Akshath Venkataraman

Kaggle Competition

Radiological Society of North America (RSNA) 2023 Breast Cancer Detection Analysis

06/01/2023

Breast Cancer Detection: Exploratory Data Analysis

As part of the exploratory data analysis using the image dataset provided by the RSNA Screening Mammography Breast Cancer Detection, I have conducted statistical analysis of both the meta data about the image dataset and the scanned images analysis. I have captured exhaustive details of the analysis in this document. This will help us better understand the dataset, various categorization of the dataset and data availability for machine learning model development purpose. I have also visualized the dataset for clarity and explanations.

Table 1

Patient metadata sample from Kaggle dataset

	site_id	patient_id	image_id	laterality	view	age	cancer	biopsy	invasive	BIRADS	implant	density	machine_id	difficult_negative_case
0	2	10006	462822612	L	CC	61.0	0	0	0	NaN	0	NaN	29	False
1	2	10006	1459541791	L	MLO	61.0	0	0	0	NaN	0	NaN	29	False
2	2	10006	1864590858	R	MLO	61.0	0	0	0	NaN	0	NaN	29	False
3	2	10006	1874946579	R	CC	61.0	0	0	0	NaN	0	NaN	29	False
4	2	10011	220375232	L	CC	55.0	0	0	0	0.0	0	NaN	21	True

Source: Kaggle dataset metadata csv file

a. site_id - ID code for the source hospital.

b. patient_id - ID code for the patient.

c. image_id - ID code for the image.

d. laterality - Whether the image is of the left or right breast.

e. view - The orientation of the image. The default for a screening exam is to capture two views per breast.

f. age - The patient's age in years.

g. implant - Whether or not the patient had breast implants. Site 1 only provides breast implant information at the patient level, not at the breast level.

h. density - A rating for how dense the breast tissue is, with A being the least dense and D being the densest. Extremely dense tissue can make diagnosis more difficult. Only provided for train.

i. machine_id - An ID code for the imaging device.

j. cancer - Whether the breast was positive for malignant cancer. The target value. Only provided for train.

k. biopsy - Whether a follow-up biopsy was performed on the breast. Only provided for train.

l. invasive - If the breast is positive for cancer, whether or not the cancer proved to be invasive. Only provided for train.

m. BIRADS - 0 if the breast required follow-up, 1 if the breast was rated as negative for cancer, and 2 if the breast was rated as normal. Only provided for train.

n. prediction_id - The ID for the matching submission row. Multiple images will share the same prediction ID. Test only.

o. difficult_negative_case - True if the case was unusually difficult. Only provided for train.

While the dataset is quite comprehensive, there is also missing data situation that needs to be handled. Here we can see the missing data counts:

site_id	0
patient_id	0
image_id	0
laterality	0
view	0
age	37
cancer	0
biopsy	0
invasive	0
BIRADS	28420
implant	0
density	25236
machine_id	0
difficult_negative_case	0

As observed above, there are missing values for age, BIRADS, and density. The missing values can be substituted with mean and mode values for those columns to establish data consistency. This is a common practice. Here are some additional observations seen from the dataset. There are 11913 different patients in the train set. The younger patient is 26 years old. The older patient is 89 years old. Here is the class distribution information for cancer dataset:

- There are 11665 patients negative to breast cancer. Ratio = 97.92%
- There are 248 patients positive to breast cancer. Ratio = 2.08%

As observed here, there is significant amount of class imbalance exists with only 2% of the image dataset classified as breast cancer.

