Journal of Applied Information Science
10 (1) 2022, 29-33
http://www.publishingindia.com/jais

# Breast Cancer Detection by using Supervised Learning Algorithm

Abarna B.[1*] and Sudha Rajesh[2]
[1]Master of Computer Application, Department of Computer Application, B. S. Abdur Rahman Crescent Institute of Science and Technology, Vandalur, Chennai, Tamil Nadu, India. Email: ananthiananthi501@gmail.com
[2]Assistant Professor, Department of Computer Application, B. S. Abdur Rahman Crescent Institute of Science and Technology, Vandalur, Chennai, Tamil Nadu, India. Email: sudharajesh@crescent.education
*Corresponding Author

**Abstract: Breast cancer is one of the largest second causes of dead among women. So, computer aided diagnosis system will be working on mammograms. In early stage breast cancer can be identified by through CAD algorithms. But, this algorithms accuracy of existing system unsatisfied result. Later, they are using to detect this cancer like microscopic images, mammography to ultrasonography and MRI. This type of prediction gave only false prediction then it took more time and cost effects. In proposed system of in this project, it will be working on big data and machine learning based on nine features in breast cancer data set from UCI Irvine Machine Learning Repository database. Big data is used to pre-processing from various fields from datasets and used to accurate detection. Machine Learning can be used to implementing the supervised algorithm's to detect cancer type based on benign and malignant breast masses. These algorithms will give us additional accurate results for detecting the breast cancer.**

**Keywords: Big data, Breast cancer detection, GUI, Machine learning.**

## I. Introduction

Breast cancer is the second largest cancer in women for world wide. Breast cancer is also principle cause of death from cancer. In early detection cannot be reduced by women death cases by using prediction method. Early diagnosis needs to require an accurate and reliable diagnosis for solution to breast cancer and clear procedure for allow physicians to distinguish between the breast cancer tumours from malignant and Belgians without going for any surgical biopsy treatment. Physicians needs to an advance level detection for breast cancer. A lot of machine learning technique involves to analysis for breast cancer detection, but it gives only 80-90% only. Big data is used to analysis by using with help of hadoop ecosystem, to pre-processing of our breast cancers dataset in set of correct data. By using big data its reduce time to process and maintains very less cost efficient and it can compatible with an all other platforms. Machine learning can be performed with supervised in naive Bayes classification to detect breast cancer model with help of UCI databases. Naive Bayes is one of the very effective classifications algorithms to improve performance and better accuracy result. GUI based predictions is very helpful to way of line treatments and helps in future with more accuracy.

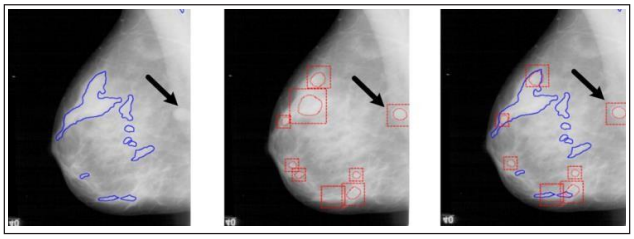


Fig. 1: CAD Result

The existing studies had shown that tests and exams are used to find a disease like in cancer patients who do not have any symptoms. This type of analysis is called on screening test. Screening Exams are such as mammograms, MRI, Microscopic images to find cancer as before they start to some cause of symptoms. Cancers are found that early, when they are small types of detection and have spread easier to treat and have minimum outcomes. Later, they are using to detect by the breast cancer through CAD algorithms. But, it gave only the number of false result to CAD prediction. The overall recall rated by using CAD algorithm detection in 80-90% results.

## II. Related Work

The proposed work is to predict in Naive Bayes classifier has been applied into the UCI Breast cancer dataset on concerning a number of patients in a binary decision on class non-recurrent-events of 151 and also Recurrent events of 47 instance. The input feature will work on 12 relevant attributes and describing by those characteristics of cells. The testing diagnosing had done by 74.24% accuracy of other well-known machine learning techniques. The goal is to be in this project using Machine learning classification (ML) was performed on nine features based on that breast cancer image data set. These data sets help us to investigation in the Breast cancer detection. In

this proposed work, 75% of overall data is used to training set the remaining 25% of data set is used to Testing Data set for using the accuracy of classification technique for our breast cancer detection. According to the result will be highest accuracy is 96.9% which belongs to Naive Bayes network with minimum number of time taken to be build the model is 0.15 seconds and low level of error is 0.2110 with compared to other classification. By using of big data to analyse the breast cancer is to improve the accurate analysis to breast cancer detection.

- A breast cancer by focusing the technique based on convex optimization has been used to imaging scenarios at millimeter waves and significant of the field inside breast model focusing both theory and practice. The breast Cancer has been designed with two antennas in manufactured and measured of focused in stratified media. The result is very good accuracy in theoretical and prediction comparing with other approach. If the optimized antenna is improvement of field level above 15db, it's providing higher sensitivity of cancer diagnostic in millimeter waves.
- Yulee Jiang, results can be observe the performance of diagnosis of malignant and benign by using CAD algorithm. By consider CAD estimated the like hood of malignant to consulting a radiologist for second opinion, mostly the radiologist should recommend more malignant and fewer benign lesions in biopsy. Finally the radiologist will identify on way of a curve and diagnostic accuracy.
- Tan and Gilbert demonstrated the cancer ensemble methods in classifying of microarray data and theoretical explanation of breast cancer. As the result will be suggest that machine learning method can be involve in the expression data for cancer sample to predict with higher accuracy.
- Delenetal, had taken by 202,932 breast cancer patients records to classify into two category of those dataset and identify by the name of survived data (93,273) and not survived data (109,659). Then, the result of predicting breast cancer data will be range of 93% accuracy.
- Stamatia Destounis and Amanda Santacroce, the accurate of false positive rates and prevalence of diagnosis will show that benefits in theoretic harms. Screening mammography treatment is effective with compared than all other screening tests. Those screening tests are beginning stage with large mortality rate and benefit in term of life.
- Shahnorbanum Sahran, Machine Learning can be improved the performance of detection and diagnosing of Cancer data set. Machine learning is one of the most powerful technique is model to understanding classification of data. Machine learning classification finds the relationship between data segregates and easy way to understand that Cancer detection. Using here pixel values in mammogram images and using this SVM classification to improve the detection and segmentation of algorithm can be classified to correct the false prediction. The result was detected with better estimation of breast tumors shapes and sizes.
- M. Tahmooresi, A. Afshar find the cancer using Machine learning with different type of classification to predict the cancer but it Result will comes different accuracy so it take to time for correct decision. This type of prediction will base on images and result is 100% accuracy with work an SVM algorithm. It is based work on only mammogram and ultrasound images and this method to be applied in performance of different data types.
- Qinghua Huqng, Yong Dong Chen results works with an computer aided diagnosis based on two attributes malignant and benign in ultrasound. It will identify by using of bi-clustering mining with an column consistent patterns of training data. The Training data patterns work with an same label as an potential diagnostic method rule. It will be resolve only problems in classification methods and avoid this performing need on a strong classifier to predict the early diagnosis.

## III. Proposed System

In this proposed work, present machine learning algorithm to detect breast cancer based of the data. Breast cancer is a common and second largest cancer for women in around the world. Early Breast cancer detection can be minimum improvement for clinical treatment to patients. Here, using to take a data from UCI machine learning repository database. By using a big data in this project for data pre-processing data from various fields set of data. Big data is mainly used to select pre-processing, this is an reduce a time and it give greater extent than others .This project is only ensure for the greater detection of breast cancer at earlier stage and the mortality rate of affected cancer patients can be reduced and it will be increase the life time of that patients. In our breast cancer detection giving more accuracy result so, can identify by the tumour stage and giving a right treatment at right stage. Machine learning classification results can be demonstrated over higher performance on accuracy level.

