Mini Project Report

**On**

# Breast Cancer Diagnosis Based on Machine Learning Algorithms

Submitted in partial fulfilment of the requirements

of the T. E ThirdYear Semester VI in degree of

Bachelor of Computer Engineering

BY

**Sakshi Shetty**

**Prishita Shrivas**

**Swati Singh**

Guide

Prof. Pravin Jangid

****

**Department Of Computer Engineering**

**Shree L. R. Tiwari College Of Engineering**

**Kanakia Park, Mira Road (E), Thane -401 107, Maharashtra.**

**April 2020**

[Mini Project Report i](#_Toc37415292)

[Breast Cancer Diagnosis Based on Machine Learning Algorithms i](#_Toc37415293)

[1. Introduction 1](#_Toc37415294)

[2. Literature Review 2](#_Toc37415295)

[2.1 Literature Survey Papers 2](#_Toc37415296)

[3. System Architecture 3](#_Toc37415297)

[3.1 System Architecture 3](#_Toc37415298)

[3.2 Algorithms Used 4](#_Toc37415299)

[3.2.1 Experiment Environment 4](#_Toc37415300)

[3.2.2 Breast cancer dataset 4](#_Toc37415301)

[3.2.3 Data Visualization 4](#_Toc37415302)

[4. Implementation 6](#_Toc37415303)

[4.1 Modules 6](#_Toc37415304)

[4.1.1 Stage 1: Preprocessing 6](#_Toc37415305)

[4.1.2 Stage 2: Features Selection 7](#_Toc37415306)

[4.1.3 Stage 3: Machine Learning Algorithm 7](#_Toc37415307)

[4.1.4 Code Implementation 7](#_Toc37415307)

[5.Results 13](#_Toc37415308)

[6. Conclusion 15](#_Toc37415309)

[Reference 16](#_Toc37415310)

# 1.Introduction

Being the most frequently occurring cancer in women, breast cancer affects around 10% of women at some point in their life. It is the second leading contributor to women’s death after lung cancer. 25% of all cancers in women including 12% of all new cases are caused by breast cancer . Big Data has seen a rise in value due to it being used in derivation of business intelligence, business analytics and data mining to obtain reports and result predictions. Topics like medial science rise rapidly when certain approaches like data mining is applied due to better possibility of prediction of diseases, reducing medicine costs, improving health of patient by revamping the quality of healthcare along with value by saving people’s lives through real time decisions. The paper provides you with a analysis of performance and comparison of accuracy in classification between the algorithms such as: Logistic Regression, SVM, Random Forest and Naïve Bayes, being the major influential algorithms of data mining used in the research community. Logistic Regression is used to perform regression analysis when the dependent variable is binary. It is a predictive analysis similar to all other regression analyses. Naïve Bayes is a powerful classification algorithm from machine learning, it is not a algorithm but a group of algorithm in which all of them have same fundamental principle. In that group of algorithms, they all classify the data independently such that no algorithm provides same classification result or analysis.

# 2. Literature Review

## 2.1 Literature Survey Papers

In machine learning and data mining, classification should be a crucial task. Researchers have already done lot of researches by applying machine learning algorithm on medical dataset for classification and data mining algorithm to find a pattern in dataset for faster calculation and prediction. Many of the approaches provide good accuracy and result.

[1]There are research about finding classification accuracy using machine learning algorithm known as k-Nearest Neighbor with different values of k. For each value of k they have received a different result.

[2]AlirezaOsarech, BitaShadgar used SVM classification technique on two different benchmark datasets for breast cancer which got 98.80% and 96.63% accuracies.

[3]Haifeng Wang and Sang Won Yoon compared Naive Bayes Classifier, Support Vector Machine (SVM), AdaBoost tree, Artificial Neural Networks (ANN), to find a powerful model for breast cancer prediction. They implemented PCA for dimensionality reduction.

[4]S.Kharya worked on breast cancer prediction and stated that artificial neural networks are widely used. The paper featured about the advantages and short comings of using machine learning methods like SVM, Naive Bayes, Neural network and Decision trees

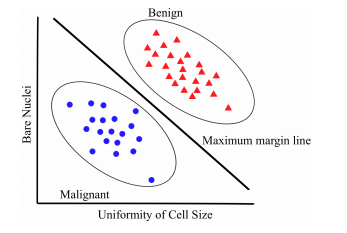
[5]Adaboost algorithm was used to predict the cause and effect of breast cancer and the reason for death. Modest Adaboost algorithm was used.

