International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-5, Issue-6, June 2019
ISSN: 2395-3470
www.ijseas.com

A Survey on Cancer Analysis Using Data Mining Techniques

Md. Moazzem Hossain[1], Dr. Md. Mamunur Rashid[2], Md. Abu Bakar Siddik[3], Md. Sydur Rahman[4]
[1, 2, 4] Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology, Saidpur, Bangladsh
[3] Department of Electrical and Electronics Engineering, Bangladesh Army University of Science and Technology, Saidpur, Bangladsh

Abstract
Cancer is nothing but abnormal cell growth with the potential to spread on the other parts of the body. There are two types of cancers Benign (noninvasive) or Malignant (cancerous). Early diagnosis and treatment helps to prevent the spread of cancer. Now a days data mining is in use for the effective and accurate diagnosis of the cancer disease by using various data mining techniques and machine learning techniques. This survey paper is about the application of different data mining techniques for analysis of cancer diseases using different data sets. Which helps the medical professionals in decision making for diagnosis of cancer diseases in order to provide proper treatments.

Keywords: *Cancer, Disease, Data Mining, Diagnosis, Treatment.*

1. Introduction

Today data mining techniques are used as a popular research tool for medical researchers to identify and exploit patterns and relationships among large number of variables, and made them able to predict the outcome of a disease using the historical datasets. It has become a fundamental methodology for computing applications in medical technology. The most important aspect is early and accurate diagnosis of any disease which helps to cure and increase the life expectancy chances. Cancer is a disease where accurate diagnosis can reduce the death rate in cancer patients. So in medical, predicting the outcome of a disease is one of the most interesting and challenging tasks where to develop data mining applications.

The objective of these predictions is to assign patients to either a benign group that is noncancerous or a malignant group that is cancerous. The prognosis problem is the long-term outlook for the disease for patients whose cancer has been surgically removed. In this problem a patient is classified as a recur if the disease is observed at some subsequent time to tumor excision and a patient for whom cancer has not recurred and may never recur.

The objective of these predictions is to handle cases for which cancer has not recurred (censored data) as well as case for which cancer has recurred at a specific time. As the use of computers powered with automated tools, large volumes of medical data are being collected and made available to the medical research groups. As a result, Knowledge Discovery in Databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers. The objective of this study is to summarize various review and technical articles on Cancer data sets. It gives an overview of the current research being carried out on these datasets using different Data Mining techniques.

2. Knowledge discovery and mining techniques

This section provides an introduction to knowledge discovery and data mining techniques and a brief discussion about data sets. Here the list of various analysis tasks have been included so that the goals of a discovery process and lists methods and research areas that is promising in solving these analysis tasks.

2.1 Knowledge Discovery Process

The terms Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably. KDD is the process of turning the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are often treated as equivalent words but in real data mining is an important step in the KDD process. The

Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps: Data cleaning and integration, Data selection and Transformation, Data mining, Evaluation and Presentation and Knowledge representation.

2.2 Data Mining Techniques

In the KDD process, the Data Mining methods are for extracting patterns from data. The patterns that can be discovered depend upon the data mining tasks applied. Generally, there are two types of data mining tasks: Descriptive data mining tasks that describe the general properties of the existing data, and Predictive data mining tasks that attempt to do predictions based on available data. Data mining can be done on data which are in quantitative, textual, or multimedia forms. Data mining involves some of the following key steps: Problem definition, Data explorations, Data preparation, Modelling, Evaluation and Deployment.

2.2 Data Sets
In this paper our concentration mainly on Tumor and cancer datasets.

1) Primary-Tumor: A primary tumor refers to a tumor or mass that is growing in the location where cancer originated. For instance, if a patient is diagnosed with stomach cancer the primary tumor would be found in the stomach itself rather than elsewhere in the body. The primary tumor is generally the easiest to remove; however, its removal does not necessarily mean that the patient is cancer-free. The first step in cancer treatment often involves the removal of the primary tumor, although this does not guarantee a recovery.

