# BREAST CANCER DETECTION USING MACHINE LEARNING

[1]Renuka Mandapalli, [2]Kuppili Vasanthi, [3]Lokesh Konna

Department of Computer Science and Engineering

Raghu Engineering College, Visakhapatnam

[1]19981a0595@raghuenggcollege.in, [2]19981a0583@raghuenggcollege.in , [3]19981a0577@raghuenggcollege.in

S.NAVYA
Assistant Professor, Department of CSE, Raghu Engineering College, Visakhapatnam

## Abstract

Breast cancer is a common and deadly disease that affects women worldwide. Early detection can significantly improve patient outcomes. Machine learning techniques have shown the potential in enhancing breast cancer detection accuracy. We discuss the various data sources and features used in machine learning models, including medical imaging, clinical data, and genetic markers. Our proposed approach's efficacy will be assessed using metrics like accuracy, sensitivity, specificity, and precision. Our study explores machine learning models for breast cancer detection and our proposed method shows superior accuracy and sensitivity compared to other algorithms. We anticipate that this approach could revolutionize breast cancer detection and improve patient outcomes by enabling early detection and treatment.

**Keywords**: Breast Cancer, SVM, Random Forest, detection, Accuracy.

## Introduction

Breast cancer is a type of cancer that develops from the cells in the breast, and it can occur in both women and men. It is a widespread type of cancer that affects individuals globally. Cancer has the potential to spread beyond the breast to other organs such as the lymph nodes, bones, lungs, liver, or brain, through the process of metastasis. The precise causes of breast cancer are not completely known, but some factors increase the chances of developing it. These include gender, age, Inherited gene mutations, Hormone levels, early menstruation onset, late menopause, and a sedentary lifestyle are some of the known risk factors.

Breast cancer symptoms can manifest in different ways and include changes in the breast or nipple area. The most frequent symptom is a lump or thickening in the breast or underarm region. Other possible symptoms include changes in breast size or shape, nipple discharge, alterations in skin texture or appearance (such as dimpling, puckering, or redness), an inverted nipple, or constant pain in the breast or nipple. A person's physical, mental, and emotional well-being, as well as their general quality of life, can all be profoundly impacted by breast cancer. It is essential for individuals to be aware of their risk factors for breast cancer and to take appropriate preventive measures. Regular breast self-exams and attending recommended screening tests, such as mammograms. Treatment of breast cancer requires early detection as it improves the chances of successful treatment and increases the likelihood of survival. Treatment for breast cancer can include various approaches such as targeted therapy, surgery, chemotherapy, radiation therapy, and hormone therapy. The treatment for breast cancer is dependent on the stage and type of cancer, as well as individual factors.

## Literature study

In this literature review, we aim to examine previous studies that have investigated the use of ML in detecting breast cancer.

[A] K. Sriram, T. Chakravarthy, and K. Anastraj performed a comparative analysis of different machine learning models, such as backpropagation network, convolutional neural network, support vector machine, and artificial neural network, using the WBC dataset. Their study aimed to determine which machine learning approach is the most effective for detecting breast cancer. According to the simulation results, the support vector machine outperformed the other algorithms, achieving an accuracy rate of 94%. The study suggests that the SVM model can be used as a trustworthy and efficient method for detecting breast cancer using machine learning methods.

[B] Raed Shubair and Dana Bazazeh compared different machine-learning methods for breast cancer detection using the Wisconsin BC dataset as their training set. The outcomes of the implementation method that the effectiveness of the models changed according to the chosen approach. The findings indicated that SVMs had the highest accuracy, specificity, and precision, making them an effective method for breast cancer detection. The analysis shows that, among the compared approaches, Random forest seemed to have the highest probability of correctly diagnosing cancers. Overall, the study emphasizes how crucial it is to choose the best machine-learning technique to detect tumors because each one has advantages and disadvantages.

[C] Nikhil Wagh, Prashant Pathak, and Kalyani Wadkar conducted a study comparing artificial neural networks and support vector machines, as well as other classifiers such as KNN, CNN, and Inception V3, and found that artificial neural networks (ANN) outperformed SVM in terms of overall performance based on experimental findings.

[D] Nikhil Wagh, Prashant Pathak, and Kalyani Wadkar compared SVM and ANN and incorporated various classifiers, such as Convolutional Neural Network (CNN), K-Nearest Neighbors, and Inception V3, to enhance dataset processing. The study came to the conclusion that Artificial Neural Network (ANN) was a better classifier than SVM based on the performance evaluation and experimental data. In terms of performance, ANN performed better and had a higher classification rate for finding breast cancer.

