EECE 544 – Final Project

Title: An Open-Source Computer-Aided Detection Software for Nodules Segmentation Using Mammograms.

Name: Holguer Andrés Becerra Daza.

Student Id: 95948155

Professor: Robert Rohling.

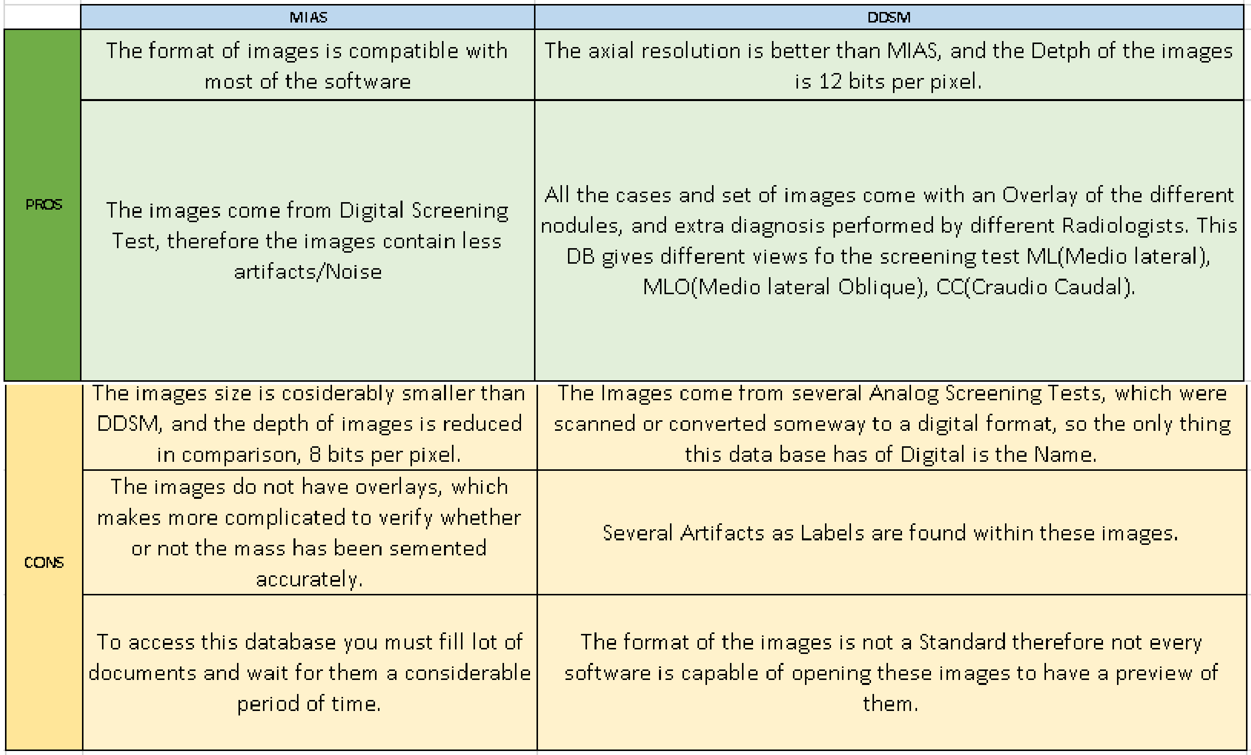
1. Introduction:

Screening mammograms are widely used for detection of cancer due to effectiveness for early detection of breast cancer, and low cost compared with other methods such as CT, and MRI [1], furthermore, the spatial resolution of x-ray which is in the order of few microns permits to visualize masses easily [6]. However, physicians have to be trained to detect and identify accurately cancer on mammograms using qualitative methods for evaluating the screening tests, which it is not 100% reliable and in some cases when a physician(Radiologist) has not been well trained or prepared to detect the type of mass within the breasts, there could be misdiagnosis(misinterpretation of the nodule), which could drive to lose valuable time in early detections therefore the chance of treating cancer earlier is reduced. Nowadays, A lot of Doctors all over the world use Computer-Aided Detection(CAD) Tools(Software) to determine if the detected mass on the mammograms is malign or benign [2], trustworthy CAD tools used by these radiologist cost between $15.000- $18.000[3]. On the other hand, in not developed countries physicians continue using traditional (qualitative ones) methods to identify early detection of breast cancer, where paying a licensed software could be a hurdle for hospitals that can barely afford utilities. This project aims to develop an Open-Source software CAD tool to help radiologist identify whether or not a nodule could be benign or malign for the patient. The software would be capable of processing a mammogram to help the radiologist segment different masses that might be found on the screening test. In this way, the radiologist could acknowledge if the shape of the mass might be malign or benign [4].

1.1. Databases:

There are few reliable Databases that contain valuable images that were used in this project to determine how effective the implemented Algorithms of the CAD tool are to segment the nodules of the screening tests. Among them we have the Mini Mammographic Database for Screening Mammography(MIAS), and the Digital Database for Screening Mammography (DDSM) distributed by the University of Florida, which provides resources with the purpose of studying and researching Breast Cancer for early detection. The DDSM has a corpus of images that are classified in three different categories, Normal breast, Breast with Malign carcinomas, and Breast with Benign carcinomas. The format of the images is LJPEG and the resolution is given in microns (42 microns). However, there are limitations with each one of those Open Source Databases. The MIAS does not have detailed and relevant information about the Nodules shape, it only contains the position of the masses but only the position, which makes harder to determine the shape of those masses to ensure that the algorithms of the CAD tool actually work properly. The shape and geometry of the mass are really important to distinguish among the different types of nodules, and especially to verify or double check whether or not the information given by the CAD tool is relevant and reliable. Furthermore, the images of the MIAS have been adjusted to the size of 1024x1024 pixels (from 2500x4000 pixels) which makes the images lose quality that could be relevant at the moment of segmenting the nodules, the format of the MIAS images is PGM. On the other hand, we have the DDSM Database, which gives better axial resolution and it is much more accurate than the MIAS due to its images resolution and image compression(LJPEG). Notwithstanding, it does not mean that this data base is perfect, there are downsides that make theses images not be easier to use for a common CAD tool.

The following table shows the different cons and pros of the 2 databases:



	MIAS	DDSM
PROS	The format of images is compatible with most of the software	The axial resolution is better than MIAS, and the Detph of the images is 12 bits per pixel.
	The images come from Digital Screening Test, therefore the images contain less artifacts/Noise	All the cases and set of images come with an Overlay of the different nodules, and extra diagnosis performed by different Radiologists. This DB gives different views fo the screening test ML(Medio lateral), MLO(Medio lateral Oblique), CC(Craudio Caudal).
CONS	The images size is cosiderably smaller than DDSM, and the depth of images is reduced in comparison, 8 bits per pixel.	The Images come from several Analog Screening Tests, which were scanned or converted someway to a digital format, so the only thing this data base has of Digital is the Name.
	The images do not have overlays, which makes more complicated to verify whether or not the mass has been semented accurately.	Several Artifacts as Labels are found within these images.
	To access this database you must fill lot of documents and wait for them a considerable period of time.	The format of the images is not a Standard therefore not every software is capable of opening these images to have a preview of them.

