ISSN NO. 2320-5407

*Journal homepage: http://www.journalijar.com*

*INTERNATIONAL JOURNAL OF ADVANCED RESEARCH*

**REVIEW ARTICLE**

# Computational Techniques in Breast Cancer Diagnosis and Prognosis: A Review

**Dr. E. S. Samundeeswari Associate Professor, Mrs. P. K. Saranya Research scholar**

## *Manuscript Info*

***Manuscript History:***

Received: 25 September 2015
Final Accepted: 15 October 2015
Published Online: November 2015

***Key words:***

Diagnosis, Prognosis, Survivability prediction, Proteomic, Genomic

***Corresponding Author***

**Dr. E. S. Samundeeswari**

## *Abstract*

Macro level information like histological, clinical and demographic details about a Breast cancer patient are used for research. Now this is supplemented with micro level information like genomic, proteomic and imaging of breast cancer. Problem domains on Breast cancer research include Diagnosis, Prognosis, Survivability and Recurrence Prediction. This paper gives a review of various types of data available and data mining techniques that are applied on different problem domain of Breast cancer research.



## INTRODUCTION

Recently the incidence of breast cancer is rising all over the world and is now the second most common cancer diagnosed in women next to cervical cancer. In last couple of decades, more emphasis is made towards cancer related research. New and innovative methods in image analysis, data and statistics driven research have been developed for early detection, diagnosis and prognosis to decrease the cancer related death rates. Data mining and image processing techniques extract novel information for making decisions regarding disease prediction, prognosis and selection of treatment methods.

### I BREAST CANCER

Treatments or tests for breast cancer are separated into three main types, screening, diagnosing and monitoring. Screening tests look for signs of cancer. If a screening test shows a breast change, an assumption of cancer, additional diagnostic tests are suggested. On detection of breast cancer, primary therapy is given to reduce or destroy the cancer cells. Primary therapy includes surgery to remove the breast (mastectomy) or to remove the tumor along with a layer of normal tissue around it (lumpectomy). Metastasize is the stage in which the cells may break away from the source tumour and spread to other internal parts of the body. Hence the doctor insists on adjuvant therapy for the patients to destroy the cancer cells that might have been possibly spread even though cancer cells are not detected through imaging or laboratory test. Adjuvant therapy is also recommended for the patients who had high risk of cancer recurrence. The main key idea behind this therapy is to give a chance of long life disease free survivability. Adjuvant therapy for breast cancer can include chemotherapy, hormonal therapy, the targeted drug, radiation therapy, or a combination of treatments. Adjuvant therapies have side effects. For example, chemotherapy is given as adjuvant treatment to prevent distant metastases or as palliation to treat patients with metastatic disease leads to side effects such as heart failure, leukemia and life threatening infections.

But the most effective way to reduce breast cancer deaths is to detect it earlier. Early diagnosis needs an accurate and reliable diagnosis system that can be used by physicians to distinguish benign tumours from malignant ones without going for surgical biopsy. Once the primary therapy is given, doctors need to decide on which patients might benefit from adjuvant treatments, using both prognostic and predictive factors. Prognosis is a medical term for predicting the likely outcome of a disease. Prognostic and predictive factors are needed to identify (i) patients with good responses for whom adjuvant systemic therapy is not much beneficial to warrant the risks; (ii) patients who

will be benefited with different form of therapy and (iii) patients with poor prognosis who may be subjected to more aggressive adjuvant therapy (Cianfrocca and Goldstein, 2004). Survival statistics which include case studies help us to make predictions about a patient's prognosis and determine the treatment options. The use of machine learning particularly the field of proteomics is part of a growing trend towards personalized predictive medicine in cancer prediction and prognosis.

## II DATA MINING

During past decades, detecting, classifying and distinguishing tumours or other malignancies were done using X-ray, CRT images and other clinical parameters. In other words, the most of the research was aimed to provide accurate cancer diagnosis and detection. Due to recent development in medical oncology, in addition to macro level information like histological (cell based), clinical (patient based) and demographic (population-based) data, molecular-scale information about patients or tumours like genomic (study of the genetic make-up of organisms), proteomic (study of proteins) and imaging systems are widely used. Biomarker testing, specifically the accurate assessment of somatic mutations in certain genes like p53, BRCA1, BRCA2, the over or misexpression of certain tumor proteins like MUC1, HER2, PSA, the estrogen, progesterone receptor (ER, PR), and Her2 status provide accurate and powerful prognostic or predictive indicators .

