Pak. J. Biotechnol. Vol. 13(4) 231-235 (2016) ISSN Print: 1812-1837
www.pjbt.org ISSN Online: 2312-7791

BREAST CANCER DIAGONOSIS IN ANALYSIS OF BRCA GENE USING MACHINE LEARNING ALGORITHMS

J. Sumitha* and T. Devi

Department of Computer Applications, Bharathiar University, India
Email*:sumivenkat2006@gmail.com

Article received 12.9.2015; Revised 26.10.2016; Accepted 2.11.2016

ABSTRACT

Worldwide, the breast cancer is the second leading cancer type which leads to death among women. Breast cancer exists due to the mutation happens in the normal growth of Breast Cancer Gene (BRCA) under certain circumstances. In this paper, we proposed the existing machine learning algorithms for finding the disease – causing BRCA gene. These existing machine learning algorithms are compared with each other to determine the efficiency in detecting the diseases from gene expression value. The results proved that the Hybrid Radial Bias Neural Network (HRBFNN) performs better than Divide and Conquer Kernel Solving Support Vector Machines (DCKSVM) and other classification algorithms than ever before.

KEYWORDS: Confusion matrix, DCKSVM, Identification of BRCA gene, HRBFNN, Sequential model.

INTRODUCTION

Cancer is the second leading cause of death and more than 1,500 people per day are affected by this disease all over the world. Approximately, there are 1479350 new cancer cases were diagnosed all over the world and analyzed that the cancer is the second most common reason of death and it accounts for nearly one of every four deaths. Breast cancer is the most common and rigorous cancer among women and it continuing to be a significant public health problem all over the world. There are about 182,000 new cases of breast cancer are diagnosed and about 46,000 women die of breast cancer each year. Nearly about 192,370 new cases of breast cancer were diagnosed among women and thus, the incidence and mortality of breast cancer are very high [1]. Hence, breast cancer is considered as the most common cause of death among women. The normal gene of BRCA exists in the nature of proto-oncogene in the human body. When the mutation occurs under certain circumstances such as habits, radiation and by hereditary aspects, this proto-oncogene is then translated into oncogene which leads to cancer.

Since breast cancer has very high incidence of death rate, the reason for breast cancer is still mysterious and there is no efficient way to prevent the existence of breast cancer. So, early detection is the first significant step to-wards treating breast cancer. The most common screening methods for earlier detection of breast cancer are mammography and sonography. Among these two methods, mammography is the most important tool that doctors prefer to diagnose [1,2]. But, mammography has some disadvantages on diagnosing breast cancer. Since it is very sensitive, it is not accurate in detecting breast cancer because mammography can identify an abnormality that looks like cancer in the human body, but it turns out to be normal which is a false positive [3,4] and such a misdiagnosis gives hectic for patients. Hence it is essential to develop software which could give reliable diagnostic results to prevent these drawbacks.

The objective of this paper is to detect BRCA gene with the help of gene expression value using machine learning algorithms. The algorithms which used in this research are the sequential algorithm, DCKSVM and HRB FNN and the prediction is done on the basis of confusion matrix method.

This paper is organized as follows: Section 2 presents proposed methods to solve the task of detecting breast cancer. Section 3 contains results obtained and discussions and section 4 for conclusions.

