Review on Breast Cancer Prediction Using Data Mining Techniques and Weka Tool

JyotiNegi[1], Sunil Kumar[2], Dr. K.L. Bansal[3]
[1, 2, 3] Department of Computer Science
[1, 2, 3] HPU

Abstract- *Nowadays Breast cancer has become a common disease among women and also the biggest reason for the high number of death among women. Early prediction of breast cancer and recurrence of breast cancer prediction will help breast cancer patients to survive. Data mining has been widely used for predicting the diagnosis. The aim of this research is to review the role of data mining techniques in the prediction of breast cancer using the WEKA tool. Various studies concentrated on the prediction of benign and malignant and on breast cancer recurrence or non-recurrence. Various studies compared different classification algorithms for the prediction of breast cancer such as Decision tree, Naïve Bayes, KNN, and SVM.*

Keywords- breast cancer, machine learning, data mining, classification algorithms, Weka.

I. INTRODUCTION

Breast cancer is one of the most common cancer in women and also, it is the second main cause of cancer death in women [1, 2]. Breast cancer occurs from a harmful tumor when the growth of cells becomes uncontrollable. Breast cancer can happen at any time in women and also it comes back between three to five years (recurrence) after the treatment. However, the recovery probability is quite low, as soon as the recurrence happens. Breast cancer can be classified into two parts: Malignant (more harmful) and Benign (less harmful) [3]. Before cancer circulates to all parts of the surface, early detection of breast cancer is very important. There are various factors related to an increased number of breast cancer seen in women like increasing women's age, inherited genes, obesity, and many more. Therefore it is important to predict cancer in the early stage. Data mining and machine learning have been widely used in the diagnosis and prognosis of breast cancer such as classification, clustering, and regression. Data mining is the process of extracting relevant knowledge from a large collection of data. There are various classification techniques were used as a solution for predicting the diagnosis.

This paper is organized as follows, Section II introduces the various data mining algorithms, Section III summarizes recent related work or literature survey and Section IV is a conclusion of the research.

II. DATA MINING ALGORITHMS

Data mining and Machine learning techniques such as classification, clustering and regression helps us in breast cancer prediction. There are numerous algorithms such as Naïve Bayes, Decision Trees, KNN, Support vector machine, Random forest, Zero R, and One R used for analyzing data. Some important algorithms are discussed here

DECISION TREE ALGORITHMS (J48)

A decision tree is a machine learning classification technique and it is a supervised learning method. A decision tree is a tree structure having nodes where nodes act as the root node, branch, and leaf node [4] each internal node indicates a test on a target attribute and each leaf node indicates the value of the target attribute or represents the class. Decision tree are commonly used in data mining to examine data and induce the tree and its rules that will be used to make predictions. The decision tree starting from the root node of the tree also split into various nodes and move through until it reached the leaf node. It split the datasets into various subsets until low-level splits are reached. A decision tree helps us in decision-making purpose. There are various types of decision tree algorithms such as C4.5, ID3, and J48. J48 is a famous Decision tree algorithm and it is an extension of ID3.

K-NEAREST NEIGHBOR (KNN)

KNN is a supervised learning technique. It classifies new data based on the similarity to the available data and also based on the majority of the number of nearest neighbors. In WEKA, the KNN classifier is called IBK. It is mostly used for solving classification problems. It is a Lazy learning model because during the training phase, it doesn’t learn from training data but learns during the testing phase.

Naïve Bayes (NB)

It is based on the Bayesian theorem. This classification technique analyses the relationship between each attribute and the class for each instance to derive a conditional probability for the relationships between the attribute values and the class.

It is a statistical and probabilistic classification algorithm. It predicts class membership probabilities based on whose values we determine to which class a sample belongs.

Bayes theorem formula

$$P(A/B) = \frac{P(B/A)*P(A)}{P(B)}$$

SUPPORT VECTOR MACHINE (SVM)

SVM is a supervised learning technique that sorts the data into two groups. SVM classifies data into two classes over a hyperplane at the same time avoiding over-fitting the data by maximizing the margin of hyperplane separation [5, 6].

