BREAST CANCER DETECTION

Bachelor of Technology
in
Information Technology

Submitted by
Riya Goyal (IIT2019096)
Maitry Jadiya (IIT2019100)
Sonal (IIT2019122)
Dev Bansal (IIT2019132)
Saloni Singla (IIB2019004)

Under the supervision of
Prof. Pritish Varadwaj

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY,
ALLAHABAD 211015
(U.P.) INDIA

Declaration

We hereby declare that the word presented in this project report entitled "**Breast Cancer Detection** " was submitted towards the fulfillment of 5th Semester Project Report (2021) of **B.Tech** in **InformationTechnology** at the INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD, U.P., INDIA is an authenticated record of my original work carried out from August 2021 to November 2021 under the supervision of **Prof. Pritish Varadwaj**. The project is being accomplished in full compliance with the requirements and constraints of the prescribed curriculum.

Place: Allahabad
Date : 02/12/2021

Submitted by:-
Riya Goyal (IIT2019096)
Maitry Jadiya (IIT2019100)
Sonal (IIT2019122)
Dev Bansal (IIT2019132)
Saloni Singla (IIB2019004)

Certificate

It is certified that the work contained in the Project titled “**Breast Cancer Detection**” by “Riya Goyal, Maitry Jadiya, Sonal, Dev Bansal, Saloni Singla" has been carried out under my/our supervision and that this work has not been submitted elsewhere.

Signature of Supervisor

Date : …………………..

Prof. Pritish Varadwaj

Acknowledgment

I would like to express my gratitude to the mercy of Almighty that gave me enough strength to complete this project work. I am very thankful to my supervisor "**Prof. Pritish Varadwaj**", Professor, Department of Applied Sciences, IIIT-Allahabad to guide me throughout the project work and always motivating me to work hard. The regular discussions and extensive comments helped to improve my project and a direct impact can be seen on the project. I am thankful to **Dr. Pawan Chakraborty, Head of Department (IT)**, IIIT-Allahabad, for providing valuable guidance at the beginning of the project. I am also grateful to all the faculty members and supporting staff who have been very supportive in providing resources to complete this project.

Abstract

Breast cancer is a deadly disease that has been emerged as the second leading cause of cancer deaths in women globally. The annual mortality rate is estimated to continue to rise. Early detection of cancer could significantly reduce breast cancer death rates in long run.

Proposing strategies to support effective treatment of the disease has undoubtedly become a priority for the government, health facilities, and the general public.

The rate of efficiency obtained on the Wisconsin UCI breast cancer dataset was 97.31%. The ANN's performance was compared with the performance of a support vector machine (SVM) model, Decision Trees (DT), naive Bayes (NB), k-nearest neighbor (k-NN).

According to our results, the SVM performed best. The algorithm of the SVM was written and executed in a Python platform, as well as the experimental and comparativeness between algorithms.

Keywords:

- Breast Cancer Detection (BCD);
- Associative Memory (AM);
- Associative Processing (AP);
- K-nearest neighbours (k-NN);
- Support Vector Machine (SVM);
- Classification and Regression Trees (CART);
- Gaussian Naive Bayes (NB).

Contents

Introduction

Breast cancer has become a common disease among women around the world and is considered the second-largest prevalent type of cancer that causes deaths among women. However, it is also considered the most curable cancer type as long as it can be diagnosed early.

Tumors are basically a lump of extra tissues which is formed by a group of rapidly dividing cells and it occurs when cells divide and multiply excessively in the body. The division and growth of cells are controlled by our body. New cells are created to replace old ones or to perform new functions and cells that are damaged or no longer need to die to give way to healthy replacement cells. Tumors can be formed if the balance of cell division and death is disturbed. Breast cancer can be of two types, the invasive or non-invasive type, and can occur in both men and women, although in men it is a hundred times less common than in women. The risk factors for developing breast cancer are many, for example, gender, followed by age, obesity, physical activity, diet, alcohol consumption, and vitamin D concentration. Although vitamin D has emerged as a potentially important determinant of breast cancer, information is still scarce. Some studies show that it can be a risk factor, while others have shown that it is not. The exact reasons for breast cancer development are unknown.

Malignant tumors are considered a dangerous group and they can penetrate and destroy healthy body tissues and the term, breast cancer, refers to a malignant tumor. Malignant tumors can penetrate and destroy healthy body tissues. World Health Organization (WHO) statistics show that there are more than 1.2 billion women around the world which are diagnosed with breast cancer. In recent years, this graph has been reduced due to effective machine learning techniques.

The advancement of data-driven techniques has introduced effective ways in the area of breast cancer diagnostics. Powerful expert and data-driven methods: Support Vector Machine (SVM), Bayesian Network, k- Nearest Neighbours (k-NN), etc. It goes without saying that data evaluation that has been attained from the patients can be considered as an important factor to develop an efficient and accurate diagnostic method. Classification algorithms have been utilized to minimize the error of humans.

Problem Statement

To be able to detect Breast Cancer using machine learning. Also, analyze how the different algorithms perform on the given dataset.