# 3. System Architecture

## 3.1 System Architecture

Throughout the decades, ML algorithms have been widely used in BC diagnosis and prognosis to gain different insights from data samples. ML is a form of Artificial Intelligence (AI) that uses a variety of statistical, probabilistic and optimization tools to learn and improve performance automatically from new data and past experiences, without explicitly programmed instructions. Both statistics and ML are used for analysing data.

The concept of SVM, which was proposed by Vapnik on the basis of the statistical learning theory [50], has become an essential component in ML techniques. The principle of SVM starts from solving linear separable problems then extends to deal with nonlinear problems. The way of solving nonlinear problems is to map training samples from the original finite dimensional space to a higher dimensional space to realize linear separability. SVM is one of the most popular approaches in BC diagnosis and prognosis [55]. An example of SVM structure [33] for BC diagnosis is shown in Figure2. 1.



**Figure 3.1.A simple example of how an SVM might work in distinguishing between benign and malignant tumor.**

In ML, DT is a predictive model that represents a mapping between object attributes and object values [56,57]. A DT is a tree-like classifier that partitions every possible outcome of data recursively into classes. DT is similar to the flowchart, in which every non-leaf node indicates a test on a particular attribute, every branch represents an outcome of that test and every leaf node expresses a classification or decision. The node at the topmost label in the tree is called root node, which corresponds to the best predictor. By using DT, both numerical and categorical data can be processed.

## 3.2 Algorithms Used

To compare the behaviors of LR, NB, SVM and Random Forest, the experiment conducted was focused on the evaluation of the algorithms. Questions raised from researchers were: Which algorithm is more effective? Which algorithm executes more efficiently? Which algorithm is more accurate in classifying?

### 3.2.1 Experiment Environment

The sickit learn python libraries were used to conduct all experiments on classifiers explained in this paper. Sklearn is a collection of data mining, machine learning and deep learning algorithms used for classification, regression, data pre-processing and clustering. The sklearn libraries were used to implement machine learning algorithms for various real-world problems. Developers and practitioners can build and evaluate suitable models with this framework. The experiment conducted in which environment is conducted is ANACONDA. It contains various applications in that we have preferred Spyder, which is a development environment that supports python. It is a powerful IDE for python compared to others. It also has introspection features. Since our problem might require those features and also debugging is easier in this platform it is preferred.

### 3.2.2 Breast cancer dataset

The UCI machine learning repository consists of The Wisconsin Breast Cancer datasets [10] which is used in this study. There are 699 instances in which 458 are benign and 241 are malignant. In addition, there are two classes malignant which contributes to 65.5% of dataset and benign 34.5%. The breast cancer dataset is obtained in a csv format from their database.

### 3.2.3 Data Visualization

Data visualization is a key aspect of data science. It helps one to comprehend and also convey the data to another person in a meaningful manner. Matplotlib and Seaborn aresome of the several python data visualization libraries. It is essential in analysing large amounts of information and to make decisions. It employs the use of pictorial elements such as maps, plots, patterns, graph trends, etc. to provide the user with an easy method of comprehending the data.



**Figure 3.2.3.Data Visualization**

# 4. Implementation

In this work, many applied techniques were tested for the subsequent stages of processing and analysis of the breast cancer dataset.

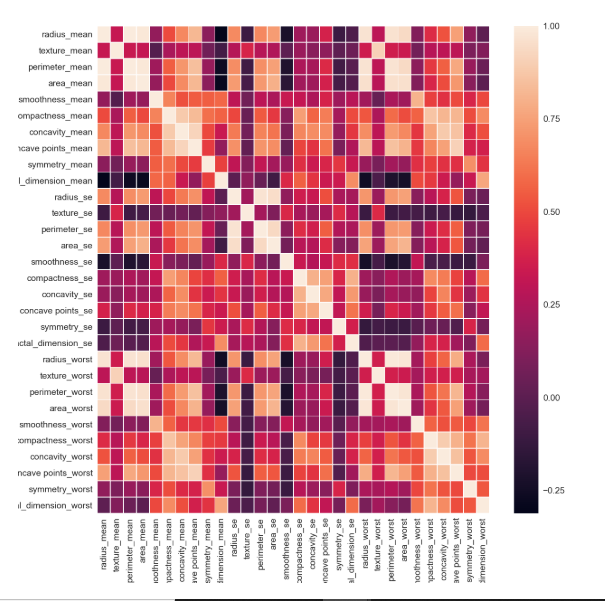
## 4.1 Modules

## 4.1.1 Stage 1: Preprocessing

Processing was performed on the raw breast cancer data to scale the features using the Standard Scaler module. Standardization of datasets is a common requirement for many machine learning estimators. It transforms the attributes to a standard Gaussian distributions based on (xi–mean(x))/stdev(x) where stdev is the standard deviation.

