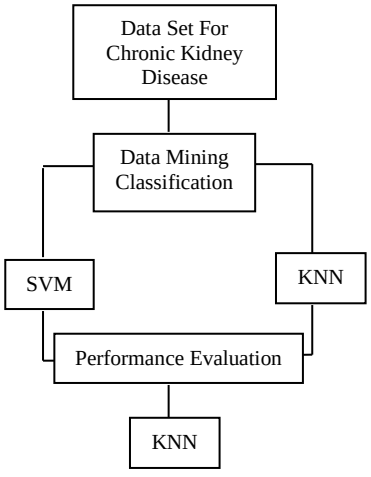

Fig3.1. Proposed System

The working of the architecture is as follows: The dataset for CKD patients have been collected and fed into the classifier named SVM and KNN. The prediction of CKD will be executed with the help of a tool known as Matlab. In this paper, the dataset is collected from UCI machine learning repository, as the input for prediction. The dataset consists of attributes and values. This tool will results the accuracy that how many patients are having the chronic kidney disease with in a particular time. In order to improve the rate of prediction, comparison of the two classifiers is done based on evaluation parameters. The experimental result is retrieved, which shows the best classifier between the two.

Evaluation parameters

Some of the data mining parameters are:

A. Sensitivity It is also called True Positive Rate. It is used for measuring the percentage of unwell people from the dataset.

Sensitivity = Number of true positives/Number of true positives + Number of false negatives

B. Specificity It is also called True Negative Rate. It measures the percentage of healthy people that are exactly recognized from the dataset.

Specificity = Number of true negatives/Number of true negatives + Number of false positives

C. Precision and recall

It is also called positive predictive value. It is defined as the average probability of relevant retrieval.

Precision = Number of true positives/Number of true positives + False positives

Recall

It is defined as the average probability of complete retrieval. Recall= True positives/True positives + False negative

D. Accuracy

Accuracy is defined in terms of correctly classified instances divided by the total number of instances present in the dataset. Accuracy=Number of correctly classified samples/Total number of samples

E. Confusion Matrix It displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. The matrix is represented in the form of n-by-n, where n is the number of classes. The accuracy of each classification algorithms can be calculated from that.

Data set

The dataset is collected from several medical labs, centres and hospitals. From this the mock kidney function test (KFT) dataset have been formed for study of kidney disease. This dataset contains four hundred instances and twenty five attributes are used in this comparative study. The attributes in this KFT dataset age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, haemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anaemia, class. This dataset consists of renal affected disease information.

This is binary classification, as we have used two classes for predicting CKD and NOT CKD.

Data Mining Techniques

- Support Vector Machines

Support Vector Machines (SVM) is a powerful, state-of-the-art algorithm based on linear and nonlinear regression. Oracle Data Mining implements SVM for binary and multiclass classification. The advantage of the SVM is that, by use of the so-called ‘‘kernel trick’’, the distance between a molecule and the hyper plane can be calculated in a transformed (nonlinear) feature space, lacking of the explicit transformation of the original descriptors. The radial basis function kernel (Gaussian kernel) which is the most commonly used was applied to this study.

- *K-nearest neighbor Classification*

In pattern recognition, the K-Nearest Neighbor algorithm (K-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the K closest training examples in the feature space. K-NN is a type of instance-based learning. In K-NN Classification, the output is a class membership. Classification is done by a majority vote of neighbours. If K = 1, then the class is single nearest neighbor. In a common weighting scheme, individual neighbour is assigned to a weight of 1/d if d is the distance to the neighbour. The shortest distance between any two neighbours is always a straight line and the distance is known as Euclidean distance [7]. The limitation of the K-NN algorithm is it’s sensitive to the local configuration of the data. The process of transforming the input data to a set of features is known as Feature extraction. In Feature space, extraction is taken place on raw data before applying K-NN algorithm. The steps involved in a K-NN algorithm:

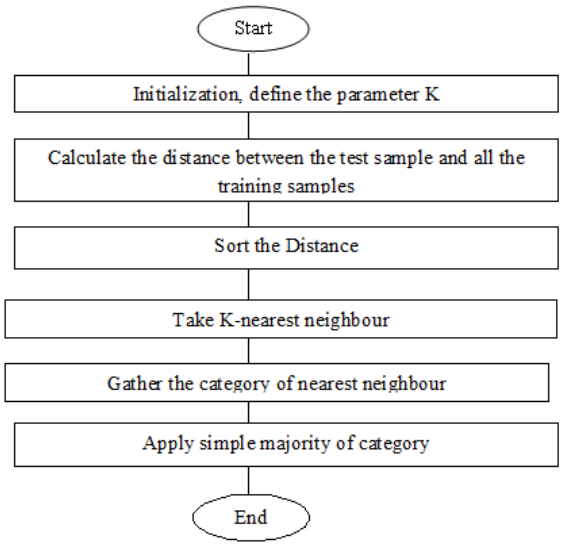

Fig. 4.1. K-NN working

IV. ALGORITHMS AND EVALUATION

Below are the figures showing performance of various evaluation parameters. The X-axis denotes the percentage of performance achieved whereas the Y-axis denotes ratio of data set taken for analysis. For evaluation the following algorithms are run:

Here *l* = number of clusters, *tp* = true positive, *fp* = false positive, *fn* = false negative, *tn* = true negative.

$$\textit{Average Accuracy} = \frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{l}$$

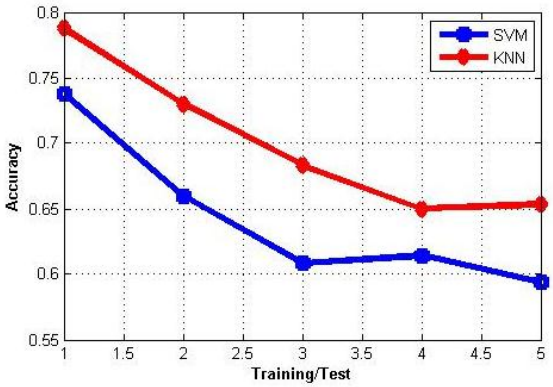

Fig. 5.1 Accuracy graph

$$\textit{F-Measure} = \frac{(\beta^2 + 1)\text{precision}_M \text{recall}_M}{\beta \text{precision}_M \text{recall}_M}$$

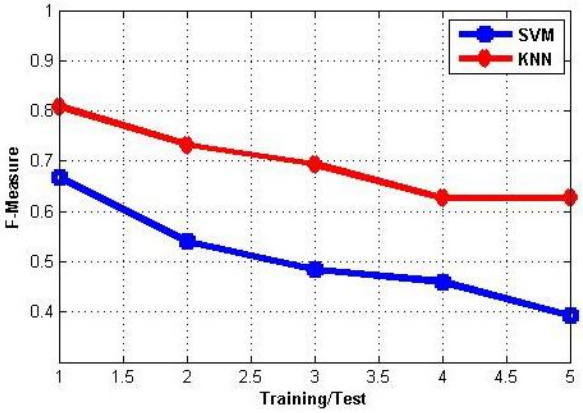

Fig. 5.2 F-measure graph

$$\textit{Precision} = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l}$$

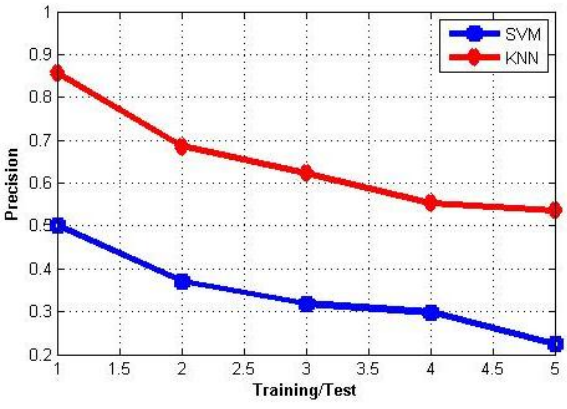

Fig. 5.3 Precision graph

$$\textbf{\textit{recall}}_M = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l}$$

