Advanced Machine Learning
Project
Breast Density Assesment and Breast Cancer Detection

Abbonato Diletta, Pezone Francesco, Testa Lucia

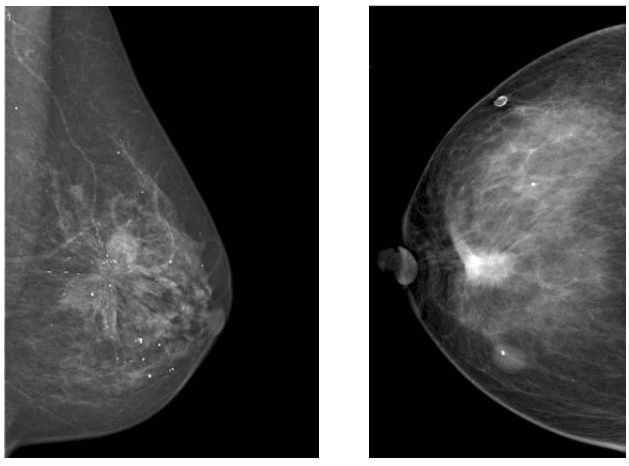
7 June 2020

1 Abstract

The main goal of this project is to make a classification through breast clinical images. In this retrospective study , a Convolutional Naural Network finetuned models were trained to asses a diagnosys based on the original interpretation by experienced radiologist. The digital screening mammograms come from the open DDSM Stanford Dataset and a private dataset given by Policlinico Umberto I just for this study. The tools used in this study has been Google Colaboratory and open python library as Pytorch, PIL, OpenCV, NumPy, Pandas and Matplotlib.

2 Introduction

The two main purposes of our work are to develop a deep learning that can accurately classify breast images according to two differents scenario: firstly throught the ue of the American College of Radiology (ACR) density index which describes four categories of breast parenchymal density, then, secondly, to classify breast images according to their diagnosys of being Malignant, Benign or Normal.



(a) MLO (b) CC

Figure 1: Representation of the Mediolateral Oblique(a) and Craniocaudal positions (b)

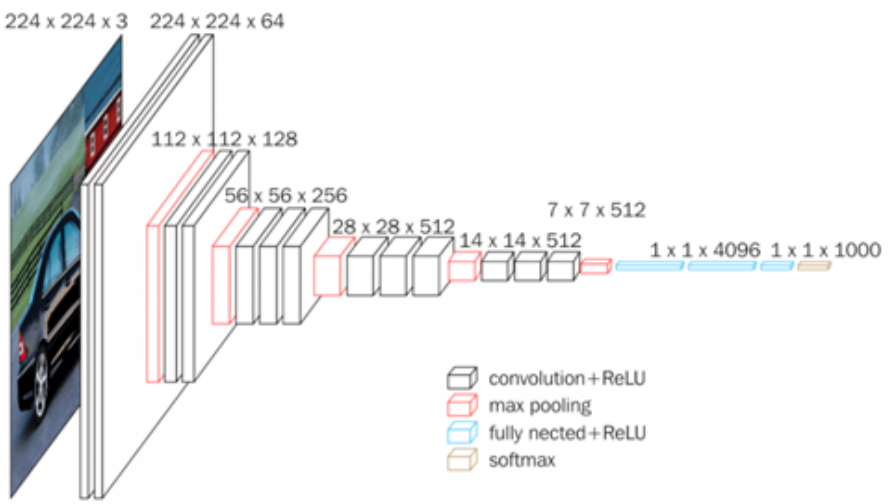
3 Related Work

In the literature the detection of cancer on mammograms has already been covered. For this reason it is possible to find multiple works on this subject. In particular, one of the many works that inspired us was 'Deep Learning to Improve Breast Cancer Detection on Screening Mammography'.
https://www.nature.com/articles/s41598-019-48995-4

