# Comparative Study on Various Machine Learning Techniques for Early Diagnosis of Breast Cancer

**[a] Keerti Yeulkar, [b]Rahila Sheikh**
[a] Mtech Scholar, Department of computer Science, Rcert, Chandrapur, India, 442402
[b] Head Of Department, Department of computer Science, Rcert, Chandrapur, India, 442402

## Abstract

Breast cancer is one of the deadliest diseases, is the most common of all cancers and is the leading cause of cancer deaths in world wide. The classification of Breast Cancer data can be useful to predict the outcome of some diseases. Despite the fact, not all general hospitals have the facilities to diagnose breast cancer through mammograms. This paper represents the comparative survey on the machine learning techniques in the Early diagnosis & Prediction of Breast Cancer & summarized the survey on the basis of accuracy obtained by implementing classification algorithm in dataset.

**KEYWORDS**—Algorithms, Breast Cancer, Decision tree, Classification, Clustering, SVM, Neural Network

## I. INTRODUCTION

Data mining is an essential step in the process of knowledge discovery in databases in which intelligent methods is applied in order to extract patterns. Data mining has become a popular technology in current research and for medical domain applications. Cancer is the result of mutation of the genes responsible for the growth of cells in the body[1].This abnormal proliferation may occur to the breast cells which is called Breast Cancer[2].Breast cancer has become one of the most common disease among women that leads to death. Breast cancer can be diagnosed by classifying tumors. Types of Breast cancer includes (a) Ductal Carcinoma (b)Lobular Carcinoma (c) Mucinous Carcinoma (d)Mixed Carcinoma (e)Inflammatory Breast Cancer. Risk of Breast cancer includes the parameters like heredity, early menstruation, late menopause, dense breast tissues, Carcinogens, excessive use of contraceptives, excessive use of alcohol.

Data mining has various techniques such as classification, clustering, prediction, association rule, Decision Trees & Neural Networks. The remaining of this paper is organized as follows. Section II discusses about the Knowledge discovery in Database with subsequently discussed data mining technique i.e Decision tree. Section III illustrates the previous study of Data mining techniques In Breast Cancer. Section IV consists of Dataset description and experiments on Weka. Section V includes conclusion and future work.

## II. KNOWLEDGE DISCOVERY AND DATABASE

Knowledge Discovery in Databases is the process of

searching for hidden knowledge in the massive amounts of
data that we are technically capable of generating and storing[3]. Data, in its raw form, is simply a collection of elements, from which little knowledge can be gleaned. With the development of data discovery techniques the value of the data is significantly improved[3].

Steps in KDD process involves Data cleaning, Data integration, Data Selection, Data transformation, Data mining, pattern evaluation, Knowledge representation.[4]

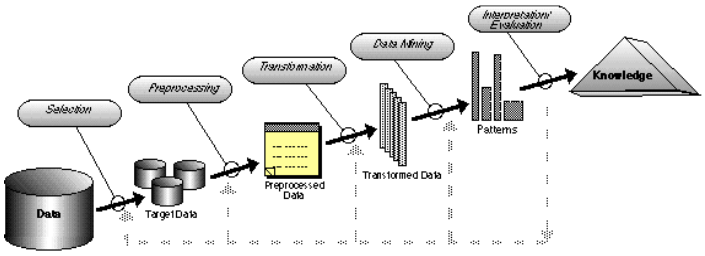


**Figure 1: KDD Process**

**(a)Machine learning** is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data[5]. Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses in prediction-making through the use of computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field[7].

**(b) Decision Tree:** A decision tree is a graph that uses a branching method to illustrate every possible outcome of a decision. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility[7].A decision tree is a predictive model that,as its name implies can be viewed as a tree. Decision trees are powerful classification algorithm that are becoming more and more popular with the growth of data mining in the field of Information systems. Popular decision tree algorithm includes Quinlan's ID3,C4.5 etc. Decision tree can be used for segmentation of the original dataset[9].

## III. PREVIOUS STUDY OF DATA MINING TECHNIQUES IN BREAST CANCER

This section consists of the reviews of various research papers and review articles on data mining techniques applied in breast cancer dataset. The various common data mining methods and techniques used for breast cancer diagnosis are mammography, Biopsy, Positron Emission, Tomography and Magnetic Resonance Imaging[8]. Data mining is a powerful and a new field having various techniques to analyses the recent real world problems. It converts the raw data into useful information in various research fields and finds the patterns to decide future trends in medical field. There are various major data mining techniques that have been developed and used in data mining projects recently for knowledge discovery from database [10]

Santi Wulan Purnami et al. in their research used Support Vector Machines (SVM)[11].They concluded that SVM is a new algorithm of data mining technique that has received popularity in machine learning community. Their paper emphasizes how 1-norm SVM can be used in feature selection and smooth SVM (SSVM) for classification.They implemented a breast

cancer diagnosis like First, feature selection for support vector machines was utilized to determine the important features. Then, SSVM was used to classify the state of disease (benign or malignant) of breast cancer. As a result, SVM can achieve the state of the art performance on feature selection and classification.

