COMPARATIVE STUDY ON DATA MINING CLASSIFICATION TECHNIQUES FOR BREAST CANCER PREDICTION

[1]J. Merlyn, [2]P. Thangaraju,
1Student, 2 Associate Professor
[1]Computer Science,
[1]Bishop Heber College, Tiruchirappalli, India

***Abstract*: In medical field we have very large database with different kinds of data stored in it. It is very difficult for us to manage, analyze and handle those data manually. Data Mining is helpful for analyzing, extracting and summarizing useful information from these large dataset. Cancer is the deadly disease that results due to the uncontrolled growth of abnormal cells. There are over 200 types of cancer. This paper deals with various Data mining techniques that can be applied for predicting breast cancer. Various algorithms are available in data mining that help us in developing new concept that can boost us to find models for prediction breast disease accurately. In this paper, an analysis of various data mining classification techniques which is used for predicting breast cancer and analysis of the prediction of survival rate of breast cancer is done.**

Index Term— Data mining, classification, Naïve Bayes, SVM, Decision Tree.

1. INTRODUCTION

Breast cancer is major problem in women and its leads to death. Breast cancer is serious hazard to the lives of people and it is the second leading cause of death in women . In Some cases men can also get affected by breast cancer but in male it accounts for less than .05%. Over the age of 50, two third of women were diagnosed with breast cancer and the majority of the women diagnosed with breast cancer are between the ages of 39 and 40.Normally breast cancer occurs when cells in the breast grows without control. Usually these cells form a tumor that can be seen on an x-ray and patients can feel it as a lump. The cancer cells grow to invade the surrounding tissues and forms a tumor. When tumor is malignant and develop into metastasize. The (AJCC) American Joint Committee on Cancer formed a TNM system to regulate way for the cancer care team to compile information about how far a cancer has been spread all over the body. The stages of breast cancer can be identified by two different stages.

CLINICAL STAGE: In which the presence of cancer cells are identified by doing self-physical exam, biopsy and imaging tests.

PATHOLOGIC STAGE: In which the pathologist performs surgery and examine the breast mass and cells in lymph nodes. The result of this stage provides more accurate result when compared with clinical staging. Pathologic stage observation is more accurate than clinical stage observation and it allows the doctor to do identify the nature of cancer. In TNM staging system which classifies breast cancers into three different stages labels T, N and M. The Fig 1 depicts the cancer tumor on breast.

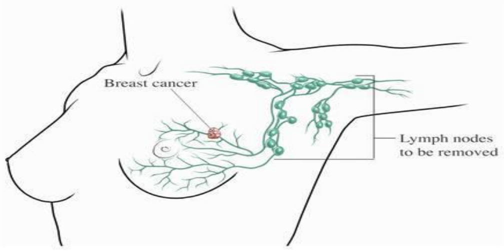

Fig1: Breast Cancer

The term T refers primary tumor in which the tumor is spread to the skin and to the chest wall under the breast, the term T is followed by a number from 0 to 4 characterize size of tumor.. The term N followed by a number from 0 to 3 indicates that cancer has spread to lymph nodes near the breast and how many lymph nodes are affect. The term M(Metastasis) followed by a 0 or 1 imply whether the cancer has spread to nearest organs for example the lungs or bones[1].

II. LITERATURE SURVEY

N.Poomani [2] et al, made a comparative study about 4 algorithms namely naïve Bayes, CART, J48Graft, JRip using breast cancer database. They give conclusion that the J48Graft algorithm gives best classification. The above algorithms are compared based on their execution time and error rate. They observed that highest accuracy of 97% with the lowest error rate 0.9587 is generated in J48 graft classifier.

Hamid Karim Khani Zand [3] et al, combined Naïve Bayes, Artificial neural network and C4.5 algorithms to diagnosis and prediction of breast cancer and survivability rate of breast cancer patients. The dataset from the SEER Public Use Database is used and made comparative study among 3 algorithms C4.5 gives highest accuracy with 86.7%.They observed from their research that the accuracy for the prognosis analysis classification techniques is highly acceptable and it can be help for medical professionals in making decision to avoid biopsy.

