Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 05, May - 2017

Categorization of Breast Cancer Cell with the Help of Information Bottleneck Theory

Himanshu Kumar Shukla[1], Dr. Vineet Saxena[2], Shruti Singh[3]
Institute of Management Science,
Lucknow University, U.P., India.

Abstract- Cancer is the most common disease in both Men and Women. It can start almost anywhere in the Human body. Our proposed system is working on Breast Cancer cell. Breast Cancer is very dangerous .Our proposed system describe how to find the growth rate of Cancer cell, categorize the Cancerous and Non-Cancerous cells and make a decision based system, with the help of Information Bottleneck principal into Breast Cancer cell Data-sets. To find the growth rate of Cancer cells using Kinetices Cells of Cancer growth formula in the given Tumor size and categorize the Cancerous and Non- Cancerous cell using S- Phase fraction Technique thereafter, find more relevant feature of cancers using IB method. This theory defines more accurate growth of Cancer according to given tumour diameter in the large Breast data sample. The purpose of this research is to develop bizarre approach of clinical problem regarding diagnosis and provide fast recovery process at early stage of the Cancer.

Keywords - Kinetic growth rate, S-phase fraction, Information Bottleneck theory, data set and cancer cell.

I.INTRODUCTION

Cancer is a broad term for a class of diseases characterized by abnormal cells that grow and invade healthy cells in the body. Breast Cancer is very dangerous disease .Breast cancer starts in the cells of the breast as a group of cancer cells that can then invade surrounding tissues or spread (metastasize) to other areas of the body[1]. Breast cancer occurs when malignant tumors develop in the breast. Breast cancer tumors can be categorized by the size, type of cells, and the characteristics that fuel its growth.

Breast cancer is the most common cancer in U.S. women, with an estimated 60,290 cases of in situ disease, 231,840 new cases of invasive disease, and 40,290 deaths expected in 2015[2]. Men account for 1% of breast cancer cases and breast cancer deaths.

Thereare many experimentsand techniques which are used on medical datasets using multiple categorization and pattern selection techniques. This paper presents a new general idea of breast cancer cell categorization and also finds out the growth rate of cancer cell data sets using the kinetices of cancer cell growth formula and make a decision based system with the help ofInformation bottleneck theory.

II. LITERATURE REVIEW

This topic was conceived after studying the various papers related on breast cancer datasets classification and information Bottleneck theory approaches.

Authors in paper [3] presentsa comparison among the different classifiers decision tree (J48), Multi-Layer Perception (MLP), Naive Bayes (NB), Sequential Minimal Optimization (SMO), and Instance Based for K-Nearest neighbor (IBK) on three different databases of breast cancer (Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC)) by using classification accuracy and confusion matrix based on 10-fold cross validation method.The experimental results is shows that using the fusion of MLP, J48, SMO and IBK is most superior to the other classifiers in WPBC dataset.using this fusion scores accuracy of 77.3196% is prposed method.

Authors in paper [4] work, using an image analysis approach and LABVIEW software. In this technique we can also count the number of defected cells and find their position with image processing. And also easily find out the defected cells,which are extracted and highlighted among the other cells, very well. This technique can be used in large scale cells because, it can improve the accuracy and speed of the examination and counting process of the defected cells.

Authors in paper [5] have been proposed to classify oral cancers using Statistical Feature Extraction technique. This paper is proposed system segments and classifies oral cancers at an earlier stage. The images are captured and the series of operations are performed to identify the classification as normal or abnormal. And the tumor is segmented using Marker Controlled Watershed segmentation and features are extracted using GLCM is following properties Energy, Contrast, Entropy, Correlation, Homogeneity. Further SVM classifier is used to identify the classification. The accuracy obtained for the proposed system is using GLCM feature extraction is 92.5%.

Authors in paper [6] present the development of CAD systems and related techniques which plays a important-role in the early detection of breast cancer and can reduce the death rate among women with breast cancer and then focus on the key CAD techniques developed recently for breast cancer,including detection of masses, calcification, architectural distortion, bilateral asymmetry in mammograms.

Authors in paper [7] present various classification rules compared to predict the best classifier.The classifier is identified to determine the nature of the disease which is highly important for finding healthy breast cancer patients.For this research we use 37 different classification algorithm for the purpose of diagnosis of healthy and sick patients.By observing the result analysis using various classifiers give more accurate result.According to using different classifier rule these research diagnosis 76% healthy and 24% sick patients.