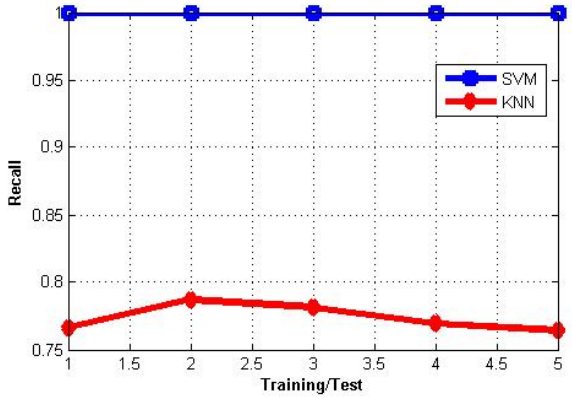

Fig. 5.4 Recall graph

The above figures shows that KNN performed better in terms of accuracy, precision and f measure over different datasets, whereas SVM shows good result in calculating recall value. Thus we can say that KNN performed better than SVM in prediction of CKD in our analysis.

V. EXPERIMENTAL RESULTS

This work is performed in Matlab tool, developed by MathWorks. MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages. The experimental comparison of KNN and SVM are done based on the performance measures of classification accuracy and precision.

Datasets and Preprocessing

The datasets are extracted from UCI Machine learning repository benchmarks. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. The archive was created as an ftp archive in 1987 by David Aha and fellow graduate students at UC Irvine. Since that time, it has been widely used by students, educators, and researchers all over the world as a primary source of machine learning data sets.

The table below shows the experimental result analysis of SVM & KNN on CKD dataset:

TABLE I.

Name of Classifier	Evaluation Parameter			
	Accuracy	*Precision*	*Recall*	*F-measure*
KNN	.7875	.8571	.7660	.8090
SVM	.7375	.5000	1	.6670

(Result Analysis)

VI. CONCLUSION

As we have already seen the applications of data mining and machine learning in medical sector. In this paper, a new decision support system is implemented for prediction of CKD. Although the classifiers worked efficiently in prediction of other diseases also. In this paper, Chronic Kidney Disease is predicted using two different classifiers and a comparative study of their performance is done. From the analysis we found that, out of two classifiers SVM and KNN, KNN classifier performed better than the other. The rate of prediction of CKD is improved.

VII. FUTURE WORK

There are other possible evolutionary techniques that may be used to improve results of the proposed classifiers. In this paper, SVM and KNN are applied to detect CKD. We can also evaluate and compare the performance of the used classifiers with other existing classifiers. CKD early detection helps in timely treatment of the patients suffering from the disease and also to avoid the disease from getting worse. Early prediction of the disease and timely treatment are the need for medical sector. New classifiers can be used and their performance can be evaluated to find better solutions of the objective function in future work.

REFERENCES

[1] Konstantina Kourou et.al, "Machine learning applications in cancer prognosis and prediction" Computational and structural bitechnology Journal, Elsevier.

[2] P.Swathi Baby, T. Panduranga Vital ,"Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms" International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-018, Vol. 4 Issue 07, July-2015,206-210.

[3] K.R.Lakshmi1, Y.Nagesh2 and M.VeeraKrishna3, "Performance Comparision of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability", International Journal of Advances in Engineering & Technology, Mar. 2014, Vol. 7, Issue 1, pp. 242-254.

[4] AndrewKusiak, Bradley Dixonb, Shital Shaha, (2005) Predicting survival time for kidney dialysis patients: a data mining approach, Elsevier Publication, Computers in Biology and Medicine 35, page no 311–327

[5] Dr. S. Vijayarani, Mr.S.Dhayanand, "Kidney Disease Prediction Using SVM and ANN Algorithms" IJCBR , ISSN (online): 2229-6166,Volume 6 Issue 2 March 2015.

[6] Mahfuzah Mustafa, Mohd Nasir Taib et.al, "Comparison between KNN and ANN Classification in Brain Balancing Application via Spectrogram Image" Journal of Computer Science & Computational Mathematics, Volume 2, Issue 4, April 2012,pp 17-22 .

[7] Ross KK Leung, Ying Wang et.al, "Using a multi-staged strategy based on machine learning and mathematical modeling to predict genotype-phenotype risk patterns in diabetic kidney disease: a prospective case–control cohort analysis" BMC Nephrology 2013, pp 1-9.

[8] Bendi Venkata Ramana, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis" International Journal of Database Management Systems(IJDMS), Vol.3, No.2, May 2011 (pp101–114).

[9] M. Bramer, Principles of Data Mining: Springer-Verlag, 2007.