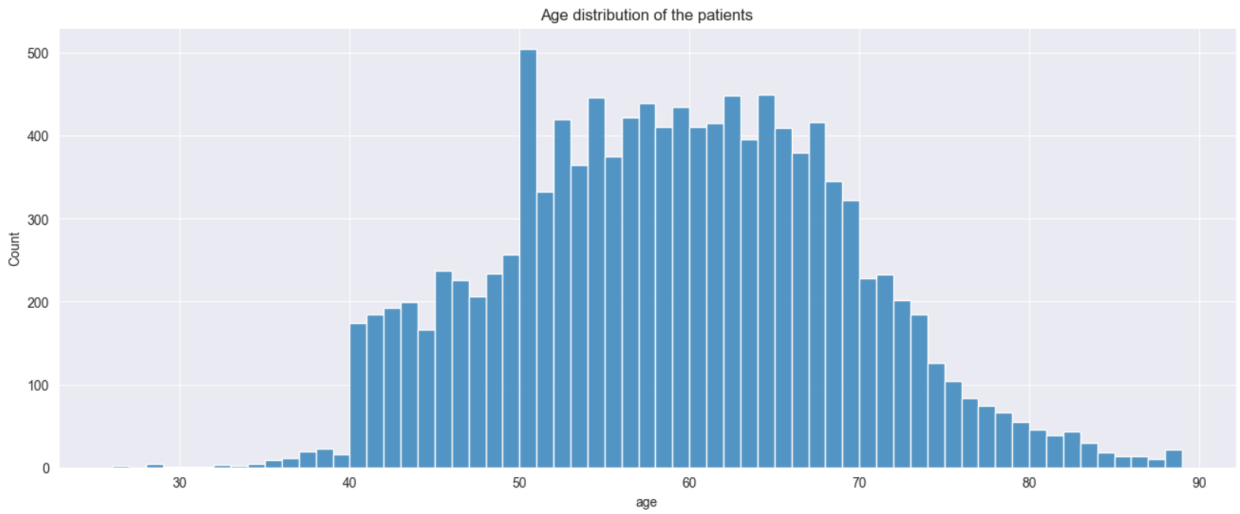

Fig. 1. Age distribution for all the images in the dataset

In addition to the age distribution seen above, here is some additional information for age in the dataset:

- Mean: 58.64
- Standard deviation: 9.89
- Quantile 1: 51.00
- Median: 59.00
- Quantile 3: 66.00
- Mode: 50.00
- Majority of the patients are older than 40 years
- Number of patients seem to peak for 50 years age
- Most patients are between the age of 50 and 70

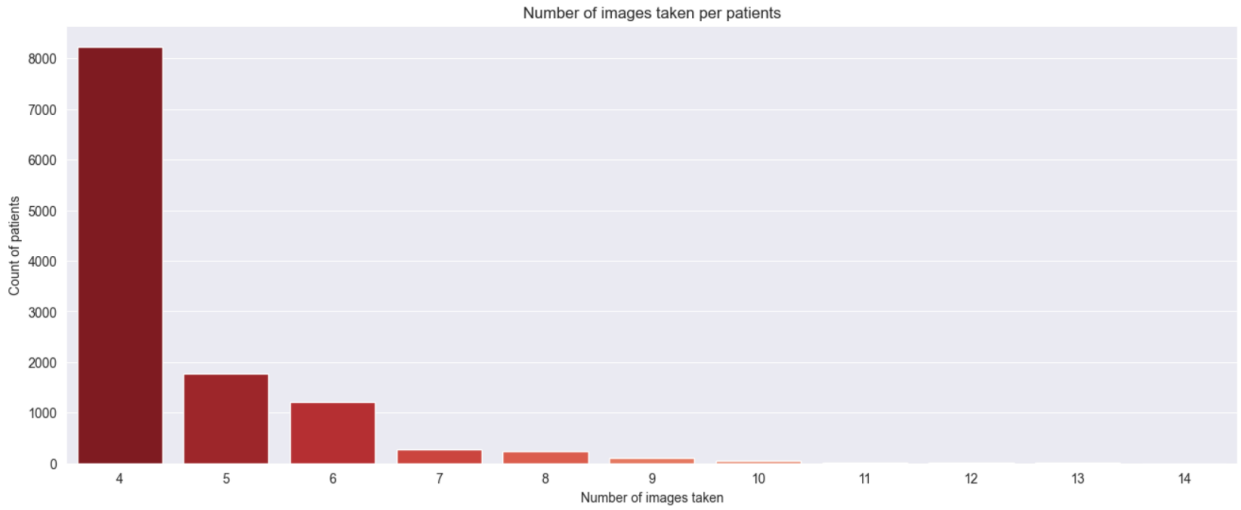

Fig. 2. Number of images taken per patient in the dataset

As observed from above, most patients have 4 images. All of patients have 4 or more images to work with.

