Breast Cancer Prediction using machine learning algorithms

Hima Bindu Desalanka
Computer Science
University of Central Missouri
hxd97040@ucmo.edu

Deepika Rani Gokavarapu
Computer Science
University of Central Missouri
dxg05990@ucmo.edu

Mounika Motupally
Computer Science
University of Central Missouri
mxm18550@ucmo.edu

Abstract:

Breast Cancer (BC) is one of the most diagnosed cancers in Women in recent years. Statistics suggest that there are more deaths in women because of BC and it is only increasing worldwide. Cancer is a disease in which the symptoms can only be seen or visible at the later stages. Having cancer detected at an early stage can reduce fatal risks. It is necessary to understand whether the group of tumors is Benign or Malignant. Knowing this can improve the rate at which the patient can be treated, reduce the treatment costs and especially reduce the risk. Machine Learning has seen a rise in value in many sectors in recent years and the health sector is one of the major sectors using a lot of machine learning algorithms. It has become of utmost importance to predict several health issues in advance to provide better health care, at least to minimize the risk involved, if not avoid it completely. It is also important to provide the right diagnosis of the issue. For example, it is always suggested to seek medical help if one feels like there is a tumor in the body, but it is important to note that not all tumors are cancerous.

In this paper, different ML algorithms such as Logistic regression classifier, KNN Classifier, Decision Tree, and Support vector Machine Classifier (SVM) are used to predict and classify the type of tumor. We took a dataset from Kaggle with information on patients diagnosed with BC and implemented the model using Python. This data contains several features of patients with tumors similar to the ones seen for breast cancer and the diagnosis (Benign/Malignant). A benign tumor is a type of tumor that is not actually cancerous and Malignant is a cancerous tumor. All the outcomes and the precision values are calculated and it is observed that the SVM classifier is providing us the most accurate results. Python's Sklearn library is used to develop the model and apply the ML algorithms and check the accuracy of each algorithm.

Keywords:
Supervised, Unsupervised, Classification, Regression, Clustering, Benign, Malignant

I. INTRODUCTION

Breast cancer is the most common cancer in women in the United States. This is the second leading cause of cancer death among women overall. It is just not a cancer that can occur only in women. There are cases of breast cancer in men as well. Statistics suggest that each year there are about 264,000 cases of breast cancer diagnosed in women and about 2,400 in men only in the United States. About 42,000 women and 500 men die each year from breast cancer. It is necessary to have an accurate diagnosis of such a disease.

Machine Learning is one of the most efficient technologies available in using the available data and drawing some conclusions and predictions. ML with its already existing algorithms has proved to be an asset in the health sector in diagnosing diseases. Several machine learning algorithms can be used based on the data type. Knowing which machine learning algorithm can be used for which type of problem depends on running the algorithms by changing the hyperparameters involved in the algorithms and normalizing the available data and comparing the results with the facts.

It is all about using the natural phenomena. Using the available data as training data and testing data and working through it by taking into consideration several features of the data, dividing the data such that we have all kinds of data available for training as well as testing is all part of Machine Learning.

Machine learning is divided into especially two categories.

1. Supervised learning
2.Unsupervised learning.

Supervised Learning:

This is a learning technique based on the labeled data. For this type of algorithm, we feed the input and we change the hyperparameters to get accurate results. An example of supervised learning include classifying the type of fruit based on the features which are already labeled or finding spam emails, etc.,

Supervised learning includes:
Classification problems
Regression/Prediction problems.

Classification Problems:

These are the kind of problems where we need to classify an object or number of objects based on its features. Some of the real-life examples of classification problems are:

1. Handwritten digit recognition: The goal in this is to recognize the exact digit from the images of handwritten digits. Everyone writes the digits differently. When a handwritten digit is scanned, the computer needs to understand the digit and put the correct number in. In the United States, this is mainly used when making an entry into the system by scanning a paper-based tax return filed.
2. Speech Recognition: The goal of this is to recognize the speech of a person and identify the correct word. A popular example of this is the automatic customer service provided by many providers such as T-Mobile. The T-Mobile automated customer service has several questions asked and popular of it is the Yes or No questions. Based on the frequency with which the customer answers the question, it recognizes whether the customer said a Yes or a No.

