ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

A Review on Prediction Of Breast Cancer Using Various Data Mining Techniques.

M Deepika*, L Mary Gladence, and R Madhu Keerthana.

Department of Information Technology, Sathyabama University, Tamil Nadu, India.

ABSTRACT

In medical diagnosis, the prediction of a disease acts as an important core in analyzing the medical images. The unwanted cell growth in any part of the organ is known as tumor. The tumor may be benign or malignant. Malignant tumor is considered to be the most dangerous tissue. So, the early diagnosis of the disease helps to prevent the cancer. In women, breast cancer is treated as the most significant issue. There are various researchers studied about the prediction of breast cancer. This paper aims to review on various data mining techniques that are specifically considered on breast cancer prediction.

Keywords: Medical diagnosis, Tumor cells, Breast Cancer, Genetic algorithms and medical images.

**Corresponding author*

INTRODUCTION

Breast cancer is one of the most dangerous diseases in women. The research done in breast cancer proved that this disease to be top cancer killers amongst women. The origin of breast cancer is the malicious tissue formed from the milk ducts. 30% of the women are suffering from the breast cancer. As indicated by [2] in the year 2008, the affected rate of breast cancer is458, 503 passings throughout the world. In the year 2009, one million new cases were analyzed [3] [8] and in the year 2010 one and half million new cases were analyzed in women. In view of right on- time discovery and treatment breast cancer patients are still alive after six a long time from the diagnosis [3] if specialists have the potential to recognize it. In this way, an early expectation of this disease can lead in decrease of death rates. The study directed in [4] proposed that the survival rate of the following five year of conclusion is 88% and if there should be an occurrence of 10 years, it is 80%. Thusly, it is vital to recognize this life-debilitating sickness at the earlier stage to increase the survivability of breast cancer patients. One of the major clinical issues of breast cancer is the prediction of malignancy at the earlier stage. The risk factor of breast cancer in women is as follows [5] [6] [7].

- Being a woman
- Growth of age
- BRCA1 and BRCA2 breast cancer gene.
- Lobular carcinoma in situ (LCIS)
- History of breast or ovarian cancer
- A family history of breast, ovarian or prostate cancer
- High tissue density on a mammogram
- Higher biopsy displaying atypical hyperplasia
- Menopause stage
- No children
- Frequent exposure of x-rays
- First child after age 35
- High bone density
- Overweight in case of menopause
- Hormone level in postmenopausal

This paper is organized as: Section I describes the basic terms in breast cancer and its present view and status in the breast cancer. Section II portrays the various studies conducted by the experts in the aspects of four approaches namely, Decision Tree (DT), Artificial Neural Networks (ANN), Genetic Algorithms (GA) and Support Vector Machines (SVM). At last, the summary about the reviews in discussed in Section III.

LITERATURE SURVEY

Several researchers studied on predicting earlier way of breast cancer diagnosis. The learning systems is of data mining techniques, machine learning techniques and the hybrid form of data mining and machine learning systems [9]. The following algorithms are widely used in breast cancer prediction:

1. Decision Trees
2. Artificial Neural Network
3. Genetic algorithms
4. Support Vector Machines

Decision Trees

The best classification algorithms widely used in medical applications is the decision trees. It is in form of graph based systems [10]. The eminent decision trees algorithms were Quinlan’s, ID3, C4.5 and C5 [11]. The representation of a generic decision tree is given by:

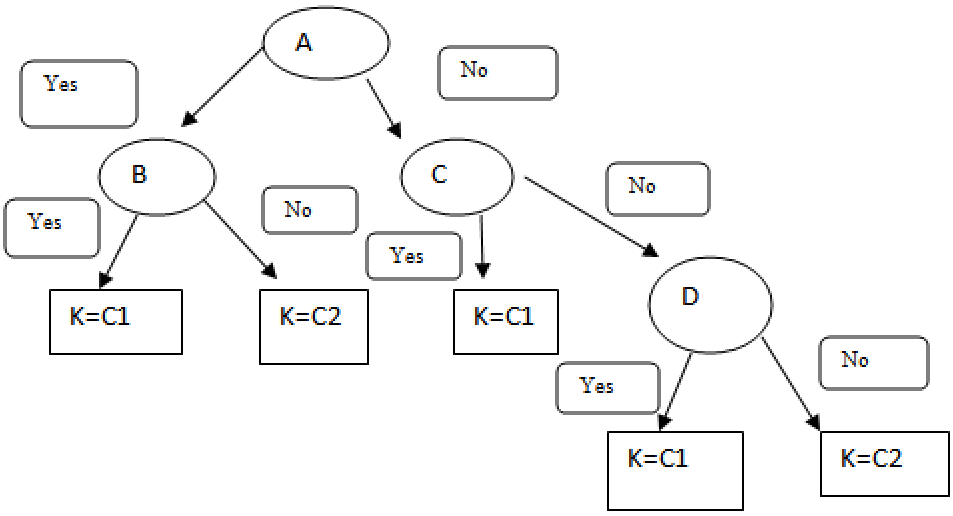

Fig 1: Representation of decision tree [1]