Arpit Bhardwaj [4] et al, proposed a new concept Genetically Optimized Neural Network (GONN) algorithm, for solving classification problems. They introduce new concept crossover and mutation operators to reduce the destructive nature of operators by using genetically optimized neural network algorithm to predict breast cancer tumors as benign or malignant. They used dataset from Wisconsin Breast Cancer Database from UCI Machine Learning repository and compared the classification accuracy, sensitivity, specificity. The proposed algorithm gave accuracy of 98.24%, 99.63% and 100% for 50-50, 60-40, 70-30 training-testing partition respectively and 100% for10 fold cross validation.

Peter Adebayo Idowu [5]et al, combined two data mining classification algorithm naïve Bayes' and the J48 decision trees algorithms to predict breast cancer risks for Nigerian patients. They observed J48 decision trees gives a higher accuracy with lower error rates when compared with the naïve Bayes' by using confusion matrices, the naïve Bayes' algorithm had 57 correct and 12 incorrect classifications giving an accuracy of 82.6% whereas the J48 algorithm 65 and 4 respectively with accuracy 94.2%

K.Sivakami [6] et al, proposed Option Extraction and Decision Tree-Support Vector Machine (DT-SVM) Hybrid Models for predictions of breast cancer. The proposed model is implemented using Weka. The Comparison of three data mining classifications techniques with DTSVM hybrid method gives prediction accuracy of 91%. Instance-based learning (IBL) 85.23%, Sequential Minimal Optimization (SMO) 72.45% and Naïve based classifiers is 89.48%. Thus DTSVM appears to be better when compared with other.

Saba Bashir[7] et al, made an analysis on research effectiveness of a group classifiers for breast cancer diagnosis. The combination of five classification Naïve Bayes, Decision tree using Gini index, Decision tree using information gain, Support vector machine and Memory based learner are used to make an entire framework. Weighted voting technique is used to achieve proposed ensemble is 97.42 %. The proposed experiment shows the average results of 85.23 % accuracy, 86.18 % Precision and 76.68 % Recall.

Vikas Chaurasia[8] et al., proposed a system for detecting breast cancer based on RepTree, RBF Network and Simple Logistic classification techniques. The dataset from University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia database is used to evaluate the performance of proposed system. Weka is used for implementation work.The10-fold cross validation method was 74.5%. The simple logistic classification can be used for reducing the dimension of feature space and proposed Rep Tree and RBF Network model can be used to obtain fast automatic diagnostic systems for other diseases. The best Simple logistic Classification algorithm with accuracy of 74.47% and the total time taken is 0.62 seconds.

S. Muthuselvan [9] et al, made a comparative analysis on five classification algorithm like Naïve Bayes, ZeroR, One R, J48 and Random Tree. The Datasets used are collected from the Arignar Anna Cancer Institute for implementing the Data Mining. WEKA Tool is used for implementation. The researchers showed J48 algorithm is best, as it correctly classified instances with accuracy 86.3636% and also the mean absolute error is 0.18 only.

Md. Nurul Amin[10] et al, made a comparative analysis of various classification technique. He used Hematology Dataset with 898 sample for this experiment analysis.he analysed the algorithm in two different phase , one with fully dataset and other by applying feature reduction techniques. On comparision he observed that J48 classifier with an accuracy of 97.16% appeared to be more suitable for hematology data. Naïve Bayes classifier has the lowest average error at 29.71% compared to other algorithm. The results suggest that Naïve Bayes classifier has given best result for use in medical field.

Shweta Kharya [11] et al, proposed a new predictive model WNBC for predicting breast cancer that is based on Naive Bayes Classifier with a new weighted approach. The experiment uses benchmark dataset to compare its performance with the other non-weighted NBC and recently available model like WAC, FWAC, and RBF etc. The accuracy is found to be 92% for WNBC. Fuzzy concepts can be introduced in WNBC.

Akinsola Adeniyi [12] et al, compared 3 algorithms namely C4.5, multi-layer perception and Naive Bayes. Experimental results showed that C4.5 proves to be the best algorithm with highest accuracy as 94% and the total time taken as 0.28 seconds and multilayer perception with an accuracy of 84% and time taken as 12.68 seconds and Bayes network classifier with an accuracy of 77% and time taken 0.03 seconds.C4.5 classifier has the potential to significantly improve the conventional classification methods for use in medical or in general, bioinformatics field.