The main objectives set to be accomplished are as follows:

- To implement an algorithm for detecting breast cancer detection.
- To examine the factors that may lead to the improvement of the detection rate of breast cancer in the given dataset.
- To analyze the impact of the selected classifiers on the algorithm's performance. Also, predict the possible future scope of the algorithm.

Literature Survey

1. **Diagnosis of Breast Cancer using Decision Tree Models and SVM:**
   By: Puneet Yadav, Rajat Varshney, Vishan Kumar Gupta
   Date: March 3, 2018

   In this study, machine learning algorithms like decision tree and Support Vector Machine (SVM) were tested using breast cancer Wisconsin data which contains records of 699 patients in which the dataset contains 458 records of benign tumors and 241 of malignant tumors. set and then compared to the result. In this study, two powerful classification algorithms, decision tree, and Artificial Neural Network have been applied for breast cancer prediction. Experimental results show that algorithms have effective results for this purpose with the overall prediction accuracy of the decision tree being from 90% to 94%, and SVM having 94.5% to 97% respectively. It shows that the Decision tree algorithm creates user-friendly rules that indicate important attributes and require less computation compared to other algorithms such as Neural Networks.

2. **Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction:**
   By: Yixuan Li, Zixuan Chen
   Date: October 18, 2018

   Firstly, in this study, they've collected the data of the BCCD dataset which contains 116 volunteers with 9 attributes as shown in the dataset, and data of the WBCD dataset another dataset on which they've run the model which contains 699 volunteers and 11 attributes. Then we've preprocessed the raw data of the WBCD dataset and obtained the info that contains 683 volunteers with 9 attributes and therefore the index shows whether the volunteer has a malignant tumor or benign tumor. Then on comparing the accuracy, RF measure metric, and Radius Of Curvature curve of 5 classification models, the result proved that RF is taken as the primary classification model. Here, This study shows there are still some limitations that need to be solved. The shortage in data has an impact on the accuracy of the models implemented. Additionally, the RF can also be merged with other data mining techniques to get more accurate, better, and efficient results.

3. **Machine learning with Applications in Breast Cancer Detection and Prognosis**

   By:-Wenbin Yue, Zidong Wang, Hongwei Chen, Annette Payne and Xiaohui Liu

   Date:- May 9, 2018

   In this paper, the authors have provided explanations of various Machine Learning approaches and their applications in Breast Cancer diagnosis and also prognosis that will analyze the information within the benchmark database WBCD that they've used. Machine Learning techniques have shown that their remarkable ability to enhance classification and prediction accuracy has taken a toll. Although some algorithms have achieved very high accuracy in the database used, the event of improved algorithms still remains necessary. Classification accuracy plays a vital assessment criterion but it's not the only one that is used. Different algorithms consider different aspects in this paper and accordingly they have different mechanisms. Further, we learned that for several decades ANNs have dominated Breast Cancer diagnosis and prognosis, but from this paper, we can see that alternative ML method when applied give better results and are more useful in intelligent healthcare systems to supply a spread of options to physicians.

4. **On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset**

   By:- Abien Fred M. Agarap

   Date:- February 9, 2019

   In this paper, six machine learning algorithms are used for the detection of cancer which is:- GRU-SVM, Linear Regression, Multilayer Perceptron, Nearest Neighbour search, Softmax Regression, and Support Vector Machine. These are performed on the Wisconsin Diagnostic Breast Cancer dataset by measuring their classification test accuracy, and their sensitivity and specificity values. For the implementation of the ML algorithms, the dataset was partitioned as 70 percent for the training phase, and 30 percent for the testing phase. The results of all the used ML algorithms exhibited high performance on the binary classification of the tumor, i.e. a benign tumor or malignant tumor. Such appliance won't only provide a much more accurate measure of model prediction performance, but it'll also help in assisting and determining the foremost optimal hyperparameters for the Machine Learning algorithms.

5. **Comparative Studies on Breast Cancer Classifications with K-Fold Cross Validations Using Machine Learning Techniques.**

   By:- Zahra Nematzadeh, Roliana Ibrahim, Ali Selamat

   Date:- January 2015

   In this paper, the author has investigated the impact of k in k-fold cross-validation. Decision tree, naïve Bayes, neural network, and Support Vector Machine algorithm with three different kernel functions is used as a classifier to classify original and prognostic Wisconsin breast cancer. It was previously suggested that the common choice is k=10 but the overall results showed we cannot always expect to have a more accurate result by increasing or decreasing the number of folds. Sometimes the increase will enhance the accuracy and sometimes it will only add to the computational cost. Also, there is not any relation between increasing or decreasing the value of K in KCV and increasing or decreasing the accuracy performance. Thus, it is an issue to do experiments with different values of K to find out the best accuracy. The exact value of K can be achieved using evolutionary algorithms such as Genetic Algorithm and Particle Swarm Optimization method.