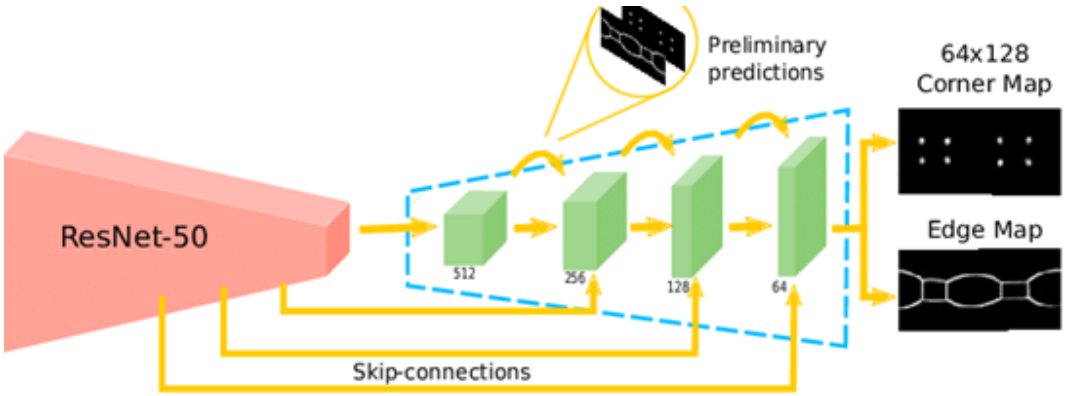
4 Proposed method explained

In choosing the pre-trained models we have considered to use the VGG16 and ResNet50 ensuring that their behavior for breast cancer classification was excellent as a support for our typology of data.

VGG was born out of the need to reduce the number of parameters in the CONV layers and improve on training time. The architecture is shown in the figure below.



ResNet addresses his network by introducing two types of **shortcut connections**: *Identity shortcut* and *Projection shortcut*. ResNet architecture makes use of shortcut connections to solve the vanishing gradient problem. And infact The basic building block of ResNet is a Residual block that is repeated throughout the network as is shown in the figure below.



4.1 ACR density assesment

The ACR index (A-B-C-D) is used to classify breast images in order to their density. We can recognize four categories.

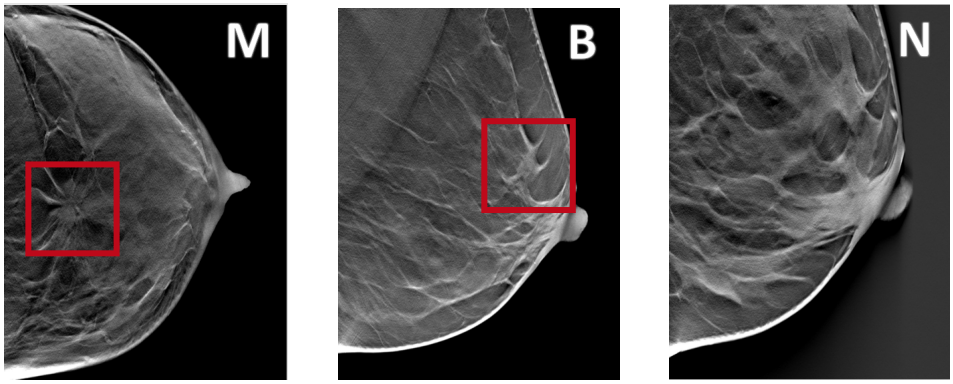
- a. The breasts are almost entirely fatty
- b. There are scattered areas of fibroglandular density
- c. The breasts are heterogeneously dense, which may obscure small masses
- d. The breasts are extremely dense, which is the sensitivity of mammography

We dropped the A from our analysis since we had just one case of it. After splitting the dataset into train, validation and set, we moved on to the training section on the models already pretrained on the ResNet50 and VGG-16 ImageNet. The amount of images available was not much, but the hope of getting good results was due to the fact that the images are easy to distinguish with the naked eye. (imagine con C,B,D) In fact,

graphically speaking, recognizing the amount of fibroglandular tissue in a breast is possible without having a trained medical eye if the image is of good quality.

