A deep Convolutional Neural Network classifier for breast density assessment: optimization and explainability

Dott.ssa Camilla Scapicchio
Relatrice: Prof.ssa Maria Evelina Fantacci

25 Marzo 2021

IN SUPREMÆ DIGNITATIS
1343
Università di Pisa

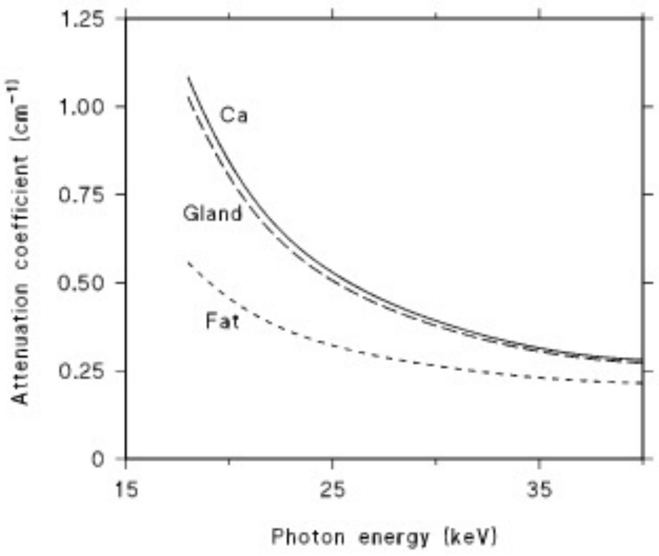
Commissione Calcolo e Reti dell' INFN

Contents

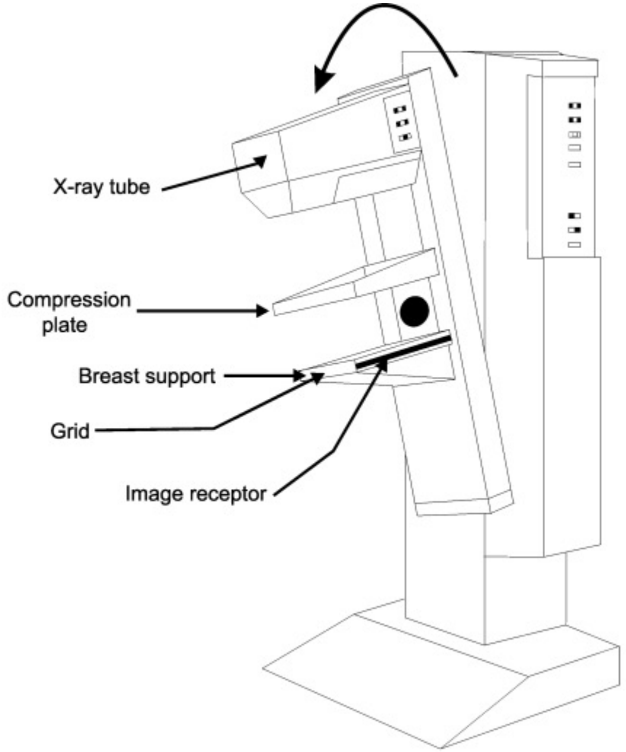
Physical principles of mammography

Mammography: low-X-ray energy examination for breast tissue characterization.
Screening program: a procedure in which asymptomatic population groups are subjected to make an early diagnosis of a high social impact and a high-risk disease.

The photons forming an x-ray beam will be absorbed in a different way depending on the tissues passing through.



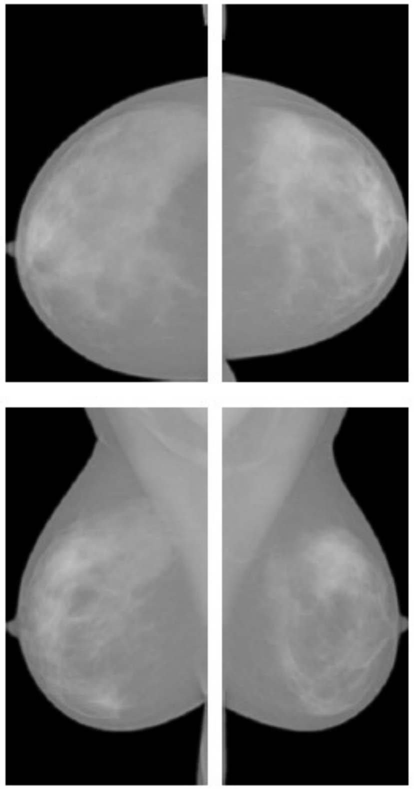
Imaging system



Mammographic exam

Right and left Cranio-Caudal (CC)

Right and left Medio-Lateral Oblique (MLO)



Breast density

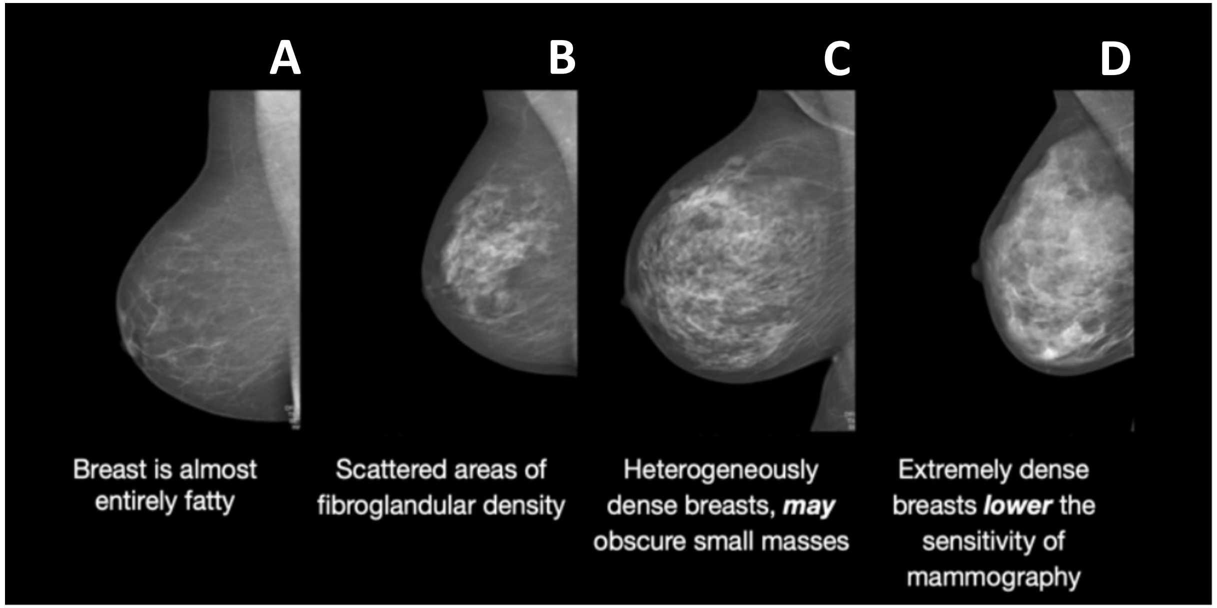
Mammographic density: the relative amount of radiodense tissue elements compared with the amount of fatty tissue elements visible on a mammographic exam.

- **Personalized dosimetric index (RADIOMA Project)**

  A. C. Traino, D. Caramella, M. E. Fantacci et al., “Average absorbed breast dose in mammography: A new possible dose index matching the requirements of the european directive 2013/59/euratom”, European radiology experimental, 2017.
- **Risk factor for breast cancer**

  K. Krishnan, L. Baglietto, et al., “Longitudinal study of mammographic density measures that predict breast cancer risk”, Cancer Epidemiology and Prevention Biomarkers, 2017.
- **Masking effect**

  BOYD, Norman F., et al. Mammographic density and the risk and detection of breast cancer, *New England Journal of Medicine*, 2007.



Density standard: fifth edition (2013) of BIRADS (Breast Imaging Reporting and Data Systems) atlas.

E. Sickles, C. D’Orsi, L. Bassett, C. Appleton, W. Berg, and E. Burnside, “Acr bi-rads atlas”, Breast Imaging Reporting and Data System, pp. 39–48, 2013.

Deep learning

Deep learning: a subset of Machine Learning (ML) and the capability of an Artificial Intelligent (AI) system to learn from experience and understand the world in terms of a hierarchy of concepts, building these concepts on top of each other in a deep graph with many layers.

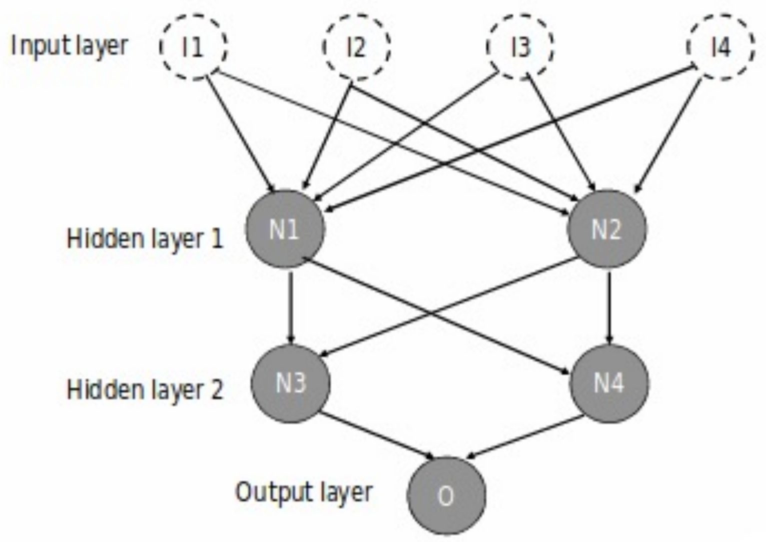
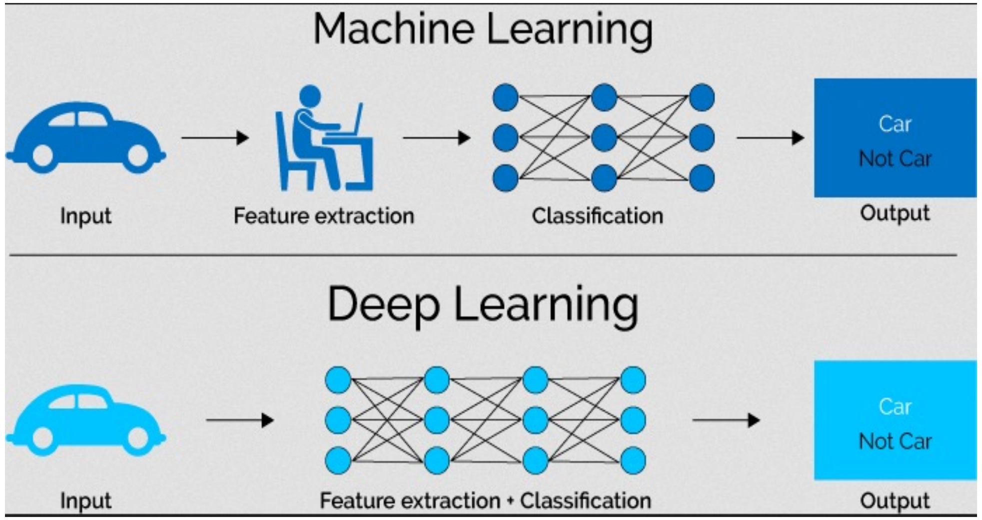


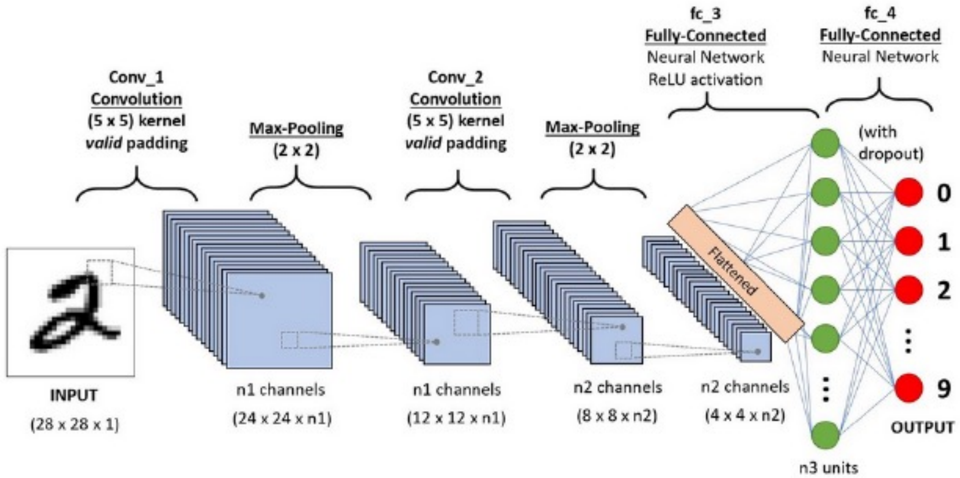
Image classification pipeline

1. Input
2. Learning
3. Evaluation



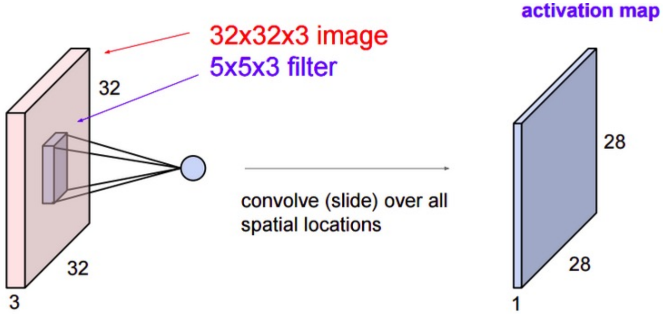
Convolutional Neural Networks

Convolutional Neural Network (CNN): a specialized kind of neural network based on convolutional layers for processing data that have a grid-like structure.

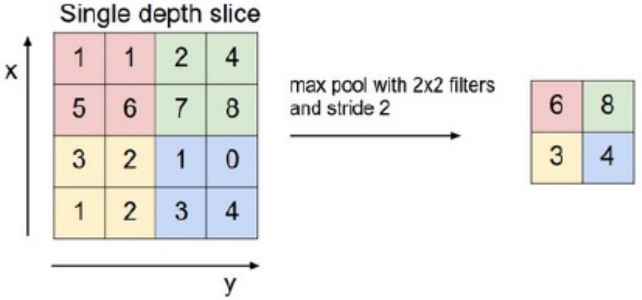


$$C_{IK} = I \otimes K = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n)$$

I: 2D image
K: 2D filter

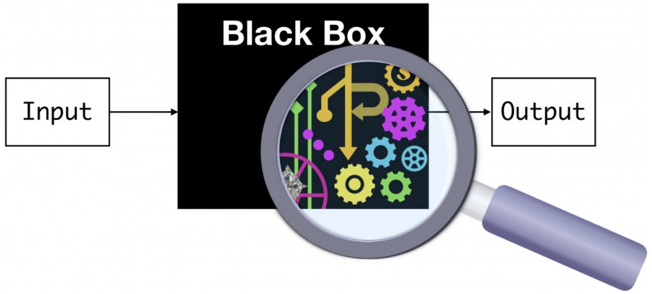


- The network will learn filters that activate when they see some type of visual feature at each layer.
- Sharing of the parameters: each neuron is connected to a part of the input. It makes the network invariant to input translation.
- Pooling layer: performs an operation (e.g. maximum) that produces a subsampled image to reduce the space dimensionality.



