# Breast Cancer Prediction Using Machine Learning Models

**Mula RamCharan Reddy [a], A.M Surendra Kumar [b], Tummala Mohith Charana Sai [c], Devanjali Relan[d]**

[a] Student, India, mula.reddy.19cse@bmu.edu.in

[b]Student,Gurugram, India, akshitala.kumar.19cse@bmu.edu.in

[c]Student,Gurugram, India, tummala.sai.19cse@bmu.edu.in

[d] Assistant Professor, Gurgaon, India, devanjali.relan@bmu.edu.in

## 1. Abstract

Cancer deaths among women are primarily caused by breast cancer. It occurs when threatening bumps start to form from the breast cells, and most diagnoses occur at the later stages, so patients have low survival rates. For early detection and prognosis, it is important to recognize if bumps are benign or threatening. The goal of this study is to forecast breast cancer, which is the second biggest cause of deaths among women globally, and may be drastically reduced with early identification and prevention. On the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which is taken from a digitized picture of an MRI, this research proposes a comparison of six machine learning (ML) algorithms: Decision Tree(DT), Random Forest(RF), X G Boost, Gaussian Naive Bayes (NB), Support Vector Machine(SVM) and K-Nearest Neighbour (KNN). The results of the experiments suggests that Random Forest has the highest accuracy of 96 percent. This discovery will aid in the selection of the optimal machine-learning classification method for breast cancer prediction.

## 2. Keywords

Breast Cancer; Classification; Machine Learning; Supervised Learning; Wisconsin Diagnostic Breast Cancer (WDBC); Decision Tree(DT); Random Forest(RF); X G Boost; Gaussian Naive Bayes (NB); Support Vector Machine(SVM); K-Nearest Neighbour (KNN); Efficiency

## 3. Introduction

Breast cancer, along with lung and bronchus cancer and pancreatic cancer, is one of the most frequent cancers. It accounts for 15% of all new cancer cases in the United States alone, making it an important study issue. Over the previous few decades, there has been a steady development in cancer research. Researchers used various procedures, such as screening, to determine the stage of cancer before symptoms appeared. Experts have also created new methods for predicting the outcome of cancer treatment early enough. With the development of new methods, precise cancer prediction has become one of the most difficult and interesting undertakings for physicians.

According to the World Health Organization (WHO), one of the most important issues in medical field research is the breast cancer diagnosis. Every year, the number of instances recorded are rising. Survivability and recurrence (return of cancer after therapy) are two features of breast cancer. These are important aspects of breast cancer behaviour that are intrinsically linked to patient death. After lung cancer, breast cancer is the second leading cause of death among women. In compared to the United States, the number of newly diagnosed women with breast cancer in India is lower, but the number of fatalities from breast cancer is much greater. As a result, early detection of breast cancer is critical.

The use of data science and machine learning techniques in medical sectors has become increasingly common, as these techniques may greatly aid medical practitioners in their decision-making. With the terrible rise in breast cancer cases, comes a large amount of data that may be used to further clinical and medical research, as well as the use of data science and machine learning in the mentioned domain. The purpose of this paper is to compare the performance of five classifiers: Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), X-G Boost, and Naives Bayes, which are among the most influential data mining algorithms and among the top 10 data mining algorithms, according to the research community. Our goal is to use machine-learning algorithms to detect and diagnose breast cancer, and to determine which is the most successful based on the confusion matrix, accuracy, precision, and sensitivity of each classifier.

The remainder of this work is structured as follows: The Section-2 introduces the methodology and findings of earlier breast cancer diagnostic studies. The recommended methodology for our research is described in Section 3. The results of the experiments are presented and explained in detail in Section 4.Section 5 concludes the paper and Section-6 consists of references.

## 4. Related Works:

This section looks at some of the most recent research publications on breast cancer prognosis and diagnosis applying various optimization techniques. Multiple classifiers and feature selection strategies are used in several experiments on medical data sets. The literature contains a lot of information on breast cancer datasets. Many of them have a good level of classification accuracy.

For the diagnosis of breast cancer, Fondon has proposed an automatic classification of tissue malignancy. The second largest cause of cancer death among women was breast cancer. To avert preventable fatalities, they needed to be diagnosed as soon as possible. The evaluation of malignancy in tissue samples, however, was challenging and subjective.