2) Colon Cancer: A Colon Tumor is an abnormal growth of cells found in the colon and can be an indication of colon cancer. If the colon tumor spreads to the bottom part of the colon, also known as the rectum, it can be an indication of colorectal cancer. As the image below shows, the colon is the large intestine or large bowel. The rectum is the passageway that connects the colon to the anus. Some Colon tumors are non-cancerous and are called benign polyps. Since benign polyps do not cause colon cancer, they are not dangerous, but if they are not identified and removed, they can change into cancerous tumors.

3) Breast Cancer: Breast cancer is the most common cancer disease among women. The information about the tumor from certain examinations and diagnostic tests are gathered using staging to determine how widespread the cancer is. The stage of a cancer is one of the most important factors in selecting treatment options, and it uses the Tumor, Nodes and Metastasis (TNM) system. When a patients T, N, and M categories have been determined then this information is combined in a process known as stage grouping to determine a woman's disease stage. This is expressed as from Stage 0 (the least advanced stage) to Stage IV (the most advanced stage). Breast cancer is a malignant tumor, grew from cells of the breast. Hence, cancer of breast tissue is called breast cancer.

4) Blood Cancer: Blood cancers, or hematologic cancers, affect the production and function of blood cells. Most of these cancers start in the bone marrow where blood is produced. Common types of blood cancer include:

Leukemia: Cancer that originates in blood-forming tissue. Non-Hodgkin lymphoma: Cancer that develops in the lymphatic system from cells called lymphocytes, a type of white blood cell that helps the body fight infections.

Hodgkin lymphoma: Cancer that develops in the lymphatic system from cells called lymphocytes. Hodgkin lymphoma is marked by the presence of an abnormal lymphocyte called the Reed-Sternberg cell (or B lymphocyte).

Multiple myeloma: Cancer that begins in the blood's plasma cells, a type of white blood cell that is made in the bone marrow.

3. Literature survey on data classification methods

In data mining, classification is one of the most

important tasks. It maps the data in to predefined targets. It is a supervised learning as targets are predefined. The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. The commonly used methods for data mining classification tasks can be classified into the following groups such as Decision Trees, Support Vector Machines, Evolutionary Programming, Fuzzy Sets, Neural Networks and Rough Sets etc.

3.1 Tumor Data Sets

Aruna sundaram developed a hybrid statistical pattern recognition algorithm with a title of, Hybrid SPR algorithm to select predictive genes for effectual cancer classification. In this study, a hybrid Statistical pattern recognition algorithm has been proposed to reduce the dimensionality and select the predictive genes for classification of cancer. Colon cancer gene expression profiles having 62 samples of 2000 genes were used for the experiment.

Bijan Moghimi-Dehkordi and Azadeh Safaee [8] conducted experiments on survival rates and prognosis in Asia with a title of an overview of colorectal cancer survival rates and prognosis in Asia. In this study, in Asia, the overall cure rate of colorectal cancer has not improved dramatically in the last decade, 5-year survival remaining at approximately 60 Srinivas Mukkamata et.al. [37] found that Computational intelligent technique that can be useful at the diagnosis stage to assist the Oncologist in identifying the malignancy of a tumor. For finding accuracy of classifications Linear genetic Programs, Multivariate Re- gression Spines (MARS), Classification and Regression Tress (CART) and Random Forests are use.K Rzysztof Fujarewicz et.al. [22] explored that use the Recursive Feature Selection (RFR) method for finding suboptimal gene subsets for tumor tissue classification. They find that RFR method is able to find the smallest gene subset that gives no misclassification in leave-one-out cross-validation for tumor colon data set.

Parvesh Kumar and Siri Krishan Wasan [38] conducted experiments on colon data set using different decision tree algorithms. In these C4.5 and Bayesian logistic regression have less value of absolute relative error? G. Sujatha et.al. [39] conducted experiments on ID3,C4.5 and CART classifiers for better accuracy and execution time to construct the tree. It is observed that C4.5 performs well for tumor datasets, if available datasets are used as it is. Among these three algorithms, C4.5 itself is the best one for enhanced data set of Primary tumor and for enhanced Colon tumor data set both ID3 and C4.5 exhibit equal classification accuracy.