Explainability and Visualization

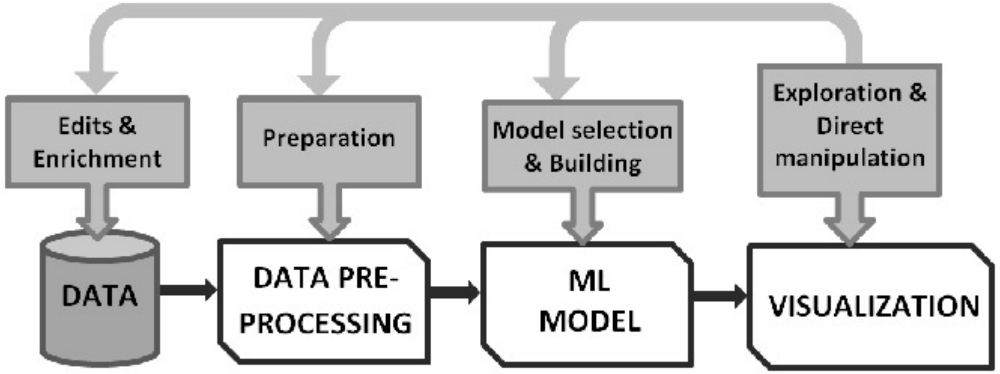
Explainability: qualitative understanding between the input and the response to build transparent models for which it is possible to explain why they predict what they predict.



Visualization: visual representation of patterns memorized by the CNN to identify salient regions that contribute most to prediction and understand the logic inside a CNN.

WHY?

- Reliability
- Algorithm selection
- Performance improvement
- Diagnosing
  - Bias in datasets
  - Misclassfied examples



Explainability and Visualization

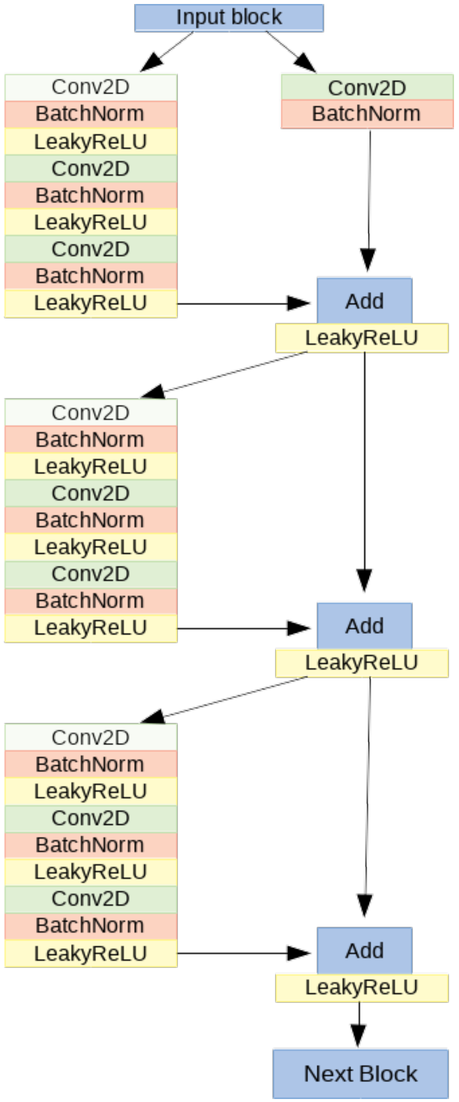
- Logistic regression classifier.
- Training set: pictures of wolves with snow in the background, pictures of huskies without snow.
- Prediction: «wolf» if there is snow, «husky» otherwise.

Correct but unreliable

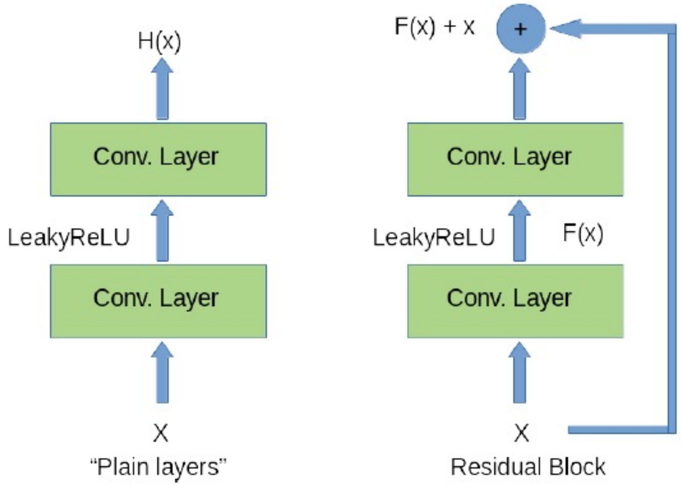
(a) Husky classified as wolf (b) Explanation

M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier", 2016.

Residual Neural Network (ResNet)



ResNet: a particular CNN made up of several residual blocks. Rather than learning a function, the residual block only learns the residual.



$$H(x) = F(x) + x \rightarrow F(x) = H(x) - x$$

Hyperparameters

- 41 conv layers in 12 similar blocks
- Training in batch of 4 images for 100 epochs
- Loss function: Categorical Cross-Entropy
- Optimizer: SGD
- Regolarization: Batch Normalization
- Learning rate = 0.1, Decay = 0,1, Patience = 15, Monitor = validation loss

F. Lizzi, et al. "Residual Convolutional Neural Networks to Automatically Extract Significant Breast Density Features." *International Conference on Computer Analysis of Images and Patterns*. Springer, Cham, 2019.

Software and Hardware

Keras: API written in Python on top of Tensorflow framework to train, fit and evaluate the CNN.

Hardware available by “Istituto Nazionale di Fisica Nucleare” (INFN) **:**

- CPUs: 2x 10 cores Intel Xeon E5-2640v4 @2.40 GHz
- RAM: 64 GB
- GPUs: 8x nVidia Tesla K80, with 2x GPUs Tesla GK210, 24 GB RAM and 2496 CUDA cores each

Data

DL : data-driven approach → Lack of huge public mammograms dataset

Collected dataset: 1962 mammographic exams made us available by the “Azienda Ospedaliero-Universitaria Pisana” (AOUP) and collected by a radiologist, specialized in mammography, and a radiology technician.

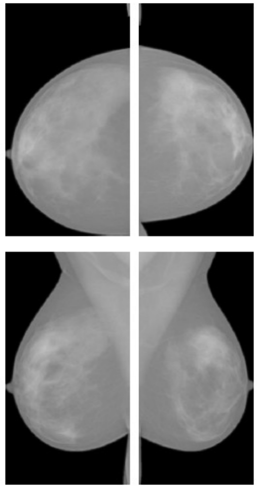
AZIENDA OSPEDALIERO UNIVERSITARIA PISANA

- Each exam composed of four images (CC, MLO) (right, left).
- Exams acquired with 4 different mammographic systems.
- Negative exam reports.
- Ground truth: density class (A, B, C, D) label assigned to each exam by a radiologist.
- DICOM format.

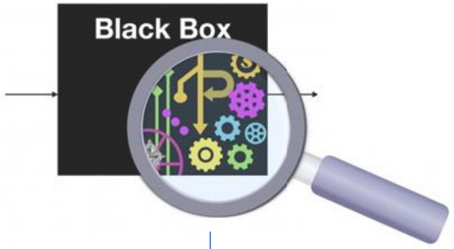
	A	B	C	D
N. of exams	264	611	888	199
Average age	67	63	58	53
Standard deviation	11	11	9	6
Median	68	62	56	52

Goals and methods

INPUT: Mammographic exam



Deep Residual CNN classifier



OUTPUT: Breast density class

A
B
C
D

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

Goal: explain the classifier behaviour and interpret its internal processes.

To assess trust and optimize the performance

Methods: Model interpreted a posteriori

1. How the output varies with the input

2. Off-line visualization

Goals and methods

1. How the output varies with the input

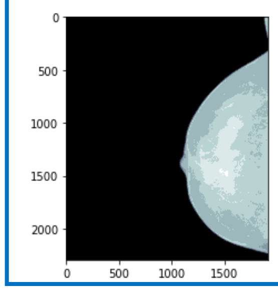
- ❑ Preprocessing
  - Preparatory steps
  - Results
  - Pectoral muscle segmentation
- ❑ CNN optimization
  - Training and generalization
  - Model fine-tuning: Dropout
  - Number of channels
  - Dataset distribution

2. Off-line visualization

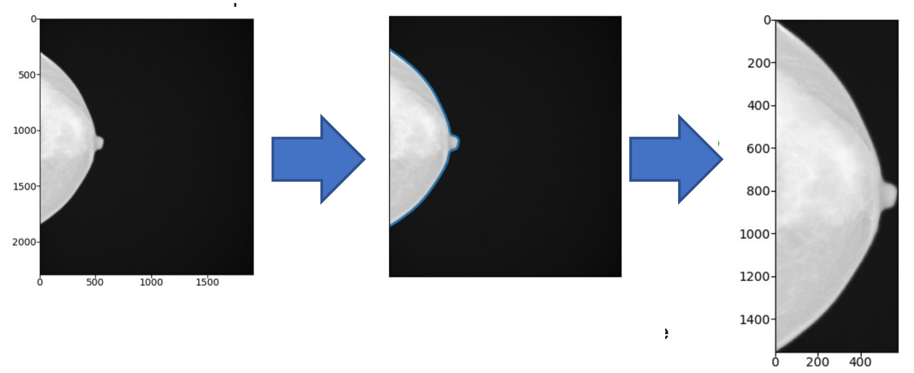
- ❑ Explainability
  - Filters and Feature maps
  - Heatmaps (grad-CAMs)

Preprocessing – Preparatory steps

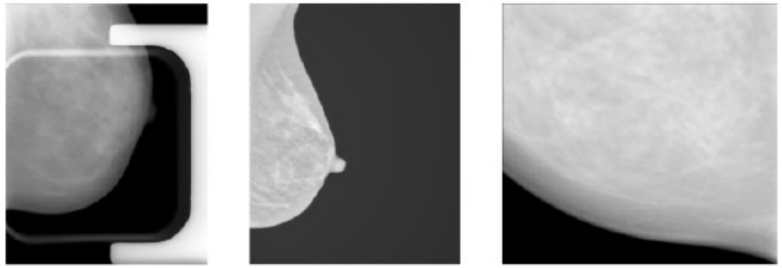
1. Decompression
2. PNG conversion
3. 8 bits conversion
4. Background removal
5. One by one inspection
6. «Real-time» Data Augmentation



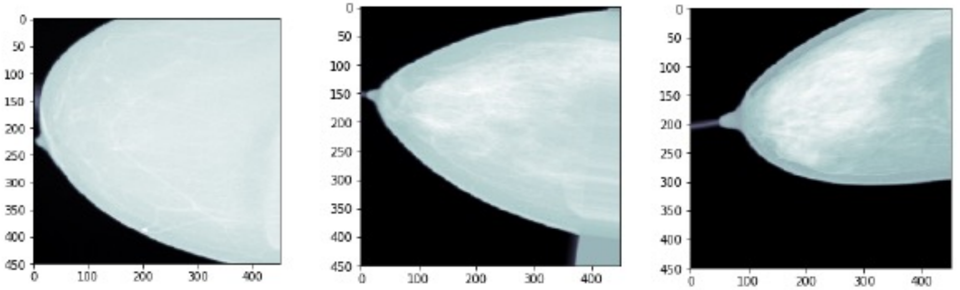
$$img8 = 256 * \frac{img - bit\_min}{bit\_max - bit\_min}$$



Problematic images



Zoom
Shift
Rotation



Preprocessing – Results

- Properly data preparation and exclusion of problematic exams.

Mammographic acquisition system	Original dataset size (No. of exams)	Pre-processed dataset size (No. of exams)
GIOTTO IMAGE SDL	232	232
SELENIA DIMENSIONS	50	49
GE Senograph DS VERSION ADS 54.11	121	116
GE Senograph DS VERSION ADS 54.11	1561	1546

- Improvement in the classifier performance.

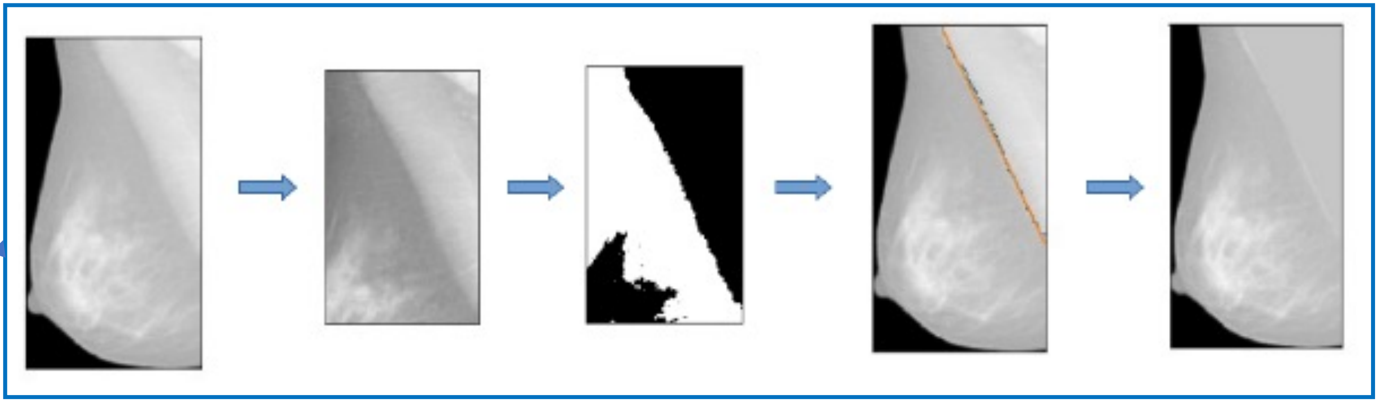
	First release (1561 exams)	With preprocessing (1546 exams)	Wu et al. (200000 exams)	Wu et al. (2000 exams)
test accuracy (%)	75.3	**83.1**	76.7	72.9
recall (%)	72.1	**80.1**		
precision (%)	76.4	**87.9**		

Wu, et al. "Breast density classification with deep convolutional neural networks.", 2018.

Pectoral muscle segmentation

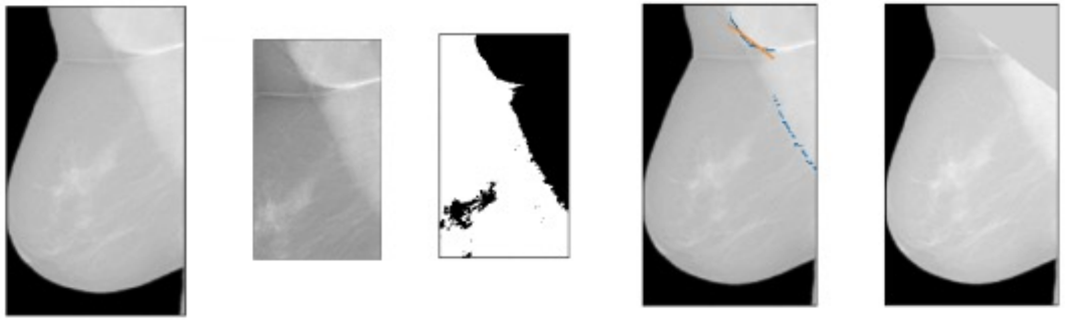
Pectoral muscle: mass of tissue on which the breast rests, it turns up in MLO mammograms views. It has pixel intensities and texture similar to that of breast dense tissues.

- Different instruments and settings.
- It doesn't happen in all the images.
- Extremely variable in size, intensity, shape and texture.



