An Efficient Automated Modeling Approach for Breast Cancer DetectionUsing Different Machine Learning Techniques

[1]Dr.V.Poornima, [2]Dr.R.Ramyadevi, [3]Mrs.R.Priya

Assistant Professor,Department of Computer Science SRMIST,Ramapuram

Assistant Professor,Department of Computer Science SRMIST,Ramapuram

Assistant Professor,Department of Computer Science SRMIST,Ramapuram

poornimv@srmist.edu.in,ramyader@srmist.edu.in, priyar3@srmist.edu.in

Abstract -The most frequent cause of cancer-related mortality is breast cancer All breast lesions are not malignant, and all benign lesions do not advance to cancer. The goal is to raise the proportion of breast cancers discovered at an early stage, allowing for the adoption of more effective treatment and lowering the risks of death. Recent study has shown that Machine Learning (ML) technology can accurately diagnose Breast Cancer (BC), because effective treatment of the illness is dependent on early detection. Based on the features, several machine learning algorithms are utilized to evaluate whether a tumor is benign or malignant. The objective of this study is to diagnose the breast cancer based on 9 features using five different classification algorithms such as KNN, SVM, Adaboost, Naive Bayes and Random Forest have been compared. To achieve this objective data were collected from UCI machine learning repository. The system was implemented using Orange tool and five different ML approaches were explored and compared on breast cancer dataset.

Index-Terms: *Machine Learning Approach, Datamining Algorithms, UCI Breast Cancer Dataset, Orange tool,Disease Prediction*

I. INTRODUCTION

Despite significant advancements in early diagnosis, screening, and patient care, breast cancer still affects one in eight women globally and accounts for the majority of cancer-related fatalities in females[1]. Both benign and malignant breast tumors do not always develop to cancer. However, a combination of preoperative tests can improve the accuracy of the diagnosis[2].

The majority of breast cancer cases, nevertheless, cannot be traced back to a single factor. Discuss your unique risk with your doctor. Early detection and accurate prognostication are fundamental to identify patients who could benefit from the treatment and reduce the mortality of cancer diseases[3].

Following is a list of some of the known Risk Factors (RF) for BC:

- Age.
- Personal history of breast cancer.
- Family history of breast cancer
- Genetic factors.
- Childbearing and menstrual history

A number of researchers have concentrated on computer-based learning models. Several researchers have concentrated on developing systems in recent years, including hybrid and completely automatic systems, that could make it easier to diagnose, prognostic, and forecast the outcomes of breast cancer by combining statistics and artificial intelligence. These systems must be developed using a variety of methods, the most popular of which being machine learning (ML) algorithms[4].

The absence of prognosis models makes it challenging for doctors to develop a therapeutic strategy that might increase patient survival time. Therefore, it takes time to design the method that delivers the minimum.

II. LITERATURE REVIEW

A variety of preoperative diagnostics, including physical examination, digital breast tomosynthesis, ultrasound, and magnetic resonance have improved breast cancer screening and detection [5]. KNN, SVM, Random Forest, and Decision Tree classification results were compared by Arpita Joshi and Dr. Ashish Mehta [6]. S. Sidhu [7] studied the effectiveness of the Support Vector Machine and found that the Artificial Neural Network algorithm was beneficial in today's medical systems. In a comparison study between ANN and SVM, Kalyani Wadkar et al. [8] came to the conclusion that ANN was a superior classifier than SVM since ANN had a greater efficiency rate. By combining machine learning and deep neural network techniques with support value, Anji Reddy Vaka et al. [9] presented a pioneer method to detect BC. The results of the simulation showed that The DNN method offered advantages in terms of potential, effectiveness, and picture quality all factors that are crucial in current medical systems.

By combining machine learning and deep learning approaches, Monica Tiwari[10] proposed a novel way to identify breast cancer. According to a comparison of ML and deep learning techniques, the accuracy achieved by CNN and ANN models (99.3 percent) was higher than that of machine learning models (97.3 percent). A novel approach to detecting BC by image categorization using machine learning approaches was proposed by Abdullah-Al Nahid and Yinan Kong [11].

According to relevant studies, many ML classification algorithms are routinely used to predict breast cancer in various investigations. The other portions of this study are organised as follows: segment 3 presented a prediction of breast cancer using various classifiers, segment 4 summarised the results of the prediction using various classifiers, and segment 5 concluded the report.