3.2 Breast Cancer Diagnosis

Clinical diagnosis of breast cancer helps in predicting the malignant cases. A lump felt during the examination roughly give clues as to the size of tumor and its texture. The various common methods used for breast cancer diagnosis are Mammography, Biopsy, Positron Emission Tomography and Magnetic Resonance Imaging. This section consists of the review of various technical and review articles on data mining techniques applied in breast cancer diagnosis.

K. Rajesh et.al. [29] attempted to classify SEER breast cancer data into the groups of Carcinoma in situ and Malignant potential using C4.5 algorithm. We obtained an accuracy of 94 accuracy of 93The authors Aruna et.al. [3] presented a comparison of classification algorithms on the Wisconsin Breast Cancer dataset. They have analyzed the classification results of only five classification algorithms namely Naive Bayes, Support Vector Machines (SVM), Radial Basis Neural Networks (RB- NN), Decision trees J48 and simple CART. Chul-Heui Lee et.al. [23] proposed a new classification method based on the hierarchical granulation structure using the rough set theory. The classification rules had minimal attributes and the knowledge reduction was accomplished by using the upper and lower approximations of rough sets. A simulation was performed on WBC dataset to show the effectiveness of the proposed method.

Aboul Ella Hassanien, and Jafar M.H.Ali[17] presented a rough set method for generating classification rules from a set of observed 360 samples of the WBC data. The attributes were

selected, normalized and then the rough set dependency rules were generated directly from the real value attribute vector. They made a comparison between the obtained results of rough sets with the well-known ID3 decision tree and concluded rough sets showed higher accuracy and generated more compact rules. Sudhir D. Sawarkar et. al. [41] applied SVM and ANN on the WBC data. The results of SVM and ANN prediction models were found comparatively more accurate than the human being. The 97these prediction models can be used to take decision to avoid biopsy. In [24], the authors D.Lavanya et.al., analyzed the performance of decision tree classifiers on various medical datasets in terms of accuracy and time complexity and proved that CART is the best.

3.3 Breast Cancer Prognosis

Once a patient is diagnosed with breast cancer, the malignant lump must be excised. During this procedure physicians must determine the prognosis of the disease. This is the prediction of the expected flow of the disease. Prognosis is important because the type and intensity of the medications are based on it. Prognosis helps in establishing a treatment plan by predicting the outcome of a disease [34].

C4.5 is a well-known decision tree induction learning technique which has been used by Abdelghani Bellaachia and Erhan Gauven [7] along with two other techniques i.e. Nave Bayes and Back-Propagated Neural Network. Authors presented an analysis of the prediction of survivability rate of breast cancer patients using above data mining techniques and used the new version of the SEER Breast Cancer Data. Authors found out that model generated by C4.5 algorithm for the given data has a much better performance than the other two techniques. W. Nick Street [35] applied ANN classification to Wisconsin Prognostic Breast Cancer and SEER datasets for the analysis of survival. Author developed a novel encoding as good and poor prognosis of censored data in ANN architecture to provide a framework for prognostic prediction.

Delen et.al.[14] compared ANN, decision tree and logistic regression techniques for breast cancer survival analysis. They used the SEER data twenty variables in the prediction models. The decision tree with 93.6accuracy and ANN with 91.2Jong Pill Choi et.al. [11] compared the performance of an Artificial Neural Network, a Bayesian Network and a Hybrid Network used to predict breast cancer prognosis. The hybrid Network combined both ANN and Bayesian Network. The Nine variables of SEER data which were clinically accepted were used as inputs for the networks. They found the proposed Hybrid model can also be useful to take decisions.

Muhammad Umer Khan et.al. [21] investigated a hybrid scheme based on fuzzy decision trees on SEER data; they performed experiments using different combinations of number of decision tree rules, types of fuzzy membership functions and inference techniques. They com- pared the performance of each for cancer prognosis and found hybrid fuzzy decision tree classification is more robust and balanced than the independently applied crisp classification.