Segmentation algorithm

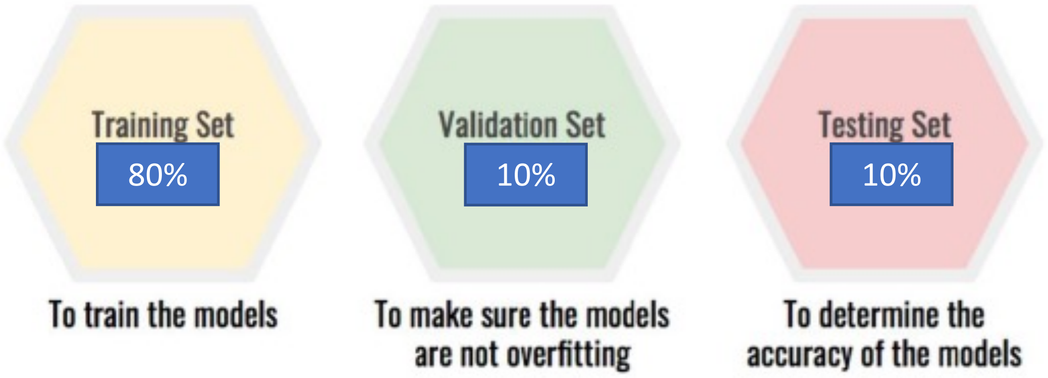
Step 1) View selection
Step 2) 8-bit transformation
Step 3) Background removal
Step 4) ROI detection
Step 5) Noise reduction
Step 6) ROI binarization and mask
Step 7) Edge coordinates and linear fitting
Step 8) Mean grey level replacement

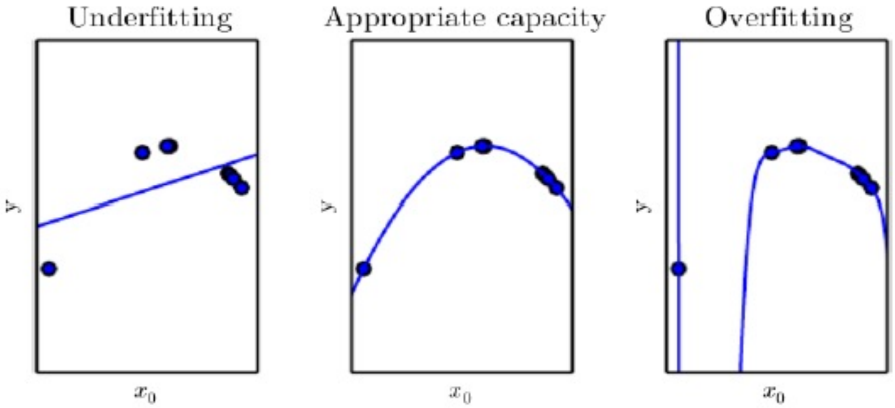


About 300 problematic images segmented «by hand»

Training and generalization

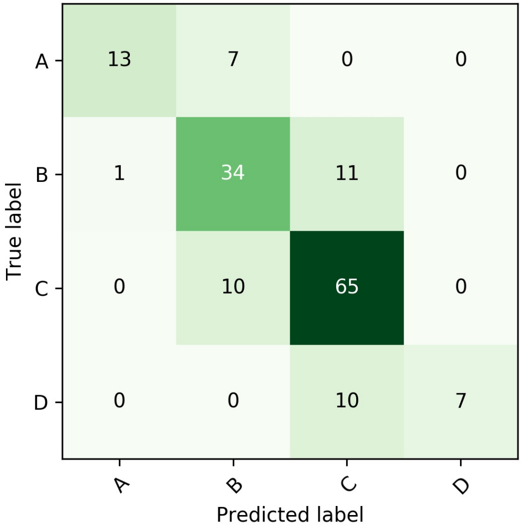
Supervised learning: each example is a *pair* consisting of an input object and a desired output value. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.





Generalization: the ability to predict the right output on unobserved inputs.

Training and generalization



Four different CNNs per projection

Average of classification scores of last layers

Performance metrics

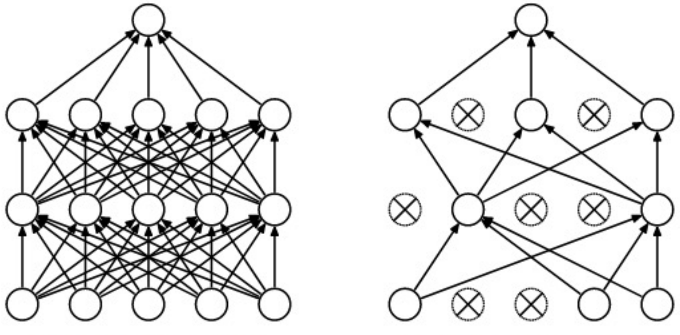
ACCURACY → ratio of correctly predicted observations to the total observations

RECALL → ratio of correctly predicted positive observations to the all observations in actual class

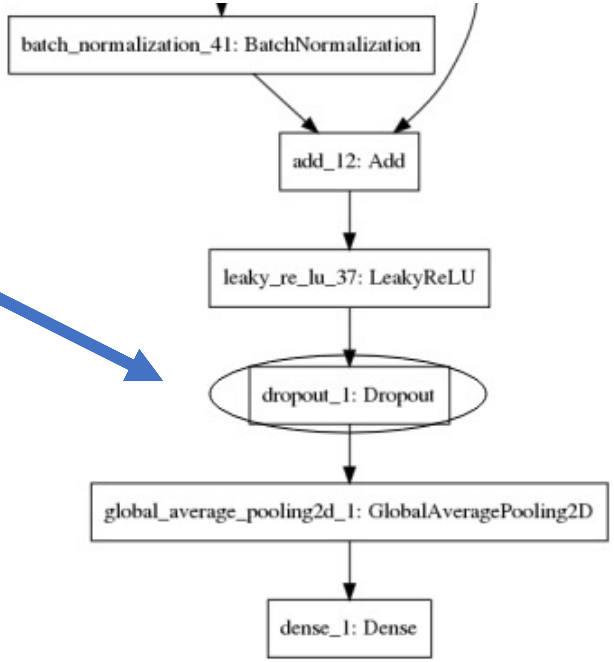
PRECISION → ratio of correctly predicted positive observations to the total predicted positive observations

Model fine-tuning: Dropout

Dropout: regularization method that consists in randomly setting of a fraction rate of input units to 0 at each update during training time, to prevent the model from overfitting.



Insertion of a Dropout layer at the end of the network.

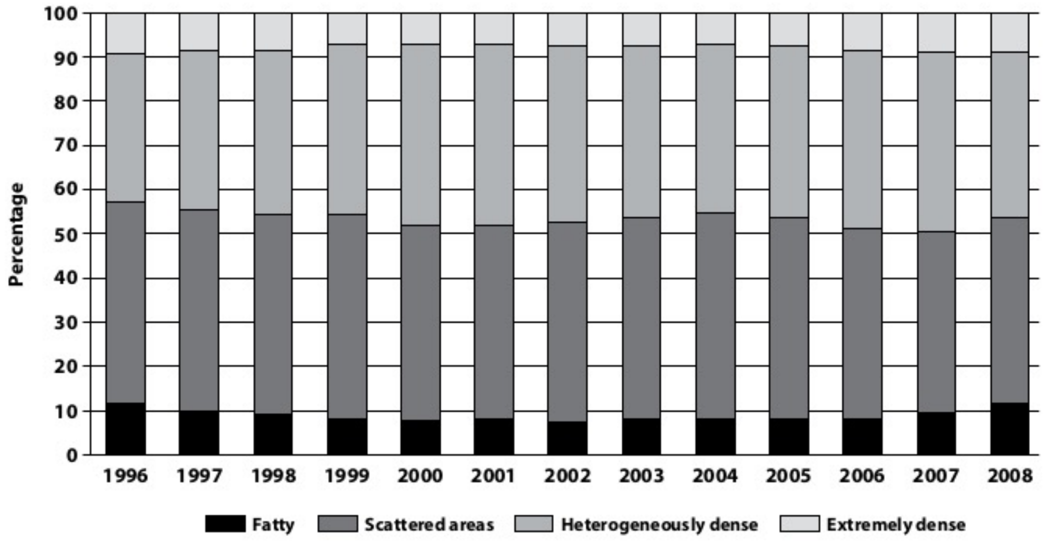


		No Dropout	With Dropout
450x450	test accuracy (%)	77.1	83.1
	recall (%)	71.7	80.1
	precision (%)	84.6	87.9
650x650	test accuracy (%)	77.1	78.8
	recall (%)	76.3	77.4
	precision (%)	74.5	79.3
850x850	test accuracy (%)	72.9	79.7
	recall (%)	72.1	76.4
	precision (%)	72.4	84.4

Dataset distribution

TRANSFERABILITY

AOUP distribution: A: 12%, B:28%, C:50%, D:10% → Clinical database
BIRADS distribution: A: 10%, B:40%, C:40%, D:10% → Screening practice
Uniform distribution: A: 25%, B:25%, C:25%, D:25% → Further comparison



BIRADS density class distribution calculated on 3,865,070 screening mammography examinations over 13 years (1996-2008).

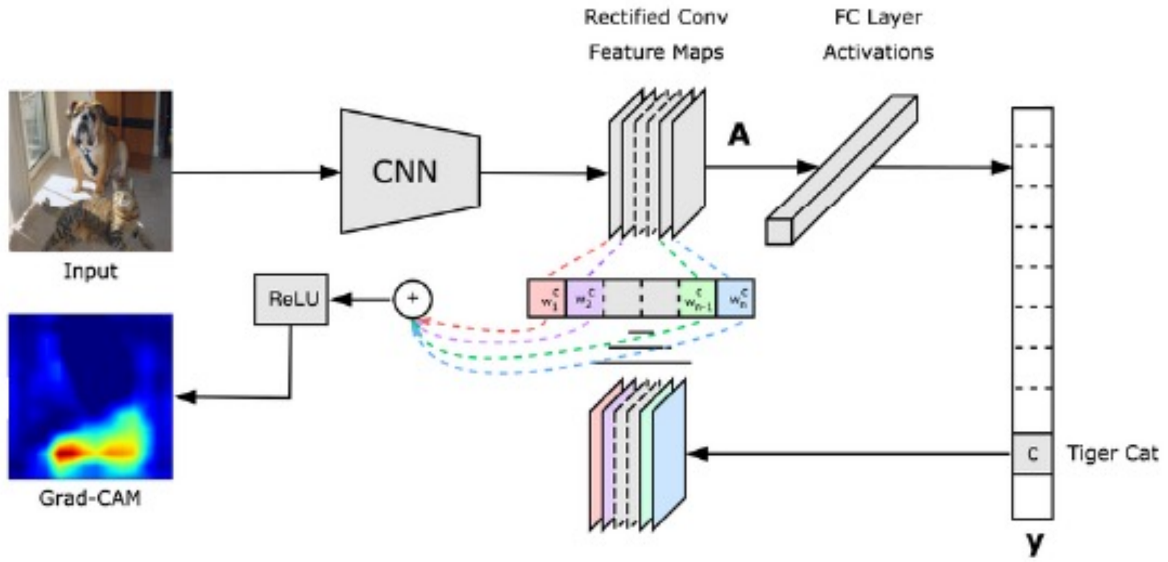
AOUP Training set				
		AOUP Test set	BIRADS Test set	Uniform Test set
ResNet1	test accuracy (%)	76.6	76.3	72.2
	recall (%)	72.0	75.8	72.2
	precision (%)	75.7	74.2	77.5
ResNet2	test accuracy (%)	75.3	73.7	73.6
	recall (%)	71.6	72.7	73.6
	precision (%)	77.2	74.6	80.3
ResNet3	test accuracy (%)	78.5	79.7	73.6
	recall (%)	74.2	77.9	73.6
	precision (%)	81.2	83.0	79.4

BIRADS Training set				
		AOUP Test set	BIRADS Test set	Uniform Test set
ResNet1	test accuracy (%)	75.3	77.1	65.3
	recall (%)	66.7	71.7	65.3
	precision (%)	83.8	84.6	74.7
ResNet2	test accuracy (%)	76.6	80.5	75.0
	recall (%)	71.9	76.9	75.0
	precision (%)	78.3	81.2	80.4
ResNet3	test accuracy (%)	79.1	83.1	73.6
	recall (%)	75.2	80.1	73.6
	precision (%)	82.6	87.9	79.0

Heatmaps (grad-CAM)

Heatmap: for a particular category indicates which regions of an image are being used by the model for discrimination among classes.

- To check the classifier behavior
- To improve the classifier performance



Gradient based Class Activation Map (grad-CAM): gradient calculation of the final classification score with respect to the final convolutional layer.

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

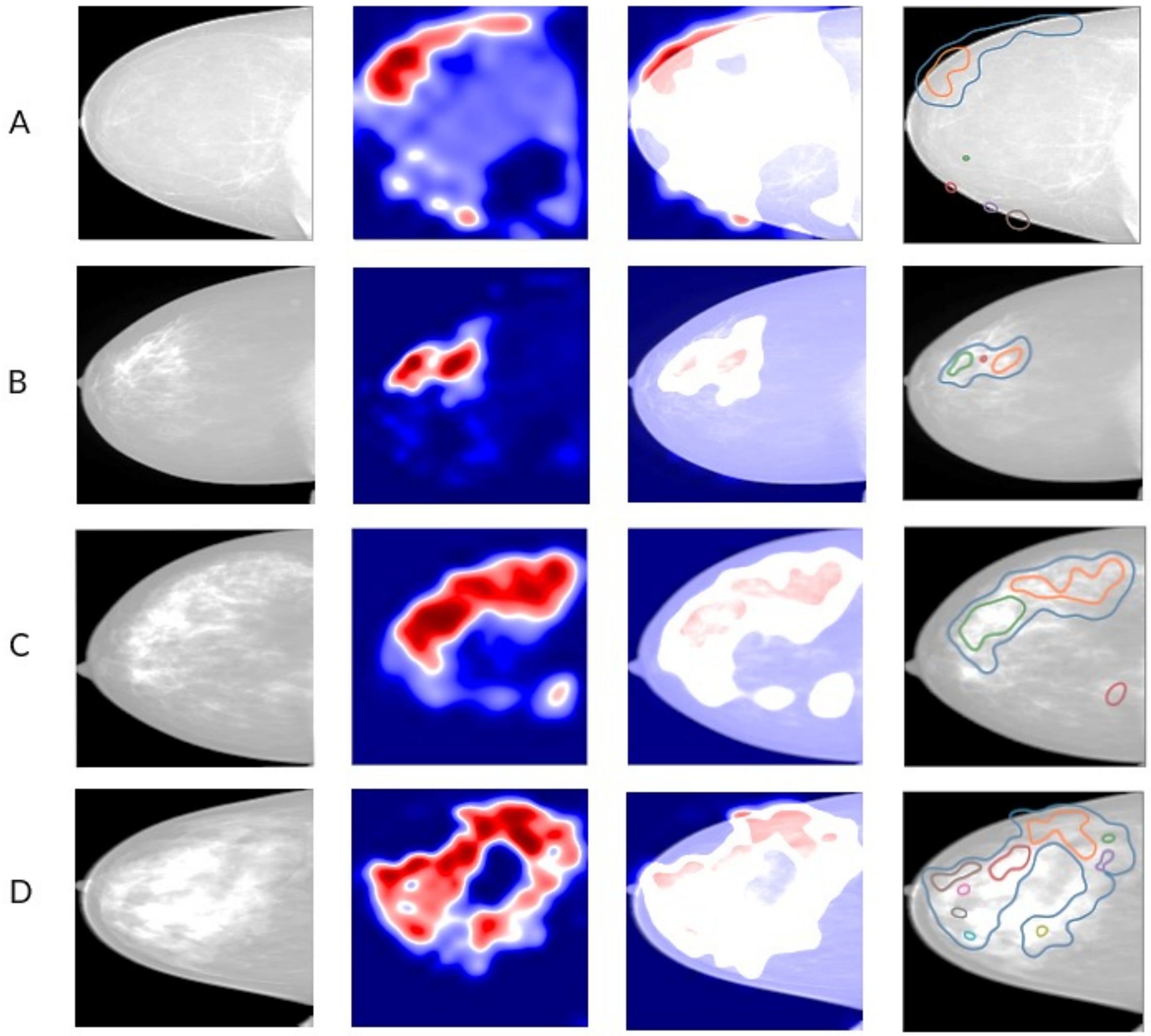
α^c_k : weights of the final dense layer

c : predicted class

A^k : feature maps of the last conv layer

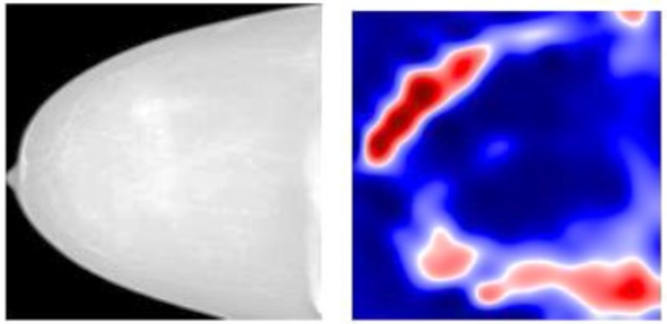
Grad-CAM: Visual Explanationsfrom Deep Networks via Gradient-based Localization - R. R. Selvaraju et al. –2019 -ArXiv

