INNOVATIONS ON BREAST CANCER CLASSIFICATION USING VISUAL DATA MINING TECHNIQUES

[1] S.Revathi ,[2] G.Jyothi
[1]Assistant Professor, [2]Assistant Professor
[1]Department of CSE, TKR College of Engineering&Technology, Hyderabad, India
[2]Department of CSE, BhojiReddy Engineering College for Women, Hyderabad, India

Abstract: The study of cancer has a long history 1000's of publications on cancer research spread over 100 years. Breast cancer is one of the diseases which affect mostly women rather than men in worldwide. For breast cancer treatment, it depends mainly on the tumor size, nodal status etc. Treatment of breast cancer is multidisciplinary, for early stage disease, firstly, simple medical dosage system will be suggested, and then radiation therapy and/or chemotherapy will be suggested depending on the tumor size, nodal status, age of patient and histological sub-type. Finally, surgery plays a crucial role in removing breast cancer in advanced stages. The study proposes breast cancer classification to assist pathologists by providing accurate statistical inference using visualization technique. The stage 0 has 65 missing (non-reactive) cancer patients, stage 1 has 894 patients, stage 2 has 236 patients and stage 3 has 12 cancer patients respectively.

Keywords: Visualization techniques, Pathological tumor size, stages and grades.

I. INTRODUCTION

Cancer is a malignant disease that has caused millions of human deaths by spreading its wings. As per ICMR (Indian Council for Medical Research) data incidence of breast cancer has nearly doubled in the last 25 years. In India, around 555 000 people died of cancer in 2010, according to estimates published in The Lancet on March 28, 2012. The study, led by Dr. Prabhat Jha, the Director of the Centre for Global Health Research at St.Michael's Hospital, Toronto, in a collaboration with Indian national institutions and the International Agency for Research on Cancer (IARC), used a unique method of projecting cancer deaths for the whole of India based on the patterns of cancer mortality in 2000-2003 in a sample [17].

It is evident from various statistics that the incidence of breast cancer is rapidly rising, and accounting to a significant percentage of all types of cancers in women. Breast cancer is the most common cancer which affects in urban areas in India and totals about 25% to 33% of all types of cancers in women. If these percentages are converted into actual numbers, the numbers are very high, also, over 50% breast cancer patients in India are having cancer stage 2 and even stage 3, due to tobacco mainly, and again it will be definitely a high risk factor leading to mortality [18]. Breast cancer cases were found to be spiraling world over, and urban India was no exception. Ten leading types of cancer that women in urban cities suffered from between 2006 and 2008 were looked at and it was found that breast cancer accounted for a high percentage in each city. In Mumbai, 30% of cancer cases among women were that of the breast, in Delhi and Bangalore it was almost 26.9% while the incidence in Chennai was marginally lower at 26.5%. In Kolkata, it accounted for 27.2% of cancer cases among women and in Pune it was 28.9% as per medical statistics [19].

The total cancer burden in Chennai is predicted to increase by 32% in 2012-2016, translating to 55,000 new cancer cases per year in Tamil Nadu. Breast cancer would also dislodge cervical cancer as the top-ranking cancer in the State. The article was authored by V. Shanta, chairperson, Cancer Institute (WIA), R. Swaminathan and S. Balasubramaniam also of the Institute, in association with J. Ferlay, F. Bray, and R. Sankaranarayanan of the International Agency for Research in Cancer, France, the WHO's nodal research body on cancer [20].

In India it is more common at middle age. But, in the Western countries, it is more common in older age, that is, after 60 years, due awareness, timely medical checkup followed by good medication. Variations in breast cancer incidence rates among different racial populations suggest that, etiologic factors differ in their biological genes. The important factors of breast carcinoma are: a) role of genetics and environment, b) reproductive experience, c) effect of endogenous and exogenous hormones in women, e) change in immune status and f) biologic determinants of breast carcinoma. The objective of this paper is to provide breast cancer classification based on breast cancer size.

The construction of a semantically annotated corpus of clinical text content for extracting clinically significant information from patient records are highly useful [1]. Mining microarray data to extract set of relevant molecular gene sequences of breast cancer is based on sequential patterns applied to classify breast cancer tumors according to their histological grade [2]. The fuzzy computations, the multi agent system computes the fuzzy probabilities of breast cancer development based on quantified by two linguistic variables of high and low[3]. The heterogeneous types of data such as molecular levels of number variants at the genome level, DNA methylation at the epigenome level, and gene expression and microRNA at the transcriptome level, have been classified to predict clinical outcomes in brain cancer (glioblastoma multiforme, GBM) and ovarian cancer (serous cysta deno carcinoma, OV) [5]. Comparison with manual registration reveals satisfactory completeness and efficiency of a computerized cancer registration system [6].

