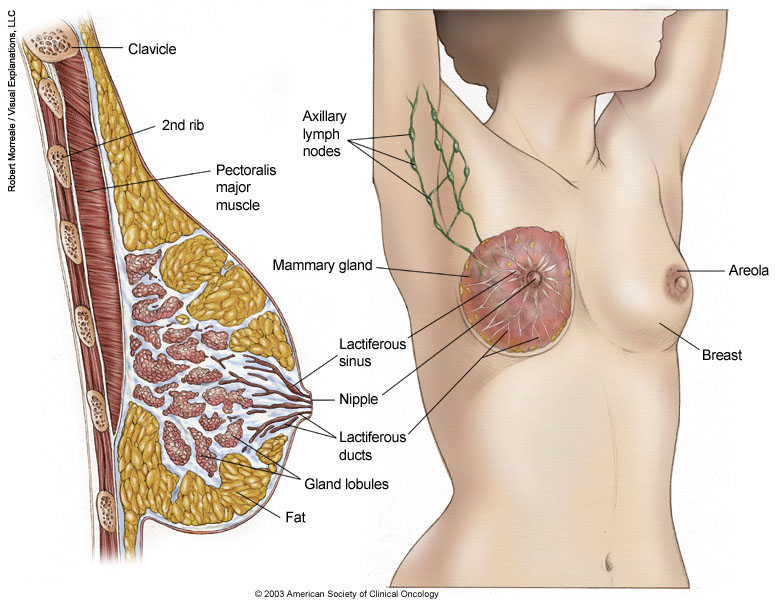
# Computer Aided Detection and Diagnosis of Breast Cancer

## Breast Cancer

According to the World Health Organization, in 2020, there were 2.3 million women diagnosed with breast cancer and 685 000 deaths globally. As of the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world’s most prevalent cancer.

Female gender is the strongest breast cancer risk factor. Approximately 0.5-1% of breast cancers occur in men.

Breast cancer arises in the lining cells (epithelium) of the ducts (85%) or lobules (15%) in the glandular tissue of the breast. Initially, the cancerous growth is confined to the duct or lobule (“in situ”) where it generally causes no symptoms and has minimal potential for spread (metastasis).



Over time, these in situ (stage 0) cancers may progress and invade the surrounding breast tissue (invasive breast cancer) then spread to the nearby lymph nodes (regional metastasis) or to other organs in the body (distant metastasis).  If a woman dies from breast cancer, it is because of widespread metastasis.  

In the past, stage number was calculated based on three clinical characteristics, T, N, and M:

* The size of the cancer tumor and whether or not it has grown into nearby tissue (T)
* Whether cancer is in the lymph nodes (N)
* Whether the cancer has spread to other parts of the body beyond the breast (M)

In 2018, the American Joint Committee on Cancer (AJCC) updated the breast cancer staging guidelines to add other cancer characteristics to the T, N, M system to determine a cancer’s stage:

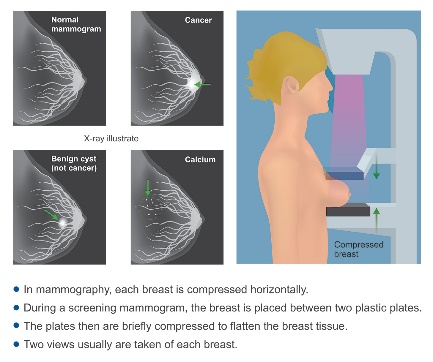
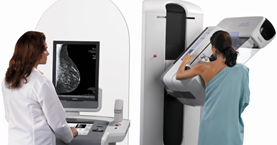
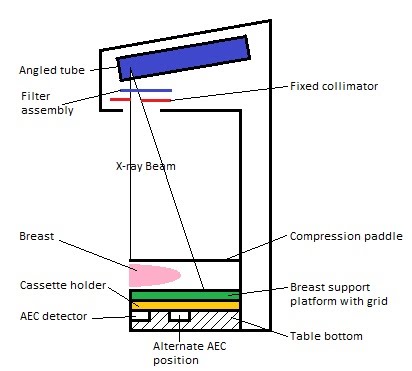
* Tumor grade: a measurement of how much the cancer cells look like normal cells
* Estrogen- and progesterone-receptor status: do the cancer cells have receptors for the hormones estrogen and progesterone?
* HER2 status: are the cancer cells making too much of the HER2 protein?
* Oncotype DX score, if the cancer is estrogen-receptor-positive, HER2-negative, and there is no cancer in the lymph nodes

Treatment of breast cancer often consists of a combination of surgical removal, radiation therapy and medication (hormonal therapy, chemotherapy and/or targeted biological therapy) to treat the microscopic cancer that has spread from the breast tumor through the blood.