Madhuri Gupta and Bharath Gupta used supervised machine learning techniques to conduct a comparison study on breast cancer diagnosis. Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Decision Tree (DT) are four extensively used machine learning approaches that are compared. MLP outperforms other methods in terms of accuracy.

Abien Fred M. Agarap has presented a study that used a variety of machine learning algorithms to diagnose breast cancer, including the proposed GRU-SVM model. The binary classification of breast cancer, i.e. determining whether the tumour is benign or malignant, was performed well by all of the given ML algorithms. As a result, the categorization problem's statistical measurements were good as well.

Raghavendra proposed that ultrasonic imaging be used to increase the papillary index of the breast in order to distinguish between benign and malignant tumours. Breast papillary lesions can be benign or cancerous. Breast papillary lesions occur with a wide range of radiographic characteristics, making it difficult to distinguish between benign and malignant lesions based on imaging characteristics. A large number of ultrasound pictures of papillary breast lesions were used to test the developed model.

Reetodeep Hazra, Megha Banerjee and Leonardo Badia have used WDBC data set to classify breast cancer using two well-known machine learning frameworks - ANN and DL. Feature selections are performed in the suggested classifiers to exclude statistical features from the data set, and performance comparisons are made to find the most acceptable methodology for conclusion.

Bhardwaj has presented a GONN (genetically optimised neural network) technique for breast cancer categorization. By incorporating the unique crossover and mutation operators of GA, this breast cancer classification system was able to optimise the artificial neural network parameter. The breast cancer classification approach becomes more complicated as the calculations become more intricate.

The automatic characterization of malignant breast lesions was presented by Acharya. Shear wave elastography (SWE) is used in this approach to measure discrete wave coefficients at three separate levels. As coefficients, these characteristics were deleted. With successive forward selection approaches, the important traits were removed and classified using the Relief classification system.

For feature weighting and parameter optimization, Phan used a hybrid model of GA and SVM. When compared to Grid Search, the GA-SVM model achieves a significant improvement in sort performance throughout the whole data sets. Genetic algorithms (GA) are effective techniques for solving nonlinear optimization problems on a large scale.

## 5. Methodology

### 5.1 Dataset:

Wisconsin Breast Cancer Database(WBCD) is a breast cancer dataset taken from the UCI machine learning repository dataset [11]. This dataset contains 569 cases that are classified as benign or malignant, with 357 cases (62.74 percent) being benign and 212 cases (37.25 percent) being malignant. The dataset is divided into two classifications, B and M, with B denoting the benign and M denoting the malignant. Breast cancer is the most common condition in medical diagnosis, and its prevalence is rising every year. Except for sample code number and class, the dataset comprises 32 features: (1) radius, (2) perimeter, (3) area, (4) texture, (5) compactness, (6) smoothness, (7) concave points, (8) concavity, (9) symmetry, and (10) fractal dimension are all factors to consider. With three pieces of information generated for each feature: (1) standard error, (2) mean, and (3) "worst" or biggest (mean of the three largest values). As a result, there are a total of 30 dataset features.. In our study, benign occurrences are represented as positive because they have no negative effects on the body, whereas malignant cases are portrayed as negative since they are cancerous cells that have negative effects on the body. In this data collection, there are 16 missing feature values. Finally, the data set is randomized to guarantee that the data is distributed correctly.