Authors in paper [8] present IB method for unsupervised clustering of image databases.The unsupervised clustering scheme is based on information theoretic principles. This section presents an investigative analysis of the IB method for image clustering. Experimental results demonstrate the IB method's ability to discover an optimal number of clusters in the database using the AIB algorithm.The AIB clustering method described was applied to our database of 1460 images. The clustering is performed on the GMM image representation. We started with 1460 clusters where each image model is a cluster. After 1459 steps all the images were grouped into a single cluster. The given database was thus arranged in a tree structure.

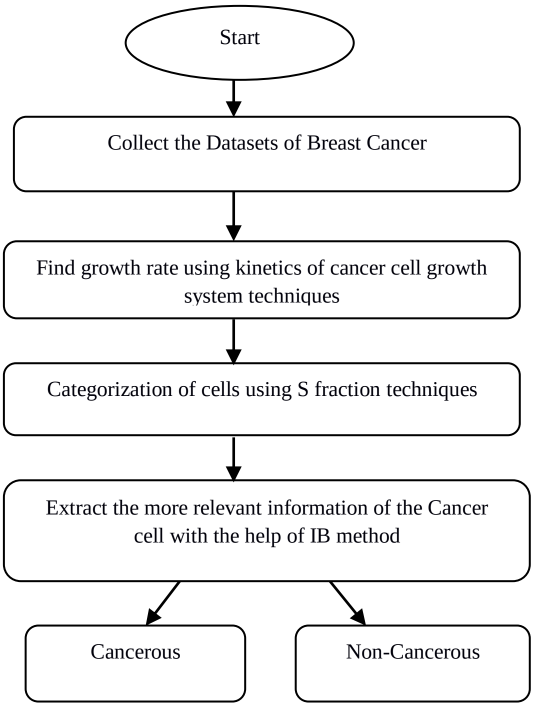
Authors in paper [9] proposed a novel and generic video/image re-ranking algorithm, IB re-ranking, which reorders results from text-only searches by discovering the salient visual patterns of relevant and irrelevant shots from the approximate relevance provided by text results. Evaluating the approach on the TRECVID 2003-2005 data sets shows significant improvement upon the text search baseline, with relative increases in average performance of up to 23% using the IB re-ranking approach and it is based on Pseudo-Labeling Strategies. We have experimented with three different strategies for such estimation of "soft" pseudo-labeling:-

- Binary approach
- Normalized Rank
- Score Stretching approach

All papers present different types of classifier and image processing technique toextract the breast cancer. The rest of paper is organized as follows: Section 2 describes the proposed frame work which describesthe formula of kinetices of cancer cell growth system techniquewhich is used to find growth rate, S fraction technique and Information Bottlneck Theroy which are used to find cancerous and non canerous cell in datasets.And section 3 shows the experimental results and Section 4 is the conclusion of the paper.

III. PROPOSEDMETHODOLOGY

The proposed Breast Cancer cell categorization method consists of a few phases as shown in figure 1.


Start
Collect the Datasets of Breast Cancer
Find growth rate using kinetics of cancer cell growth system techniques
Categorization of cells using S fraction techniques
Extract the more relevant information of the Cancer cell with the help of IB method
Cancerous
Non-Cancerous

A. DATASET EXPLANATION

The Breast cancer data-set is collected from thehttp://breast-cancer research.com/content/12/3/R25/additional,Additional file 1, to describe patient andtumor characteristics. The table is containing Excel format and it is defined clinical and experimental data on the 200 *HER2*-amplified tumors. This dataset is defined patient TAX_ID, age, tumors size, and FGA value (gain and loss) and Tumor Size Stratification.A brief description of these data-sets is presented in table 1. Each dataset define tumor size and then uses to find the growth rate of cancer in the tumor and categorization of cancer cell.

TAX_ID	AGE	TUM SIZE(mm)	FGA_GAIN	FGA_LOSS	Tumor Size Stratification
TAX577485	36	13	0.18	0.14	<=20mm
TAX577594	50	35	0.06	0.01	>=20mm
TAX577595	43	60	0.2	0.44	>20mm
TAX577002	42	5	0.1	0.03	<=20mm
TAX577601	59	20	0.13	0.16	<=20mm

TABLE 1 Description of Patient and tumor characterstices for the 200 *HER2*-amplified tumors

B. KINETICES OF CANCER CELL GROWTH SYSTEM

By the help of this technique we find "How Cancer Cells grow in a tumor?" So using this technique first of allwe are going to assume that the Cancer cells are spherical and the Tumor is also spherical and this is covered with cells. So after all radiologists measure the tumor diameter in millimeters. we can calculate the number of cancer cells, N in the tumor using this equation

$$N= (D/d)^3$$

Where D is the defined Tumor diameter givenin mm, and d is defined as Cancer cell diameter so we assume that d= 10 μm and 1mm=1000 μm. sowe have use this equation with our dataset and calculate cancer cells growth in a Tumor.