Heatmaps (grad-CAM)



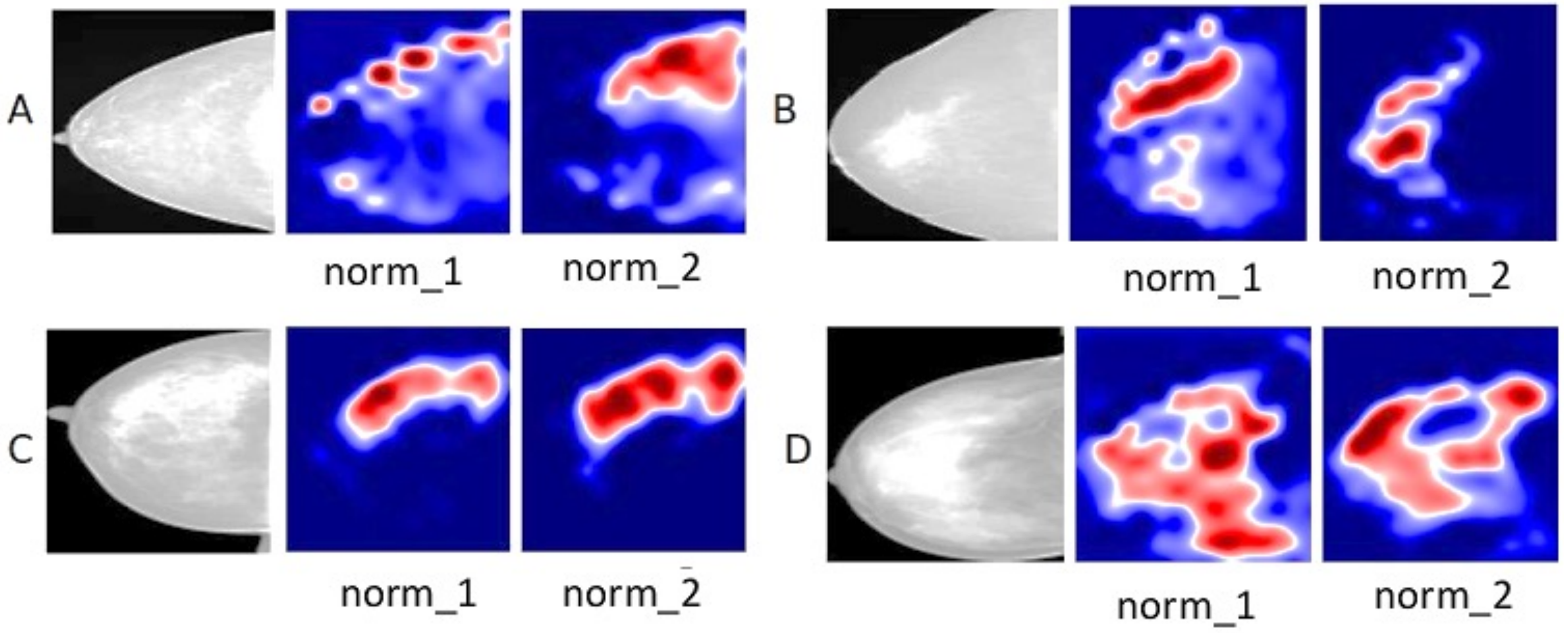
Qualitative evaluation: observing if they activate at the densest areas of the breast.

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



Explainability – Heatmaps (grad-CAM)

Normalization comparison



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.

norm_1 : test accuracy = 79.7%, recall = 79.5%, precision = 78.3%

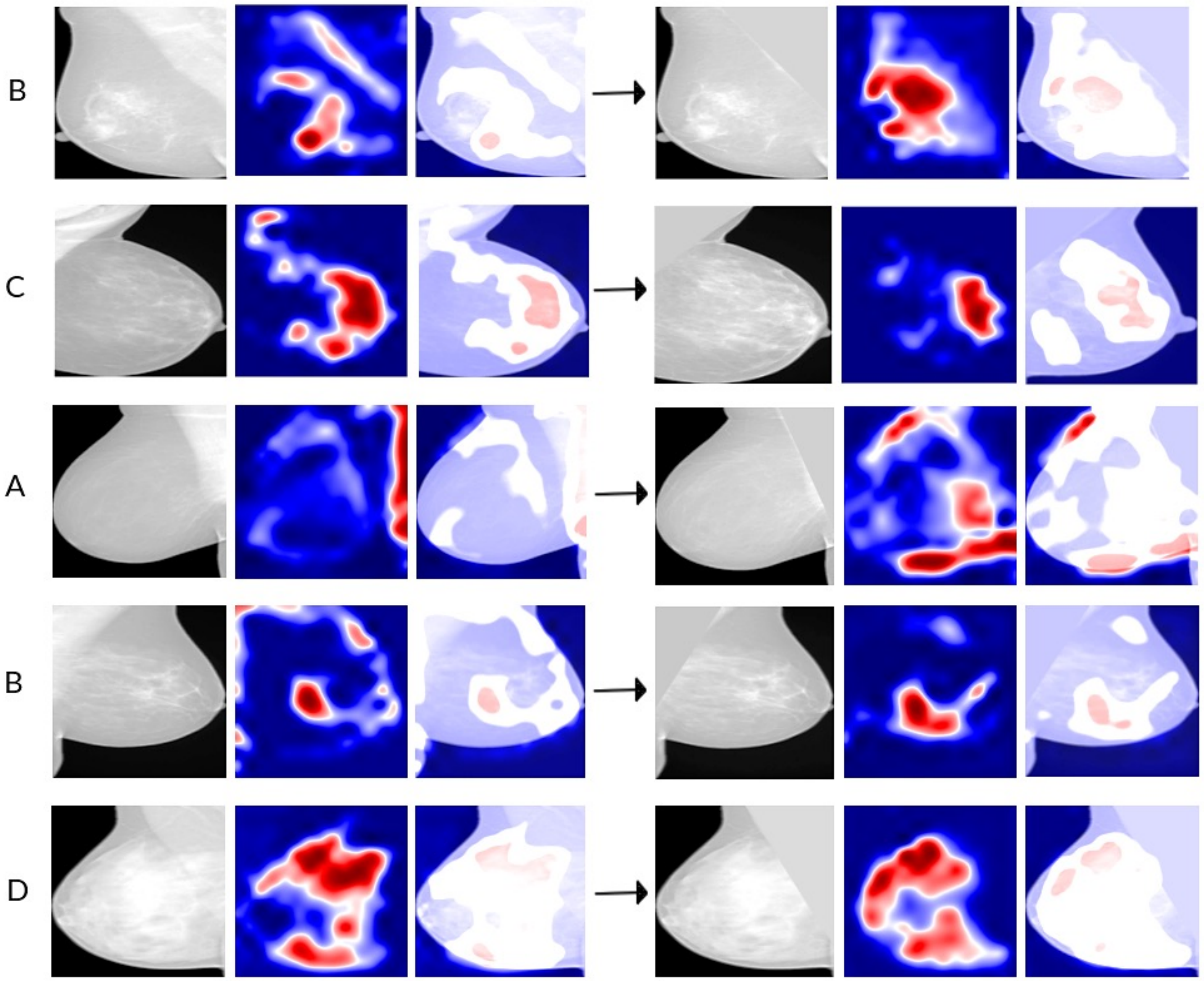
norm_2 : test accuracy = 82.2%, recall = 78.0%, precision = 89.7%

Explainability – Heatmaps (grad-CAM)

Pectoral muscle segmentation

Non-segmented mammograms:
test accuracy = 79.9%,
recall = 78.1%,
precision = 81.1%

Segmented mammograms:
test accuracy = 82.0%,
recall = 80.3%,
precision = 83.3%



Conclusions

- A better understanding of how the developed BIRADS classifier works.
- Which factors most affect the classifier performance and accuracy results (data preparation, model architecture, classes distribution)
- Systematic improvement of performance measures of accuracy, recall and precision.
- Assessed trust in the model.
- Since it does not exist a well-established method for explainability, the work here described can be a starting point for a further study.

Outlook

- Ground truth: maximum agreement between more than one radiologist and segmented images.
- Dataset: increasing dataset size and more exams acquired with different mammographic systems.
- To test other training conditions.
- To transform the model in a mixed and controlled classifier.
- CAMs used as region proposal for features calculation, also for other models of breast density classifier. This region proposal could also be used on tumors after fine-tuning.

Camilla Scapicchio - PhD Student in Physics – University of Pisa

camillascapicchio@gmail.com
camilla.scapicchio@phd.unipi.it

GRAZIE PER L'ATTENZIONE!

RADIOMA

INFN
Istituto Nazionale di Fisica Nucleare

artificial intelligence in medicine
AIM

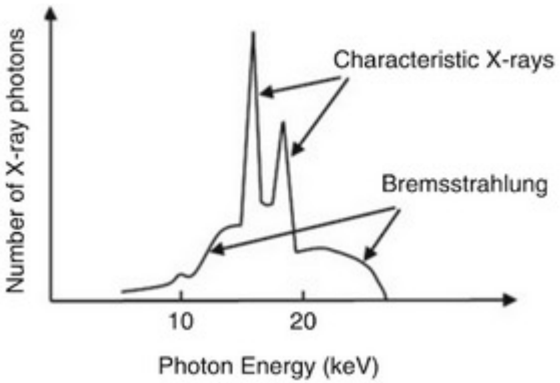
Backup Slides

Camilla Scapicchio

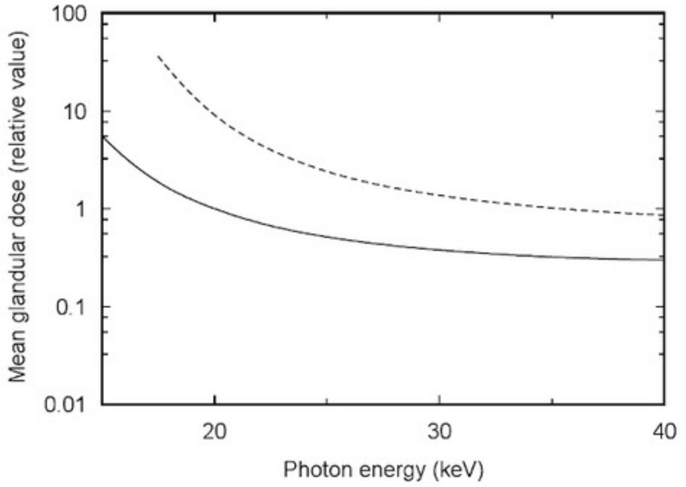
Università di Pisa

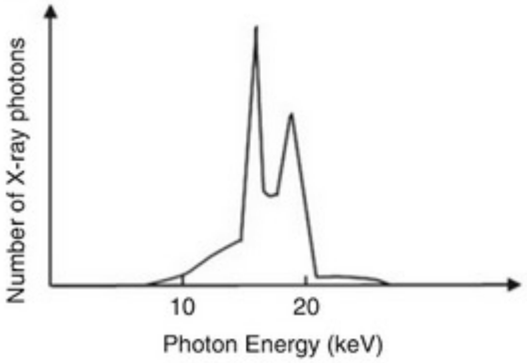
25 Marzo 2021

Photon energy and Dose

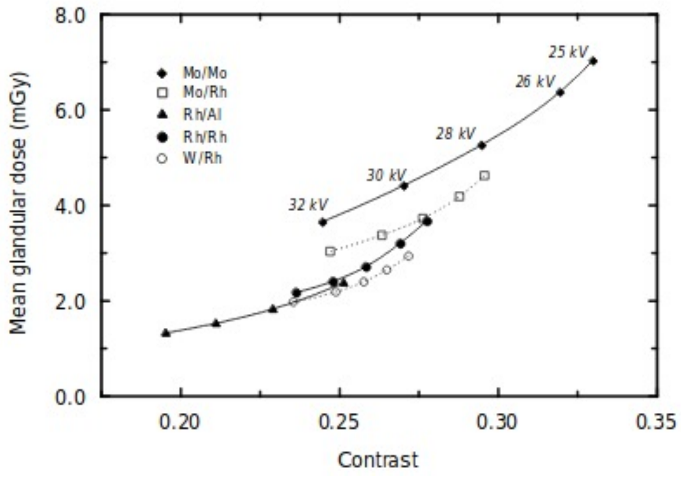


X-ray energy spectrum for a molybdenum anode





X-ray energy spectrum for a molybdenum anode with a molybdenum filter



Mammographic systems and resolution

- GIOTTO Image SDL: matrix 2816 x 3584 pixel, pixel size 85 μm,
  spatial resolution 6 lp/mm
  (230 exams)

- Selenia Dimensions (Hologic): matrix 3328 x 4096 pixel, pixel size 70 μm,
  spatial resolution 2D 7.1 lp/mm, 3D 3.5 lp/mm
  (50 exams)

- GE Senograph DS: : matrix 2294 x 1914 pixel, pixel size 100 μm,
  spatial resolution 5 lp/mm
  (VERSION ADS 54.11: 121 exams, VERSION ADS 53.40: 1561 exams)

Deep learning in medical imaging

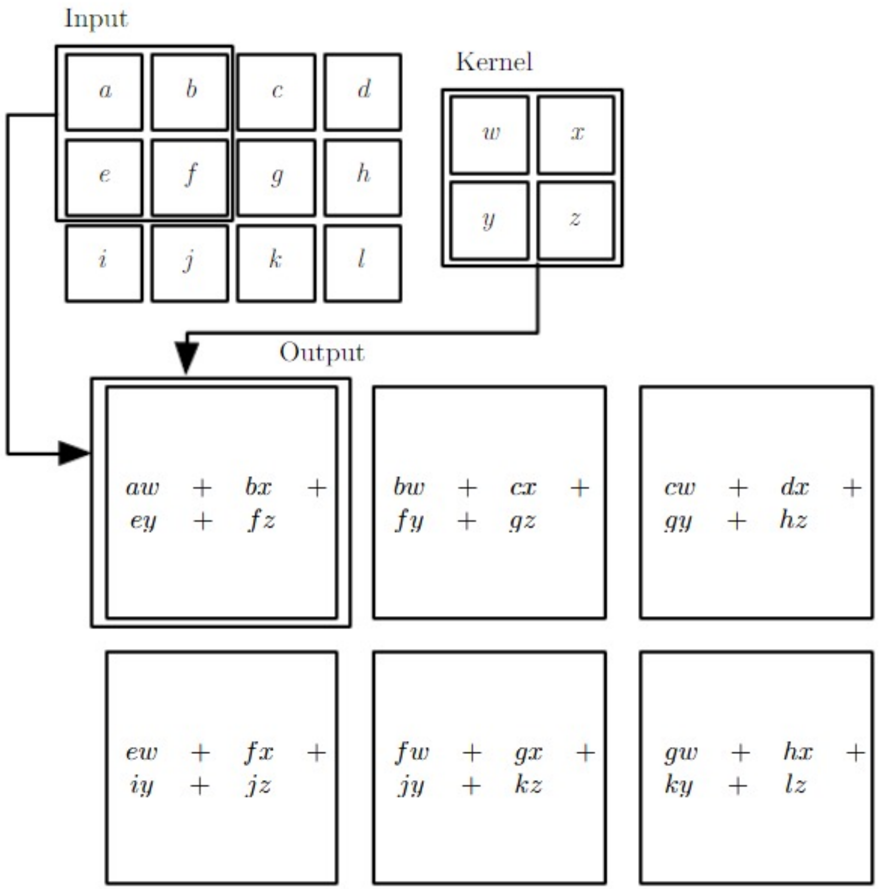
Dataset problem: Very few large public data sets available.