METHODS

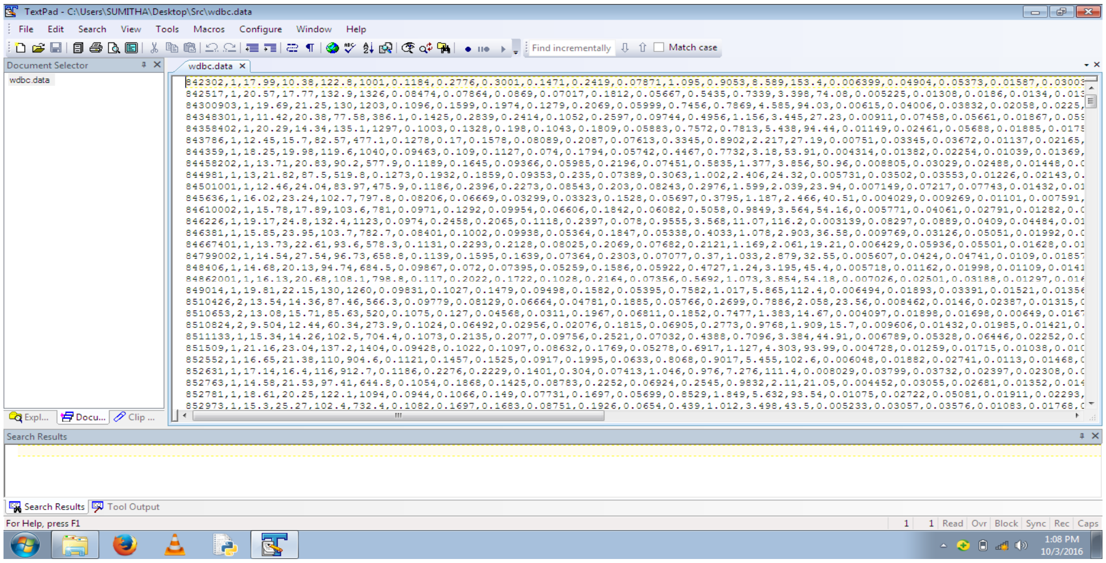
The methods used for this research is categorized as the sequential algorithm, Divide and Conquer Kernal Solving Support Vector Machine (DCKSVM) and Hybrid Radial Bias Neural Network (HRBFNN).

Data set Description: The breast cancer dataset used in this research is taken from the URI repository having 567 data with gene expression value which is shown in Figure-1. From this data, 380 data are taken as training data and remaining data are for testing data. The number of attributes taken for predicting the efficiency from this dataset is thirty-four and Matlab is the software that has been used for implementing this work. The URI repository link of this prognostic type of breast cancer dataset: http://archive.ics.uci.edu/ml/datasets/breast cancer+winconsin+%28.

The parameters used for predicting the efficiency are accuracy, precision, recall and f-measure. But this performance measure for these classifiers in identifying disease is based on confusion matrix method which has been computed from this breast cancer dataset [5]. Algorithms :

Sequential Model It is used to distinguish the differences between the pair of gene expression values in the dataset. For example, gene1, gene2. …..gene10 are genes in the dataset in which the gene pair (gene4 gene8gene2) and the another gene pair (gene8gene4gene2) are found to be similar [6]. Then it shows that this two gene pair stimulates the same disease in the human body. This sequences of each gene in breast cancer dataset are taken as an input for predicting results [7].

