*January-February 2020*
*ISSN: 0193 - 4120 Page No. 10896- 10899*

# Data Mining Technique for Cancer Prediction

**[1]A. Bharat Kumar, [2]S. Sridhar**
[1, 2]Department of Computer Science and Engineering, Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences,
[1]arunbharat98@gmail.com

*Article Info*
***Volume 82***
***Page Number: 10896- 10899***
***Publication Issue:***
***January-February 2020***

## Abstract

Recognizable proof of breast malignant growth assumes a significant job in medicinal field these days. Ladies are confronting various sorts of malignant growth and one among them is breast disease which has serious effect. Breast malignant growth is of two sorts for example Defame or Benign sort. Favorable is given as a non-repairable kind of malignancy and Malign is given as treatable sort of disease. Breast disease is symbolized by the adjustment of qualities, steady torment, changes in the estimation, change in shade(redness), skin appearance of breasts. In the beginning of distinguishing breast malignant growth is finished by utilizing various calculations specifically Support Vector Machine (SVM) calculation, K Nearest Neighbor (KNN) calculation, MLP calculation, and so forth., By utilizing these calculations the exactness of recognizing the disease isn't met the broaden. Our thought is to recognize the breast malignancy utilizing Decision Tree calculation. The choice tree calculation goes under the administered learning strategy. Our thought is to distinguish the breast malignant growth utilizing Decision Tree calculation. The tree calculation goes under the administered learning procedure. The fundamental preferred position of this choice tree calculation is distinguishing whether the anticipated disease is either insult or kind sort by delivering a 99% precision. Breast malignant growth is a kind of disease that starts in the breast. Malignancy begins when cells start to develop crazy. Breast disease cells for the most part structure a tumor that can regularly be seen on a x-beam or felt as an irregularity. Breast malignant growth happens essentially in ladies, yet men can get breast disease, as well. It's essential to comprehend that most breast knots are considerate and not disease (dangerous). Non-malignant breast tumors are unusual developments, yet they don't spread outside of the breast. They are not hazardous, however a few kinds of kindhearted breast knots can build a lady's danger of getting breast malignant growth. Any breast knot or change should be checked by a human services proficient to decide whether it is amiable or harmful (malignant growth) and in the event that it may influence your future disease hazard. See Non-dangerous Breast Conditions to find out additional. Breast disease is an exceptionally forceful kind of malignant growth with low middle endurance. Precise guess expectation of breast malignant growth can save countless patients from getting pointless adjuvant foundational treatment and its related costly medicinal expenses. Past work depends generally on chosen quality articulation information to make a prescient model. The development of profound learning strategies and multi-dimensional information offers open doors for increasingly far reaching examination of the sub-atomic attributes of breast malignant growth and along these lines can improve conclusion, treatment and counteractive action.

***Article History***
***Article Received****: 18 May 2019*
***Revised****: 14 July 2019*
***Accepted****: 22 December 2019*
***Publication****: 19 February 2020*

***Keywords:*** *Breast cancer; common among women; types; malign or benign; characterized; persistent pain, skin appearance, change in measurement; existing algorithms; SVM, KNN, MLP, etc; proposed algorithms; Decision tree algorithm; supervised learning;*

## 1. Introduction

Breast cancer is the most highly aggressive cancer and a major health problem in females, and a leading cause of cancer-related deaths worldwide. According to the estimates of American Cancer Society, more than 250,000 new cases of invasive breast cancer will be

TEST Engineering & Management

*January-February 2020*
*ISSN: 0193 - 4120 Page No. 10896- 10899*

diagnosed among females and approximately 40,000 cancer deaths expected in 2017. This heterogeneous disease is characterized by varied molecular feature, clinical behavior, morphological appearance and disparate response to therapy. Also, the complexity among invasive breast cancer and its significantly varied clinical outcomes now make it extremely difficult to predict and treat. Therefore, to the ability of predicting cancer prognosis more accurately not only could help breast cancer patients know about their life expectancy, but also help clinicians make informed decisions and further guide appropriate therapy. Meanwhile, prognostication plays an important role in clinical works for all clinicians, particularly those clinicians working with short term survivor. When a reasonably accurate estimation of prognosis is available, clinicians often utilize prognosis prediction knowledge to assist with clinical decision making, establish patients' eligibility for care programmes, design and analysis of clinical trials. In addition, when patients are predicted to be short term survivors, clinicians can provide patients with the opportunity to consider whether they want to be cared for and allow them time to take practical steps to prepare for their own deaths.