RANDOM FOREST

Random forest is a supervised machine learning technique and one of the ensemble methods which combines multiple classifiers into one and generates a single result that enhances the performance accuracy of the model. It includes a decision tree algorithm and is used for classification and regression problems.

III. LITERATURE SURVEY

The literature review has been done by studying various research papers, fifteen research articles have been studied to predict breast cancer and breast cancer recurrence. The summaries of them are below.

Ravi Kumar et al. [7] used Wisconsin Breast Cancer (WBC) dataset with 499 training sets and a test set with 200 patients. They compare six classification techniques (Decision tree, Neural Network, Naïve Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbor) using the Weka data mining tool. The result shows that the Support Vector Machine (SVM) classifier is more suitable in handling the classification problem of breast cancer prediction with an accuracy of 97.59%.

Jacob et al. [8] presented a Survey on Breast Cancer Prediction Using Data Mining Techniques. The author performs a comparison of diverse classification and clustering algorithms. The outcome shows that classification perform well as compared to the clustering algorithm.

Ahmad LG et al. [9] used Iranian Center for Breast Cancer (ICBC) data set which contained 1189 records, 22 predictor variables, and one outcome variable. They compare the performance of machine learning algorithms i.e. Decision Tree (C4.5), Support Vector Machine (SVM), and Artificial Neural Network (ANN) with accuracy, sensitivity, and specificity. The results show that the accuracy of SVM is 0.957 therefore SVM predicts breast cancer recurrence with the highest accuracy. The accuracy of the Decision Tree is the lowest 0.936.

Chaurasia et al. [10] The Wisconsin Breast Cancer Original (WBCO) data set from UCI Repository was used. Three classification techniques DT, KNN, and SMO were used and the experiment was performed on the data mining WEKA software tool. The SMO classifier outperformed with an accuracy of 96.2%.

K.Sivakami [11] used Wisconsin Breast Cancer Dataset (WBCD) from the UCI Repository which contains 699 instances, 458 benign classes, and 241 malignant classes. The main aim was to predict the model for breast cancer using the data mining technique. They compared three classification techniques using the Weka tool and the result shows that the Decision Tree-Support Vector Machine with the highest accuracy than IBL, SMO, and Naïve Bayes.

Ahmed Iqbal Pritom et al. [12] used Wisconsin Breast Cancer Dataset (WBCD) from UCI Repository which contained 35 attributes. They compare three classification techniques (Decision Tree, Naïve Bayes, and SVM). The result shows that the Support Vector Machine (SVM) classifier has the highest accuracy.

Kumar et al. [13] compared performances of DT, NB, and SVM using the voting techniques. WEKA data mining tool was used on Wisconsin dataset with 699 instances and 12 attributes. For predicting breast cancer, the best approach is a combination of the three algorithms using the voting approach.

Banu et al. [14] used the UCI breast cancer dataset with 286 instances and 10 attributes. Various classifiers J48, One R, Zero R algorithm, and decision stump were used to classify the data, and the Weka tool was used for the experiment purpose. The J48 algorithm produced better performance than all the other classification methods with an accuracy of 75.52%.

Ojha et al. [15] Compared four clustering and four classification algorithms on WPBC Wisconsin Prognostic Breast Cancer (WPBC) dataset from the UCI machine learning repository to predict the recurrence or non-recurrence of breast cancer. Their result demonstrates that classification algorithm decision tree (C5.0) and SVM produce 81% accuracy and classification algorithms is the best predictor than clustering algorithms.

Verma et al. [16] used the breast cancer and diabetes dataset from the UCI machine learning repository and discussed five classification algorithms on the Weka tool. The result shows that the naïve Bayes gives 72.70% accuracy on the breast cancer dataset and SMO gives 76.80% accuracy on the diabetes dataset.

Bharati et al. [17] presented Breast Cancer Prediction Using Different Classification Algorithm with Comparative Analysis Using Weka. The author discussed the classification algorithms Naïve Bayes, Random Forest, Logistic Regression, Multilayer Perceptron, and K nearest neighbors. The data set has been explored in terms of Kappa Statistics, TP rate, FP Rate, and precision. The result shows the KNN has the highest accuracy as compared to others.