DCKSVM: The most commonly used classification algorithm is the DCKSVM algorithm which applied to the breast cancer dataset for predicting the accuracy [9,10]. DCKSVM algorithm [8] is mainly used to divide the main clusters of data into a number of sub-clusters which makes its predictivity which becomes reliable on that clusters in this breast cancer dataset [8].



```
842302,1,17.99,10.38,122.8,1001,0.1184,0.2776,0.3001,0.1471,0.2419,0.07871,1.095,0.9053,8.589,153.4,0.006399,0.04904,0.05373,0.01587,0.03003
842517,1,20.57,17.77,132.9,1326,0.08474,0.07864,0.0869,0.07017,0.1812,0.05667,0.5435,0.7339,3.398,74.08,0.005225,0.01308,0.0186,0.0134,0.013
84300903,1,19.69,21.25,130,1203,0.1096,0.1599,0.1974,0.1279,0.2069,0.05999,0.7456,0.7869,4.585,94.03,0.00615,0.04006,0.03832,0.02058,0.0225,
84348301,1,11.42,20.38,77.58,386.1,0.1425,0.2839,0.2414,0.1052,0.2597,0.09744,0.4956,1.156,3.445,27.23,0.00911,0.07458,0.05661,0.01867,0.059
84358402,1,20.29,14.34,135.1,1297,0.1003,0.1328,0.198,0.1043,0.1809,0.05883,0.7572,0.7813,5.438,94.44,0.01149,0.02461,0.05688,0.01885,0.0175
843786,1,12.45,15.7,82.57,477.1,0.1278,0.17,0.1578,0.08089,0.2087,0.07613,0.3345,0.8902,2.217,27.19,0.00751,0.03345,0.03672,0.01137,0.02165,
844359,1,18.25,19.98,119.6,1040,0.09463,0.109,0.1127,0.074,0.1794,0.05742,0.4467,0.7732,3.18,53.91,0.004314,0.01382,0.02254,0.01039,0.01369,
84458202,1,13.71,20.83,90.2,577.9,0.1189,0.1645,0.09366,0.05985,0.2196,0.07451,0.5835,1.377,3.856,50.96,0.008805,0.03029,0.02488,0.01448,0.0
844981,1,13,21.82,87.5,519.8,0.1273,0.1932,0.1859,0.09353,0.235,0.07389,0.3063,1.002,2.406,24.32,0.005731,0.03502,0.03553,0.01226,0.02143,0.
84501001,1,12.46,24.04,83.97,475.9,0.1186,0.2396,0.2273,0.08543,0.203,0.08243,0.2976,1.599,2.039,23.94,0.007149,0.07217,0.07743,0.01432,0.01
845636,1,16.02,23.24,102.7,797.8,0.08206,0.06669,0.03299,0.03323,0.1528,0.05697,0.3795,1.187,2.466,40.51,0.004029,0.009269,0.01101,0.007591,
84610002,1,15.78,17.89,103.6,781,0.0971,0.1292,0.09954,0.06606,0.1842,0.06082,0.5058,0.9849,3.564,54.16,0.005771,0.04061,0.02791,0.01282,0.0
846226,1,19.17,24.8,132.4,1123,0.0974,0.2458,0.2065,0.1118,0.2397,0.078,0.9555,3.568,11.07,116.2,0.003139,0.08297,0.0889,0.0409,0.04484,0.01
846381,1,15.85,23.95,103.7,782.7,0.08401,0.1002,0.09938,0.05364,0.1847,0.05338,0.4033,1.078,2.903,36.58,0.009769,0.03126,0.05051,0.01992,0.0
84667401,1,13.73,22.61,93.6,578.3,0.1131,0.2293,0.2128,0.08025,0.2069,0.07682,0.2121,1.169,2.061,19.21,0.006429,0.05936,0.05501,0.01628,0.01
84799002,1,14.54,27.54,96.73,658.8,0.1139,0.1595,0.1639,0.07364,0.2303,0.07077,0.37,1.033,2.879,32.55,0.005607,0.0424,0.04741,0.0109,0.01857
848406,1,14.68,20.13,94.74,684.5,0.09867,0.072,0.07395,0.05259,0.1586,0.05922,0.4727,1.24,3.195,45.4,0.005718,0.01162,0.01998,0.01109,0.0141
84862001,1,16.13,20.68,108.1,798.8,0.117,0.2022,0.1722,0.1028,0.2164,0.07356,0.5692,1.073,3.854,54.18,0.007026,0.02501,0.03188,0.01297,0.016
849014,1,19.81,22.15,130,1260,0.09831,0.1027,0.1479,0.09498,0.1582,0.05395,0.7582,1.017,5.865,112.4,0.006494,0.01893,0.03391,0.01521,0.01356
8510426,2,13.54,14.36,87.46,566.3,0.09779,0.08129,0.06664,0.04781,0.1885,0.05766,0.2699,0.7886,2.058,23.56,0.008462,0.0146,0.02387,0.01315,0
8510653,2,13.08,15.71,85.63,520,0.1075,0.127,0.04568,0.0311,0.1967,0.06811,0.1852,0.7477,1.383,14.67,0.004097,0.01898,0.01698,0.00649,0.0167
8510824,2,9.504,12.44,60.34,273.9,0.1024,0.06492,0.02956,0.02076,0.1815,0.06905,0.2773,0.9768,1.909,15.7,0.009606,0.01432,0.01985,0.01421,0.
8511133,1,15.34,14.26,102.5,704.4,0.1073,0.2135,0.2077,0.09756,0.2521,0.07032,0.4388,0.7096,3.384,44.91,0.006789,0.05328,0.06446,0.02252,0.0
851509,1,21.16,23.04,137.2,1404,0.09428,0.1022,0.1097,0.08632,0.1769,0.05278,0.6917,1.127,4.303,93.99,0.004728,0.01259,0.01715,0.01038,0.010
852552,1,16.65,21.38,110,904.6,0.1121,0.1457,0.1525,0.0917,0.1995,0.0633,0.8068,0.9017,5.455,102.6,0.006048,0.01882,0.02741,0.0113,0.01468,0
852631,1,17.14,16.4,116,912.7,0.1186,0.2276,0.2229,0.1401,0.304,0.07413,1.046,0.976,7.276,111.4,0.008029,0.03799,0.03732,0.02397,0.02308,0.0
852763,1,14.58,21.53,97.41,644.8,0.1054,0.1868,0.1425,0.08783,0.2252,0.06924,0.2545,0.9832,2.11,21.05,0.004452,0.03055,0.02681,0.01352,0.014
852781,1,18.61,20.25,122.1,1094,0.0944,0.1066,0.149,0.07731,0.1697,0.05699,0.8529,1.849,5.632,93.54,0.01075,0.02722,0.05081,0.01911,0.02293,
852973,1,15.3,25.27,102.4,732.4,0.1082,0.1697,0.1683,0.08751,0.1926,0.0654,0.439,1.012,3.498,43.5,0.005233,0.03057,0.03576,0.01083,0.01768,0
```

