

ICT619

# A comparative study of

# AI methods on Breast Cancer dataset

Murdoch University



**Submitted by:**

33296085 Vajratiya Vajrobol

33553377 Sai kumar Vadde

33669079 Rajesh Jyothi

Artificial Intelligence

Table of Contents

[A comparative study of AI methods on Breast Cancer dataset 2](#_Toc37883572)

[Abstract 2](#_Toc37883573)

[Introduction: 2](#_Toc37883574)

[CRISP Model: 3](#_Toc37883575)

[About CRISP: 3](#_Toc37883576)

[ Business Understanding 3](#_Toc37883577)

[ Data Understanding: 4](#_Toc37883578)

[ Problem Statement: 4](#_Toc37883579)

[ Check for NULL values 6](#_Toc37883580)

[ Understanding the proportion of the classes. 6](#_Toc37883581)

[ Data Preparation 6](#_Toc37883582)

[ Remove NULL values 6](#_Toc37883583)

[ Portioning of the dataset: 6](#_Toc37883584)

[ Data Modelling 7](#_Toc37883585)

[ Why the Decision Tree: 7](#_Toc37883586)

[ Why Random Forest: 8](#_Toc37883587)

[ Why Neural networks: 9](#_Toc37883588)

[ Evaluation and Interpretation 11](#_Toc37883589)

[ Future Development Plan 11](#_Toc37883590)

[Conclusion: 12](#_Toc37883591)

[References : 13](#_Toc37883592)

# A comparative study of AI methods on Breast Cancer dataset

# Abstract

Many scientific studies proved that breast cancer detected at an early stage is curable. We collected Breast cancer Wisoncin dataset from the UCI repository to detect the cancer diagnosis. In this report, A prototype of the three Machine Learning and AI methods such as Decision Tree, Random Forest and Neural Networks are implemented on this Breast Cancer dataset and results are compared. As a prototype, It mainly focuses on the problem statement, Business and Data understanding, basic Data preparation, modelling and interpretation of the models. For further studies on Breast Cancer, “Future development plan”, is provided to give a clear overview of the project. The whole project (prototype and final project) is implemented using the CRISP-DM (Cross Industry-standard Process for Data Mining) framework for better understanding and systematic workflow of the project.

# Introduction:

Breast Cancer is the most common cancer and a leading cause of deaths in women from decades. According to World Cancer Research, 2014, 1.7 million new Breast Cancer cases were reported in 2012 alone. Among which 18.3% of cases are reported from Egypt and middle eastern countries (Abdel, 2016). According to the recent statistics from the WHO (World Health Organisation), 2.1 million are effecting with breast cancer each year. Around 627,000 women were dead in 2018 alone due to Breast Cancer, this is around 15% of total deaths caused by cancer in women (World Health Organisation, 2019). Early detection of breast cancer has 37.3% of healing it completely (Salama et al., 2012 ). From the past few decades, computer intelligence and medical science are working closely, to find a better solution, in identifying the diagnosis at an early stage. With the recent development of computational power, different Artificial Intelligence and deep learning techniques became popular. Many developed countries, reduced their Breast Cancer mortality up to 10% by detecting cancer tumour using better AI and Machine Learning techniques at an early stage (Carioli et at., 2018).

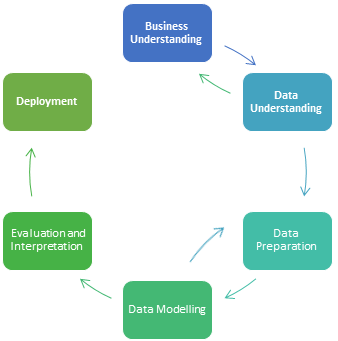
In this report, we focused mainly on Decision Tree, Random Forest and NN methods to find the Breast Cancer diagnosis. CRISP-DM framework is implemented for the systematic study.

# CRISP Model:

## About CRISP:

CRISP-DM stands for Cross Industry-standard Process for Data Mining. It is a standard data mining framework that can be implemented in any data mining applications independent of sectors (industry sectors or technology sector). Some of the advantages of this model are discussed below (Wirth et al., 2000).

* Can be implemented large, medium and small projects
* Less costly
* High reliability
* Better project manageability
* A better understanding of the flow of the project at each stage.