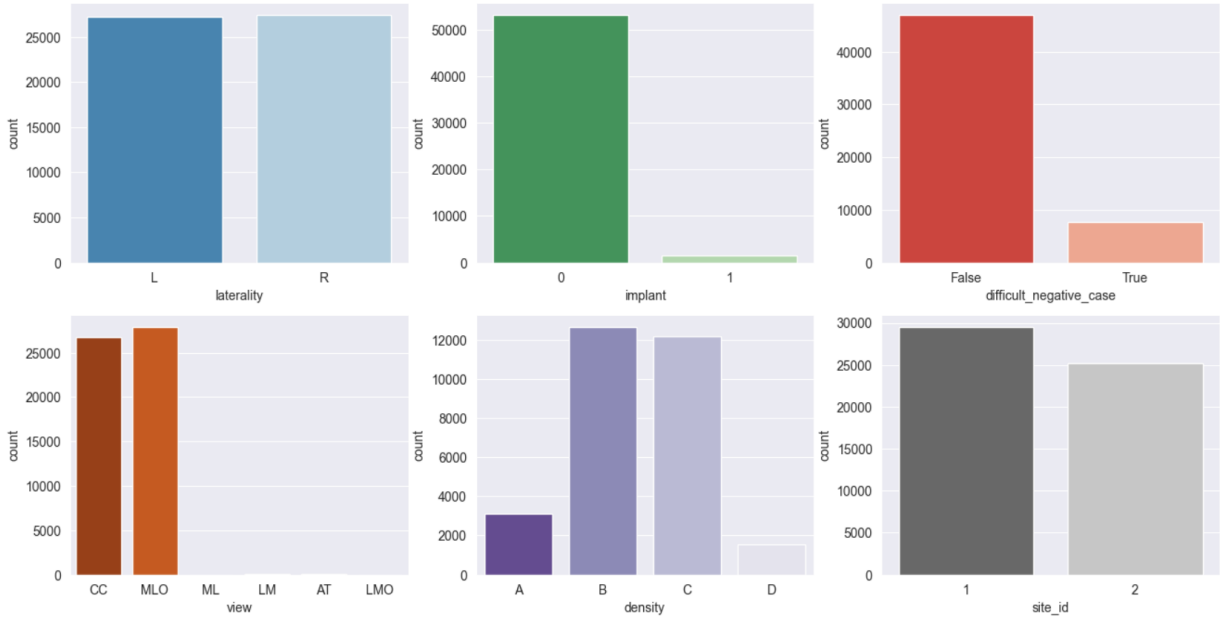

Fig. 3. Number of images across laterality, implant, difficult negative case, view, density and site id attributes

Highlights from above plots:

- Laterality - the images are balanced in terms of laterality

- Implant - Very few images are with implants
- Difficult Negative Case - these very difficult to analyze images
- View - Vast majority of the images are either CC or MLO view category while other category images are available in small numbers (ML, LM, AT and LMO)
- Density - Most of the images show density to be in the middle (B and C), while some images show larger density of A or smaller density of D
- Site ID - Images are from two different sites and they are balanced

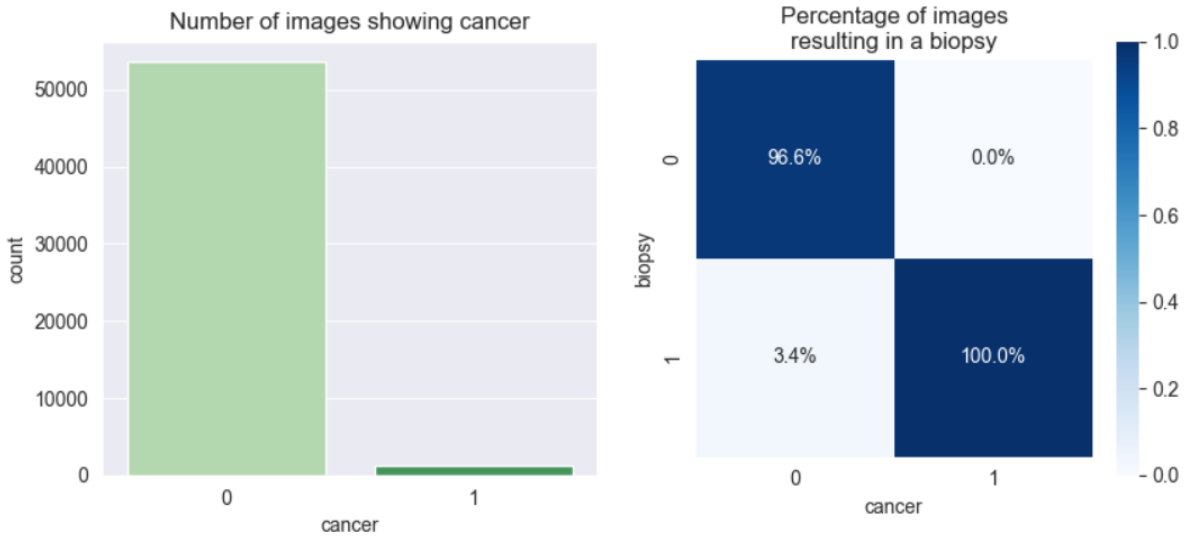

Fig. 4. Number of images showing cancer and percentage of images resulting in a biopsy