In [18] the researchers discovered three diverse data mining techniques for the forecast of breast malignancy survivability. They suggested decision tree as the best indicator with 99.6% approximation. They used an expansive dataset with 10-fold cross approval. Researchers utilized C5 learning system as the decision tree systems to show better results in terms of accuracy, sensitivity and specificity. K. Rzyszt of Fujarewicz et.al. [13] investigated that the utilization the Recursive Feature Selection (RFR) system for finding imperfect quality subsets for malignant cells classification. They discovered that RFR strategy has the capacity to discover the smallest subset for the prediction of breast cancer. Parvesh Kumar and SiriKrishan Wasan [15] analyzed on colon information set utilizing distinctive decision tree algorithms. They discovered the minimum rate of error which leads to easier breast cancer prediction. G.Sujatha et.al[16] studied on ID3,C4.5 and CART classifiers for better precision and execution time to build the tree. It is noticed that C4.5 performs well for tumor datasets, if accessible datasets are utilized as it may be. Among these three systems, C4.5 is the best one for improved information set of Primary tumor and for improved Colon tumor information set both ID3 and C4.5 exhibited the best accuracy. K. Rajesh et.al [21] endeavored to characterize SEER breast cancer information into the groupings of "Carcinoma in situ" and "Threatening potential" utilizing C4.5 systems. The prediction results showed that 94% accuracy in training phase and 93% accuracy in testing phase. In [20], 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.

Artificial Neural Networks

Artificial Neural Networks is the form of Neural Networks. A biological oriented network is formed to predict the breast cancer. The representation of the neural network is given as:

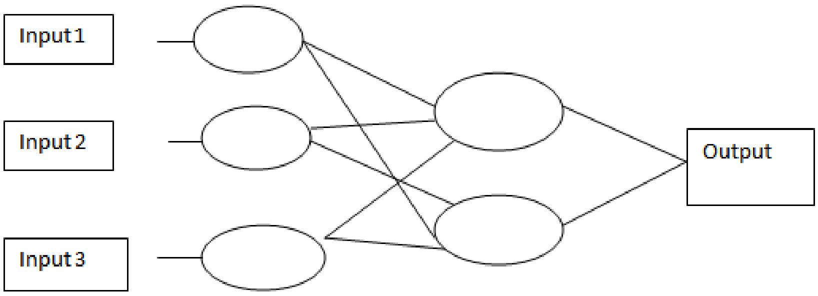

Fig 2: Representation of neural network [2]

ANN ought to be modified for every application else it leads to poor performance. The black box technology is the demerit discovered in ANN. Artificial Neural Networks (ANNs) has been utilized for the breast malignancy prediction in [21], [22]. In [21] the authors applied ANN on two distinctive breast malignancy dataset. Both of these datasets utilizes the morphometric attributes. An enhanced ANN model [22] has been utilized. Back propagation has been utilized to prepare the systems. Back propagation comprises of three layers a) Input layer b) hidden layer c) output layer. The predetermined model uses the likelihood nature and distinct the patients with the remarks good or bad. Several researchers utilized three-layer feed forwad ANNs with sigmoid function. In [1] used the ANN with multilayer perceptrons. It worked similar to the [21]. In any case, the distinction in the middle of [21] and [22] is the dataset. [22] Utilizes much greater dataset also, furnishes a decent investigation with the customary TNM neural framework. They guarantee that ANN can give vastly improved exactness when contrasted with the conventional TNM framework. ANN's accuracy achieved as high as 81%. It can even reach to 87% with the expansion of couple of more demographic variables.

Val'erie Bourd`es [17] et al., 2010 have presented the article in neural system with logistic regression. They framed Adaptive Reasoning Theory (ART) and compared with previous algorithms RBF, PNN and Multi-Layer Perceptron (MLP). They enhanced accuracy rate with lessened time. In this paper [18],neural system is extremely used for information mining in the aspect of better robustness, adaptive and fault tolerance. A reinforcement learning techniques has been proposed to predict the breast cancer. They reduced noisy in data with high accuracy. In [19], they discussed in non-linear statistical data modeling tools. They found a strong pattern between input and output which resolved higher rate of fault tolerance. In [20], the back propagation method is used as supervised neural network for breast cancer prediction using feed forward neural network. This study [21] portrayed the time series based neural networks. They studied on linear models such as Auto-Regression Moving Average (ARMA) and Auto- Regression (AR). Some parameters affect the performance of neural network such as quality of data preprocessing and neural network structure [21].

Genetic Algorithms

Genetic algorithms are dynamic heuristic techniques. A genetic fitness function is calculated to find the genetic approach. In [23] utilized the genetic systems for the forecast of breast tumor. This system is hybrid with the decision tree, ANN and logistic regression. They used 699 records acquired from the breast cancerous patients at the University of Wisconsin. They utilized 9 indicator variables and 1 result variable for the information investigation with 10-fold cross approval. The researchers asserted that their genetic prediction model gives precision as much as 99%. Lipo Wang, Feng Chu, et al., [22] proposed the cancer prediction using gene expression data. They found the minimum gene probability. Two approaches were proposed namely, gene selection and gene ranking scheme. Based on the ranking scores, the prediction of the malignant breast cancer is detected. They also employed T-test and class separablity. In [24], a semi supervised ellipsoid method was proposed to detect the multiclass cancer classification. Each attribute is labeled and similar labeled data are clustered to detect the disease. The examining methods include dartificial resampling of the information set. This can have the capacity to under examining the majority class[24],[25] and over examining the minority class[26],[27]or by consolidated the over and undersampling systems in an ordered way [28]. The class imbalance problem is solved by the binary feature selection. The mean between two classes with its threshold is estimated to solve the duplication of the data. The feature values are binarized before setting the threshold value [29]. Sometimes, there may be continuous feature selection in handling machine learning algorithms. These metrics method classified into Pearson correlation coefficient, feature assessment by sliding thresholds, FAIR and signal to noise correlation coefficient. These methods are designed to operate a continuous data and do not require any preprocessing of the data to work. All the experiments has conducted using MATLAB SPIDER package. Xiao et al distinguishing Differentially Expressed (DE) qualities from high-throughput gene expression estimations, by considering the parameters p- value and biological relevance [30]. A better gene ranking was framed using Gene Set Enrichment Analysis (GSEA).The gene expression information acquired through such advances can be valuable for some applications in bioinformatics, if legitimately examined. For example, they can be utilized to encourage gene prediction [31]. Several gene expression schemes do not provide the accuracy and efficiency of data mining systems [32]. One of the primary difficulties in characterizing gene expression information is that the number of gene is normally much higher than the quantity of analyzed examples. Likewise, it is not clear which qualities are critical and which can be overlooked without lessening the classification improvement. Numerous example characterization methods have been utilized to break down microarray information. The characterization of the recorded examples can be utilized to arrange diverse sorts of dangerous tissues as in [33], where distinctive sorts of leukemia are distinguished, or to recognize harmful tissue from ordinary tissue, as done in [34], where tumor and typical colon tissues are investigated.

Support Vector Machines

Support Vector Machines belongs to the class of supervised learning systems. It is one of the best optimization procedures. This reduces the over-flowing of the trained data. The goal is to find the optimized decision boundaries to predict the breast cancer at the earlier stage. C-Support Vector Classification Filter (C-SVCF) algorithms was used to spot and kill exceptions in breast malignancy survivability information sets. Consequences of their methodology demonstrated performance enhancement of breast tumor survivability expectation models by enhancing information quality. This script unmistakably draws consideration towards the utilization of SVM for anticipating better survival rates. [12] Gaussian bit nonlinear SVM is connected to focus Survival curves and decision on chemotherapy for new patients by allocating them to one of the three gatherings (Good Prognosis, Average Prognosis, Bad Prognosis) without the requirement for lymph node status. Xiaowei Song [35] utilized different machine learning methods also, acquired great survival rate. A least square SVM was proposed for the predictive analysis of breast cancer. They employed logistic regression using sigmoid function under ROC curve. H. Yusuff [36] proposed logistic relapse model for breast malignancy investigation, where he validated the mammogram tests. A different SVM techniques based on benign and malignant classification was framed. A hybrid SVM system was framed with the advent of linear and Radial Basis Function, logistic regression with regularization parameter and k- nearest neighbor. These systems provided a better prediction of breast cancer.

In [37] SVM was trained using linear, polynomial and radial basis function (RBF) kernels and applying PSO to each kernels for different datasets to get better accuracy for BC, Lung cancer and Heart diseases. In [39] a model that did a simultaneous optimization for SVM kernel parameters and feature subset without degrading the SVM classification accuracy was introduced. In [38] a study compared PSO based ANN, Adaptive Neuro Fuzzy Inference System (ANFIS) and a case-based reasoning (CBR) classifier with a logistic regression model and DT model. Support Vector Machine [35] utilized a nonlinear mapping to move the training data into the view of high dimensional spaces. This new dimension allowed searching for linear optimal hyper plane. The SVM discovered this hyper plane utilizing support vectors and edges. The Support Vector Machines (SVM) is a general class of learning architectures, propelled by the statistical hypothesis.

SUMMARY

Breast cancer is one of the dangerous cancerous diseases amongst women. The probability rate of breast cancer is increasing globally. An earlier detection of the tumor cells leads to decreased rate of mortality. Henceforth, an accurate and efficient prediction should be obtained. One major class of problems in medical science involves the diagnosis of disease, based upon different tests performed upon the patient. The most innovative and challenging task in medical applications is the predictive results of the disease. There are varieties of data mining techniques. Each method possesses its own merits and demerits. This paper intends to provide the reviews conducted by the various experts in the field of data mining systems. From the above study, we can infer that there is a still lack of early diagnosis, accuracy, sensitivity and specificity of the breast cancer data.

REFERENCES

[1] http://www.who.int/mediacentre/factsheets/fs297/en/

[2] American Cancer Society "Report sees 7.6 million global 2007 cancer deaths" Reuters.

[3] Marc E. Lippman, “Harrison's Principles of Internal Medicine”, 16th ed., Ch. 76, "Breast Cancer," by World Health Organization “Fact sheet No. 297: Cancer”.

[4] Umer Khan, Hyunjung Shin, Jong Pill Choi, Minkoo Kim, “wFDT - Weighted Fuzzy Decision Trees for Prognosis of Breast Cancer Survivability”, AusDM 2008.

[5] DursunDelen, Glenn Walker, AmitKadam, “Predicting breast cancer survivability: a comparison of three data mining methods”, Artificial Intelligence in Medicine, Volume 34, Issue 2, Pages 113-127, June 2005.

[6] Muhammad Umer Khan, Jong Pill Choi, Hyunjung Shin and Minkoo Kim, “Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare”, EMBS 2008.

[7] N. J. Bundred, “Prognostic and predictive factors in breast cancer”, Cancer Treatment Reviews, Vol. 27, Issue 3, Pages 137-142, June 2001.

[8] Yijun Sun, Steve Goodison, Jian Li, Li Liu and William Farmerie, “Improved breast cancer prognosis through the combination of clinical and genetic markers”, Bioinformatics 2007.

[9] Wei-Pin, Chang, Der-Ming, Liou “Comparison of Three Data Mining Techniques with Genetic Algorithm in the Analysis of Breast Cancer Data”, Journal of Telemedicine and Telecare, 2008, 9.

[10] J.R, QUINLAN, “Induction of Decision Trees”, Journal of Machine Learning, Volume 1, Number 1, March, 1986.

[11] David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams, “Learning representations by back-propagating errors”, Letter to Nature, 1986.

[12] Y.-J. Lee, O. L. Mangasarian, and W. H. Wolberg, “Breast Cancer Survival and Chemotherapy: A Support Vector Machine Analysis”, Data Mining Institute, Computer Sciences Department, University of Wisconsin, 2000.

[13] Krzysztof Fujarewicz, MalgorzataWiench, “Selecting differentially expressed genes for colon tumor classification” int.j.Appl.Math.Comput.Sci, 2003. Vol.3, No.3, Pg.no:327-335

[14] D. Lavanya, Dr.K.Usha Rani, “Performance Evaluation of Decision Tree Classifiers on Medical Datasets”,. International Journal of Computer Applications 26(4):1-4, July 2011

[15] Street W.N., “A Neural Network Model for Prognostic Prediction”, Fifteenth International Conference on Machine Learning, Madison, Wisconsin, Morgan Kaufmann, 1998.

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

[17] Val´erie Bourd`es, St´ephaneBonnevay, Paolo Lisboa, R´emyDefrance, David P´erol, Sylvie Chabaud, Thomas Bachelot, Th´er`ese Gargi,6 and Sylvie N´egrier “Comparison of Artificial Neural Network with Logistic regression as Classification Models for Variable Selection for Prediction of Breast Cancer Patient outcomes”

[18] I Guyon, J Weston, S Barnhill “Gene selection for cancer classification using support vector machines” .Machine learning, 2002 – Springer

[19] Z. Chen, “Research of Data Mining Based on Neural Network”, E-Product E-Service and Entertainment (ICEEE), 2010 International Conference on: IEEE, (2010), pp. 1-3.

[20] G. K. Dhondalay, C. Lemetre and G. R. Ball, “Modeling estrogen receptor pathways in breast cancer using an Artificial Neural Networks based inference approach”, Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on: IEEE, (2012), pp. 948-951.

[21] E. Ogasawara, L. Murta, G. Zimbrao and M. Mattoso, “Neural networks cartridges for data mining on time series”, Neural Networks, 2009.IJCNN 2009. International Joint Conference on: IEEE, (2009), pp. 2302-2309.

[22] Lipo Wang, Feng Chu, And Wei Xie, "Accurate Cancer Classification Using Expressions Of Very Few Genes", IEEE/ ACM Transactions On Computational Biology And Bioinformatics, 4, 40-52, 2007.

[23] RuiXu, Anagnostopoulos, G.C. And Wunsch, D.C.I.I.,"Multiclass Cancer Classification Using Semi supervised Ellipsoid ARTMAP and Particle Swarm Optimization with Gene Expression Data", IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol.4, No.1, Pp. 65-77, 2007.

[24] Li-Juan Zhang, Zhou-Jun Li and XiaoHua Hu “A Hybrid Gene Selection Method for Cancer Classification.

[25] M. Kubat and S. Matwin, “Addressing the Curse of Imbalanced Data Sets: One Sided Sampling,” Proc. 14th Int’l Conf. Machine Learning, pp. 179-186, 1997.

[26] X. Chen, B. Gerlach, and D. Casasent, “Pruning Support Vectors for Imbalanced Data Classification,” Proc. Int’l Joint Conf. Neural Networks, pp. 1883-1888, 2005.

[27] N. Chawla, K. Bowyer, L. Hall, and P. Kegelmeyer, “SMOTE: Synthetic Minority Over-Sampling Technique,” J. Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[28] M. Kubat and S. Matwin, “Learning When Negative Examples Abound,” Proc. Ninth European Conf. Machine Learning (ECML ’97),pp. 146-153, 1997.

[29] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” J. Machine Learning Research, vol. 3, pp. 1157-1182, 2003

[30] Xiao, Yufei, Tzu-Hung Hsiao, Uthra Suresh, Hung-I. Harry Chen, Xiaowu Wu, Steven E. Wolf, and Yidong Chen. "A novel significance score for gene selection and ranking." Bioinformatics 30, no. 6 (2014): 801-807

[31] A. Zhang, “Advanced Analysis of Gene Expression Microarray Data”. Singapore: World Scientific, 2006.

[32] F. Azuaje, W. Dubitzky, N. Black, and K. Adamson, “Discovering Relevance Knowledge in Data: A Growing Cell Structure Approach”, lEEE Transactions on Systems. Man and Cybernetics, vol. 30, issue 3, June, 2000.

[33] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, “Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring”, Science, vol. 286, pp. 531–537, 1999.

[34] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays”, In: Proceedings of. National.Academy.Science. USA, vol. 96, pp. 6745–6750, 1999

[35] XiaoweiSonga, Arnold Mitnitskib,c, JafnaCoxb, Kenneth Rockwood - Comparison of Machine Learning Techniques with Classical Statistical Models in Predicting Health Outcomes, MEDINFO 2004 M. Fieschi et al. (Eds) Amsterdam: IOS Press © 2004 IMIA

[36] H. Yusuff, N. Mohamad, U.K. Ngah& A.S. Yahaya – Breast Cancer Analysis Using Logistic Regression, www.arpapress.com/Volumes/Vol10Issue1/IJRRAS_10_1_ 02.pdf.

[37] SmrutiRekha Das, Pradeepta Kumar Panigrahi, Kaberi Das, and Debahuti Mishra.Improving RBF kernel function of support vector machine using particle swarm optimization.International Journal, 2012.

[38] Mei-Ling Huang, Yung-Hsiang Hung, Wen-Ming Lee, RK Li, and Tzu-Hao Wang. Usage of case-based reasoning, neural network and adaptive neuro-fuzzy inference system classification techniques in breast cancer dataset classification diagnosis. Journal of medical systems, 36(2):407-414, 2012.

[39] Monalisa Mandal, AnirbanMukhopadhyay, and UjjwalMaulik. Fuzzy rule-based classifier for microarray gene expression data by using a multiobjective PSO-based approach. In Fuzzy Systems (FUZZ), 2013 IEEE International Conference on, pages 1-7. IEEE, 2013