VISHWAKARMA INSTITUTE OF INFORMATION TECHNOLOGY,PUNE

## Mini Project Report

On

# Breast Cancer Detection Using Machine Learning Algorithms

## Under the Guidance of

**Professor: Mrs. F.M.Inamdar**

SUBMITTED BY


| GR.NO | ROLL.NO | NAME |
|---|---|---|
| 22010461 | 332003 | Pranav Andhale |
| 22010320 | 332004 | Ankush Chajgotra |

# Breast Cancer Detection

**Abstract:**

Breast cancer is a dangerous disease for women. If it does not identify in the early-stage then the result will be the death of the patient. It is a common cancer in women worldwide. Worldwide near about 12% of women are affected by breast cancer and the number is still increasing.

We have extracted features of breast cancer patient cells and normal person cells. As a Machine learning engineer / Data Scientist has to create an Artificial Intelligence based Machine Learning model to classify malignant and benign tumours.

To complete this project, we are using the supervised machine learning classifier algorithm. We have used different classification algorithms like Support vector classifier, Random Forest classifier, XGBoost classifier and picked the best algorithm which gives us more accuracy.

**Keywords:**

**SVC , Random Forest classifier, XGBoost classifier.**

**Introduction:**

Breast cancer is one of the major causes of death in women around the world. According to the American Cancer society, 41,760 women and more than 500 men died from breast cancer recently. Breast cancer occurs in four main types: normal, benign, in-situ carcinoma and invasive carcinoma. A benign tumour involves a minor change in the breast structure. It is not harmful and does not classify as a harmful cancer.

In cases of in-situ carcinoma, the cancer is only in the mammary duct lobule system and does not affect other organs. This type is not dangerous and can be treated if diagnosed early. Invasive carcinoma is considered to be the most dangerous type of breast cancer, as it can spread to all other organs. Breast cancer can be detected using several methods including X-ray mammography, ultrasound (US), Computed Tomography (CT), Portion Emission Tomography (PET), Magnetic Resonance Imaging (MRI)

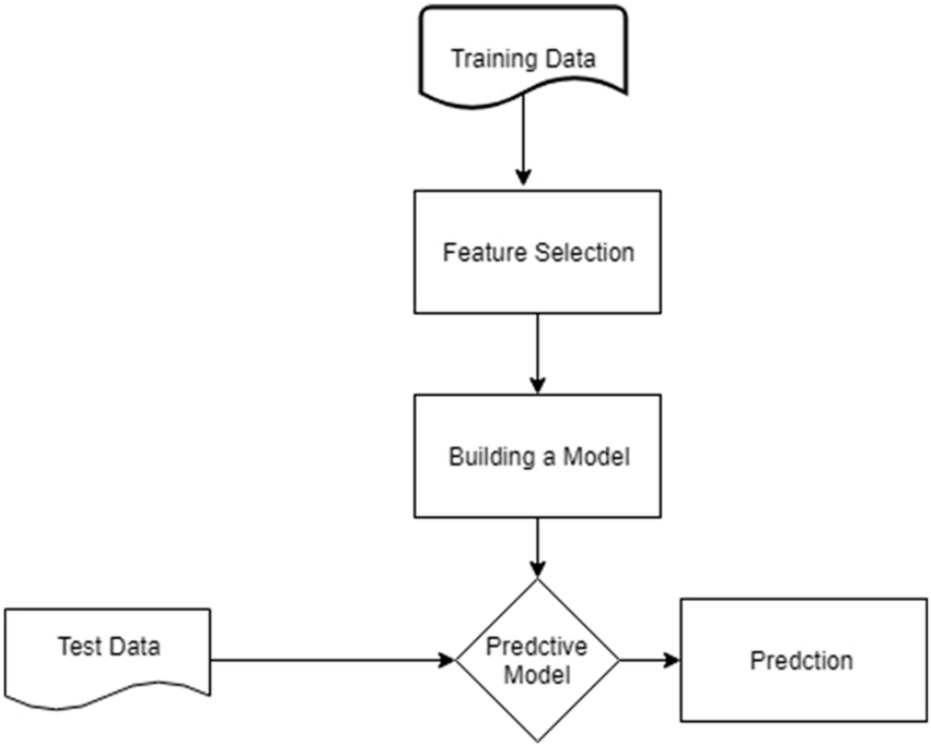
and breast temperature measurement. Usually, the golden standard is a pathological diagnosis for detecting breast cancer. This involves an image analysis of the removed tissue, which is stained in the lab to increase visibility. Hematoxylin and Eosin (H&E) are commonly used for the staining process. Breast cancer can be diagnosed using one of two approaches: histopathological image analysis or genomics. Histopathological images are microscopic images of breast tissue that are extremely useful in early treatment of cancer. As for genomics, the authors stated that radiogenomics is an emerging research field focusing on multi-scale associations between medical imaging and gene expression data. Radio-genomics provide both radiological and genetic features that can enhance diagnosis. It can analyse tissues at the molecular level, helping with prediction and early detection of cancer. The main difference between imaging information and radio-genomics is the critical knowledge gap between imaging at the tissue level and analysing the underlying molecular and genetic disease biomarkers. As imaging is less precise, it may lead to over- or under-treatment. While radio-genomics is much more effective than histopathological imaging, it is rarely used because the process involves datasets that are very expensive and require high computational power. As a result, a limited number of labs conduct radio-genomics experiments. This research paper addresses the following research questions and highlights the deep learning models, looking at their performance, the datasets used and possibilities for breast cancer classification and detection.

