COMPUTER SCIENCE AND DATA ANALYTICS

REPORT 3

COURSE: GUIDED RESEARCH I

Arzu Fatullayeva
arzu.fatullayeva@gwmail.gwu.edu

Baku 2023

Topic : Finding informative regions in grayscale mammogram images.

Description of dataset

train.csv is Metadata for each patient and image

- **site_id** - The originating hospital's ID number.
- **patient_id** - patient identification code.
- **image_id** - The image's ID number.
- **laterality**: Whether the breast is pictured on the left or right.
- **view** - The image's orientation. By default, a screening exam takes two views of each breast.
- The patient's **age,** expressed in years.
- **implant** - The presence or absence of breast implants in the patient. Site 1 only offers information about breast implants at the patient level, not the level of the individual breasts.
- **density -** Breast tissue is rated according to its density, with A being the least dense and D being the most dense. Tissue that is incredibly dense can make diagnosis more challenging. confined to trains only.
- **machine_id** - The imaging device's ID code.
- **cancer** - Whether or not malignant cancer was present in the breast. the desired amount. confined to trains only.
- **biopsy** - Whether a second breast biopsy was conducted or not. confined to trains only.
- **Invasive** - Whether or not the cancer proved to be invasive if the breast was found to be cancerous. confined to trains only.
- **BIRADS** - 0 if further testing was necessary, 1 if the breast was determined to be cancer-free, and 2 if the breast was determined to be normal. confined to trains only.
- **prediction_id** - The ID of the submission row that matches the prediction. The same prediction ID will be shared by several photos. just a test.
- **difficult_negative_case** - If the case was especially challenging, it is true. confined to trains only.

Data Measurement Strategies

A number of metrics, including accuracy, precision, recall, and F1 score, are used to assess

the model's performance. Proportion of Number of Matching Keypoints Between The Training and Query Images will be used during finding region of interest.

Performance Metrics:

o To measure the performance of the models, we will utilize several common evaluation metrics suitable for imbalanced classification problems, including:

Accuracy: Measures the overall correctness of the model's predictions.

Precision: Determines the proportion of correctly predicted positive instances out of all instances predicted as positive.

Recall: Calculates the proportion of correctly predicted positive instances out of all actual positive instances.

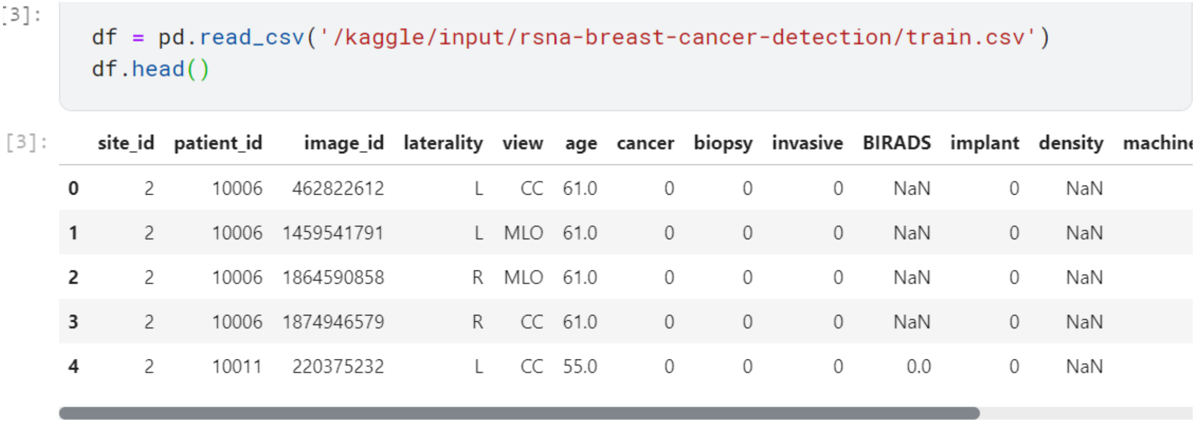
F1-Score: Harmonic mean of precision and recall, providing a balanced measure between them.

