Model explainability

Krzysztof J. Geras

NYU Langone Health
CAI²R
NYU DATA SCIENCE

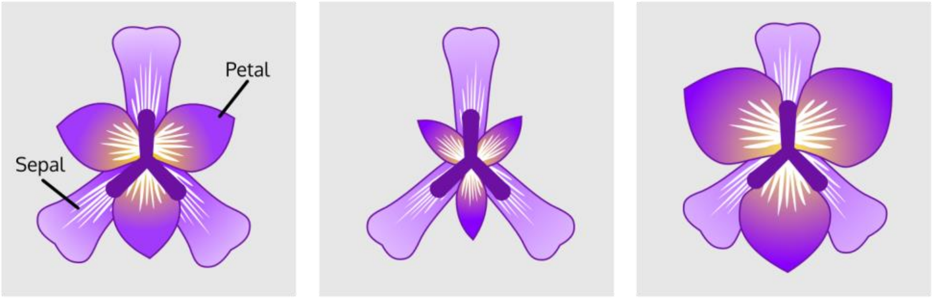
Opening notes

- The only purpose of this talk is that you learn stuff.
- If you have questions, ask. It’s more interesting for everyone that way, including me.
- We can go through the slides or we can stop and focus on what you find most interesting.
- It’s better to develop a understanding of fewer things than to have a shallow understanding of many things.
- My goal is to give an idea for what is possible and enable you to self-study effectively.
- We will focus on classification.

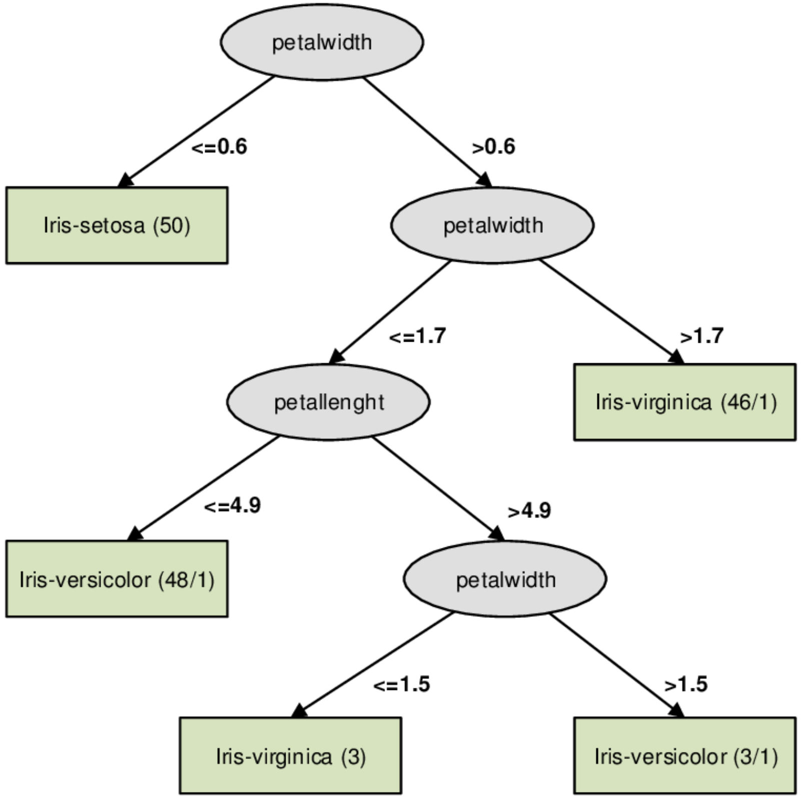
What is model explainability?

- On a high-level: “can we explain this model’s decision process in terms understandable to humans?”
- No consistent definition.
- “Explainability” exchangeable with “interpretability”, and “intelligibility” to an extent.
- This is a set of model-dependent techniques rather than consistent science.
- Model-level explanation / example-level explanation.
- Post-training / embedded into training.

Explainability in simple models – decision trees

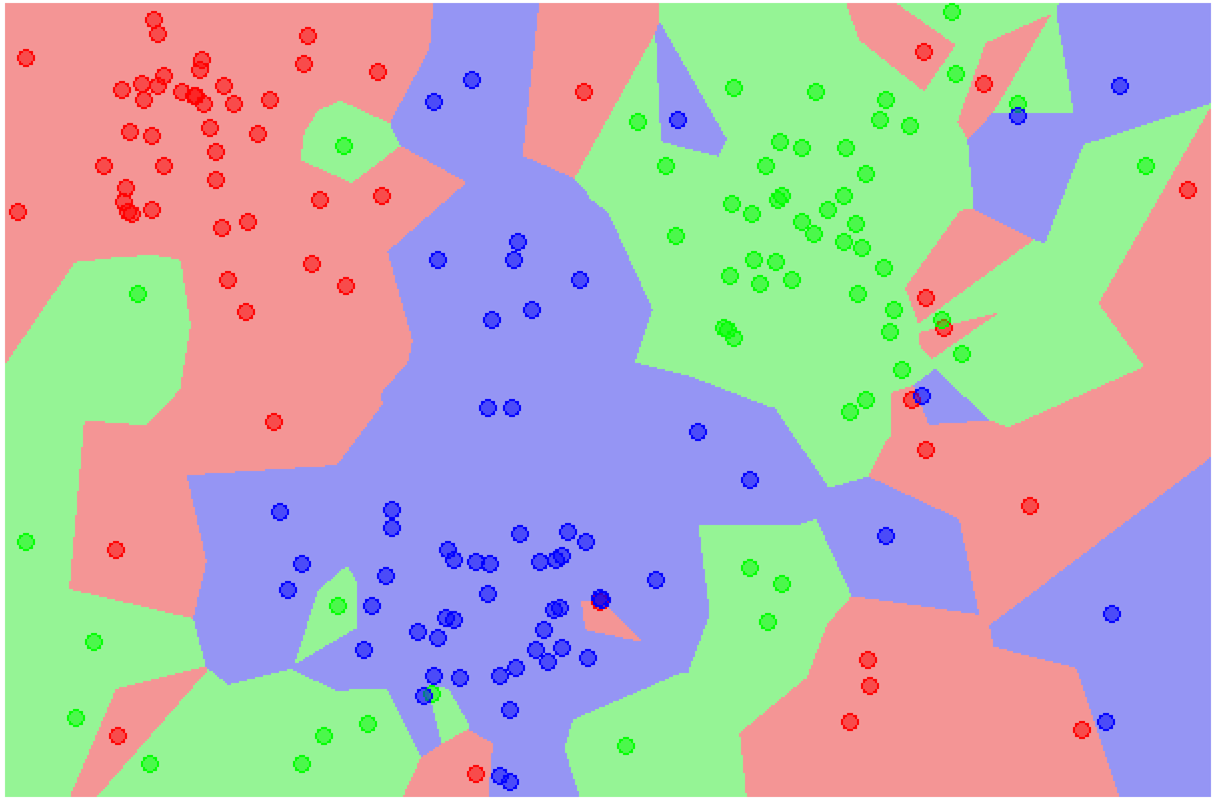


Iris Versicolor Iris Setosa Iris Virginica



Peter Grabusts, Arkady Borisov, Ludmila Aleksejeva. Ontology-Based Classification System Development Methodology. Information Technology and Management Science.

Explainability in simple models – k-NN



https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

Explainability in simple models – logistic regression

$$\hat{p}(\boldsymbol{x}, \beta) = \frac{1}{1 + e^{-(\boxed{\beta_0} + \sum_{i=1}^{K} \boxed{\beta_i} x_i)}}$$

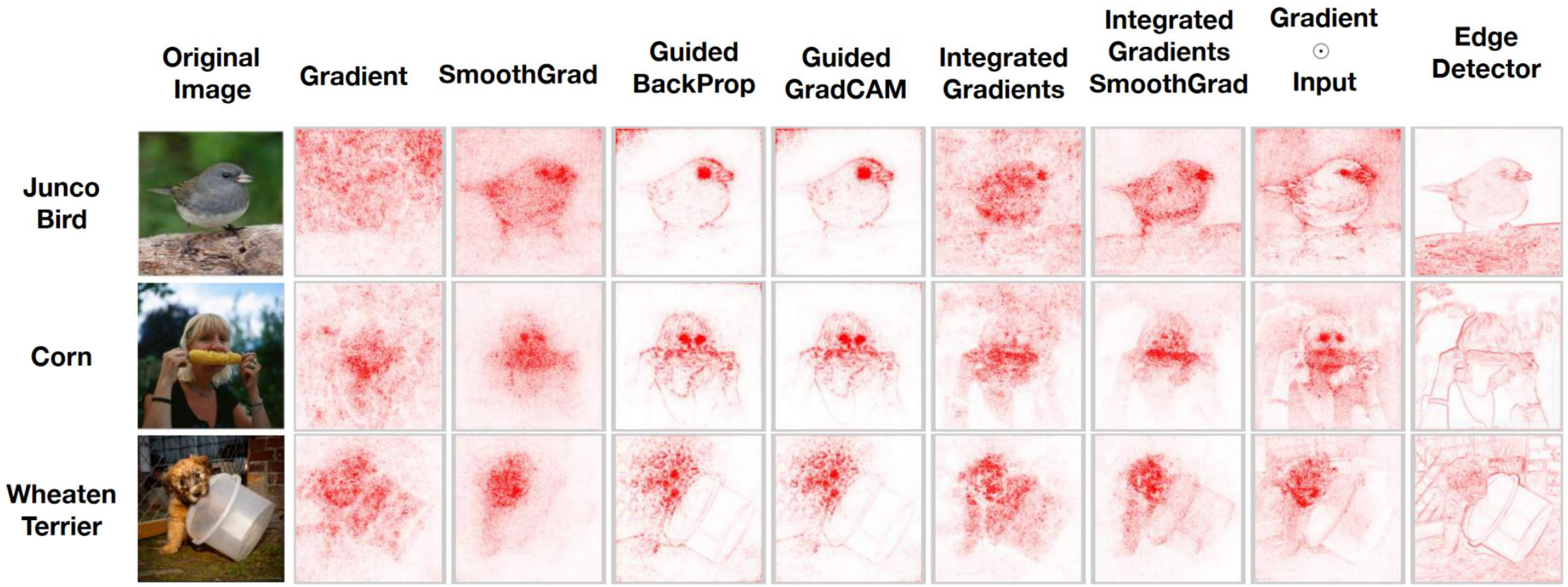
Explainability / expressivity tradeoff

- The simpler models, which are easier to explain are usually not very expressive.
- Easy to understand what knowledge is encoded in a decision tree and how it makes predictions. Typically, difficult for deep neural networks.
- This statement should be interpreted with some caution, e.g. kNN can represent any classification surface and is very easily explainable.
- Sometimes embedding explainability into the model improves generalization.

Explainability in deep learning

- These simple models are great, but it is difficult to solve any interesting learning task with them. We need something more powerful.
- We will focus on methods computer vision. Generalizations to other types of data are often straight-forward.
- There is a wide variety of different methods, mostly focusing on indicating the objects in the image that determine prediction. We will look at just a few.