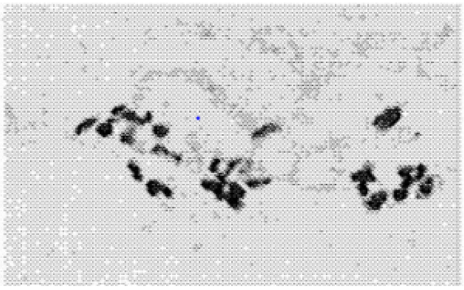


**Figure 1: Digitized image of FNA of Benign cancer.**

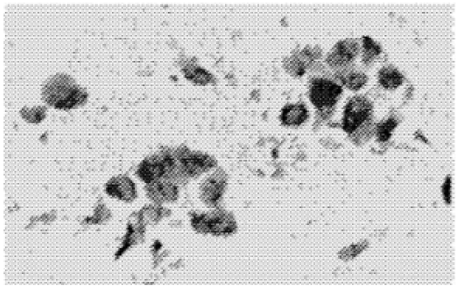


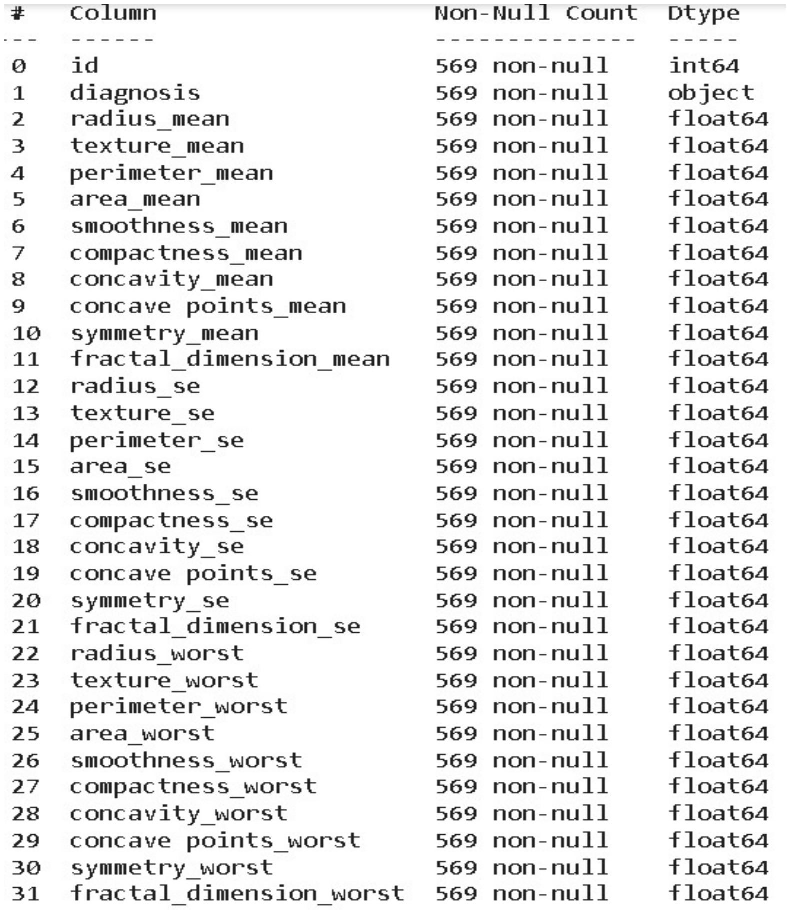
**Figure 2: Digitized image of FNA of  Malignant cancer**

Table1: Each Attribute Representation,Description and Information about the attribute.

## 5.2 Data Preprocessing:

### 5.2.1 Data Cleaning:

The data set consists of 699 samples  for each sample we are having 31 attributes. At first we checked the information of each attribute, some of them are in float type and some of them are integer type. Then we checked that weather their is any null values in our data set or not and after checking we got to know that we did not have any null values in our data.



```
#   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
0   id                       569 non-null    int64
1   diagnosis                569 non-null    object
2   radius_mean              569 non-null    float64
3   texture_mean             569 non-null    float64
4   perimeter_mean           569 non-null    float64
5   area_mean                569 non-null    float64
6   smoothness_mean          569 non-null    float64
7   compactness_mean         569 non-null    float64
8   concavity_mean           569 non-null    float64
9   concave points_mean      569 non-null    float64
10  symmetry_mean            569 non-null    float64
11  fractal_dimension_mean   569 non-null    float64
12  radius_se                569 non-null    float64
13  texture_se               569 non-null    float64
14  perimeter_se             569 non-null    float64
15  area_se                  569 non-null    float64
16  smoothness_se            569 non-null    float64
17  compactness_se           569 non-null    float64
18  concavity_se             569 non-null    float64
19  concave points_se        569 non-null    float64
20  symmetry_se              569 non-null    float64
21  fractal_dimension_se     569 non-null    float64
22  radius_worst             569 non-null    float64
23  texture_worst            569 non-null    float64
24  perimeter_worst          569 non-null    float64
25  area_worst               569 non-null    float64
26  smoothness_worst         569 non-null    float64
27  compactness_worst        569 non-null    float64
28  concavity_worst          569 non-null    float64
29  concave points_worst     569 non-null    float64
30  symmetry_worst           569 non-null    float64
31  fractal_dimension_worst  569 non-null    float64
```

**Fig-1. Checking null values.**

### 5.2.2 Ploting Correlation Matrix and Graphs:

We plotted the correlation matrix and observed that some of the attributes are highly correlated with each other and some of them are less correlated to each other. The area_mean and area_worst are highly correlated, and then we plotted the scattered plot for that and observed that data points are highly concentrated and merged. The reason is there was inconsistency over the particular range of data in area_mean and area_worst.