Knowledge acquisition from clinical documents is critical for many automated biomedical applications, including pharma covigilance and decision support system. However, acquisition of clinically meaningful relations remains challenging task, because, textual information is noisy [7]. The aphasia syndromes is a medical diagnostic using fuzzy rule-based structure which deals the symptoms of the disease [8]. Characterizing the distinct breast cancer bio molecular systems which are derived from gene expression as a protein–protein interaction network remains a key clinical challenge in breast cancer [9].

By applying the multiple normal tissues corrected differential analysis, it is identified that some genes as novel biomarker candidates [10]. An evaluative study, comparing the traditional ultrasound guided procedure with the new developed intra-operative visualization system (IVS), has been conducted with radiologists [11]. Data mining allow users to discover novelty in sequential pattern mining for biological documents using three visualization techniques: clouds, solar systems, and treemaps [12].

Clinical tests and epidemiological studies often produce large amounts of data, being multivariate in nature. The data from different sources pertaining to cancer diagnosis and incidence: (1) cytological diagnosis of breast cancer, (2) classification of breast tissues through parameters obtained from impedance spectra and (3) distribution of new cancer cases. Hierarchical cluster analysis (HCA) is needed especially in cases where there is no a priori identification of classes, suggesting a structure of the data based on clusters [13]. In the study on tissue characterization by electrical measurements, the distribution of the different types of tissues can be easily constructed. Finally, the distribution of new cancer cases possesses clear, easily unravelled, geographical patterns. To make faster and efficient the identification of mRNA targets common to more than one miRNA, and to identify new miRNAs modulated in specific pathways, a computer program identified as SID1.0 (simple String IDentifier) was developed and successfully applied in the identification of deregulated miRNAs in prostate cancer cells. Prediction data were preliminary confirmed by expression analysis of the identified miRNAs in androgen-dependent (LNCaP) and independent (PC3) prostate carcinoma cell lines and in normal prostatic epithelial cells (PrEC) [14].

A probabilistic classifiers namely, random forests and multinomial logic models are applied for defining cancer cases [15]. In a single graph, Free Viz visualization can provide a global view of the classification problem which can reveal intra-class similarities [16]. Though many papers deal with various techniques of breast cancer classification, none of the papers helps the physicians with simple visualization statistical analysis projected with graphs, which will enable to understand even by a common illiterate patient.

The rest of the paper is organized as follows: Section 2, discusses various Types of cancer, Section 3, deals with Cancer Care, Section 4, discusses Cancer Analysis, Section 5, gives results and Section 6 concludes conclusion.

II. TYPES OF CANCER

There are more than 41 types of cancers which can affect human organs are: Breast Cancer, Cervical/ Cervix, Ovarian, Blood CLL, ALL and CML, Leukemia, Lymphoma, Multiple Myloma, Bone, Oral, Mouth, Tounge, Thyroid, Throat, Larynx, Brain / Brain Tumor, Colorectal, Esophagus, Skin, Stomach, Prostate, Renal Cell Carcinoma / Kidney, Sarcoma, Submucous fibrosis, Urinary Bladder, Prostate, Hepatocellular Carcinoma, Liver, Soft Tissue Sarcoma, Blastoma, Anal, Capillary Heamenjioma, Gall Bladder, Hodgkin's Lymphoma, Non-Hodgkin's Lymphom etc [4].

2.1 Symptoms of Breast Cancer

The symptoms of breast can include: 1) a lump in the breast, 2) a change in the size or shape of the breast, 3) dimpling of the skin or thickening in the breast tissue, 4) a nipple that's turned in (inverted), 5) a rash (like eczema) on the nipple, 6) discharge from the nipple and 7) swelling or a lump in the armpit.

2.2 Cancer Myths

The focus of World Cancer Day was to dispel such damaging myths about cancer, with the theme - "Cancer – Did you know?" aimed to demystify common myths associated with cancer with an objective catalyze global prevention and treatment efforts [3]. Some of the myths which are listed below are creating confidence in patients about the disease are: Myth 1 - Cancer is not just a health issue. It has wide-reaching social, economic, development, and human rights implications. Myth 2 - Cancer is a global epidemic. It affects all ages, all income groups, and all races bearing a disproportionate burden. Myth 3 - Many cancers that were once considered a death

sentence can now be cured and can be treated effectively. Myth 4 - With the right strategies, most of the common type of cancers can be prevented.