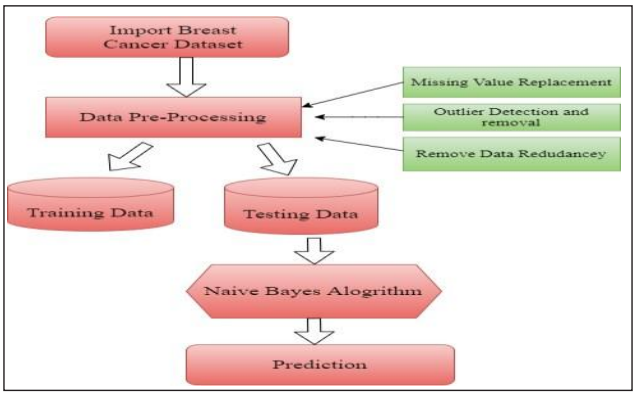



Fig. 2: System Architecture

The Fig. 2 shows the complete system architecture, the breast cancer data have to import in big data tool. Next all of our different kind of field's data can be pre-processing with help of big data. Once, pre-processing of our breast cancer data and then next split into two category in our data set. First one half of our data is a Training data and next half of data is a Testing data. If finishing those step and move on to algorithm implementation. In, our proposed work in this project to implement Naive Bayes algorithm to generate the concept of count and probability. Using this approach, classifier is applied based on testing data to predict with more accuracy and fast performance. Whenever new patient's record is added in this prediction, GUI can be easily finding the breast cancer stage with a class label.

## IV. Modules

### *A. Pre-Processing Data*

In this module, database should be analysis from various kind of fields in excel data, MySQL data, CSV format. These data can be pre-processing with the help of big data tool and finally got on set of proper breast cancer data set.

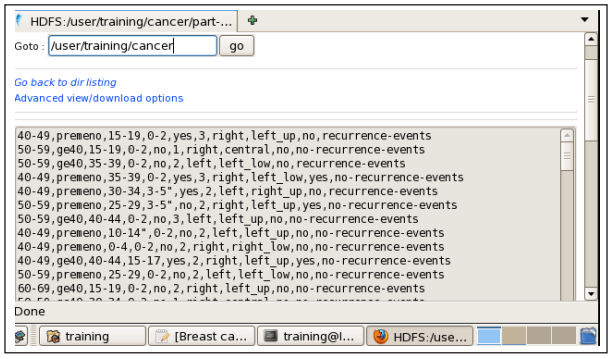



Fig. 3: Image Based Dataset

The Fig. 3 shows the Image Based Classification Dataset is collected from UCI Ml repository website. Dataset is Pre-processed to clean the unwanted data by removing the irreverent attributes and data.

### *B. Data Analysis*

In this module, after analysing the data by using hadoop tool and convert into CSV format, to import the CSV file in python and processing that data and view about all information details in data set, in this step of data analysis is helpful to processing of our next step in prediction.

### *C. Data Visualization*

In this module, after analysing the data by using hadoop tool and convert into CSV format, to import the CSV file in python to read the dataset for visualization of retrieving in that dataset. And then, analysis the breast cancer information and comparing the process for visualization to find area of breast tumour affected the patients and size of tumour.

### *D. GUI Based Prediction*

In this module, Breast cancer data with the help of Naive Bayes algorithm to predict the breast tumour accuracy. The Naive Bayes classifier works based on the presence or absence in particular feature in class unrelated or any other feature will be works in the given class variable.

ACCURACY = (TP+TN)/(TP+FP+TN+TP) SENSITIVITY = TP/(TP+FP)

SELECTIVITY = (TP+TN)/(TP+FP+TN+TP) SENSITIVITY = TP/(TP+FP)

First, importing the data set is already predefined in sklearn module it contains information all about in breast cancer Data set. Next to import any algorithm in our prediction must have split into data in training and testing format class module and load into the data variable. X prediction can identify the Feature values and Y prediction denotes in Targets value of our prediction. Fits our training data into this Naïve Bayes algorithm so the computer automatically can be trained using our data. Next to testing the breast cancer data for prediction which make it as an array to predict the target value as an output based testing and training dataset. A breast cancer dataset can be analyzed and to predict through in view of created graphs and graphical user interface results. It has been observed a good data set can be easily to detect the cancer accuracy. GUI based prediction is deciding to find the way of treatment to be followed by given database.