## Methodology

### *DATA SET*

In this work, the Wisconsin breast cancer dataset—which can be acquired from various sources, including the Machine Learning Repository at the University of Chicago and Kaggle—is being used. The dataset is provided in CSV format and contains information about breast cancer tumors, including their size, shape, and other characteristics. This methodology has the potential to improve breast cancer early detection as well as treatment, improving patient outcomes. In machine learning, the Wisconsin breast cancer dataset is frequently utilized research for breast cancer detection, as it comprises 569 instances, each with 30 features, and includes a label with a target variable indicating whether the instance is benign (non-cancerous) or malignant. The remaining 30 attributes are numerical measurements of various features, such as cell nuclei observed in the biopsy, including Bland Chromatin, Mitoses, cell nucleus smoothness, radius, texture, and fractal dimension. These 30 features of the Wisconsin breast cancer dataset are utilized to identify patterns and classify biopsy samples as either malignant or benign, based on their respective feature values.

### *PRE-PROCESSING*

Data Pre-processing involves two criteria:

**Cleaning Data:** This entails eliminating duplicates, fixing mistakes, and adding any missing data.

**Data Integration:** This step involves combining multiple datasets into a single dataset to improve the learning process.

## Proposed Method

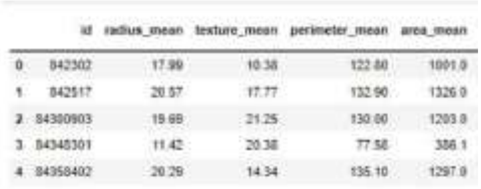
***ALGORITHM***:

This research aims to use machine learning algorithms to find breast cancer based on data. The study aims to identify the most efficient method for data classification and enhancing the precision of breast cancer detection by comparing various algorithms with the Wisconsin Breast Cancer dataset. The use of specific evaluation metrics like sensitivity, accuracy, f1-score, and precision will provide a comprehensive analysis of the algorithms' performance. Additionally, understanding how Random Forest (RF) outperforms other algorithms in diagnosing breast cancer, even with reduced variables, is essential for improving treatment and early detection of breast cancer. Overall, the research has the potential to contribute to better outcomes for patients by improving the accuracy of breast cancer detection and early intervention. The project will rely on the Python programming language and the Scikit-Learn library. Overall, the study's goal is to improve breast cancer diagnosis and treatment through the use of advanced machine-learning techniques.

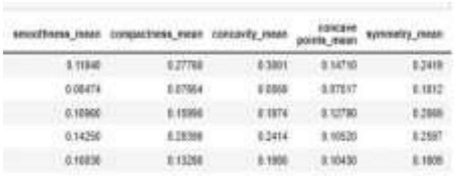
## Data Exploration

We will import libraries and datasets to create a dataset.

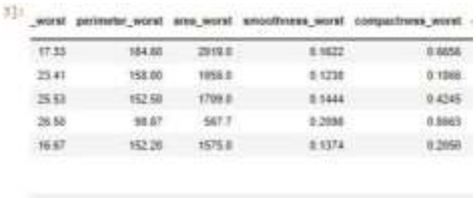
df. head(5) is used to display the first 5 rows of a Pandas DF.



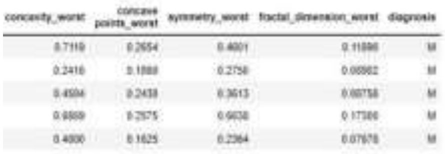
| | id | radius_mean | texture_mean | perimeter_mean | area_mean |
|---|---|---|---|---|---|
| 0 | 842302 | 17.99 | 10.38 | 122.80 | 1001.0 |
| 1 | 842517 | 20.57 | 17.77 | 132.90 | 1326.0 |
| 2 | 84300903 | 19.69 | 21.25 | 130.00 | 1203.0 |
| 3 | 84348301 | 11.42 | 20.38 | 77.58 | 386.1 |
| 4 | 84358402 | 20.29 | 14.34 | 135.10 | 1297.0 |



| smoothness_mean | compactness_mean | concavity_mean | concave points_mean | symmetry_mean |
|---|---|---|---|---|
| 0.11840 | 0.27760 | 0.3001 | 0.14710 | 0.2419 |
| 0.08474 | 0.07864 | 0.0869 | 0.07017 | 0.1812 |
| 0.10960 | 0.15990 | 0.1974 | 0.12790 | 0.2069 |
| 0.14250 | 0.28390 | 0.2414 | 0.10520 | 0.2597 |
| 0.10030 | 0.13280 | 0.1980 | 0.10430 | 0.1809 |



| _worst | perimeter_worst | area_worst | smoothness_worst | compactness_worst |
|---|---|---|---|---|
| 17.33 | 184.60 | 2019.0 | 0.1622 | 0.6656 |
| 23.41 | 158.80 | 1956.0 | 0.1238 | 0.1866 |
| 25.53 | 152.50 | 1709.0 | 0.1444 | 0.4245 |
| 26.50 | 98.87 | 567.7 | 0.2098 | 0.8663 |
| 16.67 | 152.20 | 1575.0 | 0.1374 | 0.2050 |