Table 1- Comparison between MIAS and DDSM databases

As it can be observed in the Table 1 there are different pros and cons that could lead to determine which database is more useful for this project, which its main purpose is to help the radiologist segment the different nodules within the images.

Therefore, the most suitable database for this project the DDSM because it gives an overlay and diagnosis of each one of the cases. The overlay is crucial to compare with the results of the segmentation process, furthermore, the DDSM gives different views of the screening test

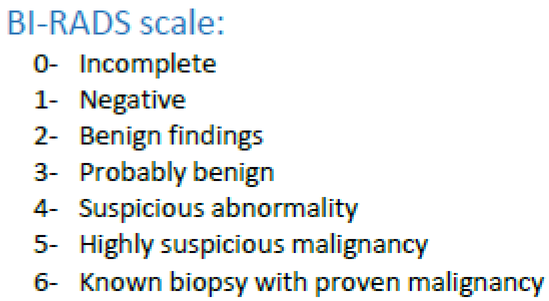
1.2. BI- RADS Breast Imaging Reporting and Data System

BI-RADs is a scale that was established by the American College of Radiology [7]. This a measurement that determine the categories for breast cancer diagnosis into a small number of well-defined levels. These units started out for use with breast screening mammography, and it was adapted also for use with Magnetic Resonance Imaging and breast ultrasound as well.

There are many benefits that this scale brings to the radiologist such as the standardization, and reliability of each performed assessment about the patient condition and the possibility of cancer, helping this way the radiologist determine thereby the statistics that could drive to a reliable diagnosis or early treatment.

1.2.1. BI-RADS Categories.

There are 7 categories (Figure 1 - BI-RADs scale) from the 0 to 6, among you can find different aspects that can help the radiologist determine the chance and probabilities that the patient actually could have cancer.



BI-RADS scale:

0- Incomplete
1- Negative
2- Benign findings
3- Probably benign
4- Suspicious abnormality
5- Highly suspicious malignancy
6- Known biopsy with proven malignancy

Figure 1 - BI-RADs scale

1.2.1.1. Category 0 (Assessment is incomplete):

There is not enough information about the patient status, therefore, the patient has to undergo Ultrasound assessment if there are round cyst to determine whether or not the these are benign. From this stage the result of the additional assessment could reach either the stage 1 or 2.

1.2.1.2. Category 1 (negative):

The patient has no macrocalcifications, and there are not relevant changes between the first assessment and the second.

1.2.1.3. Category 2 (benign):

Even though, there was something suspicious on the mammograms, the patient is diagnosed to not have any breast cancer or malignant nodule. Generally speaking, the nodule has characteristics such, roundness, and the density of the nodule is higher and observable at first sight (Figure 2). The macrocalcification are round too, and haven't change their size since the first assessment.

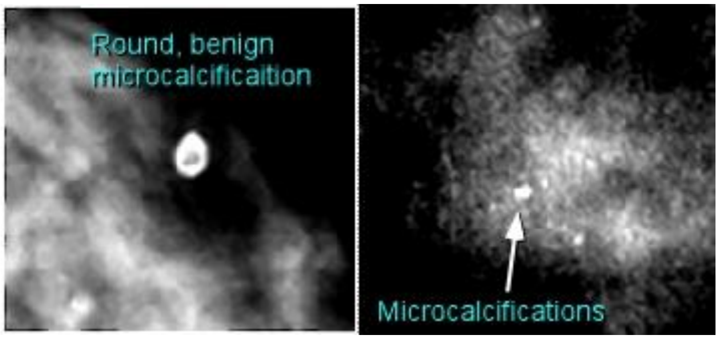


Figure 2 Benign nodules, and microcalcifications

1.2.1.4. Category 3 (Probably Benign):

The mammogram show clearly not solid nodules, there are tiny clusters of calcifications that are scattered, the edges of the found nodules do not have a defined rounded shape (Figure 3).

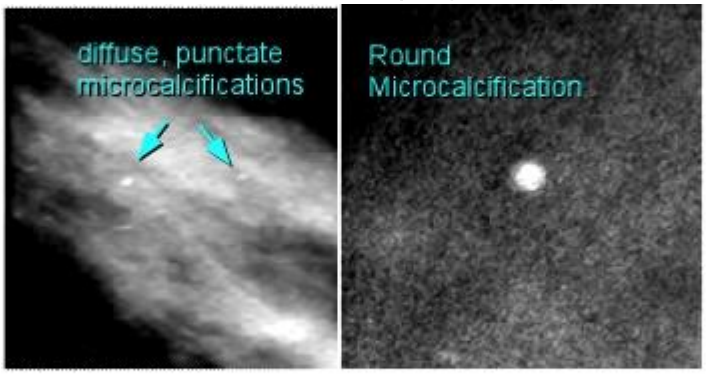


Figure 3 - Category 3

In this stage, it is really hard to determine accurately if the patient has real cancer, therefore a follow-up is performed after 6 months of determining this category, to see if there is any change at all.

1.2.1.5. Category 4 (suspicious or indeterminate abnormality):

The mammogram contains powderish macrocalcifications and it can't be determined the shape or center of the nodule, the shape is asymmetric, but there is a heterogenous density, as can be observed in the Figure 4.

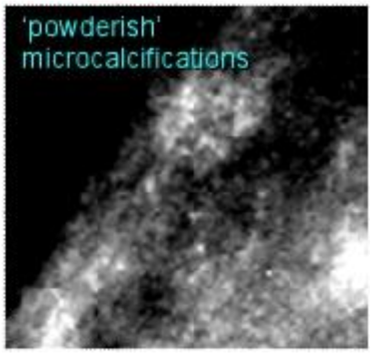


Figure 4 - Category 4, powederish Microcalcifications

It is recommended that the patient needs a biopsy, there is a 30% chance that he person has malignant nodules.

This category has within it 3 sub-categories, A, B, and C, the higher letter the higher possibility of malignancy, 13%, 36% and 79% respectively. All these categories could be diagnosed as Ductal Carcinoma In situ (DCIS).

1.2.1.6. Category 5 (Highly suggestive of malignancy):

A biopsy should be taken immediately because there is high possibility of malignancy. Crushed stone macrocalcifications, and casting macrocalcifications. The shape of the nodules is irregular and they have no rounded contours, spiculated opacities (Figure 5).

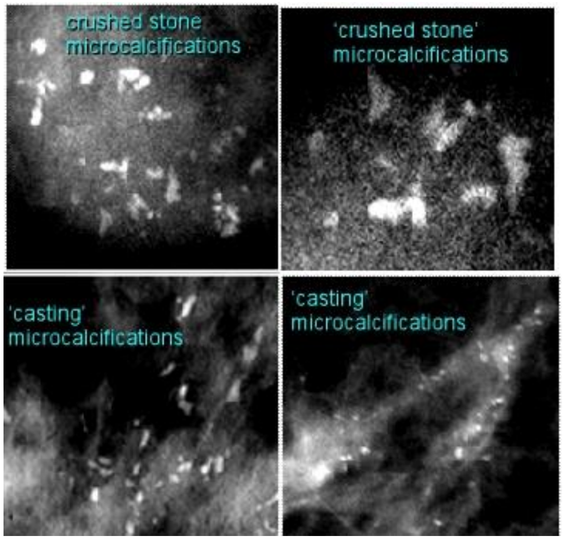


Figure 5 - Category 5, Highly suggestive of malignancy

1.2.1.7. Category 6:

This category is only reached after a biopsy, and a breast cancer treatment is mandatory. This category is not an statistical relevant for prevention, therefore there is a particular categories with people with breast cancer.