In our paper we are utilizing information science and Machine learning idea. These days innovation are building up a great deal. So as to decrease human work we are proposing an idea of Machine learning and information science. Data Science is a blend of various mechanical assemblies, estimations, and Machine learning models with the goal to discover hid models from the unrefined data. Data Science is basically used to choose decisions and gauges making use of judicious causal assessment, prescriptive examination and Machine learning. Data Science is an inexorably forward-looking procedure, an exploratory way with the consideration on breaking down the past or current data and anticipating the future outcomes with the purpose of choosing taught decisions. It reacts to the open-completed requests about "what" and "how" events occur. As information science comprises of Machine learning idea in it, we are attempting to push ahead with it to nourish the machine. Machine learning is where the machine adapts consequently with understanding. Learning is only memory and absorption of information and the choice absolutely relies upon prepared information. It is strenuous to take choice dependent on achievable information sources. To beat this issue, certain calculations were created. So as to understand a particular errand feed the calculation with increasingly explicit information. By and large a PC will utilize information as its wellspring of data and contrast its yield with an ideal yield and afterward right for it. The more information or "experience" the PC gets, the better it becomes at assigned employment, similar to a human does. At the point when Machine learning is viewed as a procedure, the accompanying definition is sagacious: "Machine learning is the procedure by which a PC can work all the more precisely as it gathers and gains from the information it is given. "By utilizing this idea we are making the machine to recognize whether the malignant growth acquired is Malign or Benign kind of breast malignancy. Machine learning is grouped into regulated and unaided learning. In that administered realizing there are order and relapse part. In our paper we use choice tree calculation which goes under the characterization part. Choice tree is spoken to utilizing a tree structure. By utilizing all the datasets the qualities are split into tree structure to recognize whether it is Malign or Benign sort of breast disease.

## 2. Literature Survey

In 2016, Moh'd Rasoul Al-hadidi, Abdulsalam Al arabeyyat, Mohannad Alhanahnah proposed a paper on recognizing Breast malignancy utilizing Machine learning calculation. In this paper they utilized picture handling ideas and two Machine learning calculations. The calculations are Logistic Regression (LR) and Back Propagation Neural Network (BPNN) to recognize the breast malignancy. They have utilized 209 pictures of 50 patients and every one of those pictures were in dark scale design. The pictures were changed from time space to recurrence area by Discrete Wavelet Transformation. By doing so the yield had 4 grids. In these networks just 3 were utilized. The estimation of the grids were given as contribution to the calculation. The yield was 0 for typical pictures and 1 for tumor images. Thus the precision rate for identifying the malignancy utilizing LR and BPNN calculation surpassed 93%.

In 2016, Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel delivered a paper on hazard forecast and determination of breast malignancy utilizing Machine Learning (ML) calculation. They have utilized the accompanying ML calculation, Bolster Vector Machine (SVM), Naïve Bayes (NB), K- Closest Neighbor (KNN), Decision Tree calculations. They characterize by information mining technique. In this paper they have utilized Wisconsin Breast Cancer dataset which contains 699 data's. For ordering and assessing the information they have utilized 10 crease cross approval test which parts the information for preparing and testing stage. Along these lines the outcome acquired is thought about between all the ML calculations utilized. They have closed saying among every one of the calculations they have utilized SVM gave the most elevated exactness pace of 97.13%.