2.3 Causes of Breast Cancer

The causes of breast are: a) family heredity, b) Consumption of alcohol, c) Consumption of Tobacco products and d) inadequate diet in phytoestrogen are directly related to breast cancer. In western countries, the higher economic people develop the disease at earlier age, whereas, the low income group people develop breast cancer at older age. The tobacco-related cancers represented around 42 % of male and 18% of female cancer deaths. In men, two of the most common fatal cancers were oral (including lip and pharynx) and lung. "A priority for cancer prevention is tobacco control, particularly through higher taxation of tobacco products to increase the low levels of cessation", said Dr Freddie Bray, the collaborating IARC scientist. Cervical, stomach and breast cancers accounted for 41% of cancer deaths in women in rural and urban areas [17].

2.4 Collection of Blood Samples

In developed countries, hospitals maintain patients' data in structured format in UCI repository. With the help of expert doctor's opinion, one year data from 1st Jan 2012 to 31st Dec 2012 of a cancer hospital in Chennai were collected to evaluate the clinic pathologic profile of breast cancer patients. Blood samples were collected from cancer patients of Adyar Cancer Hospital, Chennai, Tamil Nadu, India. As much as 1207 patients' blood samples of 3ml were collected, and stored at 80°C until further processing.

III. CANCER CURE

3.1 Cancer Nature Cure

Cow urine therapy has proved effective in reducing the suffering of patients affected by various types of cancer, cysts, tumours and neoplasm. Cow Urine is scientifically proven to enhance the anti-microbial effects of antibiotic and antifungal agents. The use of cow urine as bioactive molecules, including anti-infective agents has direct implication in reducing the dosage of antibiotics, drugs and anti-infective agent while increasing the efficiency of absorption of bio-active molecules. The cost of treatment and side-effects reduces drastically. (US Patent #6410059).

3.2 Alopathic Cancer Cure

The breast cancer size is one of the parameter to determine the effect of cancer. The cancer size is measured in centimeters (cm). The important cancer's characteristics are as follows. The growth of cancer may be directly proportional to its size.

IV. CANCER ANALYSIS

The Cancer stage depends on invasive or non-invasive cancer, whether lymph nodes are involved, and whether the cancer has spread to other places beyond the breast area. The purpose of the staging system is to organize different factors of cancer into categories in order to:

- understand prognosis based on the outcome of the disease
- guide treatment decisions based on pathology report
- Treatment can be reviewed periodically to find the effect of medicine on breast cancer

The breast cancer size can be classified into stage 0, 1, 2 and 3 based on important parameters such as size of the cancer (lump) and whether it's spread to the lymph nodes or another part of the body. The cancer is more serious, when the stage is higher. In DCIS (Ductal carcinoma in situ), which is the easiest form of diagnosis, cancer cells are in the ducts of the breast, but they haven't started to spread into the surrounding breast tissue. The DCIS shows up on a mammogram and is usually diagnosed when women go for breast screening. Normally, Breast cancer can be divided into four stages as follows:

In Stage 0, is missing (or non-reactive) used to describe non-invasive breast cancers, such as ductal carcinoma in situ (DCIS). In Stage 1, describes invasive breast cancer which is invading surrounding breast tissue may measure up to 2 centimeters and no lymph nodes are involved. Microscopic invasion is also possible; the cancer cells have only just begun to invade the tissue outside the lining of the duct or lobule. In stage 2, no tumor can be found in the breast, the tumor size ranges between 2 cm to 2.5 centimeters and may spread or might not spread to the axillary nodes. In stage 3, the cancer is above 5 centimeters found in axillary lymph nodes and may be found in skin which is clumped together or spreading to other structures of near breast bone. Also, the cancer has spread to lymph nodes above or below the collarbone. In this stage, invasive breast cancer has spread beyond the breast and nearby lymph nodes to other organs of the body such as the lungs, skin, bones, liver, brain and even distant lymph. Hence, this stage is called as advanced or metastatic stage of breast cancer.

