Explainable AI in medical imaging

Bas van der Velden, PhD

Head of Data Science
Wageningen Food Safety Research

WAGENINGEN
UNIVERSITY & RESEARCH

@basvandervelden

Black boxes

This is a horse!

This is a horse!

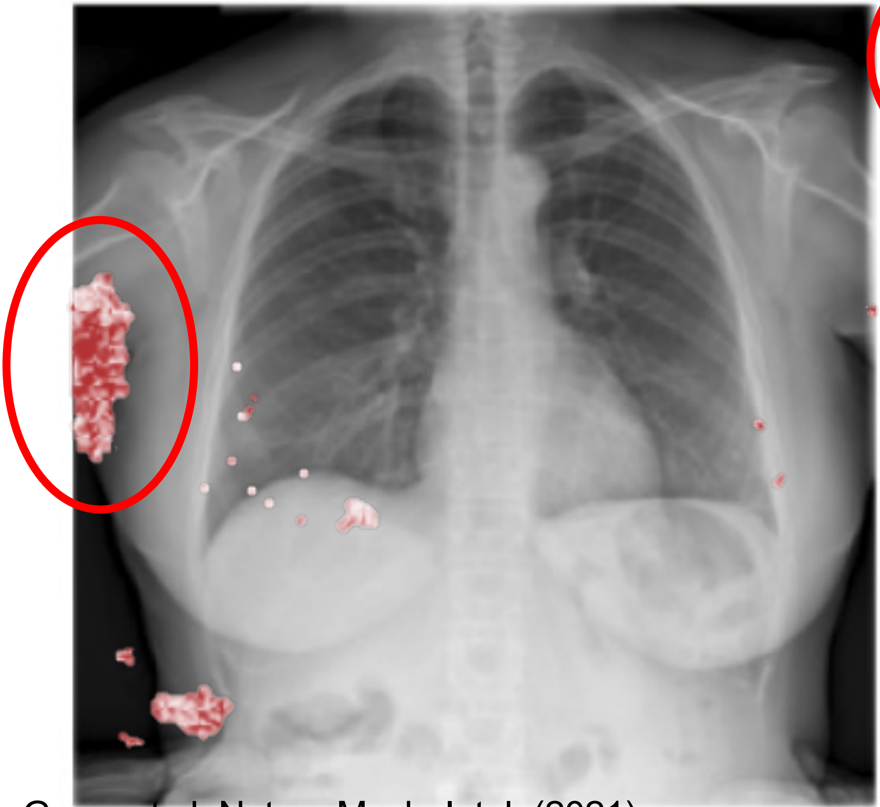
This is NOT a horse!

This is NOT a horse!

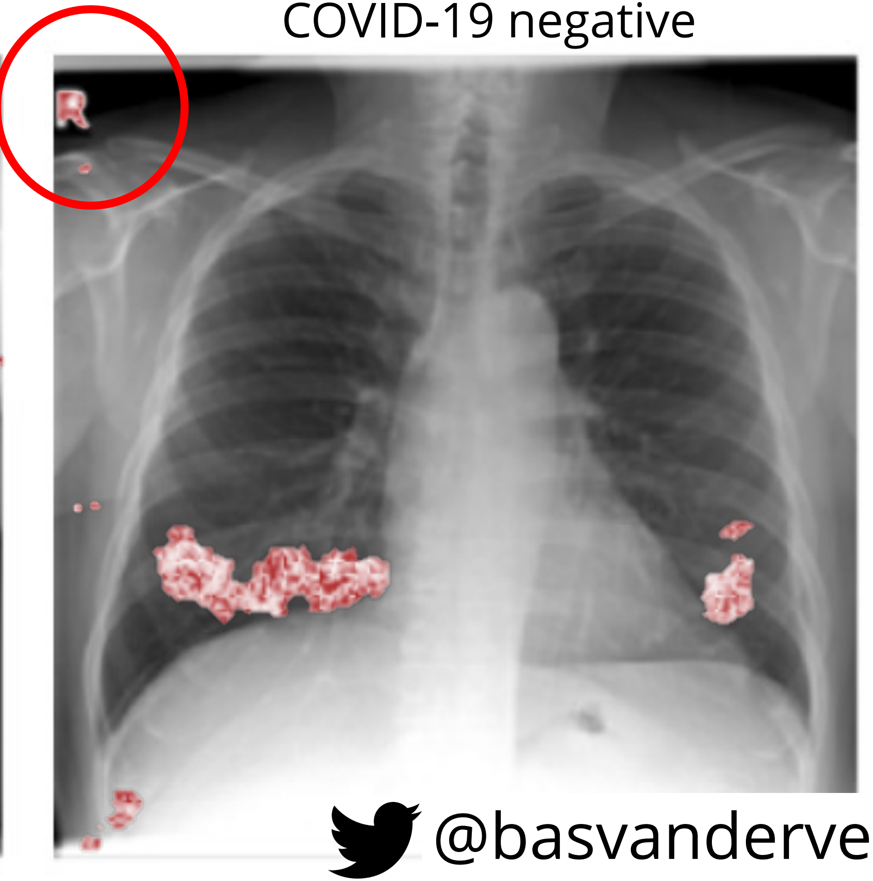
Lapuschkin et al. Nature communications 10.1 (2019): 1-8.

@basvandervelden

Black boxes in medicine



COVID-19 positive



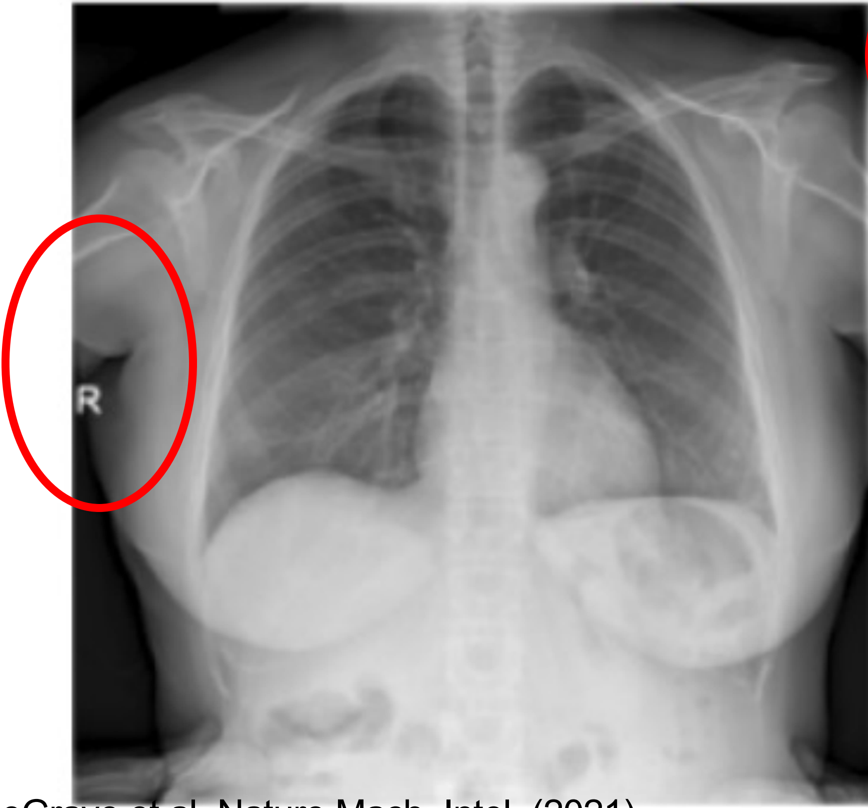
COVID-19 negative

DeGrave et al. Nature Mach. Intel. (2021)