[10] H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, 2005

[11] DSVGK Kaladhar, Krishna Apparao Rayavarapu* and Varahalarao Vadlapudi,"Statistical and Data Mining Aspects on Kidney Stones: A Systematic Review and Meta-analysis", Open Access Scientific Reports, Volume 1 • Issue 12 • 2012.

[12] Jinn-Yi Yeha, Tai-Hsi Wu et.al," Using data mining techniques to predict hospitalization of hemodialysis patients"Elsevier, Volume 50, Issue 2, January 2011, Pages 439–448.

[13] DSVGK Kaladhar, Krishna Apparao Rayavarapu* and Varahalarao Vadlapudi,"Statistical and Data Mining Aspects on Kidney Stones: A Systematic Review and Meta-analysis", Open Access Scientific Reports, Volume 1 • Issue 12 • 2012.

[14] J.Van Eyck, J.Ramon, F.Guiza, G.Meyfroidt, M.Bruynooghe, G.Van den Berghe, K.U.Leuven," Data mining techniques for predicting acute kidney injury after elective cardiac surgery", Springer, 2012.

[15] K.R.Lakshmi, Y.Nagesh and M VeeraKrishna,"Performance comparison of three data mining techniques for predicting kidney disease survivability", International Journal of Advances in Engineering & Technology, Mar. 2014.

[16] Morteza Khavanin Zadeh, Mohammad Rezapour, and Mohammad Mehdi Sepehri," Data Mining Performance in Identifying the Risk Factors of Early Arteriovenous Fistula Failure in Hemodialysis Patients", International journal of hospital research, Volume 2, Issue 1,2013, pp 49-54.

[17] Xudong Song, Zhanzhi Qiu, Jianwei Mu," Study on Data Mining Technology and its Application for Renal Failure Hemodialysis Medical Field", International Journal of Advancements in Computing Technology(IJACT) ,Volume4, Number3, February 2012.

[18] N. Sriraam, V. Natasha and H. Kaur," data mining approaches for kidney dialysis treatment",Journal of Mechanics in Medicine and Biology, Volume 06, Issue 02, June 2006.

[19] Salha M. Alzahani, Afnan Althopity et.al, "An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction", Taif University, Taif, Saudi Arabia 310-315.

[20] Anu Chaudhary, Puneet Garg,(2014) Detecting and Diagnosing a Disease by Patient Monitoring System, International Journal of Mechanical Engineering And Information Technology, Vol. 2 Issue 6 //June //Page No: 493-499.

[21] Approaches, Knowledge-Oriented Applications in Data Mining, Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-154-1,InTech,http://www.intechopen.com/books/knowledge-oriented-applications-indatamining/mining-enrollment-data-using-descriptive-and-predictive-approaches.

[22] Fadzilah Siraj, Mansour Ali Abdoulha, (2011). Mining Enrollment Data Using Descriptive and Predictive Mining.

[23] George Dimitoglou, Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability

[24] Giovanni Caocci, Roberto Baccoli, Roberto Littera, Sandro Orrù, Carlo Carcassi and Giorgio La Nasa, Comparison Between an Artificial Neural Network and Logistic Regression in Predicting Long Term Kidney Transplantation Outcome, Chapter 5, an open access article distributed under the terms of the Creative Commons Attribution License.

[25] Gualtieri. J. A, Chettri. S. R, Cromp. R. F and Johnson.L. F, (1999) Support vector machine classifiers as applied to AVIRIS data, in Summaries 8th JPL Airborne Earth Science Workshop, JPL Pub. 99-17, pp. 217–227.

[26] Ian H. Witten and Eibe Frank.(2005) Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition.

[27] B.Venkatalakshmi, M.V Shivsankar, "Heart Disease Diagnosis Using Predictive Data mining", IJIRSET Volume 3, Special Issue 3, March 2014 ,pp. 1873-1877.

[28] Suman Bala, Krishan Kumar, "A Literature Review on Kidney Disease Prediction using Data Mining Classification Technique", IJCSMC, Vol. 3, Issue. 7, July 2014, pg.960 – 967.

[29] Neha Sharma, Hari Om "Data mining models for predicting oral cancer survivability" Springer-Verlag Wien 2013, Netw Model Anal Health Inform Bioinforma (2013) 2:285–295.