Statistical Analysis

It's important to note that the image dataset used in this research was sourced from the RSNA Screening Mammography dataset of dicom images in Kaggle . Because of this, traditional statistical analysis methods typically used on structured text data are not appropriate for this kind of image data, as we discovered in class. Instead, performance indicators unique to image processing and computer vision activities will be the focus of the statistical investigation. The main goal is to evaluate the model's accuracy and reliability to see if they are adequate to complete crucial tasks for the cancer detection machine learning tool.

Visualization of the data

train.csv , which is metadata for each patient and image, used for showing some visualization techniques for breast cancer.



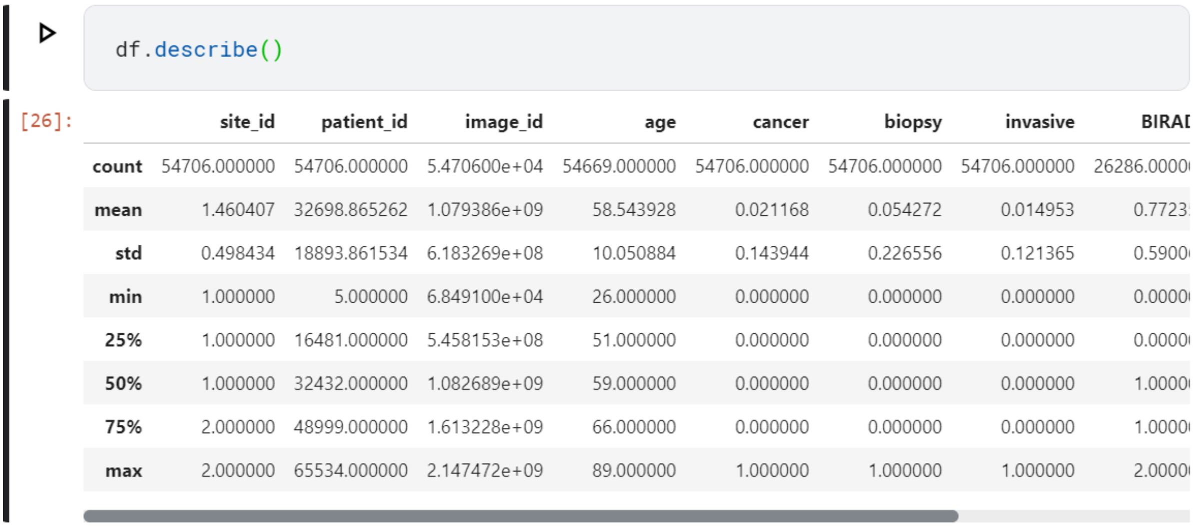
```
[3]:
df = pd.read_csv('/kaggle/input/rsna-breast-cancer-detection/train.csv')
df.head()
```

[3]:

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

Fig. 1 first several rows of dataset

Dataset consists of 14 columns and 54706 rows.



```
df.describe()
```

[26]:

	site_id	patient_id	image_id	age	cancer	biopsy	invasive	BIRAI
count	54706.000000	54706.000000	5.470600e+04	54669.000000	54706.000000	54706.000000	54706.000000	26286.0000
mean	1.460407	32698.865262	1.079386e+09	58.543928	0.021168	0.054272	0.014953	0.7723
std	0.498434	18893.861534	6.183269e+08	10.050884	0.143944	0.226556	0.121365	0.5900
min	1.000000	5.000000	6.849100e+04	26.000000	0.000000	0.000000	0.000000	0.0000
25%	1.000000	16481.000000	5.458153e+08	51.000000	0.000000	0.000000	0.000000	0.0000
50%	1.000000	32432.000000	1.082689e+09	59.000000	0.000000	0.000000	0.000000	1.0000
75%	2.000000	48999.000000	1.613228e+09	66.000000	0.000000	0.000000	0.000000	1.0000
max	2.000000	65534.000000	2.147472e+09	89.000000	1.000000	1.000000	1.000000	2.0000

