Mammography and Biopsy AI Assistant

Machine Learning Project Competition 2021

I. Introduction

This document produced within the frame of the **Machine Learning Pipo Competition** presents an AI assistant useful for cancerologists to determine whether patients suffering or likely to suffer from breast cancer.

II. Problem presentation and related works

A. Problem presentation

Also affecting men , breast cancer is the most common cancer in women worldwide, with nearly 1.7 million new cases diagnosed in 2012, representing about 25 percent of all cancers in women. Referring to Global Cancer Observatory backed by the IARC (International Agency for Research on Cancer) [4], incidence rates vary widely across the world, from 27 per 100,000 in Middle Africa and Eastern Asia to 92 per 100,000 in Northern America. It is the fifth most common cause of death from cancer in women, with an estimated 522,000 deaths (6.4 per cent of the total) each year. It is also the most frequent cause of cancer death in women from regions characterised by low indices of development and/or income (14.3 percent of deaths), and the second most frequent from regions characterised by higher indices of development and/or income (15.4 percent of deaths), after lung cancer. This growing trend has also been observed in the central African country of Cameroon, where breast cancer is now the leading cancer among women in Yaounde, comprising 18.5% of all cancers and 32.5% of female cancers. According to the International Agency for Research on Cancer, there were 2,625 new cases of breast cancer per 100,000 women in Cameroon during 2012. Hence the question: **How can we help doctors to find breast cancer tumors more efficiently and quickly?**

B. Motivation

A recent review of breast cancer data suggests that the lack of early detection programs and limited access to surgical care, the primary treatment modality for breast cancer in Sub-Saharan Africa, are likely contributors to poor overall survival among breast cancer patients in the region. A retrospective cohort study conducted in Yaounde, Cameroon, found an overall 5-year survival rate of 30% and a 10-year survival rate of 13.2% among breast cancer patients treated between 1995 and 2007. In contrast, breast cancer survival rates in high-income countries are reported to be over 80%. This disparity highlights the critical role of early detection and access to surgical treatment represent in improving breast cancer outcomes and survival, as they remain central components of breast cancer control strategies. Early diagnosis of breast cancer is critical to reducing cancer-related mortality, particularly where radiation, hormonal, and chemotherapy are not widely available. Early detection relies on breast awareness and utilization of screening methods. While mammography is the only screening modality proven to reduce breast cancer mortality, mammography is neither affordable nor feasible in many low- and middle-income countries.

C. Related works

Several works have been undertaken in this field, notably the **Panafrican medical journal** which provides a histo-epidemiological profile available at [5]. YS Sun, Z Zhao, ZN Yang, presented in 2017 in this article [8] the risk factors and means of prevention of breast cancer, the second leading cause of cancer deaths among women at the time is. We also have Y. Zou, Z. Guo with [7] where they present impedance techniques for breast cancer detection like vitro and vivo impedance measurement of human breast tissues. In November 4, 2020, Pacilè S, Lopez J, Chone P, Bertinotti T, Grouin JM, Fillard P. published [6] where they evaluate the benefits of an artificial intelligence (AI)–based tool for two-dimensional mammography in the breast cancer detection process. Several solutions have been proposed ranging from cancer detection based on a description of the X-Ray Images to breast imaging. Most of these techniques aim at telling if it is benign or malignant (assuming they already have cancer), sometimes without considering the patients' past medical results - if they have ever been suspected of having cancer, - if a parent has or had cancer.

D. Proposed solution

To overcome this problem, we proposed an AI Assistant named **breastCanDiag**, that will allow the doctor to quickly identify patients with or at risk of cancer taking into account the patient's past medical results. In practice, mammography can detect suspicious lesions in the breast and the biopsy can tell whether or not it is cancer. The detection of breast cancer is done in these 2 steps.

III. Methodology

To be able to build a reliable model allowing to determine the future presence, the presence or not of a breast cancer, we carried out several stages:

1) The search for important factors involved in mammography and biopsy.
2) The search of datasets for mammography and biopsy.