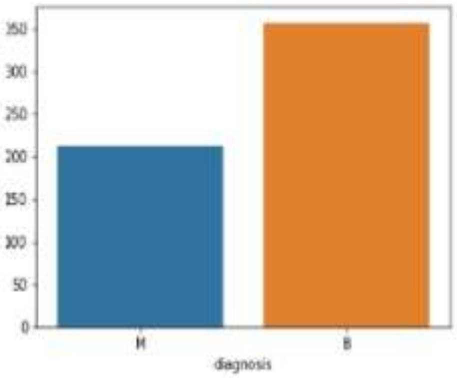
| concavity_worst | concave points_worst | symmetry_worst | fractal_dimension_worst | diagnosis |
|---|---|---|---|---|
| 0.7119 | 0.2654 | 0.4601 | 0.11890 | M |
| 0.2416 | 0.1860 | 0.2750 | 0.08902 | M |
| 0.4504 | 0.2430 | 0.3613 | 0.08758 | M |
| 0.6869 | 0.2575 | 0.6638 | 0.17300 | M |
| 0.4000 | 0.1625 | 0.2364 | 0.07678 | M |

The shape attribute in pandas returns a tuple representing the dimensions of a data frame or a Series.

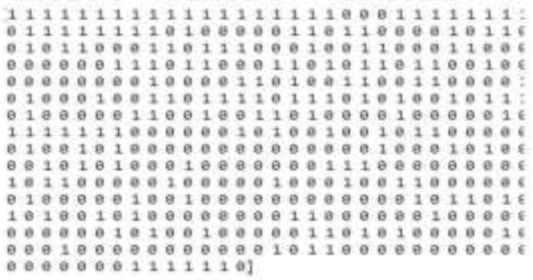
df. Shape

(569, 32)

Actually, the target variable in the Wisconsin breast cancer dataset is the 'diagnosis' column, which indicates whether the tumor is malignant (M) or benign (B). The dataset has 569 rows, each corresponding to a biopsy sample, and 31 columns, including the diagnosis column and 30 features that describe the properties of the sample's cell nuclei. A value of 1 represents malignant cancer, and 0 represents benign cancer. There are 357 benign cases and 212 malignant cases in the dataset. Each row represents a patient and their 32 characteristics. We can visualize the distribution of malignant and benign cases using a graph.

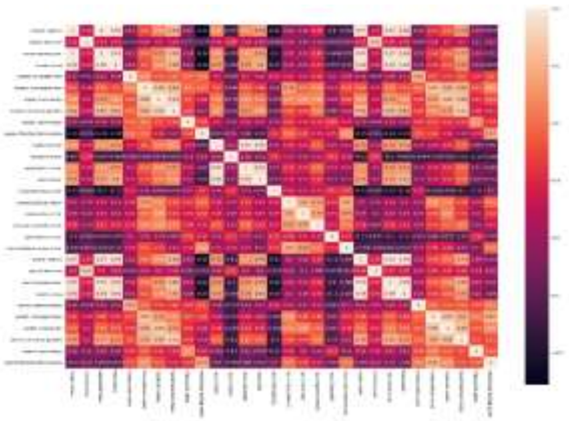


To transform categorical data to integers, we use label encoding, which assigns integer values to each category, representing the categorical data in a numerical form for machine learning algorithms to process.



```
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1
0 1 1 1 1 1 1 1 1 0 1 0 0 0 0 0 1 1 0 1 1 0 0 0 0 1 0 1 1
0 1 0 1 1 0 0 0 1 1 0 1 1 1 0 0 0 1 0 0 1 1 0 0 0 1 1 0 0
0 0 0 0 0 0 1 1 1 0 1 1 0 0 0 1 1 0 1 0 1 1 0 1 1 0 0 1 0
0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 1 0 0 1 1 0 0 1 1 0 0 0 0
0 1 0 0 0 1 0 0 1 1 0 1 1 1 1 0 1 1 1 0 1 0 1 0 0 1 0 1 1
0 1 0 0 0 0 0 1 1 0 0 1 0 0 1 1 0 1 0 0 0 0 1 0 0 0 0 0 1
1 1 1 1 1 1 1 0 0 0 0 0 0 1 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0
0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1 0
0 0 1 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0
1 0 1 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 1 1 0 0 0 0 0
0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 1
1 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0
0 0 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 1 1 0 1 0 1 0 0 0 0 0 1
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 1 1 1 1 1 1 0]
```

In the dataset, Malignant (M) cells are represented by 1, while Benign (B) cells are represented by 0. We can visualize the correlation between the different attributes in the dataset.