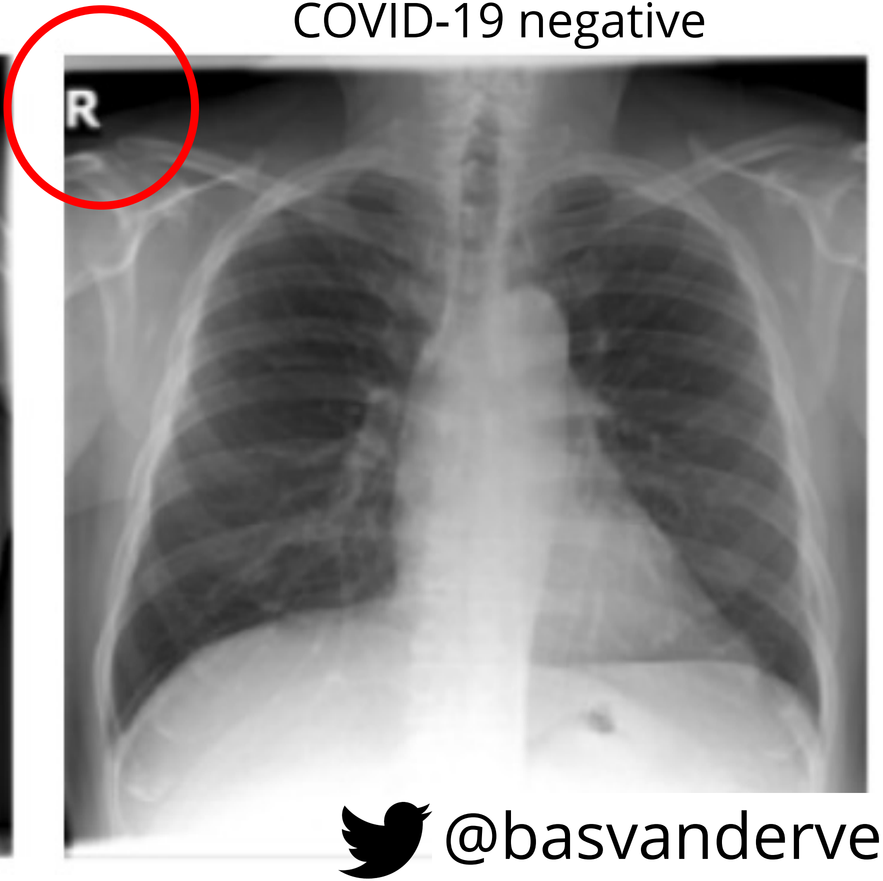
@basvandervelden

Black boxes in medicine

COVID-19 positive

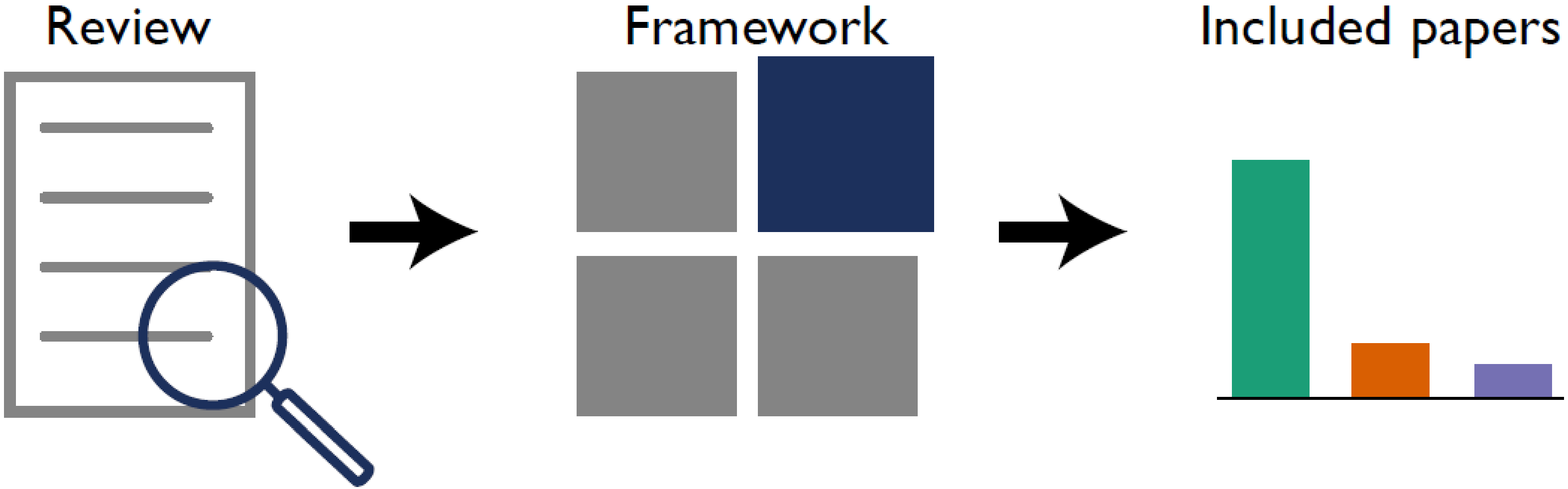


COVID-19 negative



DeGrave et al. Nature Mach. Intel. (2021)

Survey XAI in medical imaging



Review

Framework

Included papers

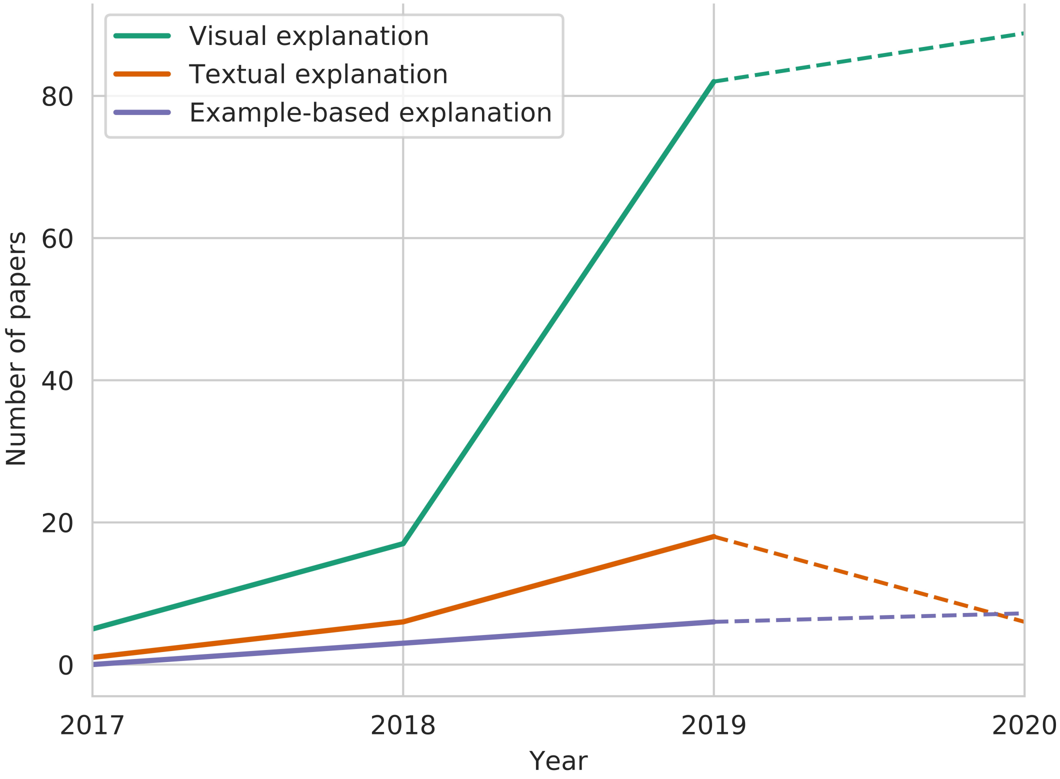
Van der Velden et al. Medical Image Analysis 2022