III. ML Algorithms for predicting BC

Machine learning (ML) use a process or algorithm to extract patterns from unstructured data. The main goal of ML is to allow systems to learn from experience without explicit programming or human involvement [12]. The goal of ML is to build more reliable, successful, and efficient machines. The doctor's mind and expertise in the medical field serve as the machine learning tool. Machine learning can be divided into three categories, as indicated in the image below.

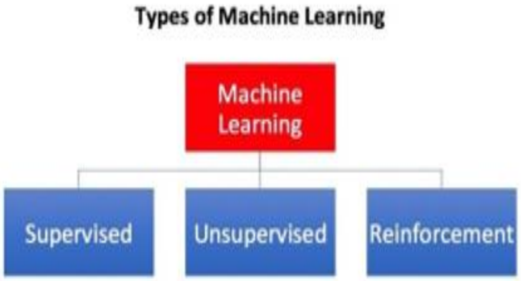


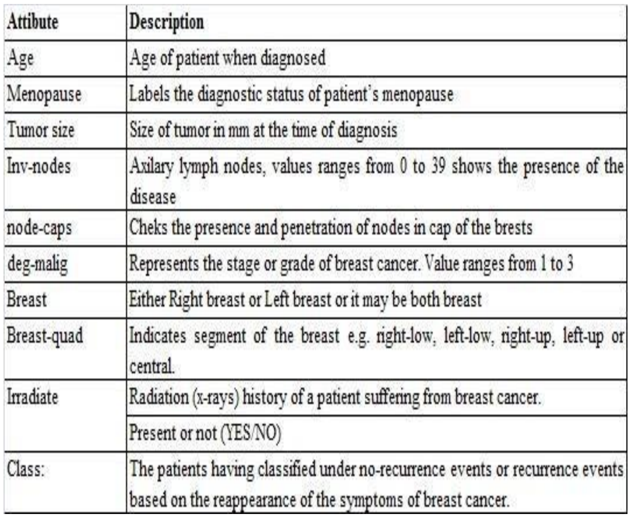
Fig. 1: Types of ML algorithms

The five various categorization methods are utilised to predict breast cancer in this study, and their implementation in the breast cancer data set is also examined.

Numerous academic research have described algorithms and nomograms for predicting the pathologic stage of patients with clinically localized cancer or Gleason score improvement. [13]. For better prognostication and stratification of patients toward personalized treatment, ML specifically enables the integration or combination of several layers of data, including those from medical pictures, laboratory results, clinical outcomes, biomarkers, and biological features[14]. Applied computational methodologies and usability issues prevent these prediction models from being widely deployed, despite the substantial scholarly interest in this area of study. Recent research has shown that ML techniques have been used to predict breast cancer prognosis using pictures of tumour tissue, diagnostic ultrasonography, and survival[15].

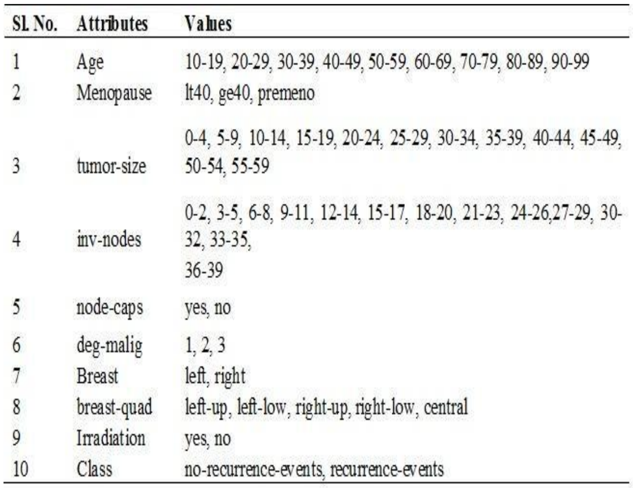
The UCI Repository provided the dataset of 286 individuals, which includes nine crucial features listed in Table I.This study's objective was to ascertain whether or not persons had a chance of developing breast cancer based on the nine attributes[16]. Table I and Table II, respectively, display the list of nine properties and their associated values.