C. S-PHASE FRACTION TECHNIQUES

S phase fraction is used to define information about the rate of Cell growth means it defines cancer cell growth as low, Intermediate and high. This S-phase stands for "Synthesis phase", so just before a cell divides into two new cells and also provide a result of growth if less than 6% is considered low, 6-10% intermediate, and more than 10% is considered high[10]. The S phase fraction technique shows the result and on the basis of this result I am going to divide cancer cell.

D. *EXTRACT MORE RELEVANT INFORMATION OF CANCER WITH THE HELP OF IB METHOD*

Our proposed method is based on the Information Bottleneck Method. This method extends elements of rate distortion theory to supervise information extraction. Rate distortion theory is applied to maximize the amount of information about Y retained for a particular length description.

The information bottleneck method is a technique introduced by Naftali Tishby et al. [11] for finding the best trade-off between accuracy and complexity (compression) when summarizing (e.g. clustering) a random variable X, given a joint probability distribution between X and an observed relevant variable Y.

Examples are include the information that face images provide about the names of the people portrayed, or the information that speech sounds provide about the words spoken. Understanding the signal x requires more than just predicting y, it also requires specifying which features of X play a role in the prediction. We formalize the problem as that of finding a short code for X that preserves the maximum information about Y. That is, we squeeze the information that X provides about Y through a `bottleneck' formed by a limited set of codeword's X.

This method is used to find the hidden and relevant or meaningful information in the large database.Algorithms thatare motivated by the IB method have already been applied to text classification,gene expression, neural code, and spectral analysis. Here, weintroduce a general principled framework for multivariate extensions ofthe IB method.

Finally we are using Parallel IB method in our dataset. In [12]this example, we are going to distribute two variables A and B. Consider the apllication of the multivariate principle with G_{in} and $G_{out.}$

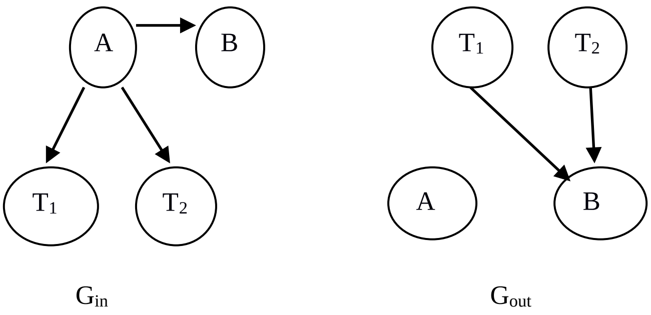


Figure 1 Parallel Information Bottleneck

In this case we intoduce two varibles T_1 and T_2 as in figure both of them are functions of A. G_{in} specifies that T_1 and T_2 compresses between A and G_{out} specifies that T_1 and T_2 should predict B. we can think of this requirement as an attempt to decompose the information A contains about B into two "Orthogonal" components. The resulting function is

$$\mathbf{I(T_1, T_2; B) - \beta^{-1} (I (T_1; A) + I (T_2; A))}(1)$$

So we are using this method to find more relevant growth rate of cancer cell. They are describing following step to calculate the more relevant Cancerous and Non- Cancerous cell into Data-set.

- Store all dataset in an array X. after storing all data-set, will sort the data in ascending order.
- Afterstoring the sorted data into two clusters T_A and $T_{B.}$
- Calculate the probability T_A and T_B with respect to G.

IV. RESULT AND ANALYSIS

All the experiments were done in MATLAB 2011.first of consider our data-sets and collect this FGA_gain and FGA_loss value in the data-sets and plots a graph using gain (green) and loss (blue) across 206 HER2-amplified tumors between frequency range and data sample respectively. Figure 2 shows the graph betweenfrequency range and different (Gain and Loss) data sample.