In the past, diagnosis based on the macro-level information with limited to number of variables is not enough to apply standard statistical evaluation methods or the physician's own intuition to predict cancer risks and outcomes. But today, even with hundreds of molecular, cellular and clinical parameters, human intuition and standard statistics alone don't generally work. Hence non-traditional, intensively computational approaches like data mining tools and soft computing methods in cancer prediction and prognosis came to exist. The objective of this paper is to summarize various review and technical articles on data mining techniques applied for diagnosis, susceptibility and prognosis of breast cancer.

## III TYPES OF BREAST CANCER DATA

This section contains the review of various applications of data mining techniques on different types of Breast Cancer data. Various data mining techniques used are variants of Artificial Neural Networks (ANN), Decision Tree, Naive Bayes, Support Vector Machines.

### *A. Clinical and histological data*

The clinic variables generally used in breast cancer diagnosis are Age, Menarche age, Menopause age, First pregnancy age, No. of miscarriages, No. of axillary nodes, Grade , Tumour size , No. of pregnancies, Estrogen and progesteron receptors, p53(a gene that codes TP53 protein which regulates the cell cycle), Ploidy (number of chromosomes), S-phase (Synthesis phase of cell cycle). Cancer stage can be classified according to four different attributes: the size of the cancer, invasive or non-invasive nature, whether cancer has spread to lymph nodes and to various parts of body. Tumour grade defines how far the breast cancer cells resemble normal cells. Poorly differentiated tumors have more chance to recur. Tumour cells that are merely like normal breast cells (well differentiated) have a tendency to have better prognosis. Currently most of the machine learning techniques are used to predict and prognosis cancer (Cruz and David, 2006).

### *B. Image data*

Medical imaging modality is one of the best way to diagnosis and evaluate the early stage of cancer. Mammogram, Magnetic Resonance Imaging (MRI), PET, Ultrasonogaphy are most frequently used imaging modality to diagnosis breast cancer. Among these, the most effective method of early detection of the breast cancer is mammograms. But certain characteristics in the mammograms imaging fail to determine whether cancer exists or not. This can be overcome by a follow-up Ultrasonography. Breast cancer often presents as a mass with or without the presence of calcifications. The location, size, shape, density and margins of the mass are useful for the radiologist in evaluating the likelihood of cancer. Research on Image data follows almost the same steps with some difference in techniques. The common steps are preprocessing, segmentation of breast image, feature extraction and feature classification.

### *C. Microarray data*

Decision on diagnosis, treatment of cancer and prognosis are generally based on macro level information. As cancer is caused by genetic aberrations, microarray technology has a great impact on cancer research. Microarray offers an efficient method of gathering data that can be used to determine the expression pattern in thousands of genes. It represents the whole genome i.e. it allows to monitor the expression levels of tens of thousands of genes simultaneously. This technology helps researchers to learn more about different diseases such as heart disease, infectious disease and especially the study of cancer. Until recently, different types of cancer have been classified on the basis of the organs in which the tumors exist. With the evolution of microarray technology, it is possible for

the researchers to further classify the types of cancer on the basis of the patterns of gene activity in the tumor cells. This helps the pharmaceutical community to develop more effective drugs as the treatment strategies that directly targeted to the specific type of cancer. Gene Expression Omnibus (GEO) is a public functional genomics data repository of National Center for Biotechnology Information (NCBI). (http://www.ncbi.nlm.nih.gov/geo/) Microarray gene expression datasets from human breast cancer is available with accession numbers: GSE2034, GSE4922, GSE6532, GSE7390, GSE11121.

### *D. Proteomic Data*

Proteomics is a branch of biotechnology to study the structure and funtions of proteins, protein complexes, localization, and their interactions. The activity and interaction of thousands of proteins can measured simultaneously using high-throughput mass spectrometry. Profiling proteomic patterns are done using biofluids like blood and urine. It can be further complemented with genomic portray of cancer. Proteomic profiling is the efficient discovery of biomarkers that reflects intrinsic genetic program (Dubitzky 2007). Cancer detection based on the application of data mining techniques to proteomic data has received a lot of attention in recent years (Srinivas et al., 2001) (Li et al., 2004).