4.2 Digital breast Tomosynthesis

The Digital Breast Tomosynthesis is a type of radiological analysis that consists of operating a radial mammographic screening, then going to "cut" the breast into "slices", which allow to analyze the breast in more depth. When the breast is very dense, the DBT becomes a significant added value as it goes to "remove" the cover constituted by the fibroglandular tissue, ensuring a better vision of the breast. The images in our study came from a further study carried out by radiologists to demonstrate how much and why tomosynthesis can represent a diagnostic improvement when dealing with dense breasts. Also in this case, a division in train, test and validation was carried out in order to better continue the training phase in the network. Out goal in this phase was to classify DBT images according to their state of being malignat, benign or normal.



(a) A malignant example (b) A benign example (c) A normal example

Figure 2: An example of DBT for a Malignant, a Benign and a Normal case

The models chosen were also in this case the ResNet50 and the VGG-16.

4.3 Digital Mammography

To classify the type of breast cancer, we decided to use two different datasets.
The first is that of the Policlinico Umberto I which however contains a limited sample of mammograms so to have a greater number of images, with which to do transfer learning from ImageNet, we decided to download the Stanford dataset, a dataset with 2620 immages.
The idea was to make a first transfer learning from ImageNet to Stanford and then another from Stanford to our dataset. In trying to do this, we ran into some problems with the Stanford dataset as it had numerous writings on the images that we realized were causing a lot of noise. For this reason we decided to filter the images before removing the writings.

We used the same approach of DBT for the digital Mammography but beyond this we implemented an image preprocessing on the Stanford dataset in order to improve the learning process.

5 Experimental results

5.1 ACR density assesment

The first training in image density classification was without data augmentation on color. The reason for this choice was to monitor the network's reaction to the few images we had. The final result at the first training, with 20, 30, 50 and 100 epochs was quite the same, a best model with 75% validation accuracy. Recognizing the evidence of the sections of fibrous tissue in the image we decided to operate a rather specific ColorJitter, operating several times on the image contrast. Moving from 0.5 to 1 (Figure 3).

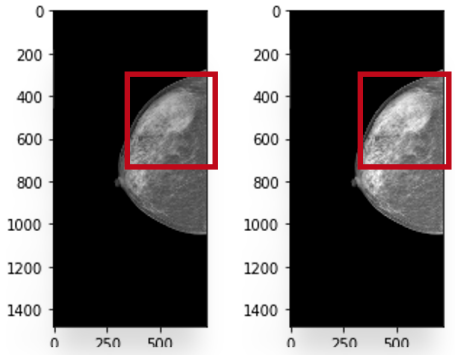
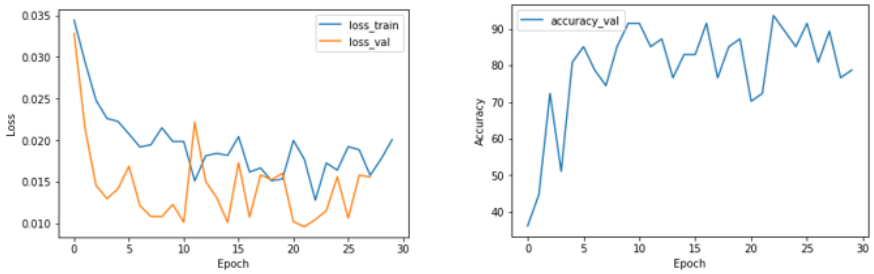


Figure 3: Example of contrast data agumentation on type C breast

Thanks to the changing in contrast, it was possible to highlight the section of fibroglandular tissue inside the breast, thus increasing the network's ability to recognize the affected area and correctly classify the image.

The results obtained were quite satisfactory. The best model in the validation phase reached 90% validation accuracy and 75% test accuracy.

The training on VGG-16 was performed in the same way as ResNet50, but the final trend of validation loss and validation accuracy showed a not inconsiderable difficulty in learning. This can certainly be due to the type of network structure and the small amount of images that have been submitted to it. In any case, the best model of the VGG has produced a test accuracy of 66%. Nevertheless, looking at the final plots, we cannot say that this is a reliable result, which we could say about the ResNet.



