# Machine Learning for Breast Cancer Diagnosis

K
Kishan Dalsania
Oct 9 · 4 min read

Can we accurately distinguish cancerous cell?

## Introduction

Machine learning is branch of Data Science which incorporates a large set of statistical techniques.

These techniques enable data scientists to create a model which can learn from past data and detect patterns from massive, noisy and complex data sets. Researchers use machine learning for cancer prediction and prognosis. Machine learning allows inferences or decisions that otherwise cannot be made using conventional statistical methodologies. With a robustly validated machine learning model, chances of right diagnosis improve. It specially helps in interpretation of results for borderline cases.

## Breast Cancer: An overview

The most common cancer in women worldwide. The principle cause of death from cancer among women globally.Early detection is the most effective way to reduce breast cancer deaths.Early diagnosis requires an accurate and reliable procedure to distinguish between benign breast tumors from malignant ones

Breast Cancer Types — three types of breast tumors: Benign breast tumors, In-situ cancers, and Invasive cancers.

The majority of breast tumors detected by mammography are benign. They are non-cancerous growths and cannot spread outside of the breast to other organs.

If the malignant cells have not gone through the basal membrane but is completely contained in the lobule or the ducts, the cancer is called in-situ or noninvasive.

If the cancer has broken through the basal membrane and spread into the surrounding tissue, it is called invasive.    This analysis assists in differentiating between benign and malignant tumors.

## Pre-Requisites

- Python 3.+
- Understanding of libraries (Scikit Learn, Numpy, Pandas, Matplotlib, Seaborn)
- Jupyter Notebook or Google Colab
- Orange Tool
- Basic understanding of classification methods or Algorithms.

## Problem Description

Breast Cancer (BC) is a common cancer for women around the world, and early detection of BC can greatly improve prognosis and survival chances by promoting clinical treatment to patients early. So it’s amazing to be able to possibly help save lives just by using data, python, and machine learning!

In some cases, it is difficult to distinguish certain benign masses from malignant lesions with mammography.So our task to classify that the beast contains benign cell or malignant cell.

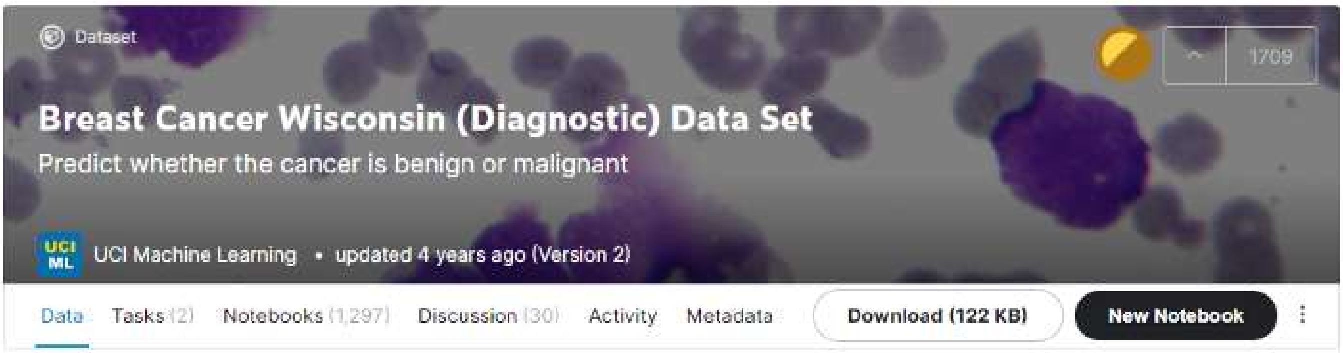
## Data Sources

Data Source(1) :- https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Data Source(2) :-https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/

