Ductal Carcinoma in Situ of the Breast - Diagnosis using Data Mining Technique

Priyanka N
B.E.,(M.Tech) [1]
CS & E dept.
AIT College, Chikkamagalur, India

Dr. Pushpa RaviKumar
B.E., M.Tech., Ph. D., L MISTE [2]
CS & E dept.
AIT College, Chikkamagalur, India

Abstract: **The incidence of cancer is increasing rapidly in recent times. Especially women are at high risk as they are more prone to various other types of cancer like breast cancer, cervical cancer etc. Breast cancer is the most common type of cancer that women are facing. Ductal carcinoma in – situ is the most frequently caused variety of breast cancer. It is a common type of non – invasive cancer occurring in general population. Among the various types of breast cancer it is difficult to diagnose the specific type. The purpose of this paper is to make the diagnosis of the type of breast cancer easier using data mining techniques – Naïve Bayes & Decision tree.**

Keywords: Breast cancer, Ductal carcinoma in – situ, Naïve Bayes, Decision tree, Classification.

I.INTRODUCTION

Medical field is growing in fast pace with introduction of newer diseases, making it difficult for a doctor to make accurate diagnosis. Obtaining assistance of computer knowledge & techniques makes this process quick & accurate. The objective of this study is to make the diagnosis precise by utilizing the tools of data mining.

Overview of the Disease:

Cancer is a deadly disease which is leading to high mortality in population. Breast cancer is the most common type of cancer in women occurring worldwide. It is the second common cancer. One among 30 women is likely to develop breast cancer in their lifetime. Previously it was occurring above the age of 50 yrs but now even young women of 30 – 40 yrs also are at risk.

Breast is composed of lobes containing sections & ducts. Cancer can occur in ducts or in lobes. The one, that occurs in ducts in the most common among breast cancer. It may manifest as painless mass in the breast which keeps growing in size. Warm, red & swollen breast are the symptoms of breast cancer. Breast scan & mammography are frequently used when the chances are suspected. In case of doubt, a portion of the mass is sent for biopsy to confirm whether it is benign or malignant. According to WHO, 1.2 million women are diagnosed with breast cancer every year. Early detection of tumor is very essential for the proper treatment. Accurate diagnosis is essential in order to distinguish between benign & malignant tumors. There are various types of breast cancers; the most common is Ductal carcinoma in situ (DCIS). Ductal carcinoma in situ (DCIS) is a type of non – invasive breast cancer. Carcinoma refers to malignant growth in any part of the body which is fatal. Ductal refers to the milk ducts inside the breast tissue. The cancerous growth occurring in the milk ducts is called Ductal carcinoma. In situ means it remains at the site i.e., it is not invasive. Invasive means spreading in nature. Ductal carcinoma in situ is not spreading type of cancer.

II. LITERATURE SURVEY

In [7] adopted neural network for the diagnosis of breast cancer. It utilized negative co-relation algorithm to solve the problem automatically. The author used two approaches – evolutionary approach & ensemble approach. Here evolutionary approach could be used to design compact neural network automatically. Ensemble approach was useful to solve large issues.

In [6] improved the Naive Bayes with the links or associations of the features such as the Tree Augmented Naive Bayes (TAN). This study showed the accuracy of a General Bayesian Network (GBN) applied with the hill-climbing learning approach, but it did not have any impositions on the structure. Here data was represented in a better way. Which did not impose any restrictions on the structure and represented the dataset in a better way? The performance of GBN against Naïve Bayes & TAN, 7 nominal datasets was used with the absence of missing values for the comparison. These nominal datasets were taken from the UCI Machine Learning Repository and they were fed into the Naive Bayes, GBN and TAN for classification with ten-fold cross validation in WEKA software using 286 instances each containing 10 attributes. Naïve Bayes model gave an accuracy of 71.68% followed by 69.58% for TAN and 74.47% for GBN.

In [4] adopted the genetic algorithm model which gave good results compared to other data mining models. It was a good tool to analyze data of breast cancer patients with respect to classification, expression & complexity. For the purpose of comparison artificial neural network, decision tree, logistic regression, and genetic algorithm were used. Accuracy & positive predictive value of each algorithm were utilized to evaluate the correct results.WBC database was included for the data analysis followed by the 10-fold cross-validation. The study showed that genetic algorithm was useful to get accurate results in the classification of breast cancer. The result was reliable& accurate.