## IV DATA MINING TECHNIQUES ON BREAST CANCER

### *A. Diagnosis*

A lot of research has been done for the early diagnosis of breast micro calcifications from the digitalized mammograms through the application of digital image processing techniques and preprocessing algorithms are proposed in papers (Alina et al., 2010, 2010; Hassanien and Ali,2004; Mohanalin et al.,2008; Shah et al., 2014; Thangavel et al., 2005). Malek et al., (2009) proposed a fuzzy logic technique to develop an automated breast cancer nuclei detection and classification system. At first, the automated system segments the nucleus in cytological image using active contour method. Then the textural features are extracted using the wavelet transform concept. Finally the fuzzy C-Means (FCM) algorithm is implemented to the extracted features to classify the images into benign and malignant category.

### *B. Prognosis*

Many studies have explored the susceptibility of various features on prognosis. Alieldin et al., (2014) analysed whether women of non-metastatic breast cancer had better prognosis based on the age at diagnosis using Chi-square test and multivariate analysis. Elkhodary et al., (2014) explored log-rank test and multivariate statistical methods for Node positive patients to calculate and predict the prognostic value of Lymph Node ratio. Clavel (2002) analysed the relationship between breast cancer and hormonal factors influencing menarche, pregnancy and menopause. Choi et al (2009) used Bayesian network model and ANN model to predict breast cancer prognosis and found that ANN and the proposed hybrid Bayesian Network model outperformed Bayesian model. Blows et al., (2010) investigated the relationships among short term, long term and subtype survival using immune histochemical markers. Six subtypes of breast cancer were defined by the markers. Glare P (2005) discussed that factors like performance status, symptoms of cancer cachexia syndrome, and patient-rated qualify of life, markers such as leukocytosis and acute phase reactants and cytokines have the potential to improve prognostic accuracy. Alexe et al., (2007) combined the Principal Component Analysis (PCA) and ensemble K-Means Clustering to cluster (group) and examine gene markers in microarray data. Giarratana et al., (2009) used various classifiers such as J48, NaiveBayes, AdaBoostM1, Bagging and Random Forests to identify the genes with levels of expression associated with a clinical prognosis for breast cancer.

### *C. Survivability Prediction*

Khan (2008) proposed a hybrid fuzzy decision tree and applied on SEER dataset to predict the survivability. Delen et al., (2005) used three data mining algorithms (ANN and decision trees, logistic regression) to develop prediction models using SEER dataset (http://www.seer.cancer.gov). They used 10-fold cross-validation methods for performance comparison. It is founded that decision tree algorithm –C5 with 93.6% accuracy rate proves better performance than artificial neural network method with 91.2% and logistic regression model with 89.2% accuracy rate. Chih-Lin Chi et al.,(2007) applied artificial neural networks to the survival analysis problem on two datasets-Wisconsin Prognostic Breast Cancer data and Love data. Endo et al. (2008) compared Artificial Neural Network, Bayes theorem, Regeression Model and Decision tree algorithm (ID3, J48) to predict breast cancer survival of SEER dataset using WEKA tool. It is concluded that Logistic Regression showed highest accuracy and J48 had the highest sensitivity and ANN had the highest specificity. Thongkam et al., (2009) used C-Support Vector classification for outlier filtering and over-sampling with replacement to solve the problem of imbalanced dataset. The performance measures such as receiver operating characteristic (ROC) curve and F-measure were used for evaluation.

### *D. Recurrence Prediction*

Eshlaghy et al. (2013) used three Machine Learning techniques (C4.5, SVM and ANN) for predicting the recurrence of breast cancer and found that SVM model predicts breast cancer recurrence with highest accuracy and least error rate. The clinical and histological data in ICBC dataset of National Cancer Institute of Tehran is used for experiment to predict the 2-year recurrence rate. Recently, microarray data have been used for predicting the outcome of cancer treatments (Van't Veer et al 2002). Gene expression signatures have been identified to classify breast tumors into sub types showing distinct expression profiles associated with specific clinical characteristics (Finak et al.,2008). Clustering is the most popular method currently used in the first step of gene expression data matrix analysis for finding co-regulated and functionally related groups **(**Selvaraj & Jeyakumar 2011). Shleeg et al., 2013 evaluated the breast cancer risk using Mamdani and Sugeno type model and found that Sugeno -type is advantageous than the other type. Sotiriou et al. (2006) investigate the examined the histological grade of gene expression profiles.

## V CONCLUSION

Breast cancer is the second most common type of cancer that is mostly found among women worldwide. Data mining, soft computing, image processing techniques, statistical and machine learning are mostly used for detection, prediction and diagnosis of breast cancer. This paper had discussed the nature of breast cancer data and various techniques applied for detection, prediction and prognosis. Recent many research are going on combining both macro and micro levels of information. Integration of gene expression signatures, clinical variables and patient related factors like co-morbidities, performance status, symptoms, psychological status, and quality of life will really provide a broader assessment of factors that are associated with cancer progression.