#### CORRELATION MAP

#### The heat map below shows the correlation between different features, where dark purple indicates negative correlation, neutral indicates no correlation, and pink/red indicates positive correlation. There appears to be a positive correlation between “compactness” and several of the features (fractal, symmetry, concavity and smoothness), as well as an inverse correlation between fractal and radius.



**Figure 4.1.1 Correlation map**

## 4.1.2 Stage 2: Features Selection

Usually, feature selection is applied as a preprocessing step before the actual learning. However, no algorithm can make good predictions without informative and discriminative features.

The module used for feature selection was implemented in using the Python scikit-learn library. All selection strategies were based to many criteria to extract the best features. In our work, feature selection was based on the following modules: removing features with low variance, univariate feature selection, and recursive feature elimination.

## 4.1.3 Stage 3: Machine Learning Algorithm

Usually, ensemble machine learning algorithms allow better predictive performance compared with a single model. This can be considered machine learning competition, where the winning solution was used as a model for breast cancer diagnosis.

The following heterogeneous ensembles machine learning algorithms were used to classify the given data set: support vector machine (SVM), K-nearest neighbor (KNN) , decision tree (DT), random forest (RF), logistic regression (LR) [34], Gaussian Naive Bayes (GNB).

# 4.1.4 Code Implementation

#import libraries

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import itertools

from sklearn.decomposition import PCA

from sklearn.preprocessing import StandardScaler

from sklearn.linear\_model import LogisticRegression

from sklearn.model\_selection import GridSearchCV, cross\_val\_score, learning\_curve, train\_test\_split

from sklearn.metrics import precision\_score, recall\_score, confusion\_matrix, roc\_curve, precision\_recall\_curve, accuracy\_score

import warnings

import plotly.offline as py

py.init\_notebook\_mode(connected=True)

import plotly.graph\_objs as go

import plotly.tools as tls

from sklearn import metrics

warnings.filterwarnings('ignore')

from google.colab import files # Use to load data on Google Colab

uploaded = files.upload() # Use to load data on Google Colab

df = pd.read\_csv('data.csv')

df.head(7)

#Count the number of rows and columns in the data set

df.shape

# describe

df.describe()

#Drop the column with all missing values (na, NAN, NaN)

#NOTE: This drops the column Unnamed

df = df.dropna(axis=1)

df.shape

#Get a count of the number of Malignant (M) (harmful) or Benign (B) cells (not harmful)

df['diagnosis'].value\_counts()

#Visualize this count

sns.countplot(df['diagnosis'],label="Count")

# Pie chart, where the slices will be ordered and plotted counter-clockwise:

labels = 'benign','malignant'

sizes = df['diagnosis'].value\_counts()

fig1, ax1 = plt.subplots()

ax1.pie(sizes, labels=labels, autopct='%1.1f%%',

        shadow=True, startangle=90)

ax1.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.

plt.show()

#Look at the data types to see which columns need to be transformed / encoded to a number

df.dtypes

from sklearn.preprocessing import LabelEncoder

labelencoder\_Y = LabelEncoder()

df.iloc[:,1]= labelencoder\_Y.fit\_transform(df.iloc[:,1].values)

sns.pairplot(df, hue='diagnosis')

#Let us find the correlation between different attributes

df.corr()

#Let us visualize with a heatmap

plt.figure(figsize=(20,20))  #This is used to change the size of the figure/ heatmap

sns.heatmap(df.corr(), annot=True, fmt='.0%')

#Split the data into independent 'X' and dependent 'Y' variables

X = df.iloc[:, 2:31].values #Notice I started from index  2 to 31, essentially removing the id column & diagnosis