**Literature Survey:**

| Research Paper | Methodology | Advantages | Limitations |
|---|---|---|---|
| Breast Cancer Detection Using Machine Learning Algorithms, December 2018 | The Wisconsin Diagnosis Breast Cancer data set has been used to compare the performance of various machine learning algorithm methodologies. | In the field of cancer research, supervised machine learning techniques will be highly helpful in determining a cancer type's prognosis and early detection. | The dataset used for methodology is Wisconsin Diagnosis Breast Cancer data set. It needs to be implemented on more such datasets. |
| Prediction of benign and malignant breast cancer using data mining techniques, February, 2018 | Uses Naive Bayes, RBF Network, and J48, three well-known data mining algorithms, for each breast cancer dataset. | The holdout sample findings showed that Naive Bayes is the best predictor with 97.36% accuracy, RBF Network was second with 96.77% accuracy, and J48 was third with 93.41% accuracy. | The algorithms used, RBF and J48 are difficult to implement on the general datasets. |
| Breast Cancer Detection using Machine Learning | After training the model on the provided dataset, the k-fold cross validation test is used to determine the model's proficiency on fresh or previously unexplored data. | The result indicates that using multidimensional data with various feature choices and classification models can produce more reliable and promising tools for usage in this field. | To forecast additional variables and improve accuracy and precision, more research must be prioritised and put into practise. |
| An investigation of XGBoost-based algorithm for breast cancer classification | Breast Cancer Histopathological Image Classification is the dataset utilised to test the suggested methodology. | The goal of the study that is being given is to investigate novel methods for classifying breast cancer using the most recent machine learning techniques. | unable to do additional research on the speed, area, footprint, etc. As a result, the publications do not assert that they are superior for greater performance within these parameters. |
| Performance Analysis of XGBoost Ensemble Methods for Survivability with the Classification. | Breast cancer prediction using the XGBoost ensemble method and known feature patterns. | With 98.20% accuracy, the findings demonstrate that XGBoost-RF performs better than other ensemble classifiers. | According to the findings, the MAE and RMSE error rates for the XGBoost-RF classifier are 0.12, 0.27. |

**Proposed Methodology:**

The likelihood of survival and cancer recurrence is significantly influenced by medical treatment as well as the precision of the diagnosis. A split of 80:20 between training and testing data was used in this experiment, which used arbitrary retrieved data. The model was trained using training sets, then test data were utilised to determine how well it worked. The dataset features 569 cases and 30 variables or traits whose values will determine if a person is likely to get breast cancer. The output variable, sometimes referred to as the target variable, is a binary variable that can either be malignant or benign. Breast cancer is present if a person has a malignant tumour; otherwise, if a benign tumour is present if a person does not have breast cancer.




Here we are using different algorithms for prediction and out of those choosing the best algorithm with accuracy score.