The study concentrates on breast cancer affected patients considering total number of population N=1207. The Table 1, shows breast cancer dataset with 10 parameters consisting of Patients' age in years, Pathologic tumor size in cm, Pathologic tumor size stages, Positive Axillary Lymph Nodes, Histologic Grade, Estrogen Receptor Status, Progesterone Receptor Status, Current health status, Lymph Nodes, Lymph Nodes and Time in month along with history of disease and investigations results were recorded. All the patients were counseled about their conditions, data collection was done with the help of automatic scanning machines which is connected to the computer system so as to record the readings through software as shown in Table 1. The analysis was performed

using SPSS version 8.0 to compute mean and standard deviation along with minimum and maximum values are recorded as range for each parameter.

Table 1 Breast Cancer dataset with 10 parameters

S.No.	X-axis				Y-axis (Count)
	Parameters	Range	Mean	Std.Dev.	
1.	Age(years)	20.0 – 90.0	56.4	13.33	0 – 200
2.	Pathologic tumor size(cm)	0.0 - 7.0	1.73	1.0	0 – 300
3.	Pathologic tumor size stages	0.0 – 7.0	9.0	2.54	0 – 1000
4.	Positive Axillary Lymph Nodes	0.0 – 20.0	2.27	0.61	0 – 1200
5.	Histologic Grade	1.0 – 3.0	0.61	0.49	0 – 600
6.	Estrogen Receptor Status	0.0 – 1.0	0.54	0.50	0 – 600
7.	Progesterone Receptor Status	0.0 – 1.0	0.06	0.24	0 – 500
8.	Current health Status	0.0 – 1.0	1.27	0.47	0 – 1200
9.	Lymph Nodes	0.0 – 1.0	0.23	0.42	0 – 1000
10.	Time(months)	0.0 – 135.0	47.0	29.64	0 - 100

V. RESULTS

In Figure 1, x-axis is plotted using age in years. From Figure 1 to Figure 10, y-axis is plotted with number of patients. The study reveals that, out of total number of 1207 patients, there were 1150 female patients and rest was male patients. The patients in the age group 20 years to 90 years in steps of 5 years is 3, 3, 25, 50, 121, 141, 167, 122, 140, 167, 131, 82, 30, 19 and 5. Among which the most affected age group is 40 years to 75 years female patients are in high risk as per statistics.

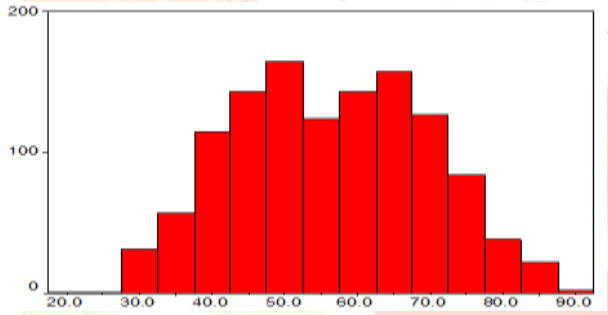

Figure 1 Age distribution

In Figure 2, x-axis is drawn taking Pathological tumor size in centimeters. The study reveals that, the pathological tumor size from 0.0cms to 7.0cms in steps of 0.5 is 5, 132, 279, 247, 228, 99, 87, 57, 45, 7, 9, 8, 2 and 2 respectively. In Figure 3, x-axis is drawn taking Pathological tumor stages in centimeters. The study reveals that, there are 4 stages namely stage 0, 1, 2, and 3. The stage 0 has 65 missing (non-reactive) cancer patients, stage 1 has 894 patients, stage 2 has 236 patients and stage 3 has 12 cancer patients respectively.