As the different categories show, a CAD tool is vital to determine in which categories from 0 to 5 the patient is, preventing in most of the cases malignancy when the images can be analyzed earlier than the stage 5.

Therefore the designed software (CAD tool) can only assist the radiologist with early identification, but other methods are involved to determined accurately the existence of breast cancer such biopsy or ultrasound.

1.3. Software Structure

The software structures and architecture for this project is not the main purpose of this work, however it is vital for future improvements and therefore to meet the requirements of an Open-Source tool that might be improved by several people all over the world.

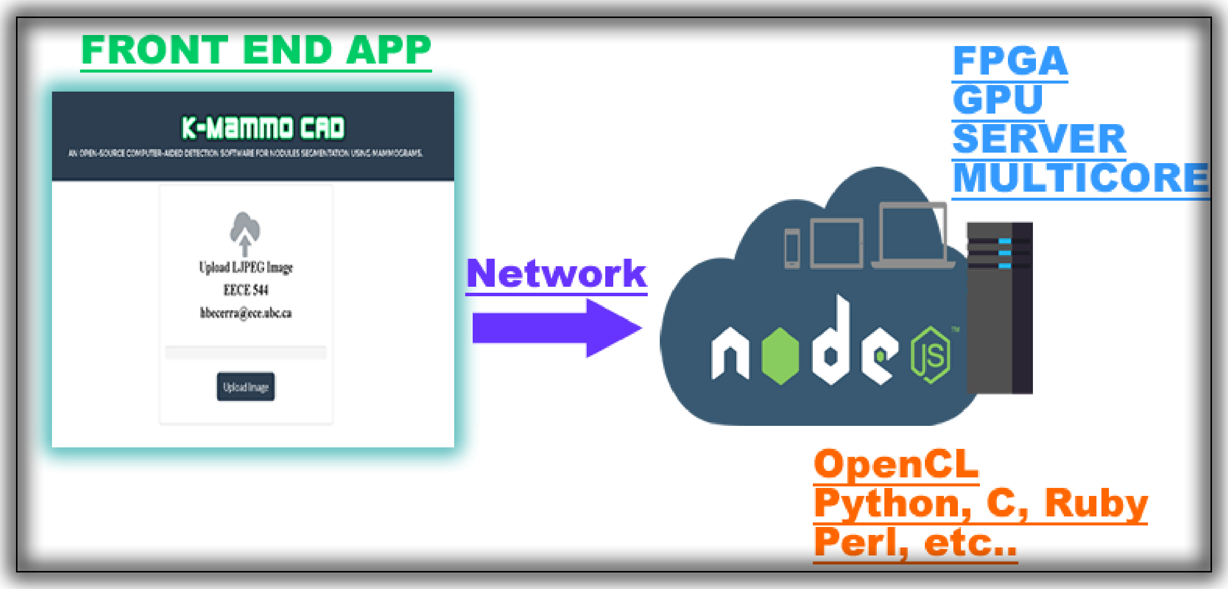


Figure 6 - Software Structure

As it can be observed on the Figure 6, the software is being designed with the purpose of being a web and open source application, that the radiologist can use from a remote computer, smart-phone, or tablet. All the algorithms are executed by an external server through a nodejs server which was written in Java Script, this server has the capability of receiving a LJPEG image and show to the user a high-quality image, and a bar of tools that help the radiologist to implement different types of segmentations process to specific regions of the Image of interest. In addition, this architecture brings several benefits for the developer such as not restriction of programming language to process the images, it is platform independent, and all the algorithms could be accelerated on the back end, giving the Radiologist the tranquility of having a tool that does not require excessive computational power. The only con this platform has for the user is that is network dependable, however, anybody can use it in a private network with in an hospital or simply in a computer. In order to fulfill the requirements of an open source application, all the sources and required documentation have been uploaded to a GitHub server in which anybody can contribute to improve the system(https://github.com/Adrizcorp/K-MAMM-CAD).

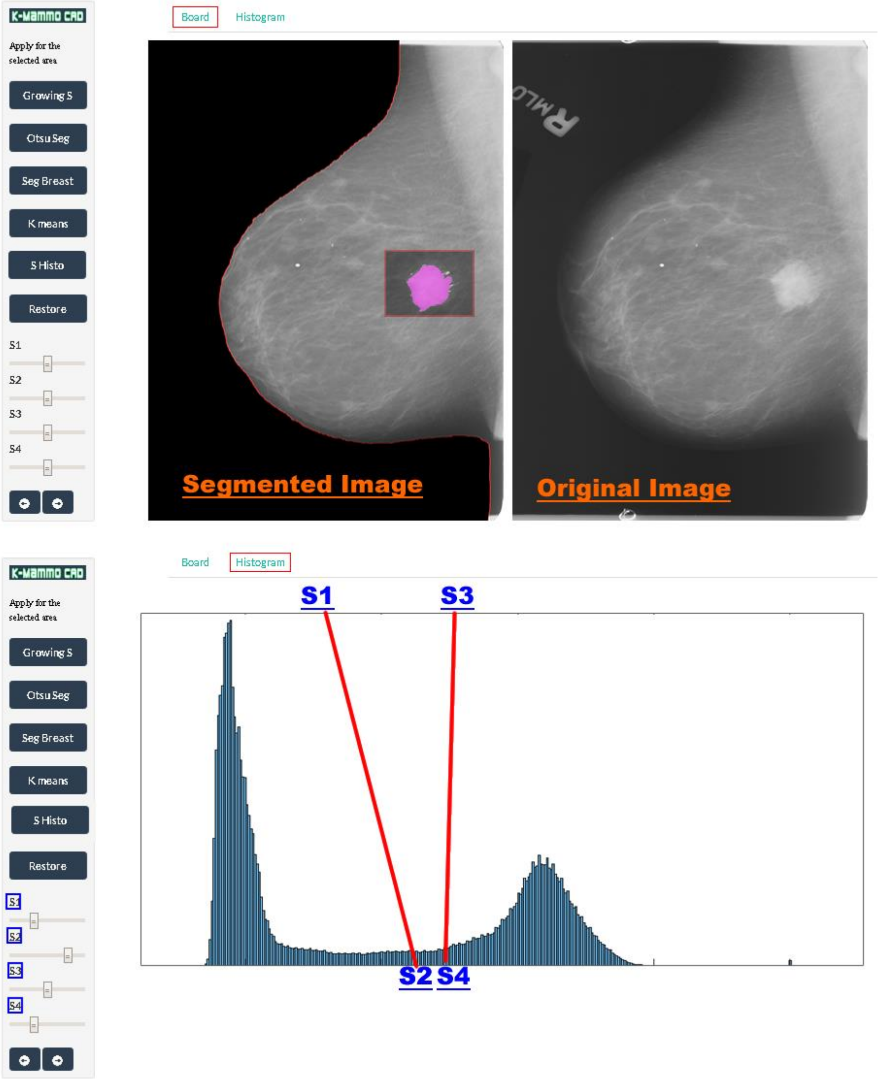


Figure 7 - Software User Interface, Front End

The Figure 7 shows the graphical user interface, in which there is a selection bar with 4 types of segmentation process among you can find, Seed growing segmentation, Otsu Segmentation using thresholding, Brest segmentation in base of Otsu method and sobel edge detection, thresholding segmentation using the histogram of an specific region or the whole image, and K-Means segmentation. Nonetheless, the software can be modified and added more types or techniques for segmentation or image processing purposes.

