International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 4, 2013. ISSN: 2348-6600
Reference ID: IJCS-022 PAGE NO: 116-122

INFORMATION EXTRACTION OF PREDICTING BLOOD CANCER

N.Sudha Bhuvaneswari [#1], S.Yamuna [*2]

[#1] Associate Professor MCA,Mphil(CS),(PhD), [#2] M.Phil(CS) Research Scholar
School Of IT&Science, Dr.G.R.Damodaran college of science,Coimbatore.
Email: sudhanarayan03@yahoo.com, yamunaa.sk@gmail.com

Abstract— **Data mining techniques have been used in medical research for many years and have been known to be effective. Cancer has become the leading cause of death worldwide. Cancer is one of the dreadful diseases in the world claiming majority of lives. Though cancer research is generally clinical and biological in nature, data driven statistical research has become a common complement. Predicting the outcome of a disease is one of the most interesting and challenging tasks where data mining techniques have to be applied[1]. Implementing accuracy in this classification by finding out the ratio of cancer in male versus female cases and also which areas record highest cancer rate and whether habits, diets, education, marital status, living area etc., which play important roles in cancer pattern.**

***Keywords*— ID3, classification, decision learning, Common Disease, Prediction.**

INTRODUCTION

Cancer begins when cells in a part of the body start to grow out of control. There are many kinds of cancer, but they all start because of out-of-control growth of abnormal cells. Cancer cell growth is different from normal cell growth. Instead of dying, cancer cells continue to grow and form new, abnormal cells. Cancer cells can also invade (grow into) other tissues, something that normal cells cannot do. Growing out of control and invading other tissues are what makes a cell a cancer cell.

In this we preprocess a data set which contain 18 attributes and approximately 4600 cases of cancer recorded in each of the following years, 1998, 2000, 2005 and 2010 and classify them with data mining techniques. Data has been useful in showing Over all trends in treatment but has only until recently been applied to evaluating how to best treat individual patients. Models built on patient level data using data mining techniques to patient level cancer datamodeling of the relationships between a patient's history,medical record, disease stage, and outcome how to better treat cancer patients. To this classification append the results of Linear Regression using the method of least squares and predict the recorded outcome for the year 2013. The Appender takes a dataset and Classifier and appends the classifiers predictions to the dataset. As a next preprocess a dataset for the year 2013 and cross examine the classification results with our predictions and analyze whether our predictions are accurate and whether ID3 truly helps in cancer prediction. ID3, Iterative Dichotomiser 3 is a decision tree learning algorithm which is used for the classification of the objects with the iterative inductive approach. In this algorithm the top to down approach is used[2]. The top node is called as the root node and others are the leaf nodes. So it"s a traversing from root node to leaf n odes. Each node requires some test on the attributes which decide the level of the leaf nodes. These decision trees are mostly used for the decision making purpose.

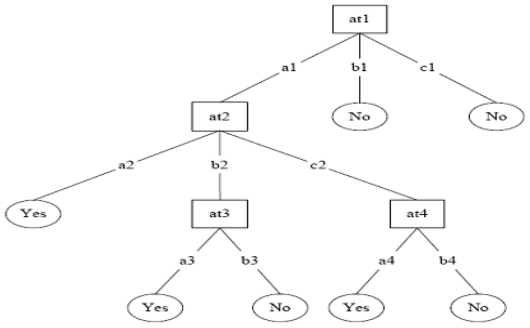


Fig.1 Decision Tree

II. LITERATURE REVIEW

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 4, 2013. ISSN: 2348-6600
Reference ID: IJCS-022 PAGE NO: 116-122