Fig 2. Some statistical measures : mean, count, standard deviation, quantiles

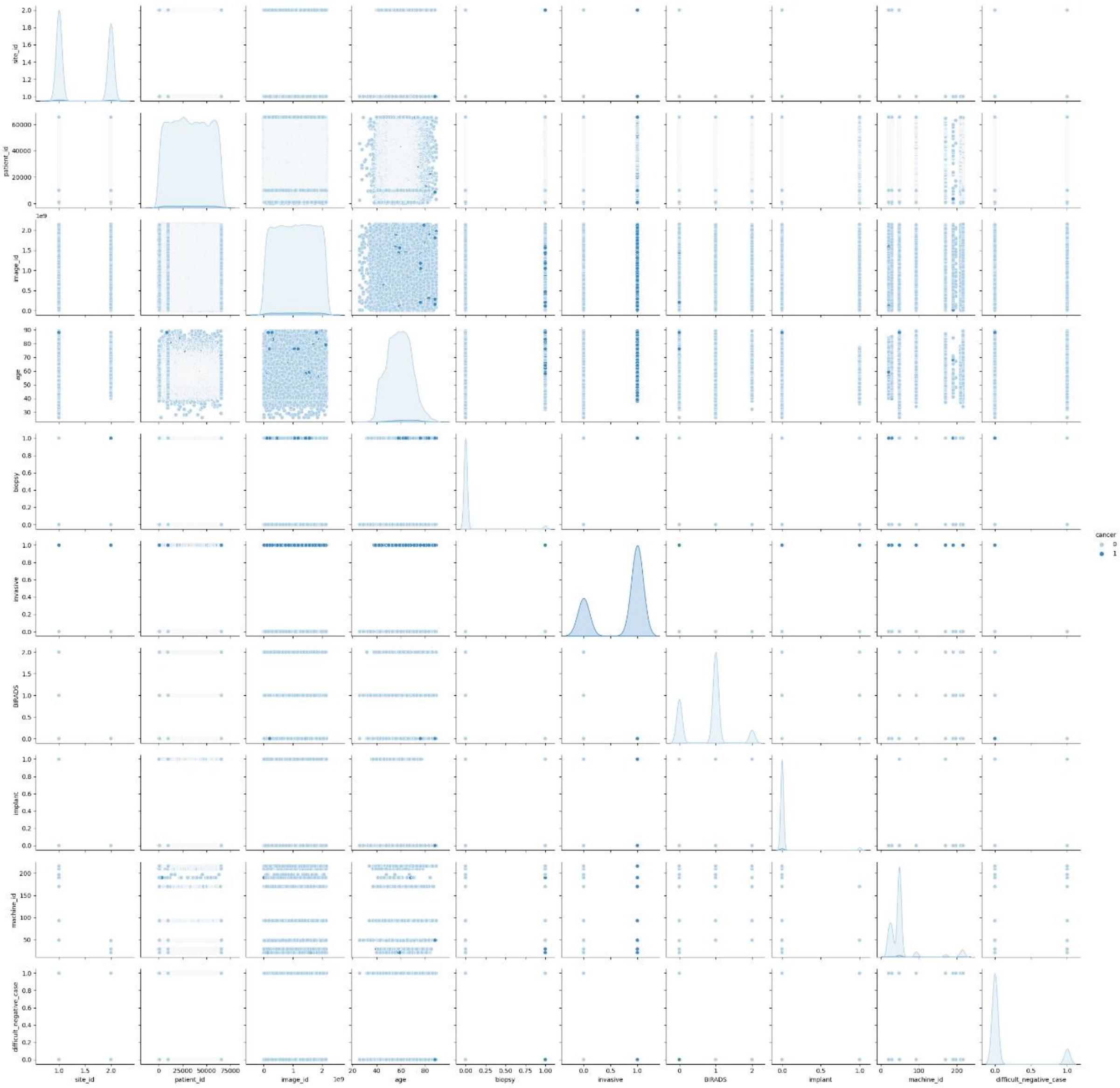


Fig.3 pairplot representation of all variables

We may visualize pairwise relationships between variables in a dataset using the Seaborn Pairplot. By condensing a lot of data into a single figure, this gives the data a pleasant visual representation and aids in our understanding of the data. This is crucial as we try to explore and become comfortable with our dataset.