2. Techniques and Results

As it was mentioned in the item 1.3 the software is capable of implementing different algorithms in which can be found techniques such as the Otsu Method to determine the threshold level among 2 different kinds of objects (Background and the Breast). The sobel edge detection method to establish the boundaries of the segmented elements in an image, which in this case it is the breast, and the nodule. Digital operation such as OR, XOR, and AND between two images are intrinsic within the segmentation processes to segment thereby masks the region of interest and the breast area. Seed Growing Region to segment the nodules within the image. Furthermore, there are methods such average filtering to reduce the noise of the segmented images (the masks, not the original image) in order to get better results at the moment of avoiding artifacts because the DDSM data base is characterized for having lots of artifacts within the images such as labels (Figure 8). Moreover, there are implemented algorithms to extract the arithmetic mean from group data, average, maximum, and minimum contrast of the region of interest or image to normalize the LJPEG images. Several algorithms were coded for the purpose of having a rich number of tools to extract different features. For this project, all the algorithms of the segmentation and visualization process were written from scratch in C/C++.

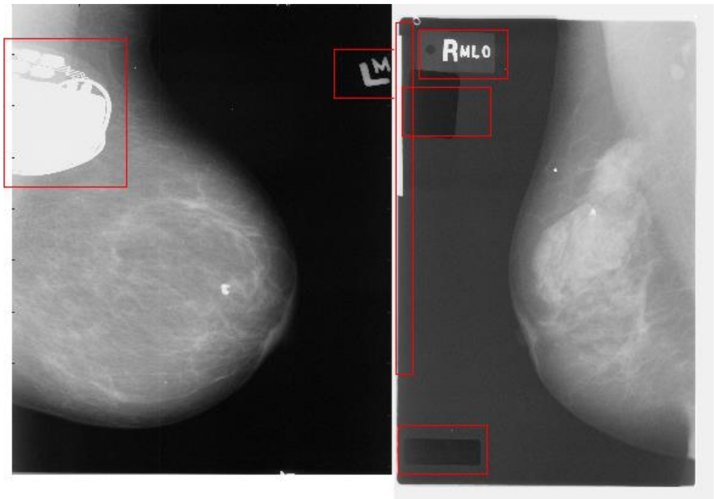


Figure 8 - Image with artifacts

2.1. LJPEG visualization and decompression.

All the images are compressed in the Lossless JPEG format, on the webserver of the University of South Florida, there was an algorithm written originally in C code to convert the images to a raw file, which after could be only shown by Matlab or Binaries that were written in the year 1999, which makes the visualization of images hard for the user, and platform dependant , therefore the architecture given in the item 1.3 was proposed to avoid dependency.

In this process, the code written in C for converting the LJPEG images was recompile, and then another code was written to convert this raw in a BMP format for the visualization purposes. However, other issues in the images were found, some of the images showed to be saturated, even Matlab did not show those images correctly, and further more the description that contains the resolution of the image sometimes has the width and height swapped, which makes even harder for the user creating misunderstanding in the interpretation of the images. That was one of the mentioned errors shown on the webpage, some of the images had inverted the resolution, and not only that, the images were taken by an analog machine, and then converted to digital, which makes the images don't have an specific range or standard contrast. The only known feature it is that the images come in chunks of 16 bits which 12 bits are used for color depth, which makes even harder to be certain about the uniform distribution of contrast among all the images.

A normalization process was required after converting the LJPEG to raw files to avoid saturation. The maximum value, the minimum value, the average value, and arithmetic mean from group data were extracted from the raw file before converting this file to a viewable BMP file. The minimum value and the maximum value were important to know the scale of the image, the average to know the trend among of the values of image, and the arithmetic mean from group data (Equation 1) due to the number of pixels in the image, it is better to have the value in proportion of the frequency (through the histogram) of each data to get a more accurate mean. After extracting these features, each pixel was divided by the maximum value minus the average and the arithmetic mean as can be seen in the Equation 2, this method has shown to have better results than just divide the value by the maximum value of 4096(12 bits), as the given example on the South Florida algorithms.

Equation 1 Mean from Grouped Data

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \ldots + x_n f_n}{N}$$

Equation 2

$$\begin{array}{l} if\ the\ range\ log2(maximun\ value) < 16\ dividedby = maximun\ value \\ else\ dividedby = maximunvalue - (average + arithmetic\ mean) \end{array}$$

The Figure 9 shows the result obtained after normalizing the image, and it can be observed the results of this normalization brings a better visualization of the image seen by the user.

