CentraleSupélec

Breast Cancer detection using different filtering methods

POULET Lancelin

1 Introduction

In the medical field, breast cancer detection is a real challenging problem. An expert is needed to analyze mammography and determine whether or not the patient is prone to breast cancer. With the arrival of Machine Learning and especially Computer Vision, we were able to create models that can automatically detect breast cancers.

With a sufficient dataset available, machine learning is capable of determining patterns or texture that are signs of breast cancer. This is really helpful for doctors, as it permits decision support for cancer diagnosis.

The paper that I choose is focusing on using different filtering methods, in order to benchmark the performances of breast cancer detection models for each of them [2]. The goal is to demonstrate if a particular combination of features is able to give better performances for a given classification model.

All the source code of this project is available on my GitHub repository

2 Method

The data are black and white mammography, obtained from the mini-BIAS dataset. It is composed of 322 images at the resolution of 1024 x 1024 pixels. Other information about the pathology is available. We'll only use the class of abnormality in our study.

In order to create a prediction of a possible cancer over an image, we need to follow multiple preprocessing steps. The preparation of data is crucial in machine learning, and can determine the results of your study. We usually say that 70% of the time is spent preparing the data, and the rest 30% are used to build the model.

In our case, the paper follows these different steps : Removing the label from the mammography; Removing the pectoral muscle; Applying a series of filters; Extracting features from images; Building Machine Learning models. The first steps are mainly extracting the Region Of Interest from the image, to focus our model on the important zones and not learn worthless information. We'll explain each part in the next paragraphs.

2.1 Removing labels

The first operation applied on our dataset is removing the labels. These are information on the mammography such as : left or right breast, etc. To remove them, we'll simply apply morphological operations. In particular, we'll use erosion and dilation. The results are given in the Figure 1.

The computation cost of morphology is quite high for each image. Thus, I computed a parallelization of this process using the function **ThreadPoolExecutor** from the **concurrent.futures** library. Even with this method, this step took about 45 min to remove labels from all 322 images.

2.2 Removing Pectoral muscle

This second part is more difficult, as we need to remove the pectoral muscle from images, to extract only the breast on the image. The paper only describes the use of a K-Means model. This is a famous and easy to use clustering technique in Machine Learning.

For my part I firstly applied a Contrast-Limited Adaptive Histogram Equalization (CLAHE) filter to enhance the image. Then, I used a threshold on the image, to keep only white parts. In fact only the pectoral muscle and the inner part of the breast is bright on the mammography.

With these bright pixels I was able to compute a K-Means algorithm. The segmentation wasn't perfect and occasionally took a few pixels from the breast interior zone, as well as the pectoral muscle. But, in general this process worked well to create a mask and remove the major part of the pectoral muscle, as we can see in Figure 2.

2.3 Applying filters

The main goal of the paper's study is finding the best combination of filters to increase the performances of Machine Learning models. The three main filters used are : Median Filter, CLAHE, Unsharp Masking [3][4]. Different combinations will be applied in this study as : CLAHE alone, MF&CLAHE and MF&CLAHE&USM.

2.3.1 Median Filter

The Median Filter is a nonlinear noise reduction filter primarily used to remove salt-and-pepper noise while preserving edges better than a simple mean filter. Instead of averaging pixel values like the mean filter, it replaces each pixel $I(i,j)$ with the median of the pixel values within a local window $W_{i,j}$, following the formula:

$$I'(i,j) = \text{median}\{I(x,y) \mid (x,y) \in W_{i,j}\}$$

This approach makes the median filter particularly effective at removing impulsive noise without blurring edges, as sharp transitions are preserved rather than being smoothed out.

2.3.2 CLAHE

CLAHE enhances local contrast in an image, making it particularly useful in low-light conditions or images with poor contrast. Unlike standard histogram equalization, it prevents over-amplification of noise by setting a contrast limit. The transformation is applied locally, adjusting each pixel $I(i,j)$ according to the histogram $h(v)$ within its neighborhood, following the formula:

$$I'(i,j) = \frac{\sum_{v=0}^{I(i,j)} h(v)}{N}(L-1)$$

Where L is the number of gray levels and N the total number of pixels. This method enhances local contrast without drastically affecting the overall brightness, making it effective for improving visibility in challenging lighting conditions.

2.3.3 Unsharp Masking

Unsharp Masking (USM) is used to sharpen an image by enhancing high-frequency details such as edges and textures. This technique enhances fine details by subtracting a blurred version of the image from the original and adding back the difference, following the formula:

$$I' = I + \alpha(I - I * G_\sigma)$$

where I is the original image, G_σ is a Gaussian-blurred version of I, and α controls the sharpening strength. The sharpening effect can be adjusted using the radius and amount parameters, allowing precise control over the enhancement of textures and details.