Figure-1:Breast Cancer Dataset

Algorithm 1: Divide and Conquer SVM

Input: Training Data

Output: The SVM solution A.

Step 1: Partitioning the variables into k subsets {v1….vk}

Step 2: Time complexity for solving sub-problems reduced to O(k*n/k)2=O(n2/k) with space complexity, where n is the variable and k is the cluster subset.

Step 3: After computation of all subproblem solutions, concatenate them to form solution for the whole problem ab=[a1….ak]

Step 4: A bound is derived on ||ab-a*||2 where ab is the optimal solution by adding cluster-kernel values.

Step 5: Minimizing the off-diagonal values of the kernel matrix with a balancing normalization.

Step 6: Go to Step 2 for partition the data in each cluster k, and calculate Step 4 for absolute scale.

HRBFNN: HRBFNN is applied to improve the capabilities of data granulation and for pre-processing the data in the breast cancer dataset [11]. It incorporates the characteristics of both data granulation and Principal Component Analysis

(PCA) method which is mainly used for preprocessing the data or neglecting the missing values in the dataset [12,13].

Algorithm 2: HRBFNN

Step 1: Preprocess the data set using PCA for dimentionality reduction.

Step 2: Training and testing data sets are formed.

Step 3: The generic parameters used in this research are decided.

Step 4: Selected inputs are determined.

Step 5: PFNs are designed. For the selecting r inputs, the number of nodes (PFNs) generated in each layer becomes equal to k = n!

n!(n−r)! r! , where, n is the number of total inputs and r stands for the number of the chosen input variables and k is the clusters [11].

Step 6: Check the termination criterion.

Step 7: Select the best predictive capability nodes and construct their corresponding layer.

It takes only a few seconds to complete 1000 iterations and generates high performance than other algorithms in terms of accuracy, precision, recall and f-measure.

RESULTS AND DISCUSSION

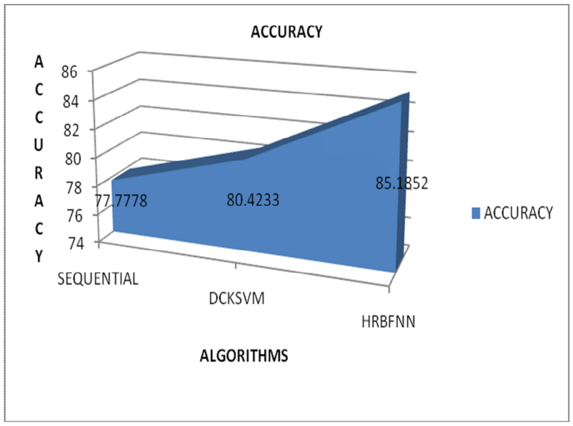
In this research, the machine learning algorithms which are applied to analyze the prognostic breast cancer dataset for detecting diseases are shown in the Table-1. HRBFNN algorithm gives better results when compared to the sequential algorithm and DCKSVM algorithm. The performance can be calculated in terms of accuracy, precision, recall and F-Measure. The criterion for these classifiers in disease detection is based on Confusion matrix. The accuracy percentage of the sequential model, DCKSVM and HRBFNN is 78, 80 and 85 respectively as depicted in Table-1. The results show that the HRBFNN gives 85.18 percentage of accuracy, 0.79 percentage of precision and 0.83 percentage of F-measure which is higher than the sequential model and the DCKSVM algorithms. However, DCKSVM algorithm shows higher recall percentage than sequential model and HRBFNN. But, the accuracy value is most significant in pridicting the disease – causing gene. Hence, HRBFNN is best in predicting the BRCA gene in the dataset. The graph plotted for these results are shown in Figure-2.