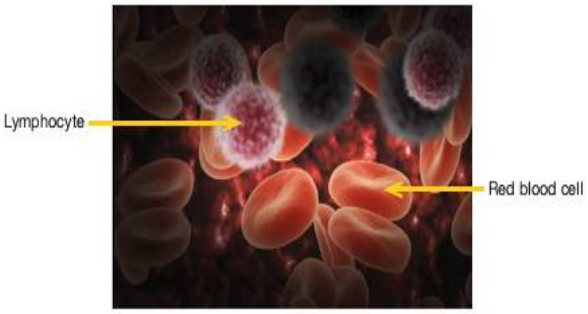
The first work on ID3 was done by J.R Quinlan in 1986.He summarizes an approach tosynthesizing decision trees that has been used in a variety of systems, and it describes one such system , ID3, in detail. Results from recent studies show ways in which the methodology can be modified to dealwith information that is noisy and/or incomplete. A model blood cancer Prediction System built with the aid of data mining techniques like Decision Trees, Naïve Bayes and Neural Network. The results illustrated the peculiar strength of each of the methodologies in comprehending the objectives of the specified mining objectives. ID3 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 blood disease. The prediction of Blood cancer, Blood Pressure and Sugar with the aid of neural networks was proposed by Niti Guru et al. Experiments were carried out on a sample database of patients' records. The Neural Network 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. Whenever unknown data was fed by the doctor, the system identified the unknown data from comparisons with the trained data and generated a list of probable diseases.The problem of identifying constrained association rules for Blood cancer prediction was studied by Carlos Ordonez. The assessed data set encompassed medical records of people having Blood cancer with attributes for risk factors.Data mining methods may aid the clinicians in the predication of the survival of patients and in the adaptation of the practices consequently. The work of Franck Le Duff et al. might be executed for each medical procedure or medical problem and it would be feasible to build a decision tree rapidly with the data of a service or a physician. Comparison of traditional analysis and data mining analysis illustrated the contribution of the data mining method in the sorting of variables and concluded the significance or the effect of the data and variables on the condition of the study[3].

III. BLOOD CELLS AND BLOOD CANCER

There are many different types of cells found in the blood. These cells are made in the bone marrow. The main types of blood cell are white blood cells, red blood cells and platelets:white blood cells are part of the immune system and help the body to fight infection, red blood cells carry oxygen around thebody and platelets are fragments of cells that help the blood to clot e.g. if you cut yourself and are bleeding.

Fig.2 Red blood cells and lymphocytes in the bloodsteam



Red blood cells and lymphocytes in the bloodstream

All blood cells start off as stem cells and develop into one of these three types. The process of producing blood cells is called haematopoiesis. Haematopoiesis makes sure that the correct amount of each type of cell is produced. This is important because different types of blood cells live for different lengths of time and need to be replaced at the end of their life cycle[5]. Also, more of one cell type might be needed at a certain time (e.g. more white blood cells are needed when you have an infection).

IV. HAEMATOPOIESIS AND BLOOD CANCER

Haematopoiesis can sometimes go wrong. Cells may be produced in abnormally high numbers, or stay in the blood for an abnormal length of time, or both. When this happens, it results in there being far too many of a certain type of blood cell in the bone marrow or bloodstream, causing serious problems for the patient. This is known as blood cancer.

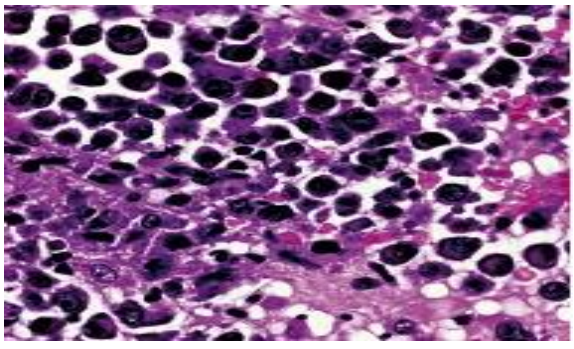


Fig.3 Lymph gland with high-grade NHL

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 4, 2013. ISSN: 2348-6600
Reference ID: IJCS-022 PAGE NO: 116-122

Diagnosis results		Blood test results		Physical exam results	
Disease Name	Classification code	Item Name	Classification code	Item Name	Classification code
Diabetes Mellitus	dm(D)	Heartbeat Number	hr_la	Age	age
Hypertension	hp(B)	HbA1c	hba1c_1	Gender	Sex
Immuno adsorption	imm(B)	Blood urea nitrogen	bun_1	Blood type	Btype
Hyper Lipemia	lip(B)	Cholesterol	cho_1	Marital condition	Mar
Ischemic heart disease	hd(H)	Triglyeride	tri_1	Smoking habit	Smoke
Myocardial infarction	mi(H)	Fasting Glucose	glu_1	Drinking habit	Drink
Arrhythmia	arr(H)	High density cholesterol	hdl_1	Body mass index	bmi_1
Cardiogenic Shock	car(H)	Low density cholesterol	ldl_1	Percentage body Fat	bfat_1
		Coagulation disorders	pt_1	Waist-and-hip ratio	wb_1
		Partial coagulation disorders	aptt_la		
		Albumin	alb_1		
		Urine Acid	ua_1		