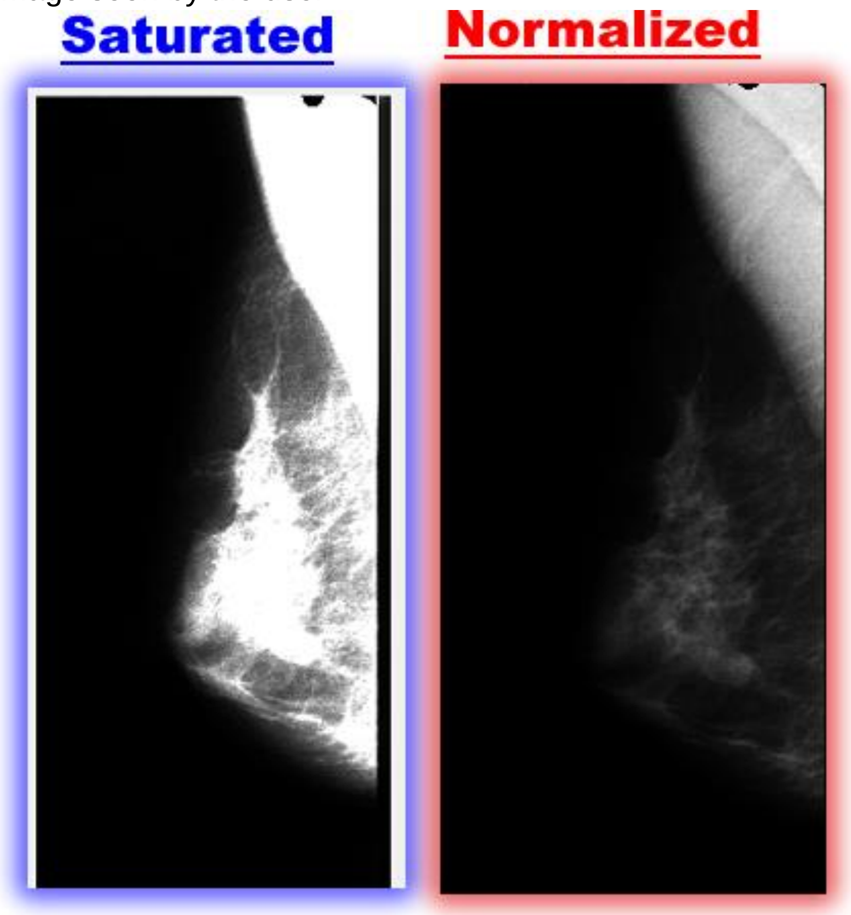


Figure 9 - Saturated Image Vs Normalized Image

2.2. Breast Segmentation

Segmenting the breast brings as benefit to the radiologist and the software, the feature of reduce artifacts and highlight the breast for further analysis such as determining the breast density which is important because depending on the type of breast it could be hard to distinguish a nodule due to the level of density [8] (Thought algorithm for further stages of this software) (Figure 10). The higher the level of density of the breast the harder to get better results. To reduce the amount of information to process the image it is really important due to the size of the image.

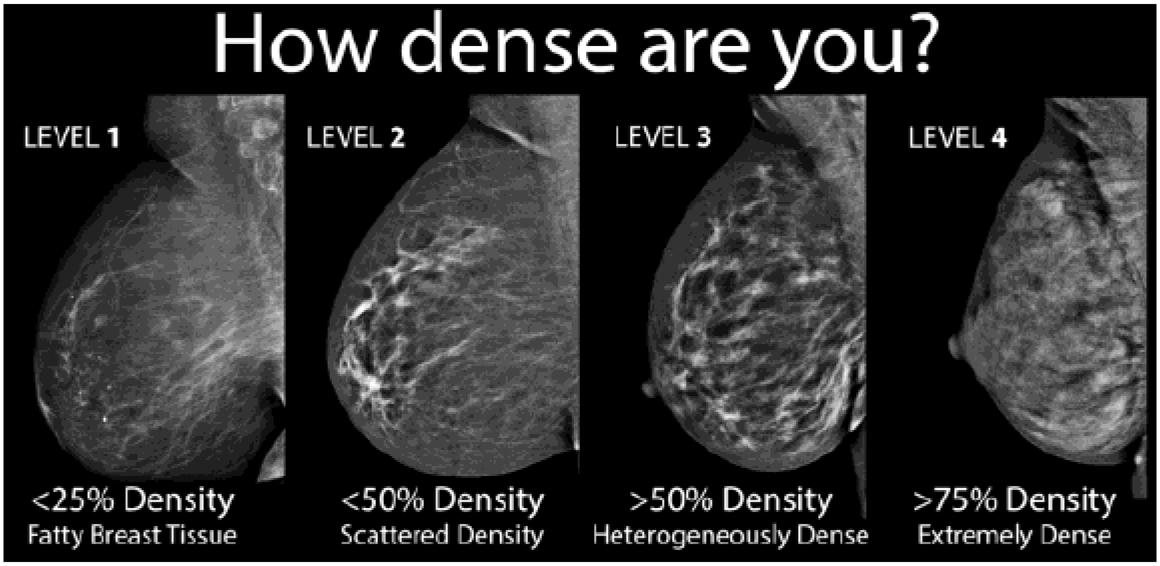


Figure 10 - Density of the breast

For segmenting the breast the image passes through the following stages:

2.2.1. Otsu Method

The Otsu method is the exhaustivity search for a threshold that minimizes the variance within an image. In the case of the breast, it has a “Black” Background and the breast area, the objective by using the Otsu method it is to determine a single Threshold (Figure 11) to then filter the image by only letting pass the pixels that are equal or greater than the found threshold.

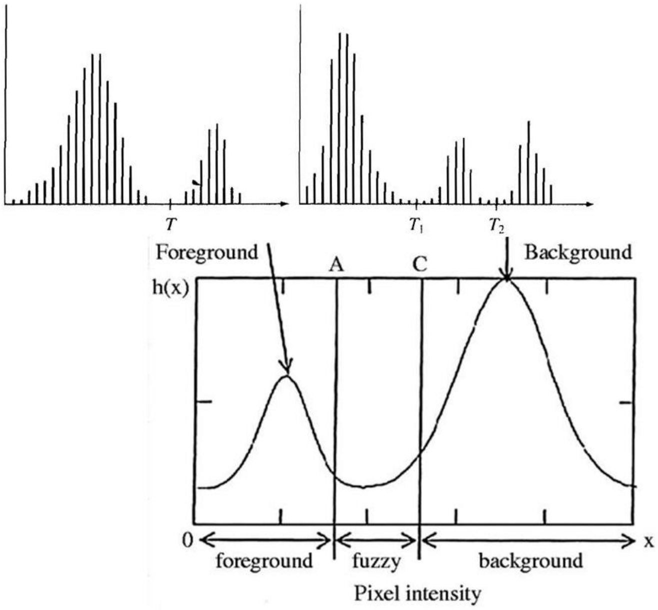


Figure 11 - Instensity Histograms that can be partitioned by single threshold, and by dual thresholds

The Otsu algorithm to find the threshold that could separate the background of the breast it is the following.

2.2.1.1.1. Compute the histogram of the image.
2.2.1.1.2. Calculate the probability for each intensity level.
2.2.1.1.3. Initialize Omega and Myu.
2.2.1.1.4. Find the optimal threshold value through the intra-class variance.
2.2.1.1.5. Binarization of the image using the Optimal Threshold.

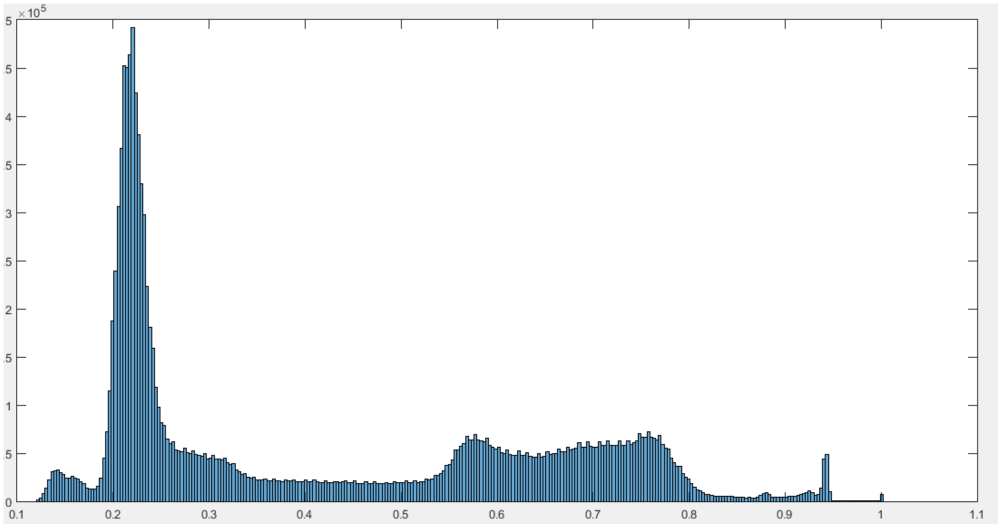


Figure 12 - Histogram of the Image C_0004_1.RIGHT_MLO DDSM Database

2.2.2. Average Filter:

After the binarization of the image, the image passes through a filter which it is a simply average filter, that takes horizontal and vertical chunks of data, and depending it removes the artifacts of the image by thresholding the incoming image through an specific threshold, in the case of the image, the argument of this filter could be adjusted to the needs of the user, however for general purposes and due to the good results the size of chunks was set to 100 samples per line. This filter is slide horizontally and Vertically through the image to removes the artifacts such a labels or not desired borders (Figure 8).