@basvandervelden

XAI framework

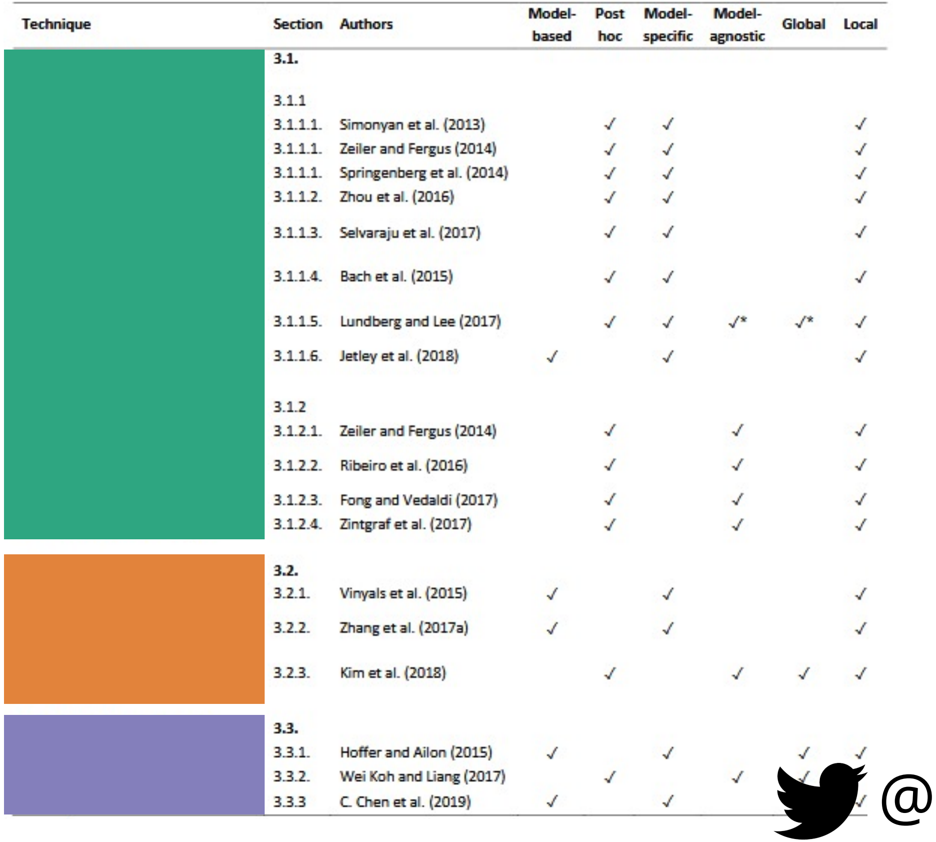
@basvandervelden

XAI in medical imaging



@basvandervelden

Classified to XAI framework



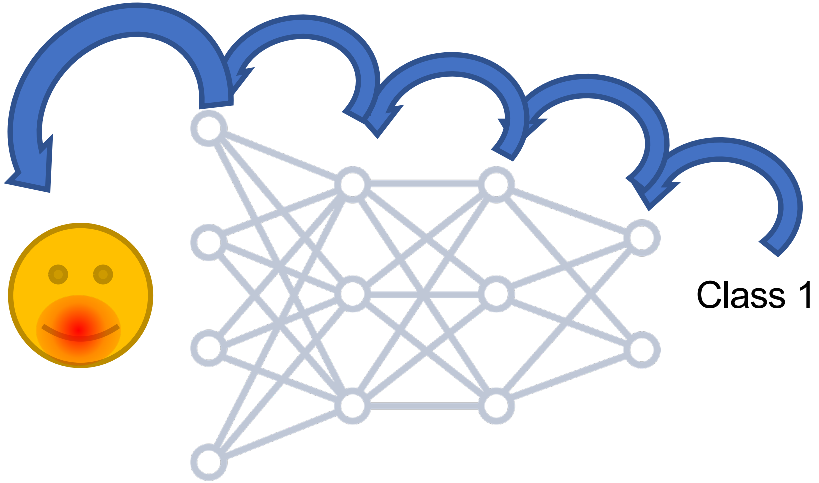
Technique	Section	Authors	Model-based	Post hoc	Model-specific	Model-agnostic	Global	Local
🟩	**3.1.**							
	3.1.1							
	3.1.1.1.	Simonyan et al. (2013)		✓	✓			✓
	3.1.1.1.	Zeiler and Fergus (2014)		✓	✓			✓
	3.1.1.1.	Springenberg et al. (2014)		✓	✓			✓
	3.1.1.2.	Zhou et al. (2016)		✓	✓			✓
	3.1.1.3.	Selvaraju et al. (2017)		✓	✓			✓
	3.1.1.4.	Bach et al. (2015)		✓	✓			✓
	3.1.1.5.	Lundberg and Lee (2017)		✓	✓	✓*	✓*	✓
	3.1.1.6.	Jetley et al. (2018)	✓		✓			✓
	3.1.2							
	3.1.2.1.	Zeiler and Fergus (2014)		✓		✓		✓
	3.1.2.2.	Ribeiro et al. (2016)		✓		✓		✓
	3.1.2.3.	Fong and Vedaldi (2017)		✓		✓		✓
	3.1.2.4.	Zintgraf et al. (2017)		✓		✓		✓
🟧	**3.2.**							
	3.2.1.	Vinyals et al. (2015)	✓		✓			✓
	3.2.2.	Zhang et al. (2017a)	✓		✓			✓
	3.2.3.	Kim et al. (2018)		✓		✓	✓	✓
🟪	**3.3.**							
	3.3.1.	Hoffer and Ailon (2015)	✓		✓		✓	✓
	3.3.2.	Wei Koh and Liang (2017)		✓		✓	✓	[illegible]
	3.3.3	C. Chen et al. (2019)	✓		✓		[illegible]	✓

@basvandervelden

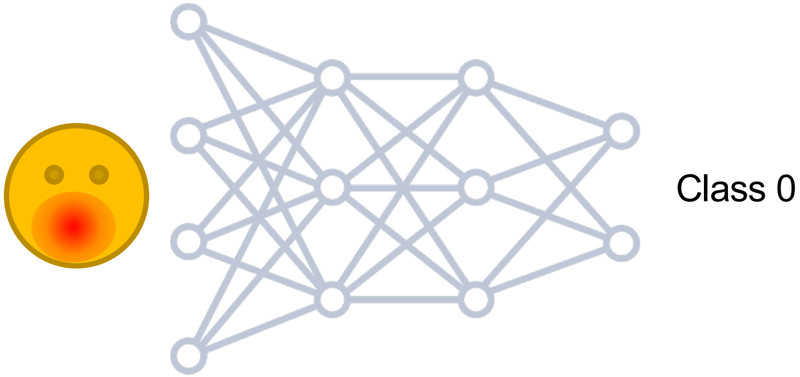
Visual explainable AI

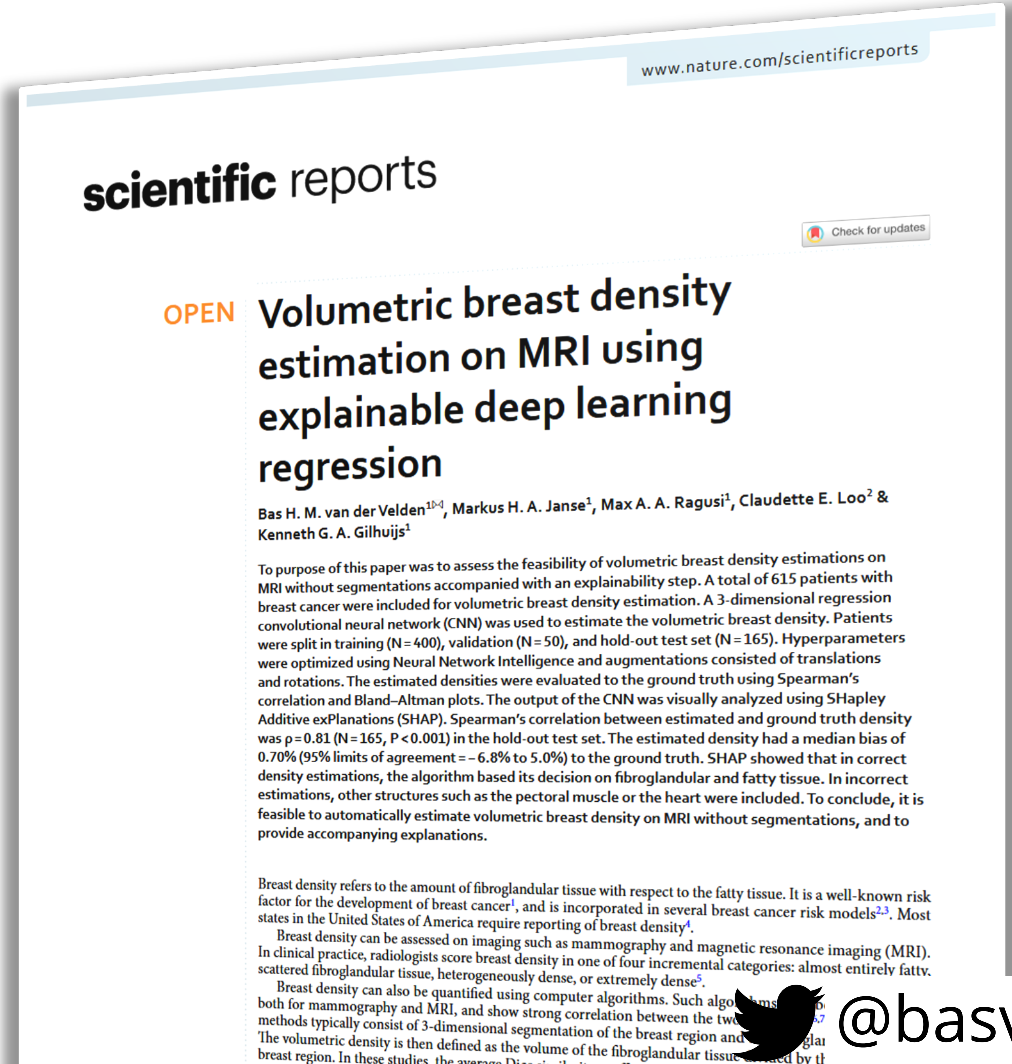
Visual explanation

Backpropagation-based



Perturbation-based





www.nature.com/scientificreports

scientific reports

OPEN

Volumetric breast density estimation on MRI using explainable deep learning regression

Bas H. M. van der Velden[1], Markus H. A. Janse[1], Max A. A. Ragusi[1], Claudette E. Loo[2] & Kenneth G. A. Gilhuijs[1]

To purpose of this paper was to assess the feasibility of volumetric breast density estimations on MRI without segmentations accompanied with an explainability step. A total of 615 patients with breast cancer were included for volumetric breast density estimation. A 3-dimensional regression convolutional neural network (CNN) was used to estimate the volumetric breast density. Patients were split in training (N = 400), validation (N = 50), and hold-out test set (N = 165). Hyperparameters were optimized using Neural Network Intelligence and augmentations consisted of translations and rotations. The estimated densities were evaluated to the ground truth using Spearman's correlation and Bland–Altman plots. The output of the CNN was visually analyzed using SHapley Additive exPlanations (SHAP). Spearman's correlation between estimated and ground truth density was ρ = 0.81 (N = 165, P < 0.001) in the hold-out test set. The estimated density had a median bias of 0.70% (95% limits of agreement = −6.8% to 5.0%) to the ground truth. SHAP showed that in correct density estimations, the algorithm based its decision on fibroglandular and fatty tissue. In incorrect estimations, other structures such as the pectoral muscle or the heart were included. To conclude, it is feasible to automatically estimate volumetric breast density on MRI without segmentations, and to provide accompanying explanations.