The heatmap displays the correlation between each column in the dataset, highlighting how strongly one column affects every other column. For training and testing purposes, the dataset was divided into independent(X) and dependent (Y) variables.

X=df.iloc[:, :-1].values
Y=df.iloc[:, 30].values

The independent (X) and dependent datasets (Y) are arrays representing the attributes used to predict the outcome and the cancer diagnosis for the patient, respectively.

The dataset was split into a 75% training portion and a 25% testing portion. The model was trained on the training dataset using a variety of machine-learning algorithms.

We now display the training data's accuracy:

[0]Logistic Regression: 95.0 %
[1]KNN Classifier : 94.0 %
[2]Linear SVC: 95.0 %
[3]Gaussian Kernel SVC: 91.0 %
[4]Decision Trees Classifier: 90.0 %
[5]Random Forest Classifier: 96.0 %
[6]BernoulliNB Classifier : 59.0 %
[7]MultinomialNB Classifier : 90.0 %
[8]Artificial Neural Network: 93.0 %

The random forest classifier achieved the highest accuracy rate of 96%. This indicates that it outperformed other models and is the most accurate model for this dataset.

## Evaluation metric

The performance of computational models is evaluated using various parameters, including F-measure, FP rate, TP rate, precision, and recall.

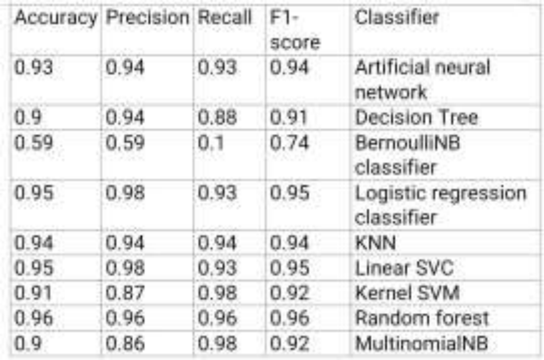
- True positive (TP): Instances where the model accurately predicts the positive class is referred to as true positives (TP).
- False negative (FN): In a confusion matrix, a false negative occurs when a positive instance is incorrectly classified as negative by a machine learning model.
- False positive (FP): A false positive occurs when a negative instance is incorrectly classified as positive by a machine learning model.
- True negative (TN): Instances where the model accurately predicts the negative class are referred to as true negatives (TN).
- Accuracy: The share of effectively categorized instances.

Accuracy = TP+TN / (FN+TP+FP+TN)

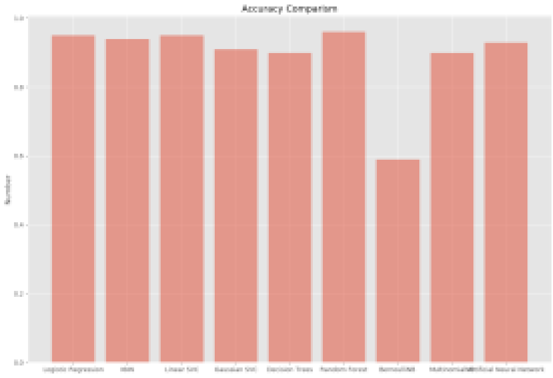
- Precision = TP / (FP+TP)
- Recall = TP / (FN+TP)
- F1 Score = 2 * ( Recall * Precision) / (Recall + Precision)