Dataset

We have taken the Wisconsin UCI breast cancer dataset available on Kaggle. It contains records of 699 patients in which the dataset contains 458 records of benign tumors and 241 of malignant tumors. Further, the dataset contains 11 features, listed down below:-

S.NO	Feature	Description	Value of Attributes
1	ID	Unique Key	Unique
2	Clump thickness	Cancerous cells are grouped often in multilayers, while benign cells are grouped in monolayers.	1-10
3	Uniformity of cell size	Cancer cells vary in size, the larger the size, the more possibility of cancer to be malignant is there.	1-10
4	Uniformity of cell shapes	Cancer cells vary in shape.	1-10
5	Marginal adhesion	Normal cells tend to stick together, while cancer cells fail to do that	1-10
6	Single Epithelial Cell Size	Epithelial cells that are enlarged may be malignant cells. In benign tumors, nuclei are often not surrounded by the rest of the cell	1-10
7	Bare Nuclei	The cells without cytoplasm coating, are mostly found in benign tumors;	1-10
8	Bland Chromatin	The texture of the nucleus in benign cells.	1-10
9	Normal Nucleoli	Nucleus small structures that are barely visible in normal cells	1-10

10	Mitoses	The process of cell division	1-10
11	Class	Indication of a tumor category	2 for Benign 4 for Malignant

Prerequisites/Learning

5.1 Machine Learning

Machine learning is an application of AI (AI) that gives systems the power to automatically learn and improve from experience without being manually programmed. Machine learning focuses and depends on the event of computer programs that will access the data provided and use it to learn for themselves. The method of learning begins with data or datasets, examples, experiences, or instructions, so they can then figure out a pattern and or improve them in the near future, if necessary

5.2 Algorithms Used

5.2.1 Naive Bayes:

A Naive Bayes classifier is a probabilistic machine learning model and it is used for the classification tasks in Machine Learning. The crux of the classifier is predicted using the Bayes theorem. Using Bayes theorem, we find the probability of an event, as long as B has occurred. Here, A is the hypothesis and B is the evidence required. The idea made here is that the features are independent and there is at least the presence of one particular feature that doesn't affect the opposite. Hence it's called Naive.

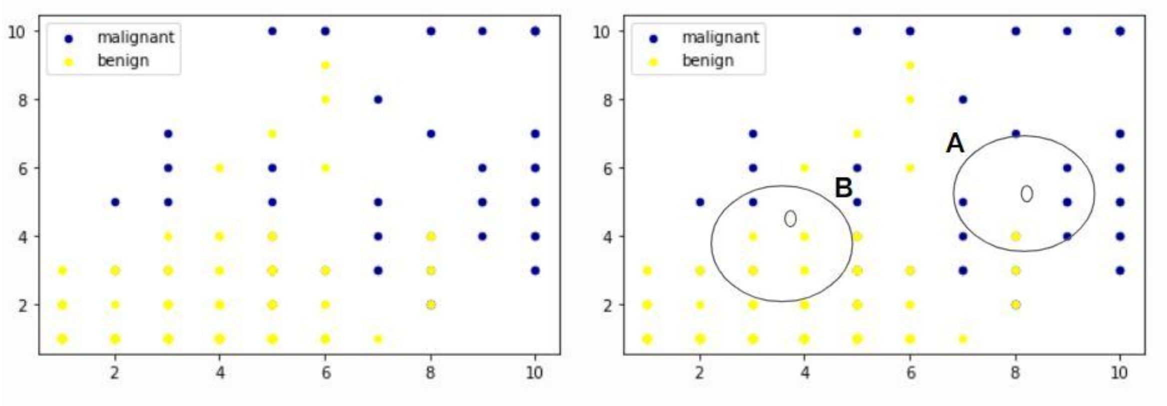


Image Taken From [colab]

Prior Probability:

P(M) = 241/699 P(B)=458/699

Likelihood Probability:

P(Mna) = 4/5	P(Bna)=1/5
P(Mnb)= 1/5	P(Bnb)=4/5

→ P(Am)=0.275 P(Ab)=0.131
White Circle is probably a Malignant tumor

→ P(Bm)=0.069 P(Bb)=0.524
White Circle is probably a Benign tumor

5.2.2 Support Vector Machine:

Support Vector Machine (SVM) is a supervised machine learning algorithm that is used for both classification and regression problems. In this algorithm, we plot each data item as a point in n-dimensional space where n is the number of features you have with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

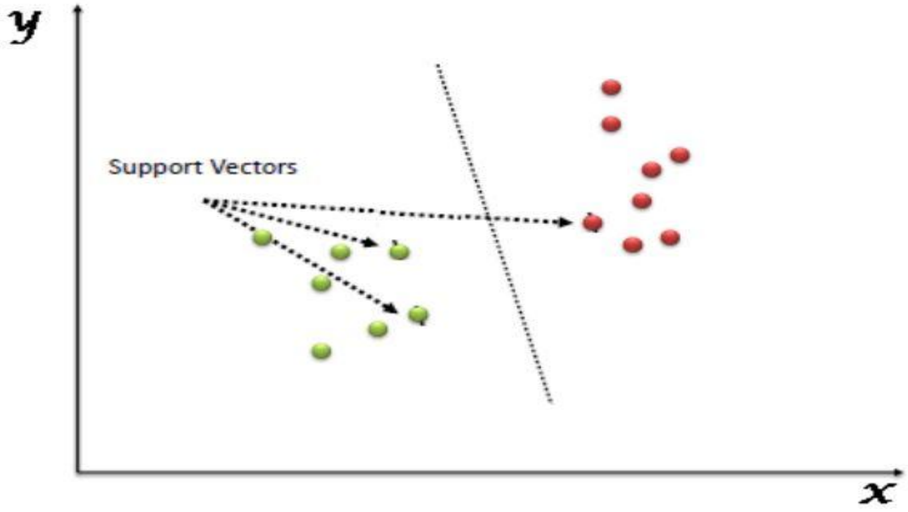


Image Taken From [1]