## Data Description

### DataSource(1)



Dataset

Breast Cancer Wisconsin (Diagnostic) Data Set

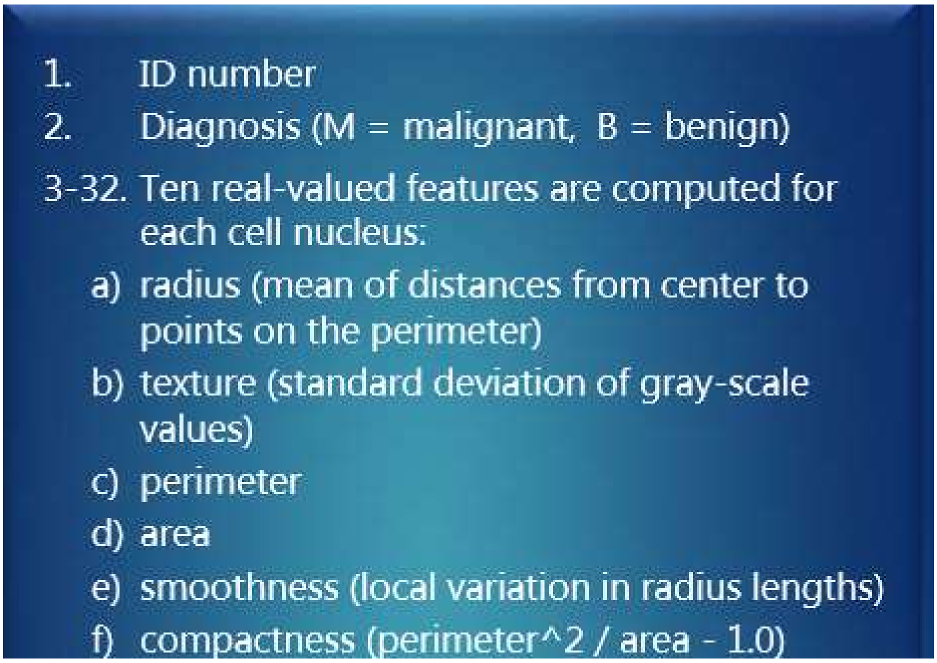
Predict whether the cancer is benign or malignant

UCI Machine Learning • updated 4 years ago (Version 2)

The data used for this Project is from University of Wisconsin.

Citation: This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.

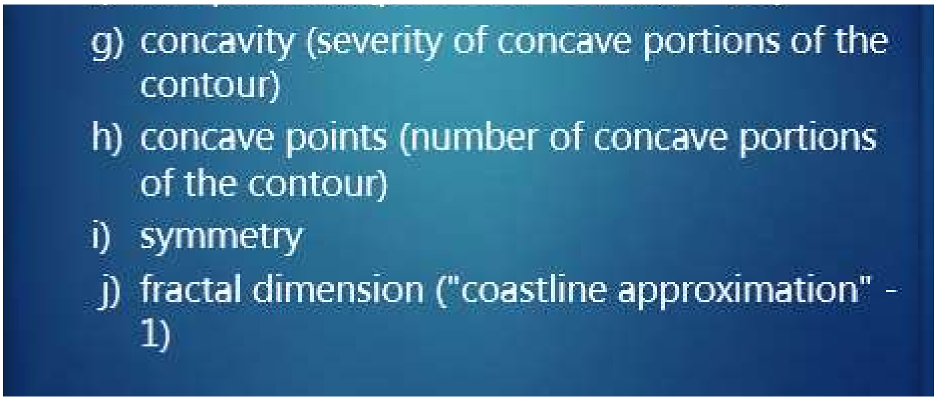
Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The mean, standard error, and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All feature values are recorded with four significant digits.