Shiv Shakti Shrivastava[8] et al. in his work he observed that neural network and decision approach mostly used by various researchers to create a predictive model and decision rules from the breast cancer data. Breast cancer data was taken from UCI machine learning data repository . This is a secondary data. Dataset consist 10 attributes and 699 instances.

J. Padmavati(2011) [12] they performed a comparative study on WBC dataset for breast cancer prediction using RBF and MLP along with logistic regression. Logistic regression was performed using logistic regression in SPSS package and MLP and RBF were constructed using MATLAB. It was observed that neural networks took slightly higher time than logistic regression

A research paper by Abdelghani Bellaachia and Erhan Guven, presents an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques [13]. In this paper, they used the SEER Public-Use Data and the preprocessed data set consists of 151,886 records, available with 16 fields from the SEER database. They have analyzed the SEER data set using three data mining techniques namely Naïve Bayes, back-propagated neural network, and the C4.5 decision tree algorithms. Several experiments were conducted using these algorithms. Finally, they conclude that C4.5 algorithm has a much better performance than the other two techniques.

A research work by Sujatha G. and K. Usha Rani, survived on effectiveness of data mining techniques on cancer data sets [14]. They state that the tumor is an abnormal cell growth that can be either Benign or Malignant. The use of machine learning and data mining techniques has revolutionized the whole process of cancer diagnosis. Many researchers contributed their effective and accurate diagnosis of the breast cancer diseases in various data mining as basic techniques and various review and technical articles on Tumor and Breast cancer data sets.
An analysis of SEER Dataset for breast cancer diagnosis using C4.5 Classification Algorithm is carried out by Rajesh et al. in [15]. In this research, the C4.5 classification algorithm has been applied to SEER breast cancer dataset to classify patients into either “Carcinoma in situ” (beginning or precancer stage) or “Malignant potential” group. Pre-processing techniques have been applied to prepare the raw dataset and identify the relevant attributes for classification. Random test samples have been selected from the pre-processed data to obtain classification rules. The rule set obtained was tested with the remaining data.

Application of data mining techniques to model breast cancer data is explored by Syed Shajahaan et al. in [16]. In this work, they explore the applicability of decision trees to predict the presence of breast cancer. Also it analyzes the performance of conventional supervised learning algorithms viz. Random tree, ID3, CART, C4.5 and Naive Bayes. Experimental results prove that Random Tree serves to be the best one with highest accuracy. It is found that among various classification techniques random tree outperforms of all other algorithms with highest accuracy rate.

Muhammad Umer Khan et al(2008) [17] they investigated a hybrid scheme based on fuzzy decision trees on SEER data, they performed experiments using different combinations of number of decision tree rules, types of fuzzy membership functions and inference techniques. They compared the performance of each for cancer prognosis and found hybrid fuzzy decision tree classification is more robust and balanced than the independently applied crisp classification.

Jong Pill Choi et al (2009) [18] they compared the performance of an Artifical Neural Network, a Bayesian Network and a Hybrid Network used to predict breast cancer prognosis. The hybrid Network combined both ANN and Bayesian Network. The Nine variables of SEER data which were clinically accepted were used as inputs for the networks. The accuracy of ANN (88.8%) and Hybrid Network (87.2%) were very similar and they both outperformed the Bayesian Network. They found the proposed Hybrid model can also be useful to take decisions.

Chih-Lin Chi et al(2007)[19] they used the Street's ANN model for Breast Cancer Prognosis on WPBC data and Love data. In their research they used recurrence at five years as a cut point to define the level of risk. The applied models successfully predicted recurrence probability and separated patients with good (>5 yrs) and bad(<5yrs) prognoses.

Sudhir D. Sawarkar et al(2006) [20] in their study they applied SVM and ANN on the WBC data .The results of SVM and ANN prediction models were found comparatively more accurate than the human being. The 97% high accuracy of these prediction models can be used to take decision to avoid biopsy.