Support Vectors are simply the coordinates of individual examination. Support Vector Machine is a borderline that best segregates the two classes. Working of SVM as we got routine to the process of segregating the two classes (malignant and benign) with a hyper-plane. Identify the right hyper-plane here we have three hyperplanes A, B, and C. Now, identify the right hyper-plane to classify star (malignant) and circle (benign). You need to remember a rule to identify the right hyper-plane Select the hyper-plane which segregates the two classes better. In this scenario, hyper-plane "B" has excellently performed this job. Identify the right hyper-plane here, we have three hyper-planes A, B, and C and all are segregating the classes well. Here, maximizing the distances between nearest data points (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called a Margin.

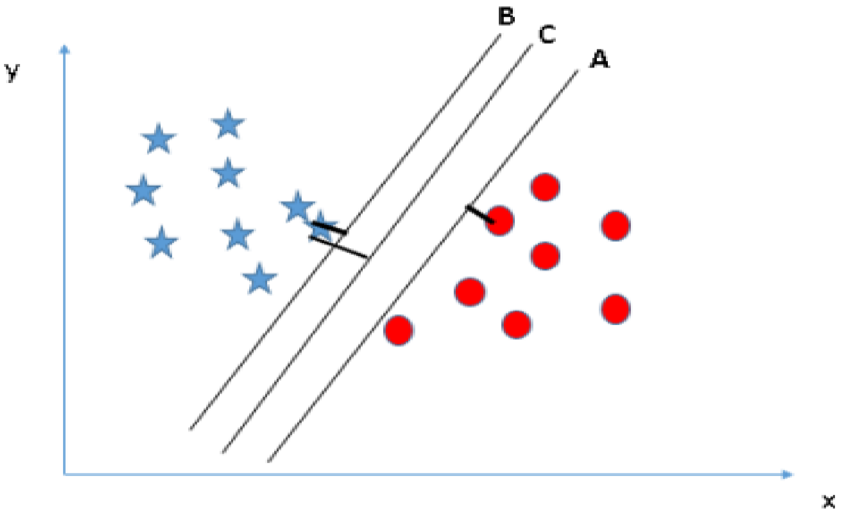


Image Taken From [1]

5.2.3 Decision Tree:

The Decision trees algorithm consists of two parts:

1. Nodes
2. rules (tests).

Using these two parts we construct the tree. Here, each node reflects a test on an attribute where the basic idea of this algorithm is to draw a diagram containing a root node on top. All other non-leaf nodes represent a test until we reach a leaf node i.e. final result, that we got.

Decision trees require less computation as compared to other algorithms. When we try to implement decision trees to detect breast cancer, the leaf nodes are divided into two categories: Benign (non-cancerous) or malignant (Cancerous). Rules will be established to determine if the tumor is benign or malignant.

Decision trees algorithm on a single attribute. Our data set contains 9 attributes as described that need to be included from the dataset. The major step in classification is to have a test set and training set from the given dataset. Otherwise, there is a chance that the evaluation results will not be reliable. here, we commonly used a ratio to split a dataset into 70% training set and 30% test set. Next, we decide which decision tree algorithm should be used for our problem. After the tree is constructed, it is applied to each row in the database. Performing initial testing on all decision tree algorithms using our dataset is shown in the figure below.