Regression Problems:

These are the problems in which we have some continuous labeled data and we need to do some predictions. Some of the real-time examples of regression problems are:

1. Weather prediction: The objective/goal in this is to predict the weather for a particular period for a specific location. In this, the data available for each day spanning for years is taken into consideration and we try to predict what the weather will be on a particular day at a particular time. Along with the vast data available, it is also necessary to consider the weather condition from the very recent days. All of these conditions are taken into account for predicting the weather.
2. Predicting the number of graduate students in a year: The objective here is to predict how many students graduate in a particular year. This is again based on the data available from so many years and the trend in the number over the years. One popular consideration of this is how the number changed after the pandemic from that before the pandemic. There is an effect of the pandemic on how many students opted to do the graduation. The job market also plays a major role in students choosing between pursuing a graduate degree in a certain year. All these variables need to be considered for predicting the number of students graduating in a particular year.

Some of the popular supervised learning algorithms are:
Simple Linear Classifier - Used for Classification and Regression problems.
Support Vector Machine (SVM) - Used for Classification problems.
Decision Trees - Used for Classification and Regression problems.
Random Forest - Used for Classification and Regression problems.
Naive Bias - Used for classification problems
K-Nearest Neighbor Algorithm (KNN) - Used for Classification and Regression problems.

Unsupervised Learning:

This is a learning technique based on unlabeled data. The algorithms in unsupervised learning work on the random data/raw data and try to group the data. This is also called clustering the data.

Unsupervised Learning includes:
Clustering
Dimensionality reduction

Clustering:

This is a technique of grouping the unlabeled data. We need to find the similarities between the data according to the characteristics seen in the data and group the same type of data into clusters/groups. A popular real-time example of clustering is retail marketing. A retail company may take the information on the income of a household, size of the household, occupation, and the distance of the house from the store and might group them into different types of clusters where the clusters for example can be one of the following:

> High income - High spenders
> Low income - high spenders
> High income - Low spenders

> Low income -Low spenders.

The retail company might have a different marketing strategy for each type of cluster/group and having the machine learning algorithms group these customers based on the information that the company gathered can help the company grow their business and provide better services.

Dimensionality Reduction:
This is a technique in which the dimension of the data is reduced by applying some machine learning algorithms. By reducing the dimension, it means out of thousands of features available in the raw data, we take only a certain number of main features with which we can get the correct output. This is needed for reducing the problem of overfitting.

A real-time example of dimensionality reduction can be the following:
Recognizing whether the person is a male or female. For example, Consider the features like height, weight, hair color, eye color, shape of the eyes, and shoe size of a person. We have raw data of say 100's of people with all these features. Out of these features, we can reduce the dimension, meaning take out irrelevant features, which doesn't add any value in figuring out whether the person is a male or a female. Applying the algorithms and reducing the dimension and calculating the output and finding the accuracy of the output and changing the dimension based on the predicted values are all part of the Dimensionality reduction.

Some of the popular unsupervised learning algorithms are:
K- means clustering - Clustering algorithm
K- Medoid clustering - Clustering algorithm
Hierarchical clustering - Clustering algorithm
Single-link hierarchical clustering
Average link hierarchical clustering
Complete link hierarchical clustering
Principal Component Analysis(PCA) - Dimensionality reduction
Linear Discriminant Analysis (LDA) - Dimensionality reduction

II. MOTIVATION

The field of study into breast cancer prediction using machine learning is quite significant and influential. We can advance the early detection, diagnosis, and treatment of breast cancer by utilizing data analytic methods. Here are some inspiring ideas to think about:

Early identification improves survival rates and prevents the loss of life from breast cancer, one of the most prevalent cancers in the world. Large-scale patient data analysis using machine learning algorithms has the ability to spot patterns, create precise prediction models, and aid in the early detection of breast cancer. You have the chance to have a positive influence on the lives of many people by making a contribution to this area. Machine learning algorithms are able to learn from enormous datasets and recognize complicated patterns that may not be immediately apparent to people. We can improve the precision and effectiveness of screening procedures by creating reliable breast cancer prediction algorithms. This can reduce the likelihood of false positives and negatives, resulting in more accurate diagnoses and fewer pointless medical procedures. Breast cancer costs both patients and healthcare institutions a lot of money. We can enable early intervention and therapy and potentially lower the total cost of care by creating precise prediction models. Additionally, machine learning can aid in resource optimization, enhancing the effectiveness and accessibility of healthcare. The important thing to remember is that breast cancer affects millions of individuals all over the world. By combining your technical knowledge with the desire to enhance healthcare outcomes, we can play a critical part in the fight against breast cancer**.**

III. Main Contributions & Objectives

The objective of our project is as follows:

- It predicts whether the patients are suffering from breast cancer or not.
- Based on the independent variables it predicts the stage of breast cancer i.e., either benign or malignant.
- Patients will be aware of breast cancer in the starting stage itself by our model.
- Improves survival and quality of life.
- Decreases the cost.