V. SYMPTOMS AND STAGING

Stage 1 The myeloma is at an early stage.
• The number of red blood cells is normal or only slightly reduced.
• The amount of calcium in the blood is normal
• There are low levels of abnormal antibodies in the blood or urine.
• The bones appear normal or there is a single cancerous plasma cell
• There may not be any symptoms.
Stage 2 The myeloma is at an intermediate stage.
Stage 3 The myeloma has caused one or more of the following to happen:
• Severe anaemia
• A high level of calcium in the blood.
• The myeloma has affected three or more bones :fractures may occur.
• High levels of abnormal antibodies in the blood or urine

VI. DATA SET

Table I
Blood Sample Report

Data set was based on data generously provided by the Alberta Cross Cancer Institute. The original data contains 3647 patients of different types of cancers. For most patients, we receive three categories of data, which includes blood test results, clinical data, and weight-loss information. At the end, 34 attributes are selected according to the suggestion of medical experts[6]. Similar to many medical data sets need to preprocess the data.

VII. CLASSIFICATION TECHNIQUE

Classification is a concept of finding a model which finds the class of unknown objects. It basically maps the data items into some predefined classes.Classification model generate set of rules based on the features of the data in the training dataset. These rules can be use for classification of unknown data items. Medical diagnosis is an important feature of classification for eg: Diagnosis of patients based on their symptoms by using the classification rules about diseases from known cases.
The classification is basically to define a
Function: f = D→ C where each ti € D is mapped to f (ti) belonging to some Cj [3].
1. D is a database of patients with tuples (x1, x2 … xn)
2. x1, x2 … xn are values of attributes A1,A2 … .An relevant to a particular disease.
3.C= {C1, C2 … Cn} is set of classes of disease depending on its severity.
Decision tree is a method of implementing the classification. Decision trees have become one of the most powerful and Datamining and knowledge discovery.Decision tree is used as a predictive model. More descriptive names for such tree models are classification trees or regression trees. Decision trees need two kinds of data: Testing data and Training data are usually the bigger part of data, are used for constructing trees. The more training data collected higher the accuracy of the results[4]. The other group of data, testing is used to get the accuracy rate and misclassification rate of the decision tree.

A.PREPROCESSING

Preprocessing of medical data was required before starting analysis. First need to match, combine, and validate data from several different data sets. Additionally, there were many

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 4, 2013. ISSN: 2348-6600
Reference ID: IJCS-022 PAGE NO: 116-122

missing values in the data. To avoid uncertainty from imputation and based on the fact that we have enough data for different ranges of survivability.

VIII. ID3 ALGORITHM APPROACH

Function ID3 (symptoms, part, symptoms set, treatment, disease, sample)
{
/* symptoms is the set of input attributes
/* part is the set of input attributes
/* symptoms set is the set of input attributes
/* treatment is the set of output attributes
/* disease is the set of output attributes
/* sample is a set of training data
/* Function ID3 returns a decision tree
*/ if (sample is empty)
{
Compute the information gain for each attribute in part relative to sample.
Let x be the attributes with largest gain (x, sample) at this attributes in parts;
Trees ID3 (part-{x}; Symptoms (Sample-1)... ID3 (test-(sample-n)} ;
Output: Dataset value will be retrieved.

B. PREDICTION PROCEDURE

1. Select the dataset for which the test to be retrieved.
2. By using the ID3 algorithm sort the specific pattern and classify the datasets based on the symptoms.
3. Then preprocess the fields of dataset based on "Symptom" field and diagnosis the causes and treatment of the disease also.