3) The creation and validation of a model.
4) The development of an interface for the doctor.

A. The research for important factors involved in mammography and biopsy.

In this phase, we had to determine all the factors involved in mammography and biopsy. After various searches carried out on sites dealing with medicine and in medical articles, it was observed that the principal means used by the doctor specialised in radiology to know the state of the breast of a patient is to examine the results of his mammogram. It consists of an X-ray tube and a breast compression system. As we said before, the aim of mammography is to detect suspicious lesions in the breast. The examination takes about 20 minutes. Several attributes can be used by the radiologist for analysis during a mammography. We have selected 8 attributes that were relevant in the searches :

- age : patient's age in years at time of mammogram.
- density : patient's BI-RADS breast density as recorded at time of mammogram.
  - 1 : Almost entirely fatty
  - 2 : Scattered fibroglandular densities
  - 3 : Heterogeneously dense
  - 4 : Extremely dense
- famhx : family history of breast cancer in a first degree relative.
- hrt : current use of hormone therapy at time of mammogram.
- prvmam : binary indicator of whether the woman had ever received a prior mammogram.
- biophx : binary indicator of whether the woman had ever received a prior breast biopsy.
- bmi : body mass index at time of mammogram.

The doctor bases his verdict on these attributes. His verdict is what we will call 'assess' which is described as follows :

- assess : Radiologist's assessment based on the [1]BI-RADS scale : used to define the abnormalities seen and to determine what should be done next, return to screening, close follow-up or biopsy[2].
  - 0 : Needs additional imaging
  - 1 : Negative
  - 2 : Benign finding(s)
  - 3 : Probably benign
  - 4 : Suspicious abnormality
  - 5 : Highly suggestive of malignancy

It is this attribute that we like to predict. So,

- X = (age , density ,famhx, hrt , prvmam, biophy, bmi)
- y = assess

Biopsy consists of taking a sample of a lesion detected by palpation or mammography to be examined under a microscope. We take the X-ray images and make the mask to reveal distinctly the cancer cells.

- X = X-ray image
- y = corresponding Image mask

B. The research of datasets for mammography and biopsy.

After having the different features that are used to predict the breast cancer, we looked for data from the mammograms and for some X-ray images of Cameroonian patients. For this purpose, We wrote a request for data collection at the regional delegation of public health of the center. But due to administrative slowness, we have not yet received a response. So we conducted our study from a dataset provided by [3] for mammography. The dataset consists of the previously 09 attributes plus 03 additionnal ones:

- binary indicator of comparison mammogram from prior mammography examination available
- binary indicator of cancer diagnosis within one year of screening mammogram
- cancer type : Ductal carcinoma in situ or invasive cancer

Regarding the biopsy, thank you to [1] who were given a dataset relatively up to date since it dates from Feb 2020. The dataset contains images and corresponding mask of benign and malign tumors.

C. Creation and validation of a model.

1) Mammography: Already, since we want to be a support to cancerologists, we give them an evaluation of the possibility that a patient is in each case (in the form of probabilities) corresponding to each value of "assess" attributes described previously. We have chosen the **Multiclass ROC AUC score** to evaluate our models. ROC AUC score is scale-invariant. It measures how well predictions are ranked, rather than their absolute values. We use 'ovo' (One Vs One) approach i.e. compute the average AUC of all possible pairwise combinations of classes. At the beginning we are 03 approaches:

- **SVM** (Support Vector Machine) : The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. We had an AUC score of **0.72**
- **LGBM Classifier** (Light Gradient Boost Model) : fast, distributed, high-performance gradient boosting framework based on decision tree algorithm. We had a worse score, **0.69**
- **Random Forest Classifier** (with criterion='gini', min_samples_split=2, n_estimators=1000, other parameters have their default value) was significantly better, which **0.87**

2) Biopsy: The objective is to do the segmentation mask of cancerous tumours. We have used Keras **U-Net** architecture, very common on computer vision. In fact, we are used **semantic segmentation** : classify all pixels on the image (is part of the tumor or not) At the figure 1 , we present an

[1] Breast Imaging-Reporting And Data System

original image, figure 2 the mask given in dataset and figure 3 predicted mask.