- Legal and ethical issues regarding the use of clinical imaging data.
- Specific annotations for the image data require domain expert.
- Label noise: there is no consensus among the radiologists.

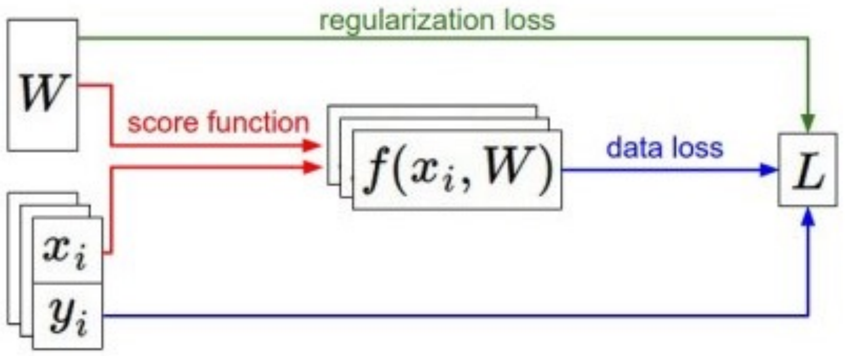
Convolution

$$C_{IK} = I \otimes K = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n)$$

Discrete convolution can be viewed as multiplication by a matrix.

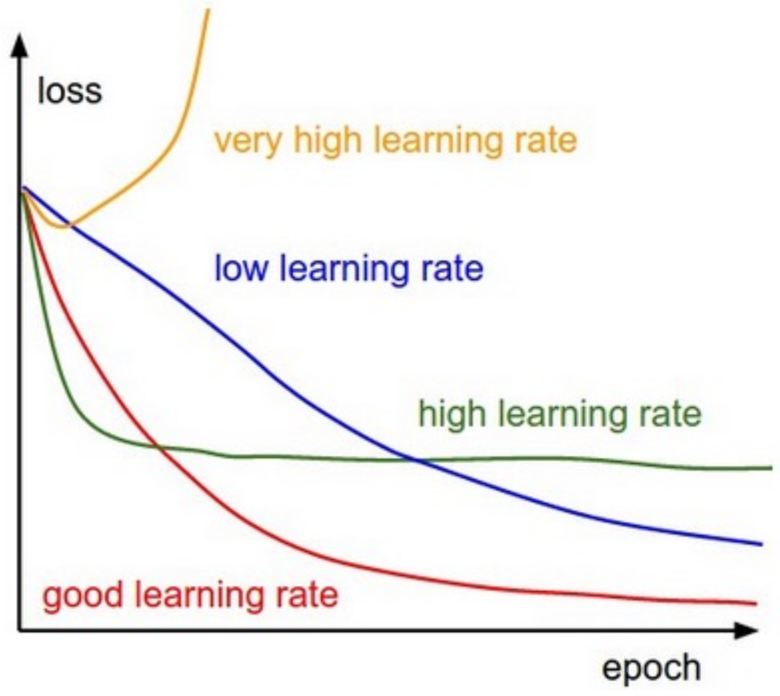


Optimization



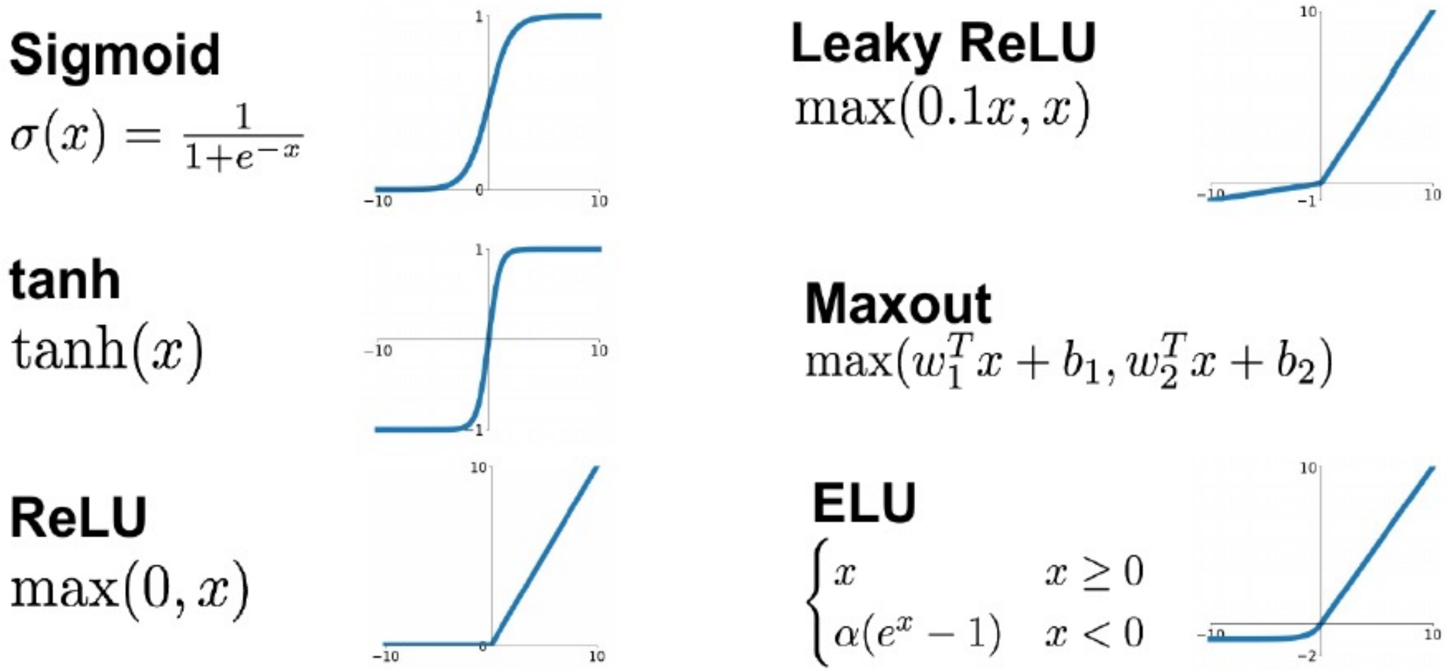
The dataset (x,y) is fixed.
The weights start out as random numbers and can change.
The score function computes class scores, stored in vector f.
The loss function contains: 1) The data loss, which computes the compatibility between the scores f and the labels y and 2) The regularization loss, which is only a function of the weights.
During Gradient Descent, we compute the gradient on the weights and use them to perform a parameter update.

Learning rate



Activation functions

Input and activation functions produce the single neuron output.



Stochastic Gradient Descent (SGD)

SGD is a Gradient Descent (GD) algorithm semplification.

$$w_{t+1} = w_t - \gamma \frac{1}{n} \sum_{i=1}^{N} \nabla_w E(z_i, w_t)$$

The gradient is calculated from a single random example z_t.

$$w_{t+1} = w_t - \gamma_t \nabla_w E(z_t, w_t)$$

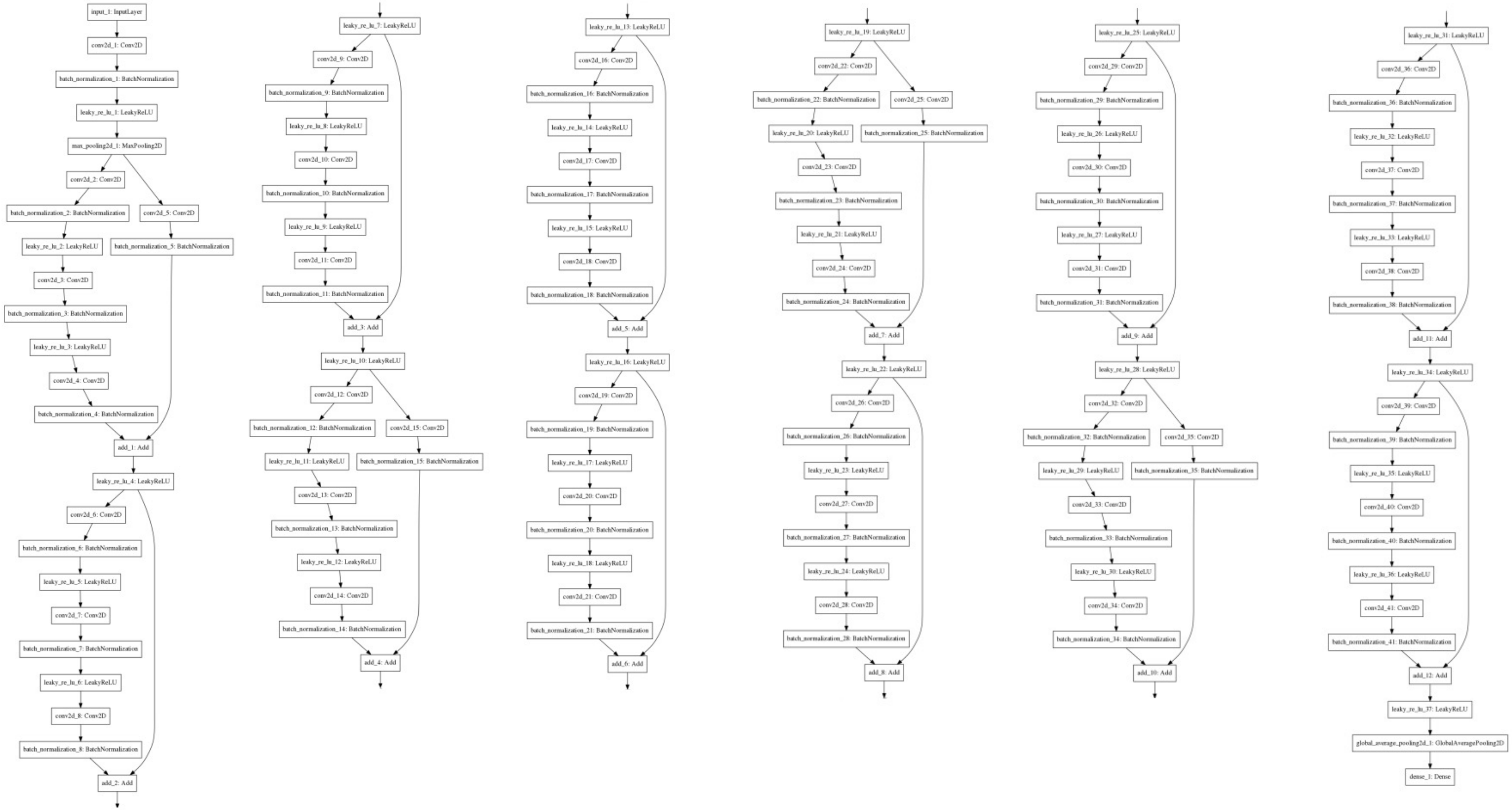
Backpropagation algorithm

Recursive application of «chain rule» in the graph.

$\boldsymbol{x} \in \mathbb{R}^m, \boldsymbol{y} \in \mathbb{R}^n, g : \mathbb{R}^m \to \mathbb{R}^n, f : \mathbb{R}^n \to \mathbb{R}$. If $\boldsymbol{y} = g(\boldsymbol{x})$ and $z = f(\boldsymbol{y})$ then:

$$\frac{\partial z}{\partial x_i} = \sum_j \frac{\partial z}{\partial y_j} \frac{\partial y_j}{\partial x_i}$$

ResNet architecture



Datasets

The exact number of exams within each dataset.

	Training set	Validation set	Test set
AOUP	1232 exams (A:142, B:337, C:611, D:142)	156 exams (A:20, B:45, C:74, D:17)	158 exams (A:20, B:46, C:75, D:17)
BIRADS	842 exams (A:84, B:337, C:337, D:84)		118 exams (A:12, B:47, C:47, D:12)
Uniform	564 exams (A:141, B:141, C:141, D:141)		72 exams (A:18, B:18, C:18, D:18)

	Training set	Validation set	Test set
Original dataset	924	156	134
New segmented dataset	910	153	128

Summary of performance improvement

Best performance metrics obtained as a result of the described analyses.

	First release	New results
test accuracy (%)	77.3	**83.1**
recall (%)	77.1	**80.1**
precision (%)	78,6	**87.9**

F. Lizzi et al., “Residual convolutional neural networks for breast density classication”, in Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies -Volume 3: BIOINFORMATICS, INSTICC, SciTePress, 2019, pp. 258–263.

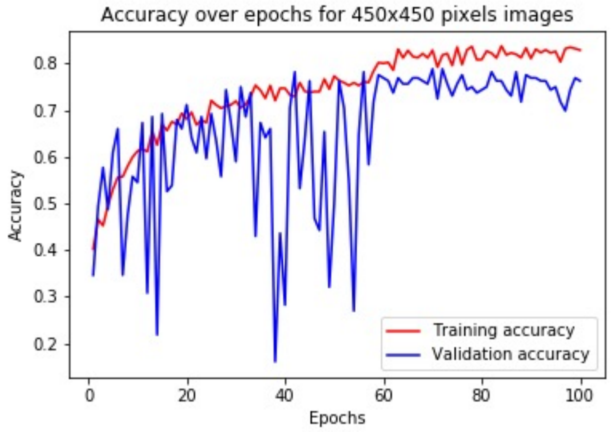
WARNING: with what reliability?

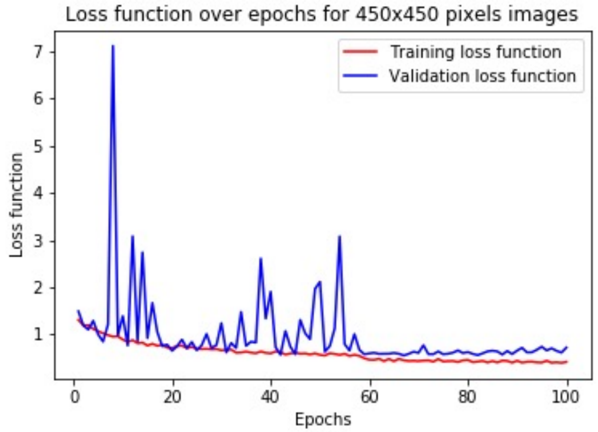
Wu, Nan, et al. "Breast density classification with deep convolutional neural networks." , 2018.

data	macAUC	top-1	top-2	top-3	superclass
1%	0.888	0.729	0.967	0.998	0.849
10%	0.907	0.745	0.976	0.999	0.856
100%	0.916	0.767	0.982	0.999	0.865

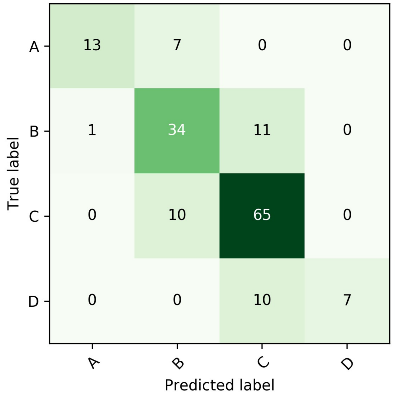
ResNet1

Input: 450x450x1
Without Dropout



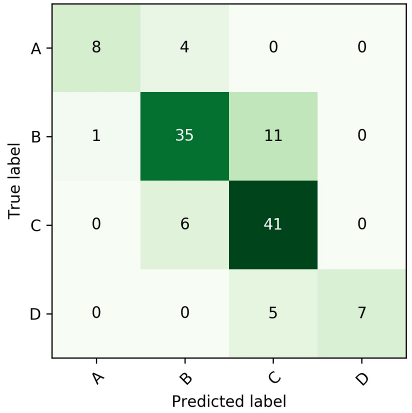


Test set: AOUP



	CC_R	MLO_R	CC_L	MLO_L	Right	Left	All
validation accuracy (%)	78.8	76.9	77.6	74.4			
test accuracy (%)	72.2	74.0	72.8	69.6	76.6	73.4	**75.3**
recall (%)	64.7	67.4	68.3	57.3	67.3	66.4	**66.7**
precision (%)	77.2	75.6	78.8	76.4	84.8	79.4	**83.8**
BV epoch	70	62	86	17			