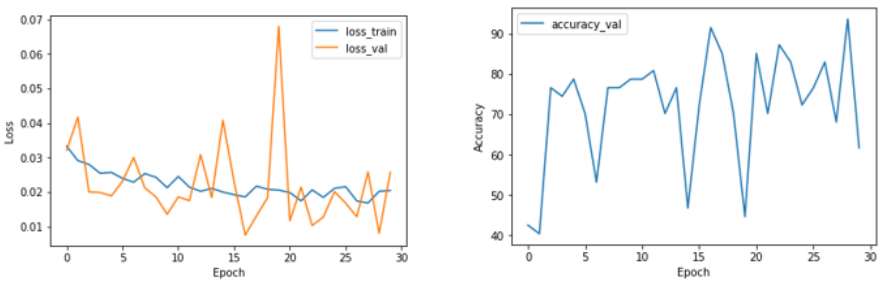
(a) Train and validation loss

(b) Validation accuracy

Figure 4: Resnet results

5.2 Digital breast Tomosynthesis

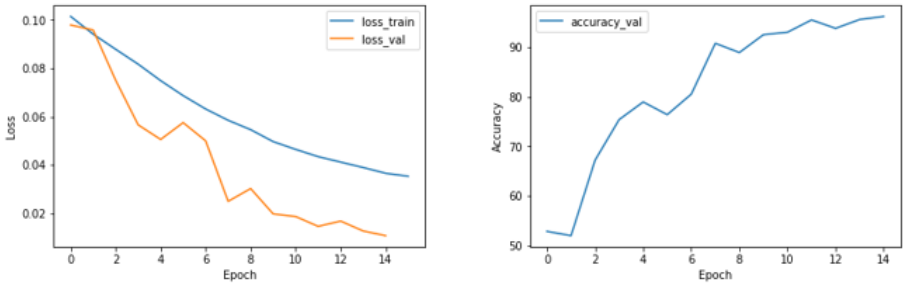
A training of 15 epochs was operated for both models, which took several days of training as the images were more than 6000, since in the DBT generates between 30 and 50 images per breast. The first training on VGG-16 produced a rather fluctuating trend, although more stable than the reaction with mammograms for ACR. Net of the fifteen epochs we obtained a best model with a validation accuracy of 64%. On the other hand the ResNet50, showed a more stable, almost monotonous increasing trend. With fifteen epochs we achieved a validation accuracy of 90%. Both models were then tested on the same set, producing results that apparently reflected the learning process outcomes. An 86% on ResNet50 and 60% on VGG-16. However, analyzing the results, we found that the VGG-16 did not perform as badly as ResNet. In fact, VGG-16 was able to correctly classify in 100% of malignant cases, while ResNet50 was only 80%. The low accuracy of the VGG lies in its



(a) Train and validation loss (b) Validation accuracy

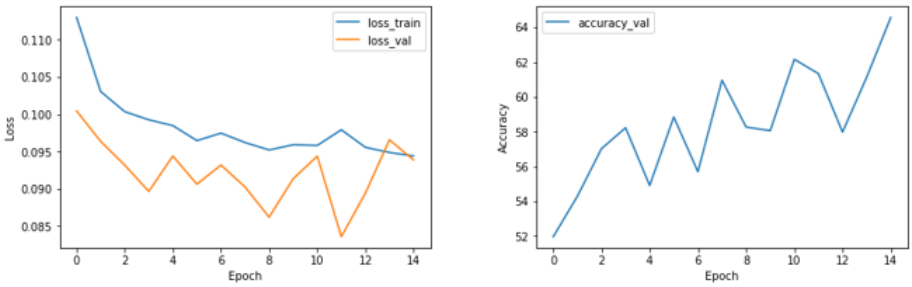
Figure 5: Vgg results

difficulty in distinguishing benign from normal cases. So we can say that with VGG we have partially achieved our goal, because when making medical diagnoses, the importance lies in correctly recognizing the really positive cases.