In [5] designed classification rules utilizing Particle Swarm Optimization Algorithm for breast cancer datasets. This study used the problem of feature subset selection as a pre-processing step to handle heavy computational efforts. The resulted datasets after feature selection were used for classification. It used particle swarm optimization algorithm.

In [2] compared three very popular different databases of breast cancer (Wisconsin Breast Cancer (WBC), Wisconsin Prognosis Breast Cancer (WPBC) and Wisconsin Diagnosis Breast Cancer (WDBC). It used confusion matrix and classification accuracy based on 10-fold cross validation method. To obtain the appropriate multi – classifier method it introduced a fusion at classification level between these classifiers for each data set. The results highlighted that the classification using fusion of J48 and MLP with the PCA was best when compared to the other classifiers using WBC data set. 92.97% accuracy was got through Naïve Bayes as classifier.

In the paper [3] used feature selection to diagnose the breast cancer. The method generated highest accuracy (99.51%, 99.02% and 98.53% for 80–20% of training-test partition, 70–30% of training-test partition and 50–50% of training-test partition respectively) for a subset that carried five features. It also used other methods to show the performance of SVM with feature selection, such as sensitivity, specificity, confusion matrix, negative predictive value and positive predictive value and ROC curves.

III. PROBLEM STATEMENT

The purpose of this study is to develop a tool to diagnose Ductal carcinoma in situ, a form of breast cancer using Naïve bayes and Decision tree algorithm. It aims to compare the accuracy between these two algorithms.

IV. METHODOLOGY

A Data mining:

Data mining is defined as a process which extracts hidden & useful information from a collection of data. Data mining uses two strategies: supervised and unsupervised learning. It utilizes various parameters to analyze the data. Potentially useful information is scanned & extracted under various headings.

B Classification:

Classification techniques in data mining are the tools to process a huge amount of data. It can classify the data under various variables which are pre- defined & thus generate the desired results. It divides the data into predefined headings. Classification contains two phase such as training phase and testing phase. In training phase every sample in the training set is assumed to belong to a predefined class. In testing phase, unknown test samples are measured to classify using the model build using the training set. It is widely adopted method in healthcare system.

The classification algorithms on Breast Cancer data can be useful to diagnose the outcome of some diseases or discover the genetic nature of tumor. This paper utilizes the classifications - Naïve Bayes & Decision tree methods.

C Naïve bayes:

The Naïve Bayes is frequently used technique used for classification Algorithms in data mining. It is a simple probability classifier. It establishes mutual relation between the attributes. It depends on number of parameters. The classifier works on the principle that variables provided are independent. This paper has used this in healthcare sector to diagnose a type of breast cancer using the data provided. It improves the accuracy of diagnosis with computational effort & quick results. It is based on Bayes theorem as following formula.

$$P(B|xa) = \frac{P(B|xa)P(B)}{P(xa)}$$

D Decision Trees

It is a type of classification where the knowledge is represented in the form of a tree. Tree shaped schematic representation is done to show the decisions. These decisions create rules for the classification of data. To classify the data, it is started with root node & divides the attributes till the terminal node is reached. The nodes are named with the attributes names, edges with positive values & leaves with different classes. The Decision tree representation is based on greedy algorithm which constructs the trees in a top down, recursive, divide and conquer method.

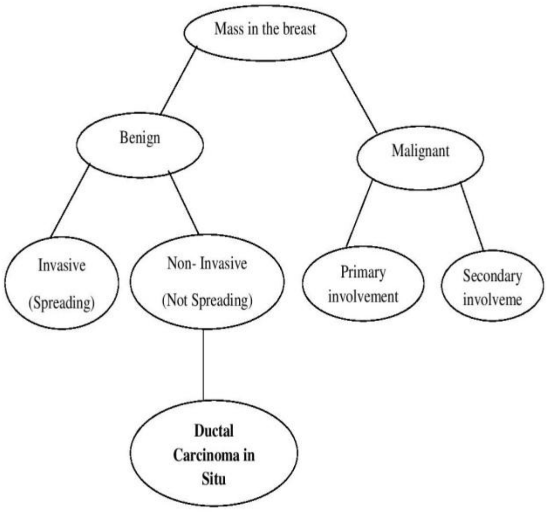

Fig 1: Decision tree showing the manifestation of Breast cancer

C4.5 Decision Tree algorithm is a software extension of the basic ID3 algorithm designed by Quinlan recursively visits each decision node selecting the optimal split. The division is continued till nu further division is possible.

