International Journal of Engineering and Management Research e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume- 9, Issue- 3 (June 2019)
www.ijemr.net https://doi.org/10.31033/ijemr.9.3.06

A Review on Data Mining Techniques for Prediction of Breast Cancer Recurrence

R.S.PadmaPriya[1] and P.Senthil Vadivu[2]
[1]Assistant Professor, Department of Computer Technology, Dr.N.G.P Arts and Science College, Coimbatore, INDIA
[2]Head of Department, Department of Computer Applications, Hindusthan College of Arts and Science, Coimbatore, INDIA

[1]Corresponding Author: priypadma@gmail.com

ABSTRACT

The most common type of cancer in women worldwide is the Breast Cancer. Breast cancer may be detected early using Mammograms, probably before it's spread. Recurrent breast cancer could occur months or years after initial treatment. The cancer could return within the same place because the original cancer (local recurrence), or it may spread to different areas of your body (distant recurrence). Early stage treatment is done not only to cure breast cancer however additionally facilitate in preventing its repetition/recurrence. Data mining algorithms provide assistance in predicting the early-stage breast cancer that continually has been difficult analysis drawback. The projected analysis can establish the most effective algorithm that predicts the recurrence of the breast cancer and improve the accuracy the algorithms. Large information like Clump, Classification, Association Rules, Prediction and Neural Networks, Decision Trees can be analyzed using data mining applications and techniques.

Keywords— Association Rules, Classification, Decision Trees, Prediction, Recurrence

I. INTRODUCTION

Data mining is playing a great role in computing applications in the domain area of medicine. Availability of data mining applications and its techniques are shown in the areas of healthcare administrations, patient care, management, and intensive care system. One of the new researches in data mining application involves analyzing Breast cancer, which is one of the deadliest disease and most common of all cancers in the leading cause of cancer deaths in women worldwide. Classification

Algorithms plays an important role in Data Mining research. Early stage treatment of this disease can give full recovery from this disease. There are several data mining applications and techniques which are used to analyze huge data, those data mining techniques are Clustering, Classification, Association Rules, Prediction and Neural Networks Decisions Trees. Among these, some classification algorithms such as Naïve Bayes (NB), Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Tree (C 5.0) and K nearest neighbor (KNN) algorithm are used to obtain most accurate result.

Data mining is currently solving a lot of real world problems. Because the main use of data mining technique is to change raw data into more meaningful information. Both male and female have chances of getting affected by breast cancer. But from the world breast cancer statics the occurrence of this disease is higher in female than males. Recurrence of breast cancer is when the cancer is come back after treated.

II. BREAST CANCER: AN OVERVIEW

The initial stage of Breast cancer begins when cells in the breast begin to grow out of control. A tumor is formed by these cells that can be seen on X-ray or it can be felt as a lump. Breast cancer not only occurs entirely in women, but men can also have the chances of affected by breast cancer.

The common symptom of breast cancer is a new swelling or accumulation. A hard mass that has irregular edges without any pain is more likely to be cancer, but breast cancers can be tender, soft, or rounded. They can even be painful.

Breast cancer is a malignant or benign tumor, inside breast, wherein cells divide and grow without control. As there are a few risk factors which increase the like hood of a woman developing breast cancer, the researchers tried to find the exact reason behind breast cancer. 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. This is indicated as a form of tumor from Stage 0 (the least advanced stage) to Stage IV (the most advanced stage).

III. FACTORS FOR BREAST CANCER IDENTIFICATION

Factors that can be used to identify breast cancer at the earlier stage are as under:

1) Age
2) Alcohol Intake

3) Life Style
4) Heredity
5) Weight
6) Hereditary
7) Child Birth

Factors that can be used for the breast cancer prediction are as follows:

Input	Description
Clump Thickness	Assesses if cells are mono- or multi-layered.
Cell Size	Evaluates the consistency in size of the cells in the sample.
Cell Shape	Estimates the equality of cell shapes and identifies marginal variances.
Marginal Adhesion	Quantifies how much cells on the outside of the epithelial tend to stick together.
Single Epithelial Cell Size	Relates to cell uniformity, determines if epithelial cells are significantly enlarged.
Bare Nuclei	Calculates the proportion of the number of cells not surrounded by cytoplasm to those that are.
Bland Chromatin	Rates the uniform "texture" of the nucleus in a range from fine to coarse.
Normal Nucleoli	Determines whether the nucleoli are small and barely visible or larger, more visible, and more plentiful.
Mitoses	Describes the level of mitotic (cell reproduction) activity.