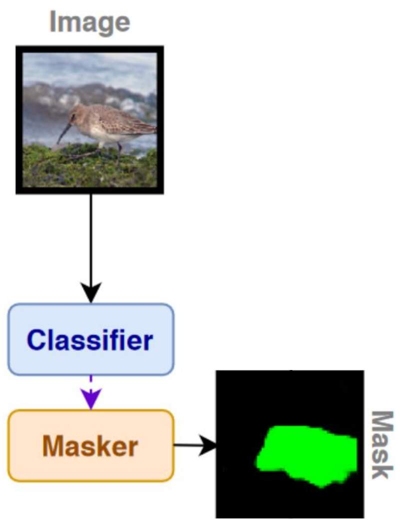
Gradient-based methods



$$\frac{\partial f(\hat{p}(\boldsymbol{x}, \beta))}{\partial \boldsymbol{x}}$$

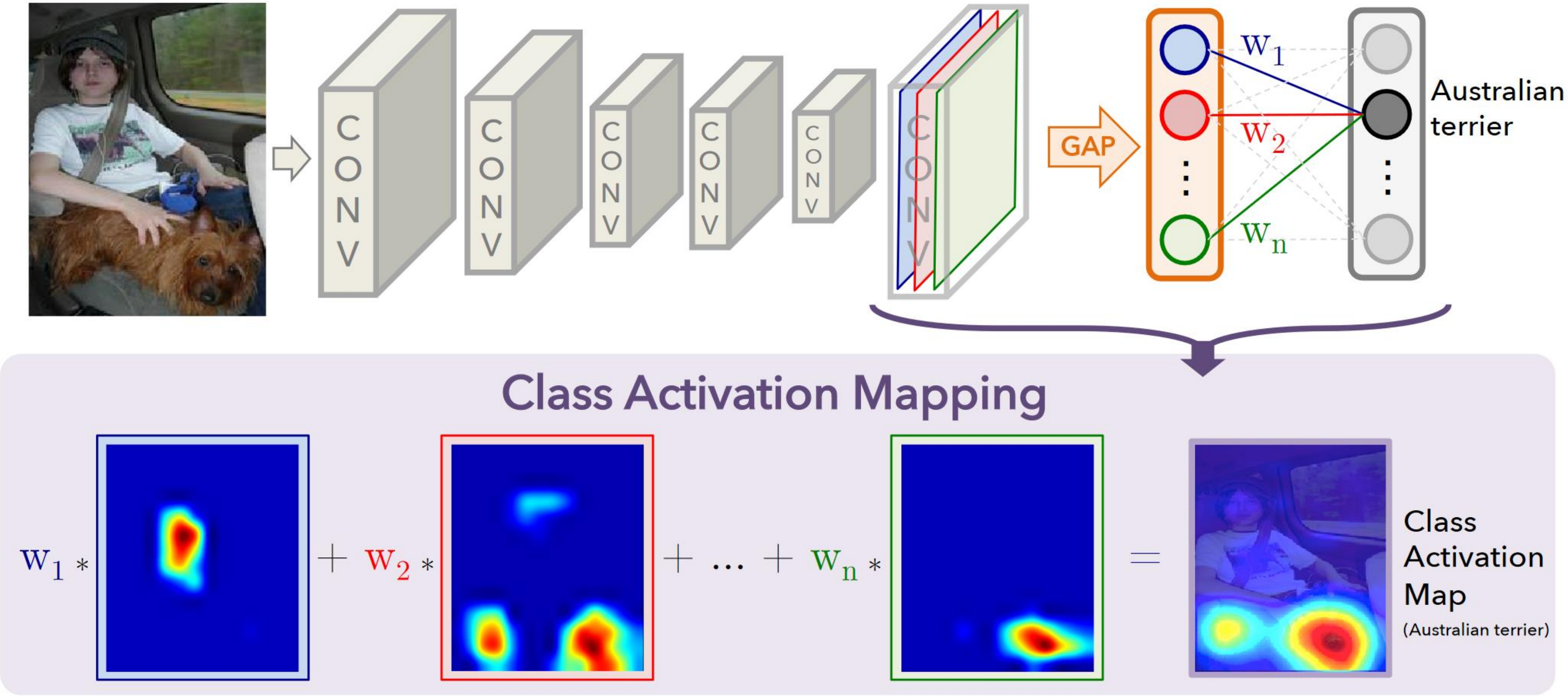
Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, Been Kim. Sanity Checks for Saliency Maps. NeurIPS 2018.

Perturbation-based methods



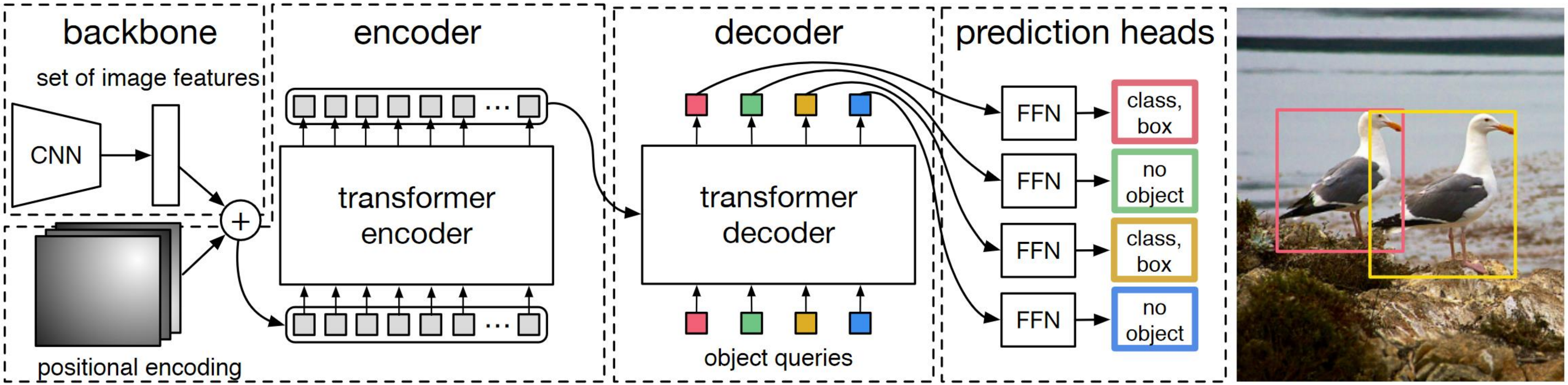
Jason Phang, Jungkyu Park, Krzysztof J. Geras. Investigating and Simplifying Masking-based Saliency Methods for Model Interpretability

“Weak” localization-based methods



Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba. Learning Deep Features for Discriminative Localization. CVPR 2016.

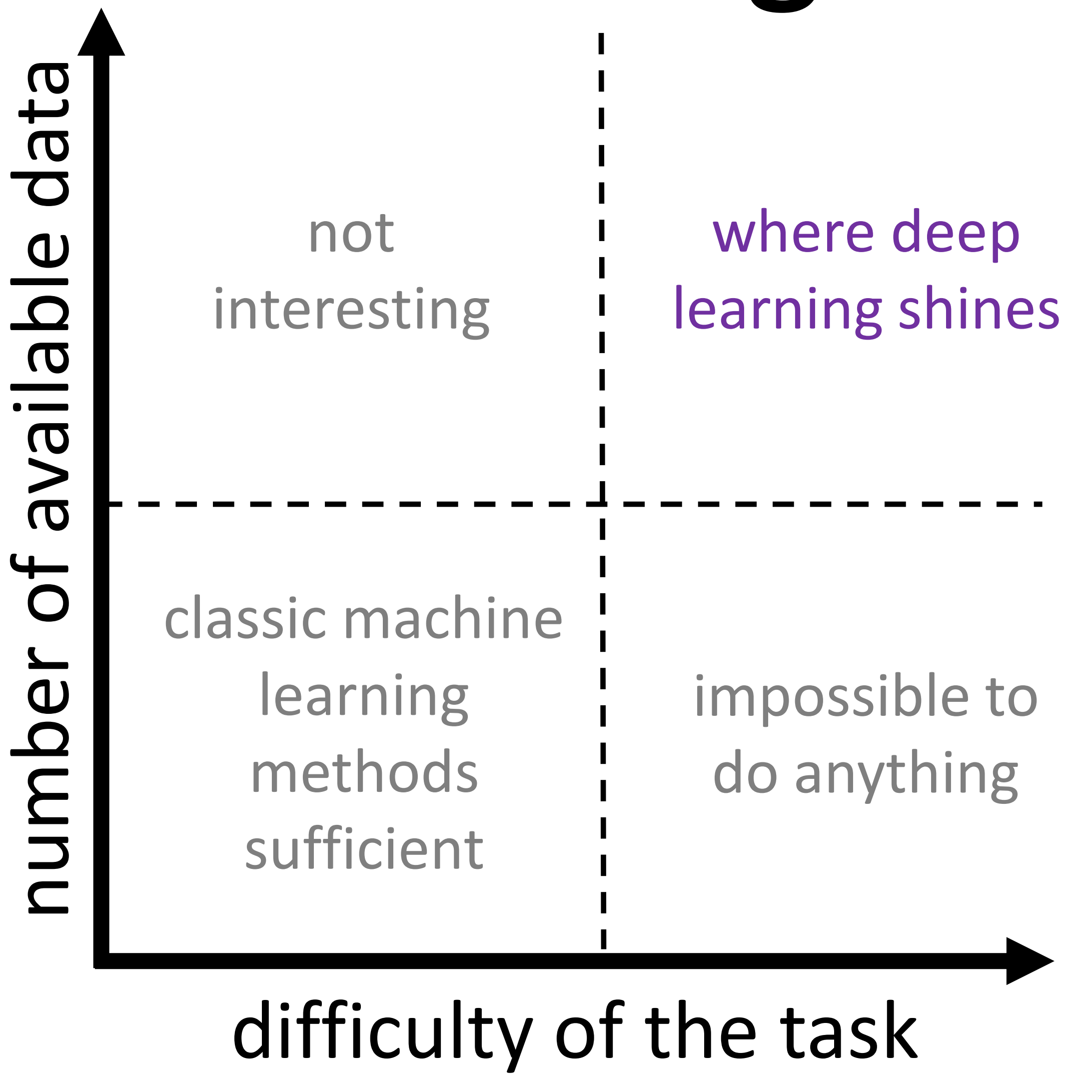
“Strong” localization-based methods



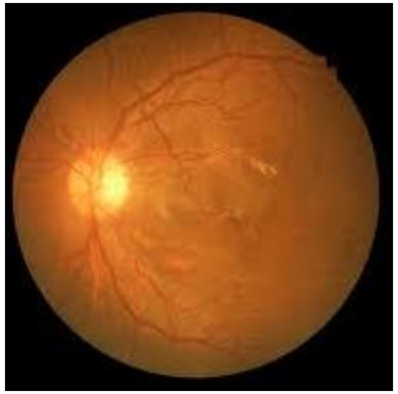
Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. ECCV 2020.