- Less time-consuming.
- The ability of machine learning to assess the massive amount of data produced by the healthcare sector and the efficacy of its decision-making processes are key factors in the industry's success.

IV. Related Work

With the development of technology, many contemporary methods for breast cancer prediction have emerged. The following is a brief description of the work in this field.

In the study [1], the obtained results demonstrate that Naive Bayes is the calculation with the best classification precision of 97%, followed by RBF Network and J48, both of which have classification precisions of 96.77% and 93.41%, respectively.

On the Wisconsin Breast Cancer datasets, RBF, SVM, SL, NB, KNN, Adaboost, Fuzzy, and J48 were applied. The accuracy and viability of those calculations that use exactness, TPR, FPR, accuracy, and F-measure to determine the calculation with precision were examined in terms of their intensity and viability. SVM had the best performance, with an associate degree in nursing accuracy of 97.07%. It ends by stating that SVM has established itself as a financially advantageous and successful algorithmic program for investigating and reaching a diagnosis of cancer

Azar et al.'s [3] innovative method for finding breast cancer was proposed. The method made use of the radial basis function (RBF), probabilistic neural networks (PNN), and multi-layer perceptron (MLP) classification methods.The method also utilized testing methods and trained the properties of the breast cancer dataset. Accuracy, specificity, sensitivity, and other machine learning performance measure indices are used to calculate the system's performance. For training and testing, MLP's accuracy was 97.80% and 97.66%, respectively.

To detect breast cancer, a system based on non-contact, non-invasive screening technology was suggested [6]. Their solution made use of the support vector machine (SVM) classifier and four (Best of Parameters) features. A wearable breast cancer screening (BCS) device was developed by many researchers in an effort to replace frequent in-hospital screening tests or CBEs, which are needed to identify the disease sooner. The accuracy of the suggested model was 83.3%.

According to Baku Bektas et al. [7], it is essential to identify breast cancer early in order to save lives and take the necessary steps to control the disease. The most common cause of mortality for women is breast cancer, which primarily affects females. They largely made use of 139 patient characteristics. The methodologies in this paper included random forest, k-stars, and selective sensor neural networks. The outcome indicated 61.85% accuracy.

Using the data mining tool WEKA, Tanaya Padhi et al. [9] examined the prediction of breast cancer. The data mining phases that were used on the dataset were defined in the paper. The system primarily processed 6 patient features. They used the J48, REP Tree, and Naive Bayes algorithms for categorization. The results showed that REP Tree has an accuracy of 72.02%.

The suggested SVM [12] is based on the division of classes. The suggested SVM model calculates F-scores to choose input features and uses grid search to optimize parameters, yielding a 5. These subset values are trained when the subset is produced using grid search using 10 fold. Up till every feature is present in the subset, the process is repeated. Applying the suggested model and feature selection resulted in the greatest accuracy of 99.51% with an 80–20% split between training and testing data.

The output is taken into account as a target class by the suggested kNN [13] algorithm. Performance Component Analysis (PCA) is used to choose features for the input characteristics, which results in a smaller data set with more relevant features. The qualified majority of the data's neighbors (k) will then treat the input as a reduced and truly valued positive integer. The output class's exposure to noise is minimized if the k value, which is calculated depending on the input, is sufficiently high. The lazy-learning algorithm known as kNN only performs calculations locally and delays the start of the learning cycle until the test step is complete. The accuracy of the suggested model PCA with kNN is 95.95%.

