International Journal of Research in Engineering, Science and Management
Volume-1, Issue-8, August 2018
www.ijresm.com
ISSN (Online): 2581-5782

Prevention and Detection of Cancer Using Data Warehousing and Data Mining Techniques in Andaman and Nicobar Islands

S. Deepa[1], R. Maruthi[2]

[1]*M.Phil Scholar, Department of Computer Science, Ponnaiyah Ramajayam Institute of Science & Technology, Chennai, India*
[2]*Dean, Department of Computer Science, Ponnaiyah Ramajayam Institute of Science & Technology, Chennai, India*

***Abstract*: Cancer is one of the leading causes of mortality and morbidity worldwide. The common sites of cancer have varied distribution in different geographical locations. This study was conducted to detect and prevent cancer using data mining and warehousing techniques in Andaman and Nicobar Islands.**

Keywords: Andaman & Nicobar Islands, Cancer patients, data mining and warehousing techniques, k- means Algorithm, OLAP, types of cancer, WAM.

I. Introduction

As per the Indian Council of Medical Research (ICMR), in India 11, 17, 2, 69 estimated cancer incidence were recorded in the year 2014 for both the sexes. Whereas the estimated mortality rate were recorded as 4, 91,5,97.

These deaths occur due to the various reasons like unawareness of the fact that cancer can be cured if detected earlier, hesitation by Potential cancer patients to do regular screening and time involved in diagnosis.

As an alternative to the tedious physical storage of resources it is important to develop a data warehouse specific to cancer disease and a data mining model to predict cancer earlier. If a machine learning technique is developed to store a person's medical and general record and predict his predisposition towards cancer, its type and exact diagnostic method, physicians can directly start treatment immediately without wasting the precious time in different methods of diagnosis. There have been multiple data mining techniques in health care and allied industries and specifically with respect to single type of cancer.

In Andaman & Nicobar Islands 335people were affected by various types of Cancer in the year 2014. The present study was carried out for 275 patients in ANIIMS, Port Blair, the only government tertiary care hospital in Andaman and Nicobar Islands.

All the patients diagnosed or suspected with cancer, registered in G.B. Pant Hospital, in the study time period - January 2015 to till 2017. This research focuses on the building of multidimensional cancer data warehouse and development of data mining model for the early detection of seven types of cancer, hence prevention is also possible.

II. Aims and Objectives of the Research

The aim of the study is to develop a multidimensional architectural cancer data warehouse built specifically to store and process cancer-related database which include patient's general and medical records.

The cancer data warehouse is proposed to be built on OLTP and OLAP technologies simultaneously, thereby retrieving necessary information using query engines. A data mining model is also proposed to be built and implemented within the cancer data warehouse which can predict a person's predisposition towards cancer and generate the risk level for a particular type of cancer and the exact method of clinical diagnosis.

III. Review of Related Literature

Extensive literature survey in the field of data mining, data warehousing ingeneral, and the application of these techniques in the medical field, especially in cancer research were done. From the literature review it was learned that use of evolving IT systems in medical sciences to eradicate, diagnose, prevent diseases like cancer and ameliorate the standard of living of patients with such life-threatening diseases has garnered the attention of IT researchers worldwide. Review of the literature on cancer related databases helped to realize the fact that suitable data ware house architecture should be implemented for the efficient functioning of the objective data analysis.

IV. Methodology of Research

Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, and analyst) to make better and faster decisions. A data warehouse (or scale data mart) is a specially prepared repository of data designed to support decision making.

The data comes from operational systems and external sources. To create the data warehouse, cancer data are extracted from source systems like questionnaire, cancer institute database, etc… cleaned (e.g., to detect and correct errors), transformed (e.g., put into subject groups or summarized), and loaded into a data store (i.e., placed into a data warehouse).

The data required for the research were collected from Andaman Nicobar Islands Institute of Medical Science (ANIIMS), Port Blair, Andaman & Nicobar Islands. The study uses data mining techniques such as classification, clustering and prediction to identify potential cancer patients. A multidimensional data warehouse specific to cancer disease is built and implemented and further used for a data mining work

International Journal of Research in Engineering, Science and Management
Volume-1, Issue-8, August 2018
www.ijresm.com
ISSN (Online): 2581-5782

to detect a person's predisposition towards cancer. Finally, a detection and prevention system is developed to analyse the risk levels which help in prognosis.