Meanwhile, in practice…

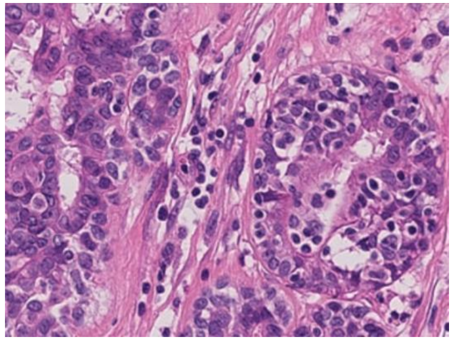
Why deep learning is the right tool for medical image analysis



Some initial successes



[Gulshan et al, JAMA 2016]



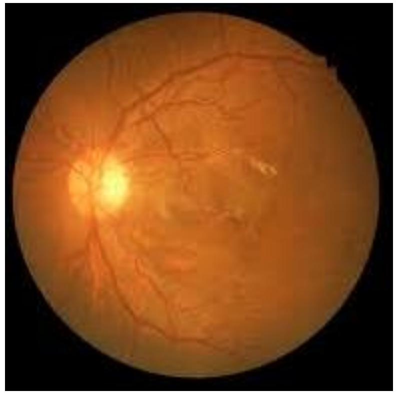
[Bejnordi et al, JAMA 2017]
[Coudray et al, Nature Medicine 2018]

[Wu et al, IEEE TMI 2019]
[McKinney et al, Nature 2020]

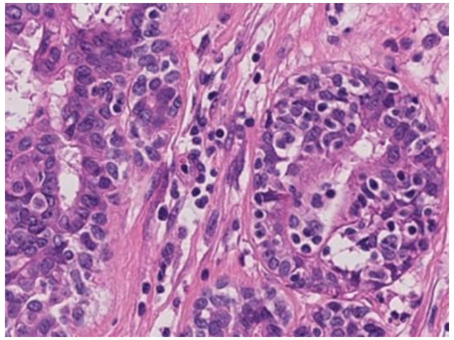
[Ardila et al, Nature Medicine 2019]

[Esteva et al, Nature 2017]

Some initial successes



[Gulshan et al, JAMA 2016]



[Bejnordi et al, JAMA 2017]
[Coudray et al, Nature Medicine 2018]

[Wu et al, IEEE TMI 2019]
[McKinney et al, Nature 2020]

[Ardila et al, Nature Medicine 2019]

[Esteva et al, Nature 2017]

Case study: breast cancer screening

- About **40 million** exams performed yearly in the US.
- About **250 thousand** women are diagnosed with cancer.
- About **40 thousand** lose their lives to cancer.

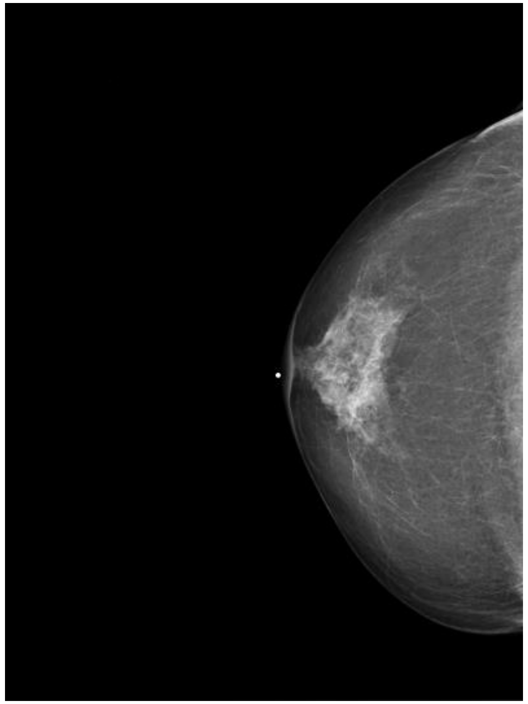
Breast cancer screening

R-MLO

(right mediolateral oblique)

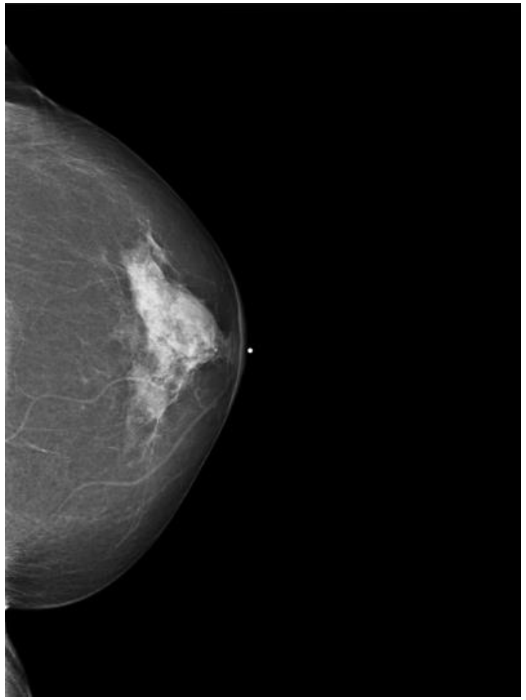
L-MLO

(left mediolateral oblique)



R-CC

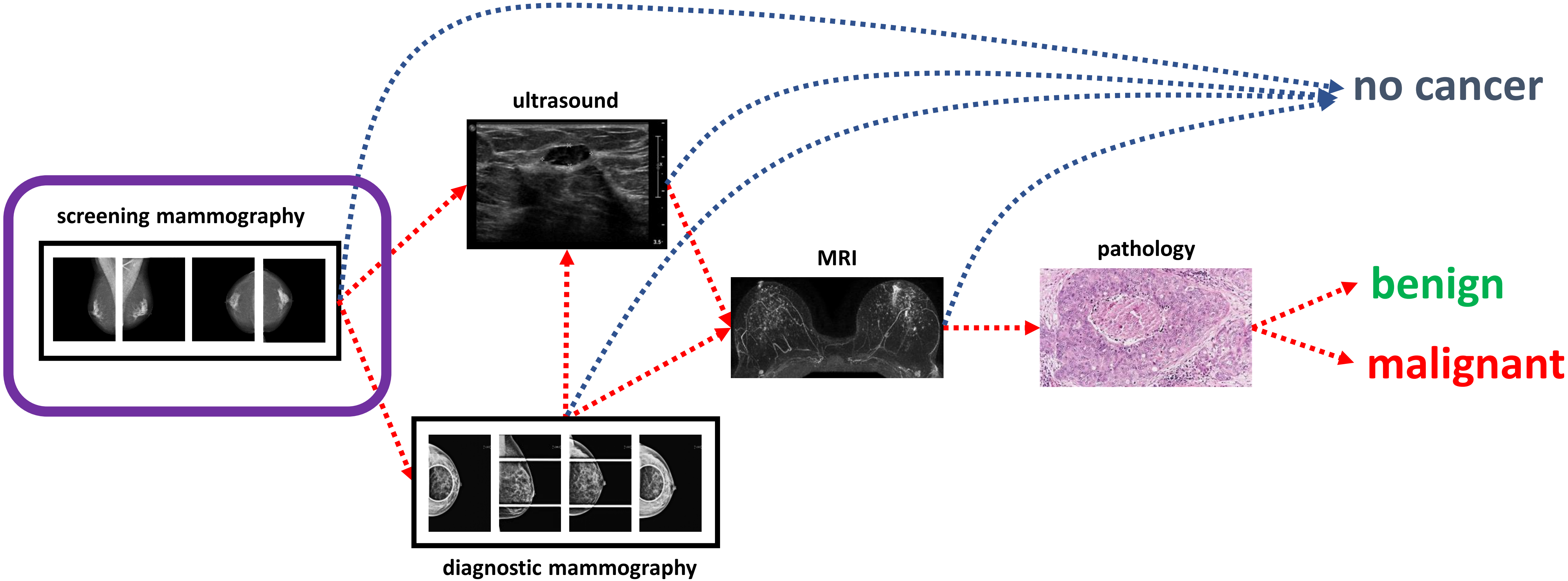
(right cranial caudal)



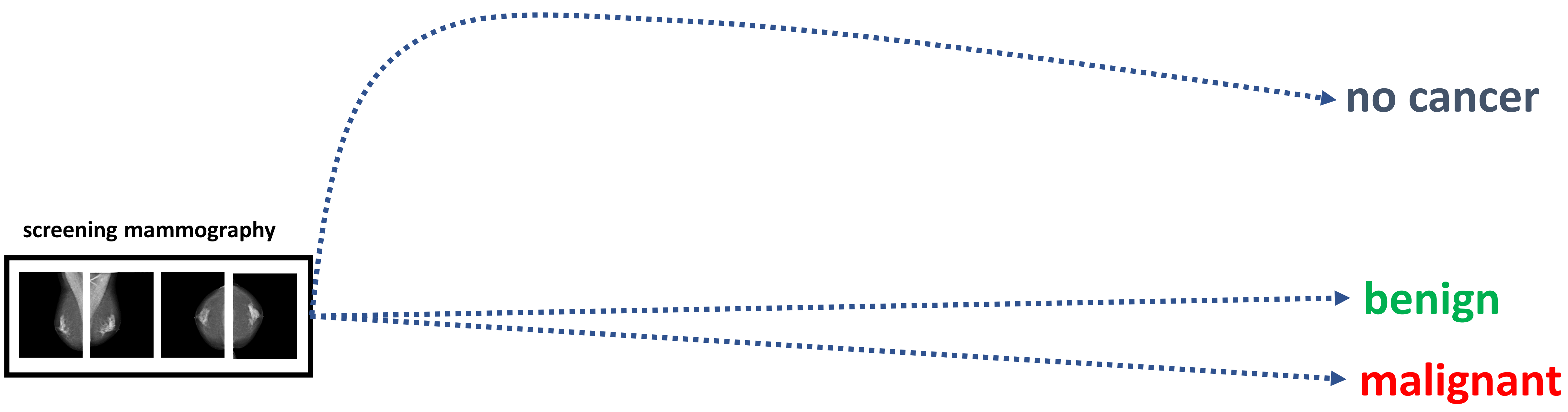
L-CC

(left cranial caudal)

