Explainable Artificial Intelligence (XAI) in mammography

Francesca Lizzi
Istituto Nazionale di Fisica Nucleare, Pisa
Scuola Normale Superiore, Pisa

Camilla Scapicchio
Istituto Nazionale di Fisica Nucleare, Pisa
Dip. di Fisica, Università di Pisa

Istituto Nazionale di Fisica Nucleare

Artificial Intelligence in Medicine
AIM

Introduction:

Why XAI? → Two main reasons

1) Get an explanation for a classifier:

"An explanation is a local linear approximation of the model's behaviour around the vicinity of a particular instance."

Result: we can obtain a vector which points out the most relevant features the classifier uses to perform the classification.

(a) Husky classified as wolf (b) Explanation

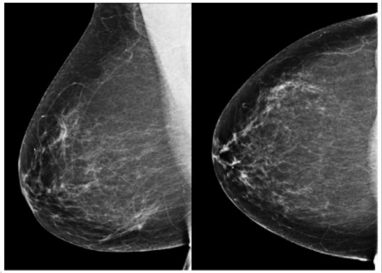
2) Allow the interpretation of models

We can interpret the model if the input is interpretable → PROBLEM FOR CNN

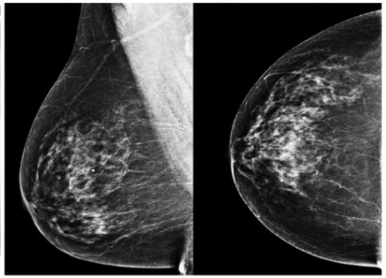
Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs, stat],February 2016. arXiv: 1602.04938

Problem characterization in mammography:

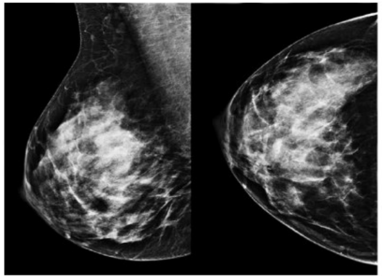
Previously on AIM3.T2 Predictive models for mammography and CESM:



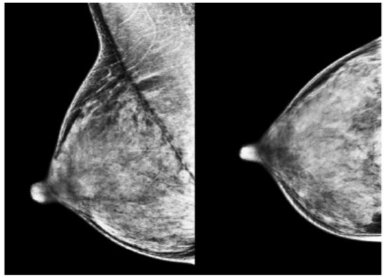
Classe A:
Seno prevalentemente composto di grasso



Classe B:
Seno che presenta aree sparpagliate di tessuto denso

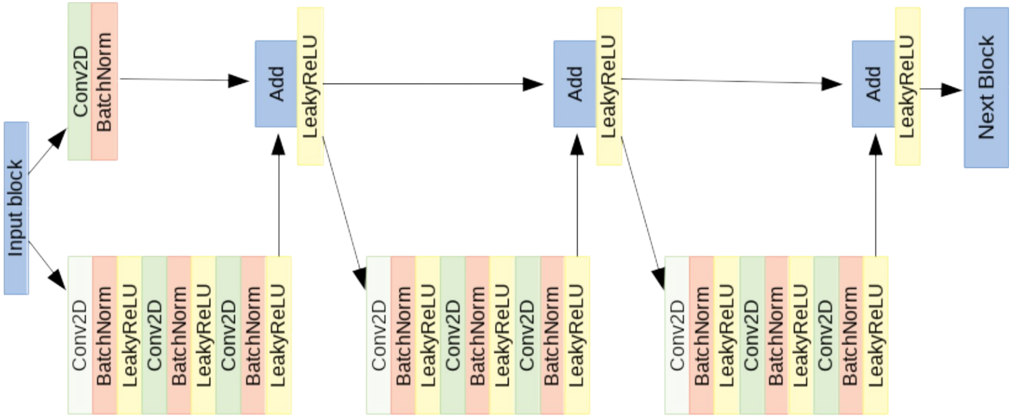


Classe C:
Seno eterogeneamente denso



Classe D:
Seno estremamente denso

We trained and evaluated a Convolutional Neural Network classifier for classifing breast density (4-class) using labeled mammograms.



Residual CNN model with 41 layers.
About 1500 mammographic exams (four images) used to train, validate and test.
Accuracy: 77.3% Recall: 77.1% Precision: 78,6%

Lizzi, F. et al. (2019). Residual Convolutional Neural Networks for Breast Density Classification:. In Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies, pages 258–263.
Lizzi, F., Laruina, F., Oliva, P., Retico, A., and Fantacci, M. E. (2019). Residual Convolutional Neural Networks to Automatically Extract Significant Breast Density Features. In Computer Analysis of Images and Patterns, pages 28–35. Springer International Publishing.

Problem characterization in mammography:

Previously on AIM3.T2:

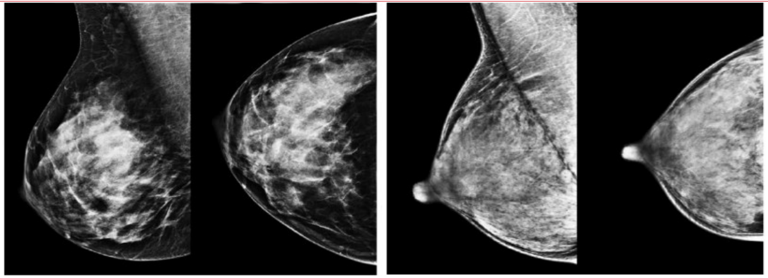
We trained and evaluated a Convolutional Neural Network based classifier for classifing

CAN WE TRUST THIS CLASSIFIER?

IS THIS CLASSIFIER LOOKING AT THE RIGHT PART OF THE IMAGE?

CAN WE EXPLAIN WHY AND HOW THE CLASSIFIER WORKS?

CAN WE AVOID THE RADIOLOGIST SEGMENTATION TO VALIDATE RESULTS?



Classe C:
Seno eterogeneamente denso

Classe D:
Seno estremamente denso

Residual CNN model with 41 layers.
About 1800 mammographic exams (about 7200 images) used to train, validate and test.
Accuracy: 77.3% Recall: 77.1% Precision: 78,6%

Lizzi, F. et al. (2019). Residual Convolutional Neural Networks for Breast Density Classification:. In Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies, pages 258–263.

Lizzi, F., Laruina, F., Oliva, P., Retico, A., and Fantacci, M. E. (2019). Residual Convolutional Neural Networks to Automatically Extract Significant Breast Density Features. In Computer Analysis of Images and Patterns, pages 28–35. Springer International Publishing.

Problem characterization in mammography:

Aggregating data from different mammographic systems

Classifier transferability on different datasets

Dataset distribution

Pre-processing evaluation

Aggregating data from different health institutions

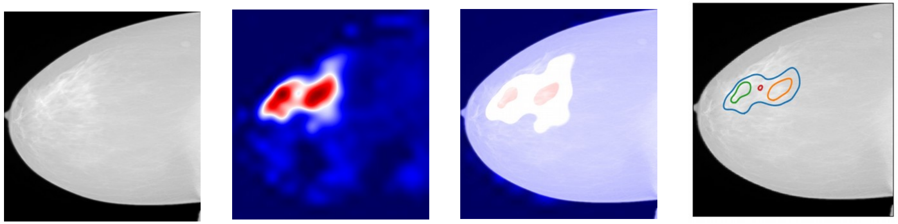
We took two approaches to the problem:

1) Train and evaluate a model that can be interpreted with explainable algorithms:

SVM → LIME → INTERPRETATION

2) Train, evaluate and test methodology to validate a CNN trained with mammograms (Camilla):

CNN → grad-CAM → INTERPRETATION



Support Vector Machine for breast density classification:

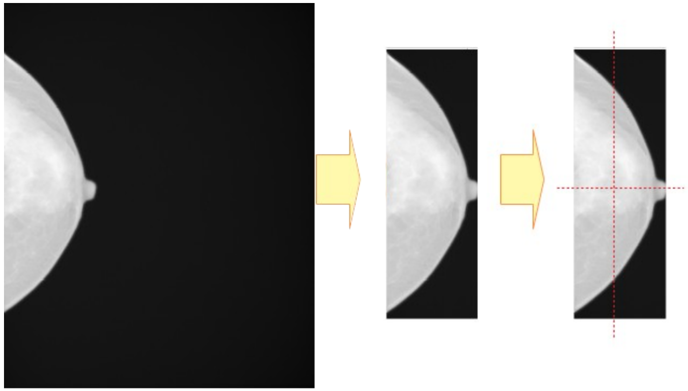
Features design:

About 1500 mammographic cases from Senograph
About 230 cases from GIOTTO

12-bit images, 4 images per patient (4 standard mammographic projection)

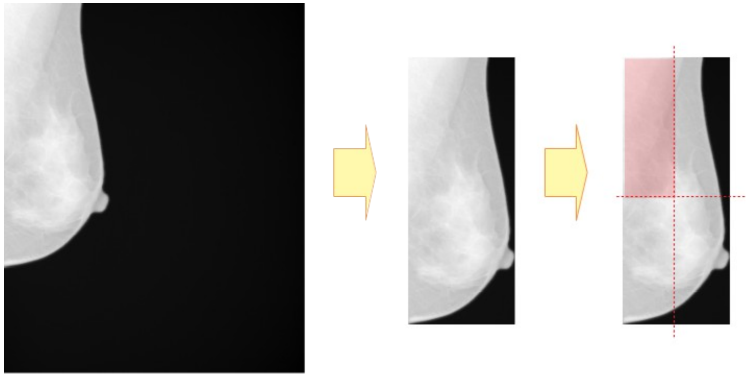
Background removal with marching square algorithm

Cranio-caudal projection:



Masks for computing the features on mammograms

Medio-lateral oblique projection:



Support Vector Machine for breast density classification:

Preliminary results:

We computed the first-order statistical features with pyradiomics:

10Percentile, 90Percentile, Energy, Entropy, Interquartile Range, Kurtosis, Maximum, Mean Absolute Deviation, Mean, Median, ~~Minimum~~, Range, Robust Mean Absolute Deviation, Root Mean Squared, Skewness, ~~Total Energy~~, Uniformity and Variance.

112 features → SVM training: scikit-learn package, linear kernel, one-versus-rest decision function:

Results:

(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

Support Vector Machine for breast density classification:

Preliminary results:

To do:

- Statistical tests (Mann-Whitney, ...)

- Feature selection: LDA, PCA, …

- Model selection.

- Test on other mammographic systems and images with malignant masses.

Manufacturer Giotto:

(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

(BREAST DENSITY CLASSIFICATION GIOTTO)

RIGHT:

ACCURACY: 0.64 (+/- 0.15)

(3 fold CROSS-VALIDATION)

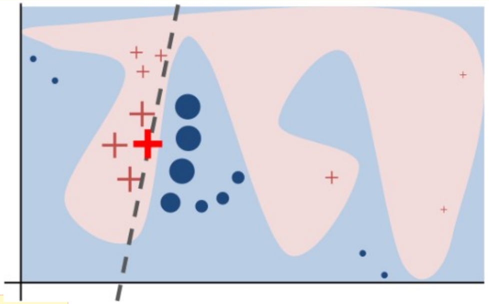
	precision	recall	f1-score	support
macro avg	0.78	0.76	0.77	110
weighted avg	0.78	0.77	0.77	110

Support Vector Machine for breast density classification:

Preliminary results (LIME):

For each patient in the test set (correctly classified), we computed an explaination with LIME algorithm.

Classifier approximation (around an instance) → perturbed samples generation.



Output:
FEATURE NAME, CLASSIFICATION CONDITIONS AND FEATURE IMPORTANCE

Variance_2_MLO > 0.04, 0.10
Variance_2 > 0.03, 0.09
Entropy_1 <= -0.64, -0.08
10Percentile_2 <= -0.28, -0.08
Robu...viation_2_MLO <= -0.48,
Kurtosis_1_MLO > 0.59, 0.07
Uniformity_1_MLO > 0.31, -0.06
Energy_2_MLO <= -0.73, -0.06
Energy_2 <= -0.71, 0.06
Robu...viation_3 > -0.01, -0.05

Class A:

Class B:

Median_1_MLO <= -0.36, 0.16
Entropy_2_MLO <= -0.76, -0.09,
Entropy_1 <= -0.64, -0.08,
Inter...Range_3_MLO <= -0.48, 0.07
Variance_2 <= -0.36, -0.06
Median_2_MLO <= -0.38, 0.06
Uniformity_1_MLO > 0.31, -0.06
90Percentile_1 <= -0.62, -0.05
Inter...Range_2_MLO <= -0.49, 0.05
Kurtosis_1_MLO > 0.59, 0.05

Variance_2_MLO > 0.04, 0.12
Median_1_MLO > 0.56, -0.12
Entropy_2_MLO > 0.90, 0.08,
90Percentile_3_MLO > 0.60, 0.
90Percentile_2_MLO > 0.80, 0.
Variance_4 > -0.05, 0.05
Maximum_1_MLO > 0.71, 0.05
Kurtosis_2_MLO <= -0.71, -0.05
Energy_2_MLO <= -0.73, -0.04
Range_2_MLO > 0.66, 0.04

Class C:

Class D:

10Percentile_3_MLO <= -0.30, -0.13
Inter...Range_2 > 0.50, 0.11
Median_4 <= -0.59, 0.11
Variance_1_MLO > 0.05, 0.10
Uniformity_1_MLO <= -0.76, -0.09
10Percentile_1_MLO <= -0.17, -0.09
90Percentile_3_MLO > 0.60, -0.08
Inter...Range_3_MLO > 0.27, 0.07
Entropy_2 > 0.89, 0.07
Variance_3_MLO > 0.08, 0.07

Support Vector Machine for breast density classification:

Conclusions:

- Linear SVM classifier seems to be promising in performing breast density classification.

- We can obtain an explaination but it is not so easy to be interpreted.

Future works:

- We need to implement the features selection (making input more interpretable)

- We are going to train a mixed classifier (Convolutional Neural Network and SVM) to increase explainability and performances.

(to be continued...)

Explainability of a Residual Convolutional Neural Network for breast density assessment

Camilla Scapicchio

Istituto Nazionale di Fisica Nucleare - Pisa (INFN)

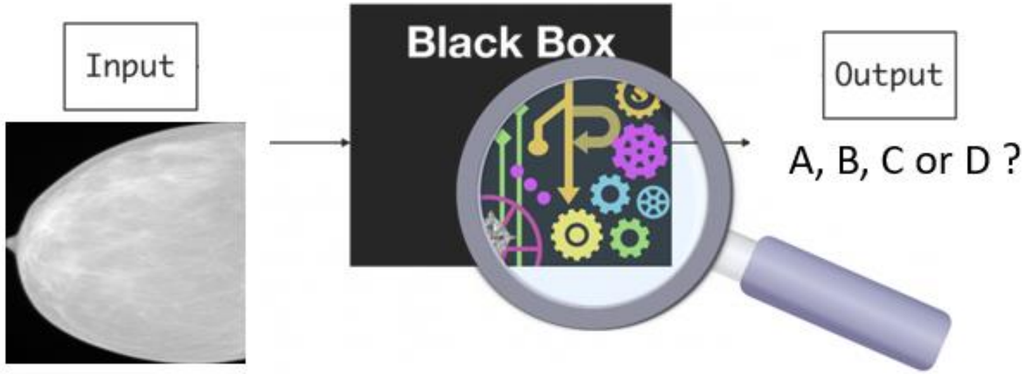
3 Febbraio 2020

Issues

Deep Residual CNN architecture

- Multi-layer nonlinear structure
- Millions of mathematical operations
- About 2 millions learnable parameters

→ Lack of transparency



Goals and Methods

Goal

Explain the classifier behaviour and interpret its internal processes → Assess trust for a potential application in clinical practice

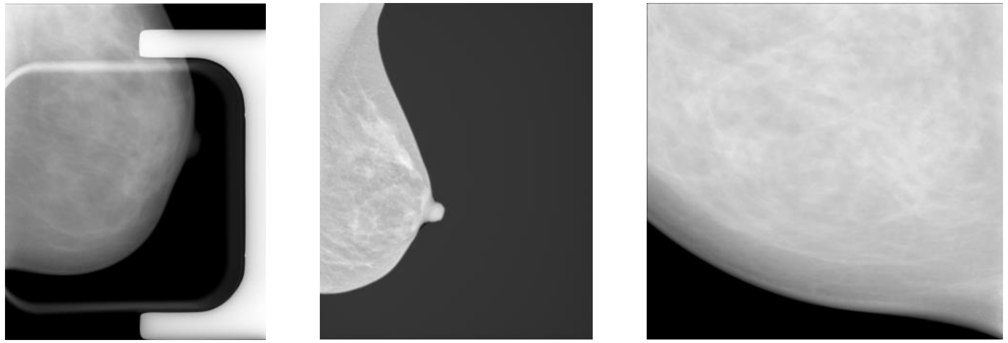
Methods

1. Train the classifier with different specifications to investigate some factors that influence the classifier performance
2. Visual representation of class activation maps (CAM) to understand the reasons behind the classifier predictions

Data preparation

Preprocessing

- 8-bit conversion
- Normalization
- Background removal
- Images inspection to exclude problematic images

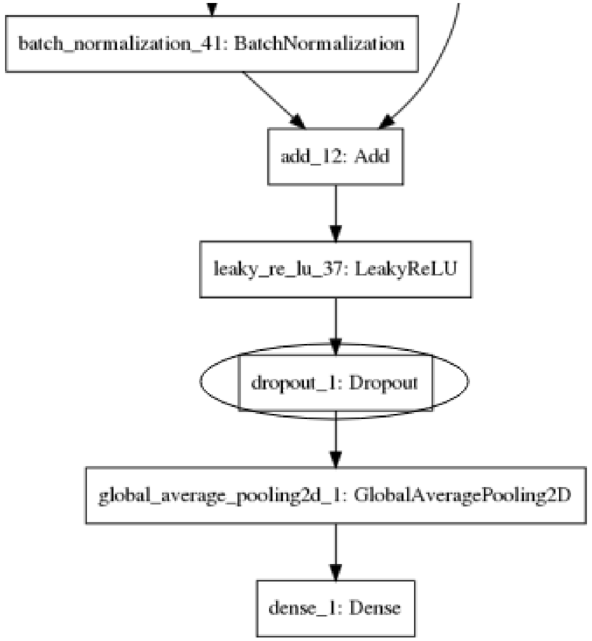


	Original data	Pre-processed data
test accuracy (%)	75.3	83.1
recall (%)	72.1	80.1
precision (%)	76.4	87.9

Model fine-tuning

Dropout

Insertion of a Dropout layer at the end of the network as an additional regularization technique to prevent the model from overfitting



		No Dropout	**With Dropout**
450x450	test accuracy (%)	77.1	83.1
	recall (%)	71.7	80.1
	precision (%)	84.6	87.9
650x650	test accuracy (%)	77.1	78.8
	recall (%)	76.3	77.4
	precision (%)	74.5	79.3
850x850	test accuracy (%)	72.9	79.7
	recall (%)	72.1	76.4
	precision (%)	72.4	84.4

Transferability in dataset distribution

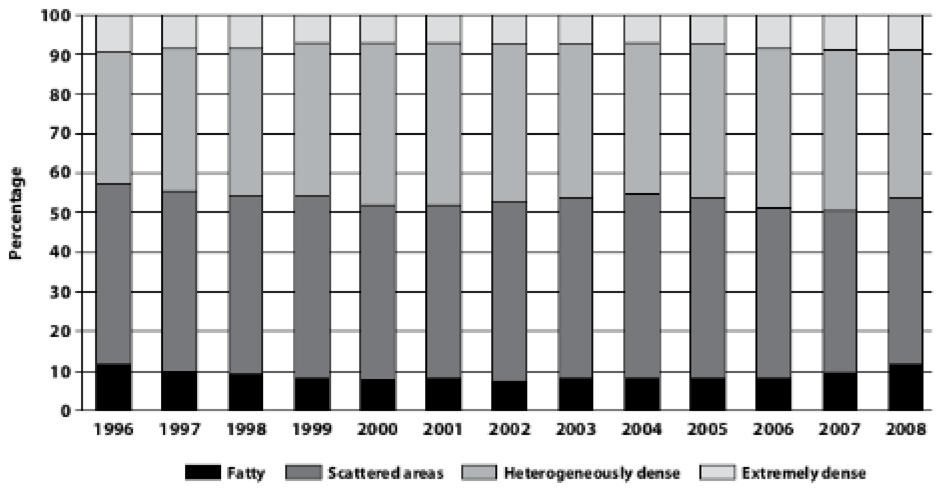
Dataset distribution

AOUP (Azienda Ospedaliera Universitaria Pisana) distribution : A: 12%, B:28%, C:50%, D:10%

BIRADS distribution: A: 10%, B:40%, C:40%, D:10%

Uniform distribution: A: 25%, B:25%, C:25%, D:25%

Training on BIRADS probability distribution and test on different probability distributions



BIRADS training set

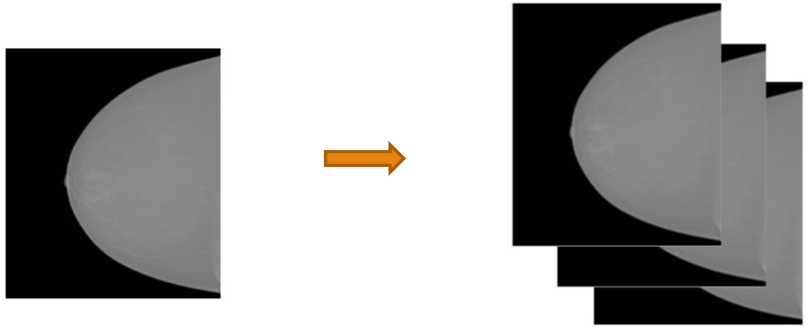
		AOUP Test set	BIRADS Test set	Uniform Test set
1 channel - No dropout	test accuracy (%)	75.3	77.1	65.3
	recall (%)	66.7	71.7	65.3
	precision (%)	83.8	84.6	74.7
3 channels - No dropout	test accuracy (%)	76.6	80.5	75.0
	recall (%)	71.9	76.9	75.0
	precision (%)	78.3	81.2	80.4
1 channel - Dropout	test accuracy (%)	79.1	83.1	73.6
	recall (%)	75.2	80.1	73.6
	precision (%)	82.6	87.9	79.0

From grayscale to RGB

Number of channels

- CNN models are designed to work on RGB images
- Our grayscale images have been fed to the network as RGB images by pasting the grayscale information to all three channels

The image is a tensor composed of three identical 2D matrices with the same pixel value identically repeated in all 3 channels



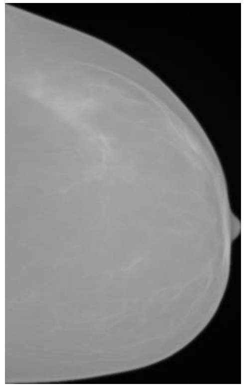
		1 channel	3 channels
450x450	test accuracy (%)	77.1	80.5
	recall (%)	71.7	76.9
	precision (%)	84.6	81.2
650x650	test accuracy (%)	77.1	77.1
	recall (%)	76.3	77.9
	precision (%)	74.5	75.2
850x850	test accuracy (%)	72.9	78.8
	recall (%)	72.1	78.9
	precision (%)	72.4	79.2

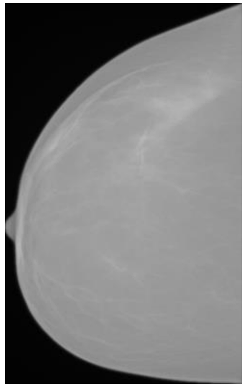
Data augmentation

Horizontal flip

Left breast flipped into right to increase the amount of data we already have

Doubled number of images for each of the two projections (CC, MLO)





Without Data augmentation (924 images)

	right CC	right MLO	All
test accuracy (%)	79.9	69.4	78.4
recall (%)	76.8	64.9	77.1
precision (%)	83.1	66.6	76.9

With Data augmentation (1848 images)

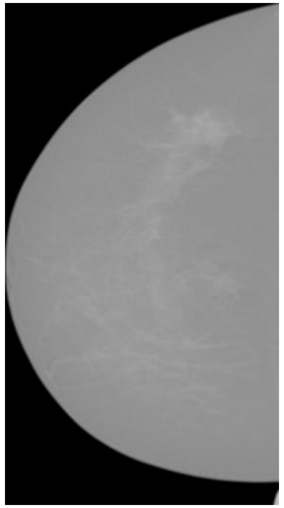
	CC	MLO	All
test accuracy (%)	77.2	72.0	76.1
recall (%)	75.8	65.9	73.7
precision (%)	78.3	70.3	76.1

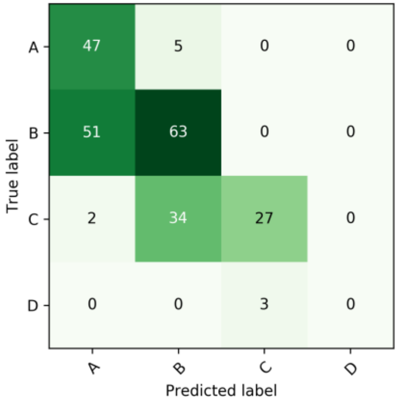
Test on a different mammographic system

GIOTTO IMAGE SDL

Classifier performance tested on mammograms acquired with a different mammographic system

- Small dataset size (232 images per projection)
- Mammograms with a different appearance from the ones used in training





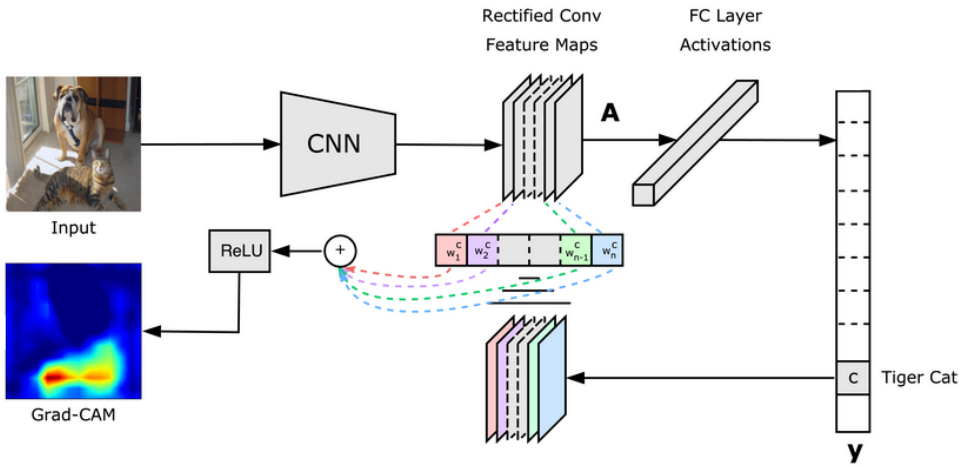
	GIOTTO
test accuracy (%)	59.1
recall (%)	47.1
precision (%)	49.7

Visualization

Heatmaps

A heatmap for a particular category indicates exactly which regions of an image are being used by the model for discrimination among classes

Gradient based Class Activation Map (Grad-CAM): gradient calculation of the final classification score with respect to the final convolutional layer. The places where this gradient is large let us exactly define the region that has a large impact on the final score decision



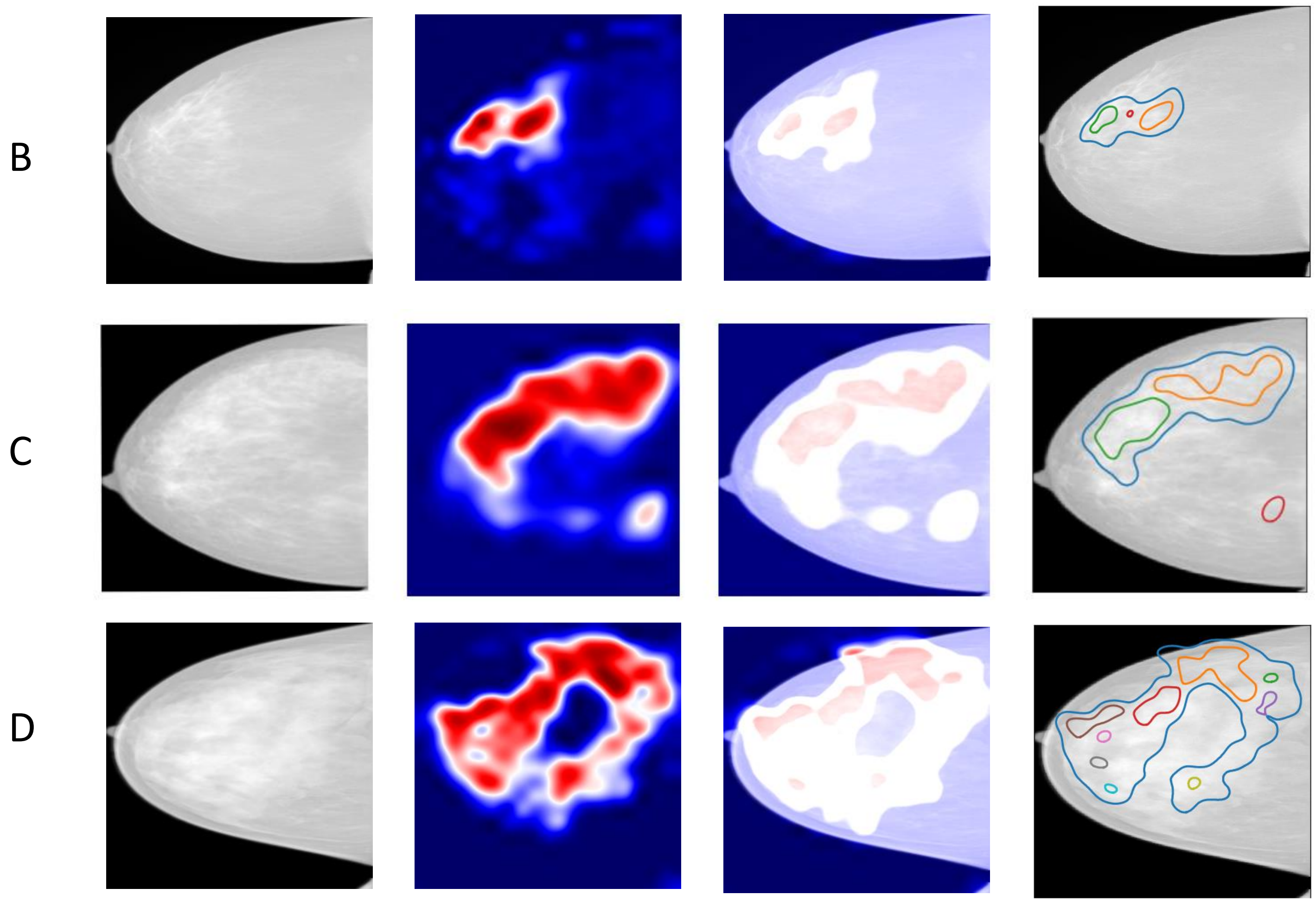
$$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha^c_k A^k\right) \qquad \alpha^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}$$

α^c_k : weights of the final dense layer

c : predicted class

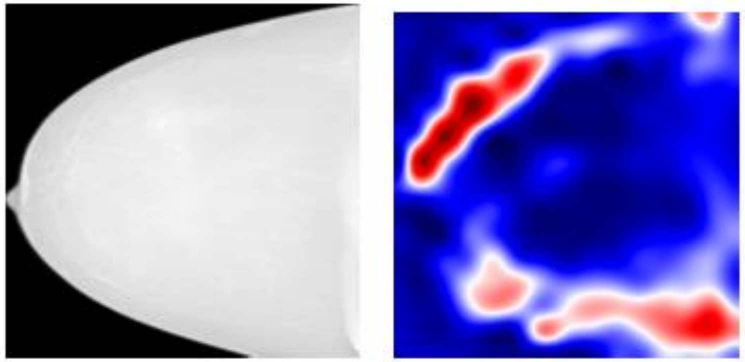
A^k : feature maps of the last conv layer

Visualization – Class Activation Maps



Maps have been evaluated observing if they activate at the densest areas of the breast

For the A class the classifier does not recognize any dense region and the maps activate almost always at the edge of the breast

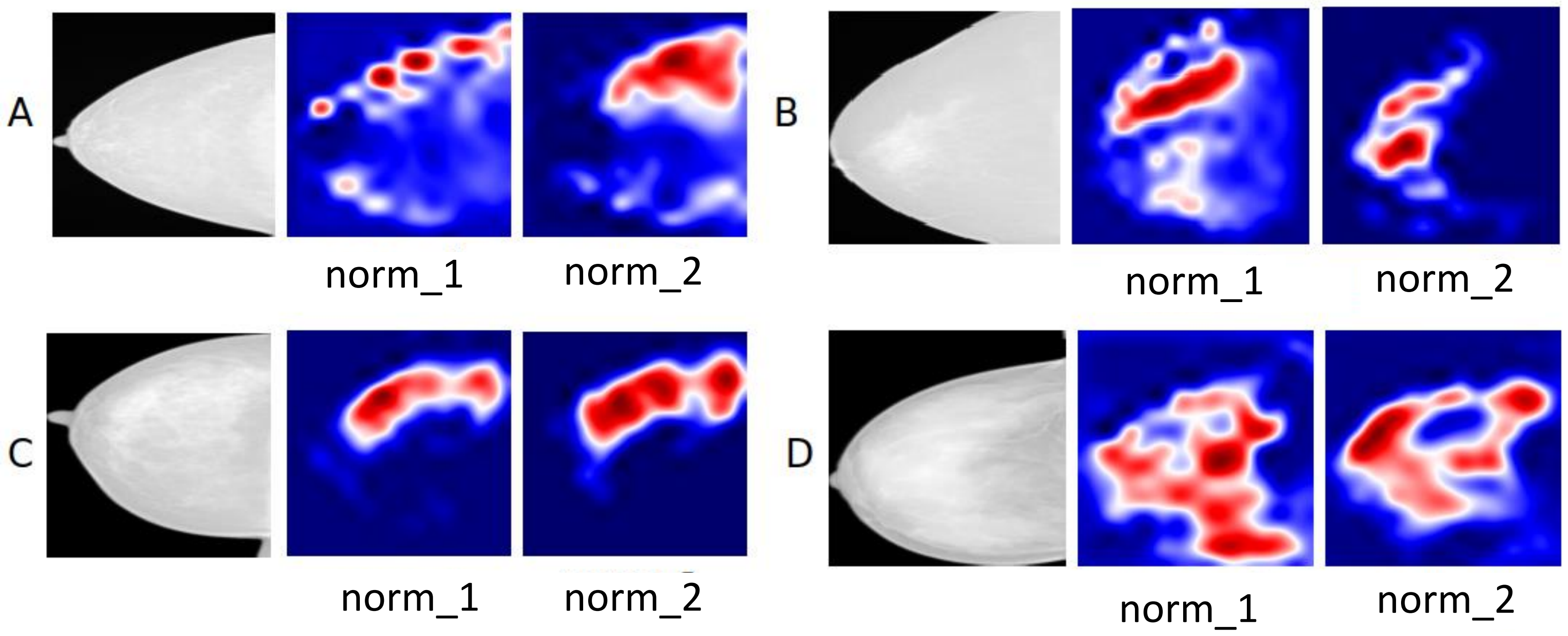


Visualization – Class Activation Maps

Grad-CAM visualization to compare two different types of image normalization

norm_1 : Sets each input mean to 0 and divides each input by its std

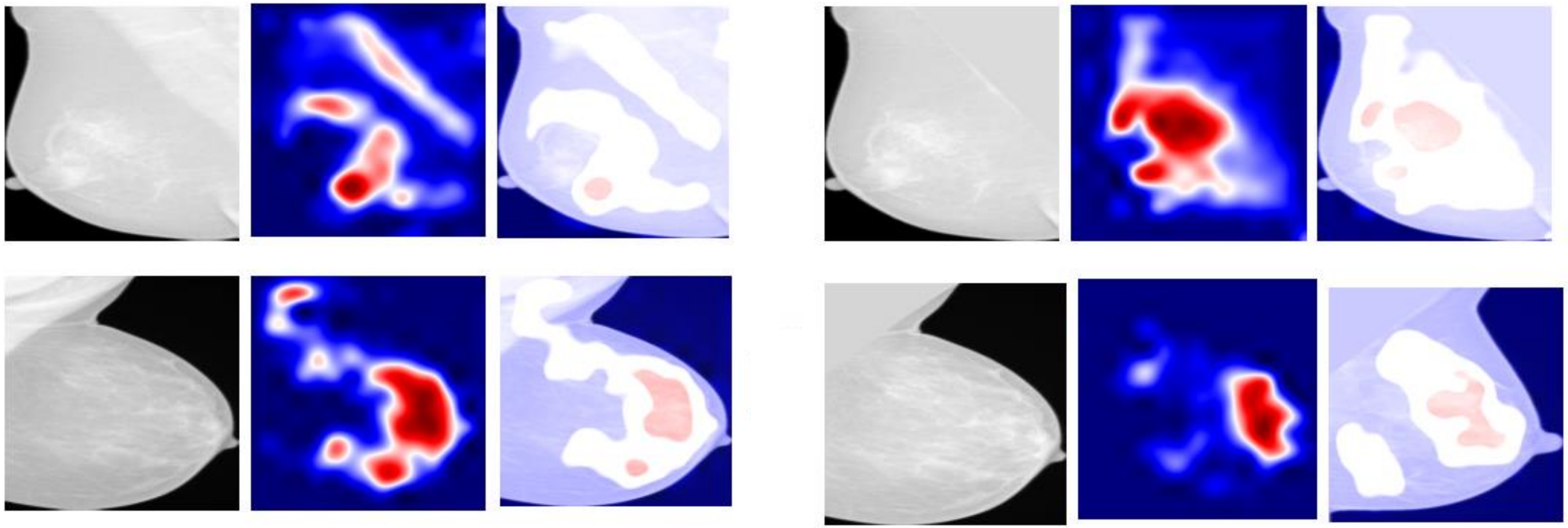
norm_2 : Rescaling factor, multiplies the data by 1./255



Visualization – Class Activation Maps

Grad-CAM visualization before and after muscle segmentation in MLO projections

Non-segmented mammograms | **Segmented mammograms**

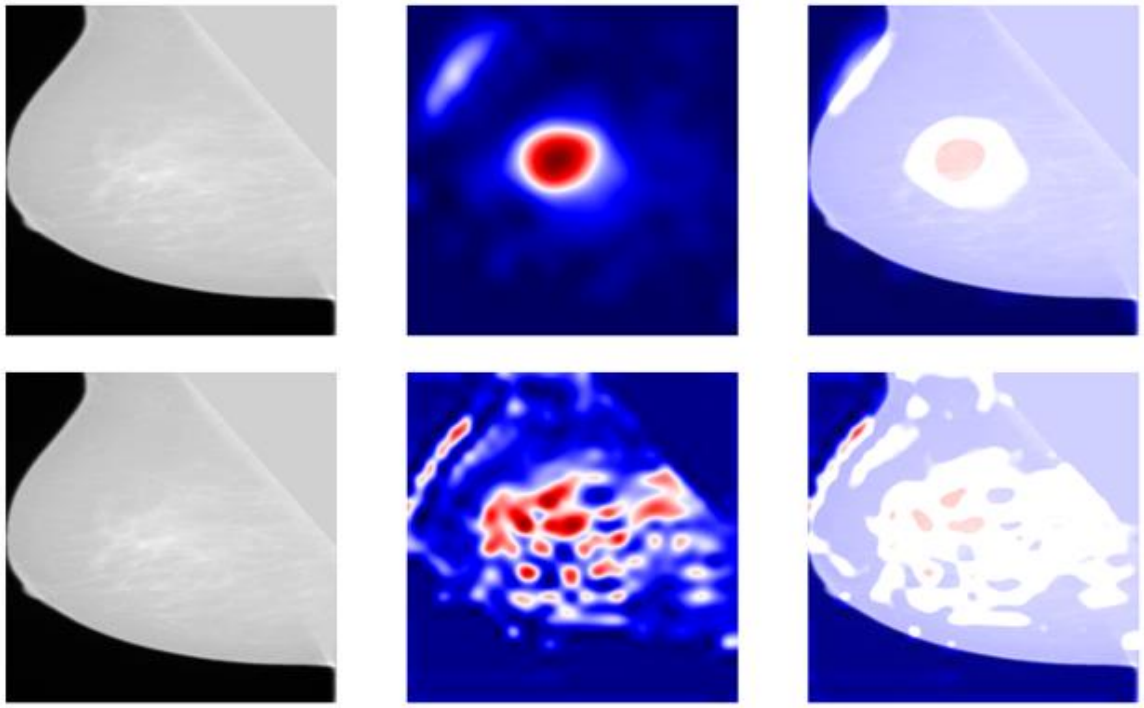


Visualization – Class Activation Maps

Grad-CAM visualization for a specific class at different convolutional layers

add_12

add_9



$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

- The last convolutional layers have the best compromize between high-level semantics and detailed spatial information
- Better resolution at the previous convolutional layers

Where do we stop?

Conclusions

- Why explainability in mammography and in medicine

- Two different approches:
  - Semi-interpretable SVM
  - CNN interpreted through grad-CAM

- Outlook:
  - Mixed and controlled classifier to maximize accuracy results
  - Grad-CAM as region proposal for features computation in SVM
  - We are collecting a new longitudinal dataset from ATNO screening, with cases and controls. Can we use the grad-CAM to track breast density for a very early diagnosis of breast cancer?

Thank you for your attention !

Questions?

Explainability of a Residual Convolutional Neural Network for breast density assessment

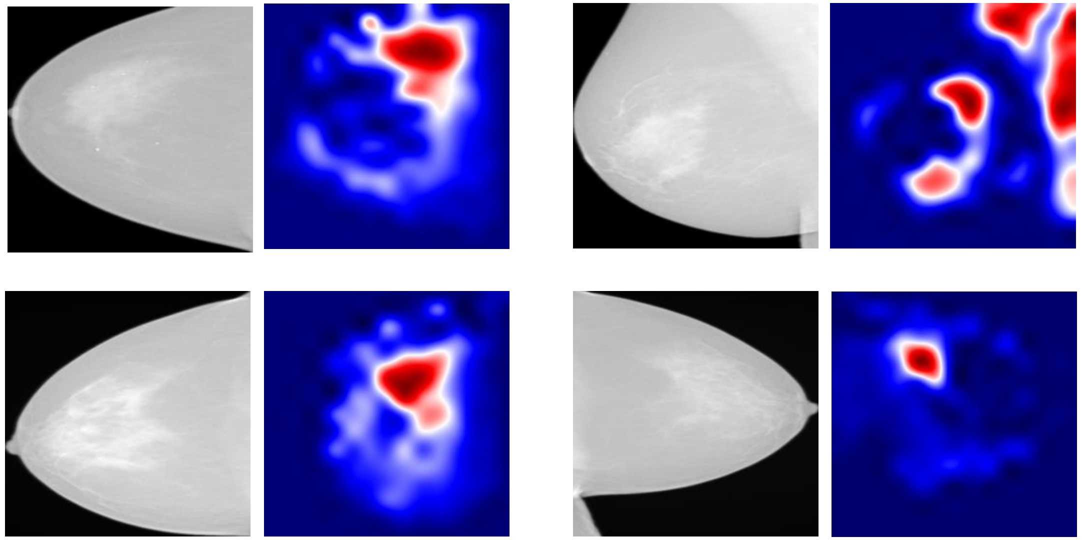
Camilla Scapicchio

Istituto Nazionale di Fisica Nucleare - Pisa (INFN)

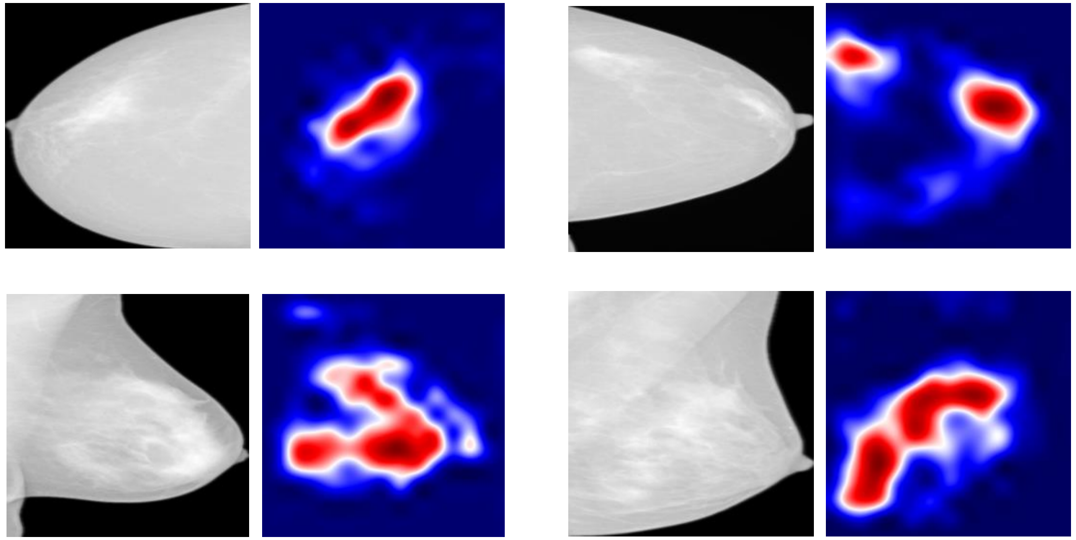
3 Febbraio 2020

Visualization – Class Activation Maps

Grad-CAM visualization to compare two different types of image normalization



norm_1



norm_2