2.4 Extracting features

The last step before computing the model is extracting features from images. To do so, the paper used Gray Level Co Occurrence Matrix (GLCM) and Gray Level Run Length Matrix (GLRLM). These two matrices are often used in image processing, to transform images into tabular data, in order to compute Machine Learning problems like a Random Forest.

The GLCM is more suited to understand the relation of recurrent patterns, whereas the GLRLM is helpful to capture a series of same pixels.

Using these two matrices, we can compute variables with mathematical formulas [5]. We obtained 9 variables : Autocorrelation, Contrast, Cluster prominence, Entropy, Short Run Emphasis, Long Run Emphasis, Gray-Level Non-Uniformity, Short Run Low Gray-level Emphasis, Long Run Low Gray-level Emphasis.

2.5 Machine Learning model

Finally, with the tabular data we can build a Machine Learning model. In the paper, multiple models are bench marked, like : Random Forest, SVM, ANN, K-NN, DT, etc.

I will simply use the Random Forest model, as it showed great performance and is easy to use. The objective of this study is to show the impact of the filter on the performances, so I only took one model to make it easier. The model will thus classify whether a breast image is normal or abnormal, as it is specified in the tabular information available with the dataset.

So, given the dataset of the 9 variables obtained previously on the 322 images, we can split the data into a training and a testing set (70% / 30%). To improve the performance of the model, the paper used a Leave One Out Cross validation. This process is really time consuming as it passes through all images. To get the best performance faster, I used the **Optuna** library. This library acts iteratively to find the best hyper parameters, using parallel processes.

To analyze the performance of our model, the metrics used are : Accuracy, Sensitivity, Specificity, PPV, NPV, AUC, BA and F1-score. These are often used metrics for classification problems.

The results of the Random Forest obtained are given in Table 1. The only problem treated is the recognition of normal or abnormal images. But, furthermore, the paper analyzes models with a benign or malign prediction.

For our case, the Accuracy as well as the AUC are increasing as we use more complex combinations of filters. This result is the same as in the paper, where the combination of MF&CLAHE&USM gave the best results.

Nevertheless, the performances of this last combination of filters gave great performances in the paper, but I didn't achieve such results. This may be due to the pectoral muscle removing process, which may have made images a little noisy.

3 Applying on new data

Furthermore, to test the performances of filter's combination I took another well known breast cancer dataset : Kaggle - DDSM Mammography Database. The images present in this dataset are a little bit different as they are zooms of breast image (Resolution of : 299 x 299 pixels) see Figure 4.

For this dataset, I didn't remove labels or pectoral muscle as they are not present in images. But, I applied the same combination of filters as before and computed a Random Forest, optimized with Optuna. The results of this study are presented in Table 2.

Here, the metrics are quite similar between the filter's combination, even if they are both well performing. This may be due to the type of images. As images are smaller and don't represent the entire breast, filters have a smaller effect, and may not enhance important parts for the prediction.

4 Discussion

The effect of combining multiple filters has been well replicated on the same dataset. But the performances are a little under what has been presented, this may be due to the pectoral muscle method or the parameters used in each filter.

Besides that, each of the three filters used in this study permits to enhance or clean the image, before applying models. The Median Filter reduces the noise to prevent artifacts. The CLAHE Filter enhances contrasts to make sure features are more visible. Finally Unsharp Masking is useful for edge sharpening, to clarify important details.

The combination of them can be theoretically beneficial for the final performances. Never a study has shown such a combination of filters. Previous studies only used a single filter or a double combination [3][6]. Thus this is a new way of pre-processing images, never seen in the literature.

Nevertheless, each of these filters have parameters that can affect their behavior. By combining them with the wrong parametrization, we could over enhance the image. Indeed, the effects of multiple filters can improve the performances, or in the opposite, completely corrupt the image with a bad parametrization. A possible solution to find the best parametrization would be for example maximizing a quality image's criteria as the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) for example.

Further discussion may be interesting about the impact of these filtering methods over different types of breast tissues. Does this approach work on all kinds of mammography (ex: different mammary density) ? The resolution of images may also affect the behavior of filters, and need to be further studied.

This new approach has shown great results with basic Machine Learning algorithms. A better image quality gives great performances. By enhancing the contrast of important details, we are able to reduce feature extraction errors. The next step would be applying it to a larger dataset, and more complex models [1].

5 Conclusion

In conclusion, combining multiple filtering techniques can increase the prediction performances of a model. The combinations MF&CLAHE&USM and CLAHE&USM give really good results. But, the replicability of these results may change with a different dataset or parametrization of the given filters.