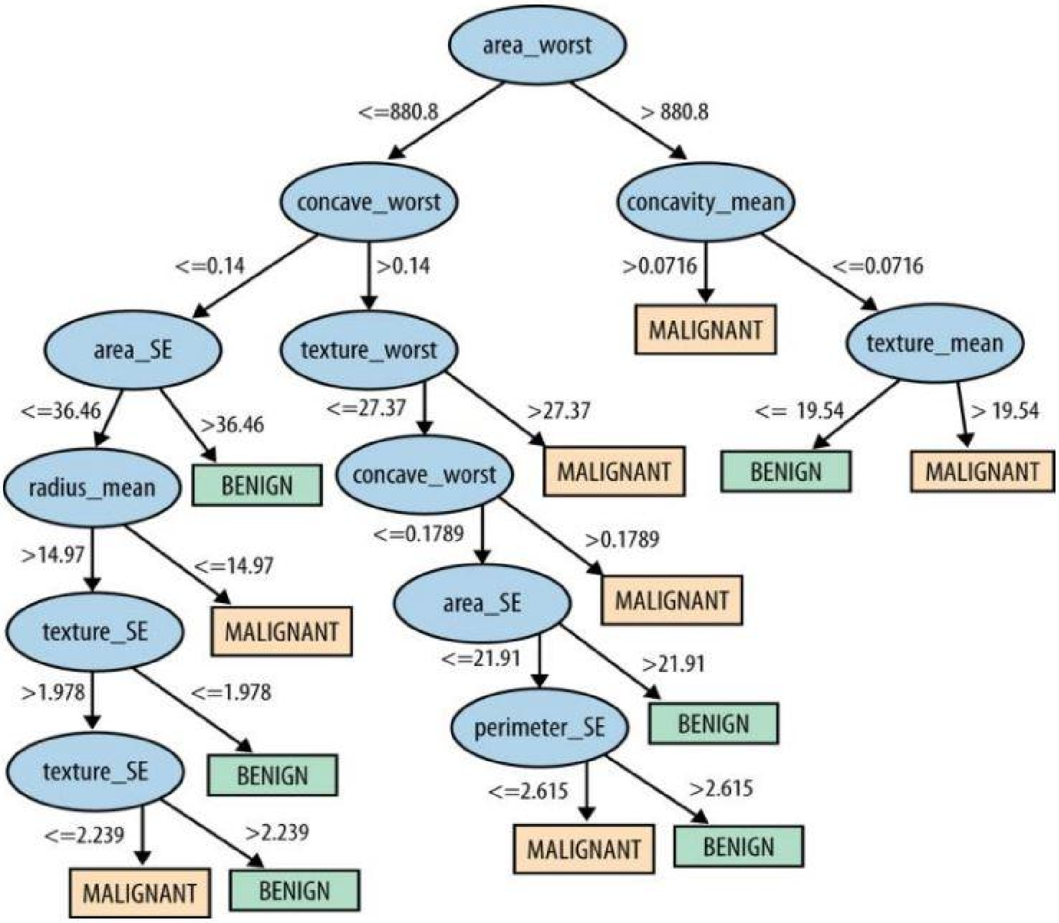


Image Taken From [1]

5.2.4 k-Nearest Neighbours:

KNN (K- Nearest Neighbours) is one among many supervised learning algorithms utilized in data processing and machine learning, it’s a classifier algorithm where the training is predicated on “how similar” may be a data from another. It is a lazy algorithm. KNN works by finding the distances between a point and all the examples within the given data. Further, it selects the required number of examples (K) closest to the given point. At last, voting for the leading frequent label.

Implementation

This proposed system presents a comparison of machine learning (ML) algorithms: Support machine vector(SVM), Decision Tree(DT), Naive Bayes (NB),k- Nearest Neighbour (k-NN) search. The data set used is obtained from the Wisconsin datasets.

Each event consists of 9 cytological features:

(1) clump thickness

(2) uniformity of cell size

(3) uniformity of cell shapes

(4) marginal adhesion, suggesting loss of adhesion.

(5) single epithelial cell size (SECS), if the SECS become larger, it may be malignant cell;

(6) bare nuclei, without cytoplasm coating
(7) bland chromatin
(8) normal nucleoli
(9) mitoses.

ALGORITHM:

- Load the dataset.
- Data Preprocessing.
- Handling the missing values.
- Data Analysis.
- Feature Selection.
- Train: Analyse and build a model to predict if a given set of symptoms lead to breast cancer. This is a binary classification problem, and a few algorithms are appropriate for use.

Now, we will do a quick test on the few appropriate algorithms to get an early indication of how each of them performs.
We will use 10 fold cross-validation for each testing. The following non-linear algorithms will be used, namely:

- ❖ Classification and Regression Trees (CART)
- ❖ Linear Support Vector Machines (SVM)
- ❖ Gaussian Naive Bayes (NB)
- ❖ k-Nearest Neighbors (KNN).

- Test: The algorithm that gave the best result is SVM.

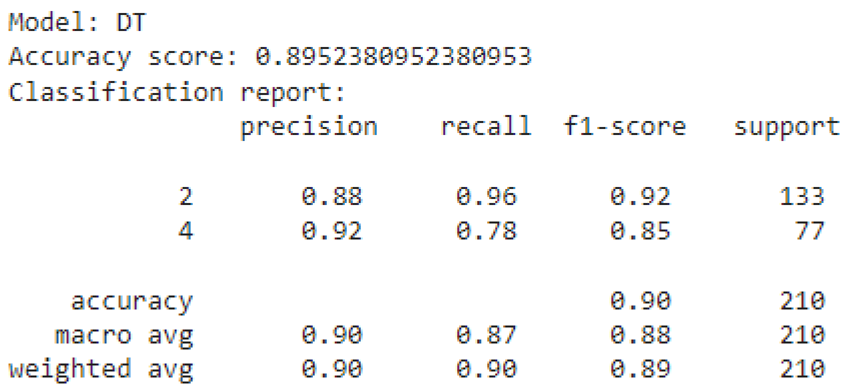
For the implementation of the ML algorithms, the dataset was partitioned into the training set and testing set. A comparison between all four algorithms will be made. The SVM model is supplied as the model for the website. The website will be made from a python framework, called a flask.