Breast cancer treatment can be highly effective, especially when the disease is identified early. Therefore, early detection and state-of-the-art cancer treatment are the most important strategies to prevent deaths from breast cancer.

Getting regular screening tests is the most reliable way to find breast cancer early.

## Mammography



Mammography is the most efficient tool to help detect breast cancer, especially at its earliest stage. It is a process of using low-energy X-rays to examine the human breast and identifying abnormalities, typically through detection of characteristic masses or microcalcifications.

In order to standardize mammography reports, American College of Radiology (ACR) introduced Breast Imaging-Reporting and Data System, **BI-RADS**. BI-RADS is a quality control system, but in day-to-day usage this term refers to the mammography assessment categories. It is used by medical professionals to communicate a patient's risk of developing breast cancer.

BI-RADS Assessment Categories are:

0: Incomplete

1: Negative

2: Benign

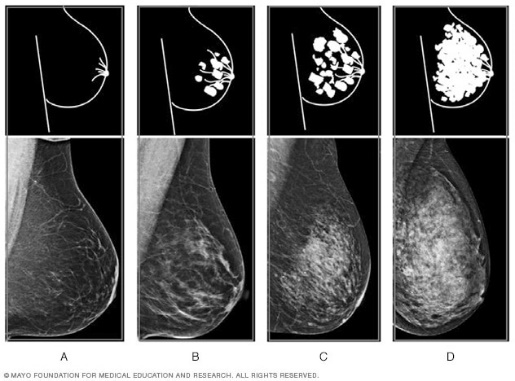
3: Probably benign

4: Suspicious

5: Highly suggestive of malignancy

6: Known biopsy – proven malignancy

Since breast are made up of several types of tissue: fat, glandular tissue (the milk ducts and lobules) and connective tissue, we use **mammographic density** (also called ‘breast density’) as a term to measure and compare the different types of breast tissue visible on a mammogram.

BI-RADS also developed a classification of the breast density levels, since breast of higher density are more likely to develop cancers. There are four categories of breast density:

* Mostly fatty
* Scattered fibroglandular density
* Heterogeneously dense
* Extremely dense

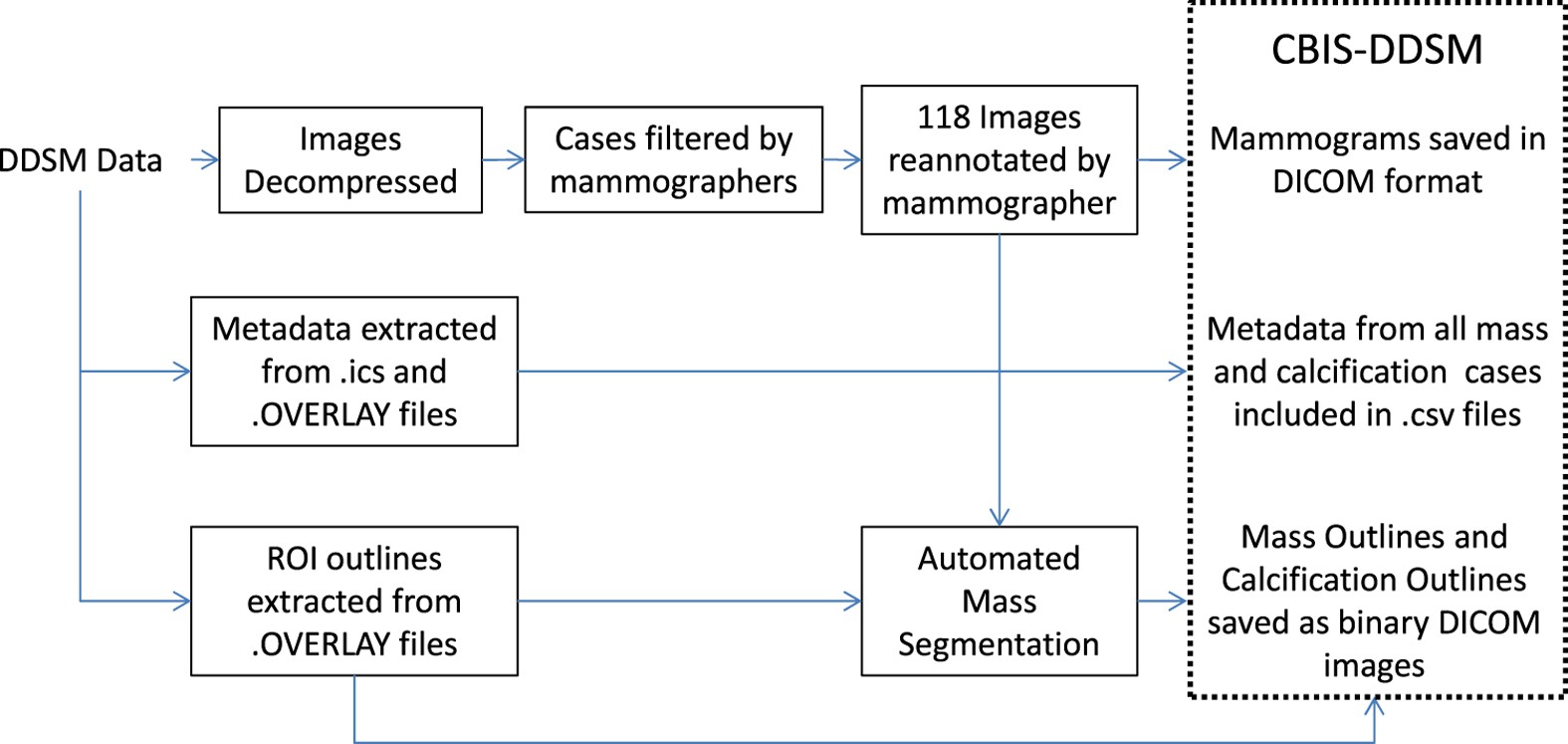
Glandular and connective tissue shows up white on a mammogram while fat shows up dark. Breasts are defined as ‘dense’ if the glandular tissue prevail.