The following are the stages in the CRISP-DM framework

## Business Understanding

Common cancer causing death in women is Breast Cancer. It is easy is to detect breast cancer by self-examination or mammography. Mammography screening is the best tool to detect cancer tumour before symptoms appear. Biopsy surgery lumpectomy is the best method to remove the cancer tumour. Fine Needle Aspiration (FNA) is the process used to evaluate malignancy. It is one of the cost-effective, invasive and non-traumatic diagnosis test. A less invasive technique is developed in the USA recently, that uses cooled-nitrogen to shrink by freezing the non-cancer cells, and destroys the blood vessels that help in the growth of the tumour. So, for detecting the cancer diagnosis, Several AI techniques are used. This helps doctors in finding a better solution to the diagnosis.

Generally, tumours in the breast can be classified into two types, Malignant and Benign. Malignant tumours are cancer cells, that grows over time and move into lymph nodes and bloodstream, where other tissues are easily spread. On the other hand, Benign are non-cancerous cells, that grow till a certain period and stop the growth. These can cause a serious problem if they grew on nerves and blood vessels. These can be detected and removed through normal surgeries.

To find the diagnosis, we used the Wisconsin Breast cancer dataset (WBCD) easily available to the public on a machine learning data sets repository called UCI, website: <http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)>

## Data Understanding:

The Breast Cancer Wisconsin Data set consist of 569 instances and 32 attributes, all the independent variables are real values and is has no missing values. Each instance is a patient record and each column defines the feature of the tumour cells.

### Problem Statement:

Machine Learning algorithms, based on independent and dependent variables in a data set, can be classified into supervised and unsupervised. Supervised learning has the attributes and the output variables (also called dependent variable), whereas Unsupervised learning data sets, do not have any output variable (Output is unknow) and has no specific definition for columns (unlabelled/no column names). Supervised can further be classified into Regression and classification based on the output variable. If the output variable/dependent variable is continuous (numeric/real) then, it falls under Regression. If it is categorical, then falls under Classification. Classification problems can be further classified into binary and multilevel classes, based on the number of classes in the output.

Breast Cancer data has 30 independent variables and 1 dependent variable (diagnosis), so, falls under supervised learning. All the independent variables are numerical values and the dependent variable is categorical. Diagnosis, which is dependent, has 2 levels (or binary), M (Malignant) and B (Benign). So, it can be categorised as a classification problem. So, we can use any classification algorithms for classifying the data into Malignant and Benign. For instance, Logistic Regression, Decision Tree, Random Forest, KNN, SVM and Neural Networks etc, can be used.

So Breast cancer data set is supervised learning and classification problem.

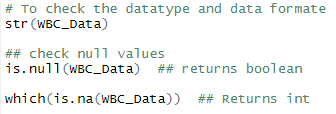
The columns in the dataset are given below.

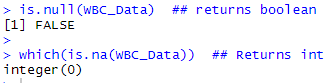
|  |  |  |
| --- | --- | --- |
| **Sno** | **Attributes** | **Description** |
| 1 | ID | ID number |
| 2 | Diagnosis | the diagnosis of breast cancer tissue (M =Malignant ,B= Benign) |
| 3 | Radius mean | mean of distances from centre points on the perimeter |
| 4 | Texture mean | the standard deviation of grey-scale values |
| 5 | Perimeter mean | the mean of core tumour size |
| 6 | Area \_mean | area |
| 7 | Smoothness mean | mean of local variation in radius length |
| 8 | Compactness \_mean | mean of perimeter ^ 2 / area - 1.0 |
| 9 | Concavity mean | mean of the severity of concave portions of the contour |
| 10 | Concave points mean | mean of concave portions of the contour |
| 11 | symmetry mean |  |
| 12 | fractal\_dimension\_mean | mean of "coastline approximation" - 1 |
| 13 | radius\_se | the standard error of the mean of distances from the centre to points on the perimeter |
| 14 | texture\_se | the standard error of standard deviation of grey-scale values |
| 15 | perimeter\_se |  |
| 16 | area\_se |  |
| 17 | smoothness\_se | the standard error of local variation in radius lengths |
| 18 | compactness\_se | standard error of perimeter^2 / area - 1.0 |
| 19 | Concavity\_se | the standard error for the severity of concave portions of the contour |
| 20 | Concave points\_se | the standard error for concave portions of the contour |
| 21 | Symmetry\_se |  |
| 22 | fractal\_dimension\_se | standard error of "coastline approximation" – 1 |
| 23 | radius\_worst | the largest mean value of the mean of distances from the centre to points on the perimeter |
| 24 | Texture\_worst | the largest mean value of the standard deviation of grey-scale values |
| 25 | Perimeter\_worst |  |
| 26 | Area\_worst |  |
| 27 | Smoothness\_worst | the largest mean value for local variation in radius lengths |
| 28 | Compactness\_worst | largest mean value for perimeter^2 / area - 1.0 |
| 29 | concavity\_worst | the largest mean value for the severity of concave portions of the contour |
| 30 | concave points\_worst | the largest mean value of concave portions of the contour |
| 31 | symmetry\_worst |  |
| 32 | fractal\_dimension\_worst | largest mean value of "coastline approximation" – 1 |
| 28 | compactness\_worst | the largest mean value of perimeter^2 / area - 1 |
| 29 | concavity\_worst | the largest mean value of severity of concave portions of the contour |
| 30 | concave points\_worst | r largest mean value of the number of concave portions of the contour |
| 31 | symmetry\_worst |  |
| 32 | fractal\_dimension\_worst | largest mean value of "coastline approximation" – 1 |