(a) Train and validation loss (b) Validation accuracy

Figure 6: Resnet results

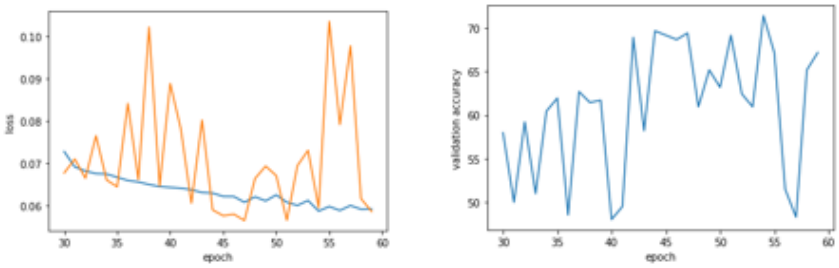


(a) Train and validation loss (b) Validation accuracy

Figure 7: Vgg results

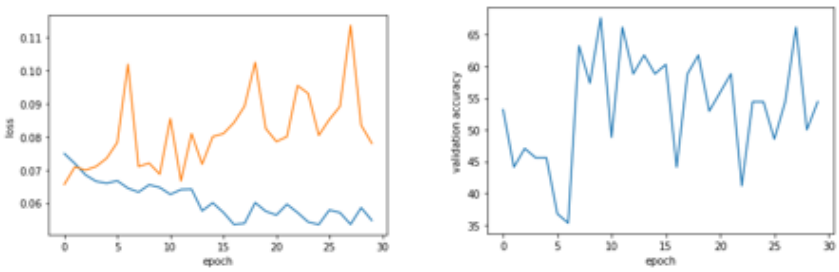
5.3 Digital Mammography

Even trying to apply all the different ways showed in Section 6.1 to modify our breast images of DDSM, the network showed the worst behaviour encountered. There's not apparence of a true learning in the plots, but just a random chance to guess the right class. Even if the Mammography is the most common type of exam that we know is also the most difficult to understand. The main problem during the training of DDSM in not only the text on the images but also the quality of mammogram itself, that does not allow a good detection of mass.



(a) Train and validation loss (b) Validation accuracy

Figure 8: Resnet results



(a) Train and validation loss (b) Validation accuracy

Figure 9: Vgg results

6 Dataset & Benchmark

We collected a set of 214 cases on digital mammography images from Polyclinic Umberto I. The digital mammography are of both breasts in the mediolateral oblique and craniocaudal position. For our analysis we decided to make the following divions:

To screen mammograms: we divided the cases into different folders, reating 3 signs for benign, normal and maglignant cases. In particular for malignant cases we identified 37 cases.

To detect the mass in the digital breast tomosynthesis: we repeat the same step to divide our 214 cases in the four ACR category in order to classify the same images for their density.

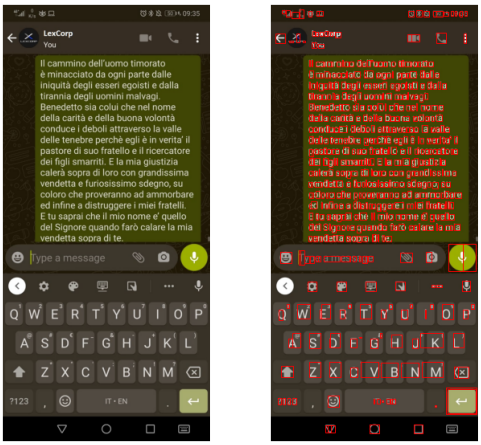
In both cases we used 200 cases of train and validation with the use of the rule regarding 85 -15 division and the remaining instead was used for the test. Since the quantity of mammograms are very low we used the Digital Database for Screening Mammography (DDSM) as support.
The DDSM is a database of 2620 scanned film mammography studies. It contains benign and malignant cases with verified pathology information.
To obtain the data we used the NBIA Data Retriever software.
The data was in DICOM format that we converted to jpg. We divided the data into 1548 for the train and validation using als in this case the 85-15 rule while the rest, that was an amount of 1328, was used for the test.
Now let's see how we filtered the images.