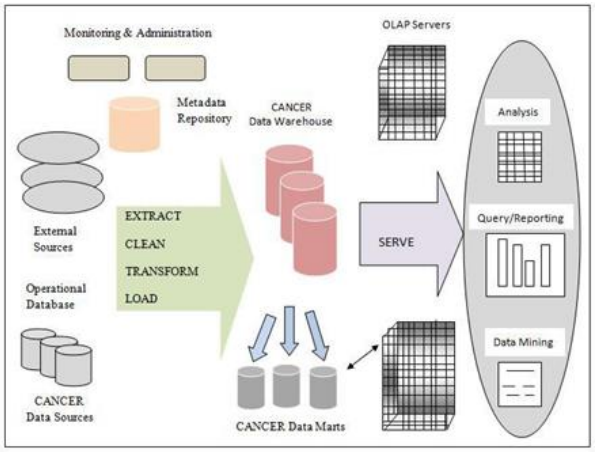

Fig. 1. Cancer data warehousing architecture using ECTL, OLTP, and OLAP servers

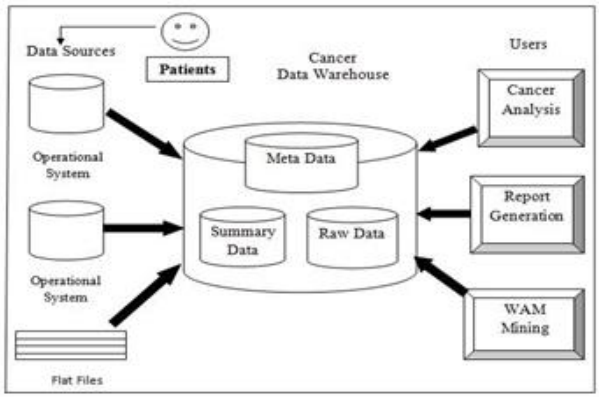

Fig. 2. Cancer data warehouse architecture

A. *Multidimensional Star Schema*

The basic building block used in dimensional modelling is the star schema. A star schema consists of one large central table called the fact table, and a number of smaller tables called dimension tables. The fact table forms the "centre" of the star, while the dimension tables form the "points" of the star. A star schema may have any number of dimensions. The fact table contains measurements (e.g. Patient History, Risk Factor, Cancer, Symptoms, Treatment, and Diagnosis) which may be aggregated in various ways.

The dimension tables provide the basis for aggregating the measurements in the fact table. The fact table is linked to all the dimension tables by one-to-many relationships. The primary key of the fact table is the concatenation of the primary keys of all the dimension tables. The advantage of using star schemas to represent data is that it reduces the number of tables in the database, the number of relationships between them and therefore the number of joins required in user queries.

B. *OLAP Operations of Medical Cancer Data Ware House*

OLAP is performed on cancer data warehouse or cancer disease data marts. The primary goal of OLAP is to support ad hoc query needed to support decision support system. The multidimensional view of cancer data is fundamental to OLAP function. OLAP is a practical view, not a data structure or schema. The complex nature of OLAP process requires a multidimensional review of the cancer data. OLAP Operations in Multidimensional Cancer Data Warehouse (MCDW),

1. Roll-up
2. Drill Down
3. Slice and Dice
4. Pivot

The following database have been created using OLAP operations.

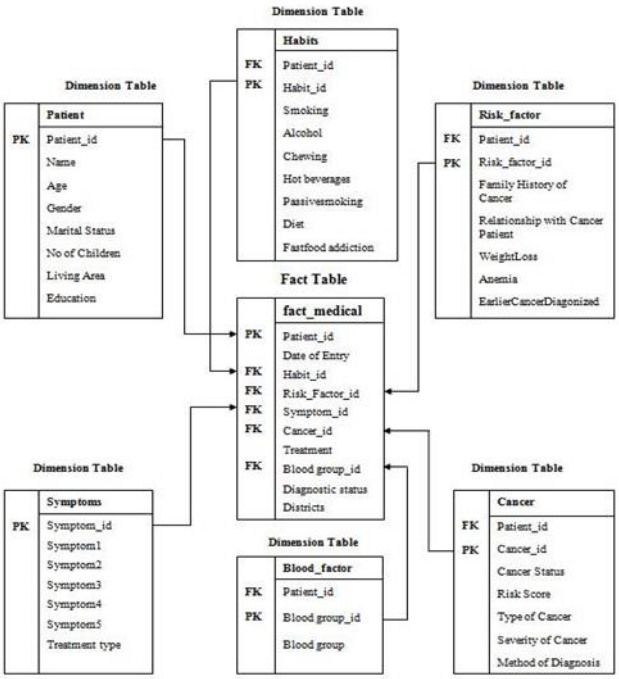

Fig. 3. Star schema representing cancer data ware house

Fig. 4. Dimension creation of specific field of cancer patients from cancer data warehouse using OLAP

The following Table-1 shows the number of cancer patients in seven major types of cancer and the correlation according to their age.