Results and Conclusion

Results

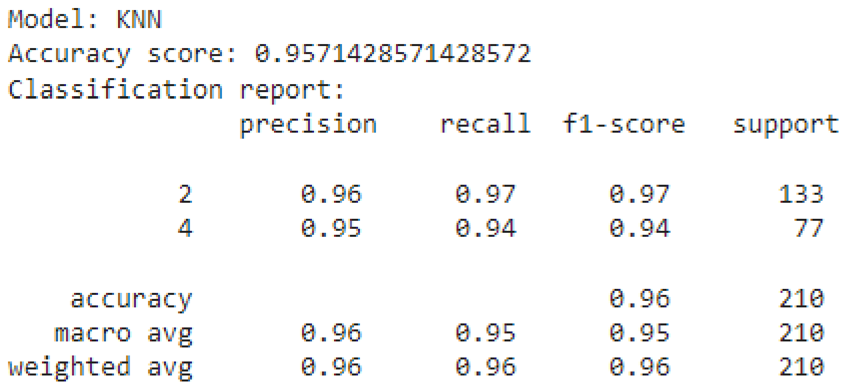
Now after training and testing different models, the accuracy of each model is:

Decision Tree



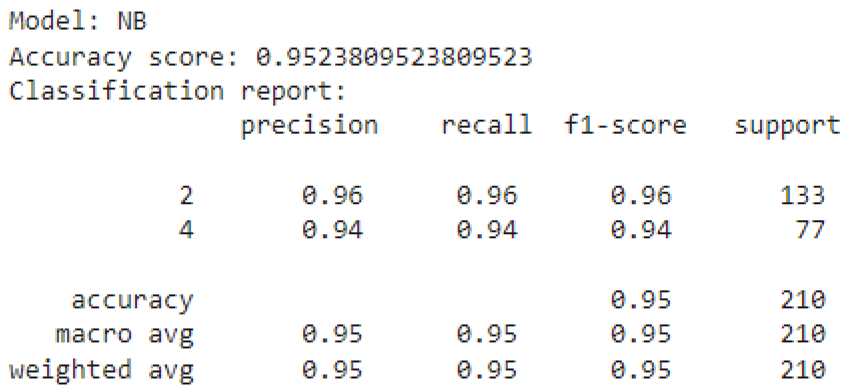
```
Model:  DT
Accuracy score: 0.8952380952380953
Classification report:
              precision    recall  f1-score   support

           2       0.88      0.96      0.92       133
           4       0.92      0.78      0.85        77

    accuracy                           0.90       210
   macro avg       0.90      0.87      0.88       210
weighted avg       0.90      0.90      0.89       210
```

K Nearest Neighbour



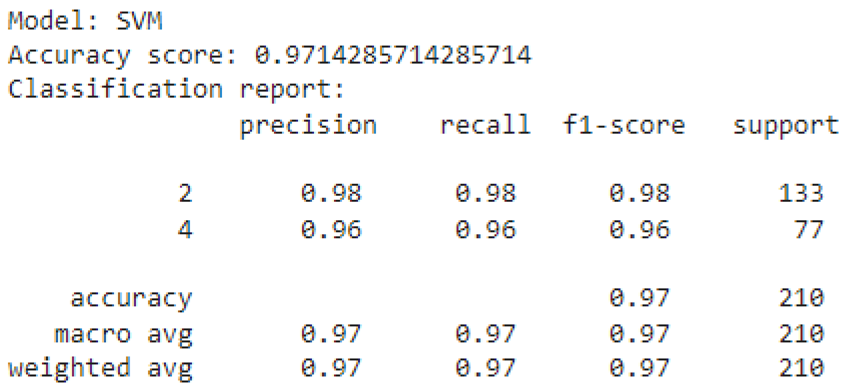
```
Model:  KNN
Accuracy score: 0.9571428571428572
Classification report:
              precision    recall  f1-score   support

           2       0.96      0.97      0.97       133
           4       0.95      0.94      0.94        77

    accuracy                           0.96       210
   macro avg       0.96      0.95      0.95       210
weighted avg       0.96      0.96      0.96       210
```

Naive Bayes



```
Model:  NB
Accuracy score: 0.9523809523809523
Classification report:
              precision    recall  f1-score   support

           2       0.96      0.96      0.96       133
           4       0.94      0.94      0.94        77

    accuracy                           0.95       210
   macro avg       0.95      0.95      0.95       210
weighted avg       0.95      0.95      0.95       210
```