P. Thangaraju [13] et al, made an analysis on various technical and review papers on disease namely lung, liver and Breast cancer dataset to find which data mining classification techniques are best that can help the medical professional in making decision early diagnosis and to avoid biopsy to predict disease . For breast cancer they show result for 3 algorithms such that C 4.5 with an accuracy of 87%, Decision tree C5 with an accuracy 93% and Decision tree with an accuracy 98.40%.

III CLASSIFICATION TECHNIQUES

3.1. NAÏVE BAYES CLASSIFIERS

Naïve Bayes classifier is one of the classification techniques in data mining which is based on Bayes' theorem with presumption of independence among predictors. In other words naïve Bayes can be defined as presence of a particular feature in a class is unrelated to the presence of any other feature. Every features of an object which depends upon each other or upon the reality of the other features and all of these properties independently contribute to the probability. Naive Bayes classifiers are simple to build and useful for very large data sets and it is known to outperform even highly sophisticated classification methods.[14]

3.2. DECISION TREE

The decision tree classification technique is important tool for classification, prediction, interpretation, and data manipulation which has been used for medical research. Using decision tree models to describe research findings has the following advantages:

- DT Simplifies complex relationships between input variables and target variables by dividing original input variables into significant subgroups and it can be easy to understand and interpret .Non-parametric approach without distributional assumptions.
- Easy to handle missing values without needing to resort to imputation.
- Easy to handle heavy skewed data without needing to resort to data transformation and it is robust to outliers [15]

The decision trees has ID3 algorithm. To construct decision tree, the entropy and Information Gain are used. There are several algorithms which comes under decision tree they are Decision stump, Hoeffiding tree,J48, LMT, Random forest, random tree,Rep tree. In ZeroR model there is no predictor, in OneR model is used to find the single best predictor and by using Bayes rule naive Bayesian includes all predictors and the independence assumptions between predictors but decision tree includes all predictors with the dependence assumptions between predictors.[16]

3.3. SUPPORT VECTOR MACHINE

A support vector machine is one of the classification techniques and it is a set of related supervised learning mode which is used for classification and regression and it is generalized linear classifiers .The regression prediction used for machine learning theory to maximize predictive accuracy and to avoid data fitness .Each data item as a point in n-dimensional space where n is number of features that have the value of each feature being the value of a particular coordinate. SVM is an example for a non-probabilistic linear classifier because the features in the new objects fully determine its location in feature space and there is no stochastic element involved. [17]

TABLE1: SUMMARY FOR CLASSIFICATION TECHNIQUES.

Reference	Technique Applied	Accuracy (%)
N.Poomani	JRip	84.4%
	CART	78.4%
	J48Graft	97.9%
	Naïve Bayes	54.6%
Hamid Karim Khani Zand	C4.5	86.7%
	Artificial neural network	86.5%
	Naïve Bayes	84.5%
Arpit Bhardwaj	Genetically Optimized Neural Network (GONN) algorithm	98.0%
Peter Adebayo Idowu	J48 decision trees	94.2%
	Naïve Bayes	82.6%
K.Sivakami	Decision Tree-Support Vector Machine	91.0%
	IBL	85.2%
	SMO	72.6%
	Naïve Bayes	89.5%

Vikas Chaurasia	Simple logistic	74.7%
	RBF Network	73.8%
	RepTree,	71.3%
S. Muthuselvan	J48	86.4%
	ZeroR	56.8%
	One R	63.6%
	Naïve Bayes	59.1%
	Random Tree	68.2%
Md. Nurul Amin	J48 classifier	97.2%
	Naïve Bayes	70.3%
	Multi layer perception	86.6%
Akinsola Adeniyi	C 4.5	94.0%
	Navie Bayes	76.5%
	Multi layer perception	83.9%
P.Thangaraju	Decision Tree	98.4%
	C4.5	86.7%
	Decision tree C5	93.0%

\

IV.CONCLUSION

This paper provides a study of various technical and survey papers on breast cancer prediction which is used to resolve the issues, algorithms, and techniques for solving problem of breast cancer using breast cancer database. From literature survey, the paper provides the accuracy and efficiency of various classification algorithms that can be used for breast cancer diagnosis and it will also help the medical professional to know, which algorithm will be suitable to develop models to predict the breast cancer.