Highlights from above plots include the following:

- Cancer images are very low as compared to healthy images
- All patients with cancer had biopsy
- Small number of images without cancer went through biopsy

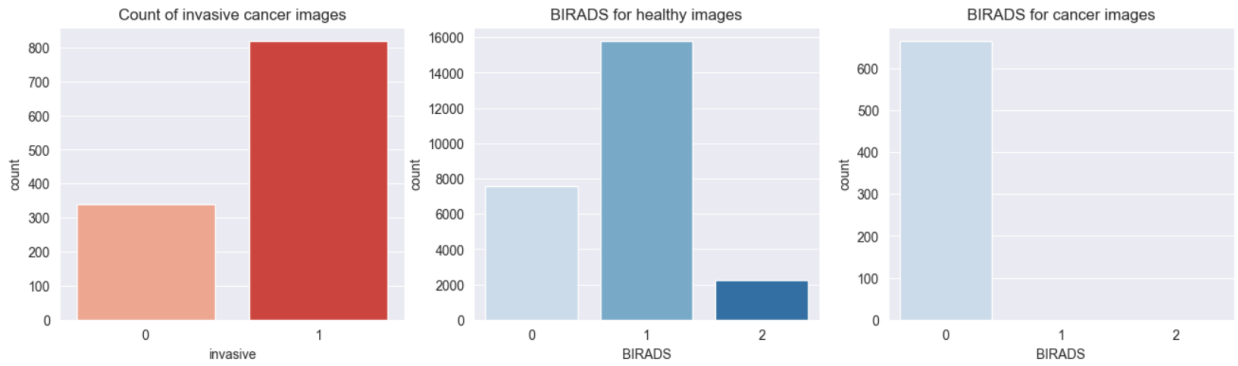

Fig. 5. Count of invasive cancer images and BIRADS for healthy and cancer images

BIRADS come with 3 indicators:

- 0 - required follow-up
- 1 - rated as negative for cancer
- 2 - rated as normal

Highlights from the above chart indicate the following:

- Most of the cancer images are invasive
- Very few healthy images led to follow-up
- All cancer images required follow-up

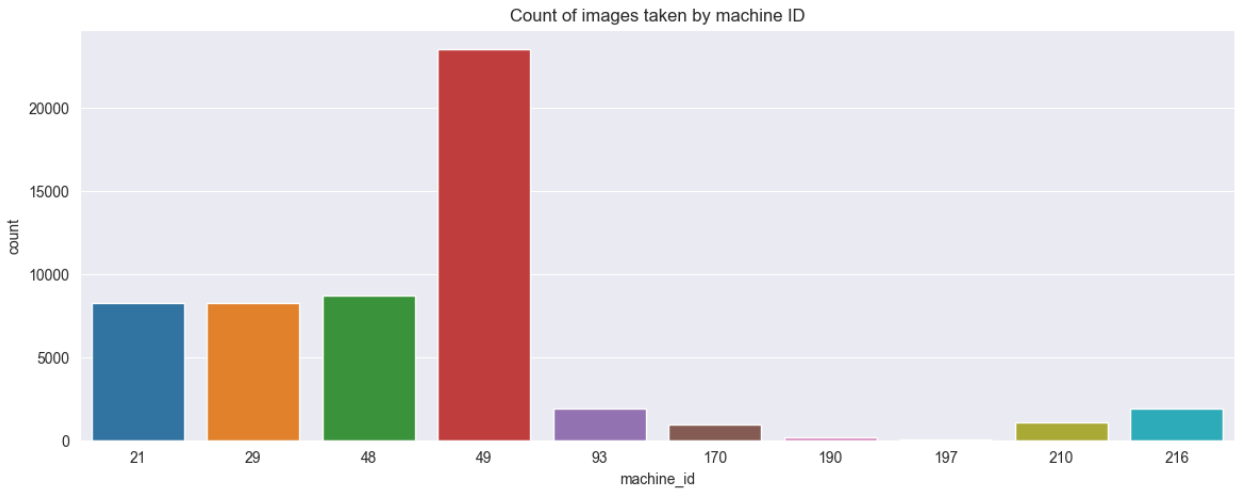

Fig. 6. Count of images taken by machine ID

Here is what is observed from above chart:

- The images are from 10 different machines
- Most of the images are from machine 21, 29, 48, and 49
- Higher machine numbers seem to have produced lower number of images

The image dataset available from Kaggle competition are DICOM image files. All images are DICOM format (Digital Imaging and Communications in Medicine). This is standard format for storing and transmitting medical images along with image related information. It consists of a set of data elements that are organized into a file structure. These data elements contain information about the medical image, such as the patient's name and medical record number, the image modality (e.g., CT, MRI, X-ray), the date and time the image was taken, and the image itself. The image data can be stored in various formats, such as 8-bit or 16-bit grayscale, or 24-bit color. In addition to the image data, the DICOM format also includes metadata that describes the characteristics of the image, such as the image resolution, the size of the image in pixels, and the orientation of the image. This metadata is important for accurately displaying and interpreting the image. The DICOM format is widely used in the medical community for storing, sharing, and analyzing medical images. It is supported by a wide range of medical devices, such as scanners, modalities, and workstations, and is used in hospitals, clinics, and research facilities around the world.

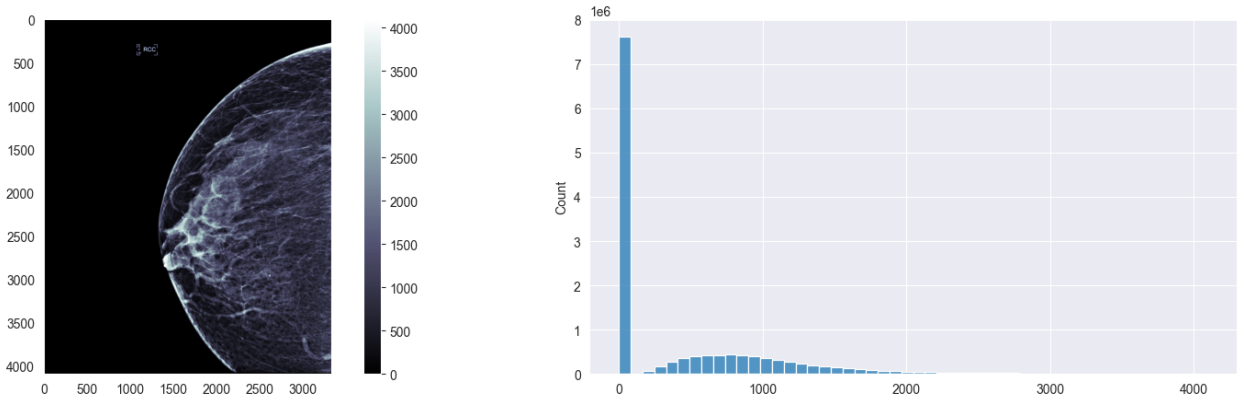

Fig. 6. Pixel distribution of all images

The value of the background is zero as seen above. The raw pixel array is in Hounsfield Units (HU). We can also observe that some image pixel values are very large. Converting the image to HU can help us get image pixel array consistency.

Table 2

Machine ID image count statistics from Kaggle dataset

Machine ID	29	21	216	93	49	48	170	210	190	197
Mode (median)	3476	0	0	0	0	0	0	1017	0	0
Mode (std)	266	0	145	286	0	0	0	2	71	176
Rows (mean)	5355	2776	2294	2862	3819	4096	3804	5072	2322	2168
Cols (mean)	4915	2082	1914	2269	3051	3328	3036	3872	1931	1811

Source: Kaggle dataset metadata csv file

We have the highlights from above table:

- Pixel distribution appears to be different based on the scanned machine ID
- The mode corresponds to background, and it is calibrated to zero for all but two machines (29 and 210)

- The value of Photometric Interpretation must be observed. There are 2 types:
    - In a MONOCHROME1 image, the pixel values represent the grayscale values of the image, with higher values corresponding to brighter pixels and lower values corresponding to darker pixels.
    - In a MONOCHROME2 image, the pixel values are reversed, with higher values corresponding to darker pixels and lower values corresponding to brighter pixels.
- The image size also depends on the machine. Again, machines 29 and 210 have a high resolution compared to others like machine IDs 216 and 197.
- It will be necessary to normalize the image size and the pixel values to train a robust model.