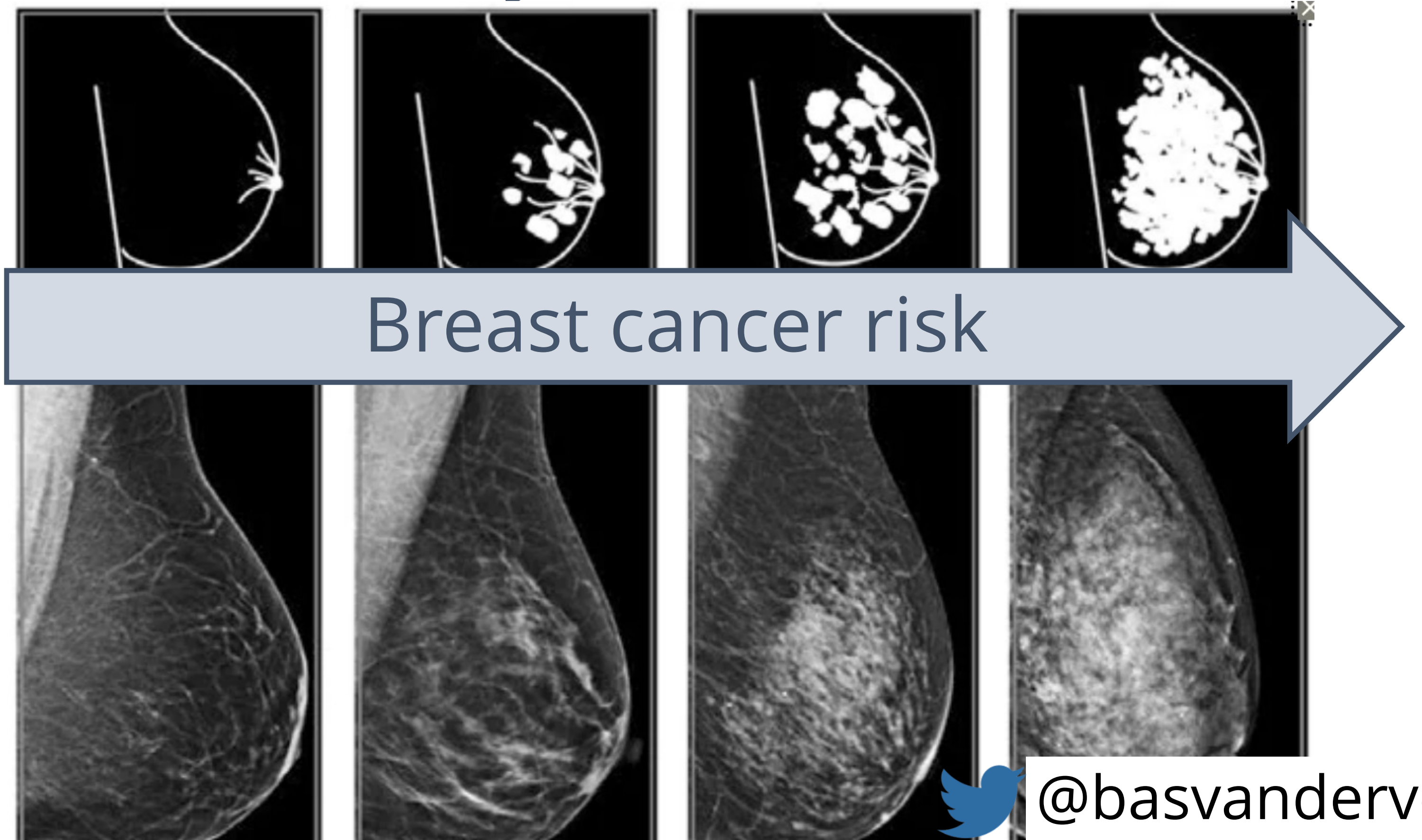
Breast density refers to the amount of fibroglandular tissue with respect to the fatty tissue. It is a well-known risk factor for the development of breast cancer[1], and is incorporated in several breast cancer risk models[2,3]. Most states in the United States of America require reporting of breast density[4].

Breast density can be assessed on imaging such as mammography and magnetic resonance imaging (MRI). In clinical practice, radiologists score breast density in one of four incremental categories: almost entirely fatty, scattered fibroglandular tissue, heterogeneously dense, or extremely dense[5].

Breast density can also be quantified using computer algorithms. Such algo
both for mammography and MRI, and show strong correlation between the tw
methods typically consist of 3-dimensional segmentation of the breast region and
The volumetric density is then defined as the volume of the fibroglandular tissue
breast region. In these studies, the average

@basvandervelden

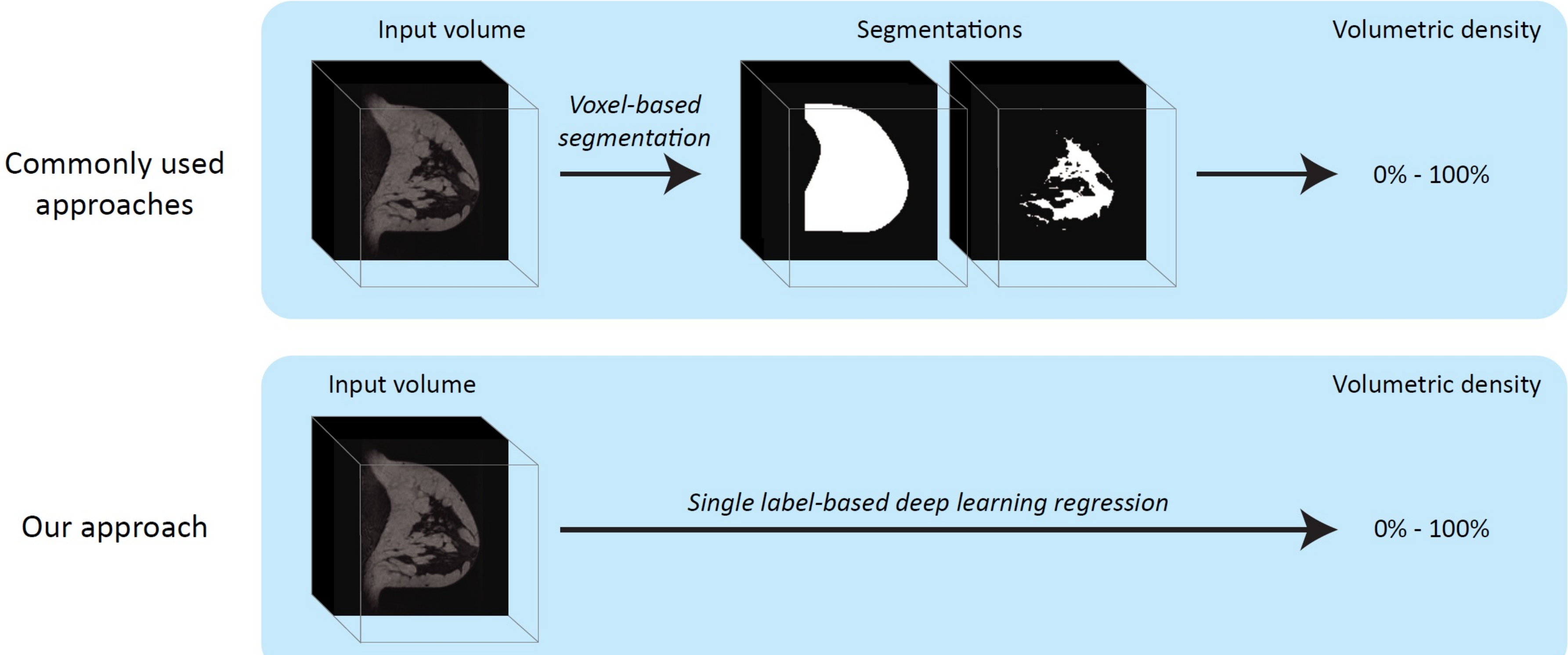
Breast density



Breast cancer risk

@basvandervelden

Regression



Commonly used approaches

Input volume

Voxel-based segmentation

Segmentations

Volumetric density

0% - 100%

Our approach

Input volume

Single label-based deep learning regression

Volumetric density

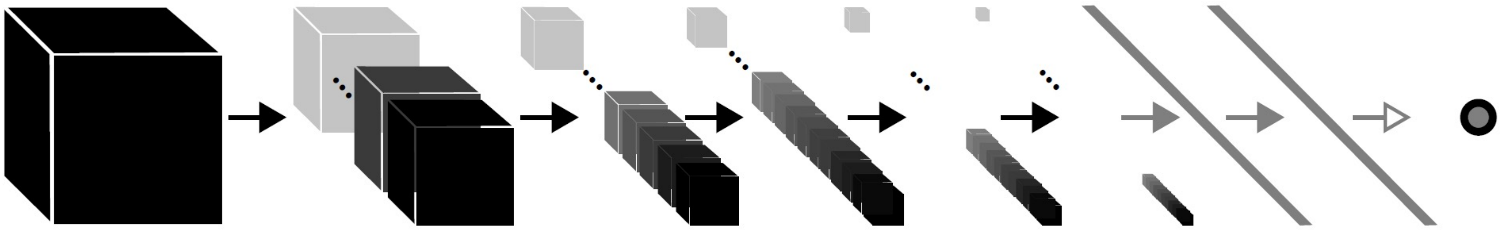
0% - 100%

Van der Velden et al. Nature Sci Rep 2020

@basvandervelden

Regression CNN

Layer	Input	1	2	3	4	5	6	7	Output
Channels	1	32	64	128	128	128	128	128	1 (linear)



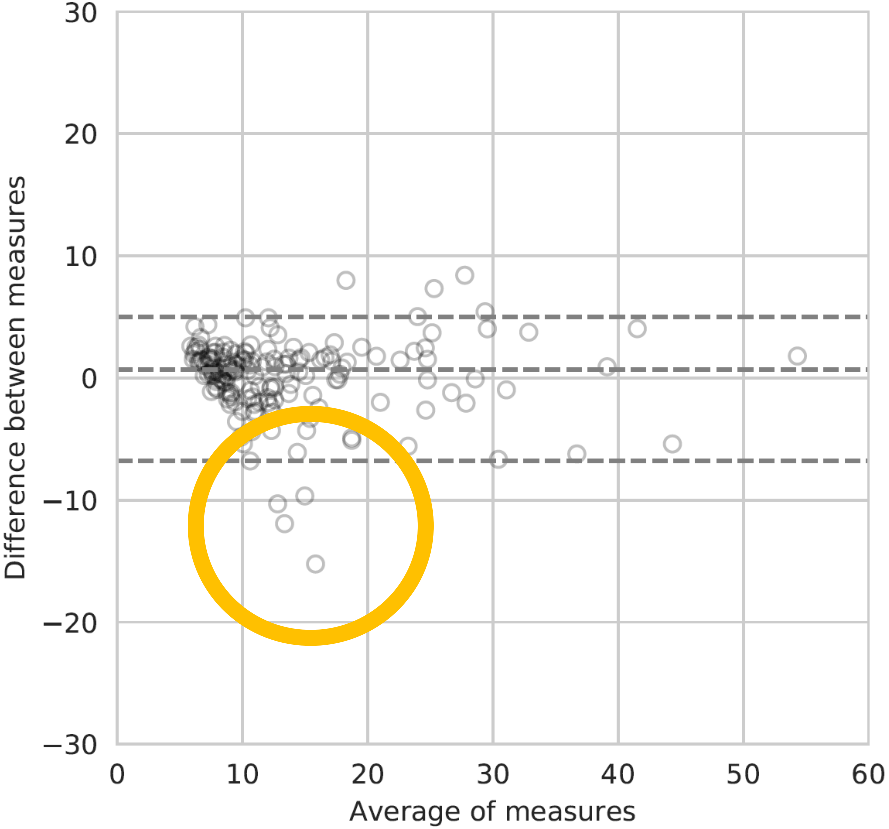
Legend

- → 3×3×3 convolution, 2×2×2 strides, PReLu activation
- → Fully connected, PReLu activation
- ▷ Fully connected, Linear activation