In 2018, Meriem Amrane, Saliha Oukid, Ikram Gagaoua, Tolga Ensar proposed a paper on Breast Cancer gathering using Machine learning computations. They have used Guileless Bayes estimation and K-Nearest Neighbor figuring to portray the threat as either affront or ideal. They utilized 11 traits in which they had id as one characteristic which has been evacuated and accordingly have 9 criteria. The breast malignant growth arrangement they done had 9 order which included Clump Thickness, Uniformity of cell size, Consistency of cell shape, Marginal Adhesion, Single Epithelial cell size, Bare

*January-February 2020*
*ISSN: 0193 - 4120 Page No. 10896- 10899*

Nuclei, Bland Chromatin, Normal Cores, Mitosis. They had 683 datasets. The order is done utilizing the separate formulae for the calculations. By looking at both the calculations they got a precision of 97.51% in K-Nearest Neighbor calculation and 96.19% in Innocent Bayes calculation.

In 2018, Arjun P. Athreya, Alan J. Gaglio, Junmei Cairns, Krishna R. Kalari, Richard M. Weinshilboum, Liewei Wang, Zbigniew T. Kalbarczyk, Ravishankar K. Iyer proposed a paper on recognizing the Breast Cancer utilizing the human genome utilizing Machine Learning (ML) calculation. In the paper they have utilized solo learning calculation. They have utilized metformin and TNBC for distinguishing proof. In unaided learning they have utilized grouping calculation. Hence they have recognized the triple negative breast malignant growth utilizing the KNN grouping calculation.

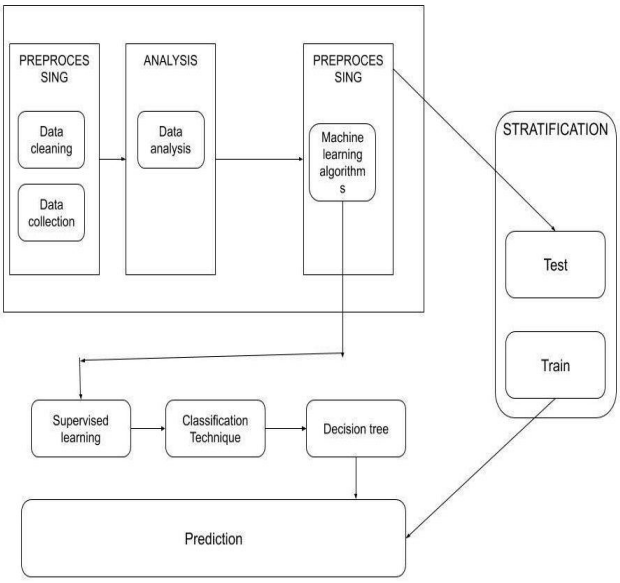
## 3. Proposed System

**Dataset:**

The dataset we are using is Wisconsin Breast Cancer dataset with 32 attributes and 569 data's. The inputs are (1) id -(ID number), (2) diagnosis-(M = malignant, B = benign), (3) radius-(mean of distances from center to points on the perimeter), (4) texture-(standard deviation of gray-scale values), (5) perimeter-(mean size of the core tumor), (6) area, (7) smoothness mean-(mean of local variation in radius lengths), (8) concavity- (mean of severity of concave portions of the contour), (9) compactness-(mean of perimeter^2 / area - 1.0), (10) concave_points mean-(mean for number of concave portions of the contour), (11) symmetry, (12) fractal_dimension mean –(mean for "coastline approximation" – 1).Here, (1) is not considered and the rest are taken into account. For these attributes they calculate mean, standard error, worst. These data's are fed into the machine. These data's are used for training the machine using Decision tree algorithm. By doing so we can get the output as either Malign or Benign. Using the Decision tree algorithm the breast cancer is identified. The decision tree algorithm is represented using a tree structure. By using this algorithm 99% accuracy is produced to identify whether the breast cancer is either Malign or Benign. The decision tree algorithm comes under the supervised learning technique.

**Stages:**

Stage 1: Dataset Collection:
Collection of dataset related to breast cancer.
Stage 2: Data Cleaning:
The attributes are checked for any null value process, if any null values are present then replace it with zero.
Stage 3: Data Analysis:
The data collected as been processed and helps in making any decision.
Stage 4: Supervised Learning:
The Machine learning errand of learning a capacity that maps a contribution to a yield dependent on model info yield sets.
Stage 5: Classification Technique:
In supervised learning, there comes a regression and classification part.
Stage 6: Decision Tree Algorithm:
This algorithm is normally represented in a tree structure. By using this algorithm we are providing an accuracy of 99% in our project.
Stage 7: Prediction:
By producing a 99% accuracy we are predicting the type of breast cancer easily.




**Requirements:**

Anaconda is data science and machine learning platform for the Python and R programming languages. It is designed to make the process of creating and distributing projects simple, stable and reproducible across systems and is available on Linux, Windows, and OSX. Anaconda is a Python based platform that curates major data science packages including pandas, scikit-learn, SciPy, NumPy and Google's machine learning platform, TensorFlow. It comes packaged with conda (a pip like install tool), Anaconda navigator for a GUI experience, and spyder for an IDE. Anaconda is an open source circulation of the Python and R programming dialects and it is utilized in information science, Machine learning, profound learning-related applications going for improving bundle the board and organization. Moreover, Python and markup language are required on the code aspect to develop the appliance.

**Software Requirements:**

Operating system: Mac or Windows or Linux
Coding language: Python, R
Tool: Anaconda

*January-February 2020*
*ISSN: 0193 - 4120 Page No. 10896- 10899*

**Hardware Requirements:**

Processor: intel pentium
RAM: 8 GB
ROM 4 GB

**Future Scope:**

This framework might be expanded by including additional information of the influenced understanding like including some more traits. Further, we will be upgrade our task by distinguishing at which organize it is present and furthermore giving the preventive measures to the tolerant.

## 4. Conclusion

Different machine-learning techniques can be used for the prediction of breast cancer. The challenge is to build accurate and computationally efficient medical data classifiers. The distinguishing proof of breast malignant growth is done here. The distinguishing proof depends on whether the patient is influenced by either Malign or Benign sort of malignancy. The kind of disease is anticipated by utilizing choice tree calculation which goes under the administered learning system, by doing so the exactness of 99% is gotten. With the great development of machine learning, precision medical treatment is put forward to effectively treat early-stage breast cancer patients and reduce recurrence risk. We need to use more effective, accurate and scientific methods to guide the treatment process. But on the other hand, more clinical or omics data should be introduced to improve the model. In future works, additional clinical data will be collected for improving accuracy of the prediction algorithm.

## References

[1] H. J. Pandya, K. Park, W. Chen, L. A. Goodell, D. J. Foran, "Toward a portable cancer diagnostic tool using a disposable MEMS-based biochip", IEEE Trans. on Biomedical Engineering, vol. 63, no. 7, pp. 1347-1353 , 2017

[2] R. Siegel, J. Ma, Z. Zou, A. Jemal, "Cancer statistics", CA Cancer J. Clin., vol. 64, no. 1, pp. 9–29, Jan.-Feb. 2014.

[3] R. Lin, P. Tripuraneni, "Radiation therapy in early-stage invasive breast cancer", Indian J. Surg. Oncol., vol 2, no. 2, pp. 101–111, Jun. 2011.

[4] M. A1-Badrashiny, A. Bellaachia, "Breast cancer survivability prediction via classifier ensemble", Inter. J. of Computer and Information Engineering, vol. 10, no. 5, pp. 833-837, 2016.

[5] R. K. Kavitha, D. Dorairangasamy, "Breast cancer survivability predictor using Adaboost and CART algorithm", Inter. J. of Innovative Research in Science, Engineering and Technology, vol. 3, no. 1, pp. 351-353, 2014. vol. 21, no. 4, pp. 1133-1145, 2017.

[6] J. kim, H. Shin, "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data", J. of the American Medical Informatics Association, vol. 20, no. 4, 2013.

[7] E. Porter, M. Coates, M. Popovic, "An early clinical study of time-domain microwave radar for breast health monitoring", IEEE Trans. on Biomedical Engineering, vol. 63, no. 3, pp. 530-539, 2016.

[8] J. Yoon, C. Davtyan, M. Schaar, "Discovery and clinical decision support for Personalized Healthcare".