C.BLOOD GROUP

Dataset shown most of the cancer patients belong to the blood group O (+ve). This pattern is observed including the year 2013.This may be due to the fact that most of the people throughout the world have O (+ve) as their blood group. It is followed by A (+ve), B (+ve), B (–ve), AB (+ve), A (–ve), and O (–ve) in decreasing order. For example in the year 2000 out of 2893 cancer records 1246 records belong to O (+ve), 518 records belong to A (+ve), 1512 records belong to B (+ve), 401 records belong to AB (+ve), 46 records belong to A (–ve), 82 records belong to B (–ve), and 19 records belong to O (-ve).

D.LINEAR REGRESSION

Linear Regression which used the formula,
$Y= a + bX$
Where a and b are constants and Y is response variable and a single predictor variable X. We found these constants using the Method of Least Squares as linear regression follows a straight line path. To find out the constants we use two more equations,
$\sum xy = a\sum x + b\sum x - 2$
$\sum y = na + b\sum x - 1$

E. THIN DATABASE

The THIN database consists of a collection of information participating general practices including information of patients such as their year of birth, gender and family connections and the demographics of the area they live in. The database which also contains temporal information detailing a patient's prescription and medical histories. For this comparison a database containing records from 875 general practices was used. This subset of the THIN database contained approximately five million patients, over 628 million prescription entries and over 522 million medical event entries From all four years of data where classified the record that is above 30 years of age as cancer[7]. Among all th e different types of cancer that is in the age range of 20-30 the top three cancers aremelanoma, Leukemia (myeloid, lymphoma), followed by lymphoma (non Hodgkin, Hodgkin) and Eye cancer.

TABLE II
Year Wise Cancer Test Report

ICD	1998	2000	2005	2010
Leukemia	10	17	12	18
Melanoma	8	14	14	11
Lymphoma	12	5	16	15
Bone	11	19	7	10
Eye	6	13	11	12
Endometrial	4	11	9	11

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 4, 2013. ISSN: 2348-6600
Reference ID: IJCS-022 PAGE NO: 116-122

From this data table we calculate and predicted the results for the year 2013 as below

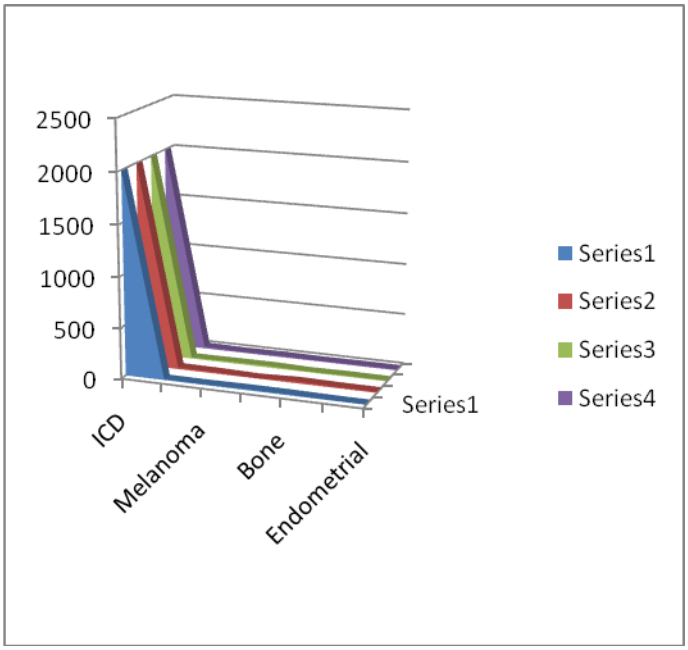
ICD	Leukemia	Melanoma	Lymphoma	Bone	Eye	Endometrial
2013	15	11	12	13	9	4

TABLE III
Cancer Report for the year 2013

ICD	Leukemia	Melanoma	Lymp Homa	Bone	Eye	Endometrial
2013	15	11	12	13	9	4

Using linear regression table were compiled for the year 2013. Based on the results we expected the number of blood cancer cases in the year 2013 to be 92 out of 2860 cases. Age group between 20-30 is the crucial period in which youth attain puberty as well as they fall into bad company acquiring bad habits such as smoking and drinking which have greater impact during old age and results in different types of cancer. According to results this is the period during which cancer instances in a steep rise. Above 50 years there is a decline of cancer rates in men and women this may be due to most of the people die of recurring cancer before they reach the age of 60 due to some factors like cervix, heart attack,breast esophagus and lungs again occupy the top positions in the age[8]. Thus we can see prediction yields approximate results in cancer dataset which is more or less similar to our observed results and we can use linear regression for our other prediction works.

