**Original Research Article** **Bioinformatics**

**International Journal of Pharma and Bio Sciences**

**ISSN 0975-6299**

# COMPUTATIONAL ANALYSIS TO ENHANCE BREAST CANCER DIAGNOSIS & PROGNOSIS

**R.RAJESHKANNAN**

*Assistant Professor (Sr), SCOPE, VIT University, Vellore, Tamilnadu, India.*

## ABSTRACT

Breast cancer is the deadliest disease and the major cause of cancer deaths in women worldwide. The milestone in breast cancer diagnosis is to distinguish malignant tumour from benign breast tumour for which reliable diagnostic procedure is required by the physicians.Prognosis of breast cancer is to predict the recurrence in breast cancer patients for whom they had their excision already. Data mining techniques have revolutionized the diagnostic and prognostic procedure in breast cancer. One of the new research's in data mining application involves analyzing. This survey work analyses the various reviews and technical articles, on breast cancer diagnosis. The main goal of this research is to explore the overview of the current research being carried out using the data mining techniques to enhance the breast cancer diagnosis.

**KEYWORDS:** **Breast cancer, Naive Bayes, C4.5 classification decision tree algorithm, Bayesian Networks**.

**R.RAJESHKANNAN**
**Assistant Professor (Sr), SCOPE, VIT University, Vellore, Tamilnadu, India.**

***Corresponding author**

# INTRODUCTION

Breast Cancer turns into a risky infection in today's time. It is only just thin tubes that convey milk from the lobules of the bosom to the little areola. Another kind of bosom growth is the lobular carcinoma, which starts in the lobules of the bosom. Intrusive breast growth is bosom malignancy that has spread from where it started in the bosom pipes or lobules to encompassing typical tissue. Breast disease happens in both men and ladies, albeit male bosom malignancy is uncommon. As indicated by the review of United States in 2014, there are 232,670 females and 2,360 guys having this kind of new cases with respect to the bosom disease. Among them, 40,000 females and 430 men were passing amid the period. Early smoldering indications of bosom growth may ingest the identification of another protuberance or an adjustment in the bosom skin. These are the signs and indications for the early discovery of the breast growth. By performing month to month bosom self-exams, a patient will have the capacity to all the more effortlessly distinguish any adjustments in her bosom. In the event that patient discovered anomalous changes in her bosom she gives a way to contact human services specialists. Bosom growth research has been one of the vital examination themes in restorative science amid the late years. The arrangement of Breast Cancer information can be helpful to foresee the consequence of a few sicknesses or find the hereditary conduct of tumours. There are numerous methods to anticipate and characterization bosom malignancy design.[1]