Van der Velden et al. Nature Sci Rep 2020

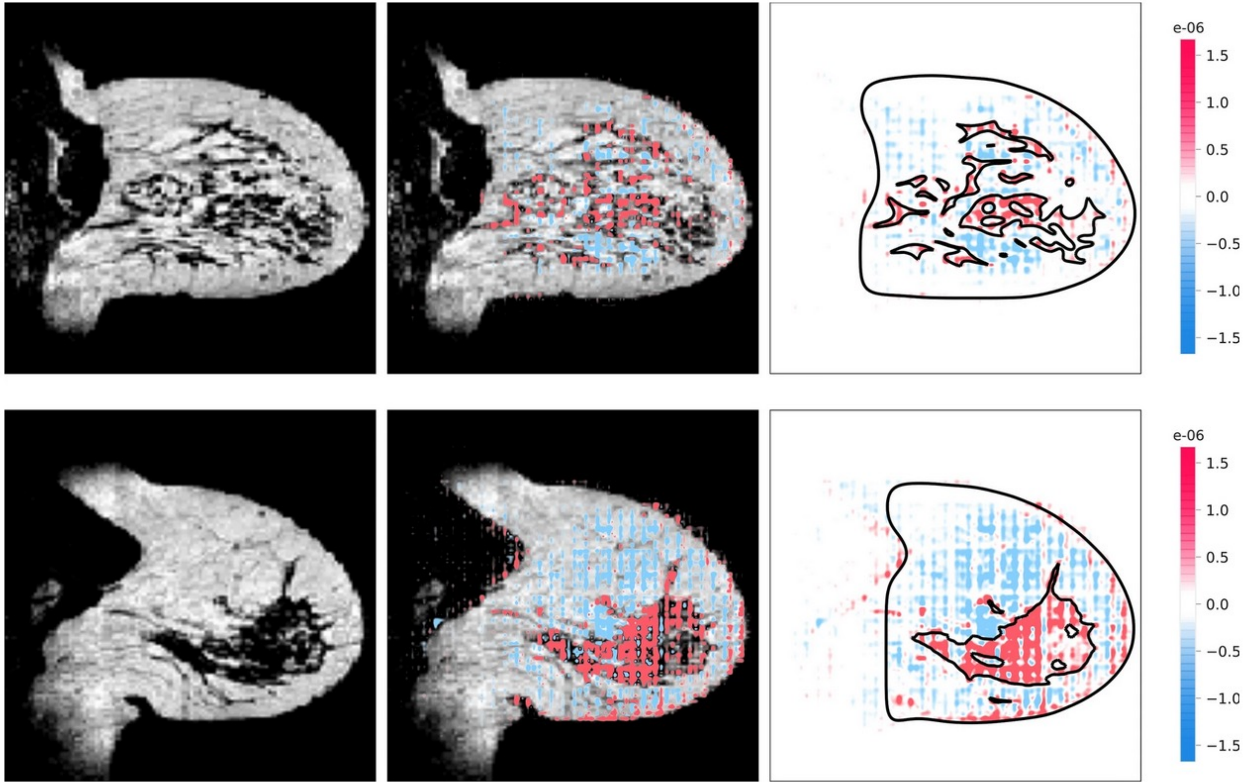
@basvandervelden

Density estimations



@basvandervelden

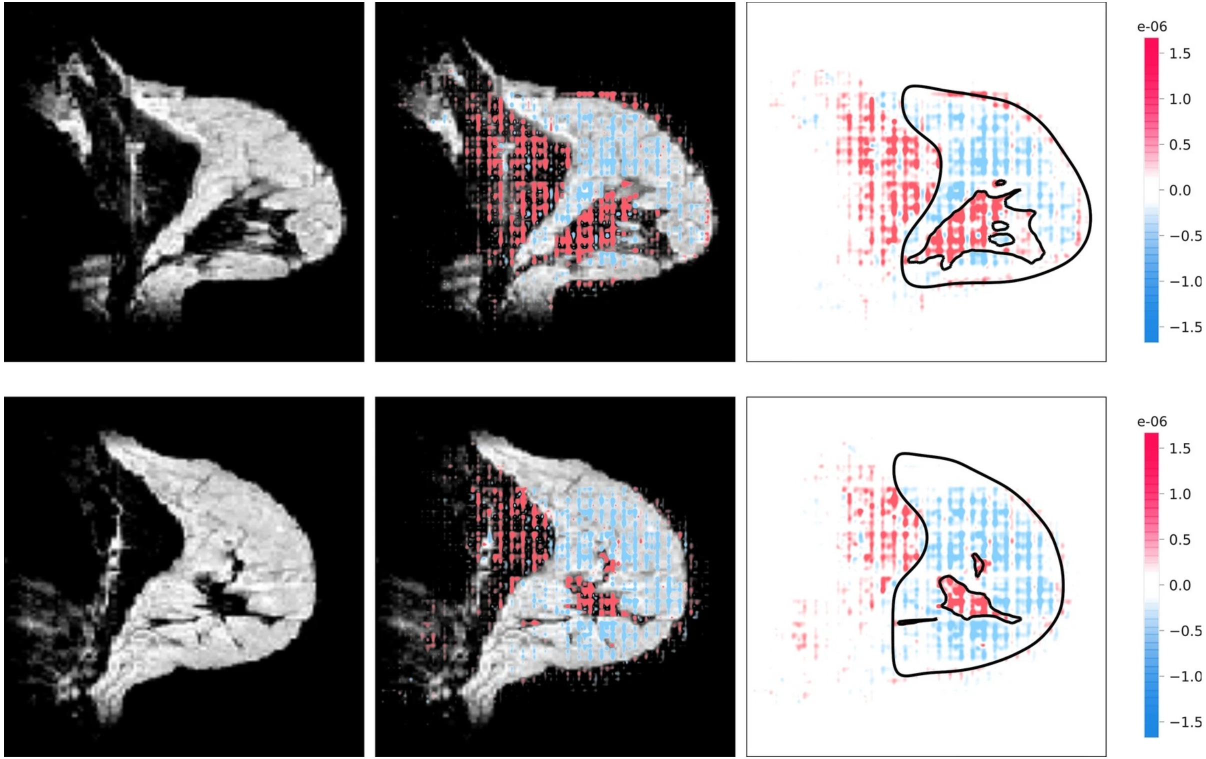
XAI: Deep SHAP



Van der Velden et al. Nature Sci Rep 2020

@basvandervelden

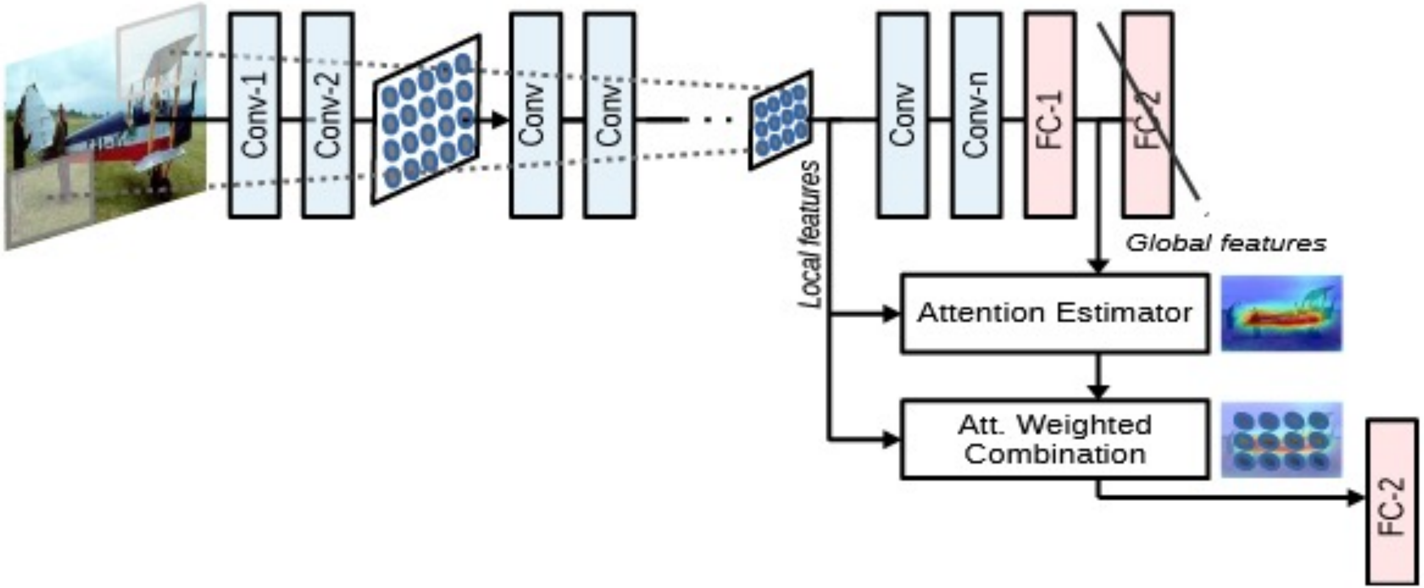
Wrong predictions? XAI!



Van der Velden et al. Nature Sci Rep 2020

@basvandervelden

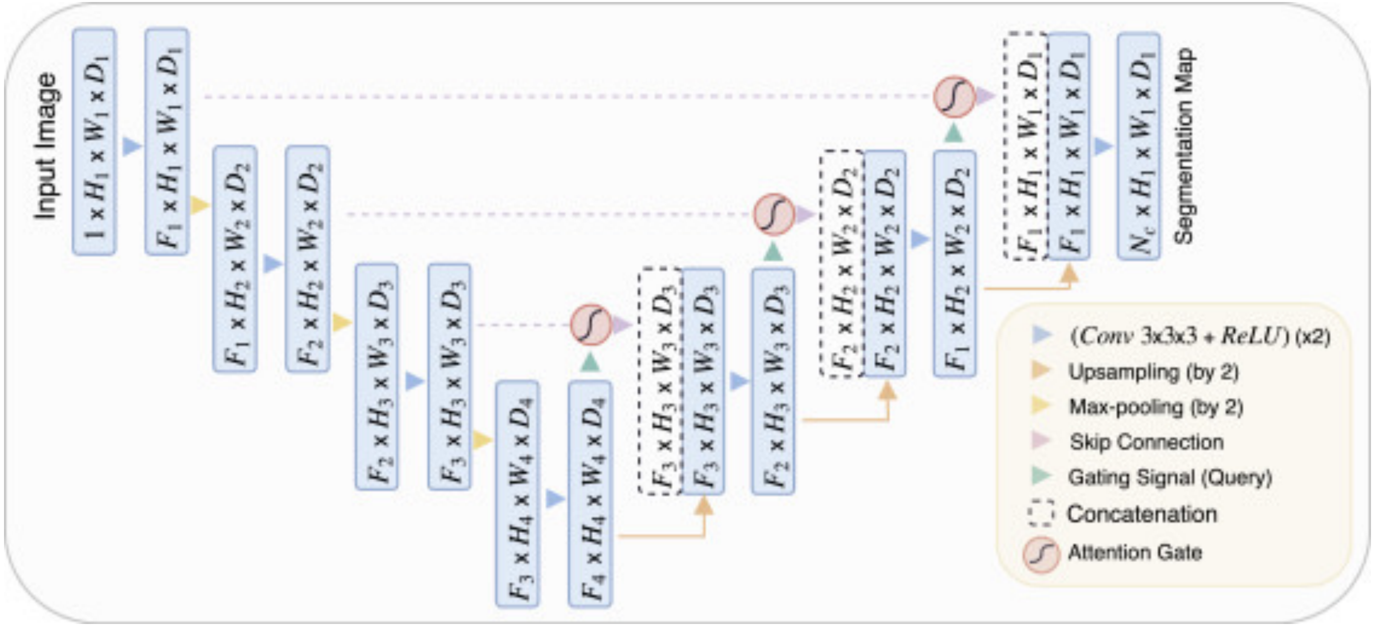
Trainable attention



Jetley et al ICLR 2018

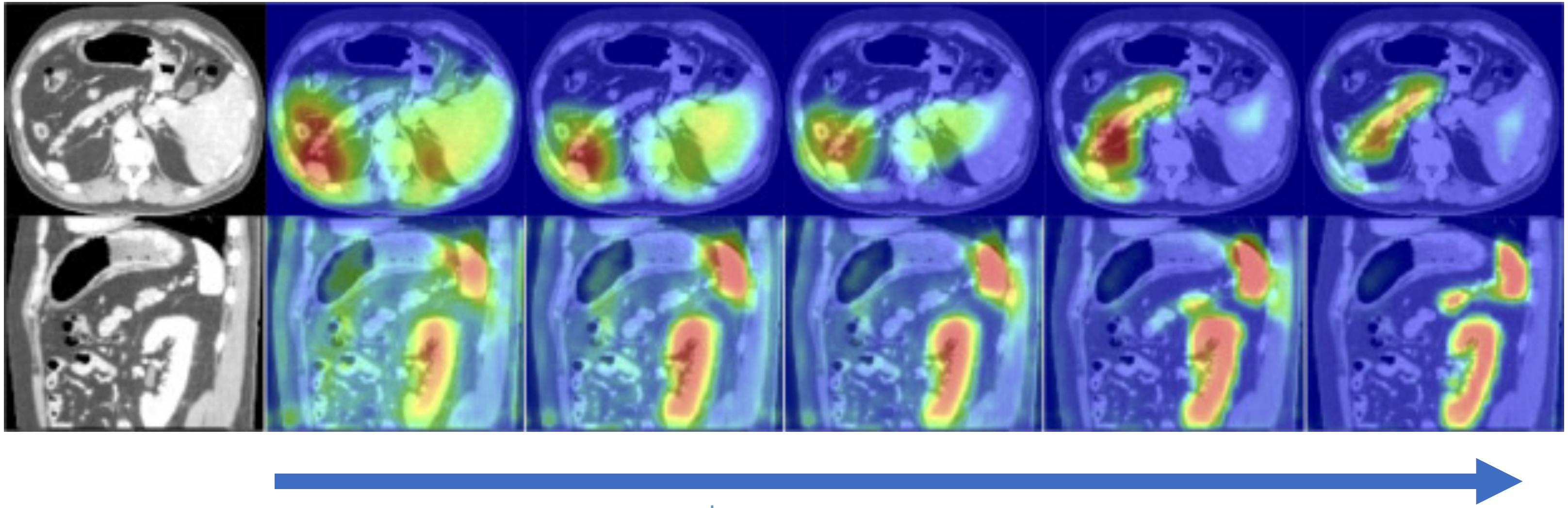
@basvandervelden

Trainable attention in MedIA


Input Image
$1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_2 \times W_2 \times D_2$
$F_2 \times H_2 \times W_2 \times D_2$
$F_2 \times H_3 \times W_3 \times D_3$
$F_3 \times H_3 \times W_3 \times D_3$
$F_3 \times H_4 \times W_4 \times D_4$
$F_4 \times H_4 \times W_4 \times D_4$
$F_3 \times H_3 \times W_3 \times D_3$
$F_3 \times H_3 \times W_3 \times D_3$
$F_2 \times H_3 \times W_3 \times D_3$
$F_2 \times H_2 \times W_2 \times D_2$
$F_2 \times H_2 \times W_2 \times D_2$
$F_1 \times H_2 \times W_2 \times D_2$
$F_1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_1 \times W_1 \times D_1$
$N_c \times H_1 \times W_1 \times D_1$
Segmentation Map
(Conv 3x3x3 + ReLU) (x2)
Upsampling (by 2)
Max-pooling (by 2)
Skip Connection
Gating Signal (Query)
Concatenation
Attention Gate

Schelmper et al MedIA 2019

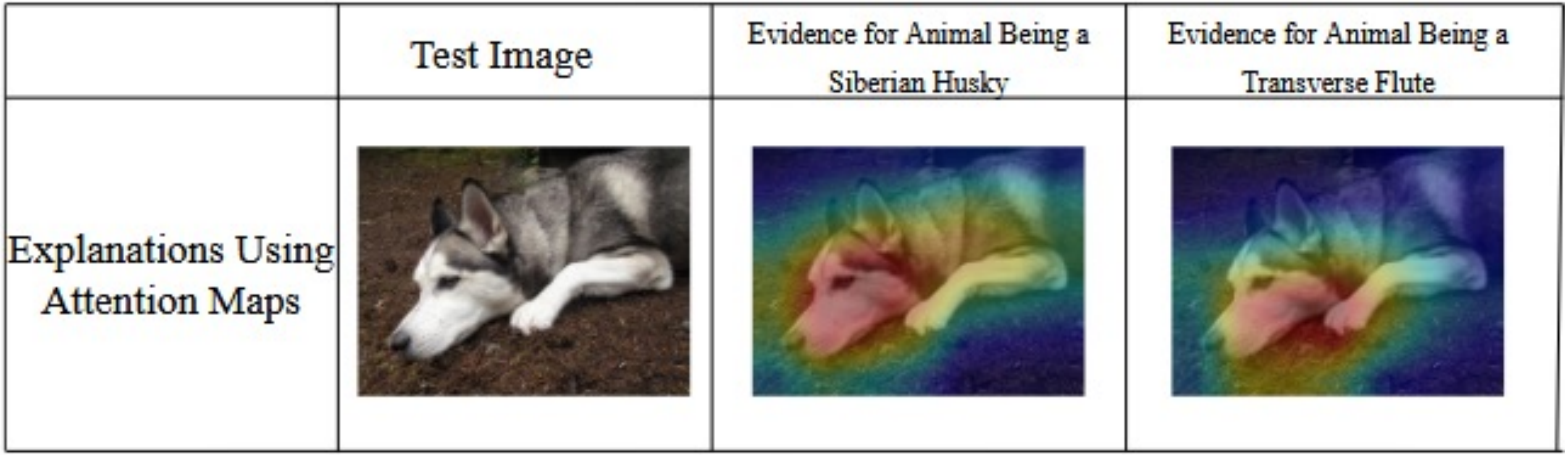
Trainable attention



epochs

Schelmper et al MedIA 2019

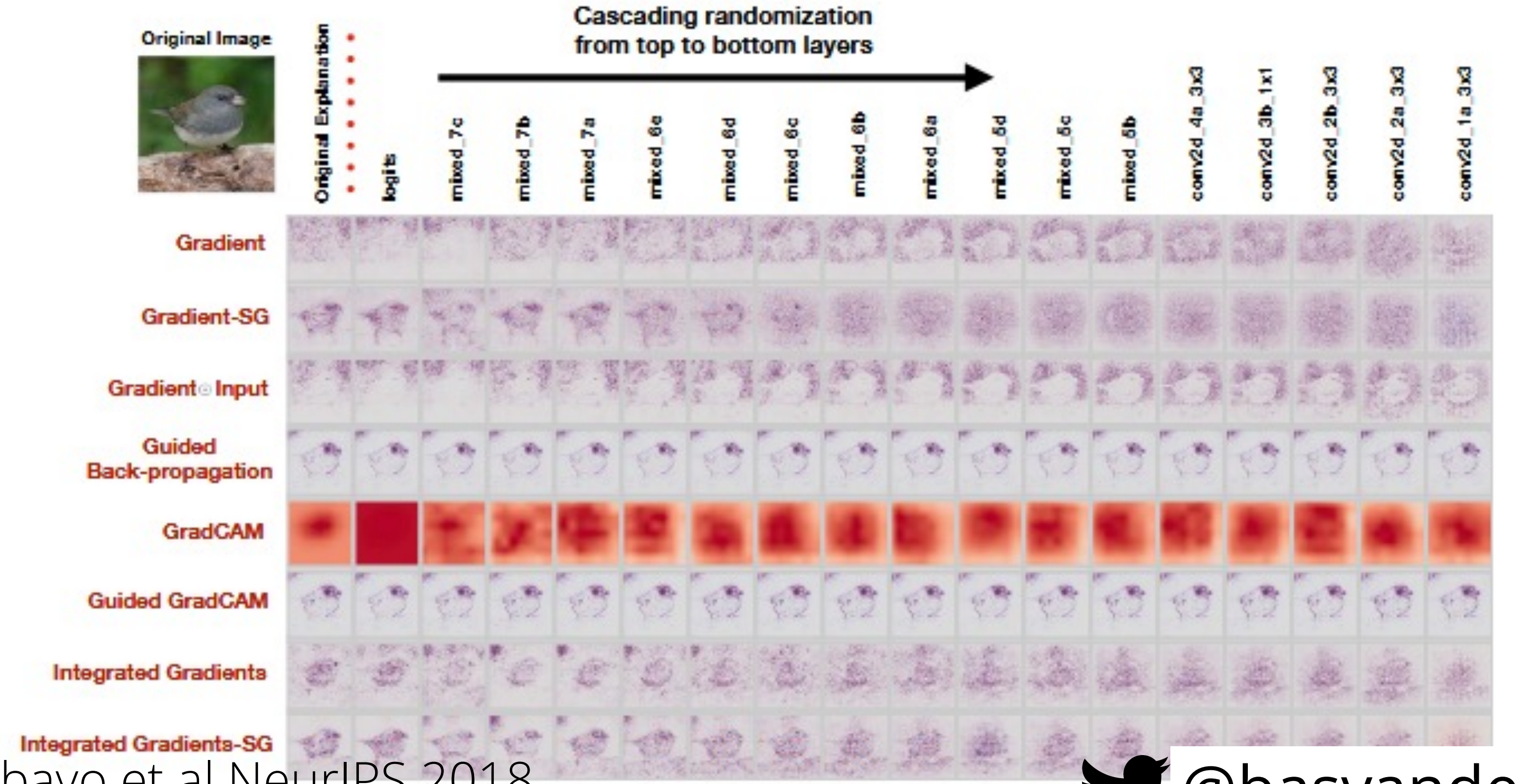
Visual may be misleading



	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

Rudin Nature Mach. Intel. 2019

Visual may focus on edges

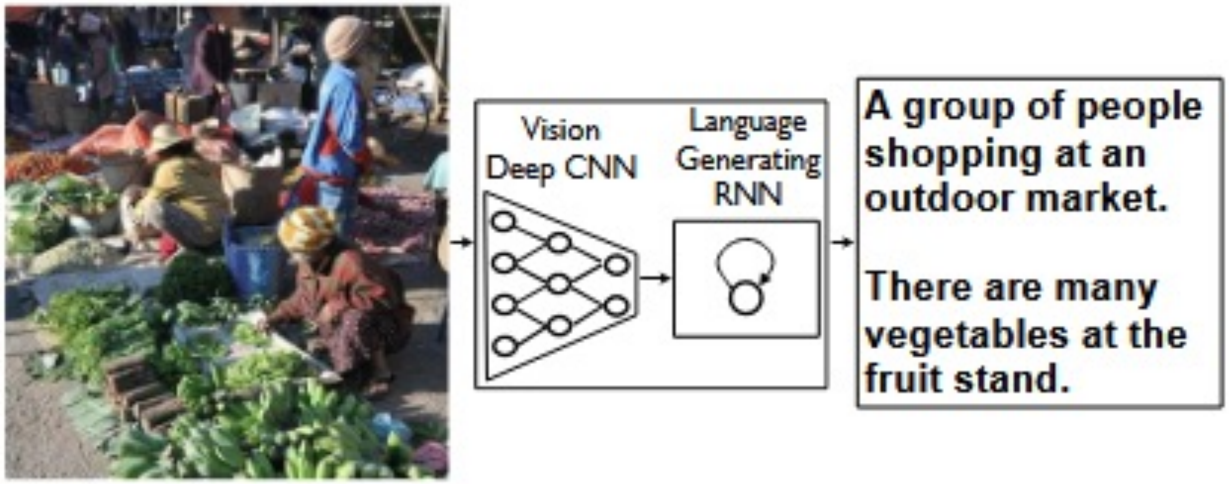


Adebayo et al NeurIPS 2018

@basvandervelden

Textual explainable AI

Image captioning

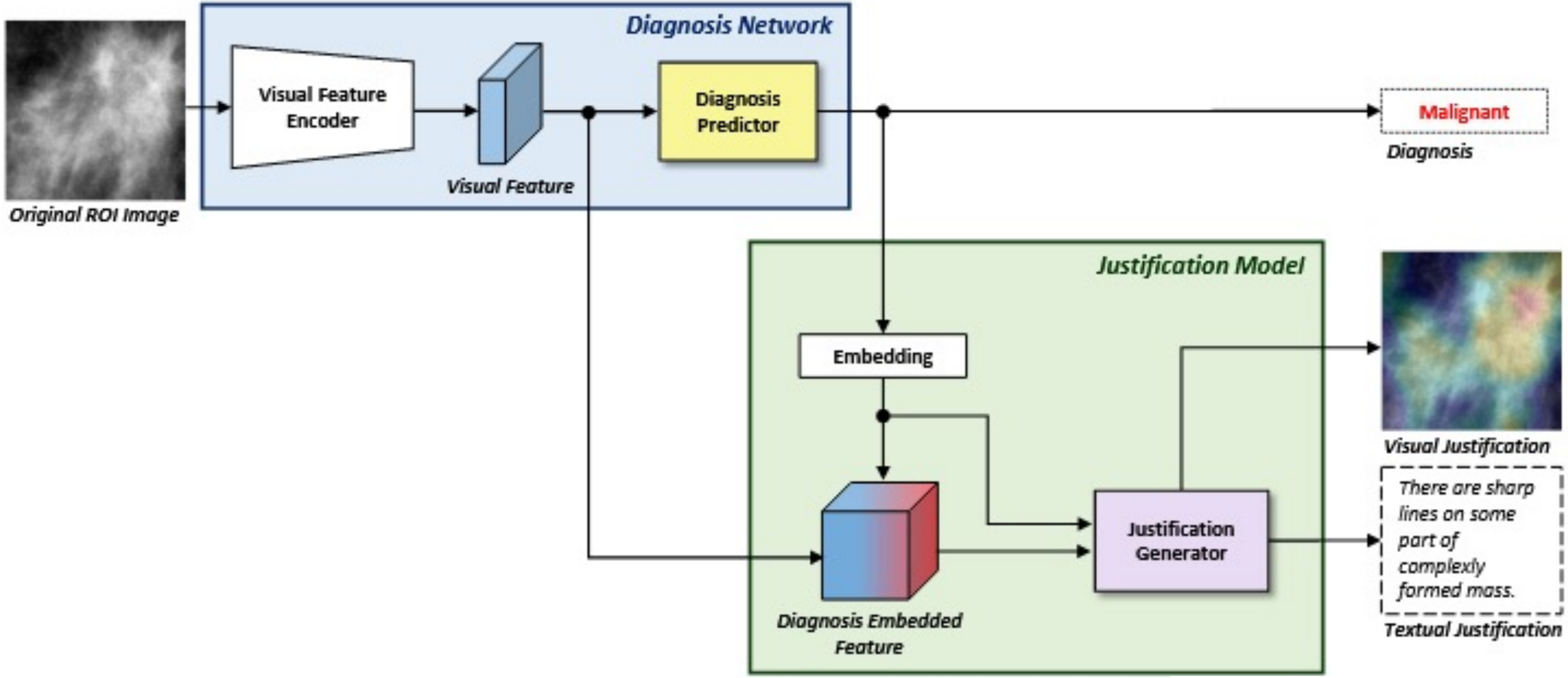


A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.