Fig.4 A Sample graph



F. RESULTS AND DISCUSSION

Training data collected over five years using ID3 algorithm we need to predict blood cancer.

TABLE IV
Training Data for finding blood cancer

S.no	Leukemia	Melanoma	Lymphoma	Bone	Eye	Endometrial
1	15	11	12	6	9	4
2	10	14	9	7	12	11
3	12	11	6	8	10	9
4	11	13	14	12	8	12
5	14	10	9	7	5	14
6	8	2	8	9	7	11
7	10	7	4	12	10	9
8	11	14	9	7	5	4
9	14	13	10	12	11	7
10	9	16	5	11	6	3

G. LEARNING SET

ID3 requires dataset to discrete.

TABLE V

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 4, 2013. ISSN: 2348-6600
Reference ID: IJCS-022 PAGE NO: 116-122

Types of Cancer and its Possible values

ATTRIBUTE	POSSIBLE VALUES		
Leukemia	10	14	7
Melanoma	14	13	11
Lymphoma	9	10	5
Bone	7	12	-
Eye	12	11	-
Endometrial	11	7	-

Assign discrete values to the attributes Partition the continuous attribute values to make them discrete. It was also possible to investigate how well depending on how common risk of each separately listing frequently, less frequently and rarely occurring known to calculated the ID3 score of each considering rarely occurring for three different cases[9].

TABLE VI
Types of Cancer

S.no	Leukemia	Melanoma	Lymphoma	Bone	Eye	Endometrial	Disease
1	15	11	12	6	9	4	Yes
2	10	14	9	7	12	11	No
3	12	11	6	8	10	9	No
4	11	13	14	12	8	12	Yes
5	14	10	9	7	5	14	No
6	8	2	8	9	7	11	No
7	10	7	4	12	10	9	Yes
8	11	14	9	7	5	4	Yes
9	14	13	10	12	11	7	Yes
10	9	16	5	11	6	3	Yes

IX. ID3 CALCULATIONS

STEP 1:"example" set s
The set s of 10examples with 6 yes and 4 no then
Entropy (S) = -(6/10) Log2 (6/10) – (4/10) Log2 (4/10) = 0.820
STEP 2:Attribute
Values can be Leukemia, Melanoma, Lymphoma, Bone, Eye and Endometrial.

Entropy (Leukemia) = -(2/4)xlog2(2/4) – (3/4)xlog2(3/4) = 0.821524
Entropy (Melanoma) = -(1/4)xlog2 (5/4) – (0/4) xlog2 (3/4) = 0.2152
Entropy (Lymphoma) = -(2/6)xlog2 (3/6) – (2/6) xlog2 (2/6) = 0.264
Entropy (Bone) = -(2/4)xlog2(3/4) – (2/4)xlog2(2/6) = 0.27514
Entropy (Eye) = -(4/6)xlog2 (4/6) – (2/6) xlog2 (0/6) = 0.01242
Entropy (Endometrial) = -(2/6)xlog2 (1/6) – (2/6) xlog2 (2/6) = 0.815245
Gain (S, Attribute) = Entropy (S) – (4/10) x Entropy (S Leukemia)
- (4/10) x Entropy (S Melanoma)
- (6/10) x Entropy (S Lymphoma) - (4/10) x Entropy (S Bone)
- (4/10) x Entropy (S Eye)
- (4/10) x Entropy (S Endometrial)
= 0.820 – (6/10) x 0.821524 – (4/10) x 0.2152 – (4/10) x 0. 617452- (6/10) x 0.27514 – (4/10) x 0.01242 – (4/10) x 0. 815245
= 0.24568716