TABLE I : LIST OF ATTRIBUTES



Attibute	Description
Age	Age of patient when diagnosed
Menopause	Labels the diagnostic status of patient's menopause
Tumor size	Size of tumor in mm at the time of diagnosis
Inv-nodes	Axilary lymph nodes, values ranges from 0 to 39 shows the presence of the disease
node-caps	Cheks the presence and penetration of nodes in cap of the brests
deg-malig	Represents the stage or grade of breast cancer. Value ranges from 1 to 3
Breast	Either Right breast or Left breast or it may be both breast
Breast-quad	Indicates segment of the breast e.g. right-low, left-low, right-up, left-up or central.
Irradiate	Radiation (x-rays) history of a patient suffering from breast cancer.
	Present or not (YES/NO)
Class:	The patients having classified under no-recurrence events or recurrence events based on the reappearance of the symptoms of breast cancer.

TABLE II : BREAST CANCER DATA SET (UCI, 1988)



Sl. No.	Attributes	Values
1	Age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
2	Menopause	lt40, ge40, premeno
3	tumor-size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
4	inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,27-29, 30-32, 33-35, 36-39
5	node-caps	yes, no
6	deg-malig	1, 2, 3
7	Breast	left, right
8	breast-quad	left-up, left-low, right-up, right-low, central
9	Irradiation	yes, no
10	Class	no-recurrence-events, recurrence-events

This study compares KNN, SVM, Adaboost, Nave Bayes, and Random Forest, five different classification algorithms, in order to diagnose breast cancer based on nine features. In order to accomplish this goal, data were gathered from the UCI machine learning repository and compared and studied using the orange tool.

IV. Results and Discussions

The Evaluation metrics are used to gauge how well the machine learning model is performing. Some of the performance indicators, including Classification Accuracy (CA), Precision, Recall, F1 Score, and ROC curve, are examined in this paper. The following table displays a comparison of ML approaches using a dataset of breast cancer cases.

• A classifier's accuracy is expressed as a percentage of all correct predictions divided by all instances.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

• Precision is a metric used to measure how many correctly positive forecasts were made.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

• Recall is a metric that measures the proportion of accurate positive predictions among all possible positive predictions.

$$Precision = \frac{TruePositives}{TruePositives + FalseNegatives}$$

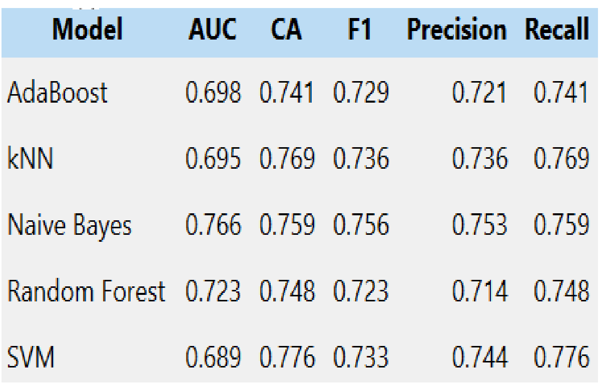
• FMeasure offers a single score that balances the issues of memory and precision in a single number.

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

• An overall assessment of performance across all potential categorization criteria is provided by AUC.

The effectiveness of breast cancer detection using five distinct algorithms is displayed in the following table.

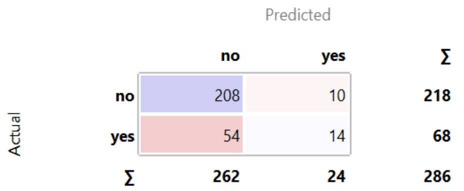
TABLE III: COMPARATIVE ANALYSIS OF ML APPROACH ON BREAST CANCER DATASET



Model	AUC	CA	F1	Precision	Recall
AdaBoost	0.698	0.741	0.729	0.721	0.741
kNN	0.695	0.769	0.736	0.736	0.769
Naive Bayes	0.766	0.759	0.756	0.753	0.759
Random Forest	0.723	0.748	0.723	0.714	0.748
SVM	0.689	0.776	0.733	0.744	0.776

The confusion matrix offers further information which classes are successfully and mistakenly predicted, and what kinds of errors are being made. The confusion matrix of five machine learning algorithms used on a dataset of breast cancer cases is displayed below.