$$SMA = \frac{p_M + p_{M-1} + \cdots + p_{M-(n-1)}}{n}$$
$$= \frac{1}{n}\sum_{i=0}^{n-1} p_{M-i}$$

Figure 13 - Moving Average Filter

The comparison between the applied Otsu method and the improvement by the filter can be seen in the following figure.

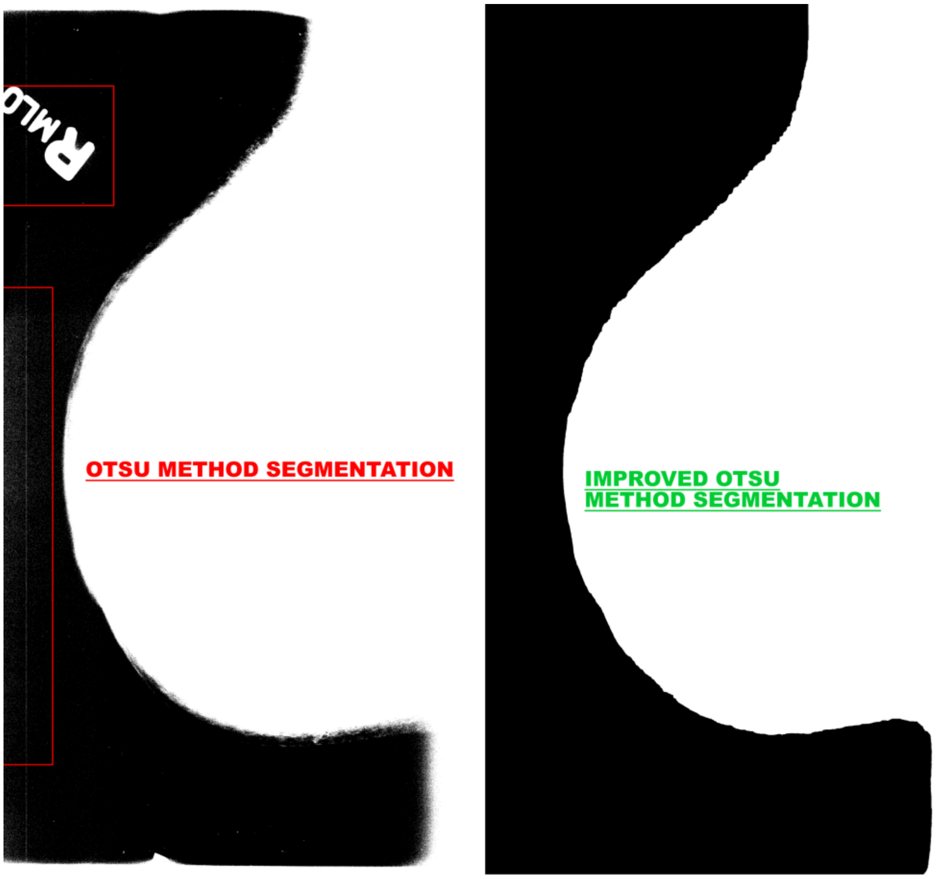


Figure 14 - Otsu Method segmentation, the noise and artifacts are shown in the left image, and the right represent s the filtered image.

2.2.3. Edge detection

Once the segmentation process is performed, the next stage it is to border the breast with a line that marks the boundary between the breast and the background, for that an edge detection algorithm, using the sobel form, was performed after the averaging stage. Consecutively, the obtained image passes through an average filter of chunks of 2 elements, to highlight more the edges, obtaining the following results.

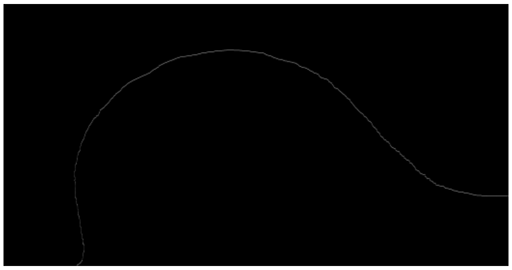
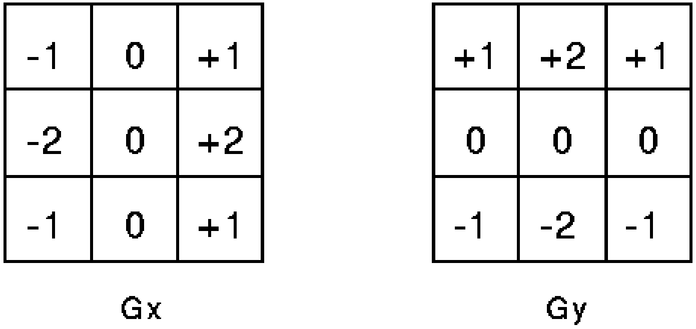


Figure 15 Sobel edge Detection

The Canny algorithm was implemented as well find the edges of the image, however it is computational too heavy (Gaussian Filter, Convolution, non-maximum suppression, Hysteresis) for the only purpose of finding edges on the image, and the image it is too big, therefore, this algorithm was applied for this first version of the CAD tool, it is not implemented.

The sobel algorithm seeks the abrupt changes among the changes of each image kernel pixels (derivative) that can be expressed as the following:

2.2.3.1. Slide over the image, multiplying each kernel by the following mask to obtain the Gx, and Gy, in term of intensity.