Harry B. Burke et.al.[9] compared the TNM staging systems predictive accuracy with that of ANN for 5years survival of patients. They made the comparison over three different datasets these are SEER data, PCE data and PCE colorectal dataset. In all cases they found ANNs more accurate than the TNM staging system.

4. Literature survey on ensemble techniques

Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a vote of their predictions. The original ensemble method is Bayesian averaging, but more recent algorithms include error correcting output coding, Bagging, and Boosting. The approach of Ensemble systems is to improve the confidence with which we are making right decision through a process in which various opinions are weighed and combined to reach a final decision. Some of the techniques are: Bagging and Boosting.

4.1 Tumor Data Sets

Jinyam LiHuiqing Liu et.al. [19] conducted experiments on ovarian tumor data to diagnose cancer using C4.5 with and without bagging. G.Sujatha et.al.[40] conducted experiments on ID3,C4.5 and CART algorithms on tumor datasets

with boosting and bagging techniques consider for the comparison of performance of accuracy and time complexity of datasets. By conducting the experiments it is observed that C4.5 with Bagging is the best algorithm for finding out whether the tumor is benign or malignant on the tumor datasets which are used as they are available. On increasing the number of instances of the data sets ID3 with boosting is best for Primary tumor data set and ID3 with bagging is best for Colon tumor data set.

4.2 Breast Cancer Data Sets

Aik Choon Tan et.al. [42] focused on three different super- vised machine learning techniques on seven publicly available microarray data in cancer classification, namely C4.5 decision tree, and bagged and boosted decision trees. They observed that ensemble learning (bagged and boosted decision trees) often performs better than single decision trees in this classi- fication task. D.Lavanya et.al. [25] proposed a hybrid method to enhance the classification accuracy of Breast cancer data sets. In this feature selection methods used to eliminate those attributes that have no significance in the application process. The experimental results of a hybrid approach with the com- bination of preprocessing, bagging with CART demonstrated the enhanced classification accuracy of the selected data sets. D.Lavanya et.al. [26] experimented on simple CART on med- ical datasets with feature selection and ensemble techniques. In this, it is clear that in the ensemble method Bagging is preferable for diagnosis of breast cancer data than Boosting.

Tan AC, et.al. [43] used C4.5 decision tree, bagged decision tree on seven publicly available cancerous micro array data and compared the prediction performance of these methods. Abdelghani Bellaachia, et.al. [1] conducted experiments on SEER (Surveillance, Epidemiology, and End Results) public- use datasets using naive-bayes, back propagated neural net- works and C4.5 decision tree algorithms. They found that C4.5 exhibited better result than remaining algorithms. In[31] S Seema et.al., used an ensemble classification technique applied to various micro array gene expression cancer datasets to distinguish between healthy samples and cancerous samples. The results showed that ensemble classifiers give a better performance in terms of accuracy over individual classifiers when applied to various cancer datasets. CaiLing Dong et.al.[10] proposed a modified Boosted Decision Tree for breast cancer detection to improve the accuracy of classification. Jaree Thangkam et.al. [18] performed a work on survivability of patients from breast cancer. The data considered for analysis was Srinagarind hospital databases during the period 1990- 2001. In this approach CART is used as a base learner.

5. Literature survey on classification with feature selection

Feature selection (FS) is a crucial part in classification. This is also one of the pre-processing techniques in data mining. Feature selection is extensively used in the fields of statistics, pattern recognition and medical domain. Feature Se- lection means reducing the number of attributes. The attributes are reduced by removing irrelevant and redundant attributes, which do not have significance in classification task. The feature selection improves the performance of the classification techniques. These are classified such as: Filter, Wrapper and Hybrid approaches.