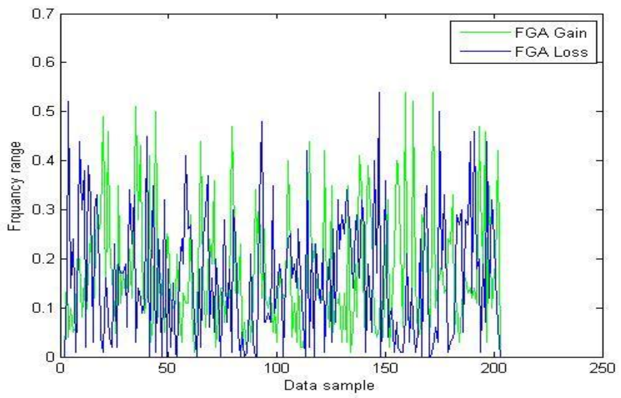


Figure 2 Graph b/w frequency and different number of data sampls.

Second find How do Cancer Cells grow in a tumor? So we are using kinetics of cancer cell growth equation in our data-sets.Data-sets define different types of tumor size according to above table. And growth rate is found between different tumor size and different time (days) interval. Table 2 defined growth rates between tumor sizes and time (days) interval.

Days	Tumor diameter,mm	Cancer cells,N= (D/d)^3
0	2	8.00e+06
20	9	7.29e+08
40	15	3.37e+09
80	25	1.56e+10
120	35	4.28e+10
160	50	1.25e+11
200	100	1.00e+12

TABLE 2Describe the grwoth rates between tumor sizes and time (days) interval.

Wwe are using 200 data sample to find the growth rate of cancer cell with the help of each tumor diameter. And after then consider each N value and plot in the graph.

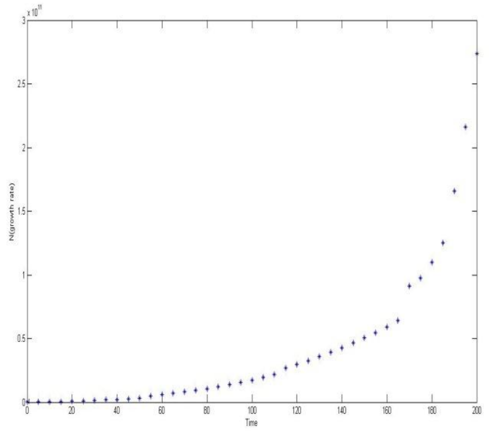


Figure 3 Graph shows the result between times and N value of Cancer cells.

This graph shows the result between times and N value of cancer cells, to observe the diameter of the tumor is directly propsnal to growth of the cancer.Afterdivide the our growth cancer cells so we using S Phase fraction processresult if Cancer growth is less than 6% is consider low, 6-10% intermediate, and more than 10% is considered high.

Low	Intermediate	High
0	7.29e+08	9.73e+10
8.00e+06	9.26e+09	4.21e+11
	5.93e+10	5.12e+11
	6.40e+10	1.72e+12

TABLE 3 Describe cancerous and non canceros cell.

According to S phase fraction divide into the200 cancerous cell between low, intermediate and high priority class.Low class is defined no cancer cell produced, intermediate class is defined cancer cell is avilable or may not and high class is defined cancer cell is growth in long time.

After categorization we again use our data-set sample we describe 200 sample of the tumor diameter and expressed of data we are taking 33 samples non-cancerous and 12 cancerous cells. Then we have express the data and measure the all tumor diameter into each sample to get a joint probability P(G,A) over tumor size and samples. We sorted all tumor size by their donation to I(G,A) and choose the 120 most informative ones it is defined highest tumor size in the data sample, which capture 70% of the original information, ending up with joint probability with |A|= 2 and |G|= 120.

Finally, to describe the performance of the parallel IB we apply it to the same data. Using the parallel IB algorithm (with β-1= 0) taking which we cluster the array A into two clusters TA and TB, that try to capture the information about G. we will sorted the data into cluster hierarchies I(TA, TB, G) = (2,45,120) and find the probability between TA and TB with respect to G. this algorithm are used to all 200 sample data sets according to table 3.

Sr. No.	T_A	T_B	$P_1(T_A/G)$	$P_2(T_B/G)$
1	2	45	2/120= 0.016	45/120=0.375
2	5	50	5/120=0.041	50/120=0.416
3	10	55	10/120=0.083	55/120=0.458
4	15	60	15/120=0.125	60/120=0.5
5	20	65	20/120=0.166	65/120=0.541
6	25	70	25/120=0.208	70/120=0.583
7	30	75	30/120=0.25	75/120=0.625
8	35	80	35/120=0.297	80/120=0.667
9	40	100	40/120=0.333	100/120=0.833

TABLE 4 Descirbe the tumor size and probability of Cancer Cell.

