IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 8, May 2022
Impact Factor: 6.252

# Breast Cancer Detection Using Machine Algorithms

**Mr. R. Ramarajesh[1], S. Divya[2], A. J. Louisa Merline[3]**
Assistant Professor, Department of Information Technology[1]
Students, Department of Information Technology [2,3]
Anjalai Ammal Mahalingam Engineering College, Thiruvarur, India

**Abstract**: *The most frequently occurring cancer among Indian women is breast cancer. There is a chance of fifty percent for fatality in a case as one of two women diagnosed with breast cancer deaths in the cases of Indian women. This paper aims to present a comparison of the largely popular machine learning algorithms and techniques commonly used for breast cancer prediction, namely Random Forest, KNN (k-Nearest-Neighbor), Support Vector Machine, and XG Boost techniques. The Wisconsin Diagnosis Breast Cancer data set was used as a training set to compare the performance of the various machine learning techniques in terms of keyparameters such as accuracy and precision. The results obtained are very competitive and can be used for detection and treatment. Breast cancer disease causes a massive number of deaths in the world. After the traditional cancer detection methods, the latest technologies enable experts with numerous adaptive methods to discover breast cancer in women. Breast cancer affects the majority of women worldwide, and it is the second most common cause of death among women. Breast cancer is among the most serious illnesses/diseases in India, causing many deaths in the current situation. Due to changes in food and lifestyle, the number of cancer cases in women is increasing day by day. Different types of machine learning are implemented for the prediction of breast cancer with a high accuracy rate. We develop different machine learning algorithms for the prediction of breast cancer.*

**Keywords**: Machine Learning, Breast Cancer, random forest, k-Nearest- Neighbor, XG Boost Technique, Support Vector Machine.

## I. INTRODUCTION

Machine learning is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks. A subset of machine learning is closely related to computational statistics which focuses on making predictions using computers, but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory, and application domains to the field of machine learning. Data Mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. Some implementations of machine learning use data and neural networks in a way that mimics the working of a biological brain. In its application across business problems, machine learning is also referred to as predictive analysis.

## II. DESIGN ISSUES

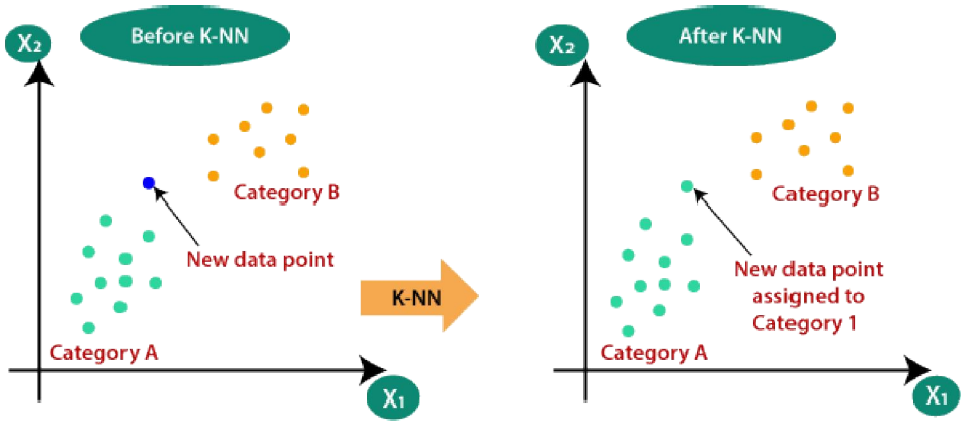
The challenge is aimed to make use of a machine learning algorithm in Breast Cancer to evaluate the accuracy and stage of cancer using a dataset. In this project, we aim to impart the ability to get rid of biases in a machine algorithm and to predict the accuracy of the datasets. By representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem health issue in today's society. Classification and data mining methods are effective way to classify data. Especially in the medical field, where those methods are widely used in diagnosis and analysis to make decisions

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 8, May 2022
Impact Factor: 6.252

### 2.1 Algorithm Used

**A. K-Nearest Neighbour**

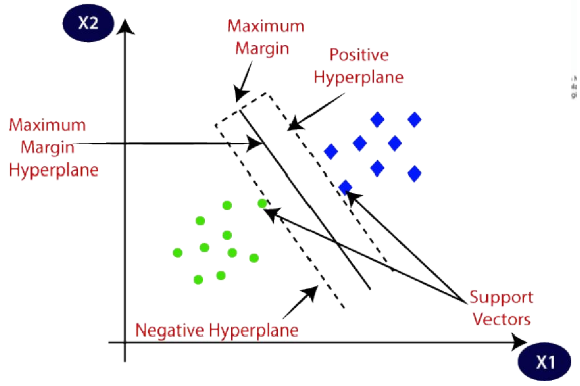
- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on the Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for classification problems.




**B. Random Forest Classification**

- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning.
- The random forest is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by the committee is more accurate than that of any individual tree.

The random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output.




## III. DEVELOPMENT MODEL

The first stage of the development of Machine Learning models in the preparation of the dataset. In this stage, the collected dataset shall be divided into two groups, training, and testing. The training and testing dataset is used for the calibration

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 8, May 2022
Impact Factor: 6.252

and validation of applied models, respectively. Depending on the simulation conditions regarding time series modeling or function fitting, the approaches to assigning a dataset for each group are different. In time series modeling, the history of collecting data shall be considered, and shuffling the dataset is not correct, whereas for function fitting using data shuffling idea is allowed. Usually, for both scenarios, about 70%–80% of the dataset is assigned for calibration and the remaining 20%–30% forvalidation. The next step for developing the ML models, such as K-Nearest Neighbor, Random forest classification, XG Boost Algorithm, and Support Vector Machine in ML is designing the architecture of the network.

## IV. TESTING ANALYSIS

We are going to implement Breast cancer detection using Machine Learning Algorithms. In this project, we are comparing four algorithms namely K-Nearest Neighbor, Random forest classification, XG Boost, and Support Vector Machine and which are efficient and have high accuracy levels in predicting breast cancer which is benign or malignant.

| ALGORITHM | ACCURACY LEVEL |
|---|---|
| K-Nearest Neighbor Algorithm | **91.5** |
| Random Forest Classification | **91.3** |
| XG Boost Techniques | **90.5** |
| Support Vector Machine | **91.8** |

## V. RESULT AND DISCUSSION

For validating the developed model, the dataset has been divided into 70% training and 30% testing subsets. We predict the accuracy of breast cancer by comparing four algorithms and which algorithm has the highest level of accuracy to predict cancer benign or malignant.

## VI. CONCLUSION

In this paper, the comparison of K-nearest neighbor, Random Forest Classification, XG Boost, and Support Vector Machine in supervised learning evaluates the accuracy of breast cancer. At the end of this paper, We came to the conclusion that XG Boost Algorithm predicts a high accuracy level to predict breast cancer.

## REFERENCES

**[1].** Epidemiology of Breast Cancer in Indian Women Author: Malvia S, Bagadi S A, Dubey U.S, Saxena S Year:2017
**[2].** Breast Cancer Identification and Diagnosis Techniques Author: Anji Reddy V, Soni Badal Year:2020
**[3].** Assessment of Ki67 in Breast Cancer: Updated Recommendations from the International Ki67 in Breast Cancer Working Group Author: Torsten O Nielsen, Samuel C. Y Leung, David L Rimm Year:2021
**[4].** The lingering mysteries of metastatic recurrence in breast cancer Author:: Alessandra I. Riggio, Katherine E. Varley & Alana L. Welm Year:2021
**[5].** Epigenetic mechanisms in breast cancer therapy and resistance Author: Liliana Garcia-Martinez, Yusheng Zhang, Yuichiro Nakata, Ho Lam Chan
**[6].** Global patterns of breast cancer incidence and mortality author: Shaoyuan Lei, Rongshou Zheng, Siwei Zhang, Shaoming Wang, Ru Chen, Kexin Sun, Hongmei Zeng, Year:2021
**[7].** Metabolic pathways in obesity-related breast cancer Author: Kristy A. Brown Year:2021
**[8].** Estimation of Breast Cancer Over diagnosis in a U.S. Breast Screening Cohort Author: Marc D. Ryser, Jane Lange.Lurdes Y.T. Inoue Year:2021