## IV. EXPERIMENTAL SETUP

In this section we have worked with experimental setup with data mining technique and data mining software. We present a use of decision tree technique on breast cancer data analysis. We can learn various decision rules from this experiment. In this part we will create a model by using decision tree technique. Detail description of decision tree has described in section

### *a) Breast Cancer Data Source and Description*

Breast cancer data was taken from UCI machine learning data repository [21]. This is a secondary data. Dataset consist 10 attributes and 699 instances. Data set description represent in following table:

| S.NO | ATTRIBUTE NAME | RANGE AND CLASS VALUE |
|---|---|---|
| 1 | Clump_Thickness | 1-10 |
| 2 | Cell_Size_Uniformity | 1-10 |
| 3 | Cell_Shape_Uniformity | 1-10 |
| 4 | Marginal_Adhesion | 1-10 |
| 5 | Single_Epi_Cell_Size | 1-10 |
| 6 | Bare_Nuclei | 1-10 |
| 7 | Bland_Chromatin | 1-10 |
| 8 | Normal_Nucleoli | 1-10 |

| | | |
|---|---|---|
| 9 | Mitoses | 1-10 |
| 10 | Class | Malignant, bengin |

**Table 1: Dataset Description**

*b) Weka*

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes[22].Familiarity was also another reason to select Weka data mining software. WEKA has proved itself to be a useful and even essential tool in the analysis of real world data sets. It reduces the level of complexity involved in getting real world data into a variety of machine learning schemes and evaluating the output of those schemes. It has also provided a flexible aid for machine learning research and a tool for introducing people to machine learning in an educational environment. Weka is developed at the University of Waikato in New Zealand. ―Weka‖ stands for the Waikato Environment of Knowledge Analysis [22]. The system is written in Java, an object-oriented programming language that is widely available for all major computer platforms, and Weka has been tested under Linux, Windows, and Macintosh operating systems. Java allows us to provide a uniform interface to many different learning algorithms, along with methods for pre and post processing and for evaluating the result of learning schemes on any given dataset. Weka expects the data to be fed into to be in ARFF format.

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: wisconsin-breast-cancer
Instances: 699
Attributes: 10
Clump_Thickness
Cell_Size_Uniformity
Cell_Shape_Uniformity
Marginal_Adhesion
Single_Epi_Cell_Size
Bare_Nuclei
Bland_Chromatin
Normal_Nucleoli
Mitoses
Class
Test mode: split 65.0% train, remainder test

=== Classifier model (full training set) ===

J48 pruned tree

---

Cell_Size_Uniformity <= 2
| Bare_Nuclei <= 3: benign (405.39/2.0)
| Bare_Nuclei > 3
| | Clump_Thickness <= 3: benign (11.55)
| | Clump_Thickness > 3
| | | Bland_Chromatin <= 2
| | | | Marginal_Adhesion <= 3: malignant (2.0)
| | | | Marginal_Adhesion > 3: benign (2.0)
| | | Bland_Chromatin > 2: malignant (8.06/0.06)
Cell_Size_Uniformity > 2
| Cell_Shape_Uniformity <= 2
| | Clump_Thickness <= 5: benign (19.0/1.0)
| | Clump_Thickness > 5: malignant (4.0)
| Cell_Shape_Uniformity > 2
| | Cell_Size_Uniformity <= 4
| | | Bare_Nuclei <= 2
| | | | Marginal_Adhesion <= 3: benign (11.41/1.21)
| | | | Marginal_Adhesion > 3: malignant (3.0)
| | | Bare_Nuclei > 2
| | | | Clump_Thickness <= 6
| | | | | Cell_Size_Uniformity <= 3: malignant (13.0/2.0)
| | | | | Cell_Size_Uniformity > 3
| | | | | | Marginal_Adhesion <= 5: benign (5.79/1.0)
| | | | | | Marginal_Adhesion > 5: malignant (5.0)
| | | | Clump_Thickness > 6: malignant (31.79/1.0)
| | Cell_Size_Uniformity > 4: malignant (177.0/5.0)

Number of Leaves : 14

Size of the tree : 27

Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on training split: 0 seconds

=== Summary ===

| | | |
|---|---|---|
| Correctly Classified Instances | 233 | 95.102 % |
| Incorrectly Classified Instances | 12 | 4.898 % |
| Kappa statistic | 0.8947 | |
| Mean absolute error | 0.0678 | |
| Root mean squared error | 0.2182 | |

Relative absolute error 14.9318 %
Root relative squared error 45.4348 %
Total Number of Instances 245

=== Detailed Accuracy By Class ===

| TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|---|---|---|---|---|---|---|---|---|
| 0.949 | 0.045 | 0.974 | 0.949 | 0.961 | 0.895 | 0.948 | 0.958 | benign |
| 0.955 | 0.051 | 0.913 | 0.955 | 0.933 | 0.895 | 0.947 | 0.876 | malignant |
| Weighted Average | 0.951 | 0.047 | 0.952 | 0.951 | 0.951 | 0.895 | 0.948 | 0.928 |

