Multi-site data harmonization in mammography

15/10/2020

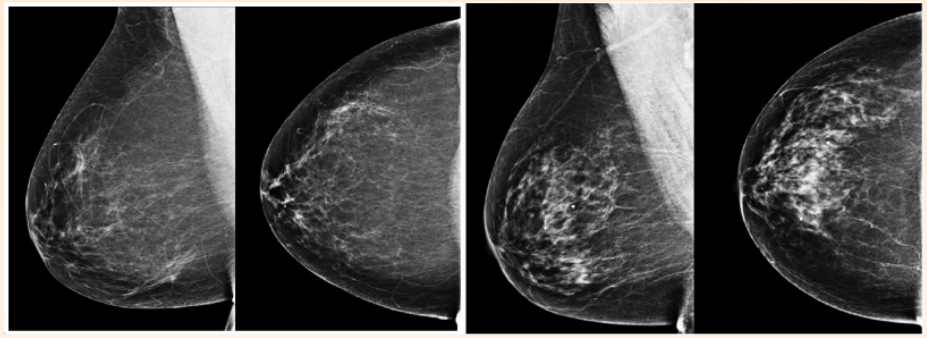
Francesca Lizzi

Istituto Nazionale di Fisica Nucleare, PI

Scuola Normale Superiore

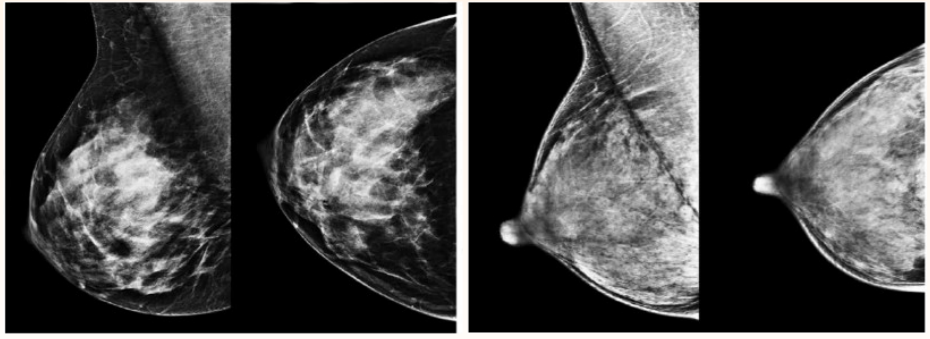
Breast density classification:

4 BI-RADS classes:



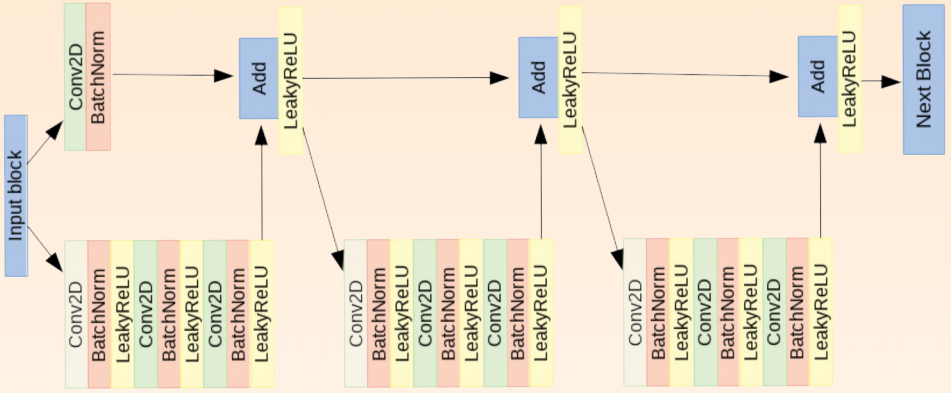
Classe A:
Seno prevalentemente composto di grasso

Classe B:
Seno che presenta aree sparpagliate di tessuto denso



Classe C:
Seno eterogeneamente denso

Classe D:
Seno estremamente denso

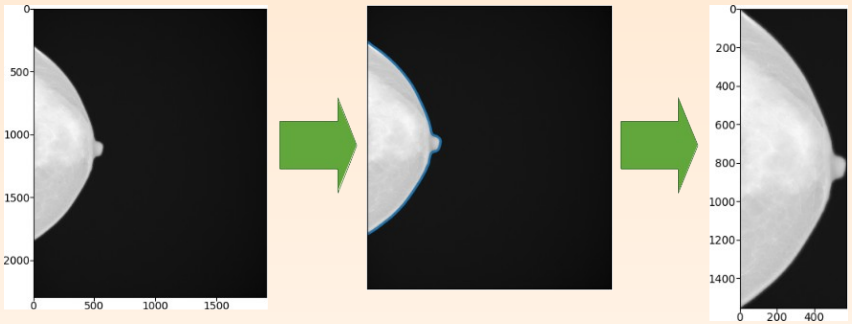


- We trained a Convolutional Neural Network from scratch to classify breast density.

- We optimized its performances studying its behaviour to input changes using grad-CAM explainability algorithm.

Optimization:

- Background removal



- Class distribution

- Best normalization

- Pectoral muscle segmentation

The criterion we used to choose the best pre-processing and performance is based on the values of accuracy, recall and precision and on the visual assessment form activation maps.

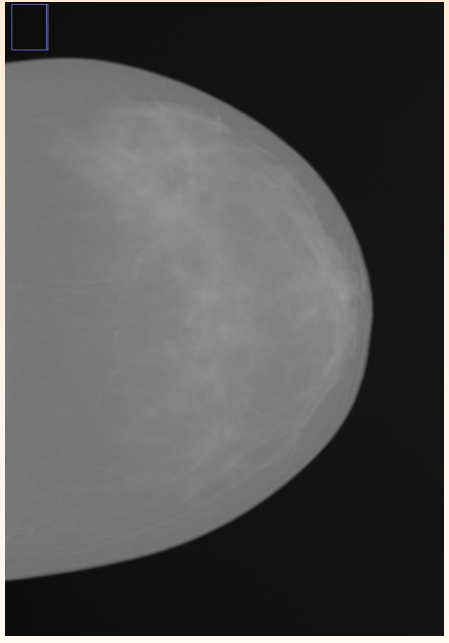
	First version	Last version
Accuracy	77,3 %	83,1 %
Recall	77,1 %	80,1 %
Precision	78,6 %	87,9 %

k = 0.79 (Cohen's coefficient)

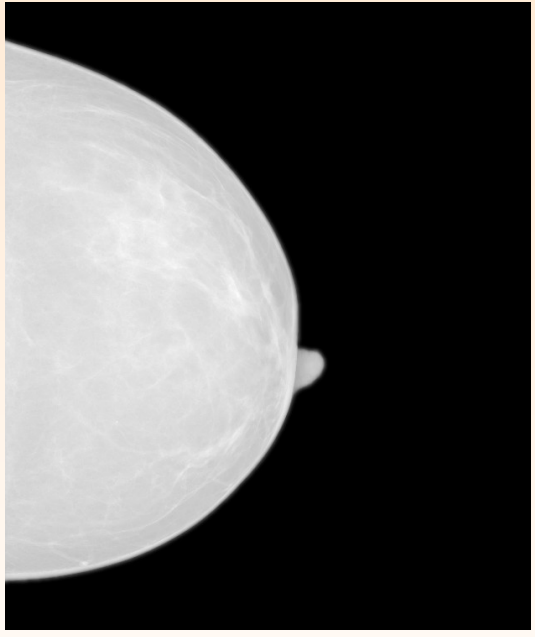
This performances are at the state-of-art in breast density classification, even considering CNN trained on more mammograms and made with HOLOGIC mammographic system. (HOLOGIC, "Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation" → k = 0.67 single radiologist, k = 0.78 consensus)

Data aggregation problems:

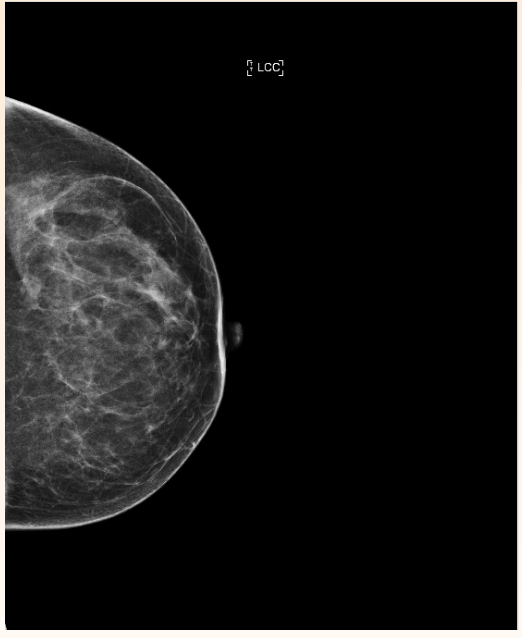
When we collect a real dataset from a hospital, especially for retrospective studies, we always deal with data made with multiple mammographic systems.



