

Breast Cancer Detection

This is a classification problem which will be used to find whether or not the breast cancer in a patient is benign or malignant, depending on other factors/symptoms shown by the patient.

Context: Breast cancer is one of the most common cancers in women in the U.S. According to statistics, 1 in 8 women suffer from breast cancer once in their lifetime (breastcancer.org). This amounts to 12% of all females in the women. This commonality calls for efficient diagnosis of breast cancer so that the stage of the cancer can be determined as early as possible in order to begin early treatment. The accuracy of detection (benign or malignant) is crucial in determining which treatment plan to follow and in the survival of the patient. The goal is to predict whether the cancer is benign or malignant with an accuracy of at least 95% using Machine Learning models. This will be done by using various classification models in sklearn.

Key Data Source: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>

Scope: The dataset consists of various parameters of the cells. These parameters were calculated from images taken from patients and they include cell perimeter, radius, texture, compactness, smoothness, etc. The appropriate features will be selected for determining the target variable (“Diagnosis”). Care must be taken to avoid model overfitting.

Criteria for success: Obtain a sufficient accuracy (more than 95%) while making sure there is no overfitting of data (by using cross-validation).