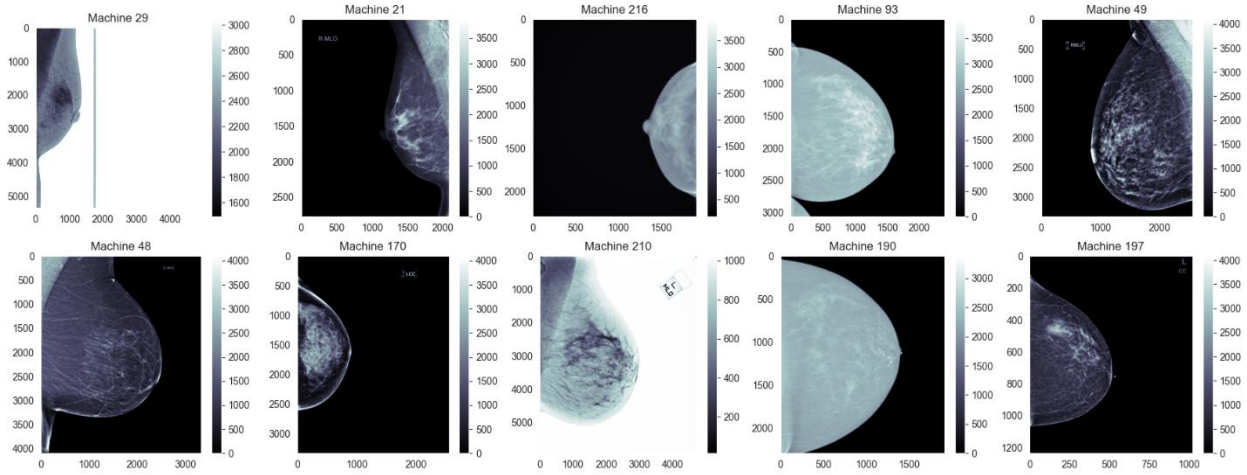

Fig. 7. Sample images from different machine IDs

Here are couple of highlights from above

- Machine 29 and 210 show the plain / white background while other machines show the dark background as observed previously

- It is possible to convert MONOCHROME1 to MONOCHROME2 by applying this formula to the pixel array (then the way all images are interpreted is the same):

  "*array = array.max() – array*"

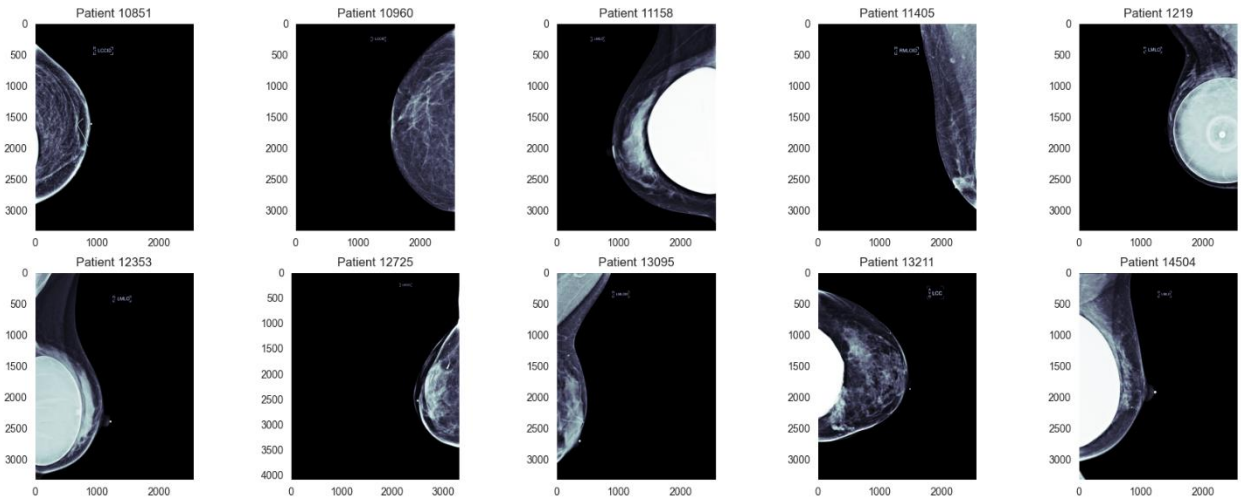

Fig. 8. Sample images with implants

The implant is distinguished by the white mass in the breast. However, it is sometimes not very easy to see it like with patients 10960 and 13095 from the dataset.