6.1 Filter the images

The huge problem with the Stanford dataset was the presence of lot of writings inside the images. In this paragraph we will explain what the approaches and problems were in removing the writing and what was the final result.

6.1.1 Text Detection

The first technique was to identify the text within the images using rectangles and then try to remove it.
To do this we used the Pytesseract library. First of all we tested teh library on some simple text and it seemed to be going well, even if it identified some symbols as text, as we can see in Figure 10.



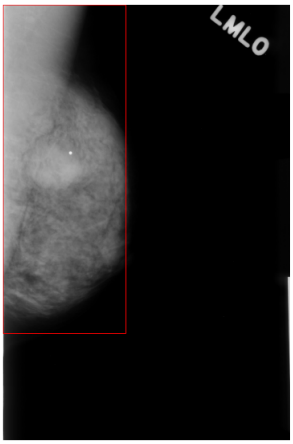
(a) A Whatsapp message and the keyboard

(b) How pytesseract detect the text

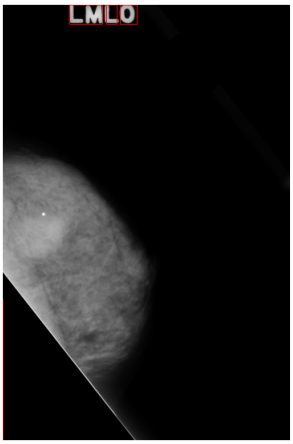
Figure 10: An example of how pytesseract is able to detect the characters in a text

After this test we started working on the images but we immediately noticed that the detection was not working well at all. In fact there weresome problems if the text was not perfectly horizontal. Is possible see that in Figure 11 where only in image b the text is correctly identified.

After this discovery we went back to testing pytesseract on inclined texts and we realized that if the text exceeded 9 degrees it was no longer correctly identified as we can see in Figure 12.

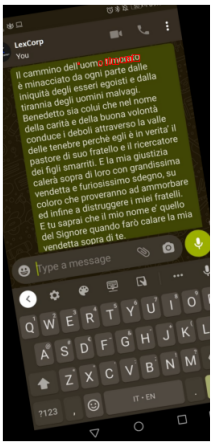


(a) The results when we try to process an image with pytesseract

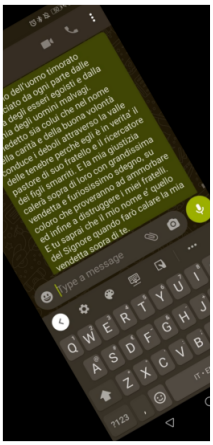


(b) What happens if the text is orizontal

Figure 11: An example of how pytesseract work on the Stanford dataset



(a) Text inclined at 9 degrees



(b) Text inclined at 30 degrees

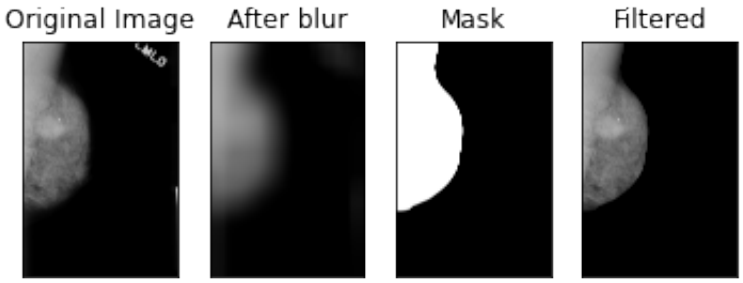
Figure 12: An example of how pytesseract work with angled texts