Cancers also show up as white on a mammogram, and therefore it is harder to detect cancers on mammograms of dense breasts than mammograms of mostly fatty breasts.

## Materials - Dataset

The dataset used in this project is CBIS-DDSM[[1]](#footnote-1) (Curated Breast Imaging Subset of DDSM) available at <https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM>. It is an updated and standardized version of the Digital Database for Screening Mammography (DDSM), a database of 2,620 scanned film mammography studies. It contains normal, benign, and malignant cases with verified pathology information. This collection is structured such that each participant has multiple patient IDs. The data set contains 753 calcification cases and 891 mass cases. Therefore it is big enough to be used in decision support systems in mammography.

The figure bellow shows the processes of upgrading the original database: images were decompressed and reformatted in order to be saved as DICOM images; 118 images reannotated by a mammographer and, based on those annotation, an automated segmentation was performed obtaining ROI masks and cropped images of lesions; metadata were also extracted.



Therefore, files with full mammographies, ROI masks of lesions and cropped images by the bounding box of their ROI were provided.

The metadata .csv files contain the following:

* Patient ID: the first 7 characters of images in the case file
* Density category
* Breast: Left or Right
* View: craniocaudal (CC) or mediolateral-oblique (MLO)
* Number of abnormality for the image (This is necessary as there are some cases containing multiple abnormalities.
* Mass shape (when applicable)
* Mass margin (when applicable)
* Calcification type (when applicable)
* Calcification distribution (when applicable)
* BI-RADS assessment
* Pathology: Benign, Benign without call-back, or Malignant
* Subtlety rating: Radiologists’ rating of difficulty in viewing the abnormality in the image
* Path to image files

This project will use for analysis only the calcification cases.

## Methods

The images in this datasets were in DICOM format, but they were digitized analog images and not digital images. Therefore the information about pixel size, pixel spacing and slice thickness were not provided. However, this database was analyzed in many papers, so as suggested in scientific literature[[2]](#footnote-2) the pixel size was set at 1 pixel = 50 µm, while slice thickness was set at 1 mm as in digital tomosynthesis.

### GUI Application for Radiologists

In order to enable radiologists to manipulate the images a Graphical User Interface Application was created.

### Machine Learning Techniques in Breast Cancer Detection and Diagnosis

## Bibliography

<https://www.who.int/news-room/fact-sheets/detail/breast-cancer>

<https://www.cancer.net/cancer-types/breast-cancer>

<https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM>

Lee, R., Gimenez, F., Hoogi, A. *et al.* A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci Data* **4,**170177 (2017).

<https://senologiadiagnostica.it/mammografia/>

<https://jordankupersmith.github.io/its_not_a_tumor/projectdetails/>

Ragab DA, Sharkas M, Marshall S, Ren J. 2019. Breast cancer detection using deep convolutional neural networks and support vector machines. PeerJ 7:e6201

1. Lee, R., Gimenez, F., Hoogi, A. *et al.* A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci Data* **4,**170177 (2017). [↑](#footnote-ref-1)
2. Ragab DA, Sharkas M, Marshall S, Ren J. 2019. Breast cancer detection using deep convolutional neural networks and support vector machines. PeerJ 7:e6201 [↑](#footnote-ref-2)