Let's check all the scans of a single patient having implant.

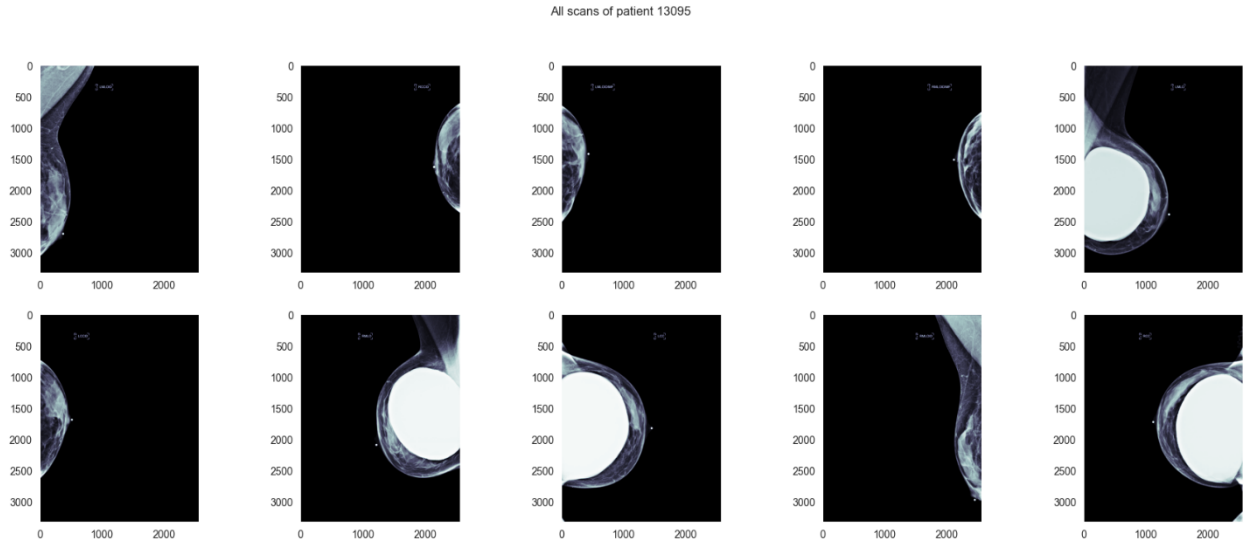

Fig. 9. All scans for patient 13095 with implant

Highlights from above:

- All the scans showing implants are from machines 49 and 170.
- The implants are sometimes only present on one side and not in both breasts for a given patient.
- The metadata does not specify the presence of implant for a scan but for a given patient. It is therefore likely that an image without an implant indicates its presence.

Finally, here are the images that are both cancerous and healthy:

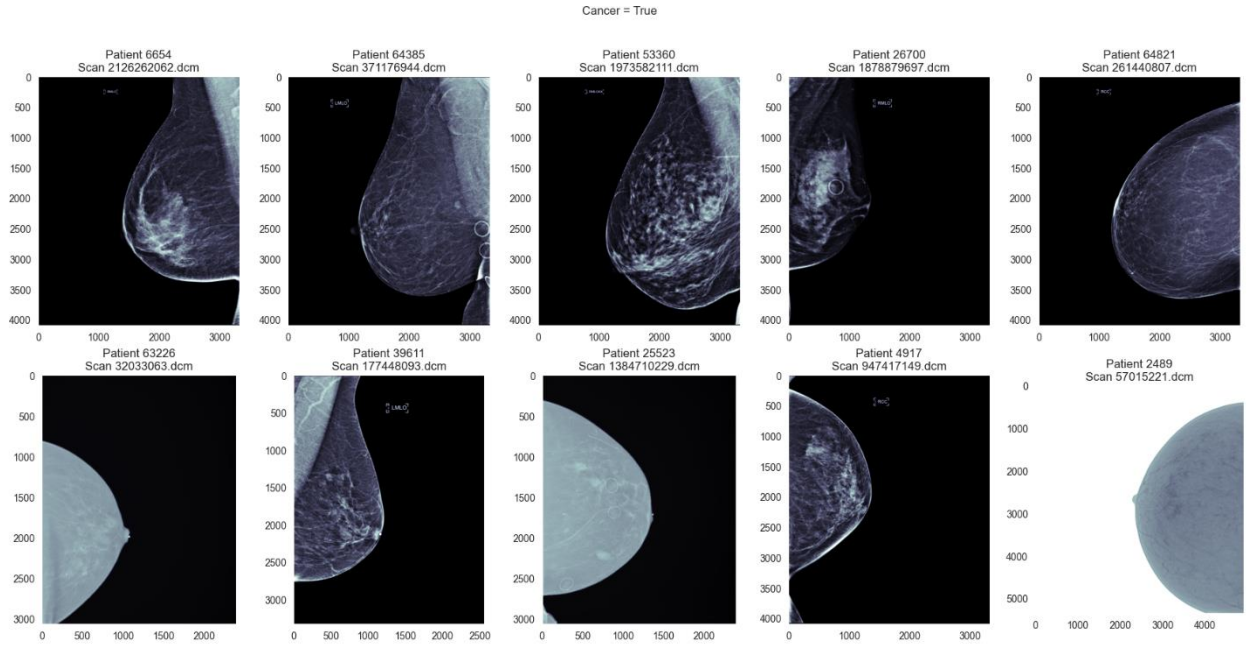

Fig. 10. Images with Cancer

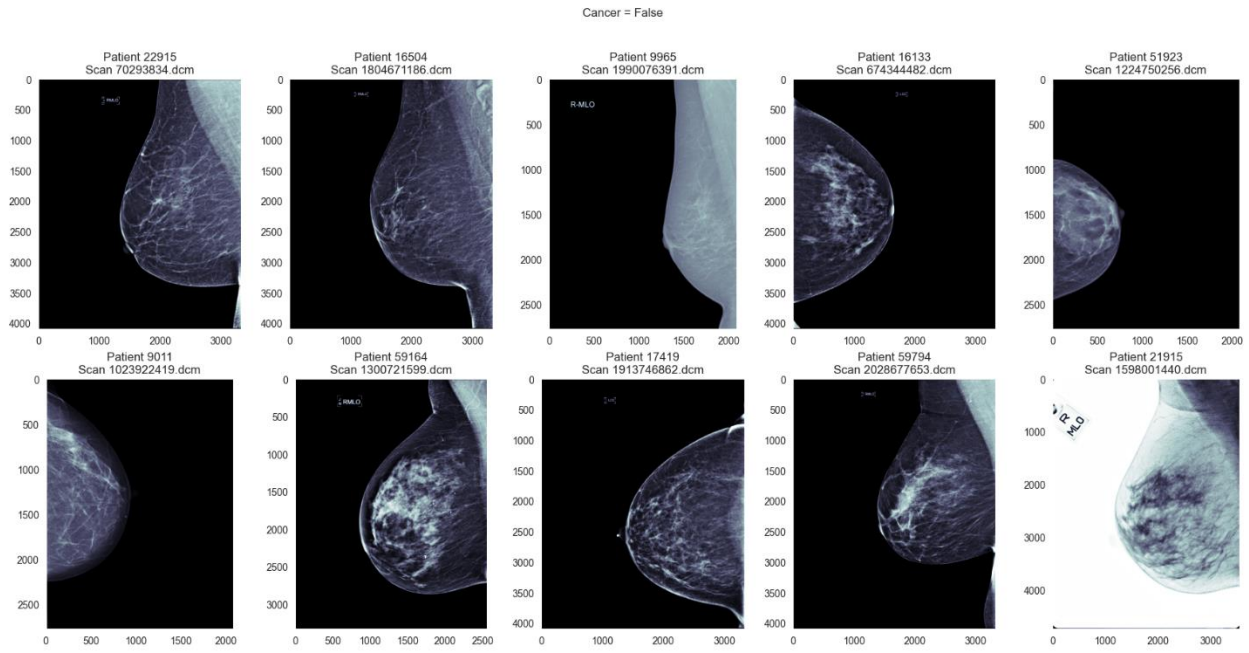

Fig. 11. Images without Cancer

Highlights from Fig 10 and Fig 11:

- It is impossible for me to distinguish by eye an image with cancer from a healthy image
- Again, there are images from different machine IDs which results in different pixel distributions

Conclusions:

- The dataset is heavily unbalanced between scans with and without cancer.
- Most of the patients are over 40 years old.
- Images are quite large and will need to be rescaled during preprocessing.
- Pixel distributions vary significantly depending on the machine ID used.
- The dataset is also unbalanced in terms of images showing implants.
- It is very difficult for a novice to distinguish a scan with cancer from a healthy one.

Performing EDA was very informative. It helped me understand what needs to be taken into consideration during the preprocessing steps.