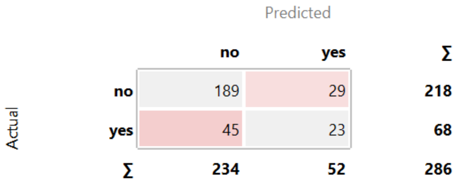
Adaboost



Predicted

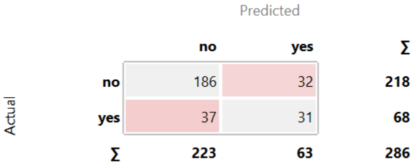
Actual	no	yes	∑
no	208	10	218
yes	54	14	68
∑	262	24	286

SVM



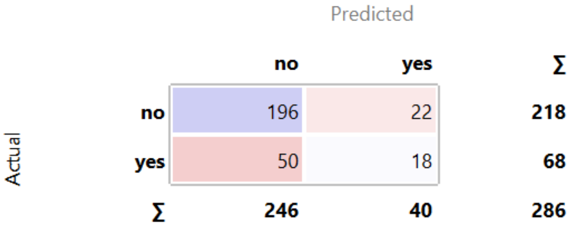
		Predicted		
		no	yes	Σ
Actual	no	189	29	218
	yes	45	23	68
	Σ	234	52	286

Random Forest



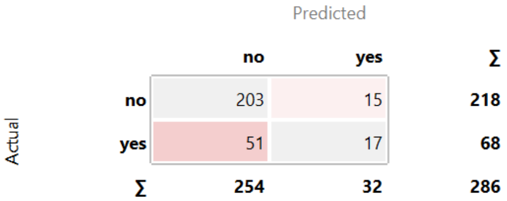
		Predicted		
		no	yes	Σ
Actual	no	186	32	218
	yes	37	31	68
	Σ	223	63	286

Naïve Bayes



		Predicted		
		no	yes	Σ
Actual	no	196	22	218
	yes	50	18	68
	Σ	246	40	286

KNN



		Predicted		
		no	yes	Σ
Actual	no	203	15	218
	yes	51	17	68
	Σ	254	32	286

Fig. 2: Confusion Matrix of ML algorithms

Based on the nine features provided by the 286 dataset and the UCI Breast Cancer dataset, Fig. 3 indicated that 201 occurrences are benign and 85 cases are malignant. Class Distribution of BC dataset is shown below:

Class Distribution:

1. No.recurrence-events:201 instances
2. Recurrence-events :85 instances

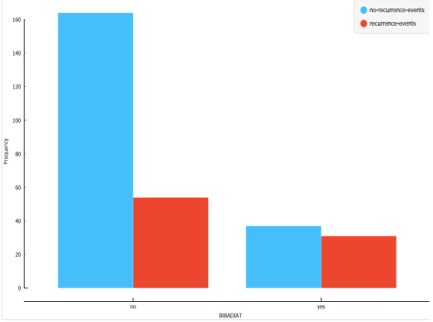


Fig 3 : Ratio of Benign and Malignant based on 2 Events

The receiver operating characteristic curve (ROC curve) displays the curve plots between two parameters on a dataset for breast cancer: True Positive and False Positives. ROC curve of ML algorithms for Benign and Malignant Data is shown in Fig 4 and Fig 5.Different colours are played by each algorithm.

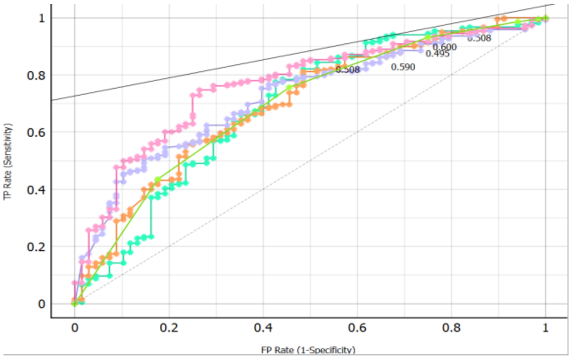


Fig 4 : Roc Curve of ML Algorithms for Benign

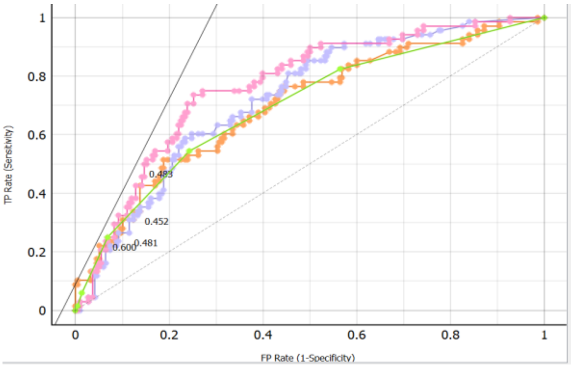


Fig 5 : Roc Curve of ML Algorithms for Malignant

The Lift Chart measures the performance of models on breast cancer dataset shown in Figure 6 and 7.