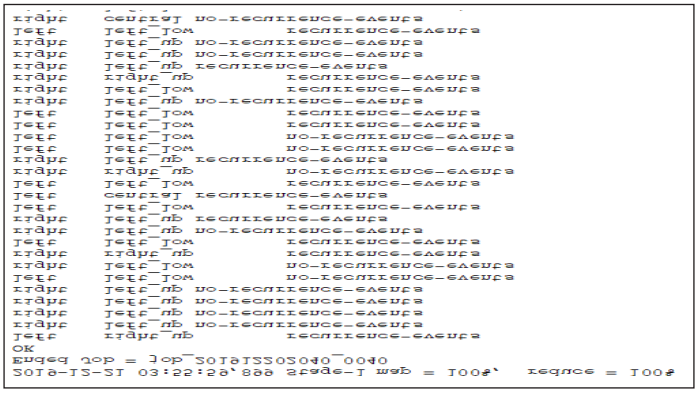


Fig. 3: Pre-Processing Database

## V. Result

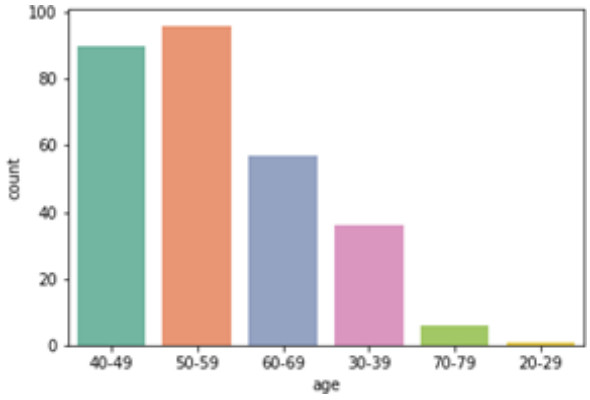
The result that we are getting from our implemented algorithm, as reflected in the figure our implemented algorithm provides better results than previous algorithm terms of accuracy,

sensitivity and specificity. Also, the implemented algorithm provides better results when compared with other existing algorithm.