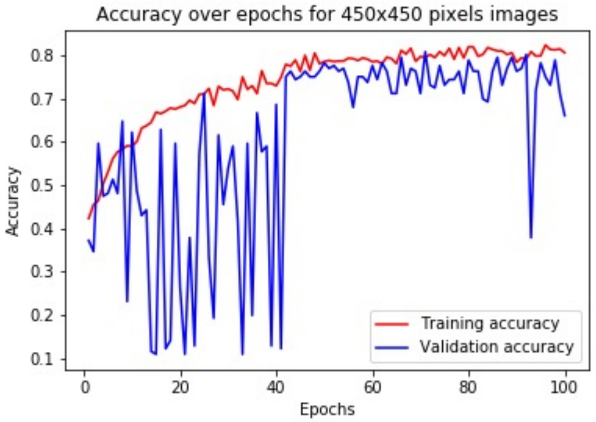
Test set: BIRADS

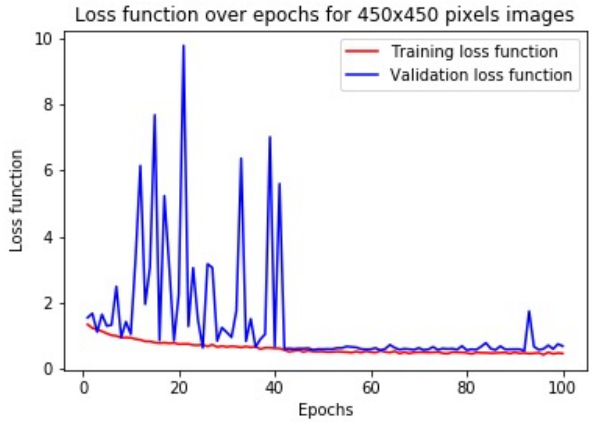


	CC_R	MLO_R	CC_L	MLO_L	Right	Left	All
validation accuracy (%)	78.8	76.9	77.6	74.4			
test accuracy (%)	72.9	75.4	74.6	68.6	79.7	74.6	**77.1**
recall (%)	70.6	70.6	71.6	61.7	74.8	70.1	**71.7**
precision (%)	77.8	76.6	78.4	76.2	86.7	78.5	**84.6**
BV epoch	70	62	86	17			

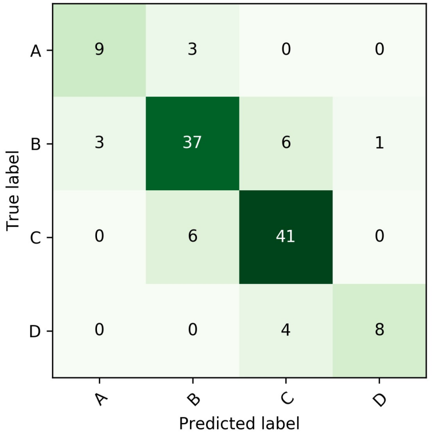
ResNet2

Input: 450x450x3
Without Dropout





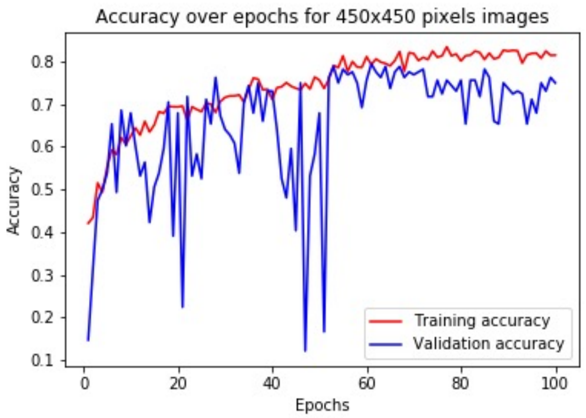
Test set: BIRADS

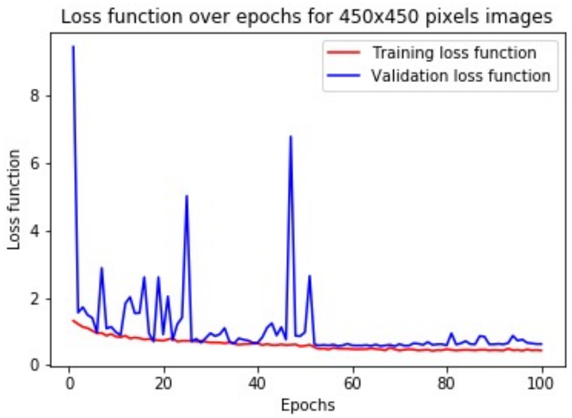


	CC_R	MLO_R	CC_L	MLO_L	Right	Left	All
validation accuracy (%)	80.8	78.8	78.8	78.2			
test accuracy (%)	77.1	71.2	74.6	76.3	77.1	78.8	**80.5**
recall (%)	77.9	71.1	74.7	74.2	76.3	75.8	**76.9**
precision (%)	78.7	69.2	76.6	80.0	78.0	83.2	**81.2**
BV epoch	71	28	88	80			

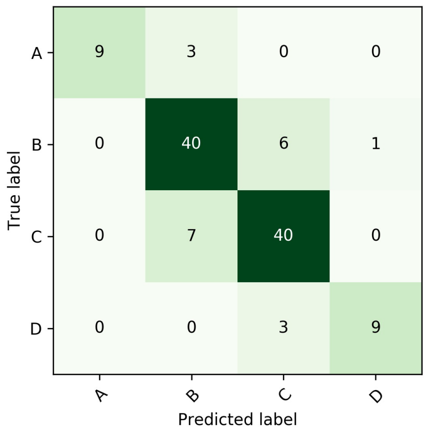
ResNet3

Input: 450x450x1
With Dropout





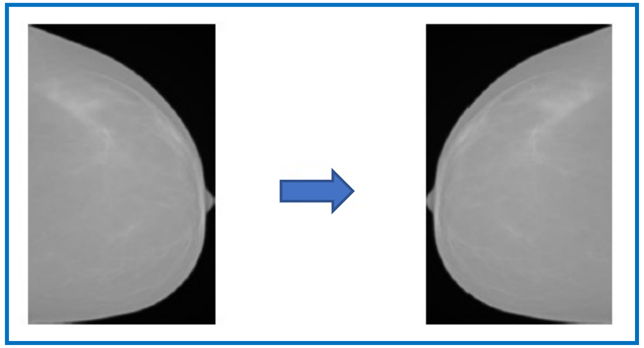
Test set: BIRADS



	CC_R	MLO_R	CC_L	MLO_L	Right	Left	All
validation accuracy (%)	79.5	76.5	76.9	76.3			
test accuracy (%)	80.5	74.6	74.6	74.6	78.8	78.0	**83.1**
recall (%)	81.6	74.7	74.7	70.1	77.4	73.8	**80.1**
precision (%)	81.7	73.0	78.3	74.9	81.5	81.6	**87.9**
BV epoch	61	73	87	98			

CNN opt. – Dataset size and Robustness

Horizontal flip for Data Augmentation



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

	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

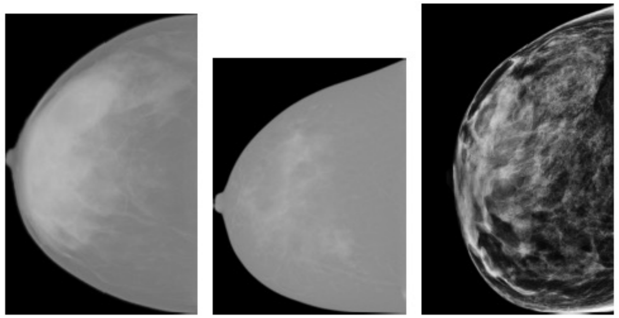
Without flip
(924 images)

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

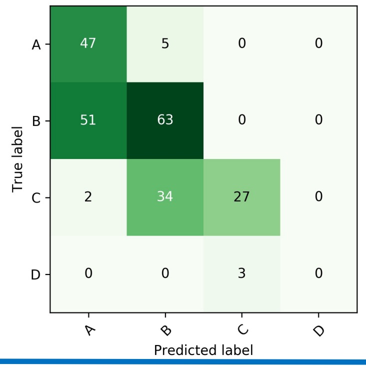
With flip
(1848 images)

Test on a different mammographic system

- Small dataset size
- Different appearance



Class C:
Senograph
Giotto
Hologic



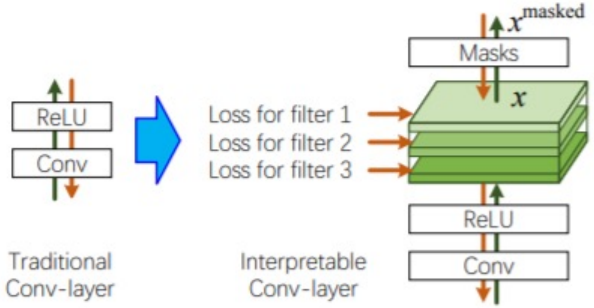
	GIOTTO
test accuracy (%)	59.1
recall (%)	47.1
precision (%)	49.7

Interpretable models

Interpretable Convolutional Neural Networks,
Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu
University of California, Los Angeles

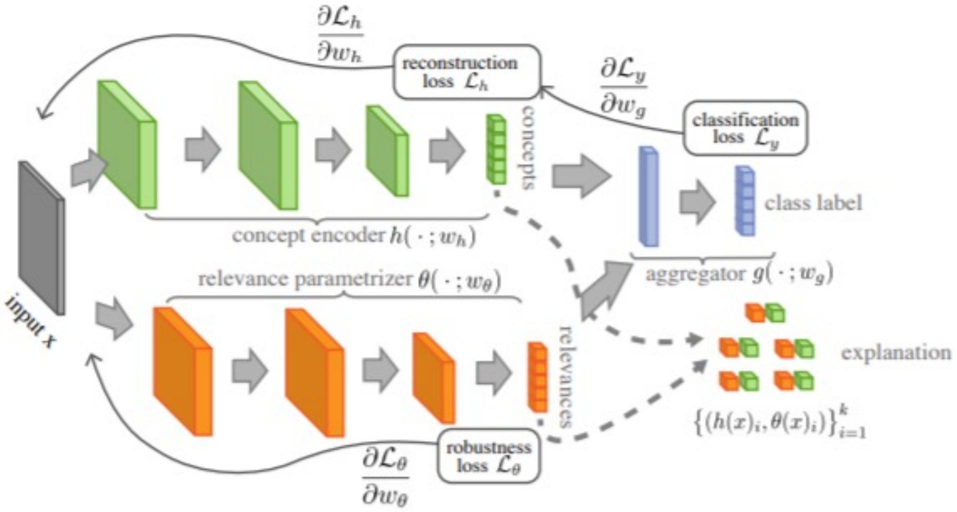
The interpretable CNN automatically assigns each filter in a high conv-layer with an object part during the learning process.

This paper proposes a method to modify a traditional convolutional neural network (CNN) into an interpretable CNN.

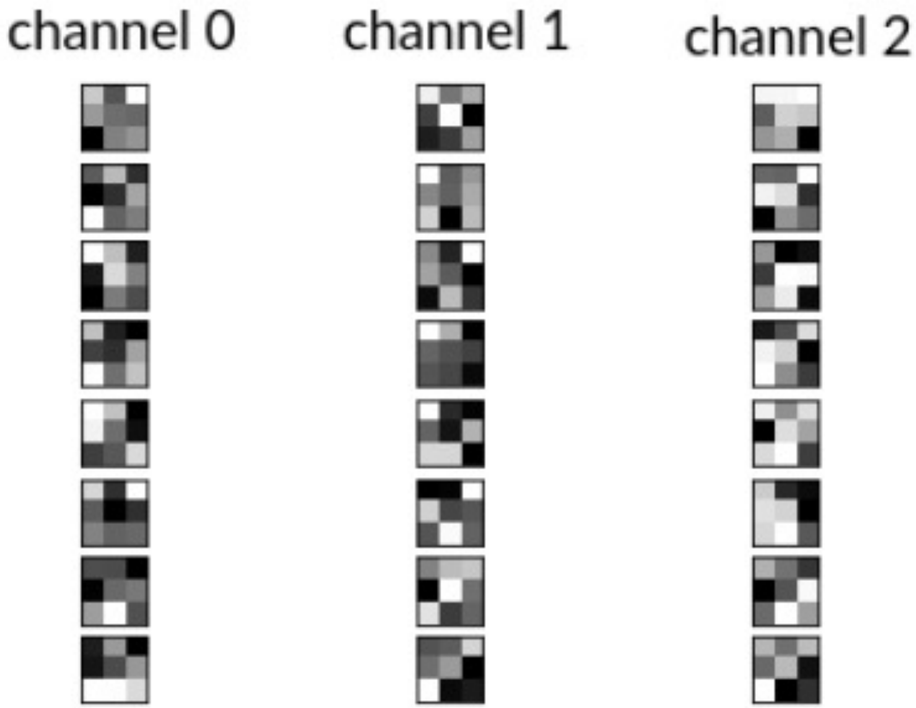


Towards Robust Interpretability with Self-Explaining Neural Networks, David Alvarez-Melis CSAIL, Tommi S. Jaakkola

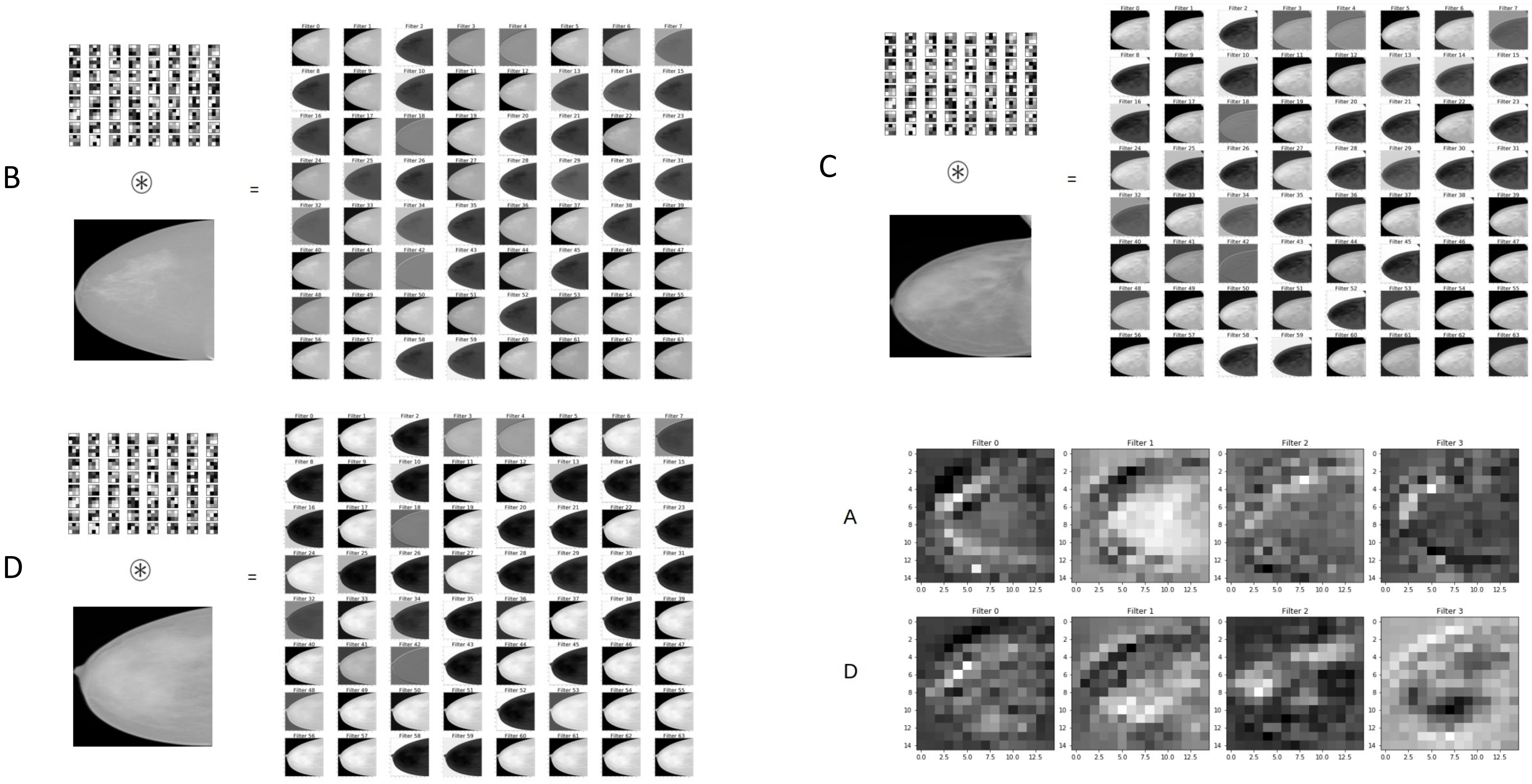
Self-explaining models in stages, progressively generalizing linear classifiers to complex yet architecturally explicit models.