International Journal of Research in Engineering, Science and Management
Volume-1, Issue-8, August 2018
www.ijresm.com
ISSN (Online): 2581-5782

Fig. 5. Dimension creation of specific field of cancer details from cancer data warehouse using OLAP

TABLE I
SITE WISE DISTRIBUTION OF CANCER

S. No.	Type of Cancer	Male (138)	Female (137)	Total (275)
1.	Cervix Cancer	0	35	35
2.	Breast Cancer	0	44	44
3.	Lung Cancer	14	07	21
4.	Stomach Cancer	24	15	39
5.	Oral Cancer	58	16	74
6.	Blood Cancer	23	11	34
7.	Genitourinary, Skin & Bone Cancer	19	9	28

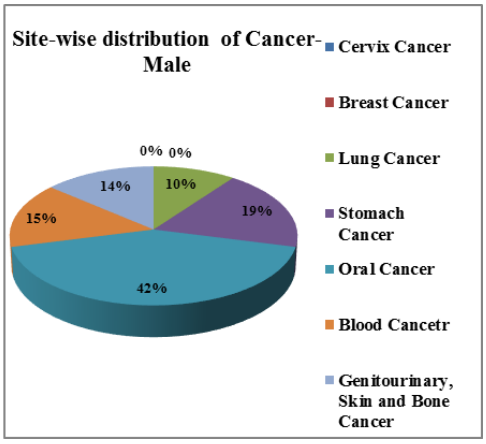

Fig. 6. Site-wise distribution of cancer- Male

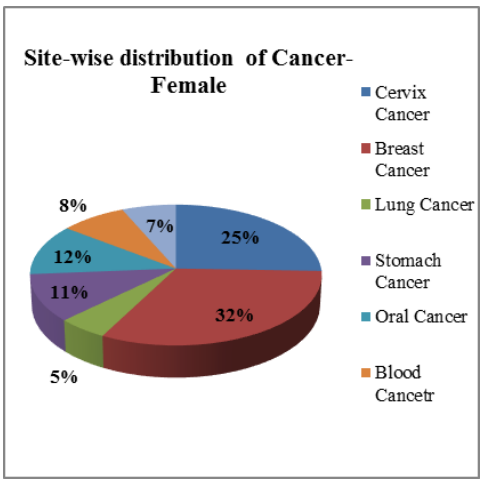

Fig. 7. Site-wise distribution of cancer - Female

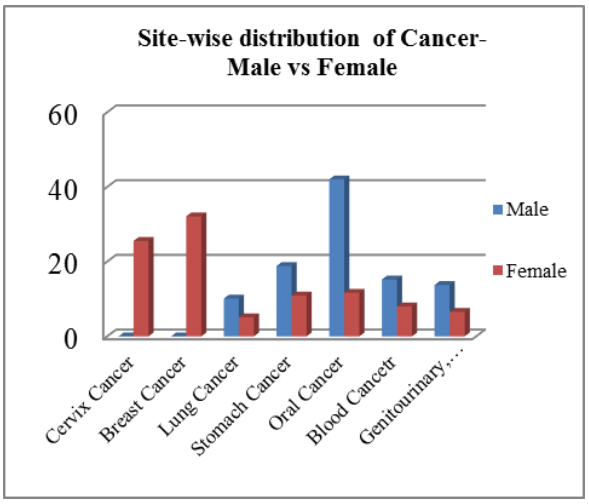

Fig. 8. Site-wise distribution of cancer - Male vs. Female

TABLE II
CORRELATION OF AGE AND SEX OF CANCER PATIENT

Age Group	Male (n=138)	Female (n=137)	Total (n=275)
<10	3	2	5
10-19	2	4	6
20-29	6	3	9
30-39	16	18	34
40-49	20	32	52
50-59	38	33	71
60-69	30	31	61
70-79	15	8	23
80-89	7	6	13
>90	1	0	1