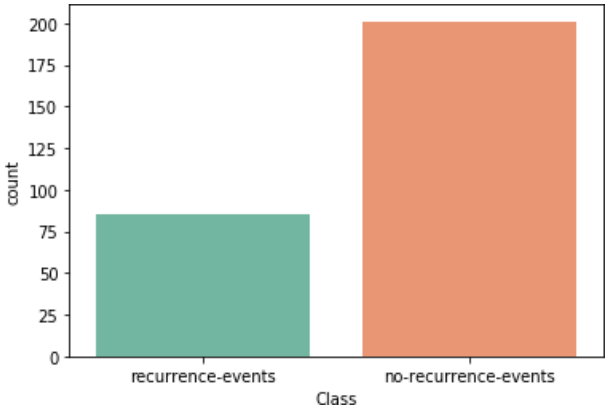
| age | menopause | tumor-size | inv-nodes | node-caps | deg-malig | breast | breast-quad | irradiat | Class |
|---|---|---|---|---|---|---|---|---|---|
| 40-49 | premeno | 15-19 | 0-2 | yes | | 3 right | left_up | no | recurrence-events |
| 50-59 | ge40 | 15-19 | 0-2 | no | | 1 right | central | no | no-recurrence-events |
| 50-59 | ge40 | 35-39 | 0-2 | no | | 2 left | left_low | no | recurrence-events |
| 40-49 | premeno | 35-39 | 0-2 | yes | | 3 right | left_low | yes | no-recurrence-events |
| 40-49 | premeno | 30-34 | | 03-May | yes | 2 left | right_up | no | recurrence-events |
| 50-59 | premeno | 25-29 | | 03-May | no | 2 right | left_up | yes | no-recurrence-events |
| 50-59 | ge40 | 40-44 | 0-2 | no | | 3 left | left_up | no | no-recurrence-events |
| 40-49 | premeno | | Oct-14 | 0-2 | no | 2 left | left_up | no | no-recurrence-events |
| 40-49 | premeno | 0-4 | 0-2 | no | | 2 right | right_low | no | no-recurrence-events |
| 40-49 | ge40 | 40-44 | 15-17 | yes | | 2 right | left_up | yes | no-recurrence-events |
| 50-59 | premeno | 25-29 | 0-2 | no | | 2 left | left_low | no | no-recurrence-events |
| 60-69 | ge40 | 15-19 | 0-2 | no | | 2 right | left_up | no | no-recurrence-events |
| 50-59 | ge40 | 30-34 | 0-2 | no | | 1 right | central | no | no-recurrence-events |
| 50-59 | ge40 | 25-29 | 0-2 | no | | 2 right | left_up | no | no-recurrence-events |
| 40-49 | premeno | 25-29 | 0-2 | no | | 2 left | left_low | yes | recurrence-events |
| 30-39 | premeno | 20-24 | 0-2 | no | | 3 left | central | no | no-recurrence-events |
| 50-59 | premeno | | Oct-14 | 03-May | no | 1 right | left_up | no | no-recurrence-events |
| 60-69 | ge40 | 15-19 | 0-2 | no | | 2 right | left_up | no | no-recurrence-events |
| 50-59 | premeno | 40-44 | 0-2 | no | | 2 left | left_up | no | no-recurrence-events |
| 50-59 | ge40 | 20-24 | 0-2 | no | | 3 left | left_up | no | no-recurrence-events |
| 50-59 | lt40 | 20-24 | 0-2 | | | 1 left | left_low | no | recurrence-events |
| 60-69 | ge40 | 40-44 | | 03-May | no | 2 right | left_up | yes | no-recurrence-events |
| 50-59 | ge40 | 15-19 | 0-2 | no | | 2 right | left_low | no | no-recurrence-events |
| 40-49 | premeno | | Oct-14 | 0-2 | no | 1 right | left_up | no | no-recurrence-events |
| 30-39 | premeno | 15-19 | | 06-Aug | yes | 3 left | left_low | yes | recurrence-events |
| 50-59 | ge40 | 20-24 | | 03-May | yes | 2 right | left_up | no | no-recurrence-events |
| 50-59 | ge40 | | Oct-14 | 0-2 | no | 2 right | left_low | no | no-recurrence-events |

Fig. 4: Dataset

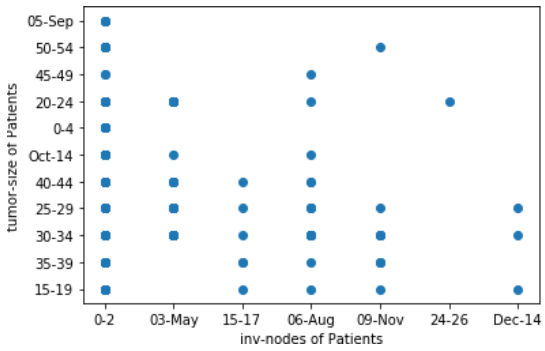
The above Fig. 4 explain the complete the data pre-processing and display to the attributes in the data set frame, so can find that to processing our next step in these project.



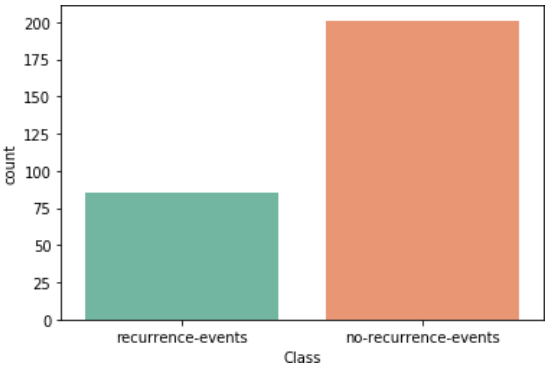

5(a)




5(b)




5(c)




5(d)

Fig. 5: Data Visualization

The above Fig. 5(a) represents the age of persons and count affected by the breast cancer 5(b) identify by the how many number of person occurred in recurrence events and non-recurrence events category. 5(c) represents which age of the persons affected by breast cancer with collected of two months of data. 5(d) compares the type of patients and inv-nodes to analysis the percentage.