Support Vector Machine

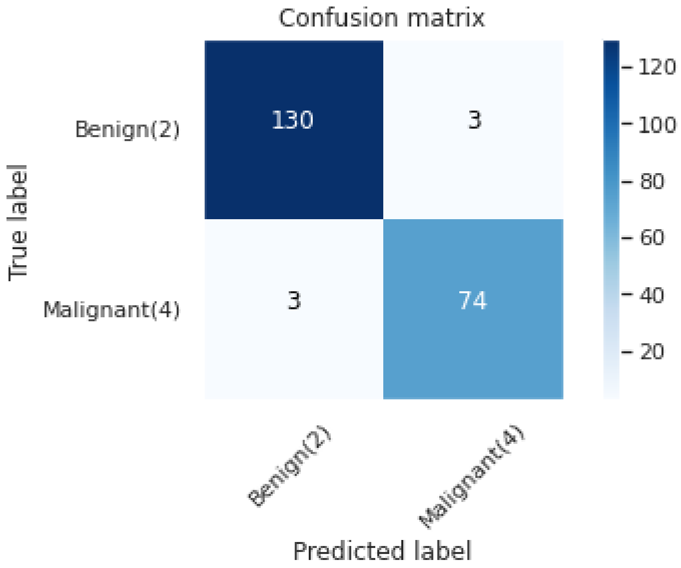


```
Model: SVM
Accuracy score: 0.9714285714285714
Classification report:
              precision    recall  f1-score   support

           2       0.98      0.98      0.98       133
           4       0.96      0.96      0.96        77

    accuracy                           0.97       210
   macro avg       0.97      0.97      0.97       210
weighted avg       0.97      0.97      0.97       210
```

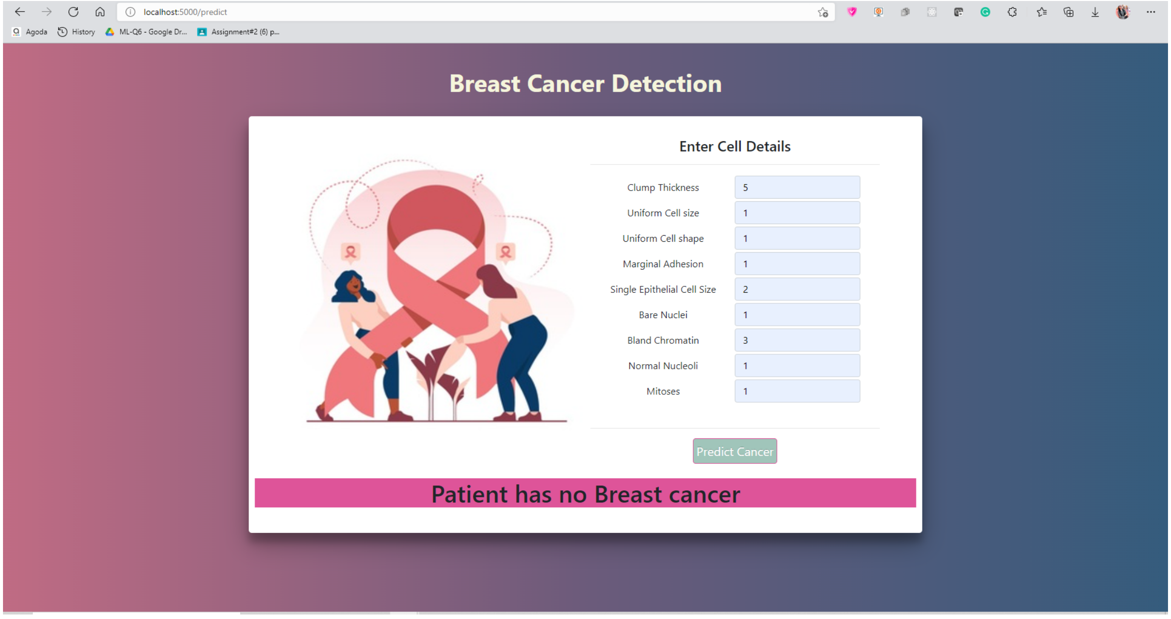
Algorithm:	**Accuracy**
Decision Tree	89.52%
Support Vector Machine	97.14%
k-Nearest Neighbour	95.71%
Naive Bayes	95.23%

So, from the above result, we can say that SVM is the best suitable algorithm for our model.

Now, seeing that SVM gives the best accuracy score, we made its Confusion Matrix:



Now, after deploying our model as a Web Application using Flask.
The output for different values is as shown :