Y = df.iloc[:, 1].values #Get the target variable 'diagnosis' located at index=1

# Split the dataset into 80% Training set and 20% Testing set

|  |
| --- |
|  |
|  | from sklearn.model\_selection import train\_test\_split  from sklearn.preprocessing import StandardScaler  X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.2, random\_state=42)  scaler = StandardScaler()  X\_train\_std = scaler.fit\_transform(X\_train)  X\_test\_std = scaler.transform(X\_test)  X\_train\_N = (X\_train-X\_train.mean())/(X\_train.max()-X\_train.min())  X\_test\_N = (X\_test-X\_train.mean())/(X\_test.max()-X\_test.min()) |
|  |  |
|  |  |
|  | # Scale the data to bring all features to the same level of magnitude |
|  | # This means the data will be within a specific range for example 0 -100 or 0 – 1 |
|  |  |
|  | #Feature Scaling |
|  | from sklearn.preprocessing import StandardScaler |
|  | sc = StandardScaler() |
|  | X\_train = sc.fit\_transform(X\_train) |
|  | X\_test = sc.transform(X\_test) |
|  |  |
|  | #Create a function within many Machine Learning Models |
|  | def models(X\_train,Y\_train): |
|  |  |
|  | #Using Logistic Regression Algorithm to the Training Set |
|  | from sklearn.linear\_model import LogisticRegression |
|  | log = LogisticRegression(random\_state = 0) |
|  | log.fit(X\_train, Y\_train) |
|  |  |
|  | #Using KNeighborsClassifier Method of neighbors class to use Nearest Neighbor algorithm |
|  | from sklearn.neighbors import KNeighborsClassifier |
|  | knn = KNeighborsClassifier(n\_neighbors = 5, metric = 'minkowski', p = 2) |
|  | knn.fit(X\_train, Y\_train) |
|  |  |
|  | #Using SVC method of svm class to use Support Vector Machine Algorithm |
|  | from sklearn.svm import SVC |
|  | svc\_lin = SVC(kernel = 'linear', random\_state = 0) |
|  | svc\_lin.fit(X\_train, Y\_train) |
|  |  |
|  | #Using SVC method of svm class to use Kernel SVM Algorithm |
|  | from sklearn.svm import SVC |
|  | svc\_rbf = SVC(kernel = 'rbf', random\_state = 0) |
|  | svc\_rbf.fit(X\_train, Y\_train) |
|  |  |
|  | #Using GaussianNB method of naïve\_bayes class to use Naïve Bayes Algorithm |
|  | from sklearn.naive\_bayes import GaussianNB |
|  | gauss = GaussianNB() |
|  | gauss.fit(X\_train, Y\_train) |
|  |  |
|  | #Using DecisionTreeClassifier of tree class to use Decision Tree Algorithm |
|  | from sklearn.tree import DecisionTreeClassifier |
|  | tree = DecisionTreeClassifier(criterion = 'entropy', random\_state = 0) |
|  | tree.fit(X\_train, Y\_train) |
|  |  |
|  | #Using RandomForestClassifier method of ensemble class to use Random Forest Classification algorithm |
|  | from sklearn.ensemble import RandomForestClassifier |
|  | forest = RandomForestClassifier(n\_estimators = 10, criterion = 'entropy', random\_state = 0) |
|  | forest.fit(X\_train, Y\_train) |
|  |  |
|  | #print model accuracy on the training data. |
|  | print('[0]Logistic Regression Training Accuracy:', log.score(X\_train, Y\_train)) |
|  | print('[1]K Nearest Neighbor Training Accuracy:', knn.score(X\_train, Y\_train)) |
|  | print('[2]Support Vector Machine (Linear Classifier) Training Accuracy:', svc\_lin.score(X\_train, Y\_train)) |
|  | print('[3]Support Vector Machine (RBF Classifier) Training Accuracy:', svc\_rbf.score(X\_train, Y\_train)) |
|  | print('[4]Gaussian Naive Bayes Training Accuracy:', gauss.score(X\_train, Y\_train)) |
|  | print('[5]Decision Tree Classifier Training Accuracy:', tree.score(X\_train, Y\_train)) |
|  | print('[6]Random Forest Classifier Training Accuracy:', forest.score(X\_train, Y\_train)) |
|  |  |
|  | return log, knn, svc\_lin, svc\_rbf, gauss, tree, forest |
|  |  |
|  | model = models(X\_train,Y\_train)  # Show the ROC Curve  y\_score = model.fit(X\_train\_N, Y\_train).decision\_function(X\_test\_N)  fpr, tpr, thresholds = roc\_curve(Y\_test, y\_score)  fig, ax = plt.subplots(1, figsize=(12, 6))  plt.plot(fpr, tpr, color='blue', label='ROC curve for SVM')  plt.plot([0, 1], [0, 1], 'k--')  plt.xlabel('False Positive Rate (1 - specificity)')  plt.ylabel('True Positive Rate (sensitivity)')  plt.title('ROC Curve for Breast Cancer Classifer')  plt.legend(loc="lower right") |
|  |  |
|  | #Show the confusion matrix and accuracy for all of the models on the test data  y\_pred = model.predict(X\_test\_N)  cm = confusion\_matrix(Y\_test, y\_pred)  df\_cm = pd.DataFrame(cm, range(2),                    range(2))  plt.figure(figsize=(10,7))  sns.set(font\_scale=1.4)#for label size  cm\_plot = sns.heatmap(df\_cm, annot=True, fmt='n', annot\_kws={"size": 12}) |
|  | #Classification accuracy is the ratio of correct predictions to total predictions made. |
|  | from sklearn.metrics import confusion\_matrix |
|  | for i in range(len(model)): |
|  | cm = confusion\_matrix(Y\_test, model[i].predict(X\_test)) |
|  |  |
|  | TN = cm[0][0] |
|  | TP = cm[1][1] |
|  | FN = cm[1][0] |
|  | FP = cm[0][1] |
|  |  |
|  | print(cm) |
|  | print('Model[{}] Testing Accuracy = "{}!"'.format(i, (TP + TN) / (TP + TN + FN + FP))) |
|  | print() |
|  |  |
|  | #Show other ways to get the classification accuracy & other metrics |
|  |  |
|  | from sklearn.metrics import classification\_report |
|  | from sklearn.metrics import accuracy\_score |
|  |  |
|  | for i in range(len(model)): |
|  | print('Model ',i) |
|  | #Check precision, recall, f1-score |
|  | print( classification\_report(Y\_test, model[i].predict(X\_test)) ) |
|  | #Another way to get the models accuracy on the test data |
|  | print( accuracy\_score(Y\_test, model[i].predict(X\_test))) |
|  | print()#Print a new line |
|  |  |
|  |  |
|  | #Print Prediction of Random Forest Classifier model |
|  | pred = model[6].predict(X\_test) |
|  | print(pred) |
|  | #Print a space |
|  | print() |
|  | #Print the actual values |
|  | print(Y\_test) |