6.1.2 Filtering the images

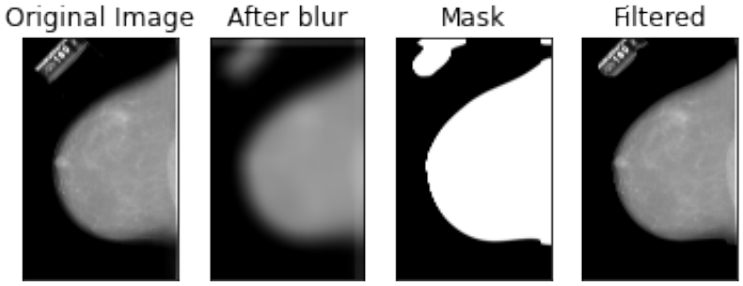
For the second approach, we noticed that the writings are generally isolated with all black around them.
The idea was therefore to blur the image with a fairly large kernel (more or less a quarter of the size of the base), in this way the tone of the writing area became very low. At this point we turned off all pixels with a value lower than 50 (for larger values it became all black, for smaller values the writings were not removed) generating the mask and then applying it on the original image.
We can see some examples in Figure 13: in the case (a) the filter has worked very well, more or less 98% of the cases are of this type, in case (b) the problem is that the writings are too large and white so the kernel cannot fade them very well, in the last case the breast is too little dense and therefore the pexels are below the threshold of 50 and are turned off.

6.1.3 Connected components

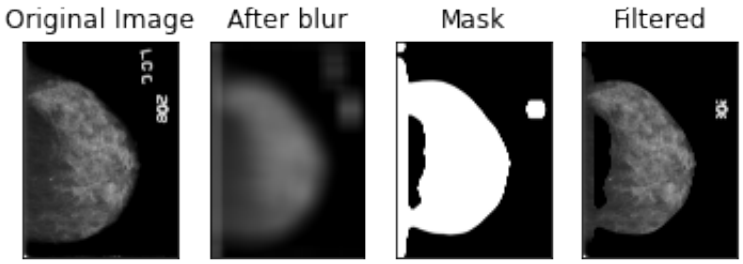
In the end we decided to proceed with the calculation of the connected components. The procedure is shown in Figure 14 and refers to the case (b) in Figure 13.
We read the image and resized it to 224x224 afterwards we subtracted 1 from all the pexels to have a white background (pixel value of 255). Subsequently we have zeroed all the pixels that were worth 255 and the edges (too avoid some problems), while to 255 all the others.
Afterwards, thanks to OpenCV, we calculated all the components present in the image, which however turned