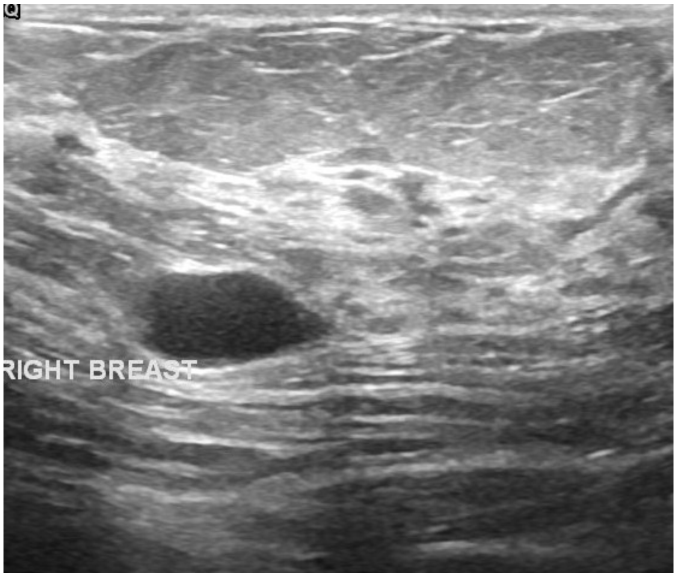


Fig. 1. Original Image

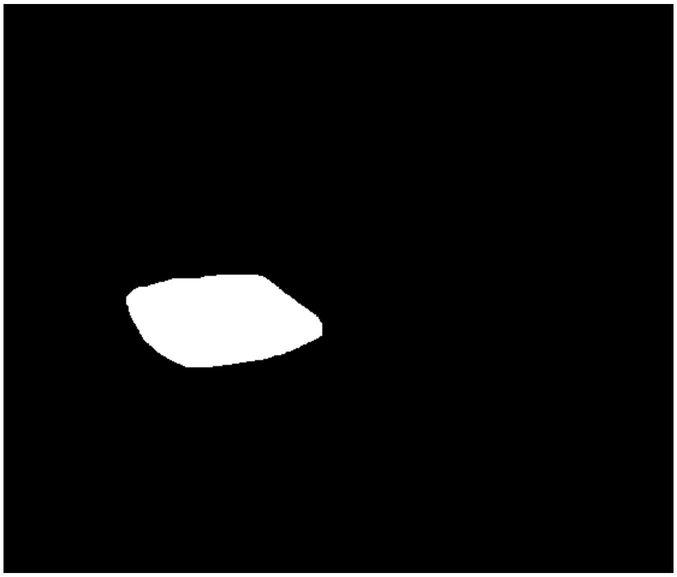


Fig. 2. Original Mask

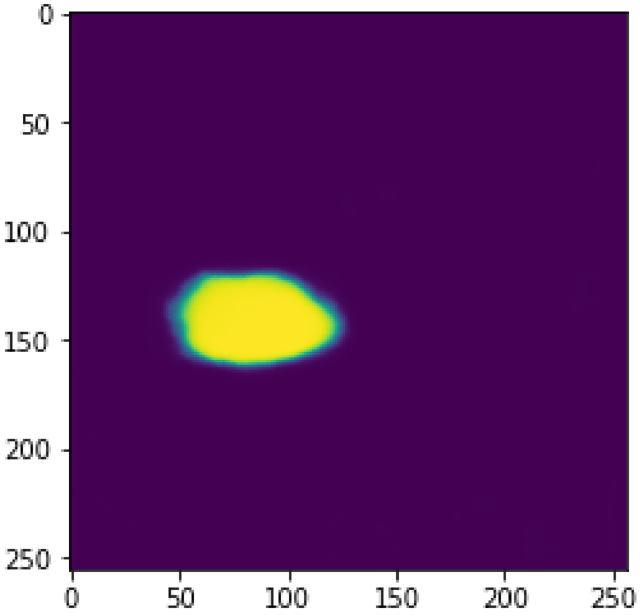
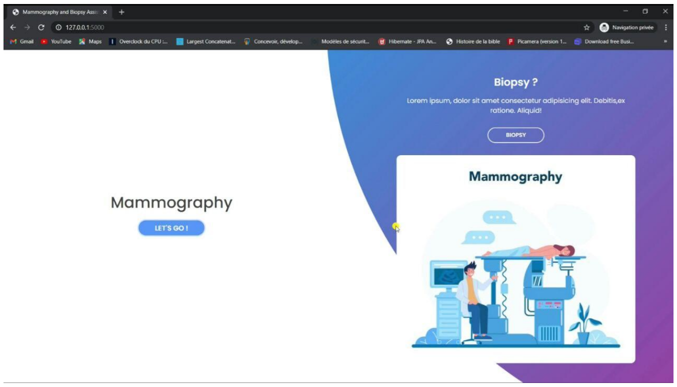


Fig. 3. Predicted Mask

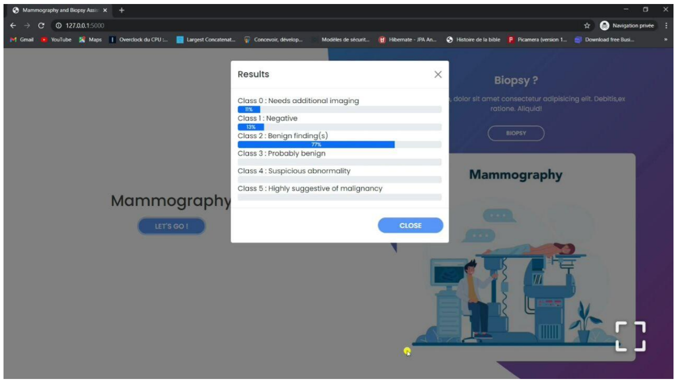
D. Interface for a doctor

We have developed a web application with Flask to allow the doctor to interact with our models. For the mammography after filling in the fields concerning the patient's information, a request is sent to us and we return the probability that the patient is in each of the cases corresponding to the 06 values of the attribute ”assess”. At figures 4 and 5 we present some interfaces for mammography.



Lorem ipsum, dolor sit amet consectetur adipisicing elit. Debitis,ex ratione. Aliquid!

Fig. 4. Mammography Home interface



Results

Class 0 : Needs additional imaging 11%
Class 1 : Negative 13%
Class 2 : Benign finding(s) 77%
Class 3 : Probably benign
Class 4 : Suspicious abnormality
Class 5 : Highly suggestive of malignancy

Fig. 5. Mammography results interface

For biopsy, an image is sent to us and we return the tumor's mask. And the two images are displayed side by side for a better comparison by the doctor. At figures 6 and 7 we present some interfaces for biopsy.