DT is so well known & widely used method as it does not need any domain knowledge or parameter setting. It can be used to explore knowledge in various fields. It is extensively used in healthcare diagnosis. Interactive Dichotomiser 3 (ID3) and C4.5 are the two well known DT algorithms proposed by Quinlan [13]. ID3 uses Entropy and Information Gain to construct a decision tree.

V. DATA SET

Data represents a group of information based on various parameters. This data includes various variables & attributes which are finally analyzed to arrive at a conclusion. The data required for this study is obtained from the case sheets of the patients suffering from breast cancer. The data was obtained from www.archive.ics.uci.edu. The dataset has information taken from 400 patients of breast cancer, of which 252 cases belong to benign class and the remaining 148 cases belong to malignant class. Data about the patient disease & details were collected & stored in database. Breast cancer in general was diagnosed & later classification was done as which type. Ductal carcinoma was diagnosed using the techniques developed.

In this study, we have conducted our study based on the Wisconsin Breast Cancer Dataset (WBCD) taken from UCI machine learning repository (UCI Repository of Machine Learning Databases).

ATTRIBUTES:

1) Age	
2) Sex	
3) Mass in the breast	
4) Pain in the breast	
5) Class	2 -Benign, 4 for malignant
6) Extension of tumor	1, 2, 3, 4, 5
7) Lymph node involvement	10
8) Stage of cancer	5
9) Multiple family members who have had breast, ovarian and/ prostate cancer	Yes = 2 No = 1
10) Birth control pills	Yes = 2 No = 1
11) Gone through menopause	Yes before 55=3 Yes after 55 = 2 No = 2

VI. SYSTEM DESIGN

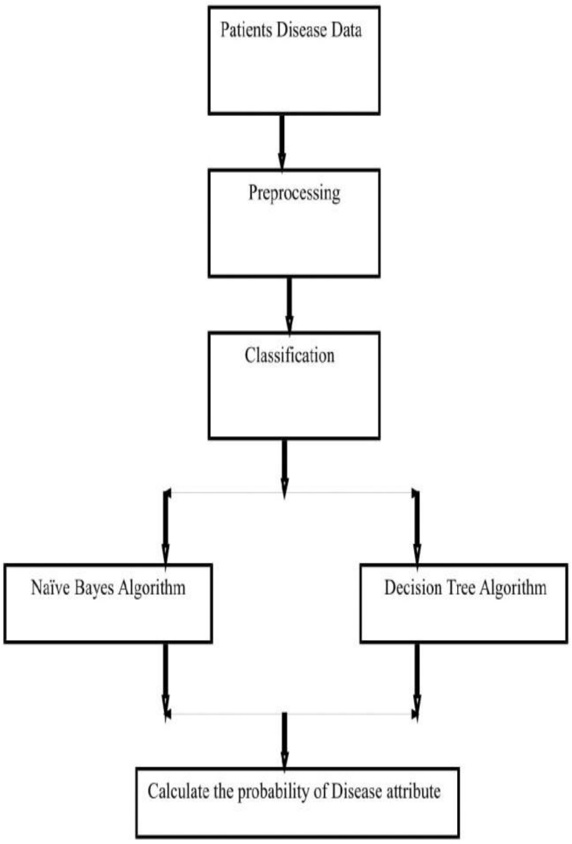

Fig 2: Workflow of Prediction of Breast Cancer Disease.

Initial data is collected from patient health records. Pre – processing is done in order to fill the missing data & differentiate between the attributes. Testing the obtained present data & the previous data using classification techniques. The classification techniques adopted are - Naïve Bayes & Decision tree methods. Using these, probable diagnosis is attained using the attributes. It is shown in the below diagram,

VII. EXPERIMENTAL RESULTS

The study goal was to obtain accurate diagnosis & prognosis of breast cancer. In this study, comparison is done between two classification techniques - Naïve Bayes & Decision tree. Decision tree was proved to be the best among the 2 techniques. It showed an accuracy of 94.85 % as in comparison with 91.48 % obtained in Naïve Bayes. Decision tree was precise in diagnosis of breast cancer. Results obtained suggest that selection of proper parameter & technique is very important to do appropriate diagnosis.