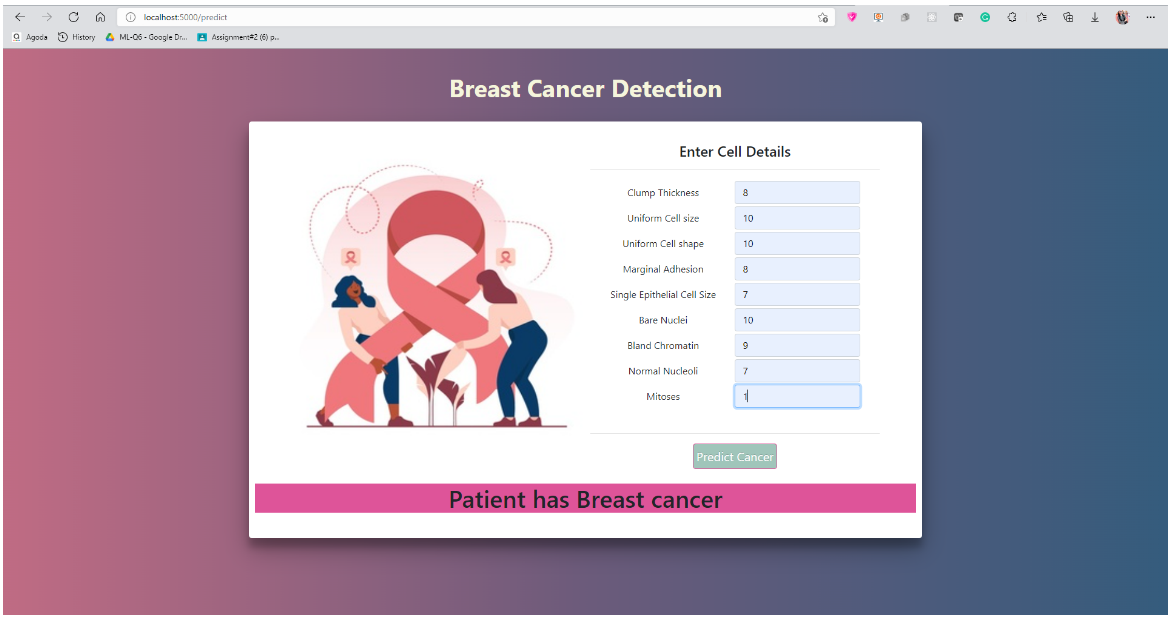


Breast Cancer Detection

Enter Cell Details

Clump Thickness	5
Uniform Cell size	1
Uniform Cell shape	1
Marginal Adhesion	1
Single Epithelial Cell Size	2
Bare Nuclei	1
Bland Chromatin	3
Normal Nucleoli	1
Mitoses	1

Patient has no Breast cancer



Breast Cancer Detection

Enter Cell Details

Clump Thickness	8
Uniform Cell size	10
Uniform Cell shape	10
Marginal Adhesion	8
Single Epithelial Cell Size	7
Bare Nuclei	10
Bland Chromatin	9
Normal Nucleoli	7
Mitoses	1

Patient has Breast cancer

Conclusion:

This paper focuses on the feasibility and effectiveness of the model from the early stage of Breast Cancer. In this paper, different algorithms are used for analyzing the best algorithm suited for the Wisconsin breast cancer dataset which consisted of values of 699 persons for 9 different features required to detect Breast Cancer. Simple classification techniques are employed for the model.

The accuracy of different algorithms we've used in our model is Decision Tree (89.52%), k-Nearest Neighbour (95.71%), Support Vector Machine (97.14%), Naive Bayes (95.23%). After analyzing the algorithms we came to the conclusion that SVM gives the best result among the algorithms used. Further, we used Flask to deploy our model as a web application.

References

[1]. Neha kumara and Khushi Verma, Bansal institute of science and technology, Volume 10, May-June 2019. “A survey on various machine learning approaches used for breast cancer detection.”

[2]. Rajesh Kumar, Rajeev Srivastava, and Subodh Srivastava Department of Computer Science and Engineering, Indian Institute of Technology, Varanasi, “ Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features”

[3]. Alaá Rateb Mahmoud Al-shamash, Ph.D. Unaizah Hanum Binti Obaidellah, Ph.D. University of Malaya, Malaysia, "Artificial Intelligence Techniques for Cancer Detection and Classification: Review Study”

[4]. Jagpreet Chhatwal, Oguzhan Alagoz, Mary J. Lindstrom, “ A Logistic Regression Model Based on the National Mammography Database Format to Aid Breast Cancer Diagnosis.”

[5]. Isabelle Guyon, Jason Weston Stephen Barnhill Bioinformatics, Savannah, Georgia, USA, “Gene Selection for Cancer Classification using Support Vector Machines”.

[6]. Naresh Khuriwal, Nidhi Mishra, Department of Computer Engineering, Poornima University Jaipur, “Breast Cancer Diagnosis Using Adaptive Voting Ensemble Machine Learning Algorithm”.

[7]. Ilia Kalogiannis,·Elia Markopoulos· Iohannis Anagnostopoulos “An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifier”.

[8]. International Journal of Computer Applications Technology and Research Volume 7–Issue 01, 23-27, 2018, ISSN:-2319–8656.

[9]. Wolberg, William. "UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set." UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set. University of Wisconsin Hospitals Madison, Wisconsin, USA, n.d. Web. Oct. 2015.’

[10]. Tike Thein1, Htet Thazin, and Khin Mo Mo Tun. "An Approach for Breast Cancer Diagnosis Classification Using Neural Network." Advanced Computing: An International Journal (ACIJ) 6 (2015)

[11]. Diagnosis of Breast Cancer using Decision Tree Models and SVM By Puneet Yadav, Rajat Varshney, Vishan Kumar, March 3, 2018

[12]. Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction, By Yixuan Li, Zixuan Chen, October 18, 2018

[13]. Machine learning with Applications in Breast Cancer Detection and Prognosis, By Wenbin Yue, Zidong Wang, Hongwei Chen, Annette Payne, and Xiaohui Liu, May 9, 2018

[14]. On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset, By:- Abien Fred M. Agarap, February 9, 2019

[15]. Comparative Studies on Breast Cancer Classifications with K-Fold Cross Validations Using Machine Learning Techniques, By:- Zahra Nematzadeh, Roliana Ibrahim, Ali Selamat, January 2015