1. ID number
2. Diagnosis (M = malignant, B = benign)

3-32. Ten real-valued features are computed for each cell nucleus:

- a) radius (mean of distances from center to points on the perimeter)
- b) texture (standard deviation of gray-scale values)
- c) perimeter
- d) area
- e) smoothness (local variation in radius lengths)
- f) compactness (perimeter^2 / area - 1.0)



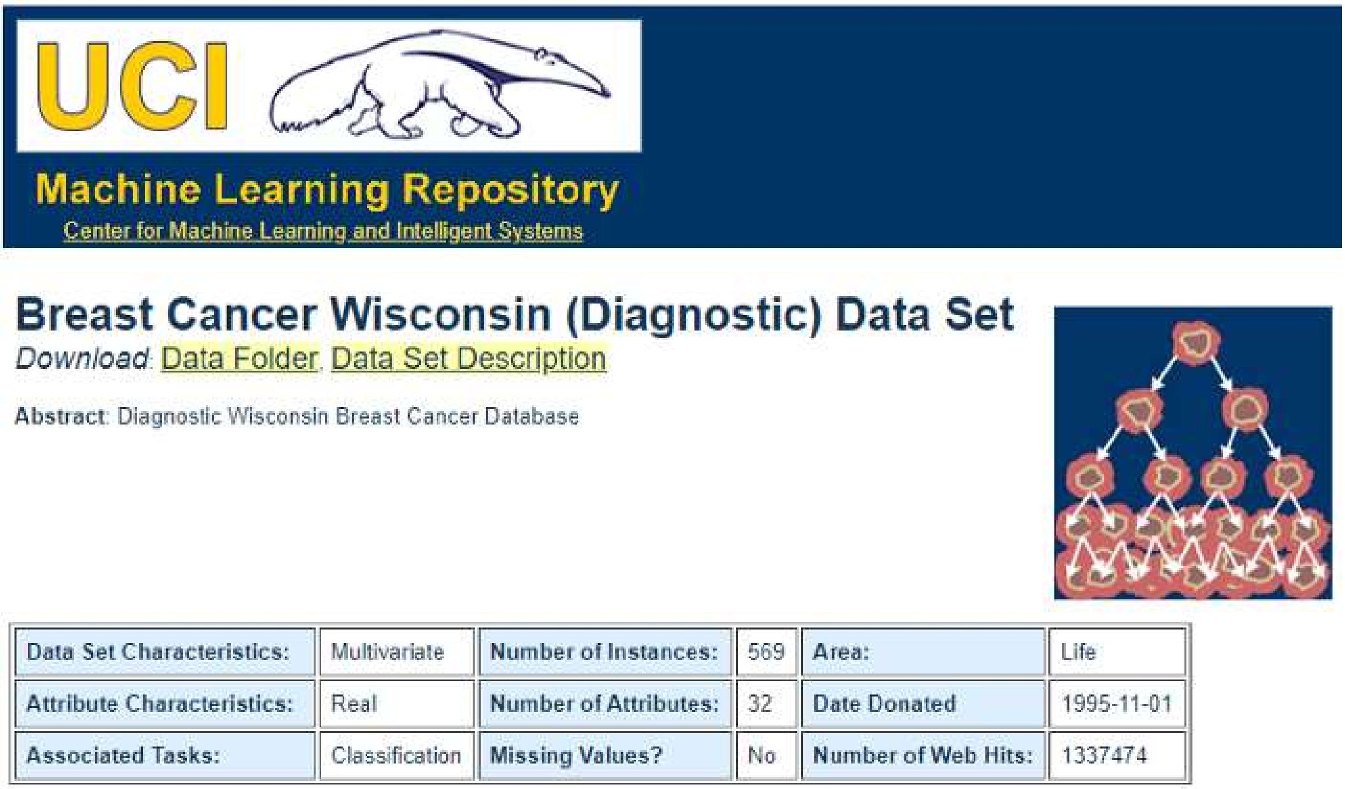
g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)