Table-1:Results of sequential, DCKSVM and HRBFNN

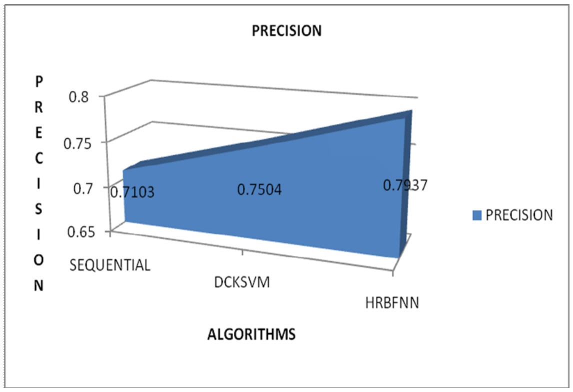
Algorithms Parameters	Sequential Model(%)	DCKSVM(%)	HRBFNN (%)
Accuracy	77.7778	80.4233	85.1852
Precision	0.7103	0.7504	0.7937
Recall	0.7659	0.8923	0.8713
F-measure	0.7371	0.7892	0.8307

Accuracy is the percentage of correct predictions from the dataset. Accuracy is the Proportion of the total number of predictions that were correct and it can be calculated using the equation,

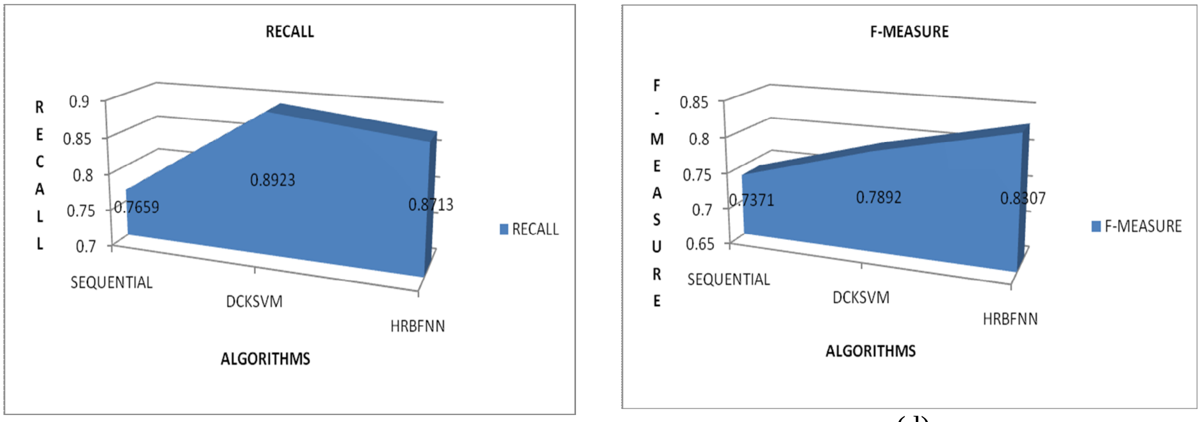
$$\text{Accuracy}= p+s/(p+q+r+s) \quad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$



(a)



(b)



(c) (d)

Figure-2: (a) Accuracy graph of Sequential, DCKSVM and HRBFNN Algorithms (b) Precision graph of Sequential, DCKSVM and HRBFNN Algorithms (c) Recall graph of Sequential, DCKSVM and HRBFNN Algorithms (d) F-measure graph of Sequential, DCKSVM and HRBFNN Algorithms

Recall is the proposition of positive cases that are correctly identified and it can be calculated using the equation,

Recall=s/(r+s) ……………...(2)

Precision is the proposition of the predicted positive cases that were correct and it can be calculated using the equation,

Precision=s/(q+s) ……………(3)

F-measure (F) can be calculated using the equation,

F=2* (Precision*Recall)/(Precision+Recall) .... (4)

Where p is the number of correctly predicted genes which are not disease-causing gene, q is the number of incorrectly predicted genes which predicted as diseases causing gene, r is the number of incorrect prediction of undiseased gene and s is the number correctly predictions of disease-causing genes in the dataset.