STEP 3:Attribute Leukemia
Leukemia value can be yes or no.
Entropy (S Leukemia) = -(2/6)xlog2 (3/6) – (2/6) xlog2 (2/6)
= 0.821524
STEP 3:Attribute Melanoma
Melanoma value can be yes or no.
Entropy (S Melanoma) = -(1/4)xlog2 (5/4) – (0/4) xlog2 (3/4)
= 0.2152
STEP 4:Attribute Lymphoma
Lymphoma value can be yes or no.
Entropy (S Lymphoma) = -(2/6)xlog2 (3/6) – (2/6) xlog2 (2/6)
= 0.264
STEP 5:Attribute Bone
Bone value can be yes or no.
Entropy (S Bone) = -(2/4)xlog2(3/4) – (2/4)xlog2(2/6)
= 0.27514
STEP 5:Attribute Eye
Eye value can be yes or no.
Entropy (S Eye) = -(4/6)xlog2 (4/6) – (2/6) xlog2 (0/6)
= 0.01242
STEP 5:Attribute Endometrial
Endometrial value can be yes or no.
Entropy (S Endometrial) = -(2/6)xlog2 (1/6) – (2/6) xlog2 (2/6)

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
Oddity...Probe...Reviste...
http://www.ijcsjournal.com Volume 1, Issue 2, No 4, 2013. ISSN: 2348-6600
Reference ID: IJCS-022 PAGE NO: 116-122

= 0.815245

H. SUMMARY OF THE RESULTS:

TABLE VII
Accuracy Table

Entropy (S)	**0.82**
Gain (S, Leukemia)	**0.82152**
Gain (S, Melanoma)	**0.2152**
Gain (S, Lymphoma)	**0.264**
Gain (S, Bone)	**0.27514**
Gain (S, Eye)	**0.01242**
Gain (S, Endometrial)	**0.81525**

Gain (S, Lymphoma) = 0.264 is highest
Therefore, "Lymphoma" attribute is the decision attribute in the root node. This process goes on until all data is classified perfectly or we run out of attributes.
Although research poses difficulties and Problems inexplanation and replication can provide an invaluable alternative to modern western medicine. Whichare only able tocope with symptoms on the entire body and its health system. Our review suggests that ID3 may actually predict cancer and help patients recover from manydifferent diseases at the sametime. This shows that attributs are enough to predict the expected survival timeof a patient over the dataset with this target.

X. CONCLUSION

This research work provides a knowledge on cancer pattern and its factors, previously unknown knowledge is mined from the dataset. Linear Regression can be used for Prediction tasks as it yields more approximate results. In would analyze a dataset comprising of five years and compare the results. The above technique can be successfully applied to the data sets for blood cancer[10]. The early and accurate detection as well as classification of cancer using the ID3 will help in the selection of treatment options.

REFERENCES

[1] Andreeva P., M. Dimitrova and A. Gegov, "Information Representation in Cardiological Knowledge Based System", SAER'06, pp: 23-25 Sept, 2006.

[2] Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer, 2009

[3] Berrar D., Dubitzky W., Granzow M. (eds.), A Practical Approach to Microarray Data Analysis, Kluwer Academic Publishers, Boston, Dec.2007

[4]N. Revathy "Accurate Cancer Classification using Expressions of Very few Genes" International Journal of Computer Applications (0975 – 8887) Volume 14– No.4, January 2011.

[6] Rajaraman Swaminathan "Cancer pattern and survival in a rural district in South India" Cancer Epidemiology 33 (2010) 325–331.

[7]T.Sakthimurugan "An Effective Retrieval of Medical Records using Data Mining Techniques" International Journal of Pharmaceutical Science and Health Care, Issue 2, Volume 2 (April 2012).

[8] Shao XM, Liu GC; ZhouQJ et.al. Effects of Qigong Waiqi on the growth and differentiation ofimplanted tumor cells. Proc. of the 3rd Nat.l Acad. Conf. onQigong. Guangzhou, China.

[9] Shelly Gupta "Performance Analysis of Various Data Mining Classification Techniques on healthcare Data" International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 4, August 2011.

[10]Neha Sharma "Framework for Early Detection and Prevention of oral Cancer Using Data Mining" International Journal of Advances in Engineering & Technology, Sept 2012 pp. 302-310.