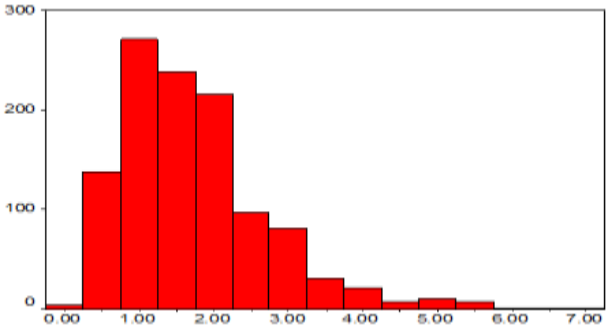

Figure 2 Pathological tumor size in cms

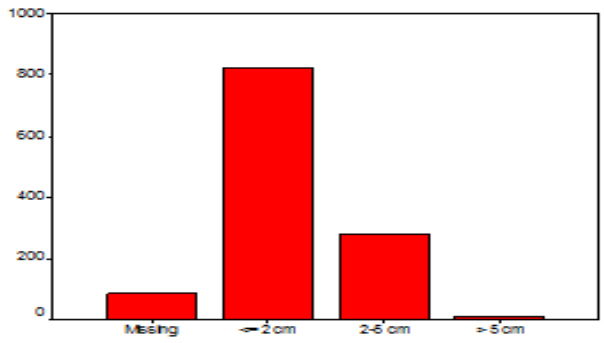

Figure 3 Pathological tumor stages in cms

In Figure 4, x-axis is drawn taking Positive Axillary Lymph Nodes. The study reveals that, the Positive Axillary Lymph Nodes from 0.0 to 35.0 in steps of 5.0 is 1081, 101, 16, 9, 0, 0, 0 and 0 respectively.

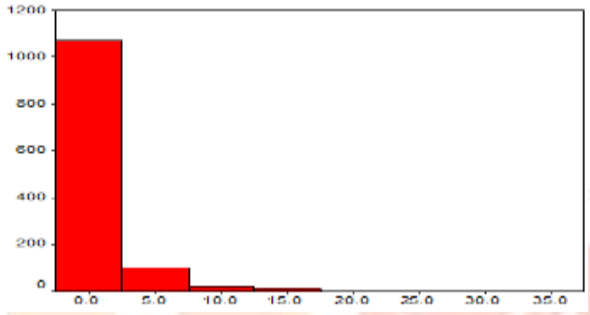

Figure 4 Positive Axillary Lymph Nodes

In Figure 5, x-axis is drawn taking Histologic Grades. The Histologic Grades are classified into grade 0 value as 187, grade 1 value as 95, and grade 2 values as 560 and grade 3 value as 365. In Figure 6, x-axis is drawn taking Estrogen Receptor Status. The Estrogen Receptor status are classified into missing receptor value as 317, Estrogen Receptor status = 0 value as 326 and Estrogen Receptor status =1 value as 564.

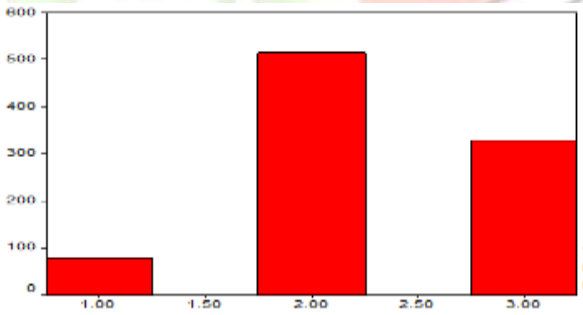

Figure 5 Histologic Grades

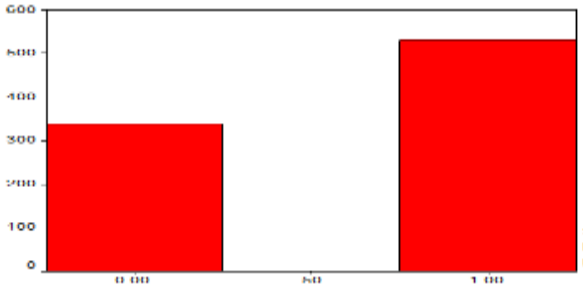

Figure 6 Estrogen Receptor Status

In Figure 7, x-axis is drawn taking Progesterone Receptor Status. The Progesterone Receptor Status are classified into missing receptor value as 325, Progesterone Receptor Status = 0 value as 395 and Progesterone Receptor Status =1 value as 487.

In Figure 8, x-axis is drawn taking Current health Status.
The Current health Status are classified into 0 and 1, the Current health Status = 0 value as 1162 and Current health Status =1 value as 45.