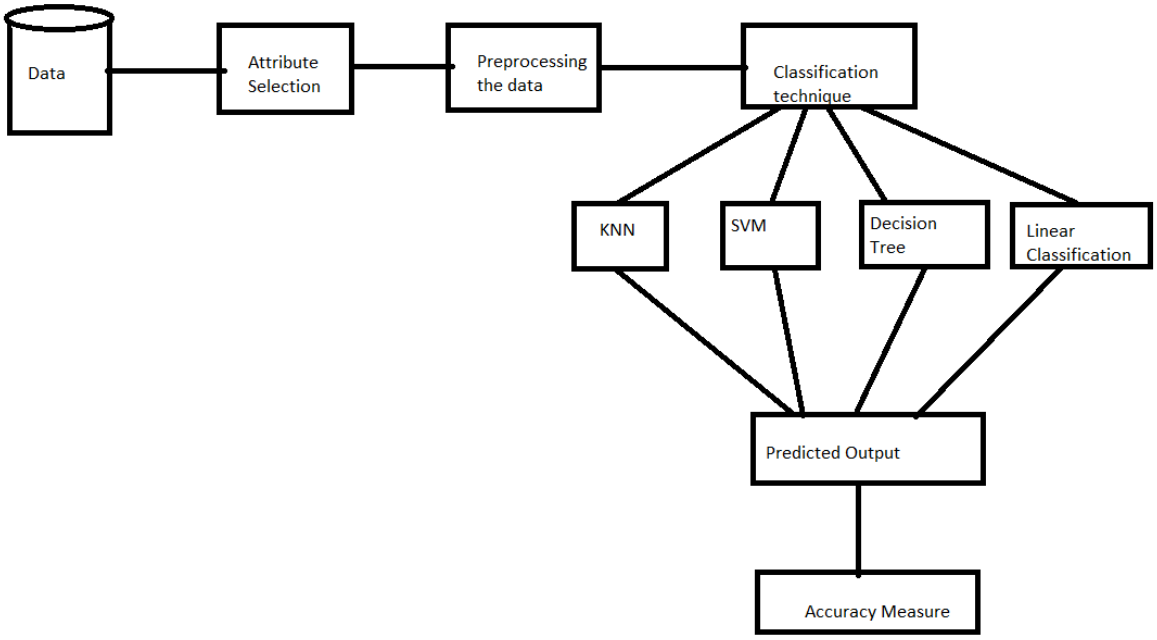
Weka is identified as the best tool for data processing approaches in this paper's conclusion.In comparison to other classification methods, J48 provides predictions that are 74.28% accurate, whereas SMO provides predictions that are 70.60% accurate for the polygenic illness dataset. It assesses every accuracy metric for the polygenic disease dataset and the carcinoma dataset, both of which contain 768 in stances and nine attributes.

Here, in our research, we aimed to demonstrate a comparison among the Machine Learning models. In order to compare the effectiveness of the analysis, we also modified a few parameters

V. Proposed Framework:

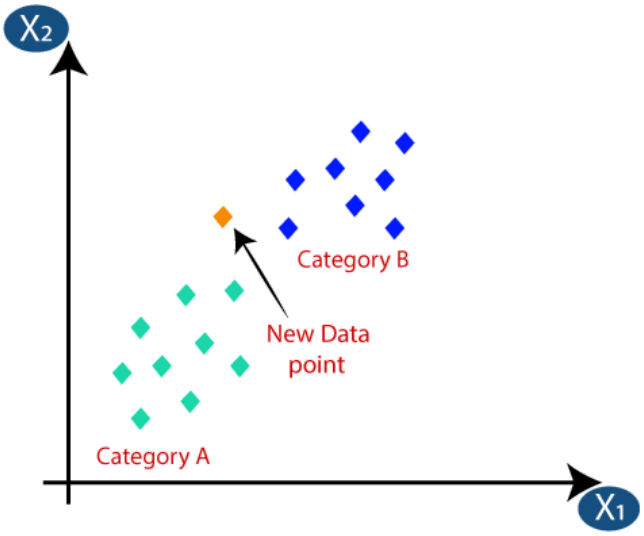
The process flow is as follows.

- The data is collected from Kaggle.
- Used Python's sklearn package to import the available machine-learning algorithms
- Used the following mentioned machine learning algorithms to analyze the data and generate outputs.
- Calculated the accuracy of the output generated.



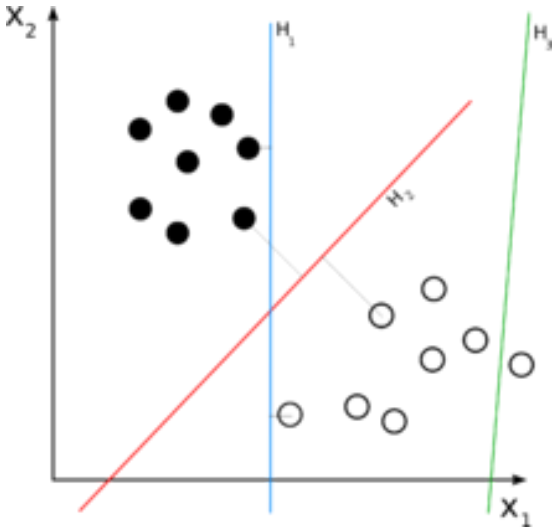
KNN Algorithm (K-Nearest Neighbor Algorithm):

KNN works on the basis of the distance between the data points. For a data point in the set, every other data point is called its neighbor. The distance between neighbors is calculated for all the data points. The class of the nearest neighbors for a data point is predicted as the class for a data point. We generally take K as an odd value. Suppose, the value of K is 3, then for a data point, the nearest 3 neighbors are checked and the class of a maximum number of the 3 nearest neighbors becomes the class for the data point. The value of K is changed to calculate the output and compared against the actual values and the K value is decided based on the level of accuracy.