## Acknowledgement

This paper was funded by UGC – Minor Research Project under grant No. 4937/14(SERO/UGC) . The authors, therefore, acknowledge with thanks UGC, India.

## References

Abdelaal, Medhat Mohamed Ahmed, Muhamed Wael Farouq, Hala Abou Sena, and A. Salem.( 2010),"Using data mining for assessing diagnosis of breast cancer." *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference* on, pp. 11-17. IEEE.

Alexe, G., G. S. Dalgin, S. Ganesan, C. Delisi, and G. Bhanot. (2007), "Analysis of breast cancer progression using principal component analysis and clustering."Journal of biosciences 32, no. 1: 1027-1039.

Alieldin, Nelly H., Omnia M. Abo-Elazm, Dalia Bilal, Salem Eid Salem, Eman Gouda, Magda Elmongy, and Amal S. Ibrahim. (2014),"Age at diagnosis in women with non-metastatic breast cancer: Is it related to prognosis?." Journal of the Egyptian National Cancer Institute 26, no. 1: 23-30.

Alina Sultana, MihaiCiuc, Rodica Strungaru and Laura Florea., (2010),"A New Approach in Breast Image Registration", International Conference on Intelligent Computer Communication and Processing, 149-154.

Antonie, Maria-Luiza, Osmar R. Zaiane, and Alexandru Coman. (2001),"Application of Data Mining Techniques for Medical Image Classification." MDM/KDD 2001: 94-101.

Blows, Fiona M., Kristy E. Driver, Marjanka K. Schmidt, Annegien Broeks, Flora E. Van Leeuwen, Jelle Wesseling, Maggie C. Cheang et al. (2010),"Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies." *PLoS medicine* 7, no. 5: e1000279.

Cianfrocca, Mary, and Lori J. Goldstein. (2004),"Prognostic and predictive factors in early-stage breast cancer." *The oncologist* 9, no. 6: 606-616.

Chi, Chih-Lin, W. Nick Street, and William H. Wolberg. (2007) "Application of artificial neural network-based survival analysis on two breast cancer datasets." In *AMIA Annual Symposium Proceedings*, vol. 2007, p. 130. American Medical Informatics Association.

Choi, Jong Pill, Tae Hwa Han, and Rae Woong Park. (2009), "A hybrid Bayesian network model for predicting breast cancer prognosis." *Journal of Korean Society of Medical Informatics* 15, no. 1: 49-57.

Clavel-Chapelon, Françoise, and Mariette Gerber. (2002), "Reproductive factors and breast cancer risk. Do they differ according to age at diagnosis?." *Breast cancer research and treatment* 72, no. 2: 107-115.

Cruz, Joseph A., and David S. Wishart. (2006) "Applications of machine learning in cancer prediction and prognosis." *Cancer informatics* 2: 59.

Daemen, Anneleen, Olivier Gevaert, and Bart De Moor. , (2007), "Integration of clinical and microarray data with kernel methods." In Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE, pp. 5411-5415. IEEE.

Dursun Delen, Glenn Walker, Amit Kadam. (2005), "Predicting breast cancer survivability: a comparison of three data mining methods." *Artificial Intelligence Medicine*, 34 :113-27.

Dubitzky, Werner, Martin Granzow, and Daniel P. Berrar. (2007),*Fundamentals of data mining in genomics and proteomics*. Springer Science & Business Media.

Elkhodary, Tawfix R., Mohamed A. Ebrahim, Elsayed E. Hatata, Nermeen A. Niazy. (2014), "Prognostic value of lymph node ratio in node-positive breast cancer in Egyptial patients." *Journal of the Egyptian National Cancer Institute*, 26 :31-35.

Endo, Arihito, Takeo Shibata, and Hiroshi Tanaka. (2008), "Comparison of Seven Algorithms to Predict Breast Cancer Survival ( Special Issue - Contribution to 21 Century Intelligent Technologies and Bioinformatics)." *Biomedical fuzzy and human sciences: the official journal of the Biomedical Fuzzy Systems Association* 13, no. 2: 11-16.

Eshlaghy, Abbas Toloie, Ali Poorebrahimi, Mandana Ebrahimi, Amir R. Razavi, and Leila Ghasem Ahmad. (2013), "Using three machine learning techniques for predicting breast cancer recurrence." *J Health Med Inform* 4, no. 2: 124.