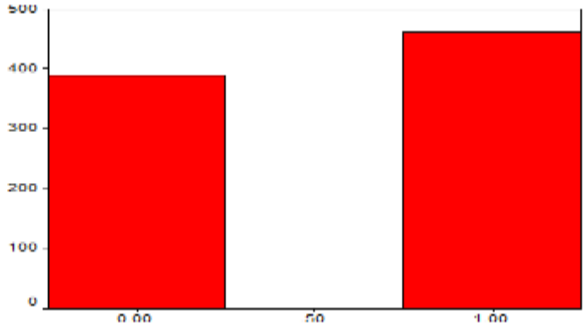

Figure 7 Progesterone Receptor Status

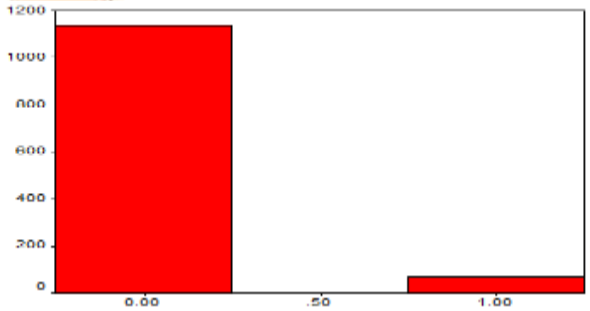

Figure 8 Current health Status

In Figure 9, x-axis is drawn taking Lymph Nodes. The Lymph Nodes are classified into 0 and 1, the Lymph Nodes = 0 value as 1060 and Lymph Nodes =1 value as 247.

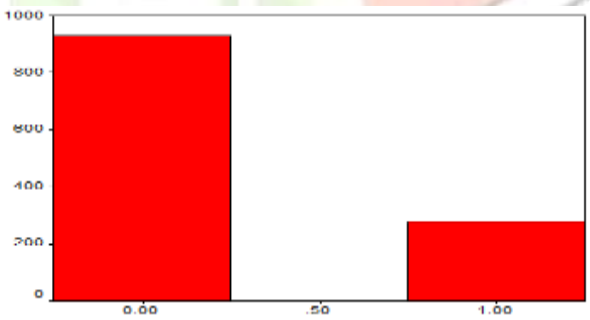

Figure 9 Lymph Nodes

In Figure 10, x-axis is drawn taking Time (months). The study reveals that, the Time (months) from 0.0 to 135.0 in steps of 10.0 is decreasing uniformly of the number of patients living is: 51, 82, 81, 63, 80, 58, 63, 78, 62, 64, 77, 60, 48, 44, 36, 37, 36, 34, 30, 20, 17, 18, 16, 12, 15, 10, 12 and 3. As months increases the life time of the cancer patients decreases.

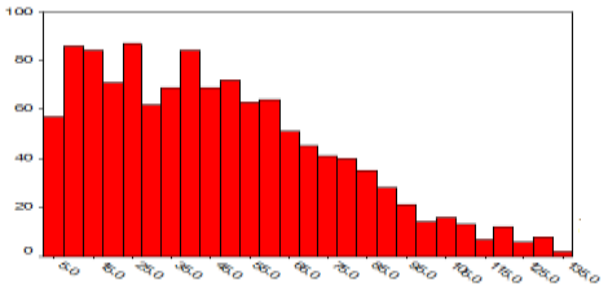

Figure 10 Time (months)

VI. CONCLUSION

Carcinoma breast cancer is still a common problem among middle age group with invasive ductal carcinoma being the commonest variant with a high grade and a late stage of presentation due to lack of screening and awareness programs. By this research, the main aim is to restrict the growth of cancer, advice regular and optimal dosage of treatment to the patients based on stages of cancer. Also, adequate dieting can be suggested to survive long life and to avoid unbearable pain by which help the patient to have better quality of life. The future scope of this paper is to compute correlation between each parameter and to suggest best dosage system along with appropriate diet. ICMR has also come out with the possibility of one in number of people developing cancer. The calculation is age range 40-70 years in Chennai women runs the risk of developing cancer. Among 28,740 cancer patients, 1207 were registered as breast cancer; the female to male ratio was 100:2. Breast cancer accounted for 23% of all female cancer patients.

VII. REFERENCES

[1] Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, Andrea Setzer, "Building a semantically annotated corpus of clinical texts", Journal of Biomedical Informatics, Vol. 42, pp. 950–966, 2009.

[2] Mickael Fabregue, Sandra Bringay, Pascal Poncelet, Maguelonne Teisseire, Béatrice Orsetti, "Mining microarray data to predict the histological grade of a breast cancer ", Journal of Biomedical Informatics Vol. 44, pp. S12–S16, 2011.

[3] Farzaneh Tatari, Mohammad-R. Akbarzadeh-T, Ahmad Sabahi, "Fuzzy-probabilistic multi agent system for breast cancer risk assessment and insurance premium assignment", Journal of Biomedical Informatics, Vol. 45, pp. 1021–1034, 2012.

