Breast Cancer Prediction
using classification models

1. Introduction

Cancer is the second biggest cause to increase human mortality rate. More than 70% of deaths due to cancer occurs in underdeveloped and developing countries. Owing to the group of diseases, there is an abnormal growth of body cells which leads to develop malign cells which are termed as cancerous cells. As per some statistics of Canadian government, breast cancer accounts for 13 to 25% of overall deaths due to cancer. Its estimated that in year 2019, around 30,000 cases were diagnosed with breast cancer and 5000 in severe cases.

2. Problem definition

The goal of this project is to use ML models which can detect cancerous cells with a higher degree of accuracy. Using different models, we will be predicting whether a person is diagnosed with cancer or not.

3. Requirement Analysis / Dataset

The dataset chosen to address this problem is obtained from Kaggle (Breast Cancer Wisconsin (Diagnostic) Data Set). The data set will be divided into two sets, one to train and the other to test the model in 80 to 20 ratio. There are 32 features in data set, including

a) ID number
b) Diagnosis (M = malignant, B = benign)

and remaining 30 features are based on real-valued features from images of cell nucleus, for example:

c) radius (mean of distances from center to points on the perimeter)
d) texture (standard deviation of gray-scale values)
e) perimeter
f) area
g) smoothness (local variation in radius lengths)
h) compactness (perimeter^2 / area - 1.0)
i) concavity (severity of concave portions of the contour)
j) concave points (number of concave portions of the contour)
k) symmetry
l) fractal dimension ("coastline approximation" - 1)

4. Methodology

Preprocessing of dataset will be done. Dataset will be trained using 3 different models to determine which gives the best output.

5. Group Members

Deep Singh	–	251122489
Sanket Salunke	–	251102392
Mandeep Singh	–	251122474