## Data Source(2)



UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

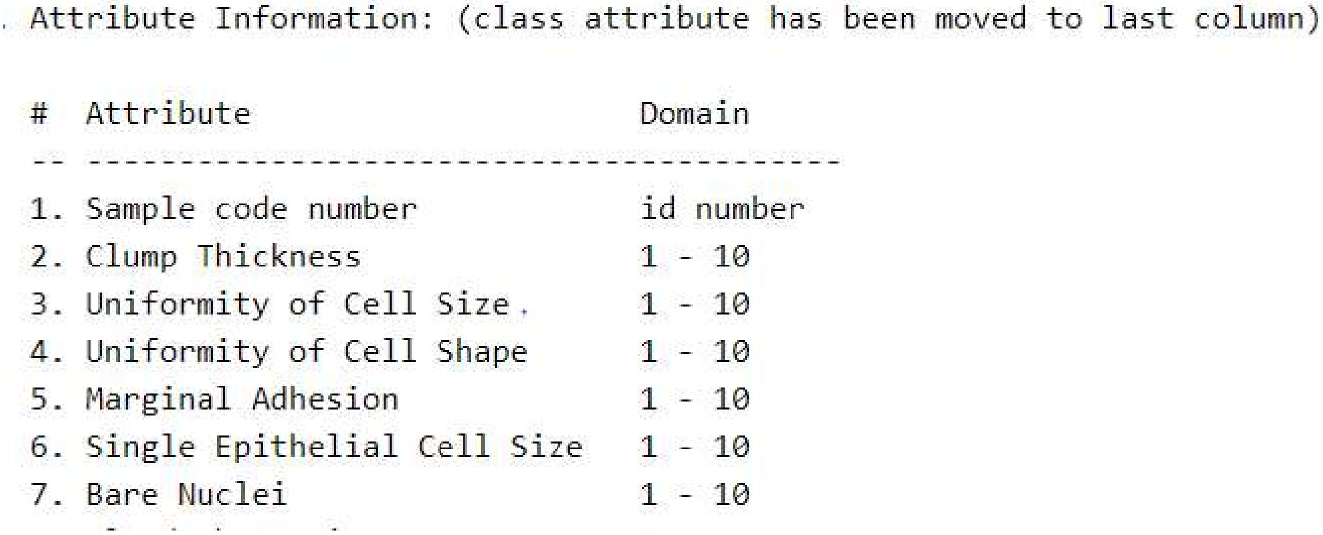
### Breast Cancer Wisconsin (Diagnostic) Data Set

*Download*: Data Folder, Data Set Description

**Abstract**: Diagnostic Wisconsin Breast Cancer Database

| | | | | | |
|---|---|---|---|---|---|
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 569 | **Area:** | Life |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 32 | **Date Donated** | 1995-11-01 |
| **Associated Tasks:** | Classification | **Missing Values?** | No | **Number of Web Hits:** | 1337474 |

From this site we have to download breast-cancer-wisconsin.data file



```
Attribute Information: (class attribute has been moved to last column)

  #  Attribute                     Domain
  -- -----------------------------------------
  1. Sample code number            id number
  2. Clump Thickness               1 - 10
  3. Uniformity of Cell Size       1 - 10
  4. Uniformity of Cell Shape      1 - 10
  5. Marginal Adhesion             1 - 10
  6. Single Epithelial Cell Size   1 - 10
  7. Bare Nuclei                   1 - 10
```