After plotting the box plot for attributes area_mean and area_worst . we understood that data points are concentrated over the particular range. Then after we plotted the Histogram to observe how many samples belong to Benign(37.3%) and malignant(62.7%).

### 5.2.3 Checking Duplicate Rows and Columns:

We checked wether we are having any duplicate rows or columns in our dataset. If we have any duplicates rows or columns it will affect our dataset,so we have deleted all the duplicate rows and column which are present in our dataset so that our model will perform well

### 5.2.4 Label Encoding:

We need to classify the diagnosis into Benign and malignant. we have done the label encoding i.e we replaced the Malignant with '1' and Benign with' 0' (Converted object type dataset to categorical data set M to 1 and B to 0) for all values in diagnosis column.

### 5.2.5 Feature selection: MI Score

The data set consists of 32 attributes we need to choose a good features for our models so that we will good accuracy for Training (70%) and Testing(30%). Then we decided to use MI Score method in our model because we need to remove so many irrelevant features which will be automatically deleted by using MI score if we set some target interval of the scores if we want to delete features in certain interval of MI scores It will tells us the relation between them each feature in our dataset with target variable(Diagnosis). we got very less scores for some attributes and very high for some attributes with the target variable. Then we decided the MI Score base value was 0.3 (taken the range from 0.3 to 0.4) .Then we have done cross verification with each and every model by changing the MI square values. Then we got the good results for MI square value with with greater than 0.303.

## 5.3 Performance Matrices:

The parameters used to measure the performance of machine learning algorithms are described in this section. To test the performance, a confusion matrix for the actual and projected class is created using the typical five values of TruePositive, FalsePositive, TrueNegative, and FalseNegative.

### 5.3.1 Accuracy:

Accuracy is an excellent predictor of how well the model was trained and how well it will perform in general. It may be defined as the proportion of true predictions to incorrect predictions. As a result, the following equation may be used to obtain the value of accuracy:

Accuracy = (TruePositive + TrueNegetive) / (TruePositive + FalsePositive +TrueNegative + FalseNegative)

### 5.3.2 Recall:

In general, recall, also known as sensitivity, may be defined as the ratio of true positive cases to all observations. Recall may be viewed as a metric for how well the system predicts positives and calculates expenses.

Recall = (TruePositive) / (TruePositive + FalseNegetive)

### 5.3.3 Precision:

Precision may be defined as the degree of accuracy in predicting good events. It's basically the proportion of true positives to the total number of positives. This shows the system's positive value handling capability but not its negative value handling capability.

Precision = TP / (TP + FP)

### 5.3.4 F1 Score:

Precision and Recall are weighted in this calculation. As a result, this metric takes into account both types of misleading values. When the F1 score is 1, it is deemed ideal, and when it is 0, it is considered a complete failure.

F1 Score = 2*(Precision*Recall) / (Precision + Recall)

## 5.4 Algorithms:

Breast cancer is the most common condition in the medical diagnosis field, with an annual increase. On the Wisconsin Breast Cancer Database (WBCD), a comparison of six commonly used machine learning approaches is undertaken to predict the occurrence of breast cancer:Random Forest

- Navies Bayes
- Decision Tree (DT),
- Support vector machine (SVM),
- K-nearest neighbor (K-NN).
- XG Boost

### 5.4.1 Random Forest:

It is a supervised machine learning algorithm . The bagging approach is used to train the system after creating an ensemble of decision trees. Recursion is the foundation on which this approach is built. In each iteration, a random sample of size N is selected from the data set.

The dataset has been separated into training and testing sets, with 398 training observations and 171 testing observations. The number of estimators is set at 72, ensuring that each observation is predicted at least twice. The importance of diagnosis, radius mean, texture mean, and perimeter mean is clear; the other factors have a mild influence, but none of them can be ignored in order to improve model accuracy.The accuracy of  Random forest is the higher which gives 96% accuracy

### 5.4.2 Navies Bayes:

Naive Bayes classifiers are probabilistic classifiers based on the application of Bayes theory. Although it is naive in that it believes that all characteristics are independent of one another, which is seldom the case in real-world circumstances, Nave Bayes has proven to be effective for a wide range of machine learning tasks. Seven benign observations and nine malignant observations are among the sixteen misclassified observations. The same 398 observations are utilised for training and 171 observations are used for testing, with a 94 percent accuracy.