1. Distribution of ages of woman with breast cancer

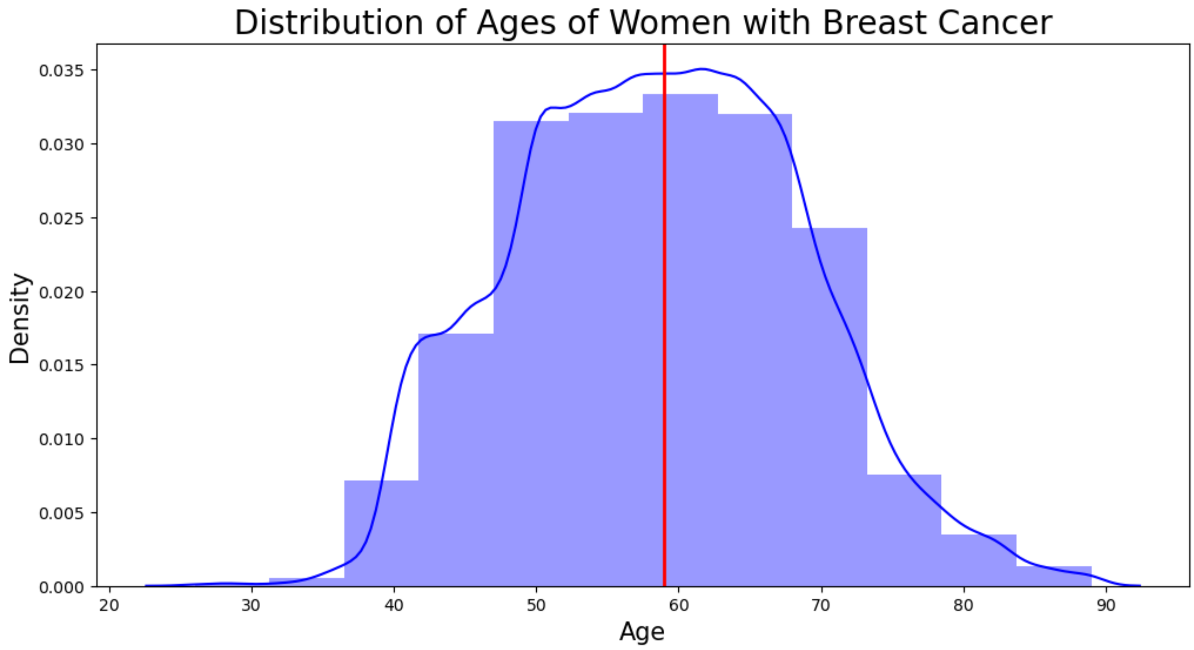
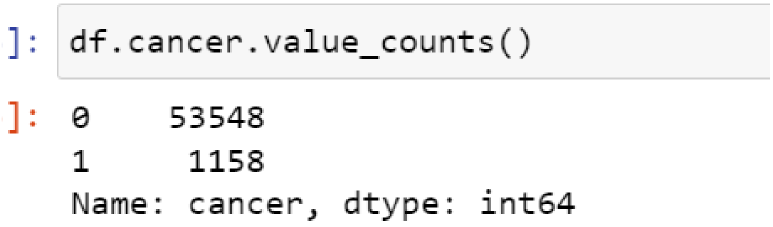


Fig. 4. distribution of women’s ages with cancer

Ages between 48 to 73 are most suffer from Brest Cancer.

The dataset of breast cancer is unbalanced , it can be seen that the number of negative ones are much higher than positives :



```
]: df.cancer.value_counts()

]: 0     53548
   1      1158
   Name: cancer, dtype: int64
```

Fig. 5. Unbalanced dataset

For the dicom images itself it can be used heatmap for further research: