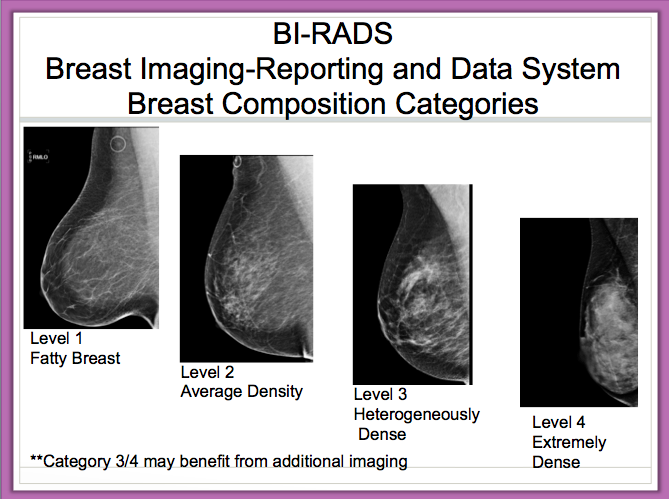
**Predict the breast cancer severity**[**¶**](http://localhost:8888/notebooks/25a-data-science/finalprojectassignment.ipynb#Predict-whether-a-mammogram-mass-is-benign-or-malignant)

Vahid Mohammadnia, June 2018

Breast cancer is the most common cancer in women and second most common cancer among both men and women in the world. There were 1.7 million new cases detected in 2012 that was 12% of all new cancer cases[[1]](#footnote-1).

With this high number of potential cancer cases, a powerful diagnosis method seems to be vital. Mammography is one the most useful methods of screening recommended to the women who are at risk of breast cancer. If positive, the second screening method, autopsy, which is more expensive will be used to categorize the potential patients into benign or malignant groups. But this process has a flaw that is the test data is skewed as the true number of malignant mass is low. That makes it difficult for Mammography to predict false positive with at a high rate and makes many breast biopsies to be unnecessary.



**Figure 1**: A typical mammogram indicating the different breast cartographies based on density, The breast is normally composed of fat and glandular tissue – the higher the proportion of glandular tissue, the denser the breast[[2]](#footnote-2). Note that the level numbers in the figure are different than the numbers assigned to density in the data set.

Accordingly, machine learning algorithms can be applied to help the physicians with a better diagnosis. If domain based data that are important in determining the severity of the cancer are detected and introduced by the physicians, the data scientists can design an algorithm being able to predict it at a high accuracy. We'll be using the "mammographic masses" public dataset from the University of California, Irvin repository[[3]](#footnote-3) in this project.

This data contains 516 benign and 445 malignant instances of masses detected in mammograms, and contains the following attributes:

1. BI-RADS assessment: 1 to 5 (ordinal)
2. Age: patient's age in years (integer)
3. Shape: mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal)
4. Margin: mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 speculated=5 (nominal)
5. Density: mass density high=1 iso=2 low=3 fat-containing=4 (ordinal)( Refer to Figure 1)
6. Severity: benign=0 or malignant=1 (binominal)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Source** | **BI-RADS** | **Age** | **Shape** | **Margin** | **Density** | **Severity (label)** |
| **Second Review** | Positive | 67 | lobular | nominal | low | malignant |
| **Transformed form** | 5 | 67 | 3 | 5 | 3 | 1 |

**Table 1:** This table shows how patient’s data are transformed into integer numbers to be usable by algorithms. Note that BI-RADS will be discarded in algorithms as it is indicative of the severity determined by physicians. 4 middle columns are the features and severity is the label**.**

BI-RADS is not a feature, it is an assessment ranges from 1 (definitely benign) to 5 (highly suggestive of malignancy) and is assigned in a double-review process by physicians. This feature will be discarded in the algorithm as it is used as a measure of algorithm’s accuracy against the physician’s second review. All other features are assigned an integer indicative of their status. This process is necessary as algorithms work only with numerical values.

A supervised algorithm is used to predict the severity level: benign or malignant. In this project I have used eight different supervised algorithms and used k-fold criterion to compare the accuracy of each method. Cleaning the data includes extracting the features, discarding the entries with NA values and scaling the features are common in all methods. Test data is considered to be %25 of data set.

**Algorithms**: Decision Tree, Random Forest, SVM with a linear kernel, K neighbor classifier, SVM with a rbf kernel, SVM with a sigmoid kernel, Logistic Regressions, Keras Neural Network

**Software**: Python

**Modules**: NumPy, Sklearn, Pandas, Keras

**Source data:** mammographic\_masses.data

**Code file:** mammographic-project.py

1. Source: https://www.wcrf.org [↑](#footnote-ref-1)
2. Source: http://drattai.com/areas-of-focus/dense-breast-tissue/ [↑](#footnote-ref-2)
3. Source: <https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass> [↑](#footnote-ref-3)