### 5.4.3 Decision Tree:

The supervised Machine Learning (ML) approach Decision Tree (DT) is used for classification and regression analysis. The divide and conquer strategy is used to create the decision tree. It divides the division into two halves using two methods: Partitions in numbers: Typically, divisions are created using discrete values with certain constraints. Nominal partitioning: Nominal characteristics are used to create divisions. It causes the tree to divide based on the values of the attributes. The experimental analysis is carried out using the decision tree technique. The same 398 observations are utilised for training and 171 observations are used for testing, with a 92 percent accuracy.

### 5.4.4 Support Vector Machine(SVM) :

SVM stands for Support Vector Machine. It is supervised learning. It is popular because of its categorization abilities. Every data item in the SVM method is displayed as a coordinate in an n-dimensional space, where n is the total number of features used for classification and the value of each feature is represented by the data point's coordinates. The decision hyper plane in SVM is used to divide data points from different classes using the largest margin. Support Vectors are data points that are near to the hyper plane. This classification method creates non-linear decision boundaries and classifies data points that aren't represented in vector space.The experimental analysis is carried out using the decision tree technique. The same 398 observations are utilised for training and 171 observations are used for testing, with a 93 percent accuracy.

### 5.4.5 K-Nearest Neighbour (KNN):

K may be thought of as a representation of the data points for training that are near to the test data point that we'll use to discover the class. The technique used to decide where a dataset belongs based on the other data sets existing around it is known as k-nearest-neighbor. The method is a regression and classification supervised learning strategy. KNN gathers all nearby data points before processing a new data point. Key criteria in defining the distance are attributes with a high degree of variance. Regardless of labels, the kNN method discovers the k closest neighbours of N training vectors. Only one observation is misclassified as benign, and four observations are misclassified as malignant, indicating that kNN is 93% accurate.

### 5.4.6 XG-Boost:

The XGBoost algorithm, also known as Extreme Gradient Boosting, is a decision tree-based machine learning technique that improves performance through boosting. It's become one of the most successful machine learning algorithms since its inception, consistently outperforming most other algorithms including the random forest model, logistic regression and conventional decision trees. The training set consists of 398 observations, whereas the testing set consists of 171 observations, with a 94 percent accuracy.

## 6. Results and Discussion

The Breast Cancer Wisconsin diagnostic dataset was analysed using Machine Learning Algorithms. Decision Tree, Random Forest, KNN, SVM, Naive Base, and XG Boost are the six algorithms we employed. We used Confusion Matrix, Accuracy, Precision, Recall, and F1 Score as performance indicators to analyse and evaluate the models and find the optimal algorithm for breast cancer prediction.

Six different machine learning models were used, with the results stated above. Each model produces a distinct set of results in terms of accuracy, precision, recall, and f1-score. Random Forest had the highest level of accuracy, with a score of 96 percent. SVM and Naive Base came in second with a score of 94 percent for XG Boost. With 92 percent accuracy, Decision Tree was the least accurate.

## 7. Conclusion

We used six algorithms on the Wisconsin Breast Cancer Diagnostic Dataset (WBCD) to calculate, compare, and evaluate different results to identify the best machine learning algorithm that is precise, reliable, and finds higher accuracy. Feature selections are performed in the suggested classifiers to remove statistical features from the data set, and models are compared based on their performance to establish the best suited methodology. After a thorough evaluation of our models, we discovered that the Random Forest Model outperforms all other methods with a 96 percent efficiency. This work can be expanded to include breast cancer classification utilising medical photographs, which are vital in cancer diagnosis.

## 8. Acknowledgments

We are very grateful to Dr. Devanjali Relan who gave us a chance to work on this project. We would like to thank her for giving us valuable suggestions and ideas. Her constant support and her wonderful guidance helped us to complete this project.

We would also like to thank BML Munjal University for including this course in our curriculum and providing us all the necessary resources for the project. All in all, we would like to thank everyone involved in this project and helped us with their suggestions to make the project better.