Giotto SDL



GE Senograph



Hologic Selenia Dimensions

The standard mammographic form needs the raw data that are usually not available.

Available data:

AOUP dataset (clinical database):

- **Only healthy women**
- **1962 mammographic cases (4 images each)**
- **BI-RADS label by one radiologist**
- **3 mammographic systems**

ATNO dataset (screening database):

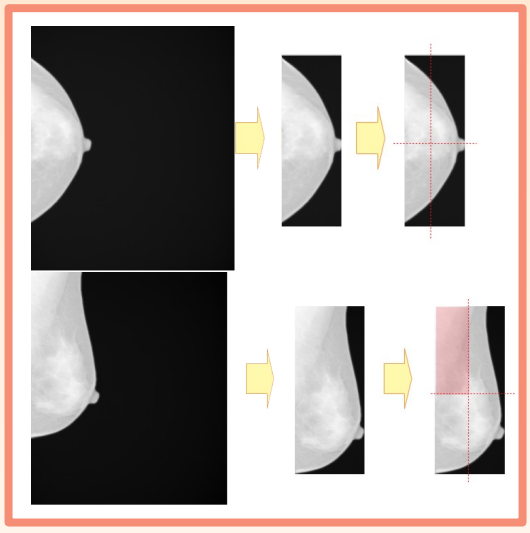
- **500 screen-detected, 72 IC, 270 controls**
- **Questionnaire with breast cancer risk fact.**
- **BI-RADS label by three radiologists**
- **2 mammographic systems**

How can this data be aggregated?

AOUP data	Left breast	Right breast
Accuracy	57,0 %	61,0 %
Recall	46,1 %	49,2 %
Precision	48,4%	51,0 %

Results for the CNN trained on GE Senograph mammograms and tested on GIOTTO mammograms.

Proposal for a radiomic approach:



Features: 10Percentile, 90Percentile, Energy, Entropy, …, Uniformity and Variance.

Support Vector Machine with first order statistical features (pyradiomics) + algoritmo di spiegabilità LIME (Ribeiro et al., “Why should I trust you?”)

(BREAST DENSITY CLASSIFICATION SENOGRAPH)

RIGHT:

ACCURACY: 0.72 (+/- 0.05)

(10 fold CROSS-VALIDATION)

	precision	recall	f1-score	support
macro avg	0.78	0.77	0.78	245
weighted avg	0.78	0.78	0.78	245

(BREAST DENSITY CLASSIFICATION SENOGRAPH)

LEFT:

ACCURACY: 0.71 (+/- 0.07)

(10 fold CROSS-VALIDATION)

	precision	recall	f1-score	support
macro avg	0.77	0.75	0.75	245
weighted avg	0.76	0.75	0.75	245

Further development:

CNN allenata sulle mammografie Senograph e testata sulle GIOTTO.

	Left breast	Right breast
Accuracy	57,0 %	61,0 %
Recall	46,1 %	49,2 %
Precision	48,4%	51,0 %

(BREAST DENSITY CLASSIFICATION GIOTTO)

RIGHT:

ACCURACY: 0.64 (+/- 0.15)

(3 fold CROSS-VALIDATION)

	precision	recall	f1-score	support
macro avg	0.78	0.76	0.77	110
weighted avg	0.78	0.77	0.77	110

(BREAST DENSITY CLASSIFICATION GIOTTO)

LEFT:

ACCURACY: 0.68 (+/- 0.12)

(3 fold CROSS-VALIDATION)

	precision	recall	f1-score	support
macro avg	0.78	0.76	0.77	110
weighted avg	0.78	0.77	0.77	110

Limitation:

- radiomics features are not easly explainable because they are not correlated to clinic.

- there are some reproducibility issues with pyradiomics.

- LIME has many problems regarding reproducibility and effective explanation of models.

Are radiomics features more roboust to mammographic system changes?