# 

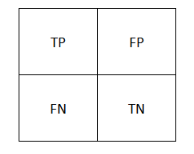
# 5.Results

Our classification metrics are prepared from the best score of accuracy (SVM algorithm).

#### 5.1CONFUSION MATRIX

Confusion Matrix is a performance measurement for machine learning classification problem, where output can be two or more classes.

It’s useful for measuring Precision, Recall, F1 score, accuracy and AUC.

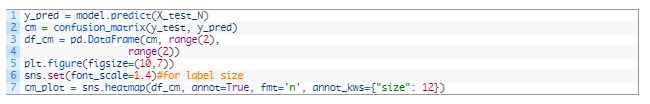


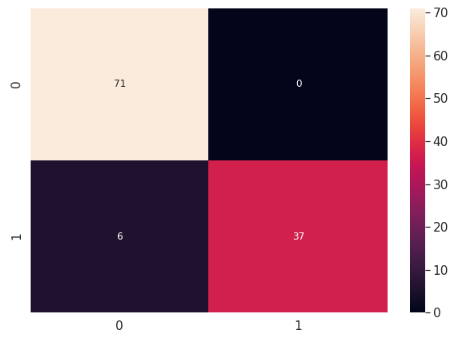
TP (True Positive) – you predicted positive and it is true,

FP (False Positive) – you predicted positive and it is false,

FN (False Negative) – you predicted negative and it is false,

TN (True Negative) – you predicted negative and it is true.