5.1 Breast Cancer Datasets

Asha Gowda Karegowda et.al.[6] proposed a wrapper ap- proach with genetic algorithm for generation of subset of attributes with different classifiers such as C4.5,Nave Bayes, Bayes Networks and Radial basis functions. The above clas- sifiers are experimented on the datasets Diabetes, Breast cancer, Heart Statlog and Wisconsin Breast cancer. Aboul Ella Hassanien [2] had experimented on breast cancer data using feature selection technique to obtain reduced number of relevant attributes, further decision tree ID3 algorithm is used to classify the data. Kemal Polat et.al.[20], proposed a new classification algorithm feature selection-Artificial Immune Recognition System (FS-AIRS) on breast cancer

data set. To reduce the data set C4.5 decision tree algorithm is used as a feature selection method. Deisy.C et.al. [13] experimented breast cancer data using three feature selection methods Fast correlation based feature selection, Multi thread based FCBF feature selection and Decision dependent -decision independent correlation further the data is classified using C4.5 decision tree algorithm. Mark A. Hall et.al. [28] have done experiments on various data sets using Correlation based filter feature selection approach further the reduced data is classified using C4.5 decision tree algorithm. Shomona Gracia Jacob et.al.[33] have considered the Wisconsin Prognostic Breast Cancer (WPBC) dataset. According to findings, Fisher Filtering, Backward Logistic Regression, Stepwise Discriminant Analysis and ReliefF filtering algorithms have performed well in terms of improving classifier accuracy on this dataset. S.Sagar Imambi et.al.[30] shows that GRW Feature selection schema used at preprocessing stage improves the performance of Pubmed abstract Classification. This algorithm shows that GRW works well in high dimension and unevenly distributed document classification. Only Bayes learning shows less accuracy, but other three learners show high accuracy rate. S.Aruna et.al.[4] Proposed CSSFFS feature selection algorithm for detecting Breast cancer. This is a greedy algorithm based on constrained search. This is a hybrid algorithm with the combination of filters and wrapper s. Attributes are ranked with the square of weights calculated by the SVM classifier. This acts as Filters to remove irrelevant features. From the remaining features SFFS with SVM is used to select the optimum subset of features. This act as a wrapper to remove the redundant features if any yields the required optimum subset. BER is used as the main criterion for selecting features. The objective of this algorithm is to select features with minimal BER. D.Lavanya [27]have considered Decision tree classifier-CART with and without feature selection in terms of accuracy, time to build a model and size of the tree on various Breast Cancer Datasets are observed. From the results it is clear that, though we considered only breast cancer datasets, a specific feat ure selection may not lead to the best accuracy for all Breast Cancer Datasets. The best feature selection method for a particular dataset depends on the number of attributes, attribute type and instances.

6. Conclusion

This paper provides a study of various technical and review papers on Tumor and Breast cancer data sets and explores that data mining techniques offer great promise to uncover patterns hidden in the data that can help the clinicians in decision making. From the above study it is observed that the accuracy for the diagnosis analysis of various applied Data mining Classification techniques, ensembling techniques are highly acceptable and can help the medical professionals in decision making for early diagnosis and to avoid biopsy. In the case of the said data sets it is observed that the best models and different types of technologies were applied.

References

[1] Abdelghani Bellaachia and Erhan Guven, “Predicting breast cancer sur- vivability using data mining techniques,” www.siam.org, mas id: 5251529, 2006.

[2] Aboul Ella Hassaneian, “Classification and feature selection of breast cancer data based on decision tree algorithm”, Studies and Informatics Control , vol12, no1,March 2003.

[3] Aruna, Dr S.P.Rajagopalan, “A Novel SVM based CSSFFS Feature Selection Algorithm for Detecting Breast Cancer”, International Journal of Computer Applications (0975 8887), Volume 31 No.8, October 2011,Pg.no:14-20.

[4] Asha Gowda Karegowda, M.A.Jayaram, A.S. Manjunath, “Feature Subset Selection Problem using Wrapper Approach in Supervised Learning”, International Journal of Computer Applications , 1(7):1317, February 2010.