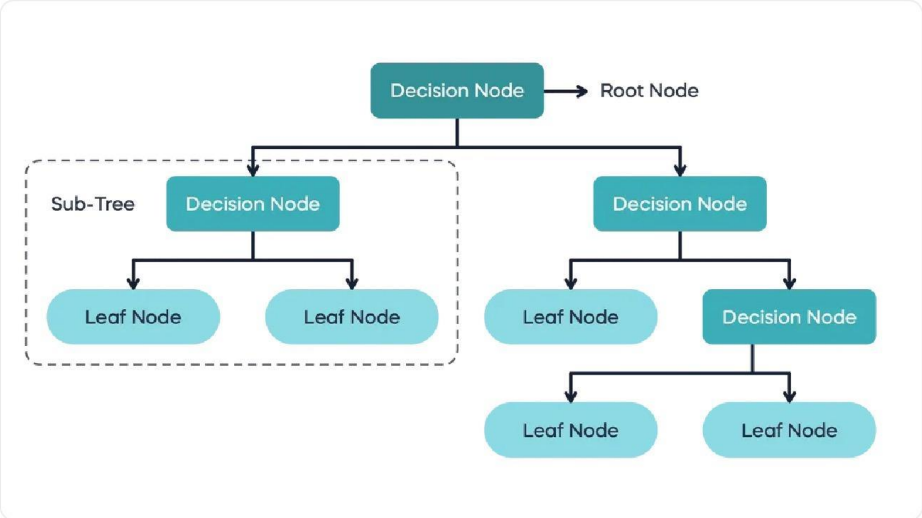
Simple Linear Classifier:

A simple linear classifier solves the data which can be linearly separable. In other words, if we have a data set where one class of data can be differentiated from the other class just by drawing a straight line between the data points then we can use Simple Linear Classifier to classify such data.

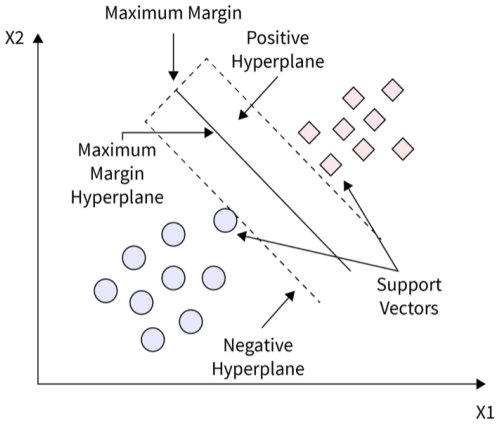


Decision Tree:

In this algorithm, a tree-like structure is created using the data. The data first is split into subsets. We pick one attribute and check if the subsets are pure. If yes, we stop there or we continue with the split. We also stop the process, when we reach a point where further splitting of data adds no value. The splitting continues based on the entropy (probability) and the information gain calculated at each level.



Support Vector Machine (SVM):
SVM is a supervised machine learning algorithm, which can be used for both classification and regression problems. In this, we draw a hyperplane between the data points of two classes. Unlike perceptron, another machine learning algorithm that also works on hyperplane, SVM tries to maximize the hyperplane, which also means that it tries to draw the hyperplane in such a way that we have maximum distance between the hyperplane and the data points of the class. This distance is also called Margin.



Area Under Curve:
Area Under Curve(AUC) is one of the most widely used metrics for evaluation.It is used for binary classification problems. AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. Before defining AUC, let us understand two basic terms They are :
1.Sensivity.
2.Specificity.

True Positive Rate (Sensitivity) :
True Positive Rate is defined as TP/ (FN+TP). True Positive Rate corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points.

False Positive Rate (Specificity):
False Positive Rate is defined as FP / (FP+TN). False Positive Rate corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points.

VI . Data Description

In our dataset we have the following attributes which are shown below.