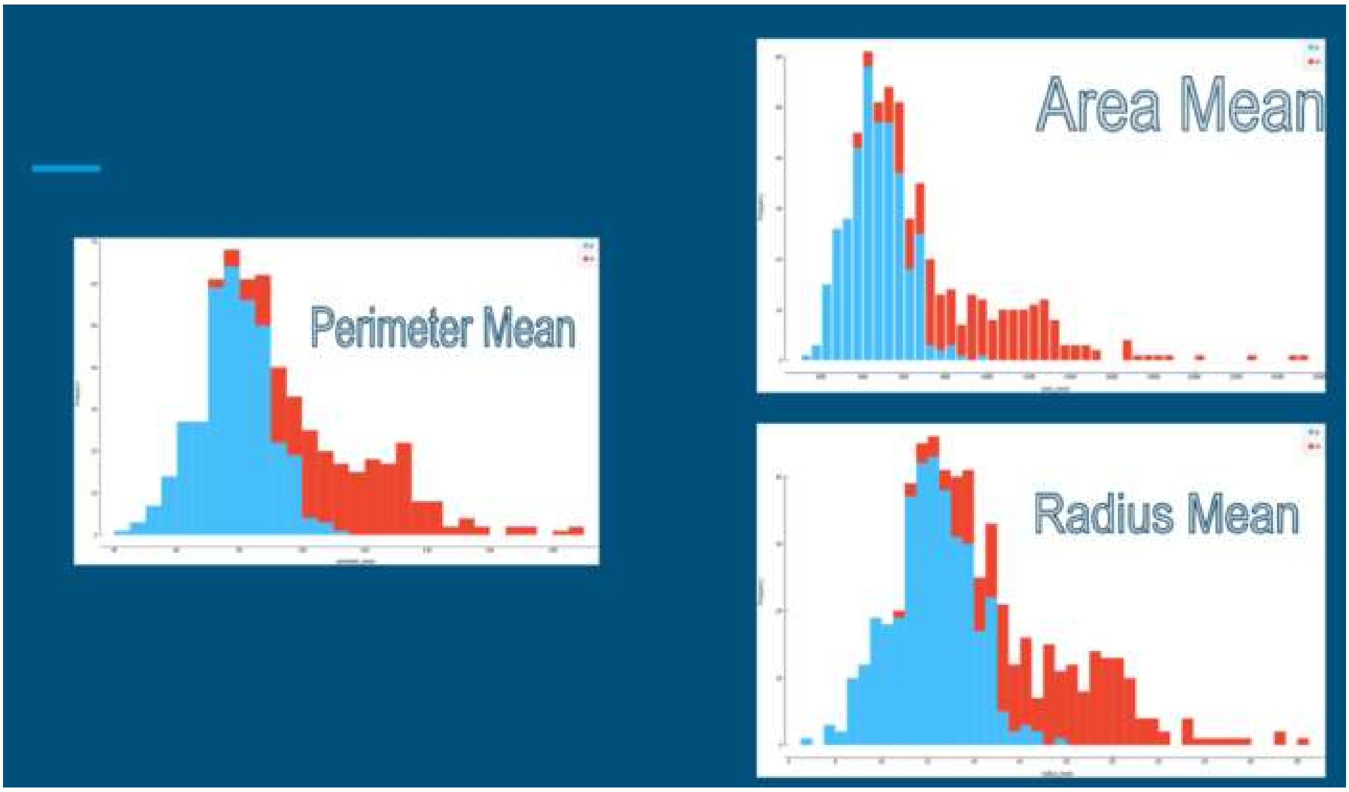
```
 8. Bland Chromatin               1 - 10
 9. Normal Nucleoli               1 - 10
10. Mitoses                       1 - 10
11. Class:                        (2 for benign, 4 for malignant)
```

```
Missing attribute values: 16

There are 16 instances in Groups 1 to 6 that contain a single missing
(i.e., unavailable) attribute value, now denoted by "?".

Class distribution:

Benign: 458 (65.5%)
Malignant: 241 (34.5%)
```

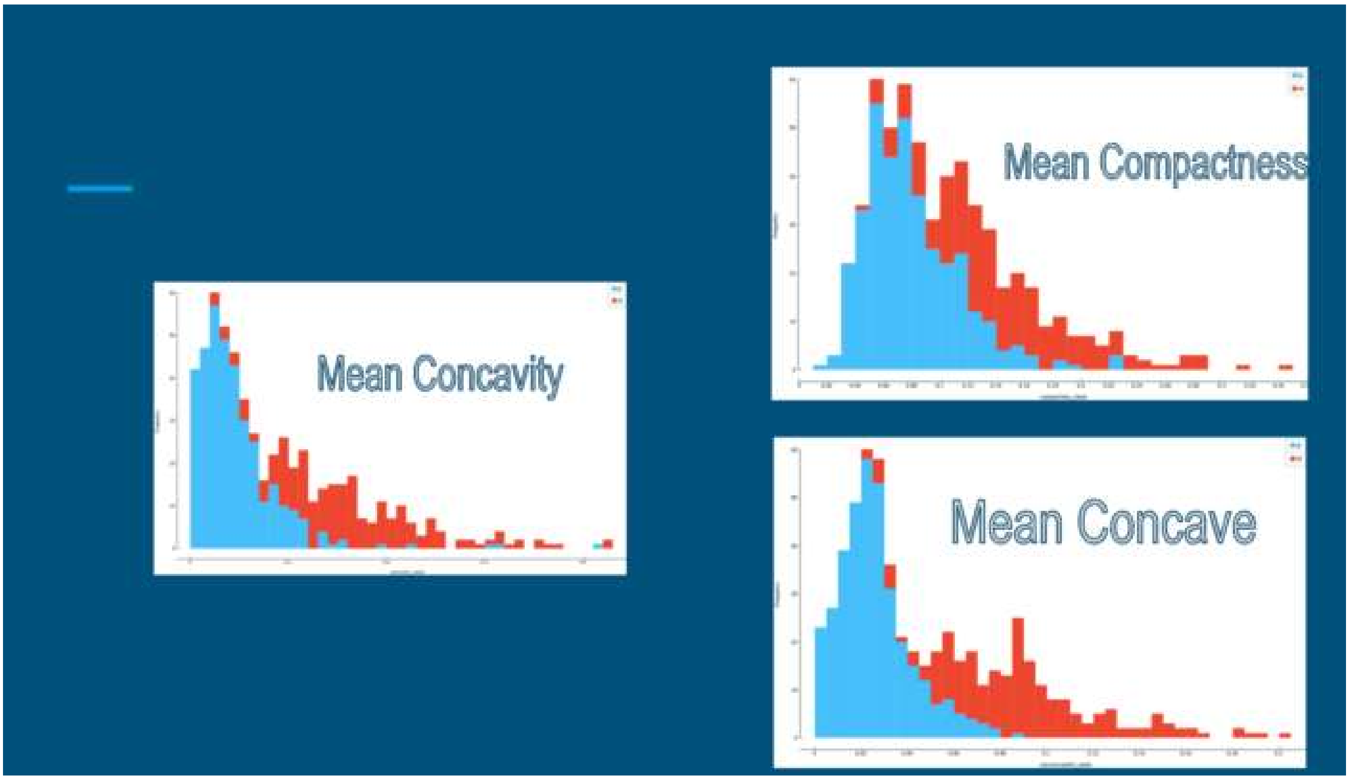
## Data Visualization

### DataSource(1)

- Mean Radius, Mean Perimeter and Mean Area appear to be helpful in classification.Higher the values of each parameter more are the chances of it being malignant.

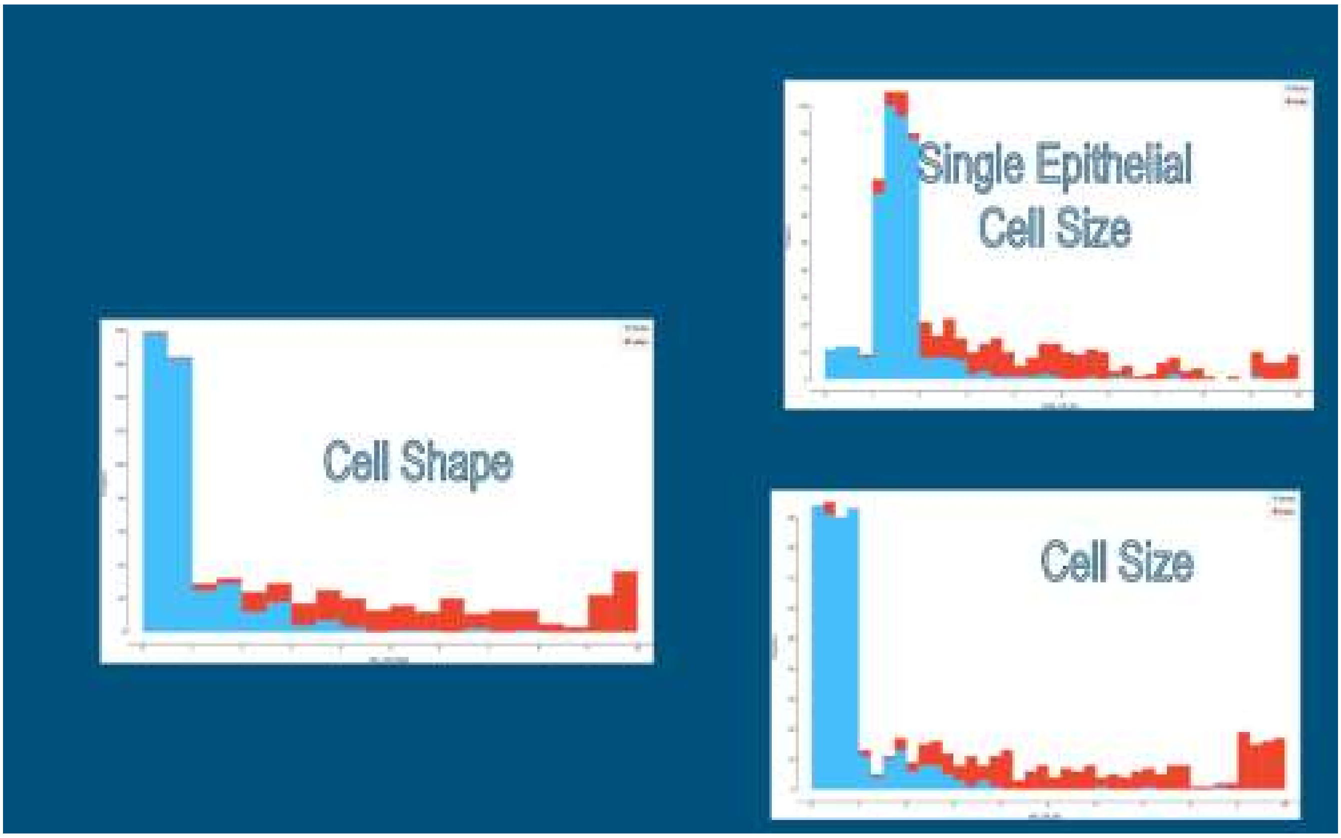


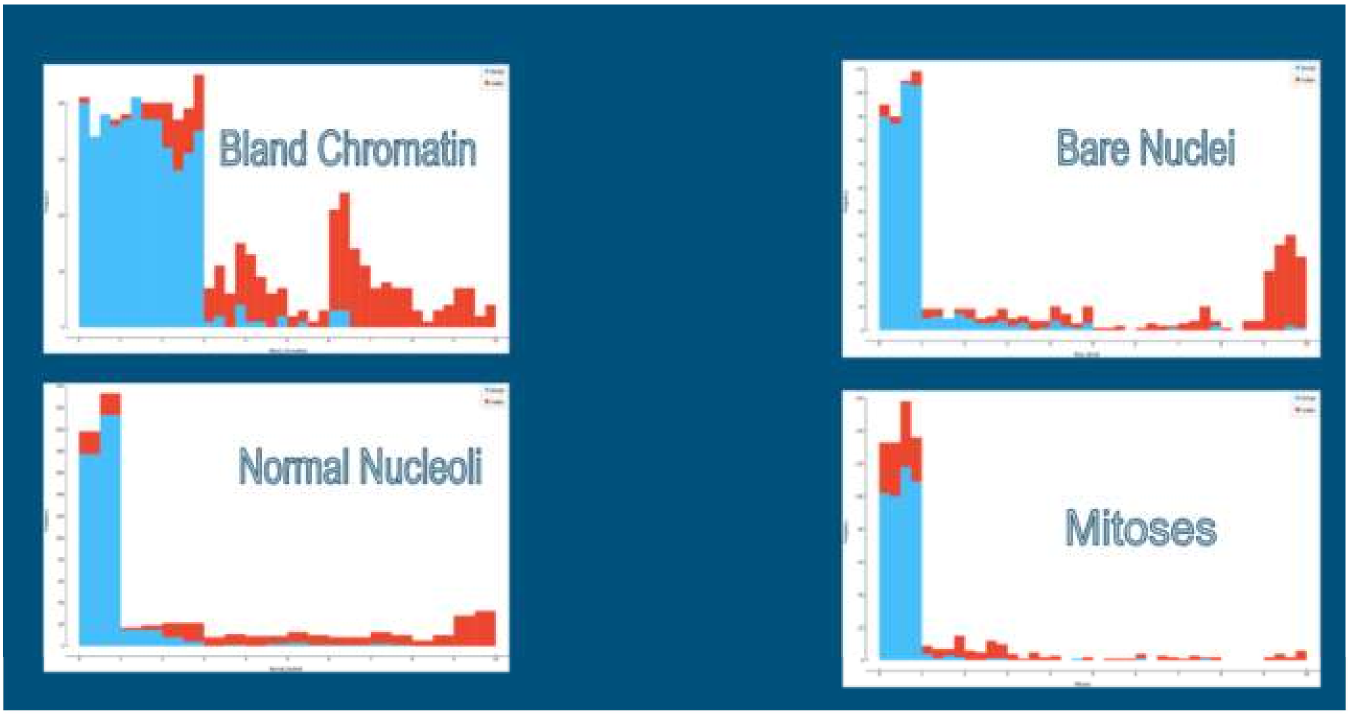
- Mean Concavity, Mean Concave , and Mean Compactness appear to be helpful in classification.Higher the values of each parameter more are the chances of it being malignant.

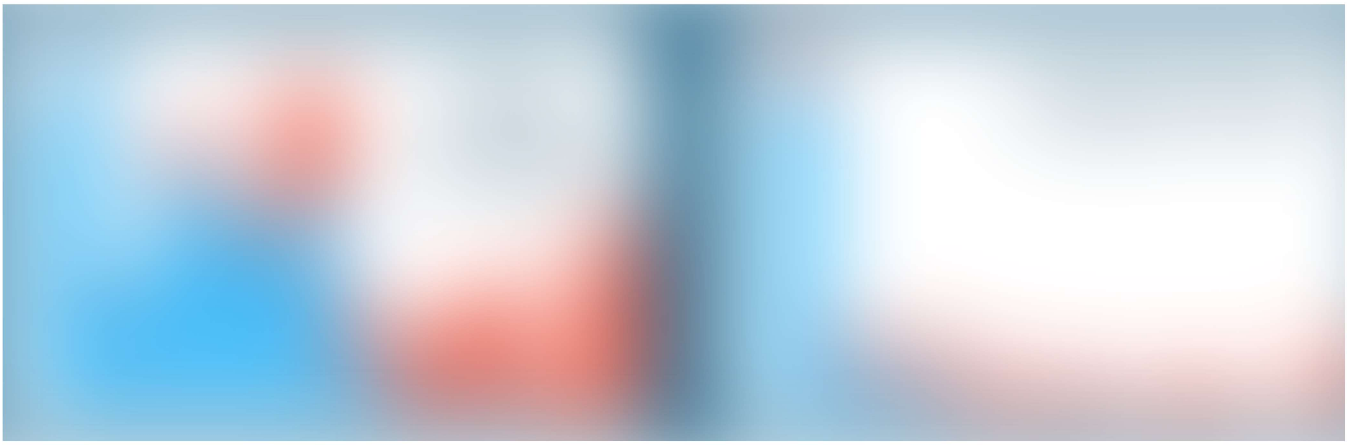


### DataSource(2)

The features distinguish between benign and Malignant fairly well.







**Co-authors : Dharmesh Chauhan**

**Guide : Sagar Patel** (Asst. Professor, KDPIT, CSPIT, CHARUSAT)

Breast Cancer Machine Learning Orange Tool