Filters with 3 channels

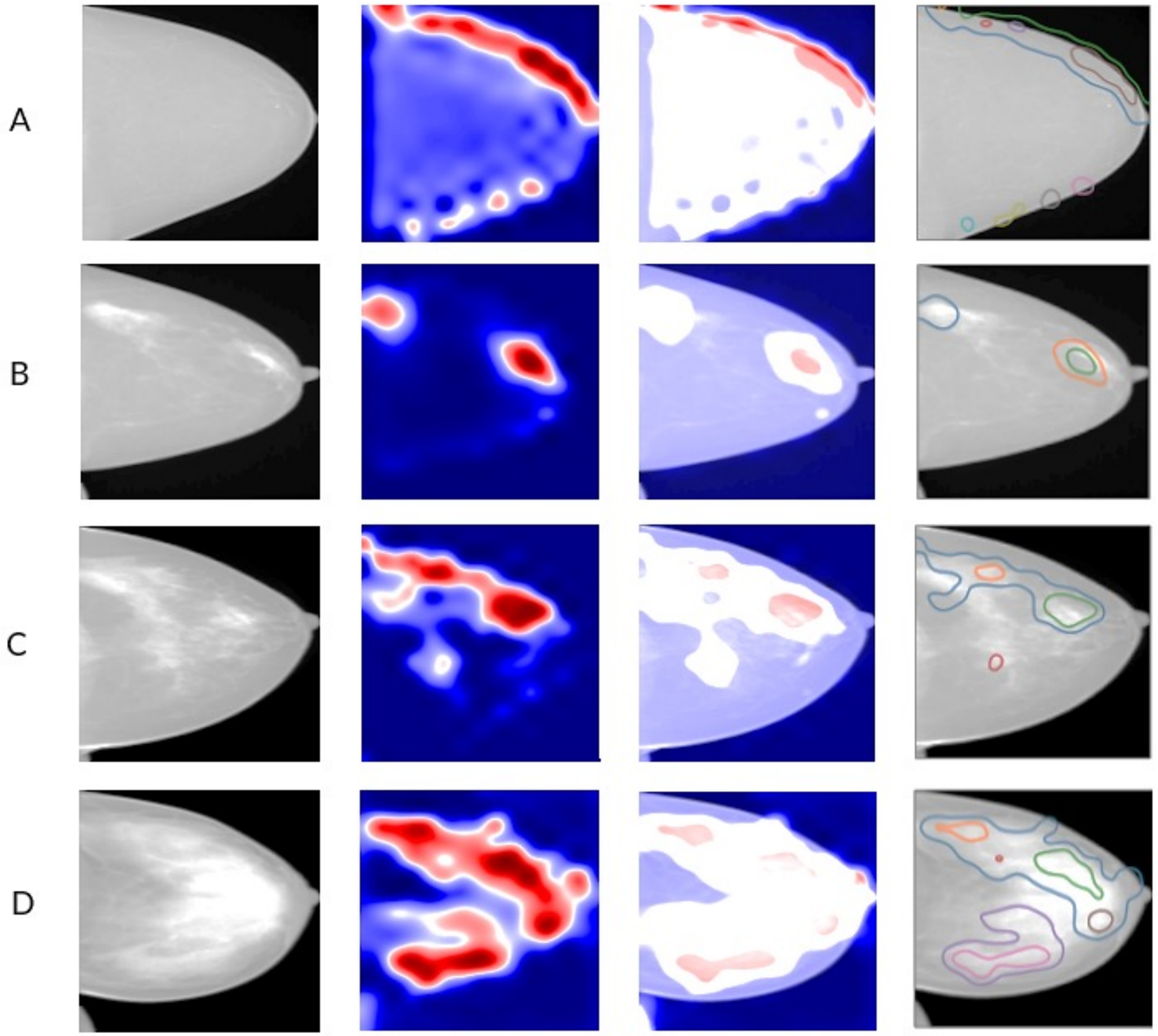


Feature maps



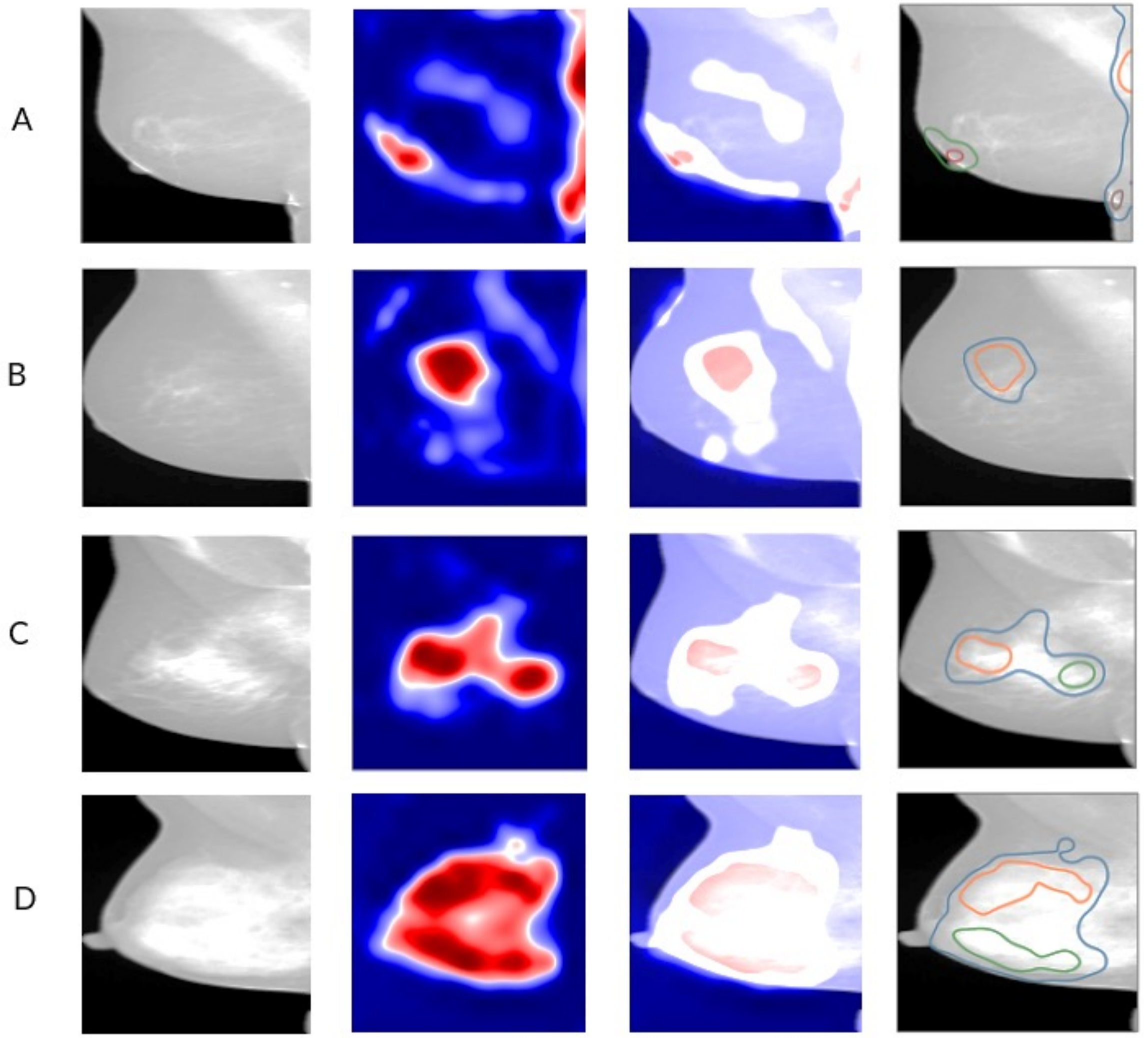
CAMs – Other projections

Left CC



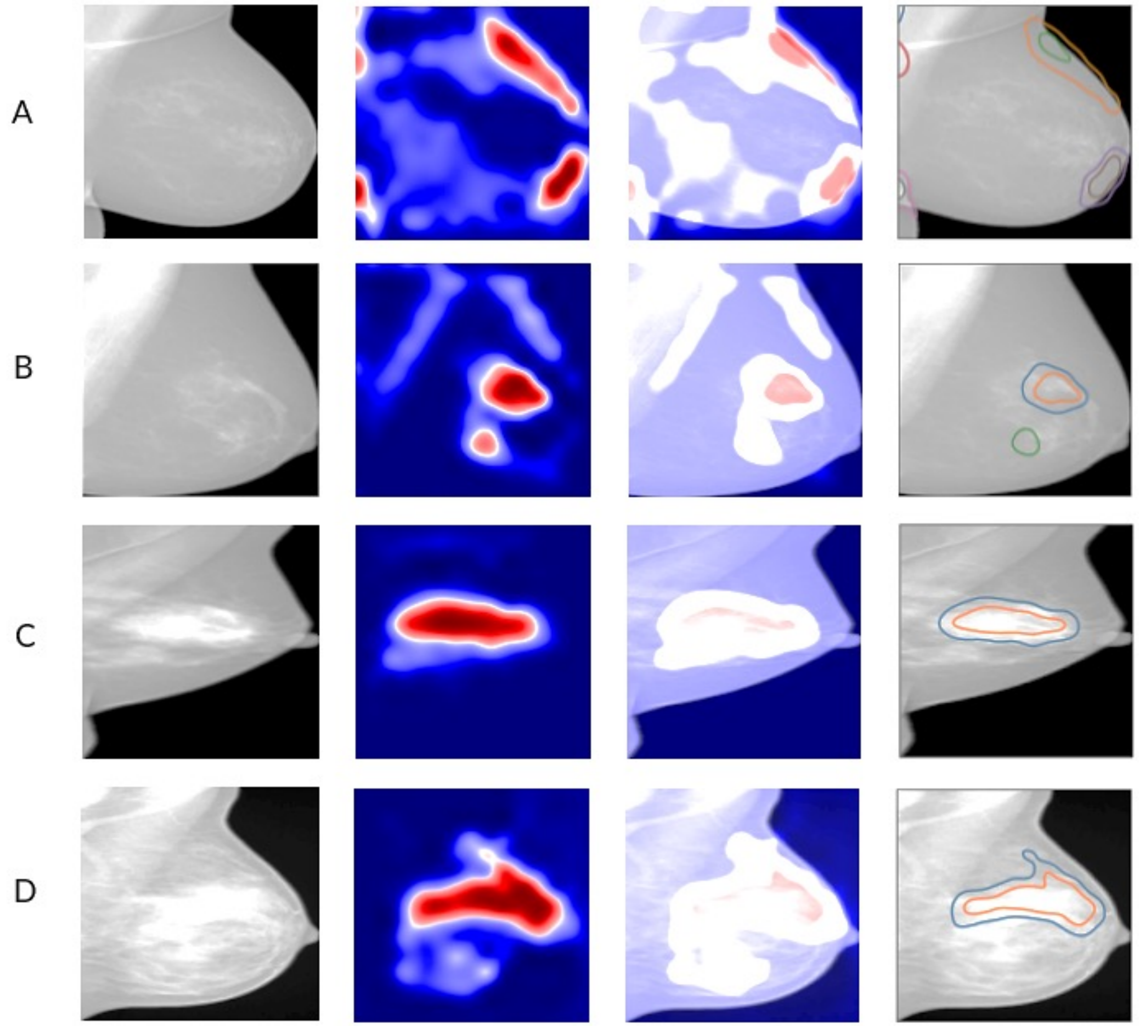
CAMs – Other projections

Right MLO

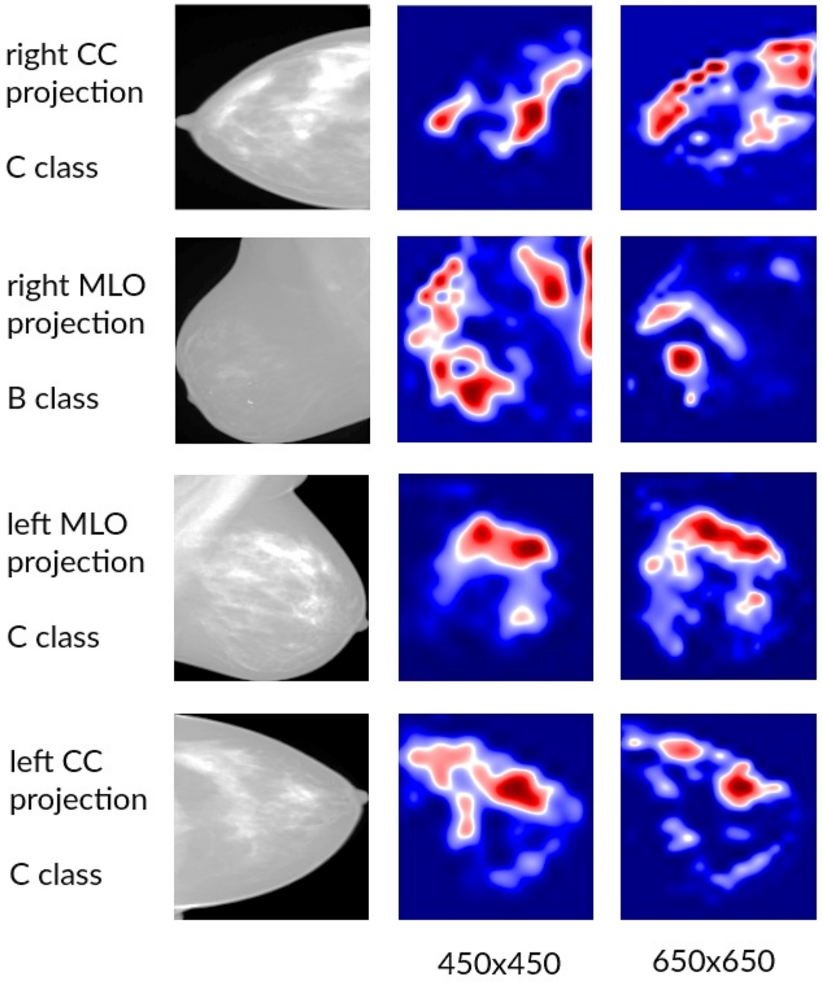


CAMs – Other projections

Left MLO

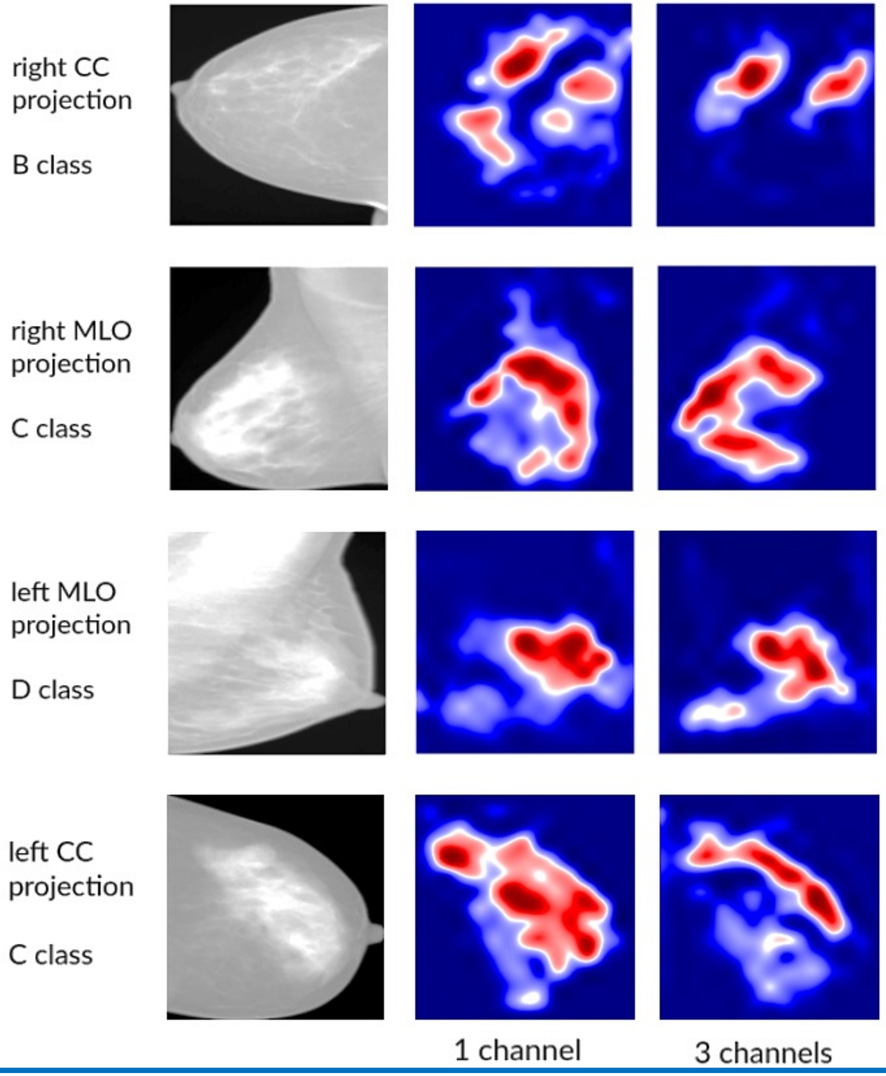


CAMs – Image size comparison

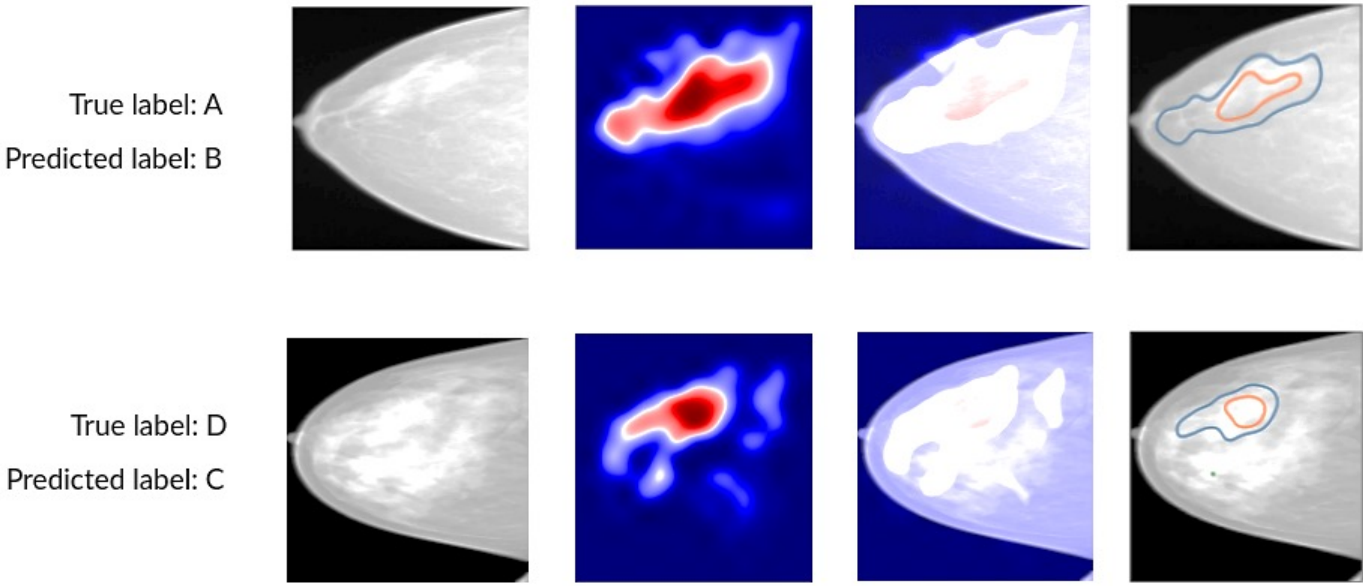


Explainability – Heatmaps (grad-CAM)

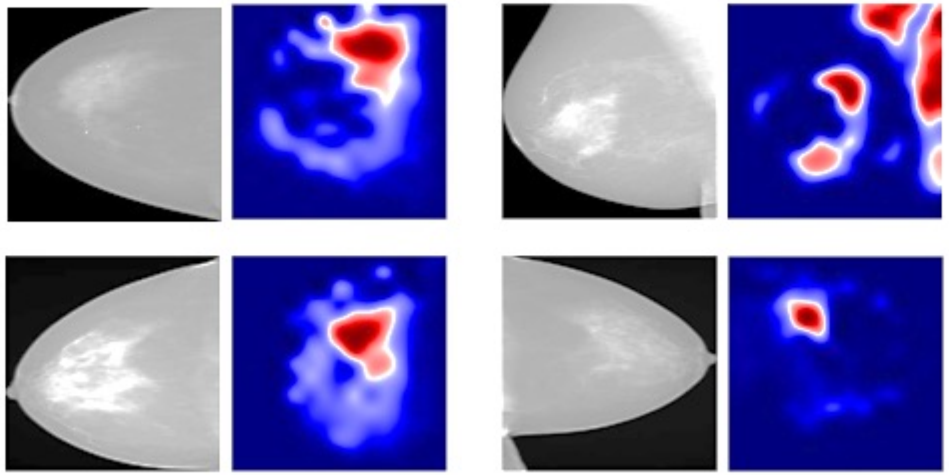
Number of channel comparison



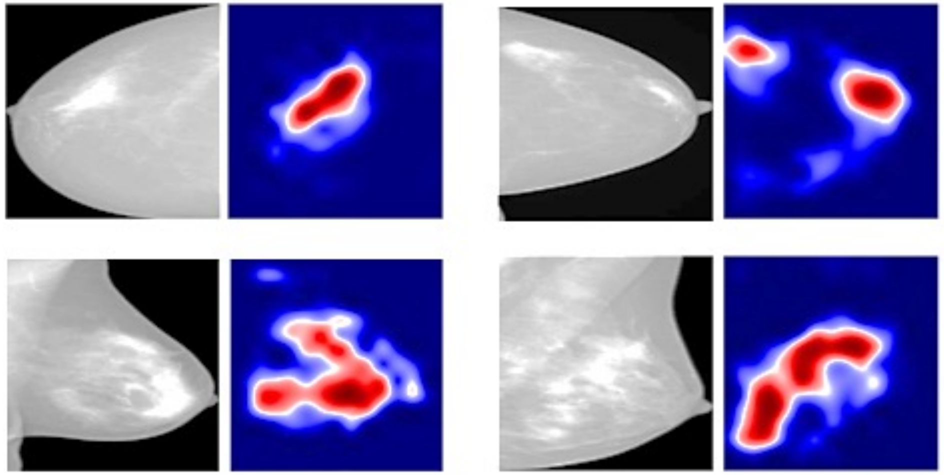
Misclassified examples



Explainability – Normalization comparison

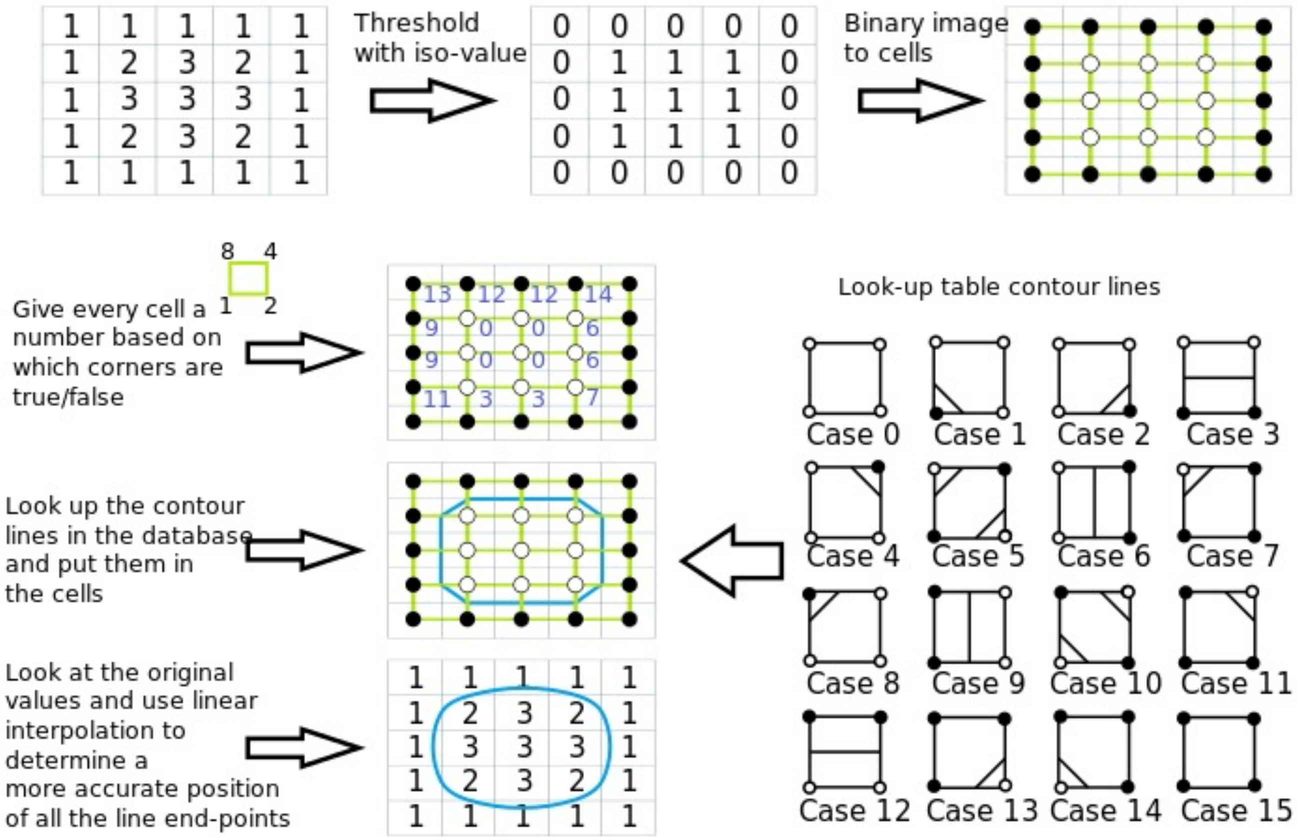


norm_1



norm_2

Marching squares algorithm



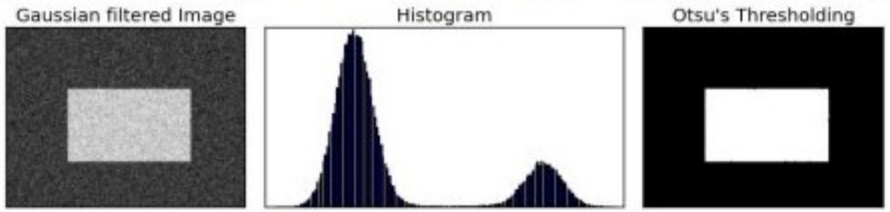
Thresholding segmentation

Inverted Binary Thresholding is an operation that transforms a grayscale image to a binary image according to the formula:

$$\mathtt{dst}(x,y) = \begin{cases} 0 & \text{if } \mathtt{src}(x,y) > \mathtt{thresh} \\ \mathtt{maxVal} & \text{otherwise} \end{cases}$$

If the intensity of the pixel src(x,y) is higher than thresh, then the new pixel intensity is set to 0. Otherwise, it is set to MaxVal.

Otsu's Binarization is a method that determines an optimal global threshold value from the image histogram.



image

Batch Normalization

Batch normalization layer normalizes the activations of the previous layer at each batch.