### *RELATED WORKS*

Delen et al. Preprocessed the Surveillance, Epidemiology, and End Results Program ( SEER) information (time of 1973 - 2000 with 433,272 records named as breast.txt) for breast cancer disease to evacuate redundancies and missing data. The subsequent dataset had 202,932 records, which then pre-ordered into two groupings of "survived" (94,273) and "not survived" (110,659) contingent upon the Survival Time Recode (STR) field. The "manage to survive" class is all records that have a quality more noteworthy than or measure up to 60 months in the STR field and the "not able survive" class have a substantial effect to the remaining records. After this stride, the information mining calculations are connected to these information sets to foresee the reliant field from 16 indicator fields.[2,3] We have seen that the quantity of "not survived" patients utilized does not corporate the numbers of "not alive" (field Vital Status Recode ) patients in the initial 60 months of survival time. Actually, the numbers of persons of "not able to survive" patients are predicted that would be around 21% in light of the bosom disease survival measurements of 80%.[4] In our discourse with the creators of, we discovered that the renaming procedure was not precise in predicting the records of the "not able to survive" tag.[10] The scope of survival analysis of a group or group of individual for each of whom there is defined point of the event often called a failure occurring after a length of time called failure time.[5] They didn't think about not one or the other the Vital Status Recode (VSR) nor the Cause of Death (COD). They accept that all patients are dead with malignancy, which is not generally genuine. Ada et al made an endeavor to distinguish the lung tumours from the malignancy pictures and steady apparatus is created to check the typical and unusual lungs and to foresee the survival rate also, years of an unusual patient so that tumour patients lives can be spared.[6] V.Krishnaiah et al built up a model lung tumour forecast framework utilizing information mining characterization procedures.[7] The best model to foresee patients with breast cancer tumour ailment seems, by all accounts, to be Credulous Bayes took after by on the off chance that then lead, Choice Trees and Neural System. For Finding of Lung Malignancy Sickness Innocent Bayes watches preferable results and fared better over Choice Trees.[5] Discuss the evolving changes in cancer mortality statistics in the U.S.Charles Edeki et al Proposes that none of the information mining, what's more, factual realizing calculations connected to a bosom malignancy dataset beat the others in such way that it could be proclaimed the ideal calculation and none of the calculation performed ineffectively as to be disposed of from the future expectation model in bosom tumour survivability undertakings.[8] This paper discusses several data mining algorithms and techniques that we have developed at the University of Arizona Artificial Intelligence Lab.[10] Sahar A. Mokhtar et al have examined three diverse grouping models for the forecast of the seriousness of bosom masses to be specific the choice tree, simulated neural system and bolster vector machine.[9] The uniqueness of medical data mining, Artificial Intelligence in Medicine the choice tree model is built utilizing the Chi-squared programmed connection location technique and pruning strategy was utilized to discover the ideal structure of the simulated neural system model and at long last, bolster vector machine have been assembled utilizing polynomial part.[10] The exhibitions of the three models have been assessed utilizing factual measures, increase and Roc outline. Bolster vector machine model outflanked the other two models on the expectation of the seriousness of bosom masses. Rajashree Dash et al a hybridized K-implies calculation has been proposed which consolidates the progressions of dimensionality decrease through Patient-controlled analgesia (PCA), a novel in statement methodology of the group focuses and the progressions of allocating information focuses to suitable groups.[11] Utilizing the proposed calculation a given information set was parceled into k groups. The trial demonstrates that the proposed calculation gives better productivity and precision correlation with unique k-implies calculation with diminished time. Impediments are the quantity of bunches (k) is required to be given as data. The strategy to locate the underlying centroids may not be dependable for extensive data set. Ritu Chauhan et al concentrates on bunching calculation such as hepatic arterial chemotherapy (HAC) and K-Implies in which, HAC is connected with K-implies to decide the quantity of groups. The nature of bunch is enhanced, if HAC is connected with K-implies.[12] Dechang Chen et al created Ensemble Algorithm for Clustering Cancer Data ( EACCD ) calculation which is a two stage grouping technique. In the initial step, a different measure is learnt by using Primary Acquired Melanosis (PAM), and in the second step, the learnt disparity is utilized with a progressive bunching calculation to get bunches of patients.[13] These bunches of patient's frame a premise

of a prognostic framework. S M Halawani et al proposes that probabilistic grouping calculations performed well than progressive grouping calculations in which all information focuses were bunched into one bunch, might be because of improper decision of separation measure.[14] Zakaria Suliman zubi et al utilized a few information mining systems, for example, neural systems for recognition and grouping of lung growths in X-beam mid-section movies to arrange issues going for distinguishing the qualities that show the gathering to which every case has a place.[16]

***FRAMEWORK***

Once a patient is determined to have breast growth, the dangerous protuberance must be extracted. This system, doctors must decide the anticipation of the infection. This is the expectation of the normal stream of the illness. Visualization is vital on the grounds that the sort and the force of the prescriptions depend on it. Anticipation issue is additionally called as "examination of survival or lifetime information". It represents a more troublesome issue than that of conclusion since the information is blue-penciled. This situation, we can characterize the patient as repeat and we know an ideal opportunity to repeat Tumour Tissue Repository (TTR). Then again, we don't watch a repeat in many patients. For these, there is no genuine time when we can consider the patient a non-intermittent case. Along these lines, the information is viewed as controlled since we don't have a clue about the season of repeat. For such patients, all known is just the season of their last registration. We call this the sickness free survival time (DFS). Forecast helps in building up a treatment arrangement by anticipating the result of an ailment. There are three prescient foci of malignancy visualization:

1) Prediction of malignancy helplessness (hazard appraisal),
2) Prediction of tumour repeat and
3) Prediction of tumour survivability.

The target of prognostic forecasts is to handle cases in which malignancy has not repeated (blue-penciled information) and in addition the case in which tumour has repeated at a particular time. In this manner, bosom disease prognostic issues are essentially in the extent of the broadly talked about characterization issues. This area comprises of the survey of different specialized and audit articles on information mining strategies connected in bosom tumour visualization.