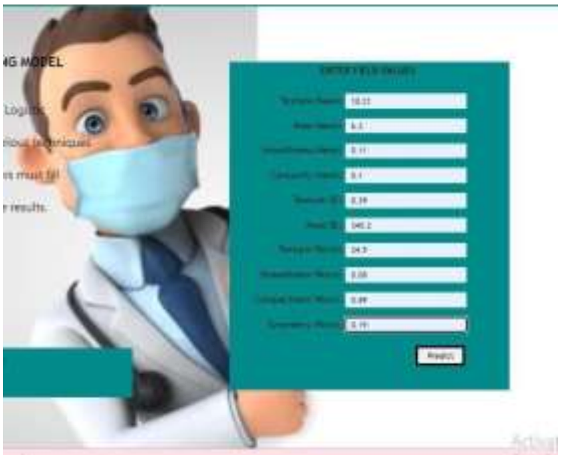
## Sample Results

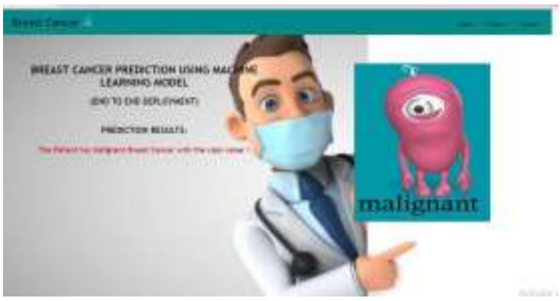
COMPARISON CLASSIFIERS PERFORMANCE



| Accuracy | Precision | Recall | F1-score | Classifier |
|---|---|---|---|---|
| 0.93 | 0.94 | 0.93 | 0.94 | Artificial neural network |
| 0.9 | 0.94 | 0.88 | 0.91 | Decision Tree |
| 0.59 | 0.59 | 0.1 | 0.74 | BernoulliNB classifier |
| 0.95 | 0.98 | 0.93 | 0.95 | Logistic regression classifier |
| 0.94 | 0.94 | 0.94 | 0.94 | KNN |
| 0.95 | 0.98 | 0.93 | 0.95 | Linear SVC |
| 0.91 | 0.87 | 0.98 | 0.92 | Kernel SVM |
| 0.96 | 0.96 | 0.96 | 0.96 | Random forest |
| 0.9 | 0.86 | 0.98 | 0.92 | MultinomialNB |







BREAST CANCER PREDICTION USING MACHINE LEARNING MODEL
(END TO END DEPLOYMENT)
PREDICTION RESULTS:
The Patient has Malignant Breast Cancer with the class value 1

## Conclusion

The paper compared the performance of several machine learning techniques for breast cancer detection, such as ANN, KNN, SVM, Logistic regression, Decision Tree, and Random forest. The study found that Random Forest outperformed other methods in terms of accuracy, precision, and scalability for large datasets. The Random Forest model achieved a high accuracy rate of 96.0%, making it a promising approach for breast cancer detection. The findings suggest that Random Forest can be a valuable tool for clinicians and researchers in accurately identifying breast cancer cases and improving patient outcomes.

## References

[1] K. Sriram, Dr. T. Chakravarthy, and K.Anastraj " Breast Cancer detection of either Benign Or Malignant Tumors using Deep Convolutional Neural Network With Machine Learning Techniques "(2019).

[2]Muhammet Fatih Ak "A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and MachineLearning Applications" ((2020).

[3] Dana Bazazeh and Raed Shubair's "Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis"(2016).

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

[5] M.A.H Akhand and M.Murase “Neural network ensemble construction fusing multiple popular methods.”. International journal of computer science 2004.

[6] Ramik Rawal “Breast Cancer Prediction Using MachineLearning”(2020).

[7] Nithya R, Santhi B. A Data Mining Techniques for Diagnosis of Breast Cancer Disease. World Applied Sci J 2014:18-23.

[8] Ilias Maglogiannis, E Zafiropoulos “An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers” Applied Intelligence, 2009 – Springer.

[9] Valerie Bourd`es, St´ephane Bonnevay, Paolo Lisboa, R´emy Defrance, David P´erol, Sylvie Chabaud, Thomas Bachelot, Th´er`ese Gargi,6 and Sylvie N´egrier “Comparison of Artificial Neural International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.3, May 2013 112 Network with Logistic regression as Classification Models for Variable Selection for Prediction of Breast Cancer Patient outcomes”

[10] I Guyon, J Weston, S Barnhill “Gene selection for cancer classification using support vector machines” … - Machine learning, 2002 – Springer

[11] RF Chang, WJ Wu, WK Moon, YH Chou “Support vector machines for diagnosis of breast tumors on US images” - Academic radiology, 2003 – Elsevier

[12] AM Bagirov, B Ferguson, S Ivkovic “New algorithms for multi-class cancer diagnosis using tumor gene expression signatures”, 2003 - Oxford Univ Press.

[13] Hiba Asria, Hajar Mousannifb, Hassan Al Moatassime, Thomas Noeld "Using Machine Learning Algorithms for Breast CancerRisk Prediction and Diagnosis" (2016).

[14] Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S "Breast Cancer Prediction using Machine Learning" (2019).