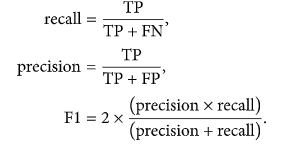


**Figure 5.1.Visualization of Confusion Matrix**

Accuracy (ACC) is the measure of correct prediction of the classifier, and it provides general information about how many samples are misclassified. It is defined as

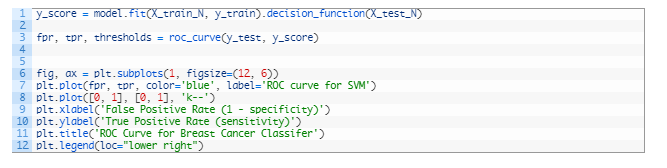
****

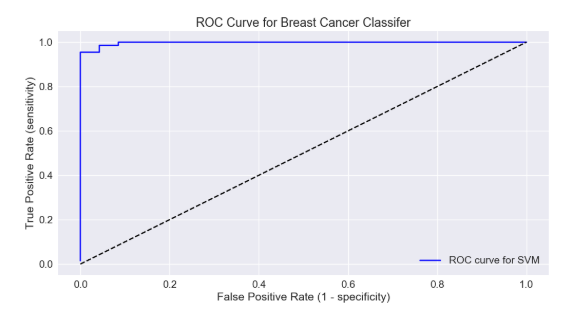
The other metrics derived from a confusion matrix are defined as follows:



#### 5.2 ROC CURVE

ROC Curve (Receiver Operating Characteristics)  is a performance measurement for classification problem at various thresholds settings. It tells how much model is capable of distinguishing between classes.





**Figure 5.2 ROC Curve**

# 6. Conclusion

Medical dataset can not only be classified with the previously mentioned algorithms from machine learning, there are many algorithms and techniques which may perform better than these. Production of accurate classifier which perform efficiently for medicinal application is the main challenge we face in machine learning. Four main algorithms were implemented in this study were NB, SVM, Random Forest and LR on Breast Cancer dataset. Our main aim for the research is to discover the algorithm which performs faster, accurate and efficiently. Random Forest surpasses all the other algorithms with an accuracy of 99.76%. In conclusion, the Random Forest algorithm achieves the lowest error rate along with highest precision which might be the best choice of algorithm for this problem and prediction of disease.

# Reference

1] Han, Jiawei, Micheline Kamber, and JianPei., “Data mining:

concepts and techniques”, publisher: Morgan Kaufmann,

2006.

[2] Jaideep Srivastava, Robert Cooley, Mukund Deshpande,

Pang-Ning Tan, “Web Usage Mining: Discovery and

Applications of Usage Patterns from Web Data”, SIGKDD

Explorations, Vol. 1, No. 2, Pp. 12-23, 2000.

[3] ShaileyMinocha, Nicola Millard, Lisa Dawson, “Integrating

Customer Relationship Management Strategies in (B2C) E-

Commerce Environments”, IFIP Conference on Human-

Computer Interaction- INTERACT, 2003.

[4] F. Johnson and S. K. Gupta, Web Content Mining

Techniques: A Survey, International journal of computer

applications (0975-888), Volume 47– No.11, June 2012.

[5] Yan LI, Boqin FENG and Qinjiao MAO, “Research on Path

Completion Technique In Web Usage Mining”, IEEE

International Symposium On Computer Science and

Computational Technology, pp. 554-559, 2008.

1. Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and Razavi AR. Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. Journal of Health & Medical Informatics

2. Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, Abdelkader Benyettou. Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. International Journal of Computer Applications (0975 - 8887)

3. Abdelghani Bellaachia, Erhan Guven Predicting Breast Cancer Survivability Using Data Mining Techniques. 2006 SIAM Conference on Data Mining

4. Cuong Nguyen, Yong Wang, Ha Nam Nguyen Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. J. Biomedical Science and Engineering, 2013, 6, 551-560

5. Diana Dumitru. Prediction of recurrent events in breast cancer using the Naive Bayesian classification. 2000 Mathematics Subject Classification

6. Turgay Ayer, MS; JagpreetChhatwal, PhD; OguzhanAlagoz, PhD; Charles E. Kahn, Jr, MD, MS; Ryan W. Woods, MD, MPH; Elizabeth S. Burnside, MD, MPH, MS. Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation. RadioGraphics 2010

7. JarceThongkam, Guandong Xu, Yanchun Zhang and Fuchun Huang. Breast Cancer Survivability via Adaboost Algorithm. HDKM '08 Proceedings of the second Australasian workshop on Health data and knowledge management

8. Rasool Fakoor, Faisal Ladhak, Azade Nazi, Manfred Huber. Using deep learning to enhance cancer diagnosis and classification. 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013

9. Cancer Statistics, 2016. CA: A Cancer Journal for Clinicians

10. H. Dhahri and A. M. Alimi, “Hierarchical learning algorithm for the beta basis function neural network,” in *Proceedings of the Third International Conference on Systems, Signals & Devices (SSD)*, Sousse, Tunesia, March 2005