### Check for NULL values

From the data set description (from the source website), we do not have any missing values.

Checking from R code:



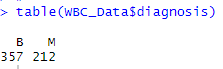


From the above R code, we found that there are no missing values in the data set.

## Understanding the proportion of the classes.

The data set is balanced and these are ignorable skewness in the data set. In total there are 357 Benign cases and 212 Malignant cases.





## Data Preparation

### Remove NULL values

As we do not have any missing values, we can ignore this step.

### Portioning of the dataset:

Before model designing, the data set is divided into two partitions. One is the training set, for training the model and other is the testing set, for testing the trained model.

Training and testing partition ratio plays a critical role in the model performance. If the training set is larger than the model is over-trained for the training set, on the other hand, if the training set is smaller, then the model is under-trained to a given training set. Both cases have an impact on model accuracy.

We portioned the training set into 80% and testing set into 20% for the prototype design.

## Data Modelling

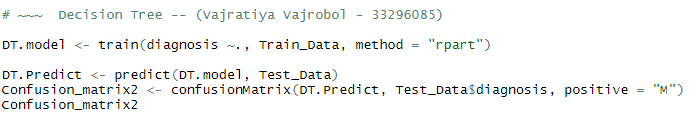
In model design, we select the models based on data understanding and train and test them by sampling the data set.

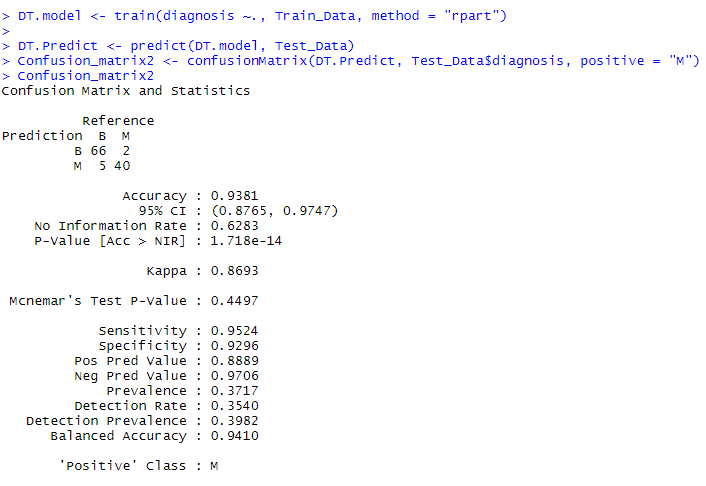
### Why the Decision Tree:

As the decision tree is compatible with the breast cancer problem which is to classify the diagnosis such as benign or malignant. Also, the decision trees have in memory classification model that doesn’t require high computation cost. Furthermore, a decision tree is easy to understand from their extreme transparency of rule-based decision making. As our data is numerical, a decision tree can handle these types of data well. The decision tree brings the capability to handle data set with a high error rate and missing value. A decision tree tends to produce the wrong result if the problem is complex but to classify two classes as benign and malignant is quite simple to generate the model.

The decision tree included a pruning technique that moving some sections of the tree to reduce the complexity of the final classifier and also improve the accuracy of prediction.

The basic Decision Tree model is designed and results are shown below. Decision Tree shows an accuracy of 93.81% for 80% training and 20% testing test