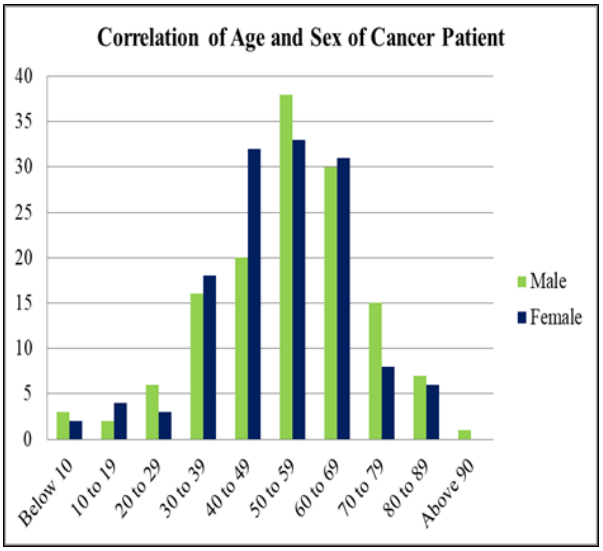

Fig. 9. Correlation of age and sex of cancer patients

District-wise data is shown in following Table-3.

TABLE III
DISTRICT WISE DATA OF CANCER PATIENTS

District	Male (n=138) No. (%)	Female (n=137) No. (%)	Total (n=275) No. (%)
North and Middle	32 (23.19)	34 (24.82)	66 (24)
South	97 (70.29)	98 (71.53)	195 (70.90)
Nicobar	9 (6.52)	5 (3.65)	14 (5.09)

International Journal of Research in Engineering, Science and Management
Volume-1, Issue-8, August 2018
www.ijresm.com
ISSN (Online): 2581-5782