IV. LITERATURE REVIEW

Shelly Gupta et al[1], observed that the accuracy for the diagnosis analysis of various applied data mining classification techniques is highly acceptable and can help the medical professionals in decision making for early diagnosis and to avoid biopsy. The problem is mainly analyzed under ANNs that result in high accuracy when compared to other classification techniques applied for the same. The authors say that the best model can be obtained after building several different types of models, or by trying different technologies and algorithms.

Ahmad. LG et al[2] studied the application of data mining techniques to develop predictive models for breast cancer recurrence in patients. They have used the machine learning techniques, such as Decision Tree (C4.5), Support Vector Machine (SVM), and Artificial Neural Network (ANN) to develop the predictive models. The main goal of this paper is to analyze the performance of these three algorithms on the data through sensitivity, specificity, and accuracy. The final outcome identified that SVM classification model predicts breast cancer recurrence with least error rate and highest accuracy. The 10-fold cross-validation approach was used for measuring the unbiased prediction accuracy of each model to obtain the results.

Chintan. Shah and Anjali. G. Jivani[3] has performed a comparative analysis between three algorithms with help of WEKA (The Waikato Environment for Knowledge Analysis), which is an open source software. It contains different type's data mining algorithms namely Decision tree, Bayesian Network and K-Nearest Neighbor algorithms. Here, for comparing the result, they used the correctly classified instances, incorrectly classified instances, time taken, kappa statistic, relative absolute error, and root relative squared error as parameters.

B.Padmapriya and T.Velmurugan[4] proposed the classification algorithms ID3 and C4.5 to identify the various categories of breast cancer. The behavior and performance of both the algorithms were analyzed through its experimental results. From their point of view, this research work concluded that the performance of C4.5 is better than the other algorithms.

Murari Kumar, Shivkumar Singh and Tomar Bhupesh Gaur **[5]** The authors have proposed a framework with optimized mining that can be a better way in breast cancer detection. There are several research directions in this field and the scope is in the direction to fulfill it in the earlier stage. The cancer symptoms are different and so it is treated differently also o that the chances is more positive. So the direction of research focuses on the identification of cancer in the early stages.

B. R. A. Cirkovic, A. M. Cvetkovic, S. M. Ninkovic, and D. Nenad [6] described the practical application of data mining methods for estimation of survival rate and disease relapse for breast cancer patients. A comparative study of prominent machine learning models was carried out and according to the achieved results it was concluded that the classifiers obviously learn some of the concepts of breast cancer survivability and recurrence. These algorithms were successfully applied to a novel breast cancer data set of the Clinical Center of Kragujevac.

G. D. Rashmi et.al,[7] used Naïve Bayes algorithm for classification and prediction to analyze whether the tumor is either benign or malignant. The dataset used is taken from the Wisconsin University database. Data sets are used to find the success rate and the error rate. The data sets were chosen randomly. Finally Naïve Bayes Classification gives 85-95% accuracy.

A. I. Pritom [8] applied Naive Bayes, C4.5 Decision Tree and Support Vector Machine (SVM) classification algorithms to calculate the prediction accuracy of breast cancer recurrence. Data were collected from Wisconsin dataset of UCI machine learning Repository with total 35 attributes. The Accuracy of each model can be improved by reducing some lower ranked

attributes with the help of an efficient feature selection algorithm.

The research paper done by Joana Diz Goreti Marreiros & Alberto Freitas[9] presents new computer based diagnosis system. By using this technique false positive diagnosis test can be reduced. After data sets analyzed Naive Bayes algorithm come with higher accuracy than Random forest.