**Machine learning algorithms used:**

**Random Forest Algorithm:** The supervised learning subcategory includes another well-known ML approach that is applied to both classification and regression problems. One of the more sophisticated and adaptable ensemble learning techniques is this one. The outcome of this algorithm mainly relies on this straightforward yet multiplex and compounded strategy, which uses numerous decision trees to create a family of trees for categorization methods. Because of this, its robustness can be shown when it is used in a big database. Even while it excels at classification, it loses effectiveness and benefit when applied to regression issues. The Random Forest Algorithm is explained in simple terms below: 1)From a given dataset, we specify random data points(k). 2)Constructing decision trees upon data points based on association as per needed. 3)Selecting N numbers for decision trees to be assembled. 4)Gathering particulars and details from N trees to forego prediction of new data, i.e assigning random data points and repeating construction of decision trees. 5)Taking prediction of each decision tree and reviewing the category and assigning it as a new data point, based on the majority vote of neighbouring trees.

**Support Vector Machine(SVM):** One of the most popular and sought-after supervised learning algorithms is SVM. The term "hyperplane" refers to a decision boundary that can separate and isolate n-dimensional space into classes, placing the data points in the appropriate category. This technique is primarily used in classification problems. It offers advantages over others, including great dimensional space and memory efficiency. Claiming that it can be applied to both linearly and nonlinearly separable data.

**XGBoost:**

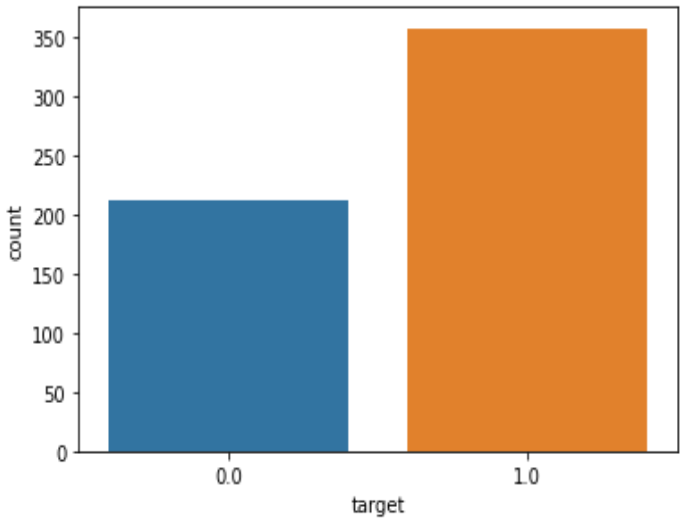
XGBoost is a distributed gradient boosting library that has been optimised to be extremely effective, adaptable, and portable. It uses the Gradient Boosting framework to implement machine learning algorithms. With the use of XGBoost, many data science issues may be quickly and accurately solved using a parallel tree boosting technique also known as GBDT or GBM. Comprehending the machine learning ideas and techniques that

supervised machine learning, decision trees, ensemble learning, and gradient boosting are built upon is essential to understanding XGBoost.

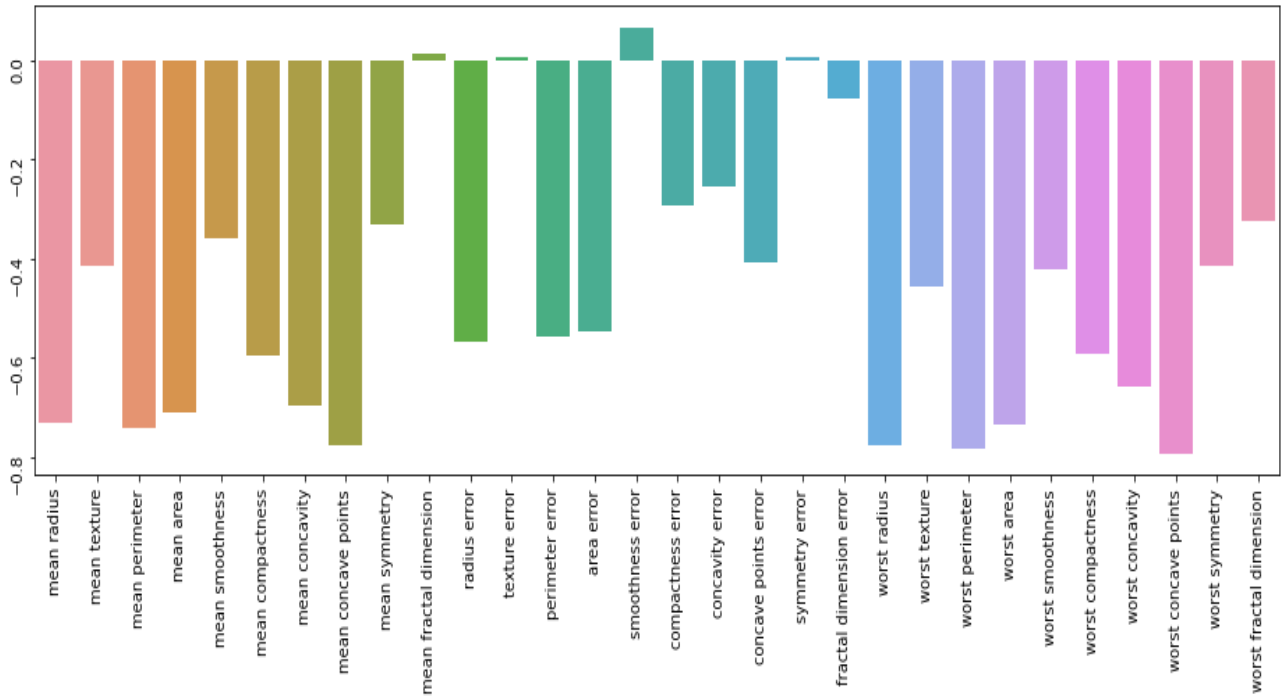
Supervised machine learning uses algorithms to train a model to find patterns in a dataset with labels and features and then uses the trained model to predict the labels on a new dataset's features.