(a) This is how this filter work for the majority of the images



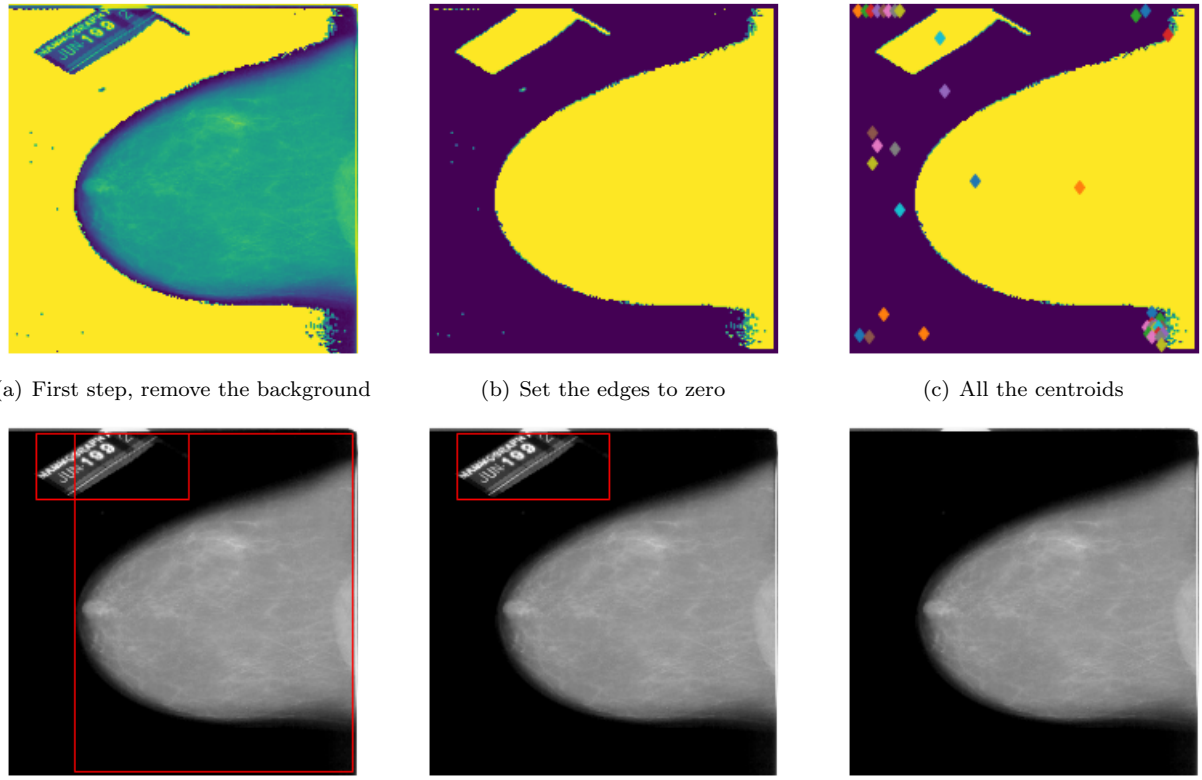
(b) This is what happens if the writing is too big



(c) This is the worst case scenario, when the breast is not very dense

Figure 13: How the new filter work on the Stanford dataset

out to be really many and with very different dimensions, as we can see in the images (c) and (d). TO avoid other problems we set a range from 10 to 100 pixels for height and width. We can see the box in (e) and the final filtered images in (f).



(a) First step, remove the background

(b) Set the edges to zero

(c) All the centroids

(d) All the boxes with at least an edge grather than 10 pixels

(e) Pick only the right connected coomponent

(f) Filtered image

Figure 14: The results of the new approach on the Stanford dataset

7 Conclusion and Future work

The main problem for our cancer detection analisys is that a mass in the breast is a 'bigger' opacity intra opacities, so the detection could be problematic when there is to much noise.
Even if the network itself could be considered as reacting well to the noise here we can say that the quality of images itself hindered to much the learning process.
Beyond this you may ask why the transfer from DDSM to Policlinico went so bad. Our breast images come from a study for demostrating how much the DBT helps in finding out masses when the density of breast is to high. So since our first learning process from ImageNet to DDSM didn't produce good results, it has been quite impossible to transfer very noised images to a bad model.

The results obtained can't be considered as the end of this reaserch. The first limit of this project has been the computing power. An image classification network needs a very powerful GPU to elaborate images in the best way. Another limit has been the structure and the quantity of the images. The next goal of this project is to make the classification more precise, also implementing mass recognition within Mammography and Tomosythesis. The best thing for us would be having more Data, especially for digital Mammography since it is the most common exam for breast cancer detection and it has been the most difficult to analyze in deep learning. One of the opened problem in radiology is also make a very reliable diagnosis for masses and microcalcifications, since they show a different topology in mammography, for this reason a deep learning model could represent a good support in diagnosis for radioligists.

8 References

- Galati,FrancescaandMarzocca,FlaminiaandBassetti,EricaandLuciani,Maria Laura and Tan, Sharon and Catalano, Carlo and Pediconi, Federica. Added Value of Digital Breast Tomosynthesis Combined with Digital Mammography According to Reader Agreement: Changes in BI-RADS Rate and Follow-Up Management. Breast Care, 2017.
- Burnside ES, Sickles EA, Bassett LW, et al. The ACR BI-RADS experience: learning from history.. J Am Coll Radiol. 2009;6(12):851–860. doi:10.1016/j.jacr.2009.07.023