## 9. References

[1] L. Dora, S. Agrawal, R. Panda, A. Abraham, Optimal breast cancer classification using Gauss–Newton representation based algorithm, Expert Syst. Appl. 85 (2017) 134–145.

[2] A.V. Phan, M. Le Nguyen, L.T. Bui, Feature weighting and SVM parameters optimization based on genetic algorithms for classification problems, Appl. Intell. 46 (2) (2017) 455–469.

[3] A. Bhardwaj, A. Tiwari, H. Bhardwaj, A. Bhardwaj, A genetically optimized neural network model for multi-class classification, Expert Syst. Appl. 60 (2016) 211–221.

[4] U.R. Acharya, W.L. Ng, K. Rahmat, V.K. Sudarshan, J.E. Koh, J.H. Tan, Y. Hagiwara, C.H. Yeong, K.H. Ng, Data mining framework for breast lesion classification in shear wave ultrasound: a hybrid feature paradigm, Biomed. Signal Process. Control 33 (2017) 400–410.

[5] M.M. Ghiasi, S. Zendehboudi, Application of decision tree-based ensemble learning in the classification of breast Cancer, Comput. Biol. Med. (2020) 104089.

[6] Y. Akbulut, A. S¸ engür, Y. Guo, K. Polat, KNCM: kernel neutrosophic c-means clustering, Appl. Soft Comput. 52 (2017) 714–724.

[7] S. Al-Mahmood, J. Sapiezynski, O.B. Garbuzenko, T. Minko, Metastatic and triplenegative breast cancer: challenges and treatment options, Drug Deliv. Transl. Res. 8 (5) (2018) 1483–1507.

[8] M.A. Mohammed, B. Al-Khateeb, A.N. Rashid, D.A. Ibrahim, M.K. Abd Ghani, S. A. Mostafa, Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images, Comput. Electr. Eng. 70 (2018) 871–882.

[9] P. Kathale and S. Thorat, “Breast cancer detection and classification," Int. Conf. Emerging Trends Inf. Tech. Engin. (ic-ETITE), 2020, pp. 1-5.

[10] S. Ghosh, S. Biswas, D. C. Sarkar, P. P. Sarkar, “Breast cancer detection using a Neuro-fuzzy based classification method," Ind. J. Sc. Tech., vol. 9, no. 14, pp. 1-15, May 2016.

[11] Y. Christobel, A., & Sivaprakasam, "An empirical comparison of data mining classification methods," Int. J. Comput. Inf. Syst., vol. 3, no. 2, pp. 24–28, 2011.

[12] H. Guo and A. K. Nandi, "Breast cancer diagnosis using genetic programming generated feature," Pattern Recognit., vol. 39, no. 5, pp. 980–987, 2006.

[13] M. Karabatak and M. C. Ince, "An expert system for detection of breast cancer based on association rules and neural network," Expert Syst. Appl., vol. 36, no. 2, Part 2, pp. 3465–3469, 2009.

[14] K.Kourou, T.Exarchos , " Machine Learning Application in Cancer Prognosis and Diagnosis",Computational and Structural Biotechnology Journal,2014.

[15] J. Guo, M.Fung, F. Iqbal, Kuppen, R. Tollenaar,J. Lebran," Revealing Early Determinants of Occurance of Breast Cancer", Information Systems Frontiers,2017 issue 6, pp.1233-1241.

[16] Karabatak M, Ince MC," An expert system for detection of breast cancer based on association rules and neural network", Expert systems with Applications",2009 March,vol 36(2),pp.3465-3469.

[17] F.Bunea,"Honest Variable Selection in Linear and Logistic Regression models" International Journal of Statistics2 (2008).

[18] Octaviani TL, Rustam Z,"Random forest for breast cancer prediction",AIP Conference Proceedings, AIP Publishing Nov 4 ,2019 ,vol. 2168 .

[19] Cancer.Net, Breast Cancer: Statistics, May 2020, Accessed on September 7, 2020 [Online], Available: https://www.cancer.net/cancertypes/breast-cancer/statistics.

[20] A. Al Nahid, Y. Kong, "Involvement of machine learning for breast cancer image classification: A Survey," Hindawi Comput. Math. Meth. Medicine, vol. 2017, pp. 1-29, 2017..

[21] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Comput. Struct. Biotechnol., vol. 13, pp. 8-17, Nov 2014.