ID number:
The unique identification number given for each patient in order to identify them easily.
Diagnosis:
(M = malignant, B = benign) The dependent variable in our dataset in diagnosis. it's two types they are benign and malignant. benign is nothing but the cancer cell is in the beginning stage and malignant means it's in severe stage.

radius_mean:
mean of distances from center to points on the perimeter texture_mean:
standard deviation of gray-scale value.

perimeter_mean:
mean size of the core tumor area_mean, smoothness_mean:
mean of local variation in radius lengths compactness_mean:
mean of perimeter^2 / area - 1.0

concavity_mean:
mean of severity of concave portions of the contour concave points_mean:
mean for number of concave portions of the contour symmetry_mean, fractal_dimension_mean:
Mean for "coastline approximation" - 1 radius_sestandard error:
For the mean of distances from center to points on the perimeter

texture_sestandard error:
For standard deviation of gray-scale values
perimeter_se, area_se, smoothness_sestandard error:
For local variation in radius length.
compactness_sestandard error:
For perimeter^2 / area - 1.0
concavity_sestandard error:
For severity of concave portions of the contour

concave points_sestandard error:
For number of concave portions of the contour symmetry_se, fractal_dimension_sestandard error:
For "coastline approximation" - 1.

radius_worst:
"Worst" or largest mean value for mean of distances from center to points on the perimeter

texture_worst:
"Worst" or largest mean value for standard deviation of gray-scale values.

perimeter_worst, area_worst, smoothness_worst:
"Worst" or largest mean value for local variation in radius lengths.

compactness_worst:
"Worst" or largest mean value for perimeter^2 / area - 1.0

concavity_worst:
"Worst" or largest mean value for severity of concave portions of the contour

concave points_worst:
"Worst" or largest mean value for number of concave portions of the contour

symmetry_worst, fractal_dimension_worst:
"Worst" or largest mean value for "coastline approximation" – 1

Dataset link: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

This dataset consists of 569 rows and 32 columns. Sample data figure is attached below:

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776
842517	M	20.57	17.77	132.9	1326	0.08474	0.07864
84300903	M	19.69	21.25	130	1203	0.1096	0.1599
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839
84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328
843786	M	12.45	15.7	82.57	477.1	0.1278	0.17
844359	M	18.25	19.98	119.6	1040	0.09463	0.109
84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645
844981	M	13	21.82	87.5	519.8	0.1273	0.1932
84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396
845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669
84610002	M	15.78	17.89	103.6	781	0.0971	0.1292
846226	M	19.17	24.8	132.4	1123	0.0974	0.2458
846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002
84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293

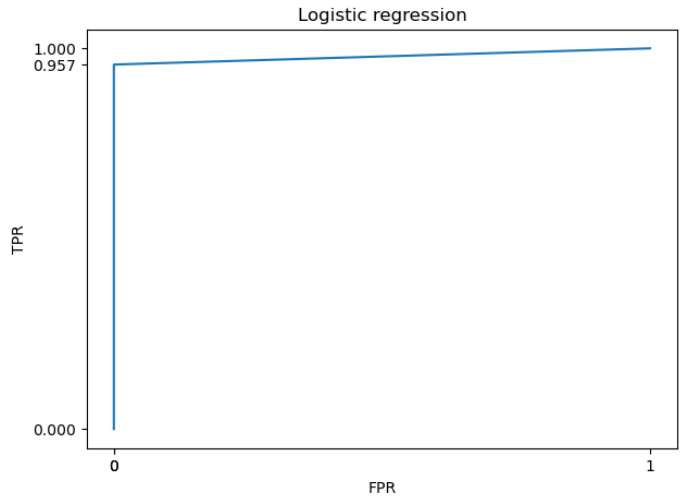
Ⅶ. Experimentation & Comparison

We train our model with different kinds of algorithms. The model with high accuracy shall be considered for deployment purpose. The below table shows the accuracy values of the different types of classification algorithms namely Logistic Regression, SVM, K-nearest neighbor, Decision Tree. By checking the accuracy of the trained data. The accuracy is more for Gaussian SVM.

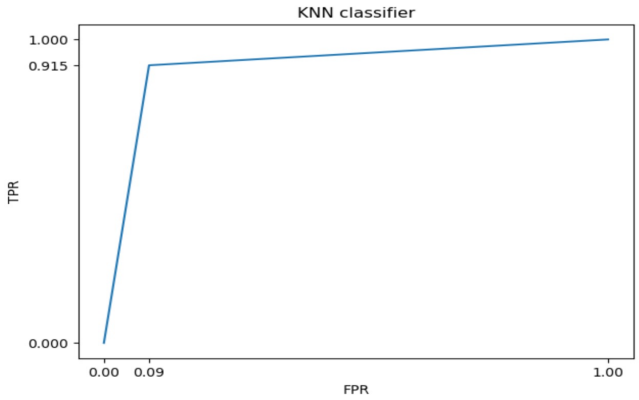
Comparison of algorithms:

To identify the best model which gives the best performance and we build different models with different algorithms. they are shown below