After finding Probability we can compare the two partitions we found the tumor size of T_B is grater than the tumor size of $T_{A.}$ The result shows as the size of tumor increasing with probability, then the occurance of Cancer cell get also increases.

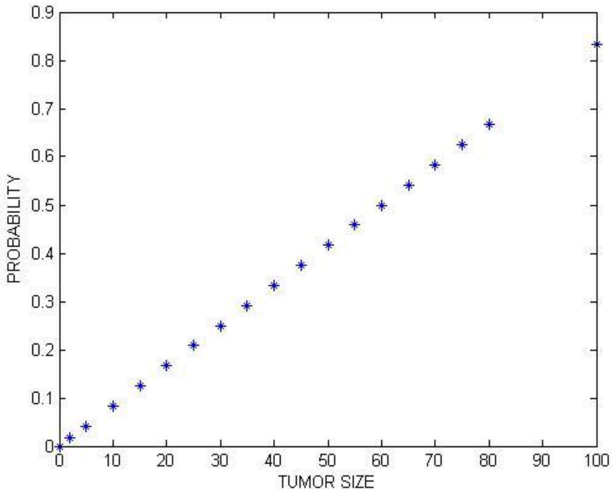


Figure 4 Graph shows the relations between Tumor size and probability of Cancer Cell.

From the above graph it is clearly visible, as the size of tumor diameter increases the probability of cancer increased linearly. This result describes the ability of the algorithm to extract in parallel different meaningful independent partitions of the data.

V.CONCLUSION AND FUTURE WORK

The selected proposed model using the kinetices cells of cancer growth techniques, S-phase fraction technique and information bottleneck theory is uesd to find the easy and better slloution of growth rate of Cancer into large breast Cancer cell data-sets and easily Categorize the Cancerous and Non-Cncerous Cells.So after using these techniques, the facts will confirm to diagnose the disease in an easy and fast manner. This new approach may lead to reduce the time of detecting the cancer and provide fast recovery process at early stage of the Cancer. In future we can implement the same techniques for other dieseses like skin Cancer and other from of Cancers. So wecan use this technique to find more accuracy between large datasets and it is helpful in medical diagnosis.

REFERNCES

[1] http://www.nationalbreastcancer.org/what-is-cancer

[2] American Cancer Society: Cancer Facts and Figure 2015.Atlanta, Ga: American Cancer Society, 2015.

[3] Gouda I. Salama, M. D. Abdelhalim, and Magdy Abd-elghany Zeild, “Breast Cancer Diagnosis on three Different Dataset using Multi-classifiers.” International Journal of Computer and Information Technology (2277 – 0764) Volume 01– Issue 01, September 2012

[4] Hossein GhayoumiZadeh, Siamak Janianpour, and Javed Haddadnia, “Recognition and Classification of the Cancer cells by Using Image processing and LABVIEW”, International Journal ofComputer Theory and Engineering, Vol 5, 1, February 2013.

[5] Anuradha. K and Dr. K. Sankaranarayanan, “Statistical feature Extraction to classify Oral Cancers”, JGRCS, Volume 4, No. 2, Feburary 2013

[6] R. Bhanumathi, G. R. Suresh, “Latest Advances in Computer Aided Detection of Breast Cancer by Mammagraphy.” International Journal in IT and Engineering (ISSN: 2321-1776) Vol.01 ISSUE-06, NOV., 2013.

[7] Miss Jahanvi Joshi, Mr. Rinal Doshi, Dr. Jigar patel, “Diagnosis and prognosis Breast Cancer using Classification Rules.” Intenational Journal of Engineering Research and General Science volume 2, Issue 6, October-November, 2014.

[8] Gordon, Hayit Greenspan, Jacob Goldberger “Unsupervised images clustering using the information bottleneck method.” Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2- Volume Set 0-7695-1950-4/03, 2003

[9] Winston H. Hsu, Lyndon S. Kennedy, Shih-Fu Chang, “Video search Re-ranking via information bottleneck principal", copyright 2006 ACM 1-59593-447-2/06/0010, October 2006.

[10] http://www.breastcancer.org/symptoms/diagnosis/rate_grade/ S-Phase fraction

[11] Naftali Tishby, Fernando C. Pereira, and William Bialek, “The Information Bottleneck method”, NEC Research Institute, 4 independence Way Princeton, New Jersey 08540,30 September 1999.

[12] N. Friedman, O. Mosenzon, N. Sionim and N. Tishby Multivariate Infonnation Bottleneck UAI,2001.