=== Confusion Matrix ===

| a | b | <-- classified as |
|---|---|---|
| 149 | 8 | a = benign |
| 4 | 84 | b = malignant |

## V. CONCLUSION AND FUTURE WORK

This paper presents the research review in the use of data mining in breast cancer. Here we observed that decision approach and Neural Network mostly used by various researchers to create a predictive model and decision rules from the breast cancer data. From the SectionIII we concluded that the most of the researcher mostly use Decision tree and Neural Network for classification. Section IV of this paper Consists of an experimental work. We found various if….then rules from decision tree which is represented.we have used J48 classifier of WEKA which is an extension form of ID3 algorithm of decision tree from which we obtained the accuracy of 95%. Future work includes the concept of implementing the various classification rule with another dataset and improving the accuracy of the result.

## REFERENCES

[1] https://www.youtube.com/watch?v=VsviAPGfPUo.
[2] http://ww5.komen.org/AboutBreastCancer/DiagnosingBreastCancer/UnderstandingaDiagnosis/TumorTypesSizesGrades.htmlC. Y. Lin, M. Wu, J. A.
[3] https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/KDD3.htm
[4] https://www.youtube.com/watch?v=a4M3GdI5UFY
[5] https://www.google.co.in/search?q=machine+learning&oq=machi&aqs=chrome.0.69i59j69i57j69i60j69i61j69i60j69i59.2456j0j9&sourceid=chrome&ie=UTF-8
[6] https://www.google.co.in/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=KDD
[7] https://en.wikipedia.org/wiki/Decision_tree
[8] http://accentsjournals.org/PaperDirectory/Journal/IJACR/2013/12/38.pdf
[9] https://www.google.co.in/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=datamining%20techniques
[10] Aarti Sharma, Rahul Sharma,Vivek Kr. Sharma,Vishal Shrivatava, “Application of Data Mining A Survey Paper”, Int. Journal of Computer Science and Information Technologies, Vol. 5, Issue 2,2014, pp. 2023- 2025.

[11]http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Santi%20Wulan%20Purnami.QT.&newsearch=true
[12]Padmavati J. (2011), ―A Comparative study on Breast Cancer Prediction Using RBF and MLP,‖ International Journal of Scientific & Engineering Research, vol. 2, Jan. 2011.
[13]Bellaachia, Abdelghani, and Erhan Guven, "Predicting breast cancer survivability using data mining techniques", Age, Vol. 58, Issue 13, 2006, pp. 10-110.
[14]Sujatha, G., And K. Usha Rani. "A Survey On Effectiveness Of Data Mining Techniques On Cancer Data Sets", Int. Journal of Engineering Sciences Research, 2013, Vol. 04, Issue 1, pp. 1298-1304.
[15]Rajesh K., and Sheila Anand, "Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm", Int. Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 2, 2012, pp. 72-77.
[16]Syed Shajahaan. S, S. Shanthi, V. ManoChitra, "Application of data mining techniques to model breast cancer data", International Journal of Emerging Technology and Advanced Engineering, Vol. 3, Issue 11, 2013, pp. 362-369
[17]Khan M.U., Choi J.P., Shin H. and Kim M (2008), ―Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare‖, Conf Proc IEEE Eng Med Biol Soc., 2008, pp. 48-51.
[18]Choi J.P., Han T.H. and Park R.W.(2009), ― A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis‖, J Korean Soc Med Inform, 2009, pp. 49-57.
[19]Chi C.L., Street W.H. and Wolberg W.H.(2007), ―Application of Artifical Neural Network- based Survival Analysis on Two Breast Cancer Datasets‖, Annual Symposium Proceedings / AMIA Symposium, 2007.
[20]Sudhir D., Ghatol Ashok A., Pande Amol P(2006)., ―Neural Network aided Breast Cancer Detection and Diagnosis‖,7 th WSEAS International Conference on Neural Networks, 2006
[21]archive.ics.uci.edu/ml/datasets.html.
[22]http://www.cs.waikato.ac.nz/ml/weka/
[23]Williams, G. and M. Hegland et al (1998). A Data Mining Tutorial. Presented at the Second IASTED International Conference on Parallel and Distributed Computing and Networks (PDCN'98).
[24]Satyanandam N., Satyanarayana Ch., Md.Riyazuddin,(2012) Data Mining Machine Learning Approaches and Medical Diagnose Systems : A Survey ,International Journal of Computer & Organization Trends – Volume2Issue3- 2012 ,PP 53-60 ISSN: 2249- 2593 http://www.internationaljournalssrg.org
[25]Anunciacao Orlando, Gomes C. Bruno, Vinga Susana, (2010) ―A Data Mining approach for detection of high-risk Breast Cancer groups,‖ Advances in Soft Computing, vol. 74, pp. 43-51, 2010.