[4] Cancer care foundation of India, No.165, RNT Marg, Indore, M.P, India. Email:mail@ccfi.in.

[5] Dokyoon Kim, Hyunjung Shin, Young Soo Song, Ju Han Kim," Synergistic effect of different levels of genomic data for cancer clinical outcome prediction", Journal of Biomedical Informatics, Vol. 45, pp. 1191–1198, 2012.

[6] Paolo Contiero, Andrea Tittarelli, Anna Maghini, Sabrina Fabiano, Emanuela Frassoldi, Enrica Costa, Daniela Gada, Tiziana Codazzi, Paolo Crosignani, Roberto Tessandori, Giovanna Tagliabue, "Cancer Registration System", Journal of Biomedical Informatics, Vol. 41, pp. 24–32, 2008.

[7] Xiaoyan Wanga, Herbert Chase, Marianthi Markatou, George Hripcsak, Carol Friedman, "Selecting information in electronic health records for knowledge acquisition", Journal of Biomedical Informatics, Vol. 43, pp. 595–601, 2010.

[8] Mohammad-R. Akbarzadeh-T, Majid Moshtagh-Khorasani," A hierarchical fuzzy rule-based approach to aphasia diagnosis", Journal of Biomedical Informatics, Vol. 40,pp. 465–475, 2007.

[9] James Chen, Lee Sam, Yong Huang, Younghee Lee, Jianrong Li, Yang Liu, H. Rosie Xing, Yves A. Lussier," Protein interaction network underpins concordant prognosis among heterogeneous breast cancer signatures", Journal of Biomedical Informatics, Vol. 43, pp. 385–396, 2010.

[10] Hoon Jin, Han-Chul Lee, Sung Sup Park, Yong-Su Jeong, Seon-Young Kim," Serum cancer biomarker discovery through analysis of gene expression data sets across multiple tumor and normal tissues", Journal of Biomedical Informatics, Vol. 44, pp. 1076–1085, 2011.

[11] Ashis Jalote-Parmar, Petra Badke-Schaub, Wajid Ali, Eigil Samset," Cognitive processes as integrative component for developing expert decision-making systems: A workflow centered framework", Journal of Biomedical Informatics, Vol. 43, pp. 60–74, 2010.

[12] Sallaberry, N. Pecheur, S. Bringay, M. Roche, M. Teisseire," Sequential patterns mining and gene sequence visualization to discover novelty from microarray data", Journal of Biomedical Informatics, Vol. 44, pp. 760–774, 2011.

[13] Tania F.G.G. Cova, Jorge L.G.F.S.C. Pereira, Alberto A.C.C. Pais .," Is standard multivariate analysis sufficient in clinical and epidemiological studies?", Journal of Biomedical Informatics, Vol. 46, pp. 75–86, 2013.

[14] Maria C. Albertini, Fabiola Olivieri, Raffaella Lazzarini, Francesca Pilolli, Francesco Galli,Giorgio Spada, Augusto Accorsi, Maria R. Rippo, Antonio D. Procopio," Predicting microRNA modulation in human prostate cancer using a simple String IDentifier (SID1.0)", Journal of Biomedical Informatics, Vol. 44, pp. 615–620, 2011.

[15] Sandro Tognazzo, Bovo Emanuela, Fiore Anna Rita, Guzzinati Stefano, Monetti Daniele, Stocco Cramen Fiorella, Zambon Paola,"Probabilistic classifiers and automated cancer registration: An exploratory application", Journal of Biomedical Informatics, Vol. 42, pp. 1–10, 2009.

[16] Janez Dems, Gregor Leban, Blazˇ Zupan, ”FreeViz—An intelligent multivariate visualization approach to explorative analysis of biomedical data”, Journal of Biomedical Informatics, Vol.40, pp. 661–671, 2007.
[17] International agency for Research on cancer, WHO, Press Release, N° 210, March 28th , 2012.
[18] Sumeet Shah, “ Breast Cancer in India”, dr.sumeetshah@gmail.com,http://in.linkedin.com/in/drsumeet
[19] ICMR Study revealed Doubling of Incidence of Breast Cancer in Metropolitan Cities, Current affairs Report, October 19th, 2011.
[20] V. Shanta, R. Swaminathan and S. Balasubramaniam, J. Ferlay, F. Bray, and R. Sankaranarayanan, WIA, International Agency for Research in Cancer, France, the WHO's nodal research body on cancer.