-1	0	+1
-2	0	+2
-1	0	+1

Gx

+1	+2	+1
0	0	0
-1	-2	-1

Gy

Figure 16 - Sobel Operator

2.2.3.2. Calculate the gradient Magnitude

$$\mathbf{G} = \sqrt{{\mathbf{G}_x}^2 + {\mathbf{G}_y}^2}$$

Figure 17 Sobel Gradient Magnitude

2.2.3.3. Normalize the magnitude, dividing this magnitude by a constant.

2.2.4. Digital operations

After the previous process, a simple and logic operation is performed between the original image and the obtained masks by the Otsu method, and the edge detection method. As it is illustrated in the Figure 18,

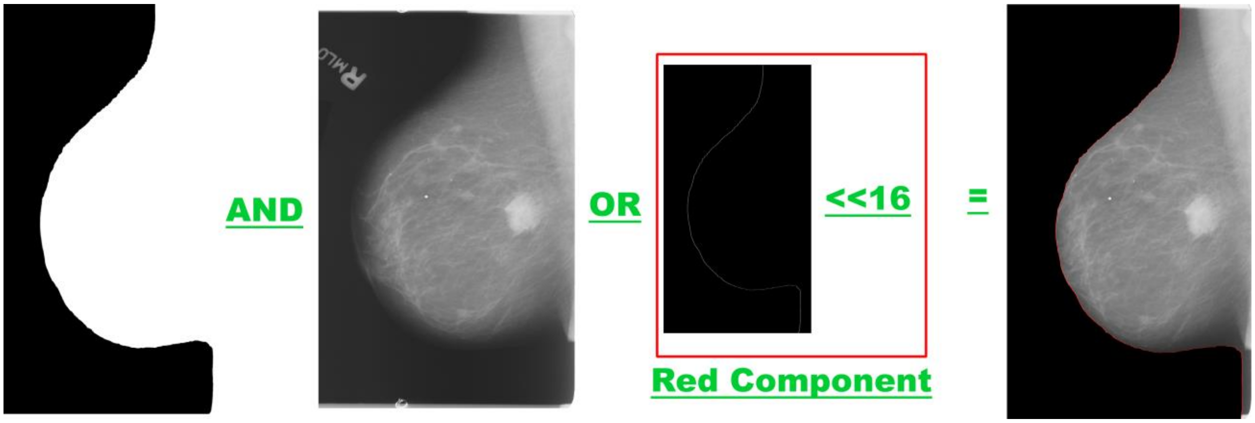


Figure 18 Digital operation to obtain the final breast Segmentation

2.3. Mass Segmentation.

To segment the specific mass, the user must choose by painting a square on the image the area of interest, once the user has completed this selection, a Region Growing Algorithm is performed to extract the possible mass within the small portion of the image (Figure 19).

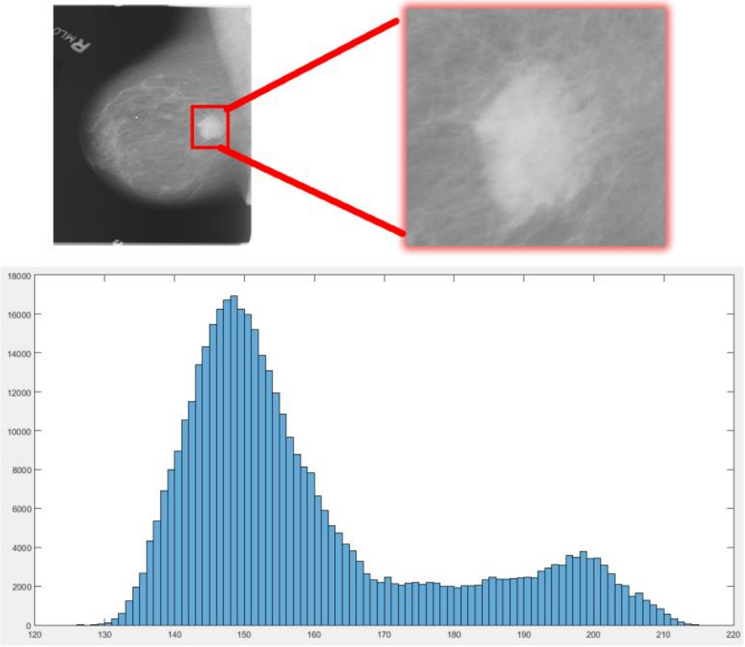


Figure 19 - Selection of the specific region

2.3.1. Region Growing Algorithm.

This algorithm examines an initial kernel of pixel which is called the seed points, this seed is compared with its neighbors (pixels around) to determine whether or not the pixel belongs to the region of interest. This algorithm it is an intensive and iterative method to extract similar parts of an image. This region grows until this is block by the stop criteria, which in this case it is the difference between the outside(contour) pixel’s intensity and the region’s mean, then the minimum difference of the mean and the contour pixel is compared with a threshold, and if this difference is greater than a given thresholds, the algorithms stops. Otherwise the pixels are added to a mask (other image with the same size area of the original initialized with 0s). This algorithm could be seen as a snake going through the contour (the shadow pixels or the minimum value of the found differences) of the mask to determine whether or not the pixel belongs to the region of interest. The type of connectivity for this algorithm is 4-connected pixels, well-known as the Von Neumann Neighborhood [10] and can be seen in the Figure 20, however, any other configuration of kernels could be used, as the Moore neighborhood but it must be taking into account that the computational cost is increased.

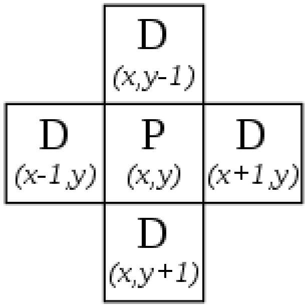


Figure 20 - Von Neumann Nighborhood

One of the advantages of this algorithm, it is that multiple criteria could be evaluated at the same time to generate different clusters, with a few number of seed the algorithm can evaluate and segment a region of interest. Notwithstanding, this algorithm is sensitive to noise, and computationally expensive (in future stages this algorithm should be accelerated).

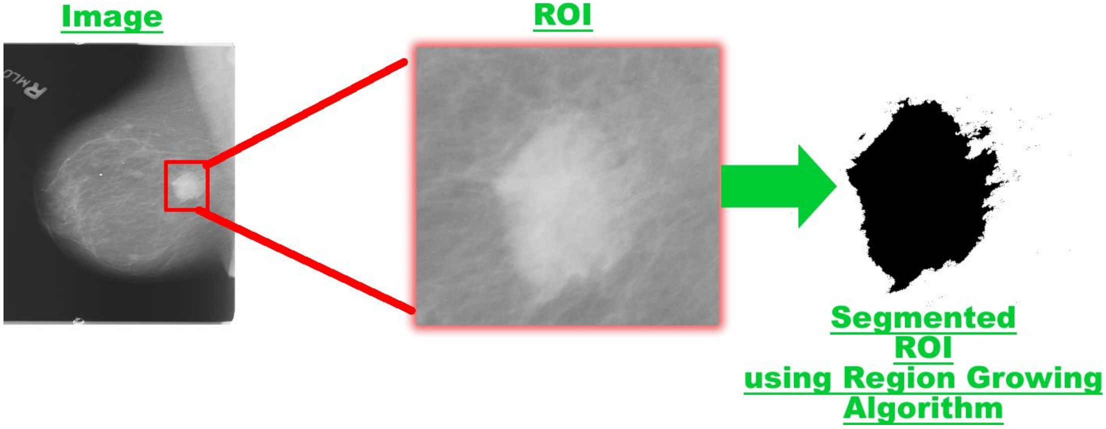


Figure 21 - Results of the segmentation using Region Growing Algorithm, using the image B_3084_1.RIGHT_MLO of the DDSM