Singh et al. [18] used Wisconsin Breast Cancer (original) dataset taken from UCI Machine learning Repository. The author compared the Decision tree classifier (J4.8, Simple CART), Bayes classifier (Naïve Bayes, Bayesian Logistic Regression). The experimental result shows that the decision tree classifier i.e. Simple CART (98.13%) gives higher accuracy.

Chaurasia et al. [19] used Wisconsin breast cancer data set. The dataset contained 699 instances, two classes (malignant and benign). Naive Bayes, RBF Network, and J48 are the three most popular data mining algorithms that were used to develop the prediction models and the experiment was done on the Weka tool. Results showed that the Naive Bayes performed best with an accuracy of 97.36%.

Shaikh et al. [20] used two cancer datasets and also used the four data mining classifiers, J48, k-NN, Naive Bayes, and SVM. The experiment was done on Weka and on MATLAB.

Nour A. AbouElNadar et al. [21] used the UCI repository dataset and used Classification techniques to classify whether breast cancer is recurrent or non-recurrent. K- Nearest Neighbor (KNN), Decision Trees (DT), Naïve Bayes (NB), Support Vector Machines (SVM), and ensemble techniques Bagging, Voting and Random Forest (RF) were compared using the WEKA tool for the experiment purpose. The result comes into two phases, in the first phase the Random Forest classifier gives the best results with 84.3% accuracy and in the second phase, the voting ensemble classifier gives the best results with 89.9% accuracy.

eltalhi et al. [22] presented a review on Breast cancer diagnosis and prediction using machine learning and data mining techniques. Breast cancer is one of the causes of death among women and early prediction of breast cancer will help breast cancer patients to survive. Data mining and machine learning have been widely used in the early detection of breast cancer. Studies compared various classification algorithms for the prediction of breast cancer such as Decision tree, Naïve Bayes, and Artificial Neural Network.

Ref.	Year	Technique	Accuracy
[7]	2013	DT, SVM, ANN, NB, LR, KNN	SVM-97.5%
[10]	2014	DT, SMO, KNN	SMO-96.2%
[11]	2015	DT,SVM,SMO,IBL,NB	DT+SVM (91%)
[12]	2016	DT, NB, SVM	SVM-75.7%
[14]	2017	J48, One R, Zero R, Decision stump.	J48-75.5%
[15]	2017	SVM, DT, NB, KNN	SVM-81%
[16].	2017	NB,SMO, REP Tree,J48, MLP	SMO-76.8%
[17]	2018	NB,RF,LR,ML,KNN	KNN-97.9%
[19]	2018	NB, RBF Network, J48	NB-97.3%
[20]	2019	J48,kNN,NB, SVM	SVM-97.9% KNN-97.9%
[21]	2019	KNN,DT,NB, SVM, RF, Voting	Vote-89.9%

Table 1.Shows a summary of the most recent papers

IV. CONCLUSION

Breast cancer is one of the most common cancer in women and also the leading cause of death in women. Breast cancer can be prevented if cancer identifies at an early stage. The aim of this research is to review the various data mining techniques to predict breast cancer and breast cancer recurrence. This research summarizes some of the recent studies were done in data mining about breast cancer prediction in the early stages. Many researchers used the Weka tool for their research purpose. There are other popular tools Tanagra, MATLAB, etc. are also available for data analysis. The Result depends on the selection of algorithms i.e. selection of good algorithms gives an effective result. In some cases, the results are better when applying more than one classification algorithm instead of relying on a single classification algorithm on a dataset also in some cases ensemble method gives highest accuracy. Many studies did comparisons among different classification techniques to classify the breast cancer recurrence or non-recurrence and also patients who have benign or malignant breast cancer. There are many classification techniques, commonly used are Decision tree, K-nearest neighbor, Naïve Bayes, support vector machine. It has been observed that a good dataset multiple algorithms provides better results.

REFERENCES

[1] N. Fatima, L. Liu, S. Hong and H. Ahmed, "Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis," in IEEE Access, vol. 8, 2020.