Cho-Jui et al., [8] calculated accuracy and time of DCKSVM with LIBSVM in covtype dataset. But it is compared within the SVM classifiers itself and predicted the results in absence of recall value. In this research, DCKSVM is compared with other classifiers rather than SVM classifiers and predicted the accuracy as 80.4 percentage and recall as 0.89 percentage. Wei et al., [11] estimated the performance of HRBFNN with the use of PCA and without the use of PCA for determining the predicting capabilities which show that the use of PCA with HRBFNN gives better performance than the without use of PCA experimented on using a series of numerical data. Hence this algorithm is selected for this research and predicted the accuracy of 85.18 percentage which is the highest performance than other classifiers in diagnosing of breast cancer.

Machine learning algorithm is most commonly used in all the research related to this work, but particularly in SVM classifiers, DCK SVM is implemented for the purpose of dividing the problem into sub-problems which aid for independent solutions with high efficiency. Table-1 concludes that HRBFNN predicts high efficiency than other classifiers. Based on the confusion matrix calculation method, the parameters used for predicting the results are shown in equation (1) to (4).

CONCLUSION

In this research, HRBFNN algorithm proved that it is effective in analyzing the BRCA gene from the breast cancer dataset. The performance analysis of the sequential algorithm with DC KSVM and HRBFNN shows that the HRBFNN gives better efficiency than the DCKSVM and sequential algorithm. It can also be further enhanced with some other computer-assisted algorithms to improve the efficiency.

REFERENCES

[1]. Erin L. Linnenbringer, Social Constructions, Biological Implications: A Structural Examination of Racial Disparities In Breast Cancer Subtype, University of Michigan, (2014)

[2]. Laura J.Van, M.Mao, L.Hans, M.J. Marton, P.S. Linsley, Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature* 15: 415 (2012) www.nature.com,

[3]. Guha S., A.Meyerson, N.Mishra, R.Motwani and O'Callaghan, Clustering Data Streams: Theory and Practice. *IEEE Transactions*

on Knowledge and Data Engineering 15: 515-528 (2003) http://dx. doi. org/ 10.1109/TKDE.2003.119 8387.

[4]. Ingrid A. Hedenfalk, Markus Ringner, Jerey M. Trent and Ake Borg, Gene Expression in Inherited Breast Cancer. *Advances in Cancer Research* 4: 1-34, (2002)

[5]. Isabelle G., W.Jason, B.Stephen and V. Vladimir, Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* 16(1): 389-422 (2014)

[6]. Ken Kaneiwa, A sequential pattern mining algorithm using rough set theory. *International Journal of Approximate Reasoning* 52: 881–893 (2011)

[7]. Carl H. Mooney and John F. Roddick, Sequential Pattern Mining – Approaches and Algorithms. *ACM Computing Surveys* (2013)

[8]. Cho-Jui, S. Si and Inderjit S. Dhillon, A Divide and Conquer Solver for Kernel Support Vector Machines. *Proceedings of the* 31st *International Conference on Machine Learning,* Beijing, China (2014)

[9]. Minseung K. and K.Sung-Hou, Empirical prediction of genomic susceptibilities for multiple cancer classes. *PNAS* 11(5): 1921–1926 (2015)

[10]. Oana F., I.Diana and T.Thomas, A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts. *IEEE Transactions on Knowledge and Data Engineerin* 23(6): (2016)

[11]. Wei H., Sung-Kwun and P. Witold, Design of hybrid radial basis function neural networks (HRBFNNs) realized with the aid of hybridization of fuzzy clustering method (FCM) and polynomial neural networks (PNNs). *Neural Networks* 60: 166–181 (2015)

[12]. Hilmi Berk Celikoglu, Application of radial basis functions and generalized regression neural networks in non-linear utility function specification for travel mode choice modeling. *Mathematical and Computer Modelling* 44: 640–658 (2006)

[13]. Seema Singh and H. Sushmitha, An Efficient Neural Network Based System for Diagnosis of Breast Cancer. *International Journal of Computer Science and Information Technologies* 5(3): .4354-4360 (2014)