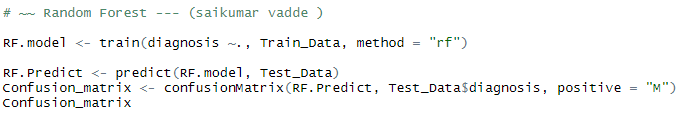


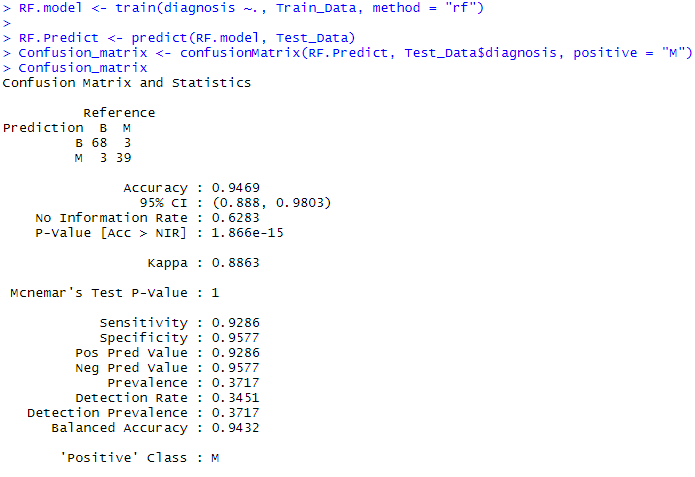
### Why Random Forest:

A forest is a combination of trees, similarly, the random forest is the combination of individual decision trees. Some of the key reasons for taking Random Forest are mentioned below.

* As these are tree-based structures, No need to worry about scaling of the data.
* Random Forest algorithm prevents overfitting automatically, by using random subspace and bagging methods.
* Handling missing data is build-in in this algorithm.
* Feature section is easy to implement as it in build.

Basic Random forest model is designed and results are shown below. Randon Forest shows an accuracy of 94.69% for 80% training and 20% testing test



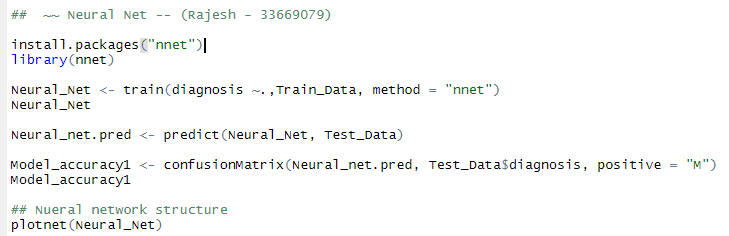


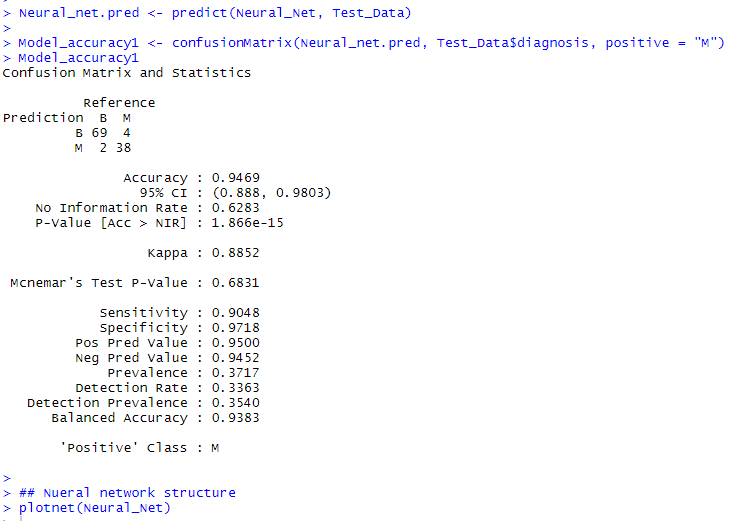
### Why Neural networks:

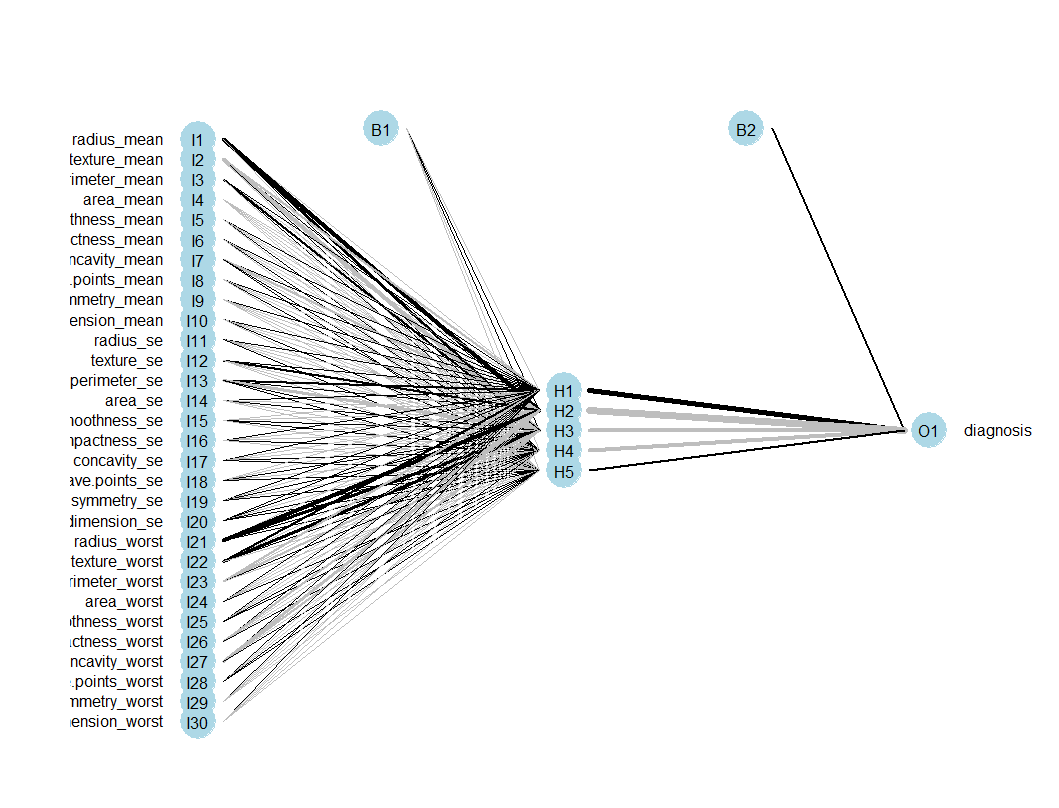
The neural network is similar to the work done by the neurons in the brain. It includes technologies such as Artificial Intelligence, Machine learning and deep learning. Neural networks work in layers format, namely the input layer, hidden layer (optional) and the output layer. The input to the model is given at the input layer and the output is extracted from the output layer. The hidden layer in the neural network is optional. Based on the requirement we can add the hidden layers. Few reasons for considering the neural networks for breast cancer prediction are discussed below.

* Neural networks generate information on their own during the time of learning.
* These models have adaptive learning (can learn based on training data set given)
* If need to change the functionality of the model, one can change the activation function in the model, rather changing the whole model.
* Scalability of hidden layers (adding hidden layers is flexible).
* Increases the model accuracy.

Basic Neural Network model is designed and results are shown below. Neural Network shows an accuracy of 94.69% for 80% training and 20% testing test







Basic neural network diagram is shown above. It has an input layer and an output layer.

## Evaluation and Interpretation

All the three models namely Decision Tree, Random Forest and Neural Networks are implemented and below table shows the accuracy comparison.

|  |  |  |
| --- | --- | --- |
| S.no | Models | Accuracy |
| 1 | Decision Tree | 93.81 |
| 2 | Random Forest | 94.69 |
| 3 | Neural Networks | 94.69 |

The training set and testing test for these models are taken in 0.8 and 0.2 proportion respectively. All the models performed well in classifying the Malignant and Benign patients. Interestingly, Random forest and Neural Networks has shown the same accuracy and Decision tree shown slightly low accuracy.

## Future Development Plan

This is a simple prototype designed initial stage. As a part of the future project, we would like to check the outlier in the data set. as we have 30 dependent variables, we would implement dimensionality reduction techniques for better model performance. Try implementing better sample techniques like random sampling and cross-validation. We shall implement PCA and LDA for feature selection. In neural networks, we would like to check the accuracy of the model by adding the hidden layers. and finally, compare the results

# Conclusion:

Medical diagnosis process can be easily simplified using better computer algorithms. Based on the business requirement and data availability, the techniques to be implemented changes. For finding the better model, trial and error method best choice, rather than just assuming the models to be working fine. A better model gives a better solution to makes life easier. In diagnosing breast cancer patients to affected with cancer, if the model gives better results, then better treatment could be provided at the initial stage and can save the lives. All three models performed well in this case. Random Forest and Neural networks gave the same performance with 94.69% accuracy.

# References :

Abdel-Zaher, A. M., & Eldeib, A. M. (2016). Breast cancer classification using deep belief networks. *Expert Systems with Applications*, *46*, 139-144.

Carioli, G., Malvezzi, M., Rodriguez, T., Bertuccio, P., Negri, E., & La Vecchia, C. (2018). Trends and predictions to 2020 in breast cancer mortality: Americas and Australasia. *The Breast*, *37*, 163-169.

Salama, G. I., Abdelhalim, M., & Zeid, M. A. E. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. *Breast Cancer (WDBC)*, *32*(569), 2.

Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29-39). London, UK: Springer-Verlag.

World Health Organisation (n.d.). *Breast Cancer*. Retrieved from WHO: <https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/>