V.REFERENCE

[1] https://www.babymed.com/cancer/breast-cancer-diagnosis-and-staging.

[2] N.Poomani and Dr.R.Porkodi“ A Comparison of Data Mining Classification Algorithms using Breast Cancer Microarray Dataset: A study” International Journal for Scientific Research & Development, 2015, Vol. 2, No. 12,pp.543-545,ISSN (online): 2321-0613

[3] Hamid Karim Khani Zand,” A Comparitive survey on data mining techniques for breast cancer diagnosis and prediction” Indian Journal of Fundamental and Applied Life Sciences, 2015, Vol.5 , pp. 4330-4339 ISSN: 2231– 6345 2015

[4] Arpit Bhardwaj and Aruna Tiwari “Breast Cancer Diagnosis Using Genetically Optimized Neural Network Model” Elsevier, 2015, Vol.42, No.10, pp: 4611-4620, S0957-4174.

[5] Peter Adebayo Idowu, Kehinde Oladipo Williams, Jeremiah Ademola Balogun and Adeniran Ishola Oluwaranti, “Breast cancer risk prediction using data mining classification techniques” transaction on networks and communication,2015, Vol. 3 No. 2 , pp: 1-11 ,ISSN:2054-7420.

[6] K.Sivakami,“Mining Big Data: Breast Cancer Prediction using DT - SVM Hybrid Model”, International Journal of Scientific Engineering and Applied Science, 2015, Vol.1, No5, ISSN: 2395-3470.

[7]Saba Bashir · Usman Qamar · Farhan Hassan Khan, “Heterogeneous classifiers fusion for dynamic breast
Cancer diagnosis using weighted vote based ensemble” Springer, 2014, Vol.49, No. 5, pp. 2061–2076.

[8] Vikas Chaurasia1, Saurabh Pal “Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability” A Monthly Journal of Computer Science and Information Technology, 2014, Vol. 3, No. 1, pp.10 – 22,ISSN 2320–088X IJCSMC.

[9] S. Muthuselvan, Dr. K. Soma Sundaram, Dr. Prabasheela[10] “Prediction of Breast Cancer Using Classification Rule Mining Techniques in Blood Test Datasets” International Conference On Information Communication And Embedded System,2016, ISBN: 978-1-5090-2552-7.

[10] Md. Nurul Amin, Md. Ahsan Habib "Comparison of Different Classification TechniquesUsing WEKA for Hematological Data" American Journal of Engineering Research, 2015, Vol.4, No.3, pp. 55-61, e-ISSN :2320-0847 p-ISSN : 2320-0936.

[11] Shweta Kharya, Sunita Soni[13]," Weighted Naive Bayes Classifier: A Predictive Model for Breast Cancer Detection" International Journal of Computer Applicatio, 2016, Vol.133 ,No.9,IJCA 0975 – 8887.

[12] Akinsola Adeniyi F1, Sokunbi M.A2, Okikiola F.M 3, Onadokun I.O, "Data Mining For Breast Cancer Classification", International Journal Of Engineering And Computer Science, 2017, Vol. 6, No. 8 August 2017, pp. 22250-22258,ISSN:2319-7242.

[13] Mr. P. Thangaraju, R. Mehala, "Novel Classification based approaches over Cancer Diseases" International Journal of Advanced Research in Computer and Communication Engineering, 2015, Vol. 4, No. 3 ,ISSN (Online) 2278-1021.

[14] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained.

[15] Yan-yan SONG, Ying LU, "Decision tree methods: applications for classification and prediction" Shanghai Archives of Psychiatry, 2015, Vol. 27, No. 2, j.issn.1002-0829.215044.

[16] http://www.saedsayad.com/decision_tree.htm

[17]https://www.quantstart.com/articles/Support-Vector-Machines-A-Guide-for-Beginners.