## MATERIALS AND METHODS

Naive Bayes Classifier is a probabilistic model based on the Baye's theorem. It is defined as a statistical classifier. It is one of the frequently used method for supervised learning. It provides an efficient way of handling any number of attributes or classes which is purely based on probabilistic theory. Bayesian classification provides practical learning algorithms and prior knowledge on observed data.[17]

- Let X be a data sample: class label is unknown
- Let H be a hypothesis that X belongs to class C
- Classification is to determine P (H|X), (i.e., posteriori probability): the probability that the hypothesis holds given the
  observed data sample X
- P (H) (prior probability): the initial probability
- P(X): probability that the sample data is observed
- P (X|H) (likelihood): the probability of observing the sample X, given that the hypothesis holds
- Training data X, the posterior probability of a hypothesis H, P(H|X), follows the Bayes' theorem
  P (H|X) = (P (X|H) * P (H))/ P(X)

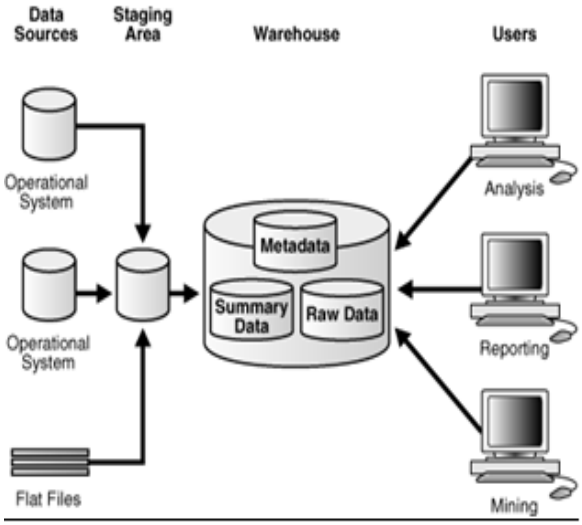


**Figure 1**
***Typical Data Mining System***

***Figure 1 illustrates the data collected remains the central part and the Users do their Analysis, Reporting and Mining. The sources of data come from preprocessing done on the data coming from operational systems and flat files. The data to be processed is kept as raw data, Training data is known as the Meta Data and the mined data are called the Summary data.***

### *4.1 Decision Tree*

A decision tree was developed by Quinlan Ross which is an extension to ID3.[17](Iterative Dichotomiser 3). Decision Tree classification uses entropy and information gain for tree splitting. It is suitable for handling both categories as well as continuous data. A threshold value is fixed such that all the values above the threshold are not taken into consideration. The initial step is to calculate the information gain for each attribute.

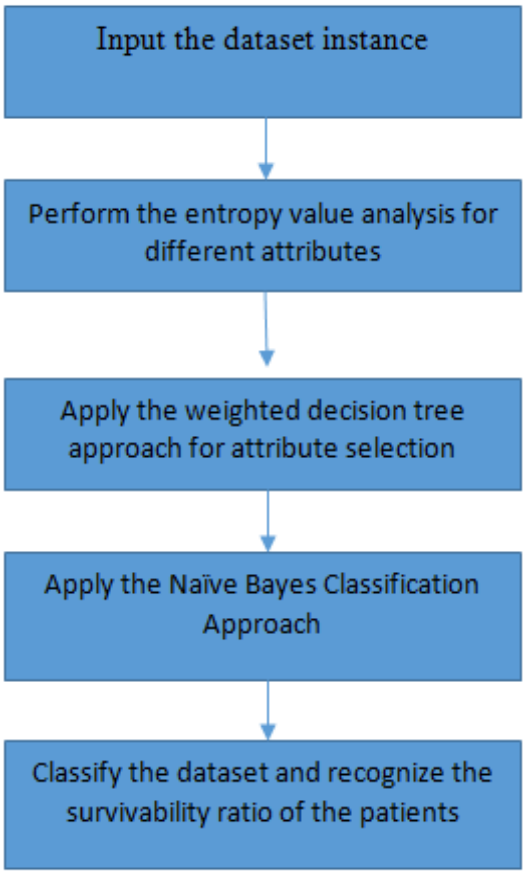


**Figure 2**
***Naïve Bayes Tree***

Naive Bayes Classifier is a probabilistic model characterized as a factual classifier and one of the most utilized technique for regulated learning. It gives a proficient method for taking care of any number of properties or classes which is simply in view of probabilistic hypothesis. Bayesian characterization gives reasonable learning calculations and earlier information on checking information.[15]