- Pair Plot of sample feature:

- Count the target class:



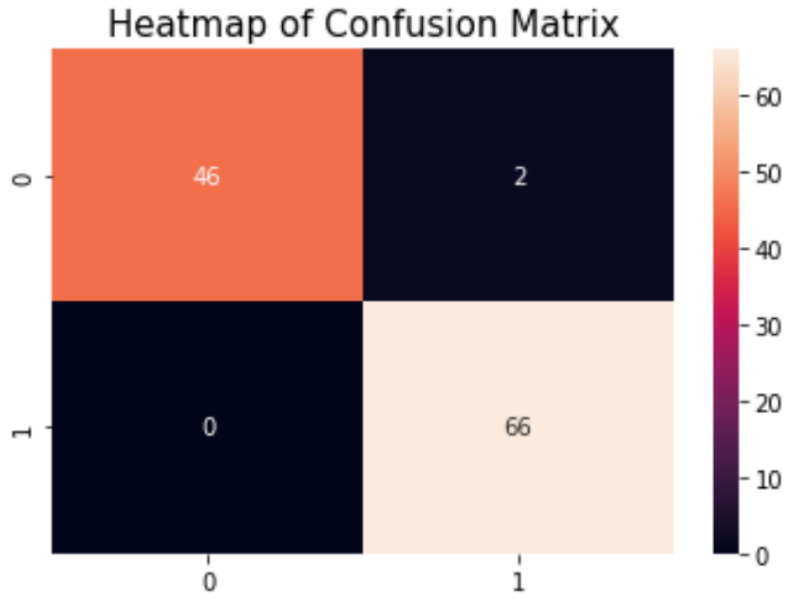

- Correlation barplot:

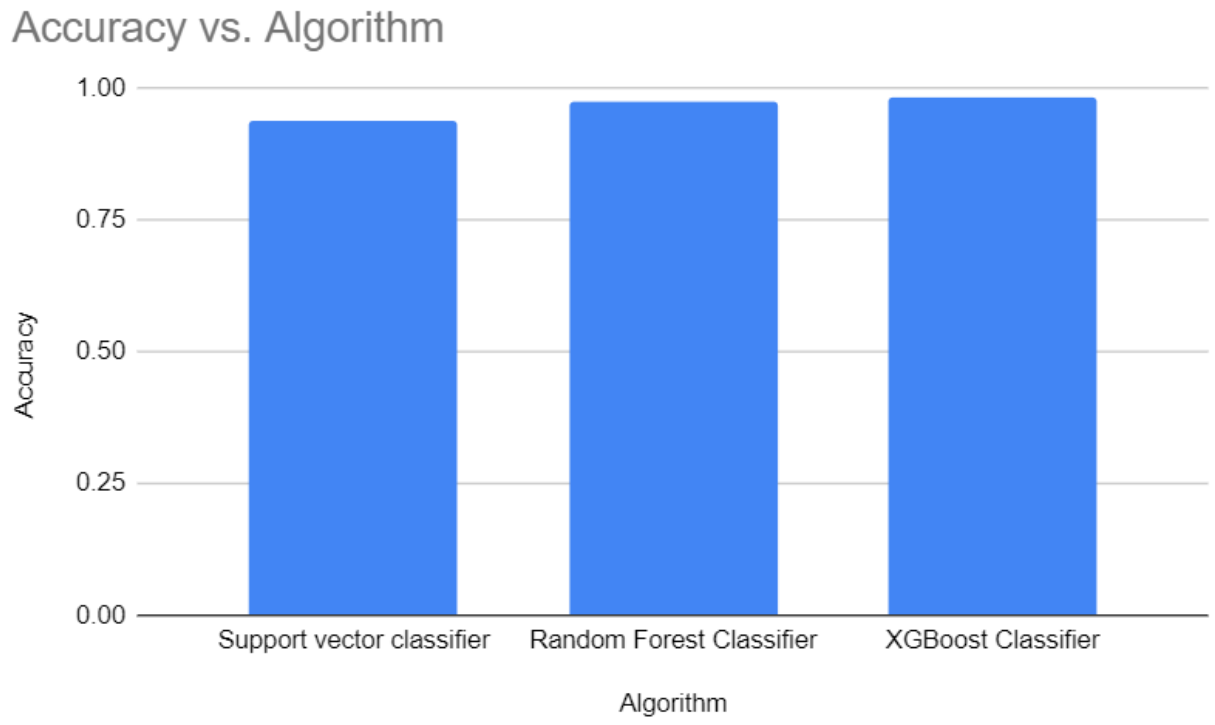



- It is important to define the target for the model.
- The output variable, also known as the target variable, is a binary variable that can be either malignant or benign.

**Result & Discussion:**

The individual's target variable determines the result. When the target score is "0," it means the person has either a malignant tumour or breast cancer; when it's "1, it means the person doesn't have either of those conditions." Three approaches were employed in the comparison: Support Vector Machine (SVM), Random Forest Classifier, and XGBoost. When compared to other calculations, XGBoost delivers the most notable exactness of 98.2%. In this sense, we contend that the XGBoost computation is the most appropriate technique for determining the likelihood of developing breast cancer in large datasets.






**Conclusion:**

Breast cancer is the form of cancer that occurs most frequently. A randomly selected woman has a 12% chance of being diagnosed with the disease. As a result, many precious lives can be saved through early detection of breast cancer. This paper proposes a model that compares various machine learning algorithms for the detection of breast cancer. Performance comparison of the machine learning algorithms techniques has been carried out using the 'UCI ML Breast Cancer Wisconsin (Diagnostic)' dataset.It has been observed that each of the algorithms had an accuracy of more than 94%, to determine benign tumour or malignant tumour. Due to its superior accuracy, precision, and F1 score over the other algorithms, it was discovered that XGBoost is the most efficient in the detection of breast cancer.

**References:**

[1]Breast Cancer Detection Using Machine Learning Algorithms
Authors: Shubham Sharma, Archit Aggarwal, Tanupriya Choudhury
University of Petroleum & Energy Studies (UPES)
Publication Date: July 2019
https://ieeexplore.ieee.org/document/8769187

[2]Prediction of benign and malignant breast cancer using data mining techniques
Authors: Vikas Chaurasia Saurabh Pal, and BB Tiwari
Publication Date: February 20, 2018
https://journals.sagepub.com/doi/10.1177/1748301818756225

[3]Breast Cancer Detection using Machine Learning
Authors: Vanlalhmangaihsanga, P.C. Vanlalbeiseia, Laledenthara
Publication Date:June-2021
https://www.jetir.org/view?paper=JETIR2106248

[4]An investigation of the XGBoost-based algorithm for breast cancer classification.
Author: Xin YuLiew
Publication Date:December 2021
https://reader.elsevier.com/reader/sd/pii/S2666827021000773?token=6050FD040083F665F89766CB1E269D6C12B38D7F8124B7F21364509E5299B2CAE58E7E000863837DF475F83860A72326&originRegion=eu-west-1&originCreation=20221129192156

[5]Performance Analysis of XGBoost Ensemble Methods for Survivability with the Classification of Breast Cancer
Author: T. R. Mahesh, V. Vinoth Kumar
Publication Date:Sept 2022
https://www.hindawi.com/journals/js/2022/4649510/