Finak, Greg, Nicholas Bertos, Francois Pepin, Svetlana Sadekova, Margarita Souleimanova, Hong Zhao, Haiying Chen et al. (2008), "Stromal gene expression predicts clinical outcome in breast cancer." *Nature medicine* 14, no. 5: 518-527.

Glare, Paul. "Clinical predictors of survival in advanced cancer. (2005)," J Support Oncol 3, no. 5: 331-339.

Hassanien, A.E., and Ali, J.M., (2004), "Enhanced Rough Sets Rule Reduction Algorithm for Classification Digital Mammography", Intelligent System journal, UK, Freund & Pettman, 13(2): 151-171.

Khan, Muhammad Umer, Jong Pill Choi, Hyunjung Shin, and Minkoo Kim. (2008),"Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare." In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pp. 5148-5151. IEEE.

Li, Lihua, Hong Tang, Zuobao Wu, Jianli Gong, Michael Gruidl, Jun Zou, Melvyn Tockman, and Robert A. Clark. (2004),"Data mining techniques for cancer detection using serum proteomic profiling." *Artificial intelligence in medicine* 32, no. 2: 71-83.

Longadge, Rushi, and Snehalata Dongre.(2013), "Class Imbalance Problem in Data Mining Review.". International Journal of Computer Science and Network (IJCSN) Volume 2 , Issue 1, February.

Malek Jihene, Abderrahim Sebri, Souhir Mabrouk, Kholdoun Torki, and Rached Tourki. (2009),"Automated breast cancer diagnosis based on GVF-snake segmentation, wavelet features extraction and fuzzy classification." Journal of Signal Processing Systems 55, no. 1-3: 49-66.

Mohanalin, J.,Kalra, P.K., Kumar, N., (2008), " Fuzzy based micro calcification segmentation", Electrical and Computer Engineering ICECE: 49-52.

Saar-Tsechansky, Maytal, and Foster Provost. (2007), "Handling missing values when applying classification models."

Selvaraj, Saravanakumar, and Jeyakumar Natarajan. (2011),"Microarray data analysis and mining tools." *Bioinformation* 6, no. 3 :95.

Shah, Naishil N., Tushar V. Ratanpara, and C. K. Bhensdadia. (2014), "Early Breast Cancer Tumor Detection on Mammogram Images." *International Journal of Computer Applications* 87, no. 14: 14-18.

Shleeg, Alshalaa A., and Issmail M. Ellabib. (2013),"Comparison of Mamdani and Sugeno Fuzzy Interference Systems for the Breast Cancer Risk." *International Journal of Computer, Information Science and Engineering* 7, no. 10 :387-391.

Sommer, Christoph, Luca Fiaschi, Fred A. Hamprecht, and D. W. Gerlich. (2012)"Learning-based mitotic cell detection in histopathological images." In Pattern Recognition (ICPR), 2012 21st International Conference on, pp. 2306-2309. IEEE.

Sotiriou, Christos, Pratyaksha Wirapati, Sherene Loi, Adrian Harris, Steve Fox, Johanna Smeds, Hans Nordgren et al. (2006), "Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis." *Journal of the National Cancer Institute 98*, no. 4: 262-272.

Srinivas, Pothur R., Sudhir Srivastava, Sam Hanash, and George L. Wright. (2001),"Proteomics in early detection of cancer." *Clinical Chemistry* 47, no. 10 :1901-1911.

Thangavel, K., Karnan, M., Siva Kumar, R., and KajaMohideen, A., (2005) "Segmentation and Classification of Micro calcification in Mammograms Using the Ant Colony System", International Journal on Artificial Intelligence and Machine Learning, 5, :29-40

Thongkam, Jaree, Guandong Xu, Yanchun Zhang, and Fuchun Huang. (2009), "Toward breast cancer survivability prediction models through improving training space." *Expert Systems with Applications* 36, no. 10 :12200-12209.

Van't Veer, Laura J., Hongyue Dai, Marc J. Van De Vijver, Yudong D. He, Augustinus AM Hart, Mao Mao, Hans L. Peterse et al. (2002),"Gene expression profiling predicts clinical outcome of breast cancer." nature 415, no. 6871 :530-536.

Wang, Yixin, Jan GM Klijn, Yi Zhang, Anieta M. Sieuwerts, Maxime P. Look, Fei Yang, Dmitri Talantov et al. "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer." *The Lancet* 365, no. 9460 (2005): 671-679.

Wang, Shuo, and Xin Yao. (2012), "Multiclass imbalance problems: Analysis and potential solutions." *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 42, no. 4 :1119-1130.