The paper brings a novel approach of preprocessing images never used in the literature. This may open a new way of using filters to enhance images. At the end, with the help of appropriate filters, computer vision models can show great performances, now in the medical field, but in the future maybe in broader applications.

Appendix

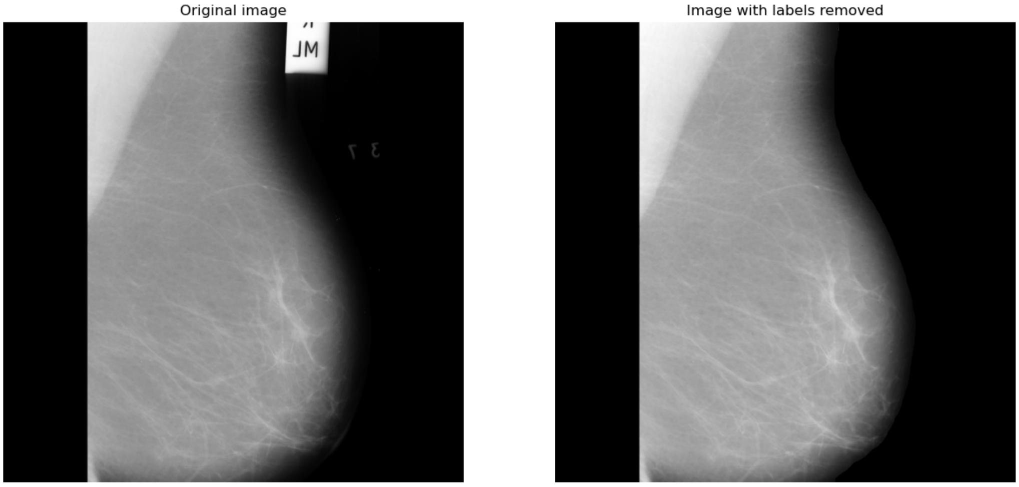

Figure 1: Removing labels from mdb006

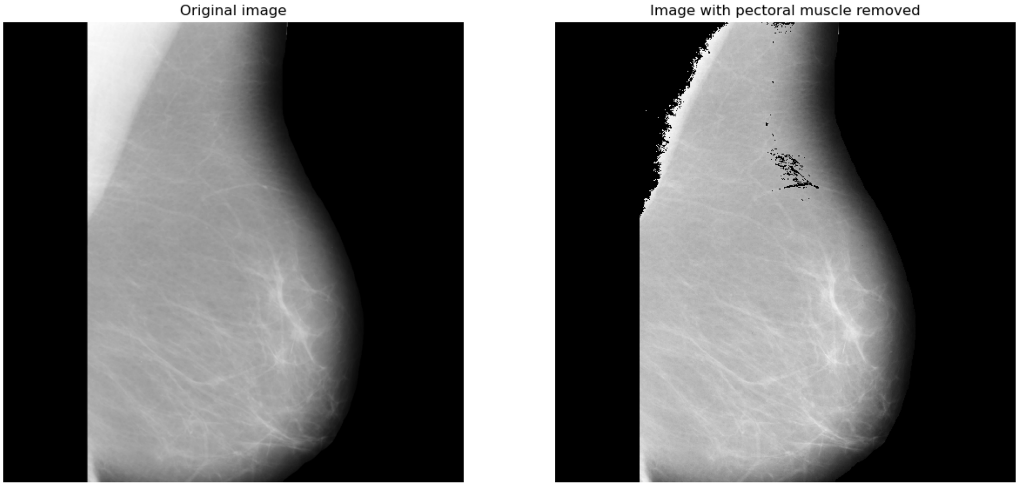

Figure 2: Removing pectoral muscle from mdb006

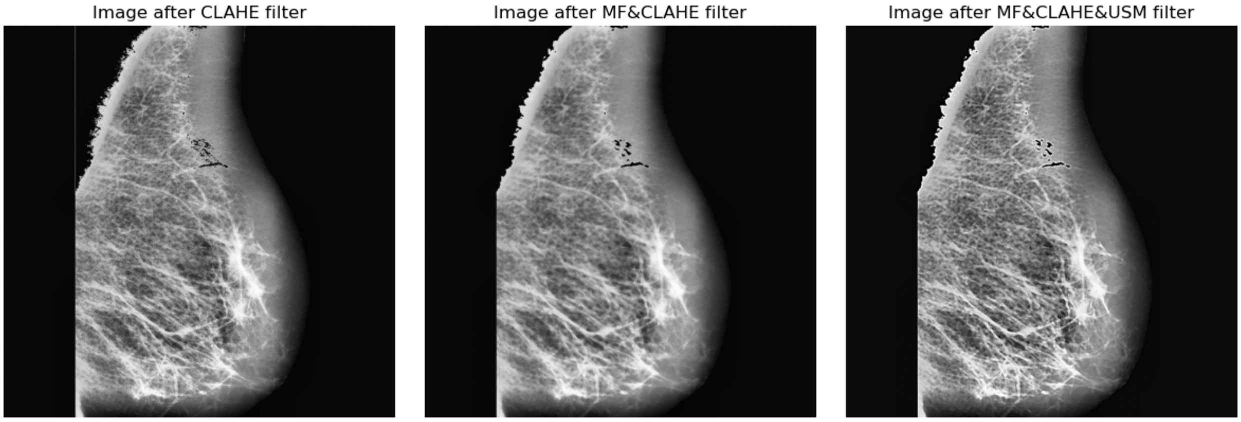

Figure 3: Different filters applied on mdb006

Filtre(s)	Accuracy	Sensitivity (Recall)	Specificity	PPV (Precision)	NPV	F1-score	AUC	Balanced Accuracy
CLAHE	0.598	0.905	0.029	0.633	0.143	0.745	0.467	0.467
MF&CLAHE	0.608	0.937	0.000	0.634	0.000	0.756	0.468	0.468
MF&CLAHE&USM	0.649	0.841	0.294	0.688	0.500	0.757	0.568	0.568

Table 1: Performances on mini-MIAS

Filtre(s)	Accuracy	Sensitivity (Recall)	Specificity	PPV (Precision)	NPV	F1-score	AUC	Balanced Accuracy
CLAHE	0.912	0.463	0.977	0.742	0.927	0.571	0.720	0.720
MF&CLAHE&USM	0.901	0.423	0.970	0.673	0.921	0.520	0.697	0.697

Table 2: Performances on DDSM

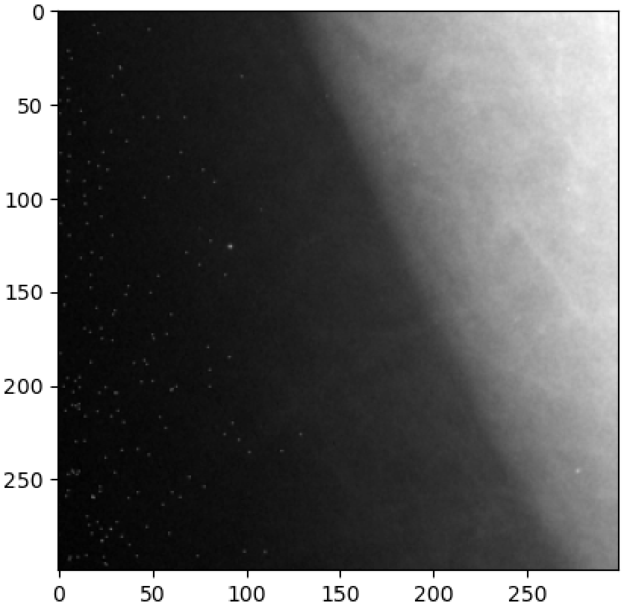

Figure 4: Example of an image from DDSM database

References