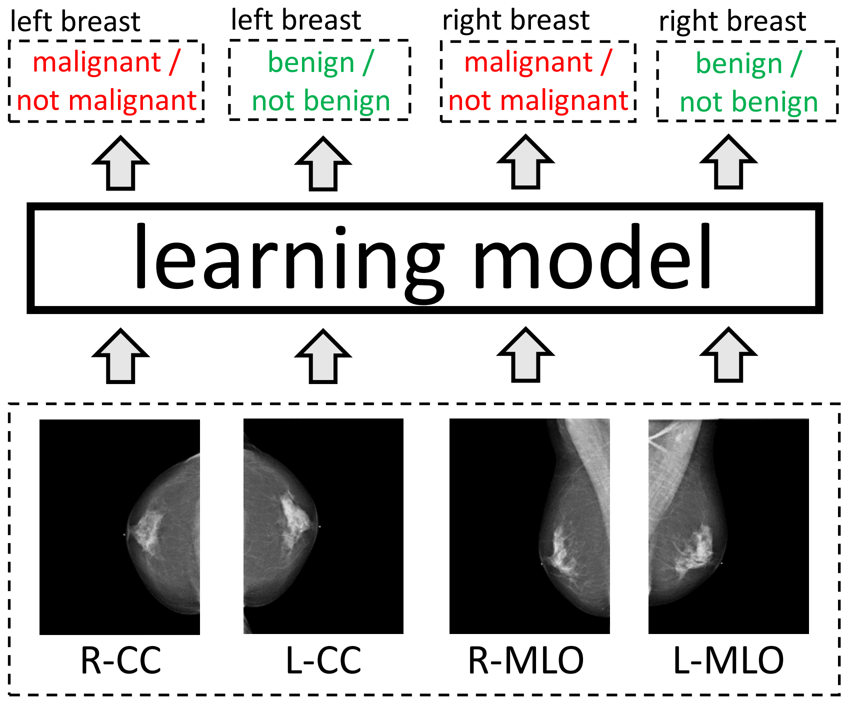
Diagnostic workflow



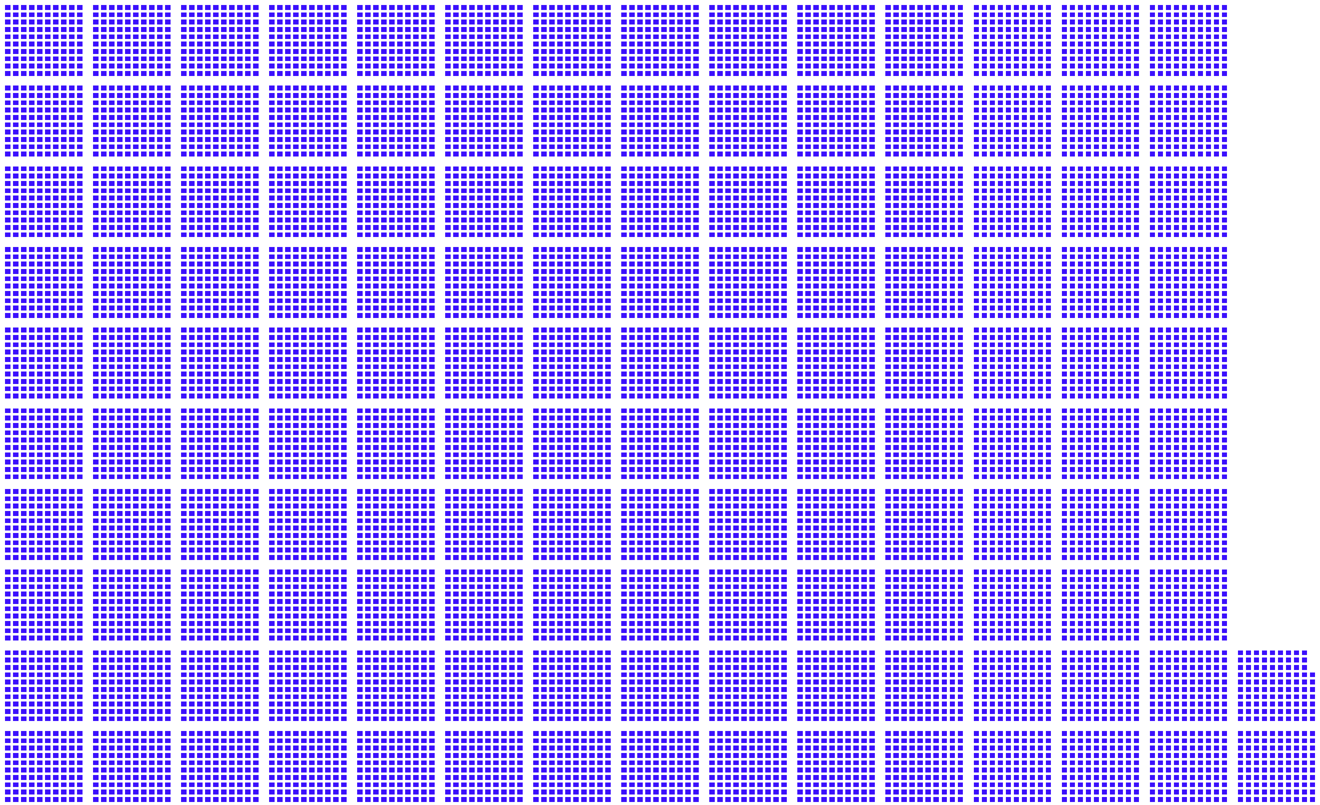
Cancer prediction



Cancer prediction task

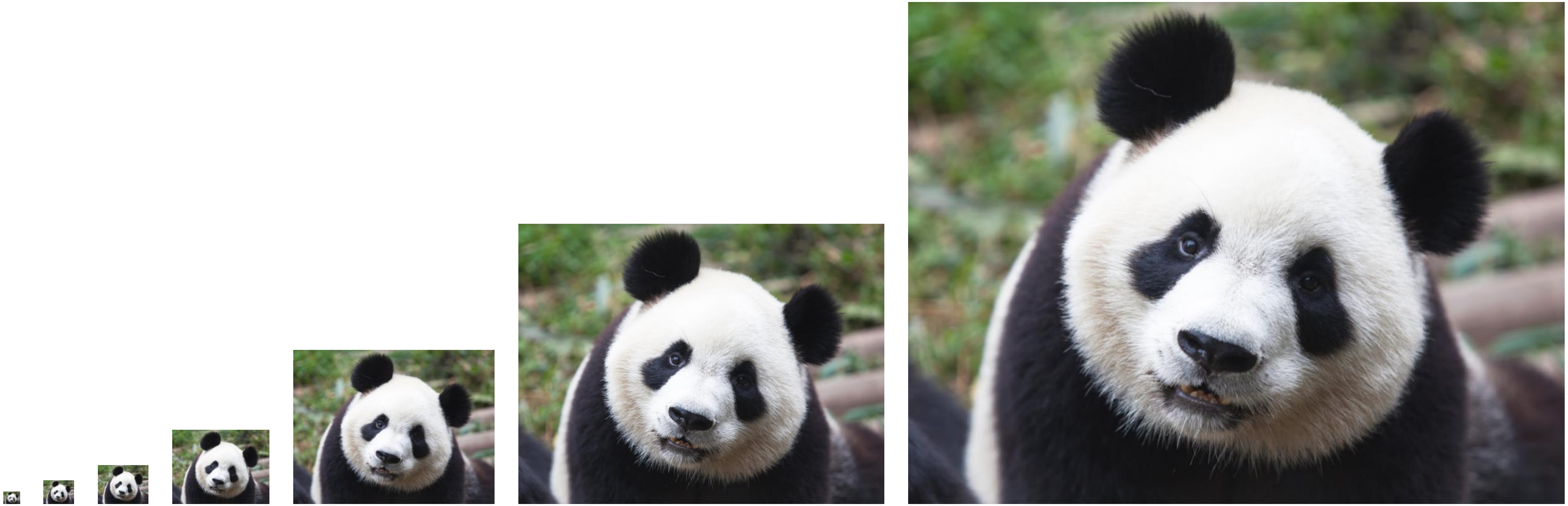


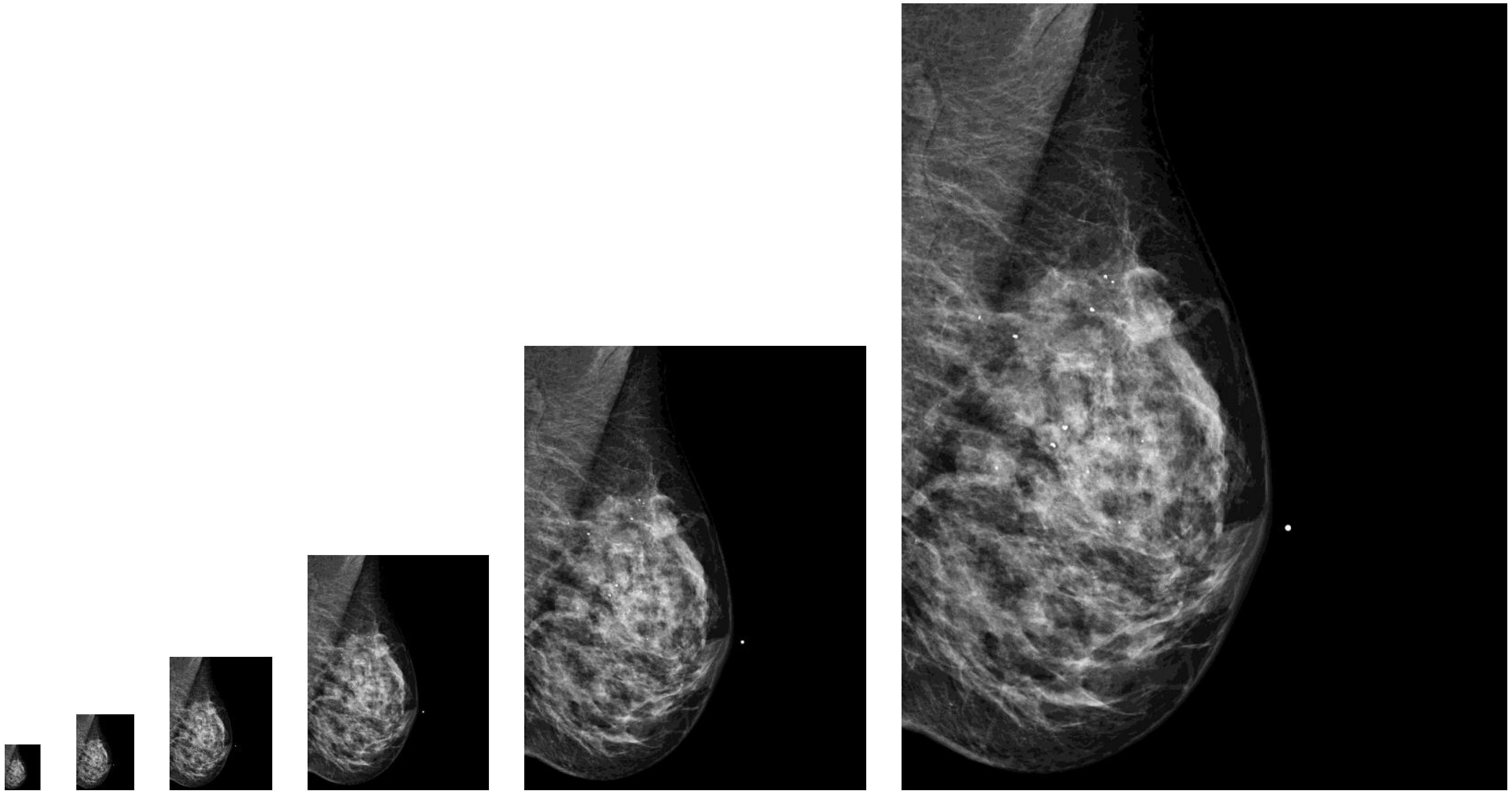
Why classification of medical images is hard