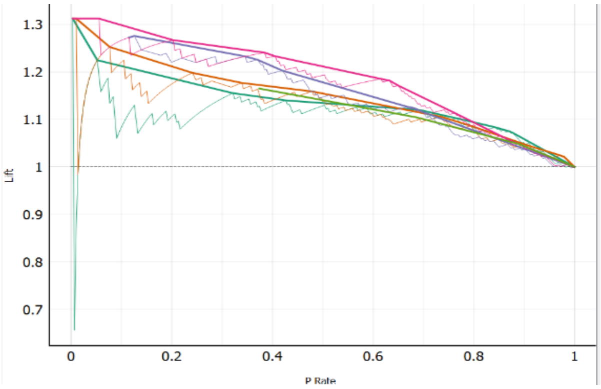


Fig 6 : Lift Curve of ML Algorithms for Benign

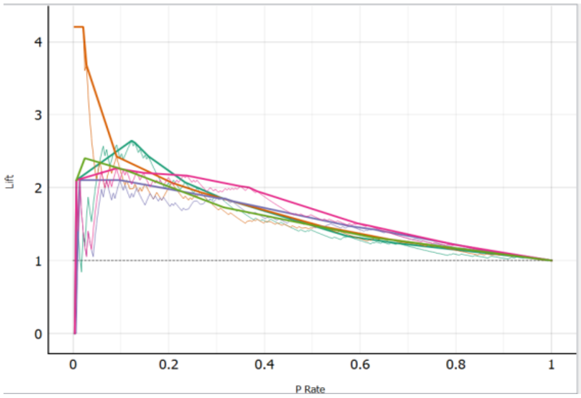


Fig 7 : Lift Curve of ML Algorithms for Malignant

Calibration curve of ML algorithms for Benign and Malignant Data is shown in Fig 8 and Fig 9.

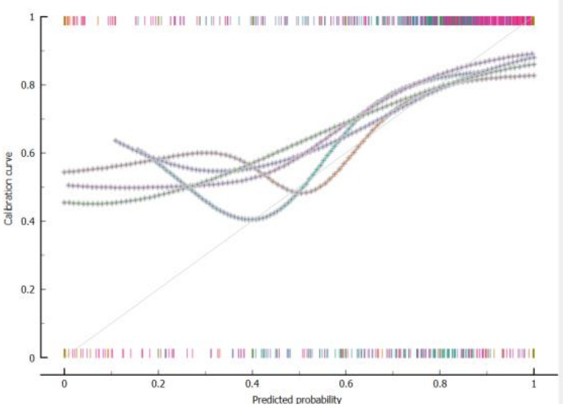


Fig 8 : Calibration Curve of ML Algorithms for Benign

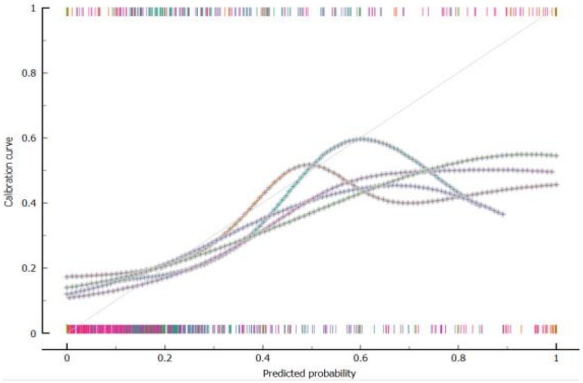


Fig 9 : Calibration Curve of ML Algorithms for Malignant

Based on the Performance of experiment the following figure concluded that KNN achieved 77% accuracy of detection and Accuracy of remaining Approaches using UCI dataset is shown below:

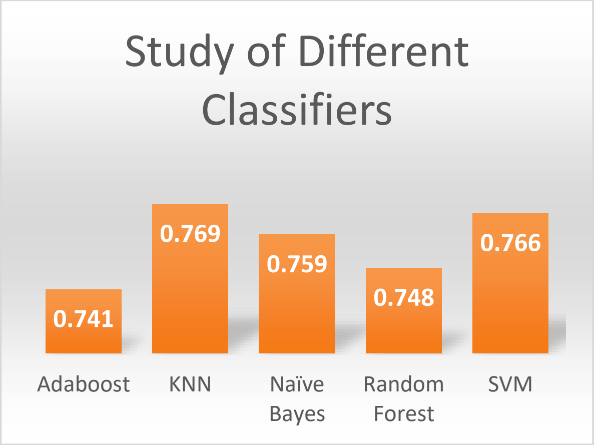


Fig 10 : Accuracy of Different Classifier

V.CONCLUSION

The major goal of this work is to compare Adaboost, KNN, Naive Bayes, Random Forest, and SVM for BC detection using several machine learning approaches based on nine characteristics. On a dataset of breast cancer cases, it was found that KNN outperforms other classifiers. Data were gathered from the UCI machine learning repository and examined and contrasted using the orange tool in order to achieve this detection. Based on the nine necessary features, these systems can help a patient regardless of whether they have breast cancer. Based on a patient's stage of breast cancer, there are numerous treatments available; data mining and machine learning can be very helpful in determining the course of treatment to be taken by collecting knowledge from such suitable databases.

Reference

[1]. Nolan J, Dunne SS, Mustafa W, Sivananthan L, Kiely PA, Dunne CP. Proposed hypothesis and rationale for association between mastitis and breast cancer. *Med Hypotheses*. 2020;144. doi: 10.1016/j.mehy.2020.110057 [PubMed] [CrossRef] [Google Scholar]

[2]. DeSantis CE, Ma J, Gaudet MM, et al. Breast cancer statistics, 2019. *CA Cancer J Clin*. 2019;69(6). doi: 10.3322/caac.21583 [PubMed] [CrossRef] [Google Scholar

[3].Fiorica JV Breast Cancer Screening, Mammography, and Other Modalities. *Clin Obstet Gynecol*. 2016;59(4). doi: 10.1097/GRF.0000000000000246 [PubMed] [CrossRef] [Google Scholar]

[4].Lughezzani G, Briganti A, Karakiewicz PI, et al. Predictive and prognostic models in radical prostatectomy candidates: A critical analysis of the literature. *Eur Urol*. 2010;58(5). doi: 10.1016/j.eururo.2010.07.034 [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[5].Cai D, Lin T, Jiang K, Sun Z. Diagnostic value of MRI combined with ultrasound for lymph node metastasis in breast cancer: Protocol for a meta-analysis. *Med (United States)*. 2019;98(30). doi: 10.1097/MD.0000000000016528 [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[6]. Arpita Joshi and Dr. Ashish Mehta Comparative Analysis of Various Machine Learning Techniques for Diagnosis of Breast Cancer (2017).

[7]. David A. Omondiagbe, Shanmugam Veeramani and Amandeep S. Sidhu Machine Learning Classification Techniques for Breast CancerDiagnosis (2019).

[8].Kalyani Wadkar, Prashant Pathak and Nikhil Wagh Breast Cancer Detection Using ANN Network and Performance Analysis with SVM (2019).

[9].Anji Reddy Vaka, Badal Soni and Sudheer Reddy Breast Cancer Detection by Leveraging Machine Learning (2020).

[10]. Monika Tiwari, Rashi Bharuka, Praditi Shah and Reena Lokare Breast Cancer Prediction using Deep learning and Machine Learning Techniques.

[11].Abdullah-Al Nahid and Yinan Kong Involvement of Machine Learning for Breast Cancer Image Classification: Asurvey (2017).

[12].https://www.tutorialspoint.com/machine_learning_with_python/index.htm

[13].Lughezzani G, Briganti A, Karakiewicz PI, et al. Predictive and prognostic models in radical prostatectomy candidates: A critical analysis of the literature. *Eur Urol*. 2010;58(5). doi: 10.1016/j.eururo.2010.07.034 [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[14]. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Inf Fusion*. 2019;50. doi: 10.1016/j.inffus.2018.09.012 [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[15].Turkki R, Byckhov D, Lundin M, et al. Breast cancer outcome prediction with tumour tissue images and machine learning. *Breast Cancer Res Treat*. 2019;177(1). doi: 10.1007/s10549-019-05281-1 [PMC free article] [PubMed] [CrossRef] [Google Scholar]

16.https://archive.ics.uci.edu/ml/datasets/Breast+Cancer