[1] Syed Jamal Safdar Gardezi et al. “Breast Cancer Detection and Diagnosis Using Mammographic Data: Systematic Review”. In: *J Med Internet Res* 21.7 (2019). DOI: https://doi.org/10.2196/14464.

[2] Avcı Hanife and Jale Karakaya. “A Novel Medical Image Enhancement Algorithm for Breast Cancer Detection on Mammography Images Using Machine Learning”. In: *Diagnostics* 13.3 (2023). DOI: https://doi.org/10.3390/diagnostics13030348.

[3] Nijad Al-Najdawi, Mariam Biltawi, and Sara Tedmori. “Mammogram image visual enhancement, mass segmentation and classification”. In: *Applied Soft Computing* 35 (2015), pp. 175–185. DOI: https://doi.org/10.1016/j.asoc.2015.06.029.

[4] S. Valarmathy R. Ramani N.Suthanthira Vanitha. “The Pre-Processing Techniques for Breast Cancer Detection in Mammography Images”. In: *Int. J. Image Graph. Signal Process* 5 (2013). DOI: https://doi.org/10.5815/ijigsp.2013.05.06.

[5] Milos Radovic et al. “Parameter optimization of a computer-aided diagnosis system for detection of masses on digitized mammograms”. In: *Technology and Health Care* 23.6 (2015), pp. 757–774. DOI: https://doi.org/10.3233/THC-151034.

[6] Alain Tiedeu et al. “Texture-based analysis of clustered microcalcifications detected on mammograms”. In: *Digital Signal Processing* 22.1 (2012), pp. 124–132. DOI: https://doi.org/10.1016/j.dsp.2011.09.004.