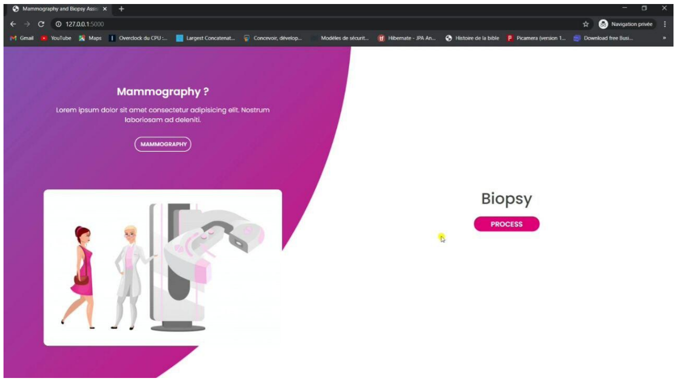


Fig. 6. biopsy Home interface

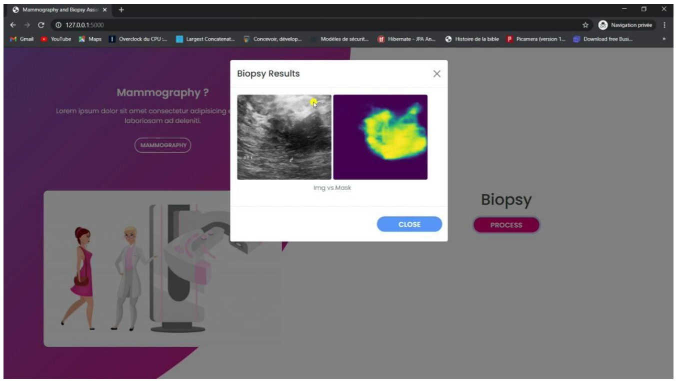


Fig. 7. Biopsy Results interface

IV. Conclusion and Future works

The biopsy is already quite satisfying but the mammography is not. It will be a question for us to improve our model with the data that we will receive from the different hospitals of the place after the acceptance of our request of data collection to the delegation of health. Also, our model is still heavy enough to be deployed online with free accounts on platforms like Heroku. It will therefore be necessary to make it as light as possible (we should not sacrifice performance neither).

References

[1] Khaled H Fahmy A. Al-Dhabyani W Gomaa M. *Dataset of breast ultrasound images. Data in Brief. 2020 Feb;28:104863.* URL: DOI:10.1016/j.dib.2019.104863..

[2] Cancer.org. *Understanding Your Mammogram Report*. URL: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms/understanding-your-mammogram-report.html.

[3] Data collection and sharing was supported by. *the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C). You can learn more about the BCSC at:* URL: http://www.bcsc-research.org/..

[4] IARC. *Cancer Today*. URL: https://gco.iarc.fr/today/.

[5] Panafrican Medical Journal. *Breast cancer in Cameroon, histo-epidemiological profile: about 3044 cases*. URL: http://www.panafrican-med-journal.com/content/article/21/242/full/.

[6] Chone P Bertinotti T Grouin JM Fillard P. Pacilè S Lopez J. *Improving breast cancer detection accuracy of mammography with the concurrent use of an artificial intelligence tool*. URL: https://doi.org/10.1148/ryai.2020190208.

[7] Z. Guo Y. Zou. *A review of electrical impedance techniques for breast cancerdetection*. URL: doi:10.1016/S1350-4533(02)00194-7.

[8] ZN Yang YS Sun Z Zhao. *Risk Factors and Preventions of Breast Cancer*. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5715522/.

Appendix

A. Linked Files

Because of the large size of dump of our models, we preferred to share the google drive folder of the project. The folder is accessible **here.**

B. Description of linked files

In these parts we will describe the elements that are in the shared folder. The folder contains 03 folders :

1) App: In this folder, we have

- project_paper.pdf : this paper.
- "How to deploy + tests.mp4" : a brief video (66 MB) of how can deploy locally the code contains on "code source" folder and some tests
- "code source" (23 MB) contains source code of web app designed for interaction between your model and the doctor
- team_members.xlsx : description of team members

2) dataset: We have :

- breast_dataset.csv : the dataset for the mammography and
- a folder ”biopsy” with contains 02 folders
  - ”benign” for images of benign tumor and corresponding mask
  - ”malign” for images of malign tumor and corresponding mask

3) trained_model: we are

- biopsy.h5 (25 MB) : the saved model for biopsy
- random_forest_model.pkl (2 GB) : the saved model for mammography

Google Drive Project Folder link **https://drive.google.com/drive/folders/1lR51CchPuPIORJMASlQgOhl0CWxHasBC?usp=sharing**