ImageNet
~14m images

Why classification of medical images is hard

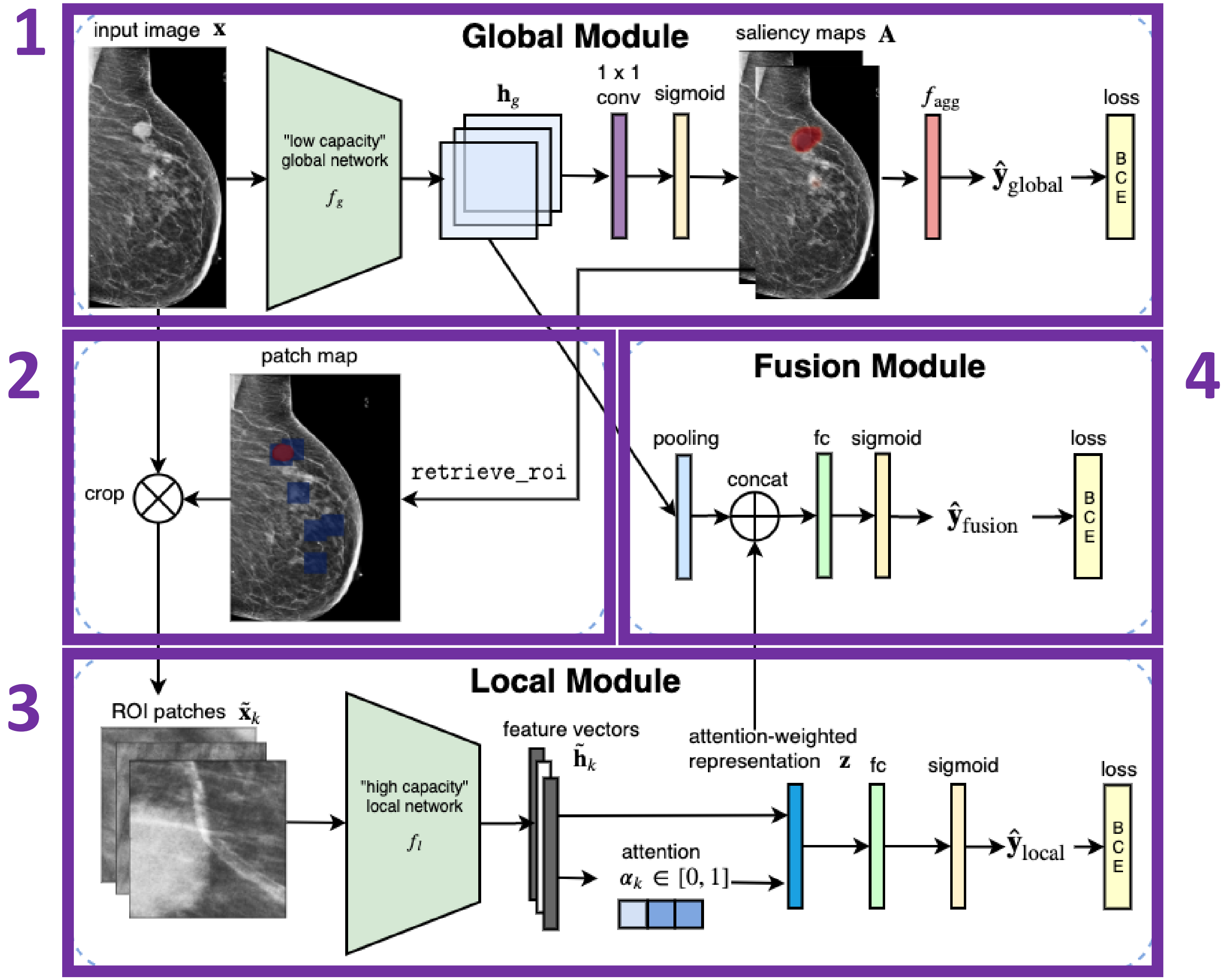




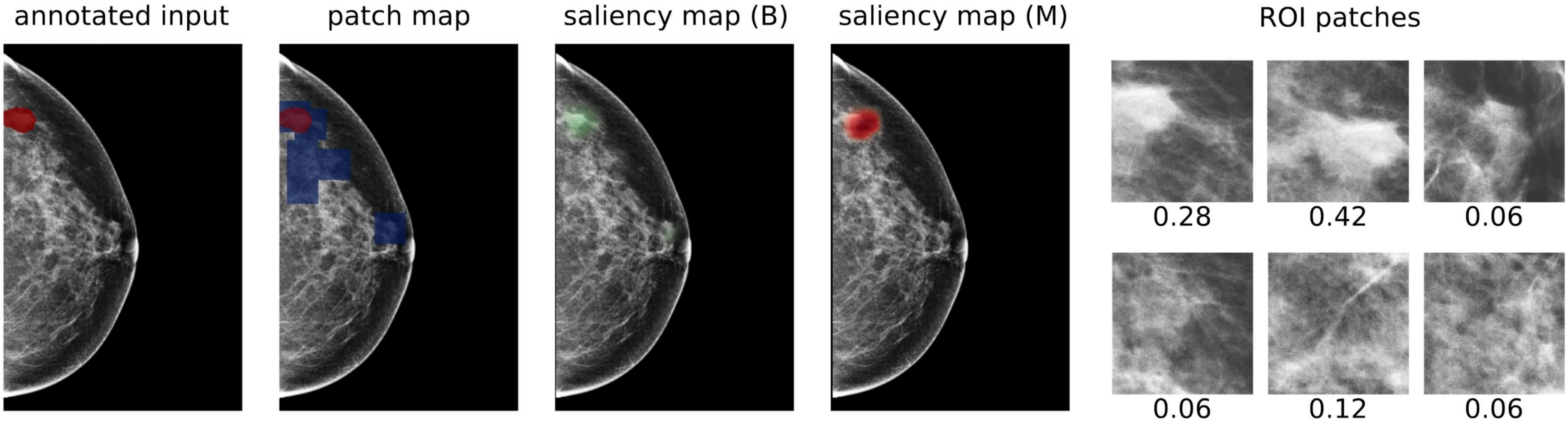
Challenges of learning from medical imaging data

- Public data sets are very tiny. Hospitals are not keen to share data between themselves.
- Labeling medical imaging data on a pixel level is difficult.
- Medical image data has very different properties than natural images for which standard neural networks are designed.
- The standard neural network architectures do not have any direct mechanism to explain their predictions.
- Evaulating the impact of machine learning is difficult.

Globally-Aware Multiple Instance Classifier



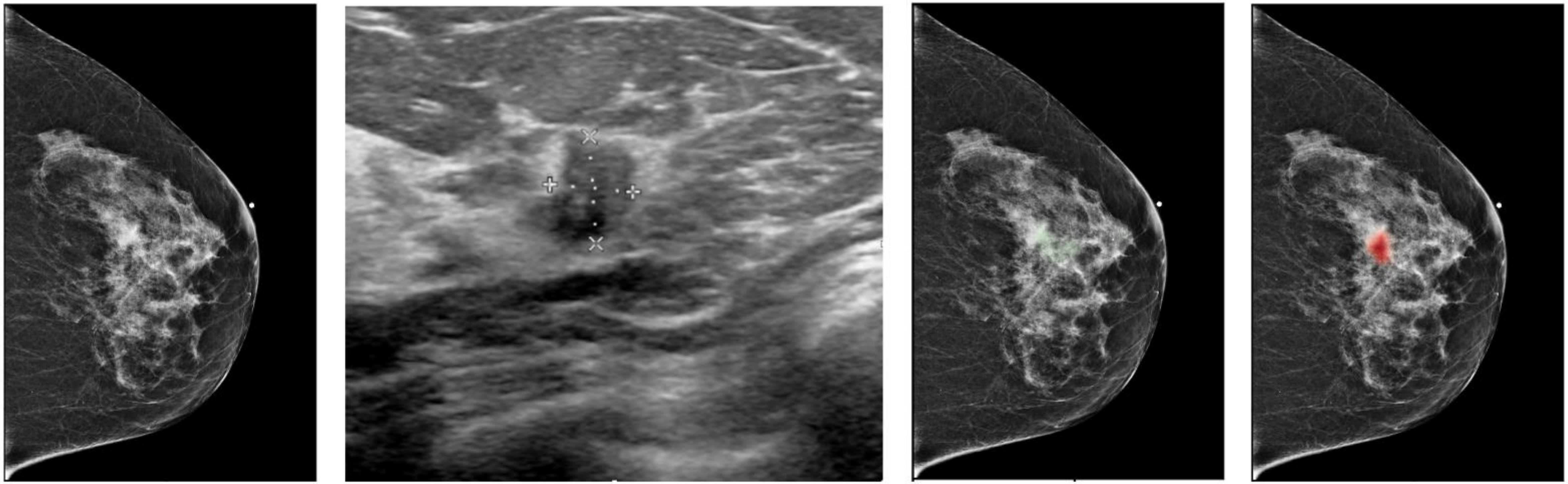
Shen et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Medical Image Analysis 2021


annotated input
patch map
saliency map (B)
saliency map (M)
ROI patches
0.28
0.42
0.06
0.06
0.12
0.06

Comparison to prior models

Model	AUC(M)	AUC(B)	#Param	Mem(GB)	Fwd/Bwd (ms)	FLOPs
ResNet 34 + fc	0.736 ± 0.026	0.684 ± 0.015	21.30M	13.95	189/459	1622B
ResNet 34 + 1×1 conv	0.889 ± 0.015	0.772 ± 0.008	21.30M	12.58	201/450	1625B
DMV-CNN (w/o heatmaps)	0.827 ± 0.008	0.731 ± 0.004	6.13M	2.4	38/86	65B
DMV-CNN (w/ heatmaps)	0.886 ± 0.003	0.747 ± 0.002	6.13M	2.4	38/86	65B
Faster R-CNN	0.908 ± 0.014	0.761 ± 0.008	104.8M	25.75	920/2019	-[3]
GMIC-ResNet-18	0.913 ± 0.007	0.791 ± 0.005	15.17M	3.01	46/82	122B
GMIC-ResNet-34	0.909 ± 0.005	0.790 ± 0.006	25.29M	3.45	58/94	180B
GMIC-ResNet-50	$\mathbf{0.915} \pm 0.005$	$\mathbf{0.797} \pm 0.003$	27.95M	5.05	66/131	194B
GMIC-ResNet-18-ensemble	**0.930**	0.800	-	-	-	-
GMIC-ResNet-34-ensemble	0.920	0.795	-	-	-	-
GMIC-ResNet-50-ensemble	0.927	**0.805**	-	-	-	-

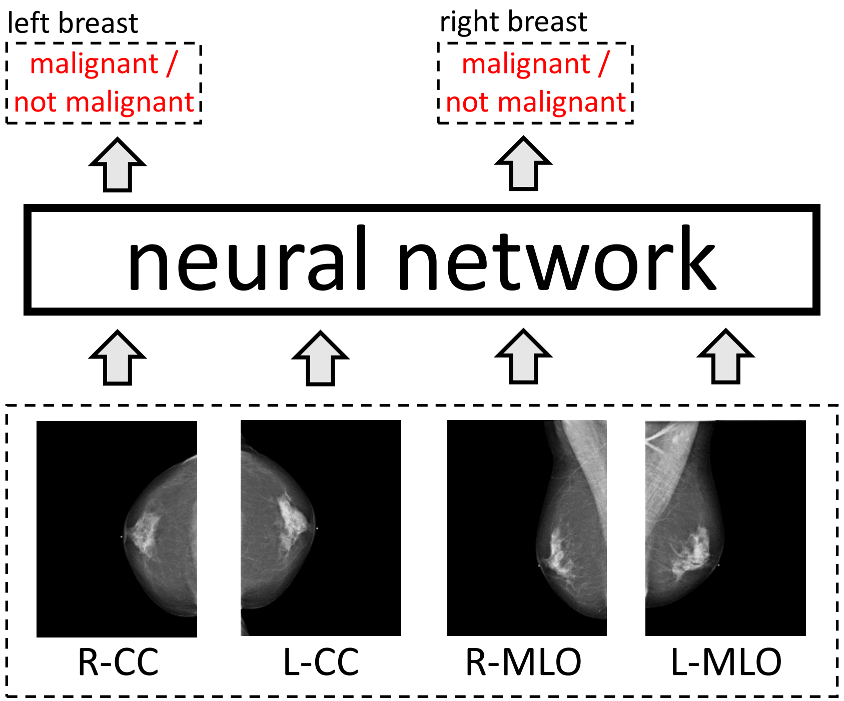
Interpreting mammographically occult cases



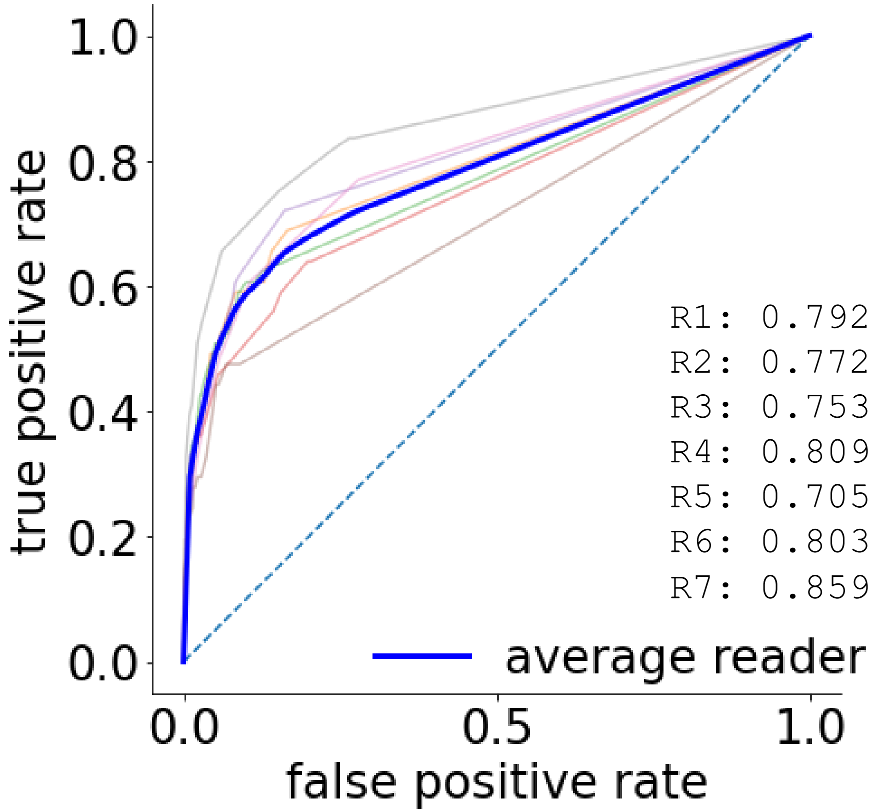
Comparison to human performance

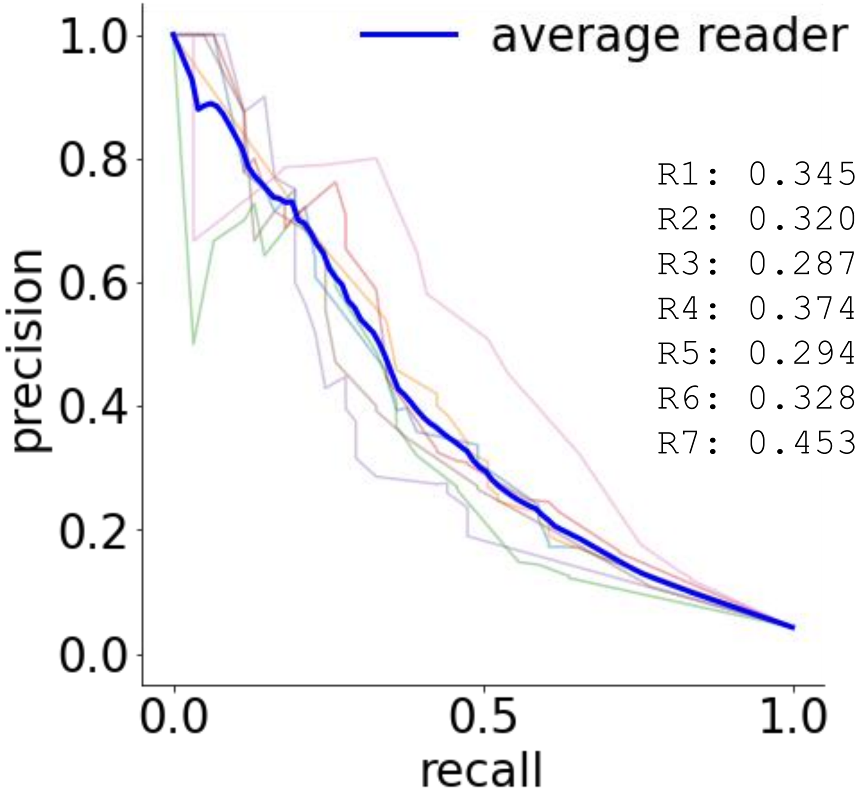
Reader study:

- 360 exams with a biopsy, 360 negative exams.
- 7 attending radiologists.
- Radiologists asked for a prediction of probability of malignancy.

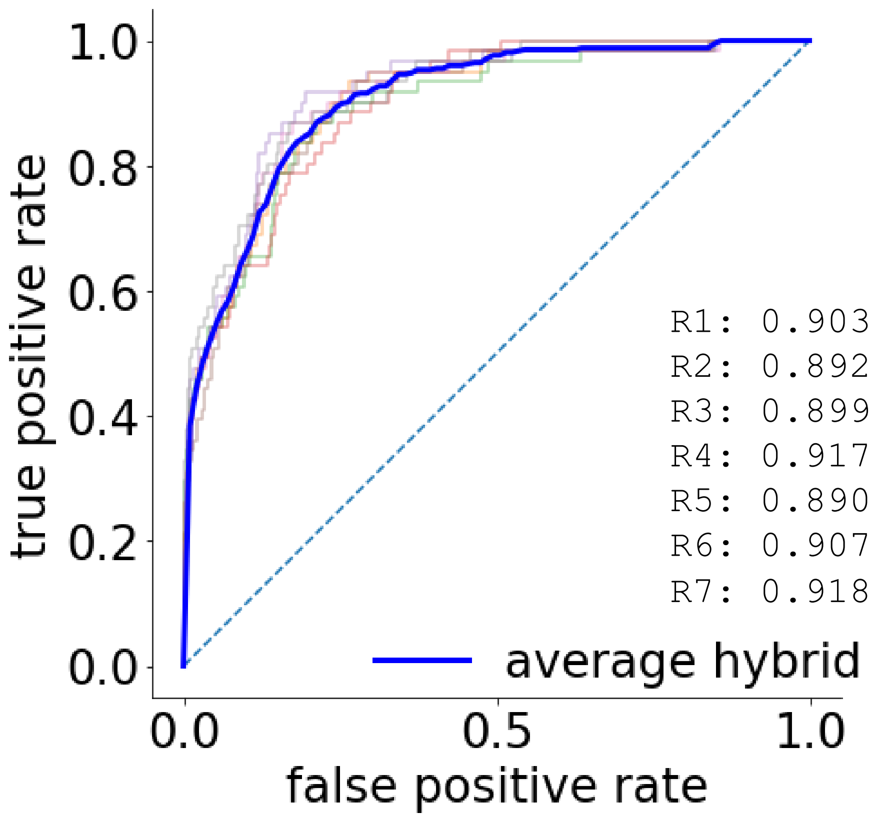


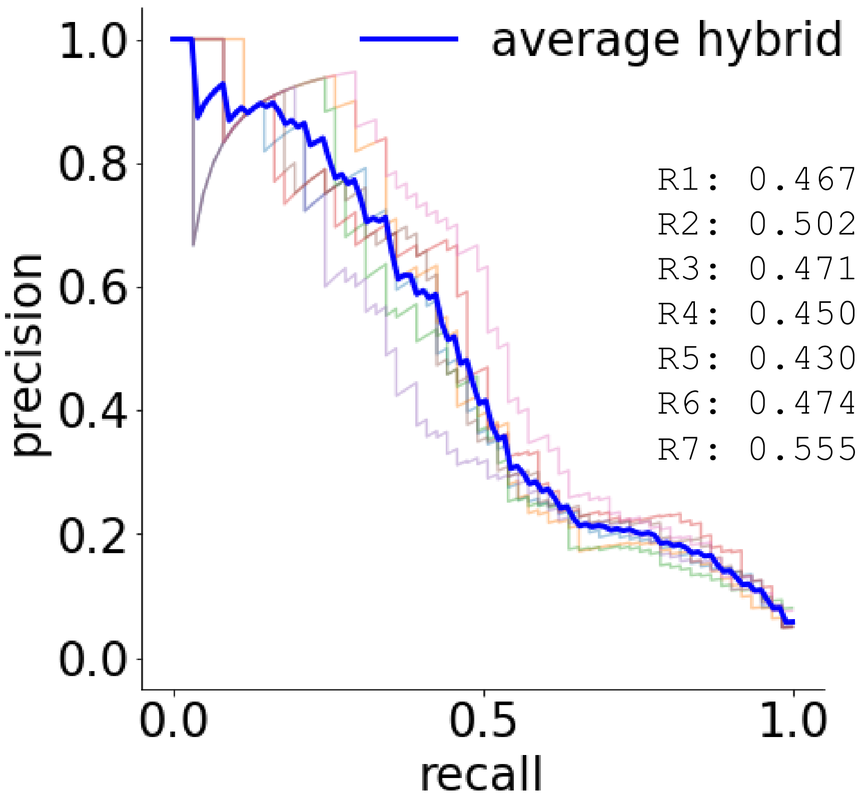
Improving radiologist performance





Improving radiologist performance





Code and models

https://github.com/nyukat/GMIC

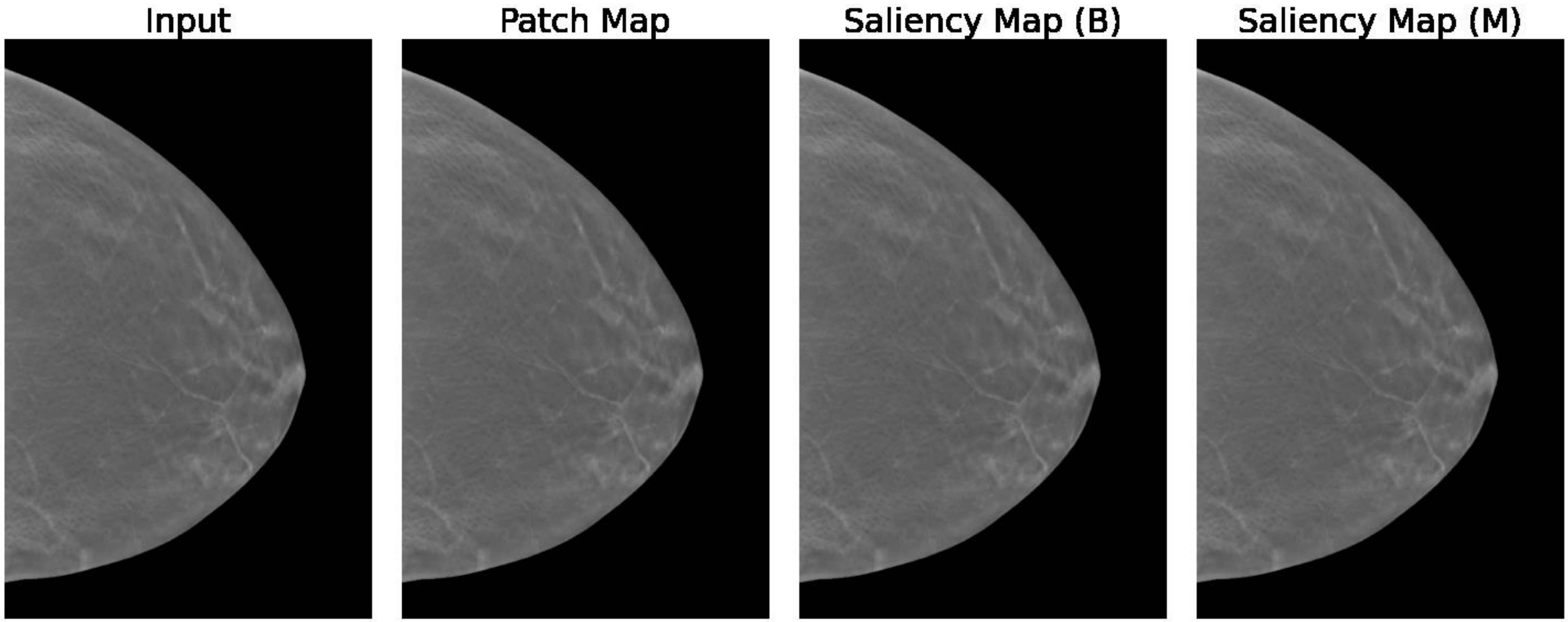
GMIC

An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization

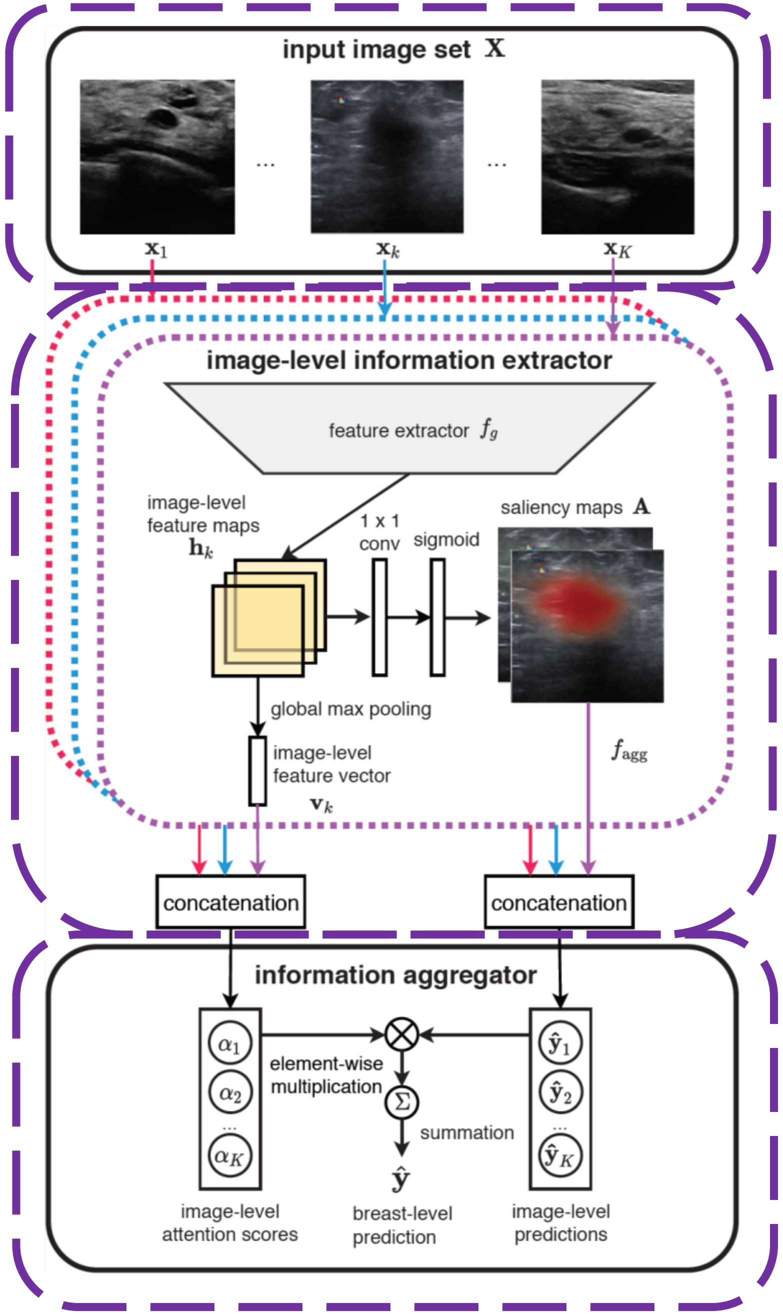
deep-learning pytorch medical-imaging breast-cancer breast-cancer-diagnosis breast-cancer-screening