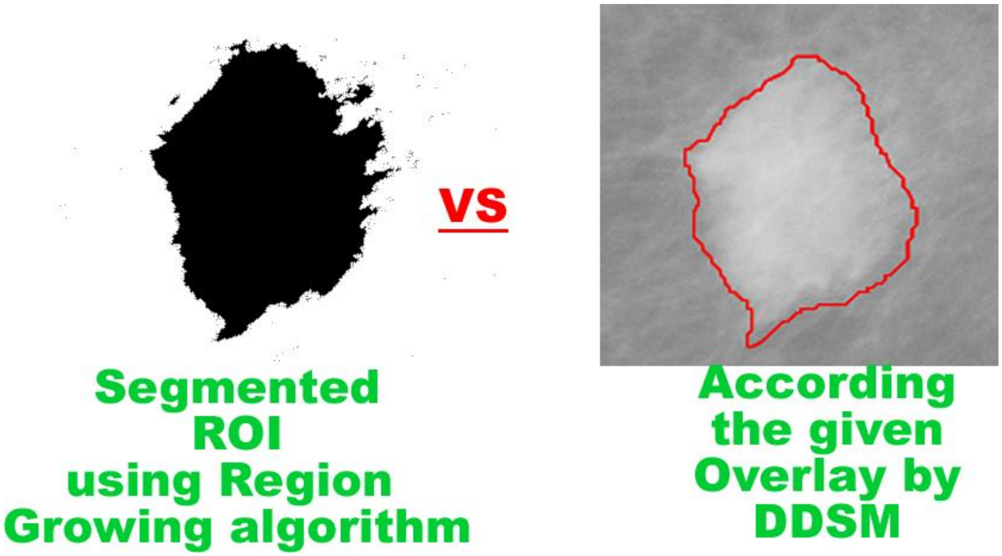

VS
Segmented ROI using Region Growing algorithm
According the given Overlay by DDSM

Figure 22 - Comparison between Region Growing and the Given overlay by the Database for the Image, B_3084_1.RIGHT_MLO.

As it can be seen in the Figure 21 and Figure 22, the obtained results are not perfect, and have noise in comparison with the overlay, however, this could be solved by applying a filter as mentioned in 2.2.2. Nevertheless, in the Figure 23, the shape has been affected in the left side of the nodule.

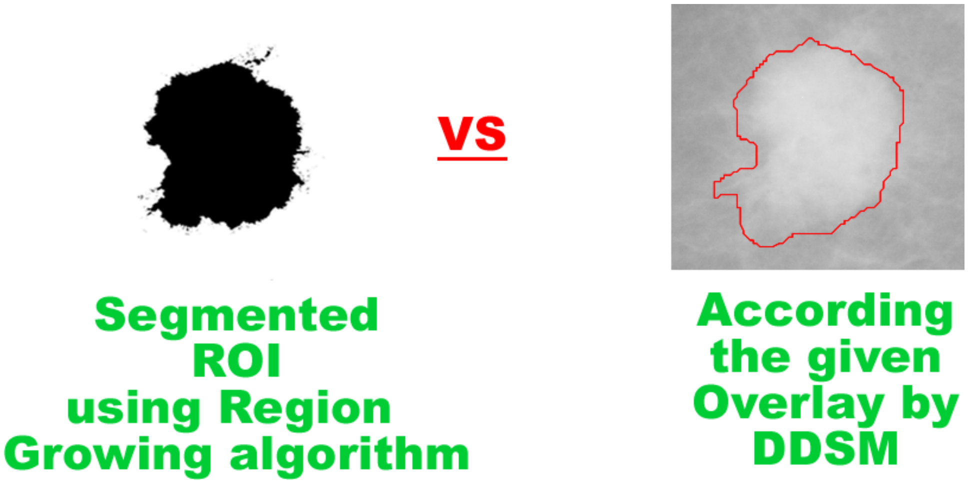

VS
Segmented ROI using Region Growing algorithm
According the given Overlay by DDSM

Figure 23 Comparison between Region Growing and the Given overlay by the Database for the Image, B_3084_1.RIGHT_CC

3. Conclusions:

As a first conclusion, in base of the obtained results, is that there is a considerably number of other methods that could be applied, and wanted to be applied in the project, but for the time it was not possible, such as the K-Means Segmentation, masses classification extracting morphological features that can help the radiologist determine the type of mass (benign, malign), these features could be extracted in base of these first results using mathematical models, Autoencoder or Deep Neural networks. As a first version of a CAD tool, it can be said that the obtained results are satisfying, it is a start that can springboard a new generation of Open-Source CAD tools with the main objective of assisting radiologist in identifying the different kinds of masses and its shapes. The project still has a lot work to do, and the help of lot of people that might be interested in continue this work, as I am. All the implemented methods show that the variation of the threshold, a good selection and interpretation of the intensities, it is crucial to perform a good segmentation, and get better results. The constant research in this area and development could take this CAD tool to get better results in the future.

4. Bibliography:

[1]"SEGMENTATION OF BREAST CANCER MASS IN MAMMOGRAMS AND DETECTION USING MAGNETIC RESONANCE IMAGING". [Online]. Available: http://www3.ntu.edu.sg/eee/urop/Congress2003/.../yao%20yao.pdf. [Accessed: 06- Oct- 2016].

[2] J. Block, "Digital Mammography Equipment Price/Cost Info [2016 Update]", *Info.blockimaging.com*, 2016. [Online]. Available: https://info.blockimaging.com/bid/95356/digital-mammography-equipment-price-cost-info. [Accessed: 06- Oct- 2016].

[3] "CAD for Mammography: Importance of Computer Aided Detection (CAD) In Treatment", *Radiology-info.org*, 2016. [Online]. Available: http://www.radiology-info.org/computer-aided-detection.html. [Accessed: 06- Oct- 2016].

[4]"Construcción de una base de datos de imágenes de mamografía para la identificación de microcalcificaciones", *Repositorio.utp.edu.co*, 2016. [Online]. Available: http://repositorio.utp.edu.co/dspace/bitstream/handle/11059/4236/621367S232.pdf?sequence=1. [Accessed: 06- Oct- 2016].

[5]"http://www.sersc.org/journals/IJSIP/vol6_no1/2.pdf", 2013. [Online]. Available: http://www.sersc.org/journals/IJSIP/vol6_no1/2.pdf. [Accessed: 06- Oct- 2016].

[6]"Mammogram Image Features Extraction and Classification for Breast Cancer Detection.", *International Research Journal of Engineering and Technology (IRJET)*, vol. 02, no. 07, 2015.

[7] S. Halls, "BI-RADS category scale 2 3 4 5 score", *Breast Cancer - Moose and Doc*, 2016. [Online]. Available: http://breast-cancer.ca/bi-rads/. [Accessed: 07- Nov- 2016].

[8] Hogg, Peter, Judith Kelly, and Claire Mercer. *Digital Mammography: A Holistic Approach*. 1st ed. Cham: Springer, 2015. Print.

[9] Quesson B, Sabel J, Boumna H, Dekker R, Lengrand-Lambert J. Automatic target recognition in synthetic aperture sonar images for autonomus mine hunting. TNO Defensive en Veiligheid, 10th European Conference on Underwater Acoustics. 2010;.

[10] *Wilson, Joseph N.; Ritter, Gerhard X. (2000), Handbook of Computer Vision Algorithms in Image Algebra (2nd ed.), CRC Press, p. 177, ISBN 9781420042382*.