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1 \dots m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m} \sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

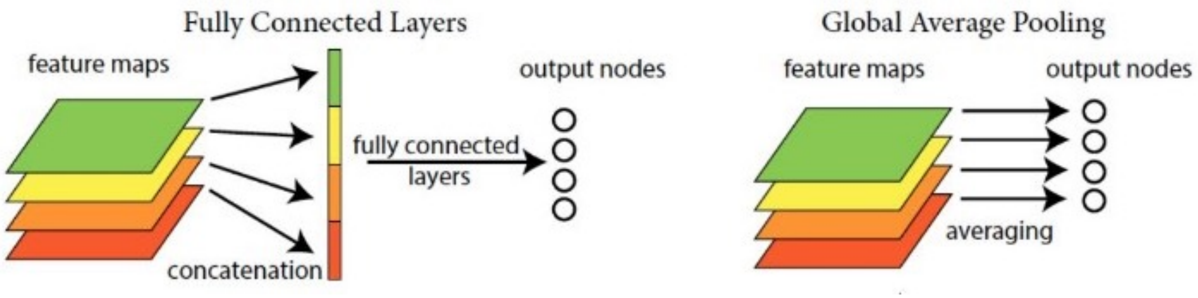
$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Global Average Pooling (GAP)

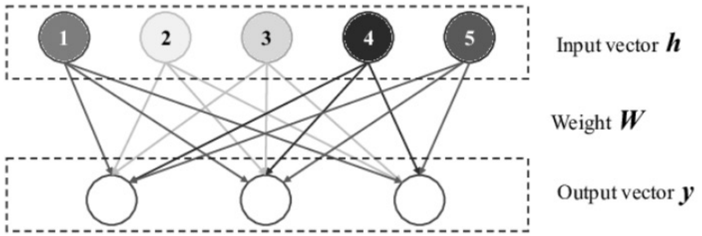
Global Average Pooling is an operation that calculates the average output of each feature map in the previous layer.



Dense layer

A Dense layer is a regular densely-connected NN layer. It implements the operation:

output = activation((input x kernel) + bias)



Input: GAP output tensor
Output: 4 classes

Softmax function: converts K-dimensional vector x containing real values to the same shaped vector of real values in the range of (0; 1), whose sum is 1.

$$\sigma(x)_i = \frac{\exp x_i}{\sum_{j=1}^{K} \exp x_j}$$

We apply the softmax function to the output of our convolutional network in order to convert the output to the probability for each class.

Categorical Cross-Entropy

Categorical crossentropy is a loss function.

$$CE = -\sum_{i}^{C} t_i log(f(s_i))$$

where t_i and s_i are the ground truth and the CNN score for each class i in C.
$f(s_i)$ refers to the activations applied to the scores before the CE Loss computation.

Categorical cross-entropy will compare the distribution of the predictions (the activations in the output layer, one for each class) with the true distribution, where the probability of the class is a value between 0 and 1.

Cross-entropy loss increases as the predicted probability diverges from the actual label.