[2] G. I. Salama, M. B. Abdelhalim and M. A. Zeid, "Breast Cancer Diagnosis on Three Different Datasets Using Multi – Classifiers", International Journal of Computer and Information Technology, Vol. 01, 2012.

[3] Ajay kumar, R. Sushil, A. K. Tiwari," Comparative study of Classification Techniques for Breast cancer Diagnosis", International Journal of Computer Science and Engineering, vol. 7, 2019.

[4] Ali Idri, E. O. Bouchra, M. Hosni, I. Abnane," Assessing the impact of parameters tuning in ensemble based breast cancer classification", Heath and technology, part of springer nature 2020.

[5] B.Prabadevi, N.Deepa, Krithika L.B, Vani Vinod, "Analysis Of Machine Learning Algorithm On Cancer Dataset", International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020.

[6] R. Srinivas, "Managing Large Data Sets Using Support Vector Machines", 2010.

[7] G. Ravi Kumar, "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques" International Journal of Innova tions in Engineering and Technology, vol. 2, 2013

[8] D. S. Jacob, R. Viswan, V. Manju, L. PadmaSuresh, and S. Raj, "A Survey on Breast Cancer Prediction Using Data Mining Techniques", 2018 Conference on Emerging Devices and Smart Systems (ICEDSS),2018,pp.256-258,

[9] Ahmad et al., " Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence", Journal of Health and Medical Informatics, 4.124-130, 2013

[10] V. Chaurasia, S. Pal, "A Novel Approach for Detection using Data Mining Techniques", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 1, pp.2464, 2014

[11] K. sivakami, "Mining Big Data: Breast Cancer Prediction using DT - SVM Hybrid Model", International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, 2015.

[12] A. I. Pritom, M. A. R. Munshi, S. A. Sabab and S. Shihab, "Predicting breast cancer recurrence using effective classification and feature selection technique", 2016 19th International Conference on Computer and Information technology (ICCIT), 2016, pp.310- 340.

[13] U. K. Kumar, M. B. Sai Nikhil and K. Sumangali, "Prediction of Breast Cancer using Voting Classifier Technique", 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), pp.108-114, 2017.

[14] G.R. Banu, Prakash, I. Bashier, Summera,"Applications of data mining classification techniques on predicting breast cancer disease", International Journal of Latest Trends in Engineering and Technology, Vol. (8), pp.321-325.

[15] Uma Ojha, Savita Goel, "A study on Prediction of Breast cancer Recurrence using data mining technique", 2017 7th International Conference on Cloud Computing , Data Science & Engineering – Confluence, 2017,pp.527-530.

[16] DeepikaVerma, Nidhi Mishra, " Analysis and Prediction of Breast cancer and Diabetes disease datasets using Data mining classification Techniques", 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 533-538.

[17] S. Bharati, M. A. Rahman and P. Podder, "Breast Cancer Prediction Applying Different Classification Algorithm with Comparative Analysis using WEKA", 2018 4th International Conference on Electrical Engineering and Information and Communication Technology, 2018, pp. 581-584.

[18] S.N. Singh and S. Thakral, "Using Data Mining Tools for Breast Cancer Prediction and Analysis", 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp.1-4.

[19] V. Chaurasia, S. Pal, BB. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques", Journal of Algorithms & Computational Technology, 2018.

[20] T. A. Shaikh, R. Ali, "Applying Machine Learning Algorithms for Early Diagnosis and Prediction of Breast Cancer Risk", C. R. Krishna et al. (eds.), Proceedings of 2nd International Conference on Communication, Computing and Networking, 2019

[21] N. A. AbouEINadar, A. Saad, "Towards a Better Model for Predicting Cancer Recurrence in Breast Cancer Patients", In: Arai, K., Bhatia, R., Kapoor, S. (eds) Intelligent computing. Compcom 2019. Advances in Intelligent Systems and Computing, Vol 997, Springer, Cham.,pp. 887–899.

[22] S. Eltalhi, H. Kutrani, "Breast Cancer Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", IOSR Journal of Dental and Medical Sciences (IOSRJDMS), Volume 18, 2019, PP 85-94.