[5] Bellaachia Abdelghani and Erhan Guven, “Predicting Breast Cancer Survivability using Data Mining Techniques”, Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining, 2006.

[6] Bijan Moghimi-Dehkordi,Azadeh Safaee,“An overview of colorectal can- cer survival rates and prognosis in Asia”, World Gastrointest Oncol 2012 April15, 4(4):Pg.no:71-75.

[7] Burke H. B. Et al , Artificial Neural Networks Improve the Accuracy of Cancer Survival Prediction, Cancer, 1997, vol.79, Pg.no:857-862.
[8] CaiLing Dong, YiLong Yin and XiuKun Yang, Detecting Malignant Patients via Modified Boosted tree, Science China Information Sciences, 2010.
[9] Choi J.P., Han T.H. and Park R.W., A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis, J Korean Soc Med Inform, 2009, Pg.no: 49-57.
[10] ColonTumor, http://www.wisegeek.com/what-is-a- colon-tumor.htm
[11] Deisy.C, Subbulakshmi.B, Baskar S, Ramaraj.N, Efficient Dimensional- ity Reduction Approaches for Feature Selection, Conference on Compu-tational Intelligence and Multimedia Applications, 2007.
[12] Delen Dursun , Walker Glenn and Kadam Amit , Predicting breast cancer survivability: a comparison of three data mining methods, Artificial Intelligence in Medicine ,vol. 34, Pg.no:. 113-127 , June 2005.
[13] EI-Sebakhy A. Emad, Faisal Abed Kanaan, Helmy T, Azzedin F and AI-Sushaim F, Evaluation of Breast Cancer tumor classification with unconstrained functional networks classifier, Computer Systems and Applications, IEEE, International Conference, 2006, Pg.no: 281-287.
[14] Han J. and Kamber M., Data Mining: Concepts and Techniques, 2nd ed., San Francisco, Morgan Kauffmann Publishers, 2001
[15] Hassanien Ella Aboul and Ali H.M. Jafar, Rough set approach for gen- eration of classsification rules of Breast cancer data, Journal Informatica, 2004, vol. 15, Pg.no: 2338.
[16] Jaree Thangkam Guandong Xu, Yanchun Zang and Fuchun Huang,HDKM08 Proceedings of the second Australian workshop on Health data and Knowledge Management, Vol 80.
[17] Jinyan LiHuiqing Liu, See-Kiong Ng and Limsoon Wong, Discovery of significant rules for classifying cancer diagnosis data, Bioinformatics 19(Suppl. 2) Oxford University Press 2003.
[18] Kemal Polat, Seral Sahan, Halife Kodaz and Salih Gnes, A New Classification Method for Breast Cancer Diagnosis: Feature Selection Artificial Immune Recognition System (FS-AIRS), In Proceedings of ICNC (2)'2005. Pg.no:830-838
[19] Khan M.U., Choi J.P., Shin H. and Kim M, Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare, Conf Proc IEEE Eng Med Biol Soc., 2008, Pg.no: 48-51.
[20] Krzysztof Fujarewicz, Malgorzata Wiench, Selecting differentially ex- pressed genes for colon tumor classification int.j.Appl.Math.Comput.Sci, 2003. Vol.3, No.3, Pg.no:327 -335.
[21] Lee Heui Chul, Seo Hak Seon and Choi Chul Sang, Rule discovery using hierarchial classification structure with rough sets, IFSA World Congress and 20th NAFIPS International Conference, 2001, vol.1 , Pg.no: 447-452.
[22] D.Lavanya, Dr.K.Usha Rani, Performance Evaluation of Decision Tree Classifiers on Medical Datasets,. International Journal of Computer Ap- plications 26(4):1-4, July 2011.
[23] D.Lavanya and Dr.K.Usha Rani, Ensemble Decision T ree Classifiers for Brest Cancer Data, International Journal of Information Technology Convergence and Services, Vol.2, No.1, Feb2012, Pg.no:17 -24
[24] D.Lavanya and Dr.K.Usha Rani, Ensemble Decision Making System for Breast Cancer Data, International Journal of Computer Applications, Vol.51-No.17, August 2012, Pg.no:19-23.
[25] D.Lavanya and Dr.K.Usha Rani, Analysis of feature selection with clas- sification :Breast cancer datasets, Vol.2-No.5, oct-nov: 2011, Pg.no:756- 763.
[26] Mark A. Hall, Lloyd A. Smith, Feature Subset Selection: A Correlation Based Filter Approach, In 1997 International Conference on Neural In-formation Processing and Intelligent Information Systems (1997), Pg.no: 855-858.
[27] Osmar R. Zaane, Principles of Knowledge Discovery in Databases. [Online]. Available webdocs.cs.ualberta.ca/zariane/courses/cmput690/notes/Chapter1/ch1.pdf
[28] Pantel Patrick , Breast Cancer Diagnosis and Prognosis.[Online].Available:http://citeseer.nj.nec.com/pantel98breast.html.
[29] Parvesh Kumar, Siri Krishan Wasan,Analysis Of Cancer Datasets Using Classification Algorithms,IJCSN,Vol10 No.6,June 2010 Pg.no:175-182.
[30] Primary tumor, http://www.wisegeek.com/what-is-a-primary-tumor.htm