Jupyter Notebook AGPL-3.0 21 75 0 0 Updated on Mar 8

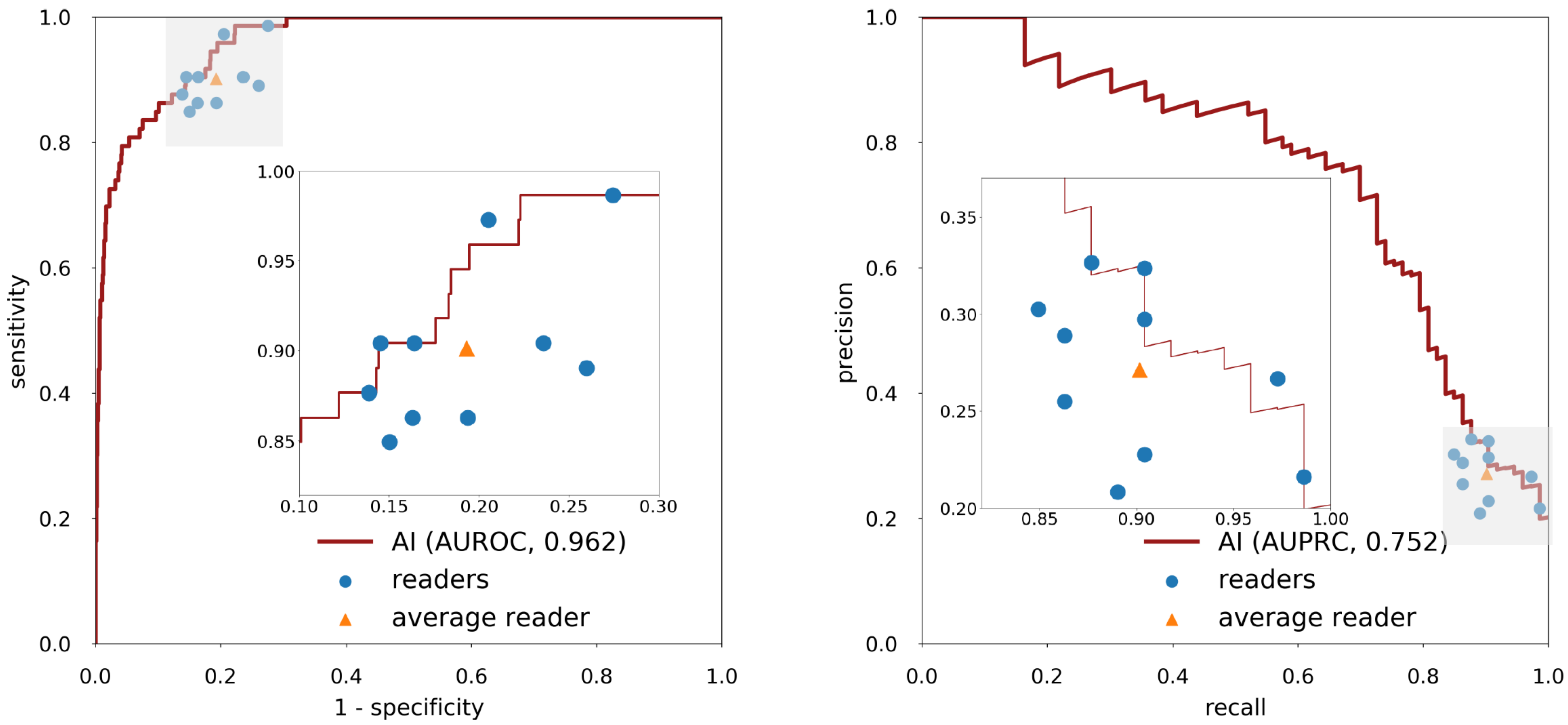
Similar ideas can be generalized to 3D…



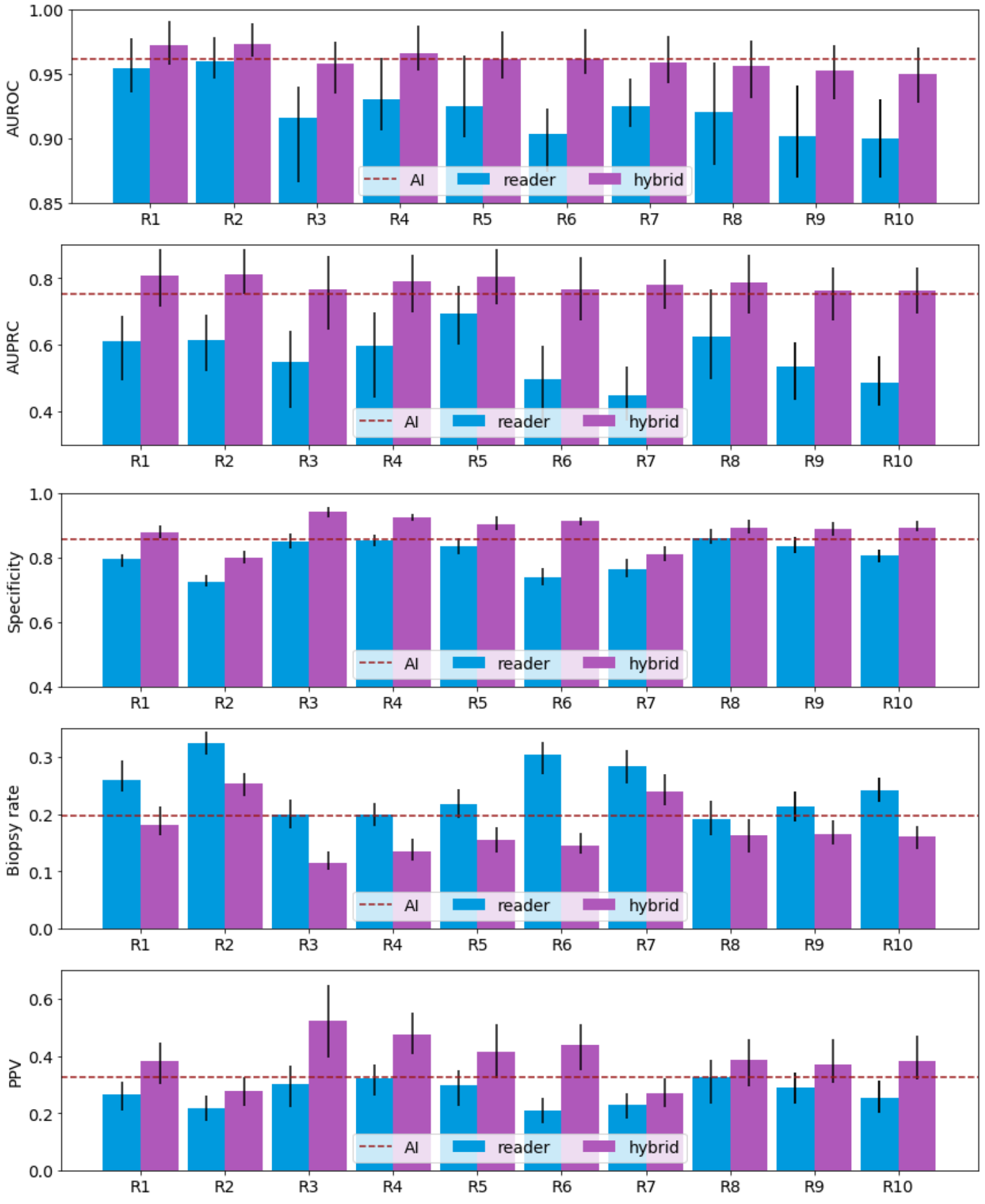
… and breast ultrasound



Shen et al. Artificial Intelligence System Reduces False-Positive Findings in the Interpretation of Breast Ultrasound Exams. 2021



sensitivity
1 - specificity
AI (AUROC, 0.962)
readers
average reader
precision
recall
AI (AUPRC, 0.752)
readers
average reader



Specificity and sensitivity.

Radiologists: 80.7% and 90.1%.

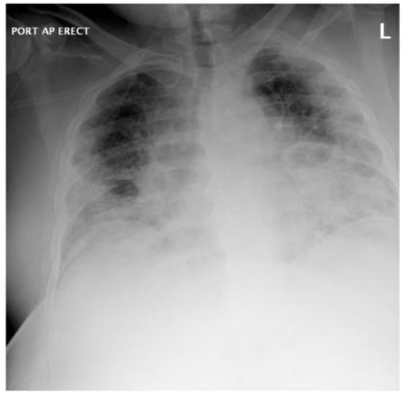
AI's specificity at radiologists' sensitivity: 85.6%

AI's sensitivity at radiologists' specificity: 94.5%

… and COVID-19 deterioration

Example 1

Example 2



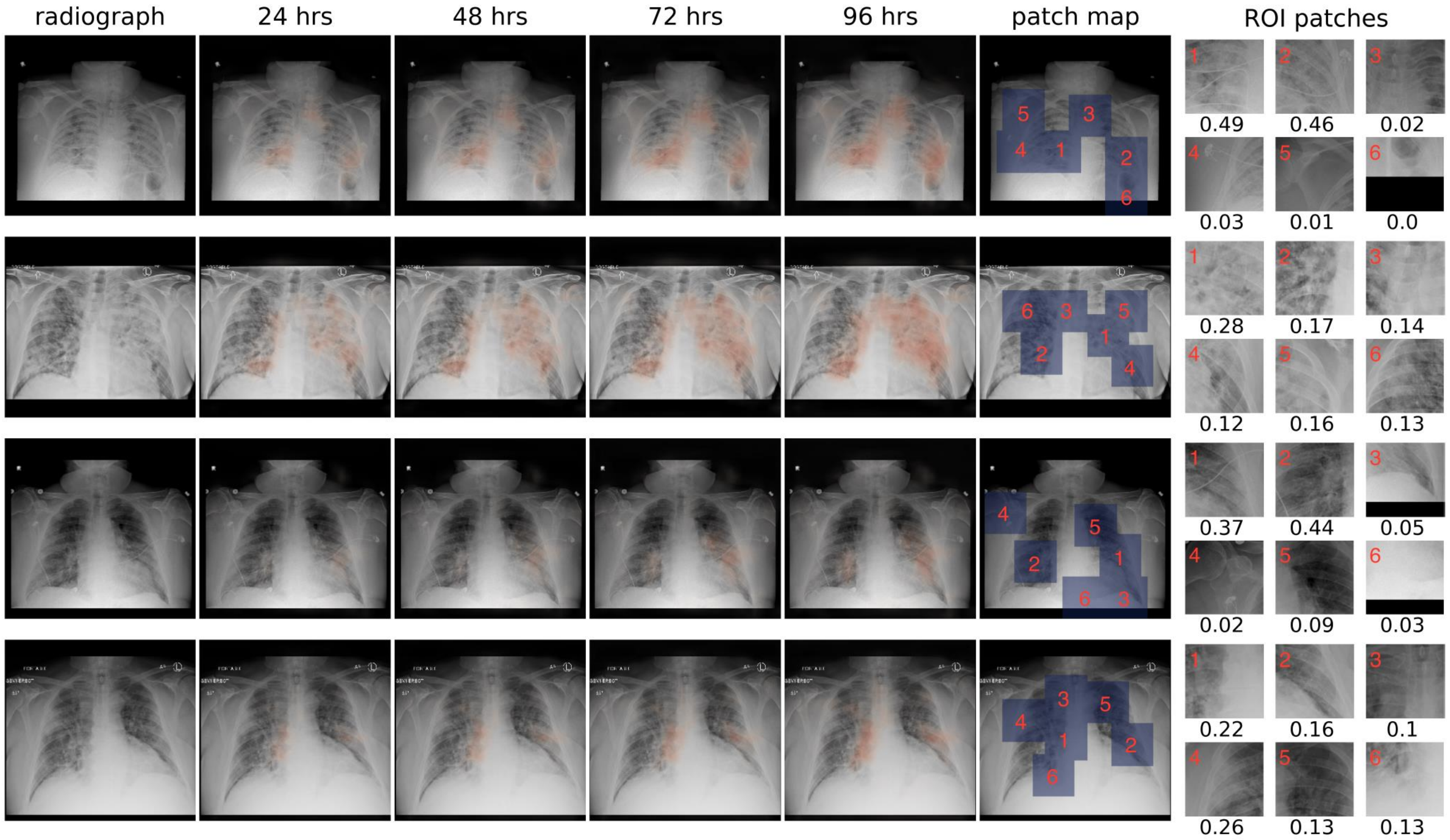
Example 3

Example 4

Example 5

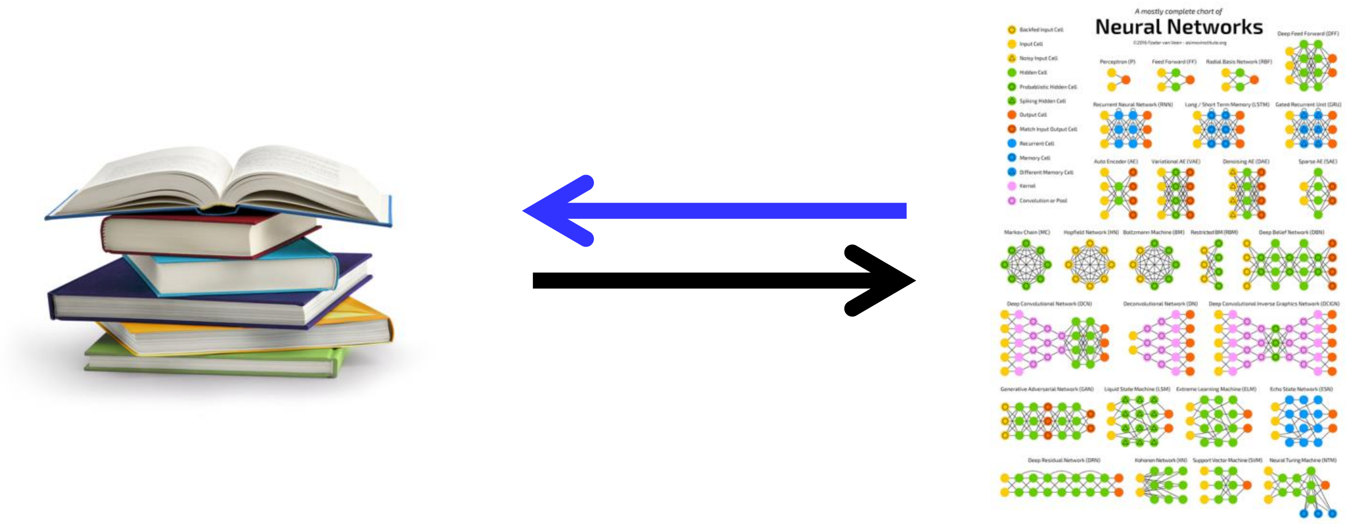
Example 6

Will this person deteriorate in 24, 48, 72 or 96 hours?



Shamout et al. An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department. npj Digital Medicine 2021

Discovering knowledge through machine learning



Conclusions

- Explainability is important in applications such as life sciences and medicine.
- Explainable models, such as the Globally-Aware Multiple Instance Classifier (GMIC) will be used a lot in the future.
- AI is soon going to be very clearly superhuman in medical imaging tasks.
- Discovering knowledge on biological and physical processes through explainable neural networks will be a hot topic soon.

Thank you!

kjgeras

k.j.geras@nyu.edu