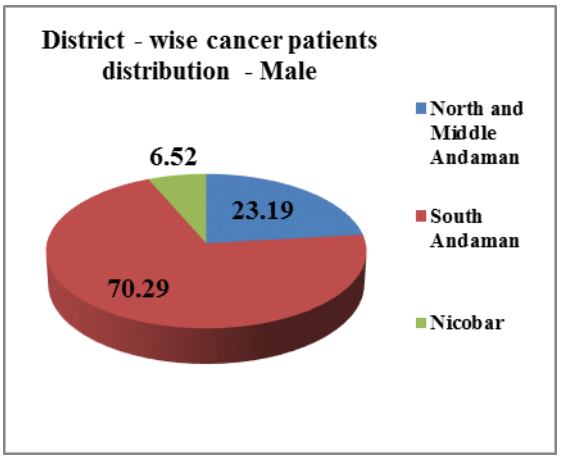

Fig. 10. District wise data of cancer patients- Male

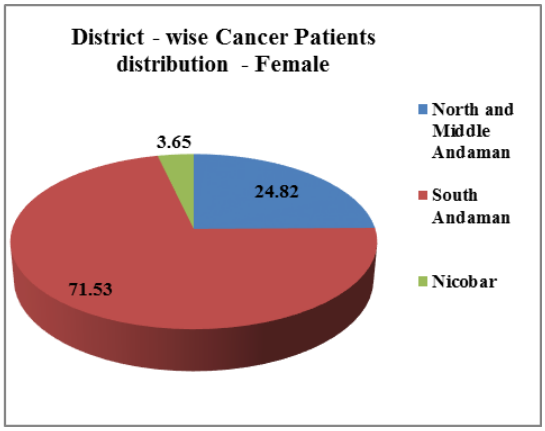

Fig. 11. District wise data of cancer patients- Female

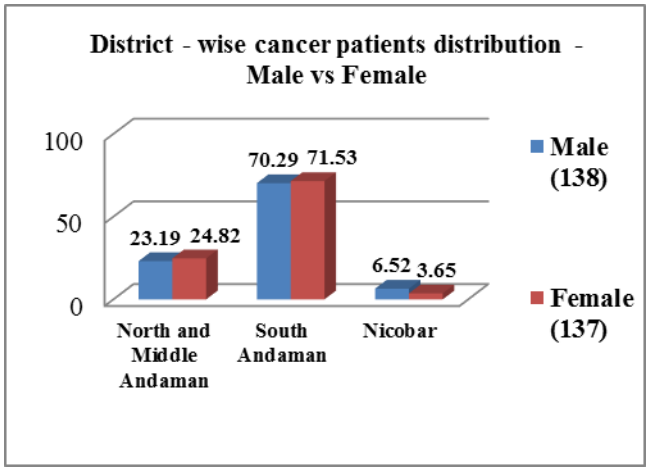

Fig. 12. District wise data of cancer patients - Male vs. Female

V. Developing a Data Mining Model to Diagnose Cancer Disease

Cancer is the most important cause of death globally. The disease diagnosis is a major process to treat the patients who are affected by cancer disease. The diagnosis process is more difficult comparatively known about the cancer disease detection. Developing a proposed data mining model is useful to diagnose the cancer disease once the cancer detection is accomplished.

In this study, a proposed data mining model has been separated into two different techniques, but it performs consecutively. The techniques are classification and clustering method of conceptual modeling. Thus the cancer data has to be converted into a knowledge base which is called as training data.

A. K-means Clustering for Classified Significant Pattern

The instances are now clustered into a number of classes where each class is identified by a unique feature based on the significant patterns mined by the decision tree algorithm.

The aim of clustering is that the data object is assigned to unknown classes that has a unique feature and hence maximize the intra class similarity and minimize the interclass similarity. The weightage scores of the significant patterns mined are fed into k-means clustering algorithm to cluster and divide it into cancer and non-cancer groups. The cancer group is further subdivided into six groups with each cluster representing a type of cancer.

The data in the cluster is again fed into k-means clustering algorithm to further subdivide it. The resulting six clusters are separated based on particular symptoms associated with any one type of cancer i.e. lung, cervix, breast, stomach, oral and blood. Finally all the data is partitioned into two types of clusters and six sub-clusters of the cancer cluster. The k-means clustering algorithm is used for partitioning the data into cancer and non-cancer clusters, where the initial cluster center is represented by the mean value of the weightage of significant patterns.

Algorithm 1: k-means

$$W(S, C) = \sum_{k=1}^{K_i} \epsilon S_k \, d(i, k) \ldots.$$

Input: k: The number of clusters. D: training dataset containing n objects.
Output: A set of group of clusters.
Step-1: Chooses two mean values from weightage of significant patterns as initial cluster centers.
Step-2: Assign each object to the cluster to which it is most similar based on the mean value of the weightage.
Step-3: Update the cluster means by calculating the mean value of all the objects in the cluster.
Given k: Number of clusters of a dataset containing a set of N entities. I, and M measurements; Here S = Seven cancer types; C – Class;

B. Weighted Average Method k-means Clustering Based Cancer Detection

Weighted Average Method (WAM) is used to improve the accuracy of analytic predictive performance models for cancer prevention systems with more number of new patients. WAM considers the patient population distribution at a system to reflect the impact of behavior/genetic factors (family history).

International Journal of Research in Engineering, Science and Management
Volume-1, Issue-8, August 2018
www.ijresm.com
ISSN (Online): 2581-5782

The WAM k-means algorithm follows an iterative optimization similar to k-means, and by consequence it is affected by some of its strengths, such as its convergence in a finite number of iterations to improve the centroid of clusters.

Weighted Average cluster k-means could cluster the patient data of medical records to different groups, and divided into six groups mapped as cancer types based on Db1, Db2,Db3…The subset of cancer types of instances with clusters will be processed towards weighted of k-means in specific features, parameter range. The sum of all the predicted values can be averaged by set of instances.

VI. Summary of Findings

Through this research a novel multilayered method combines Data warehouse and Data Mining techniques to build the cancer risk detection and prevention system is developed. The most effective way to reduce the cancer deaths is to detect and prevent cancer disease. The developing of detection and prevention system may provide an easy and a cost-effective way for screening cancer and may play a pivotal role in disease diagnosis process for different types of cancer and provide useful, preventive strategy.

Fig. 13. User interface for patients registering their data

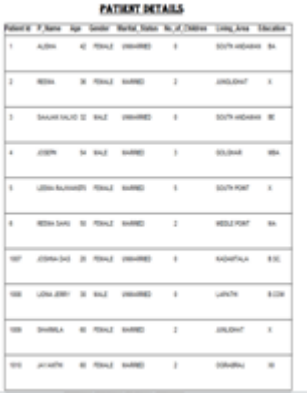

Fig. 14. Report generated for cancer patients

As a result, it is concluded that the highest no. of cancer patients i.e., 71 belong to the age group of 50 to 60 in which 38 male and 33 female patients lie in that group. Whereas the lowest no. of people belongs to the age group of above 90 in which only one male patient is there. The correlation coefficient (r) between the attributes of Age and Sex of patients is 0.96 which clearly depicts that there is a strong relationship between these two variables or attributes.