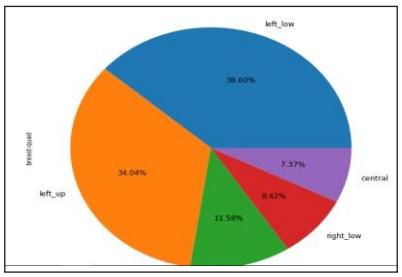



Fig. 6: Visualization of Breast Cancer

The above Fig. 6 shows details information in breast Cancer data set for visualization and find out that patient affected range in the visualization. Once analysis the range of tumour it can be to prediction will based on that process.

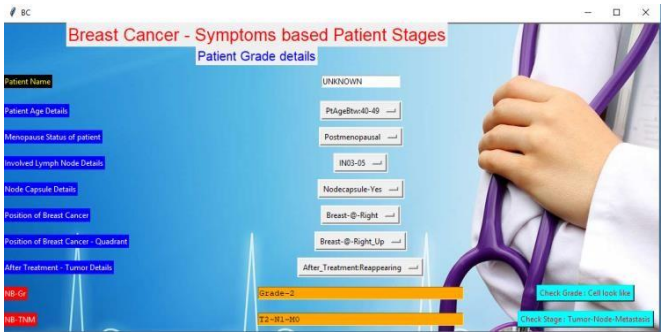



Fig. 7: GUI Based Prediction

The above Fig. 7 shows the GUI based result in this output contains about breast cancer details information will be working with an our machine learning classification and these type of output can be using to detect the cancer in very fast and decision to take way of treatment to affected patients. This prediction will be working with an age of details, menopause status of patients, lymph node details, node capsule details, position of breast cancer, treatment stage of patients to predict and finally give output as an in grade and stage of the tumour and then machine learning classification by using Naïve Bayes theorem prediction result is an 96% accuracy.

## VI. Conclusion and Future Enhancement

In proposed work implemented using machine learning technique is helpful in diagnosing cancer type into assist oncologist in decision taking for cancer patient. The experimental result shows that approach performs better predicting the cancer type as benign and malignant. The future research in this Breast Cancer data detection is solving to the regarding research paper using by Machine learning method. And, to analyze the Breast Cancer data in Hadoop ecosystem and to improve the accurate analysis of breast cancer. The work will include experiment by using machine learning with Hadoop tool to make java interface by accurate prediction or involve with one or two more algorithm combination can predictive the breast cancer detection with great higher accuracy.

## References

[1] I. Iliopoulos, S. D. Meo, M. Pasian, M. Zhadobov, P. Pouliguen, P. Potier, L. Perregrini, R. Sauleau, and M. Ettorre, "Enhancement of penetration of millimetre waves by field focusing applied to breast cancer detection," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 3, pp. 959-966, Mar. 2021, doi: 10.1109/TBME.2020.3014277.

[2] Y. Jiang, "Computer-aided diagnosis of breast cancer in mammography: Evidence and potential," *Technology in Cancer Research & Treatment*, vol. 1, no. 3, pp. 211-216, Jun. 2002.

[3] A. C. Tan, and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," *Applied Bioinformatics,* vol. 2, no. sp.3*,* pp. S75-S83, 2003,

[4] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113-127, 2005.

[5] S. Destounis, and A. Santacroce, "Age to begin and intervals for breast cancer screening: Balancing benefits and harms," *American Journal of Roentgenology*, vol. 210, no. 2, pp. 279-284, 2018.

[6] S. Sahran, A. Qasem, K. Omar, D. Albashih, A. Adam, ....., and A. Shukor, "Machine learning methods for breast cancer diagnostic," Nov. 2018.

[7] M. Tahmooresi, A. Afshar, B. B. Rad, K. B. Nowshath, and M. Bamiah, "Early detection of breast cancer using machine learning techniques," *Journal of Telecommunication, Electronic and Computer Engineering* , vol. 10, pp. 21-27, 2018.

[8] Q. Huqng, Y. Chen, L. Liu, D. Tao, and X. Li, "On combining biclustering mining and Adaboost breast tumour classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, 2019.