- Let X be an information test: class name is obscure.
- Let H be a speculation that X has a place with class C.
- Classification is to decide P (H|X), (i.e., posteriori likelihood): the likelihood that the speculation holds given the watched information test X.
- P (H) (earlier probability): the underlying probability.
- P(X): probability that specimen information is observed.
- P (X|H) (probability): the probability of observing the specimen X, given that the speculation holds.
- Training information X, back probability of a speculation H, P(H|X), takes after the Bayes' hypothesis
  P (H|X) = (P (X|H) * P (H))/P(X)

## EXPERIMENT AND RESULTS

| | age | menopause | tumor-size | inv-nodes | node-caps | deg-malig | breast | breast-quad | irradiat | recurrence |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 30-39 | premeno | ? | 0-2 | no | 3 | left | left_low | no | no-recurrence-... |
| 2 | 40-49 | premeno | ? | 0-2 | no | 2 | right | right_up | no | no-recurrence-... |
| 3 | 40-49 | premeno | ? | 0-2 | no | 2 | left | left_low | no | no-recurrence-... |
| 4 | 60-69 | ge40 | ? | 0-2 | no | 2 | right | left_up | no | no-recurrence-... |
| 5 | 40-49 | premeno | 0-4 | 0-2 | no | 2 | right | right_low | no | no-recurrence-... |
| 6 | 60-69 | ge40 | ? | 0-2 | no | 2 | left | left_low | no | no-recurrence-... |
| 7 | 50-59 | premeno | ? | 0-2 | no | 2 | left | left_low | no | no-recurrence-... |
| 8 | 60-69 | ge40 | ? | 0-2 | no | 1 | left | left_low | no | no-recurrence-... |
| 9 | 40-49 | premeno | ? | 0-2 | no | 2 | left | left_low | no | no-recurrence-... |
| 10 | 40-49 | premeno | ? | 0-2 | no | 2 | right | left_up | no | no-recurrence-... |
| 11 | 40-49 | premeno | 0-4 | 0-2 | no | 3 | left | central | no | no-recurrence-... |
| 12 | 50-59 | ge40 | ? | 0-2 | no | 2 | left | left_low | no | no-recurrence-... |
| 13 | 60-69 | lt40 | ? | 0-2 | no | 1 | left | right_up | no | no-recurrence-... |
| 14 | 50-59 | ge40 | ? | 0-2 | no | 3 | left | right_up | no | no-recurrence-... |
| 15 | 40-49 | premeno | ? | 0-2 | no | 3 | left | left_up | no | no-recurrence-... |
| 16 | 60-69 | lt40 | ? | 0-2 | no | 1 | left | left_low | no | no-recurrence-... |
| 17 | 40-49 | premeno | ? | 0-2 | no | 2 | left | left_low | no | no-recurrence-... |
| 18 | 50-59 | premeno | ? | 0-2 | no | 3 | left | left_low | no | no-recurrence-... |
| 19 | 60-69 | ge40 | ? | 0-2 | no | 3 | left | left_low | no | no-recurrence-... |
| 20 | 50-59 | ge40 | ? | 0-2 | no | 1 | right | right_up | no | no-recurrence-... |

**Table 1.1**
***Breast Cancer data set***

This sample data set was taken from the breast cancer dataset provided by the University of California Irvine (UCI). Our whole analysis comprised of four similar data sets. The description about each column used in the data set is provided in Table 1.2.

| Attribute Name | Description |
|---|---|
| Age | Patient's Age in years |
| Menopause | The period in a woman's life when menstruation ceases |
| Tumour-size | Patient's tumour-size of her breast |
| Invasion –nodes | Node size in main portion of the breast |
| Node-caps | The Node is present or not in cap of the breast |
| Malignant degree | Stage of breast cancer |
| Breast | Left breast or Right breast or both breasts |
| Breast-quad | A Portion of the breast for example left-up, left-low, Right-up, Right-low, central |
| Irradiate | Present or not(YES/NO) |
| Class | No-recurrence-events, recurrence-events (reduce the risk of breast cancer) |