Classification Technique	Accuracy (%)
Naïve Bayes	91.48
Decision Tree	94.85

Table 1: Performance Measures of Algorithms

VIII. CONCLUSION

Cancer is a deadly disease which if not detected & diagnosed earlier can prove fatal to patient’s life. Detection & diagnosis of cancer can be made accurate using software techniques. The purpose of this paper was to adopt two classification techniques - Naïve Bayes & Decision tree in the diagnosis of breast cancer specifically Ductal carcinoma in situ & compare their results. The accuracy obtained with Decision tree was more precise. This can be utilized in healthcare system to make the diagnosis quick & accurate. The system can prove to be of great help in the process.

Future scope

- This system can be utilized to diagnose even other common types of cancer.
- The prognosis can be decided based on the technique so that treatment can be effective.
- The technique can be experimented by changing the attributes & variables to make the diagnosis more precise.

REFERENCES

[1] Gouda I. Salama, M. B. Abdelhalim and Magdy Abd-elghany Zeid, "Breast cancer diagnosis on three different datasets using multi-classifiers." Breast Cancer (WDBC) 32.569 (2012): 2.

[2] Kim W, Kim KS, Lee JE, Noh DY, Kim SW, Jung YS, Park MY, Park RW, "Development of novel breast cancer recurence prediction model using support vector machine." Journal of breast cancer 15.2 (2012): 230 238.

[3] Mehmet Fatih Akay,“Support vector machines combined with feature selection for breast cancer diagnosis.” Expert Systems with Applications 36 (2009) 3240–3247.

[4] Chang Pin Wei and Liou Ming Der, “Comparision of three Data Mining techniques with Ginetic Algorithm in analysis of Breast cancer data”.

[5] Gandhi Rajiv K., Karnan Marcus and Kannan S., “Classification rule construction using particle swarm optimization algorithm for breast cancer datasets,” Signal Acquisition and Processing. ICSAP, International Conference, 2010, pp. 233 – 237.

[6] Sau Loong Ang, Hong Choon Ong and Heng ChinLow, "Classification Using the General BayesianNetwork." Pertanika Journal of Science & Technology24.1 (2016).

[7] Xin Yao, Yong Liu “Neural Networks for Breast Cancer Diagnosis” 01999 IEEE.

[8] V.Kroshnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra (2013) Diagnosis of Lung Cancer Prediction System Using Data mining Classification Techniques International Journal Of Computer Science And Information Technologies, Vol 4(1), 39-45.

[9] Ada Ranjneet Kaur (2013) A Study of Lung Cancer Using Data mining Classification Techniques. International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3, Issue 3.

[10] D.Lavanya and Dr.K.Usha Rani,” Analysis of feature selection with classification Breast cancer datasets”, Vol.2-No.5, oct-nov: 2011, Pg.no:756-763.

[11] Delen Dursun, Walker Glenn and Kadam Amit, “Predicting breast cancer survivability: a comparison of three data mining methods,” Artificial Intelligence in Medicine, vol. 34, pgno. 113-127, June 2005.

[12] Hassanien Ella Aboul and Ali H.M. Jafar, “Rough set approach for generation of classsification rules of Breast cnacer data,” *Journal Informatica*, 2004, vol. 15, pp. 23–38.

[13] Sudhir D., Ghatol Ashok A., Pande Amol P., “Neural Network aided Breast Cancer Detection and Diagnosis”.

BIOGRAPHIES

Ms. Priyanka N is a student of Computer Science from, Adhichunchanagiri Institute of Technology, Chikkamagalur , Presently pursuing M.Tech (CS) from this college. She received B.E from Malnad college of Engineering, affiliated to VTU University, Hassan in the year 2015.

Dr. Pushpa RaviKumar B.E., M.Tech., Ph.D

is working as professor and Head, Department of computer science & Engineering, AIT college,Chikkamagalur. She had 16yr of teaching experience & 5yr of research experience. She had completed her Ph.D from VTU in 2014 as full time research scholar at R.V College of engineering Bangalore. She had completed her M.Tech from R.V. College of Engineering in the year 2007.She has published many research paper in National & International conferences & journals. Her research interest in Data mining, Neural networks, Social network analysis & Computer networks.