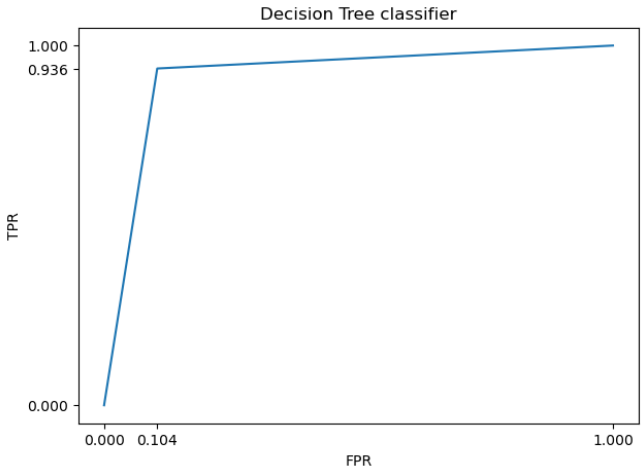
Logistic Regression:



K-NN Algorithm:



Decision Tree:



SVM:

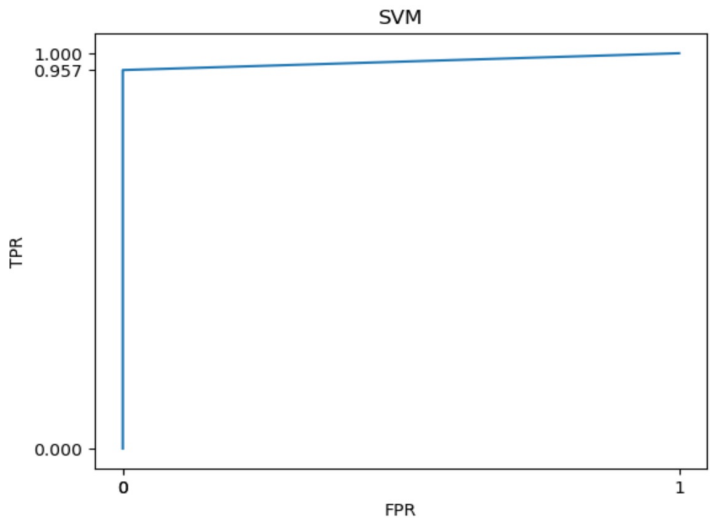


Table 1: Details about the classification algorithm and their accuracies.

Classification Algorithm	Accuracy
Logistic Regression	0.94
Decision Tree	0.91
K-NN	0.91
SVM	0.98

References

1. Vikas Chaurasia1, Saurabh Pal1 and BB Tiwari2, "Prediction of benign and malignant breast cancer using data mining techniques", Journal of Algorithms & Computational Technology 2018, Vol. 12(2) 119–126 ! The Author(s) 2018,pp. 119-126.

2. Prerita; Nidhi Sindhwani; Ajay Rana; Alka Chaudhary: Breast Cancer Detection using Machine Learning Algorithms(2021). https://ieeexplore.ieee.org/document/9596295

3. Deepika Verma(M. Tech Scholar), Dr. Nidhi Mishra(AssociateProfessor), Poornima University Jaipur, India, "Analysis and Prediction of Breast cancer and Diabetes disease datasets using Data mining classification Techniques", IEEE Xplore Compliant – Part Number:CFP17M19-ART, ISBN:978-1-5386-1959-9, pp. 533-538.

4. Mamatha Sai Yarabarla; Lakshmi Kavya Ravi; A. Sivasangari: Breast Cancer Prediction via Machine Learning(2019). https://ieeexplore.ieee.org/document/8862533

5. Alok Chauhan; Harshwardhan Kharpate; Yogesh Narekar; Sakshi Gulhane; Tanvi Virulkar; Yamini Hedau: Breast Cancer Detection and Prediction using Machine Learning(2021). https://ieeexplore.ieee.org/document/9544687

6. A. JT, "Breast cancer in sub-Saharan African women", PubMed - NCBI, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/7839882. [Accessed: 08-Mar- 2020].

7. B. Majeed, H. T. Iqbal, U. Khan and M. A. Bin Altaf, "A Portable Thermogram based Non-contact Non-invasive Early Breast-Cancer

Screening Device"in 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS), Cleveland, OH, 2018, pp. 1-4.

8. Nidhi Mishra, Naresh Khuriwal "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm", 2018 IEEMA Engineer Infinite Conference (TechNxT), 2018

9. A. Bharat, N. Pooja and R. A. Reddy, "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis", in 20183rd International Conference on Circuits, Control, Communication and Computing (I4C), Bangalore, India, 2018, pp. 1-4

10. "What Is Breast Cancer? — CDC", CDC, 2020. [Online]. Available: https://www.cdc.gov/cancer/breast/basic info/what-is-breastcancer.htm. [Accessed: 01- Feb- 2020].