The highest mean value (0.84) occupies male patients who lie in blood cancer type whereas female patients hold 0.04 mean value. On the other hand, the lowest mean value (0.025) occupies female patients who lie in lung cancer type whereas male patients hold 0.051 mean value.

The South Andaman District has the highest mean value i.e., 0.71 of female patient and 0.70 of male patient and the total is 0.71. Whereas the male and 0.036 of female patients the total is 0.051. The mean value of male and female of North & Middle Andaman District is 0.23 and 0.25 respectively the total is 0.24.

VII. Conclusion

In the past, our dependency on macro-scale information like cancer types, patient, population, environment factor, behavioral factor and disease factor generally kept the number of variables is enough so that the standard statistical methods or even a physician's own intuition could be used to predict cancer risk. Four basic components of cancer control–prevention, early detection, diagnosis, treatment and painkilling care–thus avoid and cure many cancers, as well as palliative the suffered patients. Cancer control aims to reduce the incidence or instance, morbidity and mortality of cancer and to improve the quality of life of cancer patients in a defined population, through the systematic implementation of evidence. An implementation of our new system to expose the cancer risk factors and to ensure that people are provided with the information and support they need to adopt in a healthy lifestyles.

Cancer detection and prevention is still a challenging for the upgraded and modern medical technology. After researching a lot of statistical analyses which is based on those people who are affected in various cancer types are based on some general risk factors and symptoms have been discovered. More significantly, a globalization of unhealthy lifestyles, which particularly smoking and alcohol the adoption of many features of the bad diet habit will increase cancer incidence.

References

[1] Pujari K. Arun (2013), "Data Mining Techniques", University Press (India) Private Ltd, Hyderabad: Pp 7-23, 34-52, Pg. 69, Pp. 98 – 100, 106-110.

[2] Chattamvelli Rajan (2016), "Data Mining Methods",Narosa Publishing House Pvt. Ltd., New Delhi: Pp. 1.14 – 1.16, 4.1 – 4.24, 5.1 – 5.23, 5.27 – 5.35, Pg. 9.1, Pp. 9.8 – 9-11.

[3] Chattamvelli Rajan (2011), "Data Mining Algorithm",Narosa Publishing House Pvt. Ltd., New Delhi: Pp. 1 – 24, Pp. 25-26.

[4] Panda M, et.al. (2016), "Modern Approaches of Data Mining – Theory and Practices", Narosa Publishing House Pvt. Ltd., New Delhi: Pp. 1.1 – 1.7, Pg. 4.1, 4.6, 4.8.

[5] Jiawei Han, Micheline Kamber, Jian Pei (2012), "Data Mining Concepts and Techniques",Morgan Kaufmann Publishers is an imprint of Elsevier, Wyman Street, Waltham, USA: Pp. 5-10.

[6] P.P. Abdul Shahid, PoojaGogia B, Nagma Rafi (2017) "Comprehensive Analysis of Cancer burden in Andaman and Nicobar Islands: a Descriptive Study" International Journal of Contemporary Medical Research, ISSN (Online): 2393-915X, Volume 4, no. 2, Pp.357-359.

[7] Abubakar Ado and Ahmed Aliyu (2014) "Building a Diabetes Data Warehouse to Support Decision Making in Healthcare Industry" Journal of Computer Engineering, Volume 16, no. 2, Ver. IX, pp. 138-143.
[8] Ada and RajneetKaur (2013) "Using Some Data Mining Techniques toPredict the Survival Year of Lung Cancer Patient" International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 2, no. 4, pg.1 – 6.
[9] Alaa Khalaf Hamoud and Talib A. S. Obaid (2013) "Building Data Warehouse for Diseases Registry: First step for Clinical Data Warehouse" International Journal of Scientific & Engineering Research, Volume 4, no. 11.
[10] Alaa M. Elsayad (2010) "Diagnosis of Breast Tumour using Boosted Decision Trees" ICGST-AIML Journal, Volume 10, no. 1, pp. 01-11