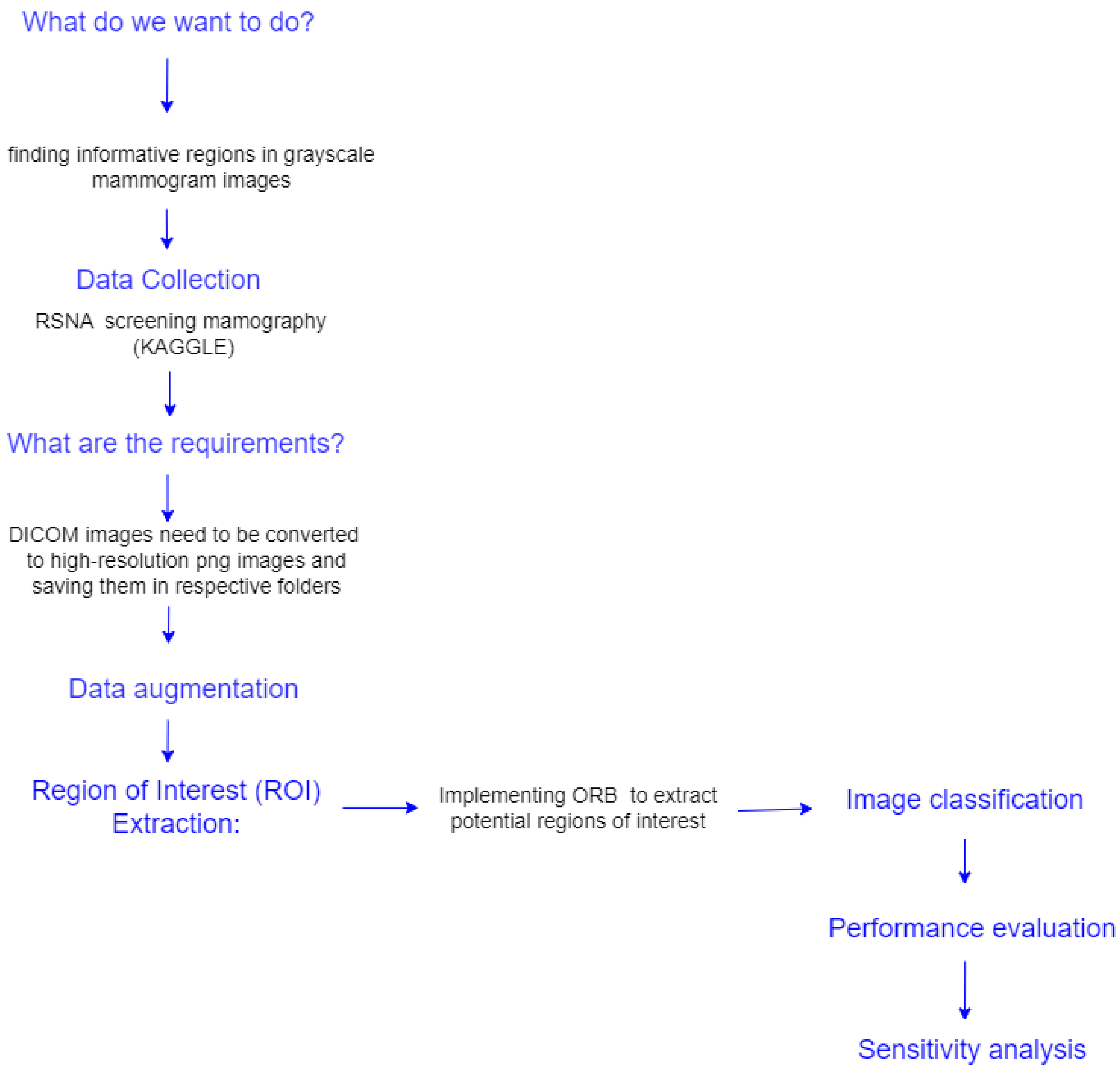
Student: Arzu Fatullayeva

Project: Finding informative regions in grayscale mammogram images.

Date: 25 June, 2023

Selected strategy for the research

For this research, a quantitative strategy is used to develop and evaluate algorithm for. The main idea is on using quantitative data analysis approaches to get objective and measurable results.


What do we want to do?
finding informative regions in grayscale mammogram images
Data Collection
RSNA screening mamography (KAGGLE)
What are the requirements?
DICOM images need to be converted to high-resolution png images and saving them in respective folders
Data augmentation
Region of Interest (ROI) Extraction:
Implementing ORB to extract potential regions of interest
Image classification
Performance evaluation
Sensitivity analysis

Objective: The objective of the research is to develop an algorithm that can automatically identify informative regions in grayscale mammogram images.

Data Collection: A dataset of grayscale mammogram images is collected, consisting of images with known informative regions and corresponding ground truth annotations. The images may be obtained from existing databases or collected specifically for the research.

Preprocessing: The mammogram images are preprocessed to enhance the relevant features and remove noise or artifacts.

Region of Interest (ROI) Extraction: The algorithm(ORB) is applied to the preprocessed mammogram images to extract potential regions of interest that are likely to contain informative features. This step may involve the use of image segmentation, feature extraction, or other computer vision techniques.

Strategy for the Data Collection

The data is collected from the RSNA Screening Mammography, which is Kaggle dataset. The mammograms there in **dicom** format. It can be expected roughly 11,900 patients in the train set. There are usually but not always 4 images per patient.

However, using all this images for research purpose requires computational and storage power. So for this research I will use 1000 images, 500 positive and 500 negative ones. For this aim the train.csv file is used, and the column "cancer", which shows whether or not the breast was positive for malignant cancer, helped to extract 1000 images .

https://www.kaggle.com/competitions/rsna-breast-cancer-detection/data?select=train_images

The collected dataset consists of 1000 images and it will be uploaded to github /data folder.

Data cleansing approaches

When it comes to data cleansing, the first step was converting dicom images to png.

In computer vision, we frequently work with medical images, and nearly all of our databases feature DICOM images. The Dicom image is more than just an image; it contains the pixels information, the patient information, and other things.

In this situation, all that is required of us is to view the image or save it as a PNG file so that we can access it using any software we like.

We require extra Python libraries to accomplish that. The first one is Pillow, which I prefer to use for showing and saving JPG images, and the second one is Pydicom, a library designed specifically for Dicom images. The third option is Numpy, which manipulates arrays.