Two various data sets from Wisconsin breast cancer have been evaluated by different data mining algorithms. The outcome that Rotation Forest model shows the highest classification accuracy (99.48 %) and when compared with the previous works, the new approach and methodology have come with highest performance and accuracy [10].

Jimin Guo, Benjamin C. M. Fung, Farkhund Iqbal [11] implemented decision tree algorithm with breast cancer data sets collected from Leiden University Medical Center. The data sets have 574 patients who have got surgery at that hospital. So they generate the recurrence of breast cancer by a decision tree algorithm within three years of initial diagnosis. The classifier predicted 70% accuracy. For the independent classifier of 65 patients the classifier exactly predicts the recurrence of the disease in 55 patients. The classifier also separates patient into two based on their disease characteristic and their relevance of early relapse.

Rohit J Kate and Ramya Nadig[12] used different algorithm in order predict the breast cancer solvability. The evaluation was done based the stage of the breast cancer. Three machine learning algorithms were applied in order to predict breast cancer survivability. These data sets were evaluated by classification algorithms such as Naïve Bayes, Logistic Regression and Decision Tree to predict breast cancer survivability.

M.R.Mohebian[13] et al., has defined breast cancer recurrence prediction with clinicopathologic characteristics of 579 breast cancer patients (recurrence prevalence of 19.3%) Datas were analyzed and discriminative features were selected using statistical feature selection methods. They were additionally distinguished by Particle Swarm Optimization (PSO) as the inputs of the classification system with ensemble learning (Bagged Decision Tree: BDT). The PSO algorithm was used to identify the combination of selected categorical features and also the weight (importance) of the selected interval-measurement-scale features. The performance of HPBCR (hybrid predictor of breast cancer recurrence) was assessed as a final outcome.

Akinsola Adeniyi F and et al, used Breast cancer data set derived from Federal government hospital Lagos. The authors used three supervised learning algorithms: Multilayer Perceptron, Naïve Bayes and C 4.5 which were executed using WEKA tool. Out of three algorithms the authors concluded that C 4.5 produced more accuracy of 93.9854% with an execution time of 0.28 seconds.

Hamid Karim Khani Zan[16] provided a study of different technical and review papers on breast cancer diagnosis and prognosis problems, outlined and resolved the issues, algorithms, and techniques for the breast cancer survivability prediction in SEER database.

Siyabend Turgut, Mustafa Datekin and Tolga Ensari[17] used microarray breast cancer data for classification of the patients using machine learning methods. First, 8 different machine learning algorithms were applied to the data, with two different feature selection methods. The methods applied are SVM, KNN, MLP, Decision Trees, Random Forest, Logistic Regression, Adaboost and Gradient Boosting Machines. After applying the two different feature selection methods with the best 50 features are applied, SVM gave the best results. MLP is applied using different number of layers and neurons to examine the effect of the number of layers and neurons on the classification accuracy. It is determined that the increase in the number of layers sometimes decreased, sometimes didn't change the accuracy.

V. PROPOSED RESEARCH

Breast cancer starts to grow in the human body when cells in the breast are growing most in an unexpected manner. After these cells grow, it can be seen by x- ray. Basically, there are two types of breast cancer, cancer that spread into another area and cancer that can't spread into another area. Among the world women breast cancer is the first and the most leading of death of women and the accurate diagnosis have lots of advantage to prevent and detection of the disease. Data mining is a technique can support doctors in the decision making process. As breast cancer recurrence is high, good diagnosis is important. This research is going to be implemented by different data mining algorithms like Bayes net, support vector machine and Decision tree (j48). So to get a more accurate value about the recurrence of breast cancer we are going to use data sets which were taken from the UCI machine learning repository.

Table1. Review of Algorithm & Approaches

Title	Methodology	Results
Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis [1]	Artificial Neural Network Bayesian Network	ANN has highest Accuracy
Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence [2]	Decision Tree, SVM, ANN	SVM classification model predicts breast cancer recurrence with least error rate

e-ISSN: 2250-0758 | p-ISSN: 2394-6962
Volume- 9, Issue- 3 (June 2019)
https://doi.org/10.31033/ijemr.9.3.06

		and highest accuracy
Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction [3]	KNN, Decision Tree, Naïve Bayes	Naïve Bayes gives high accuracy with low execution time.
A survey on breast cancer analysis using data mining techniques[4]	Decision Tree C 4.5 , ID3	Performance of C4.5 is better than the other algorithms
Mining based Optimization for Breast Cancer Analysis [5]	SVM, Neural Network, Fuzzy Theory, Genetic Algorithm	In future a framework with optimized mining can be a better way in breast cancer detection.
Prediction Models for Estimation of Survival Rate and Relapse for Breast Cancer Patients.[6]	SVM, Decision Tree	The classifiers obviously learn some of the concepts of breast cancer survivability and recurrence
Analysis of Efficiency of Classification and Prediction Algorithms (Naïve Bayes) for Breast Cancer Data set [7]	Naïve Bayes classification algorithm and Naïve Bayes prediction algorithms	The Naïve Bayes classification algorithm has highest accuracy value 89 -95%
Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique[8]	Decision tree C 4.5 , Support vector machine, Naïve Bayes algorithms	Support vector machine provide better performance before and after attribute selection.
Applying Data Mining Techniques to Improve Breast Cancer Diagnosis[9]	Naïve Bayes, Random Forest	Naive Bayes algorithm come with higher accuracy than Random forest.
Breast cancer diagnosis using GA feature selection and Rotation Forest[10]	Rotation forest model	Rotation forest shows highest classification accuracy (99.48%)
Revealing determinant factors for early breast cancer recurrence by a decision tree [11]	Decision tree	The classifier predicts for whether a patient developed early disease recurrence; and is estimated to be about 70% accurate
Stage-specific predictive models for breast cancer survivability[12]	Decision tree,(DT) Naïve Bayes , Logistic regression	Decision Tree Results in high accuracy
Using three machine learning techniques for predicting breast cancer recurrence [9]	Decision tree, Artificial neural network, Support vector machine	DT =0.936, ANN =0.947, SVM = 0.957. so that SVM have higher accuracy than other algorithms
Hybrid computer aided diagnosis system for prediction of breast cancer recurrence using optimized ensemble learning [13]	SVM Decision tree multilayer perception (MLP)	Support vector machine give a more accurate value
Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction [11]	K Nearest Neighbor algorithm Decision Tree, and Naïve Bayes algorithm	Better performance was provided by naïve Bayes algorithm 95.9%

VI. CONCLUSION

As discussed in this survey breast cancer recurrence is the most challenge of researchers for a lot of years. The actual cause of breast cancer is unknown, but early treatment may be a better way for prevention and detection of breast cancer. In the above survey/review, it is concluded that the data mining algorithms like Decision Tree, Support Vector Machine, and Naïve Bayes are giving more accurate results.

REFERENCES

[1] Shelly Gupta, Dharminder Kumar, & Anand Sharma. (2011). Data mining classification techniques applied for

breast cancer diagnosis and prognosis. *Indian Journal of Computer Science and Engineering*, *2*(2), 188-195.
[2] Ahmad. LG & Eshlagy. AT. (2013). Using three machine learning techniques for predicting breast cancer recurrence. *Journal of Health and Medical Informatics*, *4*(2). Available at: https://www.researchgate.net/publication/259583297_using_three_machine_learning_techniques_for_predicting_breast_cancer.
[3] Chintan. Shah & Anjali. G. Jivani. (2013). Comparison of data mining classification algorithms for breast cancer prediction. *Fourth International Conference on Computing, Communications and Network Technologies*, *1*, 1-4
[4] B.Padmapriya & T.Velmurugan. (2014). A survey on breast cancer analysis using data mining techniques. *IEEE International Conference on Computational Intelligence and Computing Research*. Available at: https://ieeexplore.ieee.org/document/7238530.
[5] Murari Kumar, Shivkumar Singh Tomar, & Bhupesh Gaur. (2015). Mining based optimization for breast cancer analysis. *International Journal of Computer Applications*, *119*(13), 1-6.
[6] B. R. A. Cirkovic, A. M. Cvetkovic, S. M. Ninkovic, & D. Nenad. (2015). Prediction models for estimation of survival rate and relapse for breast cancer patients. *IEEE International Conference on Bioinformatics and Bioengineering*. DOI:10.1109/BIBE.2015.7367658
[7] G. D. Rashmi, A. Lekha, & N. Bawane. (2015). Analysis of efficiency of classification and prediction algorithms (Naïve Bayes) for breast cancer dataset. *International Conference on Emerging Research in Electronics, Computer Science and Technology*, pp. 310–314.
[8] A. I. Pritom. (2016). *Predicting breast cancer recurrence using effective classification and feature selection technique*. Available at: https://ijrier.com/published-papers/volume-3/issue-5/predicting-breast-cancer-recurrence-using-effective-classification-and-feature-selection-technique.pdf..
[9] J. Diz, G. Marreiros & A. Freitas. (2016). Applying data mining techniques to improve breast cancer diagnosis. *Journal of Medical Systems*, *40*(9). Available at: https://link.springer.com/journal/10916.
[10] E. Aličković & A. Subasi. (2017). Breast cancer diagnosis using GA feature selection and Rotation Forest. *Neural Computing and Applications*, *28*(4), 753–763.
[11] J. Guo *et al.* (2017). Revealing determinant factors for early breast cancer recurrence by decision tree. *Information Systems Frontiers*, *19*(6), DOI: https://doi.org/10.1007/s10796-017-9764-0.
[12] R. J. Kate & R. Nadig. (2017). Stage-specific predictive models for breast cancer survivability. *International Journal of Medical Informatics*, *97*, 304–311.
[13] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. A. Mañanas, & F. Mokarian. (2017). A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (HPBCR) using optimized ensemble learning. *Computational and Structural Biotechnology Journal*, *15*, 75–85.
[14] Akinsola Adeniyi F, Sokunbi M.A, Okikiola F.M, & Onadokun I.O. (2017). Data mining for breast cancer classification. *International Journal of Engineering and Computer Science*, *6*(8), 22250-22258.
[15] Nisreen I.R. Yassin, Shaimaa Omran, Enas M.F. El Houby, & Hemat Allam. (2018). *Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review*. Available at: https://europepmc.org/abstract/med/29428074.
[16] Hamid Karim Khani Zand. (2015). A comparitive survey on data mining techniques for breast cancer diagnosis and prediction. *Indian Journal of Fundamental and Applied Life Sciences*, *5*, 4330-4339.
[17] Siyabend Turgut, Mustafa Datekin, & Tolga Ensari. (2018). Microarray breast cancer data classification using machine learning methods. *IEEE Transactions on Computational Biology*. Available at: https://www.semanticscholar.org/paper/Microarray-breast-cancer-data-classification-using-Turgut-Da%C4%9Ftekin/160fd6477a64ce89a65d74485b83cbe8a92efcf5.
[18] Meriem Amrane, Saliha Oukid Ikram, & Gagaoua Tolga Ensar. (2018). Breast cancer classification using machine learning. *IEEE Transactions on Computational Biology*, 1-4.
[19] Dona Sara Jacob, Rakhi Viswan, V Manju, L Padma Suresh, & Shine Raj. (2018). A survey on breast cancer prediction using data mining techniques. *IEEE Conference on Emerging Devices and Smart Systems*. Available at: http://toc.proceedings.com/41848webtoc.pdf.
[20] Sapiah Sakri, Nuraini Abdul Rashid, & Zuhaira Muhammad Zain. (2017). Particle swarm optimization feature selection for breast cancer recurrence prediction. *IEEE Access*, pp 1-1. 10.1109/ACCESS.2018.2843443.
[21] Amrita Sanjay, H Vinayak Nair, Sruthy Murali, & Krishnaveni K S. (2018). A data mining model to predict breast cancer using improved feature selection method on real time data. *International Conference on Advances in Computing, Communication and Informatics*. Available at: https://ieeexplore.ieee.org/document/8554450.