**Table 1.2**
***Brest Cancer Dataset Attribute Description***

***Here the class attributes used in the recurrence of the event. i.e. whether they occur recursively or not. The other column attributes are basic attributes required for accurate analysis.***

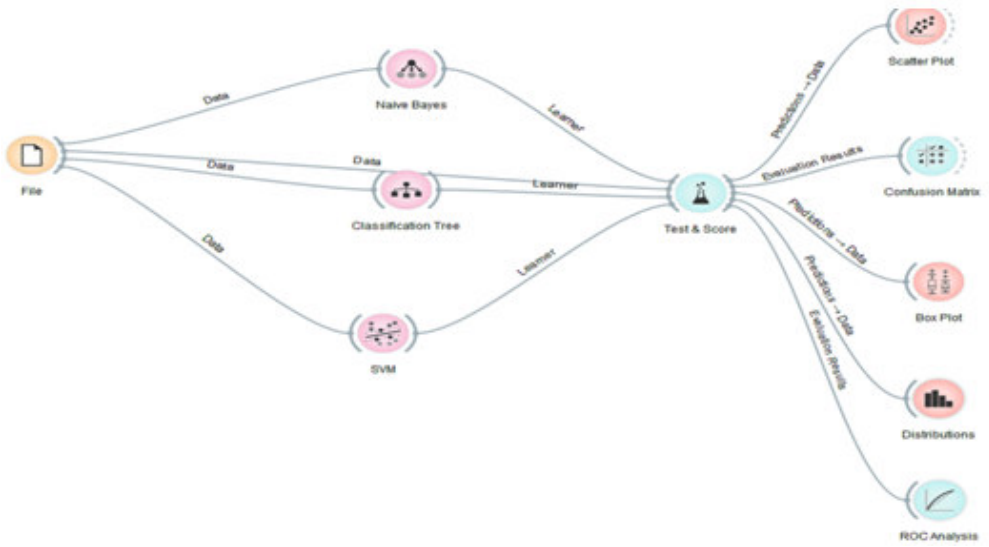


**Figure 3**
***Orange implementation circuit***

The usage circuit given in Figure 3 involves a document containing the information set . This document is associated with different arrangement modules for further order. In our usage, we have utilized Naïve Bayes, Classification Tree and Support Vector Machines **(**SVM) for characterization reason. The characterized information comes to Test and Score to change over them to noticeable yields. The Test and Score broke down information further goes to different representation procedures. In our examination, we have utilized Scatter plot, Confusion Matrix, Box Plot, Distributions and ROC investigation.

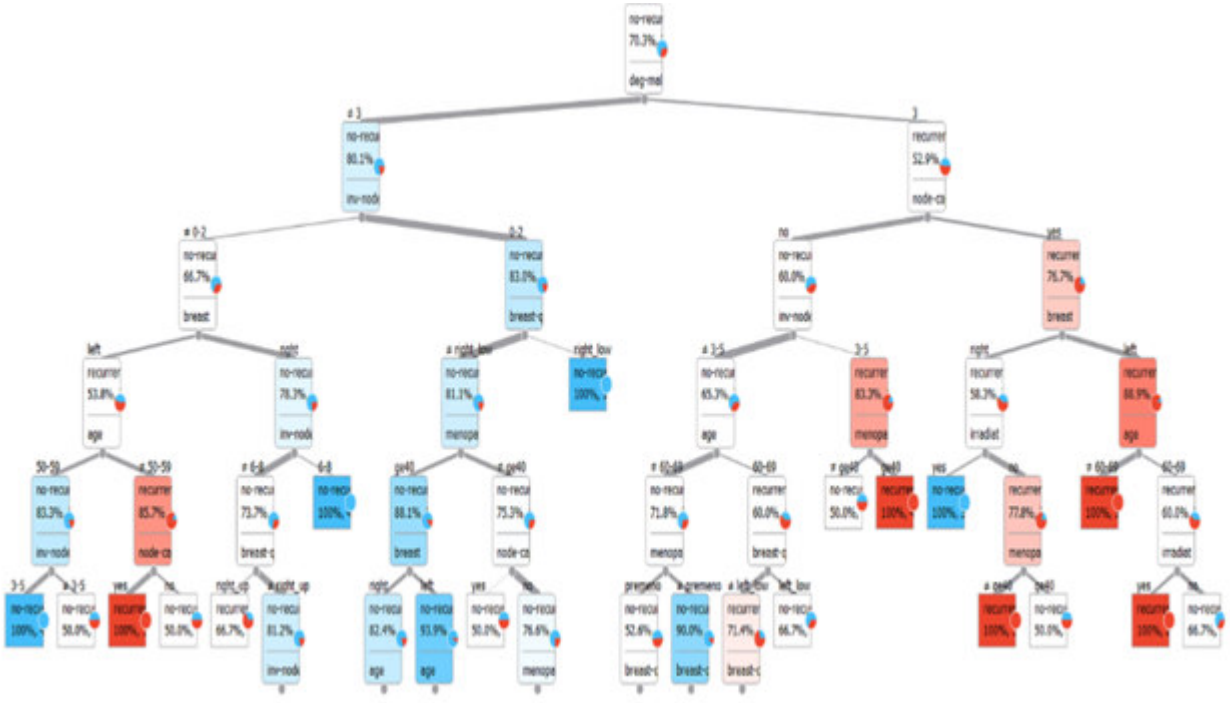

**Figure 4**
***Classification tree analysis of the dataset***

A Classification tree is a basic representation for grouping illustrations. In this area, the majority of the elements has limited discrete areas, and there is a solitary target highlight called the grouping. Every component of the area of the arrangement is known as a class. A choice tree or a characterization tree is a tree in which each interior (non-leaf) hub is marked with a data highlight. The bends originating from a hub named with an element are named with each of the conceivable estimations of the component. Every leaf of the tree is named in a class or a likelihood circulation over the classes. A tree can be "scholarly" by part the source set into subsets in view of a property estimation test. This procedure is rehashed on each inferred subset in a recursive way called recursive apportioning. The recursion is finished when the subset at a hub has all the same estimation of the objective variable, or while part no more increases the value of the forecasts. This procedure of top-down affectation of choice trees is a case of an eager calculation.

### *5.1.1 Comparative analysis between Naïve Bayes, SVM and Classification tree*

#### *5.1.1.1 Confusion Matrix*

A confusion Matrix, otherwise called an error framework, is a particular table format that permits representation of the execution of a calculation, ordinarily a directed learning one. Every segment of the framework speaks to the cases in an anticipated class while every line speaks to the occasions in a genuine class (or the other way around). It is a unique sort of possibility table, with two measurements – "genuine" and "anticipated", and indistinguishable arrangements of "classes" in both measurements. All right conjectures are situated in the corner to corner of the table, so it's anything but difficult to outwardly assess the table for mistakes, as they will be spoken to by qualities outside the slanting.

| | | Predicted | | |
|---|---|---|---|---|
| | | no-recurrence-events | recurrence-events | ∑ |
| Actual | no-recurrence-events | 174 | 27 | **201** |
| | recurrence-events | 61 | 24 | **85** |
| | ∑ | **235** | **51** | **286** |

**Table 2.1**
***Confusion Matrix for Classification Tree***
***This confusion matrix says that out of 201 non-recurring events, 174 were predicted correctly. And out of 85 recurring events, only 24 were predicted correctly.***

| | | Predicted | | |
|---|---|---|---|---|
| | | no-recurrence-events | recurrence-events | ∑ |
| Actual | no-recurrence-events | 173 | 28 | **201** |
| | recurrence-events | 48 | 37 | **85** |
| | ∑ | **221** | **65** | **286** |

**Table 2.2**
***Confusion Matrix for Naïve Bayes***
***This confusion matrix says that out of 201 non-recurring events, 173 were predicted correctly. And out of 85 recurring events, 37 were predicted correctly.***

| | | Predicted | | |
|---|---|---|---|---|
| | | no-recurrence-events | recurrence-events | ∑ |
| Actual | no-recurrence-events | 190 | 11 | **201** |
| | recurrence-events | 79 | 6 | **85** |
| | ∑ | **269** | **17** | **286** |

**Table 3.2**
***Confusion Matrix for SVM***

This disarray grid, says that out of 201 non-repeating occasions, 190 were anticipated accurately. What's more, out of 85 repeating occasions, just 6 were anticipated effectively. As can be seen from the disarray grid examination, SVM Predicts the best non repeat occasions, yet it is a disappointment at identifying repeat occasions, while Naïve Bayes predicts the best repeat occasions, however is a stage ahead to order tree on distinguishing non - repeat occasions.

### *5.1.2 ROC ANALYSIS*

A receiver operating characteristics (ROC), or ROC bend, is a graphical plot that represents the execution of a parallel classifier framework as its segregation limit is fluctuated. The bend is made by plotting the genuine positive rate (GPR) against the false positive rate (FPR) at different limit settings.

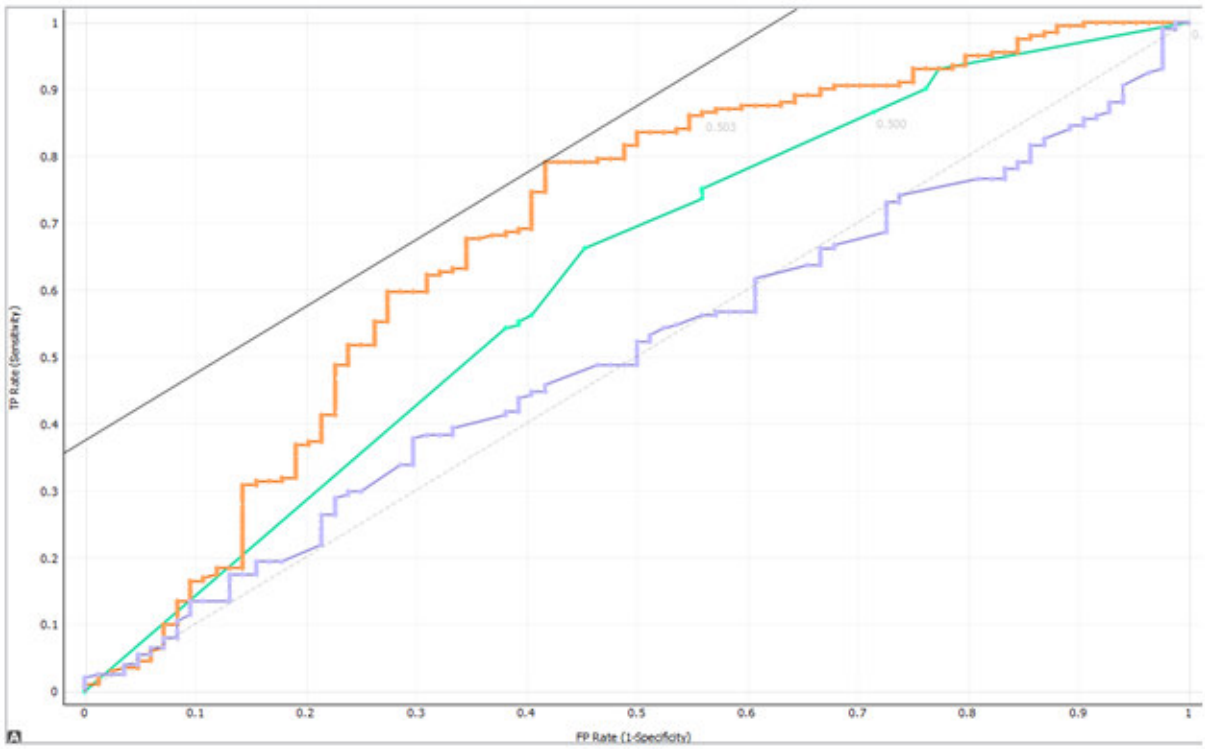


**Figure 5**
***ROC comparative analysis of SVM, Naïve Bayes and Classification Tree***

| Classifier Name | Error Rate | Accuracy |
|---|---|---|
| Naïve Bayes | 0.0558 | 90.42 |
| Classification Tree | 0.0236 | 93.62 |
| SVM | 0.0596 | 89.97 |

**Table 3**
***Error and Accuracy analysis***
***This table looks at the different calculations, we have utilized as a part of our examination in light of their Error Rate and Accuracy. As can be seen the Classification Tree turns out to be best in precision.***

## DISCUSSION

Early detection of breast cancer is the challenging task of physicians in most of the developing countries like India. Various data mining techniques are now available to detect cancer at its earliest stage without going for surgical biopsy. It is also very effective in cancer prognosis in such a way it can predict its recurrence in future. In this paper classification tree, SVM and Naïve Bayes were used for cancer diagnosis and prognosis. The decision tree classifies the patients according to their stage which is exactly like that of human knowledge in cancer diagnosis. Choice tree is developed with best indicator in cancer diagnosis, which can be useful in electronic application development in the future. Bayesian analysis is deliberately used in the medical field than other techniques.Diagnostic procedures prevail for cancer diagnosis is time consuming and common people cannot afford it.They can easily avail these techniques since it is less time consuming and easy to afford compare to existing techniques.

## CONCLUSION

Enhancement in breast cancer diagnosis and prognosis is achieved using computational symptomatic devices. Different information mining procedures have been broadly utilized for better bosom disease decision making. Choice tree is observed to be the best indicator with 93.62%. In future the indicator can be utilized to plan an electronic application to acknowledge the indicator variables and mechanized framework. The decision tree is deliberately used to classify the patients based upon their disease stage and it imitates typically like human beings in disease prediction. The Bayesian system is likewise observed to be a well-known procedure in medicinal forecast Particular it has been effectively used for Breast tumour guess and determination. In the future, we mean to plan and actualize such framework for online applications.

## ACKNOWLEDGEMENT

The work is partially supported by Himanshu Singh and Aishwarya Singh of VIT students. The authors would like to thank the project partners for their valuable support.

## REFERENCES


1. Tarver T. Cancer Facts & Figures 2012. American Cancer Society (ACS) Atlanta, GA: American Cancer Society, 2012. 66 p., pdf. Available from. Journal of Consumer Health on the Internet. 2012 Jul 1;16(3):366-7.
2. Young JL, Percy CL, Asire AJ. Surveillance epidemiology and end results: incidence and mortality data 1973-77.
3. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine. 2005 Jun 30;34(2):113-27.
4. Bellaachia A, Guven E. Predicting breast cancer survivability using data mining techniques. Age. 2006;58(13):10-10.
5. Cox DR, Oakes D. Analysis of survival data. CRC Press; 1984 Jun 1.
6. Ada RK. Using Some Data Mining Techniques to Predict the Survival Year of Lung Cancer Patient.2013.
7. Krishnaiah V, Narsimha DG, Chandra DN. Diagnosis of lung cancer prediction system using data mining classification techniques. International Journal of Computer Science and Information Technologies. 2013;4(1):39-45.
8. Edeki C, Pandya S. Comparative Study of Data Mining and Statistical Learning Techniques for Prediction of Cancer Survivability. School of Business and Technology. 2012 Nov.
9. Mokhtar SA, Elsayad A. Predicting the Severity of Breast Masses with Data Mining Methods. arXiv preprint arXiv:1305.7057. 2013 May 30.
10. Cios KJ, Moore GW. Uniqueness of medical data mining. Artificial intelligence in medicine. 2002 Oct 31;26(1):1-24.
11. Dash B, Mishra D, Rath A, Acharya M. A hybridized K-means clustering approach for high dimensional dataset. International Journal of Engineering, Science and Technology. 2010;2(2):59-66.
12. Chauhan R, Kaur H, Alam MA. Data clustering method for discovering clusters in spatial cancer databases. International Journal of Computer Applications (0975–8887) Volume. 2010 Nov.
13. Chen D, Xing K, Henson D, Sheng L, Schwartz AM, Cheng X. Developing prognostic systems of cancer patients by ensemble clustering. BioMed Research International. 2009 Jun 23;2009.
14. Halawani SM, Alhaddad M, Ahmad A. A study of digital mammograms by using clustering algorithms. Journal of Scientific and Industrial Research. 2012 Sep 1;71(9):594
15. Zubi ZS, Saad RA. Improves Treatment Programs of Lung Cancer Using Data Mining Techniques. Journal of Software Engineering and Applications. 2014 Feb 1;7(2):69.
16. Hankey BF, Ries LA, Edwards BK. The surveillance, epidemiology, and end results program a national resource. Cancer Epidemiology Biomarkers & Prevention. 1999 Dec 1;8(12):1117-21.
17. Xiong X, Kim Y, Baek Y, Rhee DW, Kim SH. Analysis of breast cancer using data mining & statistical techniques. InSoftware Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2005 and First ACIS International Workshop on Self-Assembling Wireless Networks. SNPD/SAWN 2005. Sixth International Conference on 2005 May 23 (pp. 82-87). IEEE.