11. "Breast Cancer", Wikipedia, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Breast cancer. [Accessed: 02-Feb2020].

12. Md.Milon Islam, Md. Rezwanul Haque, Hasib Iqbal, Mahmudul Hasan, "Breast cancer prediction: A comparative study using machine learning techniques", SN Computer Science: A Springer Nature Journal, 2020

13. Harikumar Rajaguru, Sannasi Chakravarthy S R, "Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer", Asian Pacific Journal of Cancer Prevention, Vol 20(12), pp.3777–3781, 2019.

14. A. JT, "Breast cancer in sub-Saharan African women", PubMed - NCBI, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/7839882. [Accessed: 08- Mar- 2020].

15. B. Majeed, H. T. Iqbal, U. Khan and M. A. Bin Altaf, “A Portable Thermogram based Non-contact Non-invasive Early Breast-Cancer Screening Device”in 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS), Cleveland, OH, 2018, pp. 1-4.

16. T. Padhi and P. Kumar, “Breast Cancer Analysis Using WEKA”in 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 229-232.

17. “How the Trimmed Mean Is Used”, Investopedia, 2020. [Online]. Available: https://www.investopedia.com/terms/t/trimmed mean.asp. [Accessed: 14- Mar- 2020].

18. “How Random Forest Algorithm Works in Machine Learning”, Medium, 2020. [Online]. Available: https://medium.com/@Synced/how-randomforest-algorithm-works-in-machine-learning-3c0fe15b6674. [Accessed: 14- Mar- 2020].

19. “XGBoost Algorithm: Long May She Reign!”, Medium, 2020. [Online]. Available: https://towardsdatascience.com/https-medium-comvishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d. [Accessed: 14- Mar- 2020]. 20. “Accuracy, Recall, Precision, F-Score & Specificity, which to optimize on?”, Medium, 2020. [Online]. Available: https://towardsdatascience.com/accuracy-recall-precision-f-scorespecificity-which-to-optimize-on-867d3f11124. [Accessed: 14- Mar2020].

21. M. Islam, M. Raihan, S. Akash, F. Farzana and N. Aktar, “Diabetes Mellitus Prediction Using Ensemble Machine Learning Techniques”, Advances in Computational Intelligence, Security and Internet of Things, vol. 1192, pp. 453-467, 2020. Available: 10.1007/978-981-15- 3666-3 37 [Accessed 10 April 2020].

22. Dada Emmanuel Gbenga*1, Ngene Christopher1, Daramola Comfort Yetunde 2, 1Department of Computer Engineering, University of Maiduguri, Maiduguri, Nigeria 2Department of Computer Science,University of Oye Ekiti, Oye Ekiti, Ekiti State, Nigeria *Corresponding Author: Dada Emmanuel Gbenga, Department of Computer Engineering, University of Maiduguri, Maiduguri, Nigeria. “Performance Comparison of Machine Learning Techniques for Breast Cancer Detection”, Nova Journal of Engineering and Applied Sciences, pp. 1-8

23. Anuj Mangal; Vinod Jain: Prediction of Breast Cancer using Machine Learning Algorithms(2021). https://ieeexplore.ieee.org/document/9640813

24. M. Raihan, Muhammad Muinul Islam, Promila Ghosh, Shakil Ahmed Shaj, Mubtasim Rafid Chowdhury, Saikat Mondal, Arun More, “A Comprehensive Analysis on Risk Prediction of Acute Coronary Syndrome Using Machine Learning Approaches”, in 2018 21st International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2018, pp. 1 - 6.

25. “Everything You Need to Know About Breast Cancer”, Healthline, 2020. [Online]. Available: https://www.healthline.com/health/breastcancer#prevention. [Accessed: 08- Feb- 2020].

26. “Breast Cancer - Stages”, Cancer.Net, 2020. [Online]. Available: https://www.cancer.net/cancer-types/breast-cancer/stages. [Accessed: 08- Feb- 2020].

27. Vinod Jain; Mayank Agrawal. Breast Cancer Prediction Using Advance Machine Learning Algorithms (2022). https://ieeexplore.ieee.org/document/9785112

28.Sajib Kabiraj; M. Raihan; Nasif Alvi; Marina Afrin; Laboni Akter; Shawmi Akhter Sohagi; Etu Podder: Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm (2020). https://ieeexplore.ieee.org/document/9225451

30. Udit Pratap; Saurabh Chhabra: Breast Cancer Prediction using Different Machine Learning Algorithms(2021). https://ieeexplore.ieee.org/document/9725688