[31] Rajesh,Dr.Sheila Anand,Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm, IJARCCE,Vol.1,Issue 2,April 2012

[32] S.Sagar Imambi ,T.Sudha ,A Novel Feature Selection Method for Classification of Medical Documents from Pubmed, International Journal of Computer Applications (0975 8887) ,Pg.no:29-33,Volume 26 No.9, July 2011.

[33] S seema, Srinivas KG et.al., Ensemble Classifiers with Stepwise Fea- ture Selection For Classification of Cancer Data, International Jour- nal of Pharamaceutical science and Health care,Issue 2, Pg.no:48 - 61,Vol6,December 2012.

[34] Shomona Gracia Jacob, R.Geetha Ramani ,Efficient Classifier for Clas- sification of Prognostic Breast Cancer Data through Data Mining Tech- niques, P roceedings of the World Congress on Engineering and Computer Science 2012 Vol I WCECS 2012, October 24-26, 2012, San Francisco, USA

[35] Srinivas Mukkamata,Qing Zhang Liu, Rajeev Verraghattam, Andrew H.sung, Computational Intelligent Techniques for Tumor Classification (Ubibs Microarray Gene Expression Data) Dept of Computer Science, New Maxico Tech, Socorro NM , USA 2002

[36] Street W.N., A Neural Network Model for Prognostic Prediction, Fif- teenth International Conference on Machine Learning, Madison, Wiscon- sin, Morgan Kaufmann, 1998.

[37] Sujatha, Dr.K.Usha Rani, Evaluation of Decision Tree Classifiers on Tumor Data sets,IJETTCS,Vol2,Issue4,July-aug2013,Pg.no:418-423

[38] Sujatha,Dr.K.Usha Rani,An Experimental Study on Ensemble of Deci- sion Tree Classifiers,IJAIEM,Vol 2,Issue 8,August 2013,Pg.no:300-306

[39] Sudhir D., Ghatol Ashok A., Pande Amol P., Neural Network aided Breast Cancer Detection and Diagnosis,7th WSEAS International Con- ference on Neural Networks, 2006.

[40] Tan,Gilbert, Ensembling machine learning on gene expression data for cancer classification, Proceedings of New Zealand Bioinformatics Conference, Te Papa, Wellington, New Zealand, ,13-14 February 2003.

[41] Tan AC, Gilbert D. Ensemble machine learning on gene expres- sion data for cancer classification, Appl Bioinformatics. 2003; 2 (3 Suppl):Pg.no:75-83.

[42] Wen-Jai Kuo, Ruey-Feng Chag, Dar-Ren Chen and Cheng Chun Lee, Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images, Breast Cancer Research and Treatment 66: 51-57, 2001, Kulwer Academic Publishers, Printed in the Netherlands.