Vinyals et al. CVPR 2015

Image captioning with visual XAI



Lee et al. iMIMIC 2019

@basvandervelden

Example-based explainable AI

Case-based reasoning

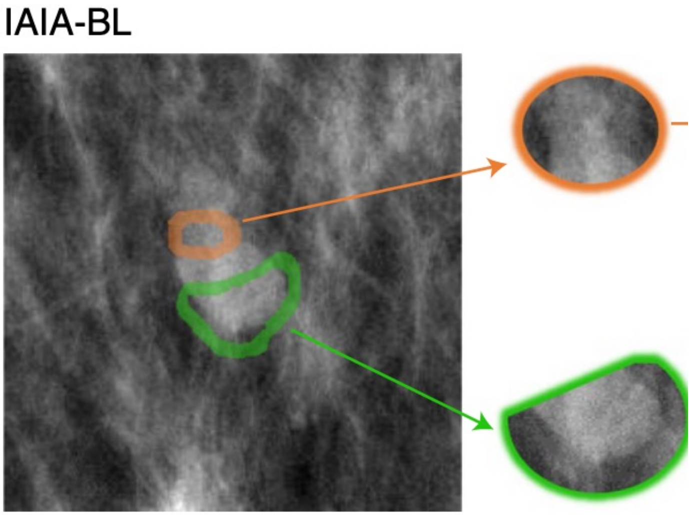
@basvandervelden

Case-based reasoning

Chen et al. NeurIPS 2019

@basvandervelden

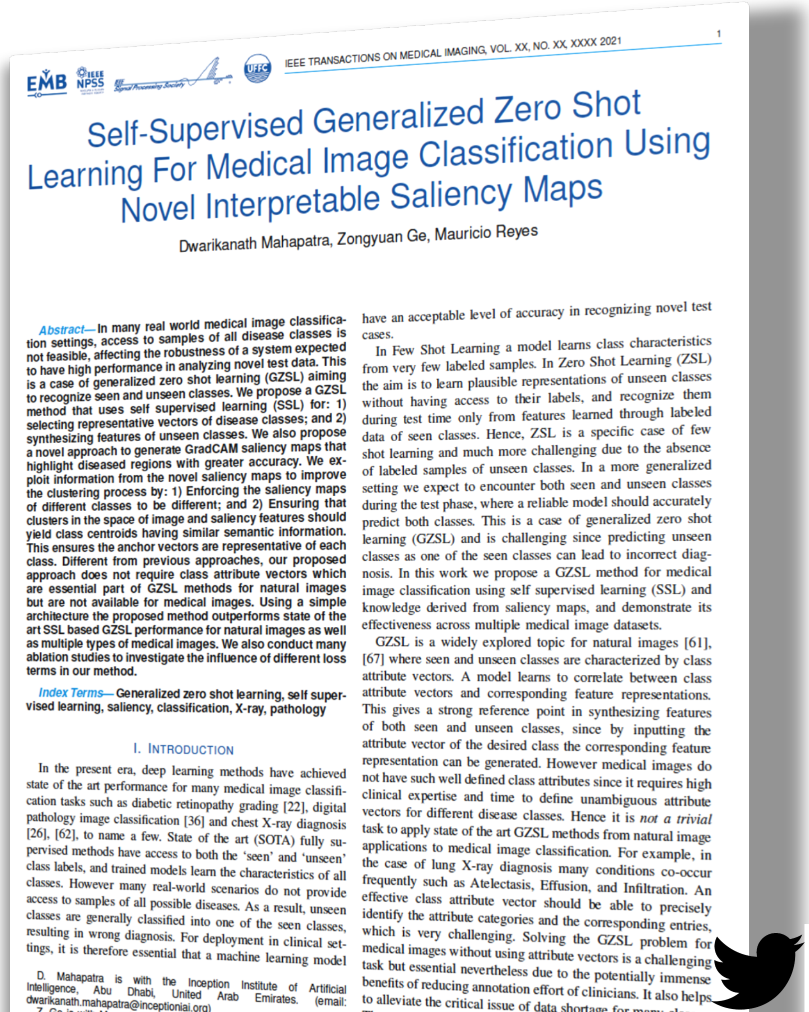
Case-based reasoning



Barnett et al. Nature Mach. Intell. 2021

Explainable AI to improve learning

@basvandervelden



IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. XX, NO. XX, XXXX 2021 1

Self-Supervised Generalized Zero Shot Learning For Medical Image Classification Using Novel Interpretable Saliency Maps

Dwarikanath Mahapatra, Zongyuan Ge, Mauricio Reyes

Abstract—**In many real world medical image classification settings, access to samples of all disease classes is not feasible, affecting the robustness of a system expected to have high performance in analyzing novel test data. This is a case of generalized zero shot learning (GZSL) aiming to recognize seen and unseen classes. We propose a GZSL method that uses self supervised learning (SSL) for: 1) selecting representative vectors of disease classes; and 2) synthesizing features of unseen classes. We also propose a novel approach to generate GradCAM saliency maps that highlight diseased regions with greater accuracy. We exploit information from the novel saliency maps to improve the clustering process by: 1) Enforcing the saliency maps of different classes to be different; and 2) Ensuring that clusters in the space of image and saliency features should yield class centroids having similar semantic information. This ensures the anchor vectors are representative of each class. Different from previous approaches, our proposed approach does not require class attribute vectors which are essential part of GZSL methods for natural images but are not available for medical images. Using a simple architecture the proposed method outperforms state of the art SSL based GZSL performance for natural images as well as multiple types of medical images. We also conduct many ablation studies to investigate the influence of different loss terms in our method.**

Index Terms—**Generalized zero shot learning, self supervised learning, saliency, classification, X-ray, pathology**

I. Introduction

In the present era, deep learning methods have achieved state of the art performance for many medical image classification tasks such as diabetic retinopathy grading [22], digital pathology image classification [36] and chest X-ray diagnosis [26], [62], to name a few. State of the art (SOTA) fully supervised methods have access to both the 'seen' and 'unseen' class labels, and trained models learn the characteristics of all classes. However many real-world scenarios do not provide access to samples of all possible diseases. As a result, unseen classes are generally classified into one of the seen classes, resulting in wrong diagnosis. For deployment in clinical settings, it is therefore essential that a machine learning model have an acceptable level of accuracy in recognizing novel test cases.

In Few Shot Learning a model learns class characteristics from very few labeled samples. In Zero Shot Learning (ZSL) the aim is to learn plausible representations of unseen classes without having access to their labels, and recognize them during test time only from features learned through labeled data of seen classes. Hence, ZSL is a specific case of few shot learning and much more challenging due to the absence of labeled samples of unseen classes. In a more generalized setting we expect to encounter both seen and unseen classes during the test phase, where a reliable model should accurately predict both classes. This is a case of generalized zero shot learning (GZSL) and is challenging since predicting unseen classes as one of the seen classes can lead to incorrect diagnosis. In this work we propose a GZSL method for medical image classification using self supervised learning (SSL) and knowledge derived from saliency maps, and demonstrate its effectiveness across multiple medical image datasets.

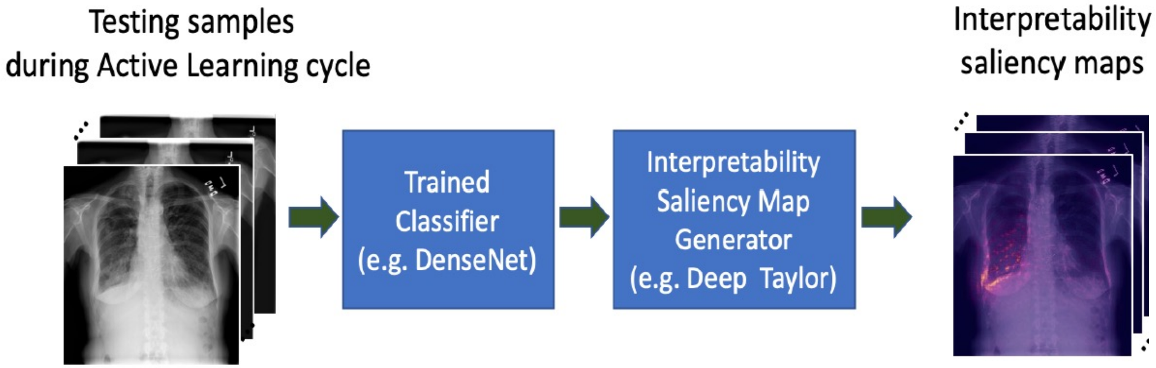
GZSL is a widely explored topic for natural images [61], [67] where seen and unseen classes are characterized by class attribute vectors. A model learns to correlate between class attribute vectors and corresponding feature representations. This gives a strong reference point in synthesizing features of both seen and unseen classes, since by inputting the attribute vector of the desired class the corresponding feature representation can be generated. However medical images do not have such well defined class attributes since it requires high clinical expertise and time to define unambiguous attribute vectors for different disease classes. Hence it is *not a trivial* task to apply state of the art GZSL methods from natural image applications to medical image classification. For example, in the case of lung X-ray diagnosis many conditions co-occur frequently such as Atelectasis, Effusion, and Infiltration. An effective class attribute vector should be able to precisely identify the attribute categories and the corresponding entries, which is very challenging. Solving the GZSL problem for medical images without using attribute vectors is a challenging task but essential nevertheless due to the potentially immense benefits of reducing annotation effort of clinicians. It also helps to alleviate the critical issue of data shortage for many

D. Mahapatra is with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. (email: dwarikanath.mahapatra@inceptionai.org)

Z. Ge is with

@basvandervelden

Active learning using XAI

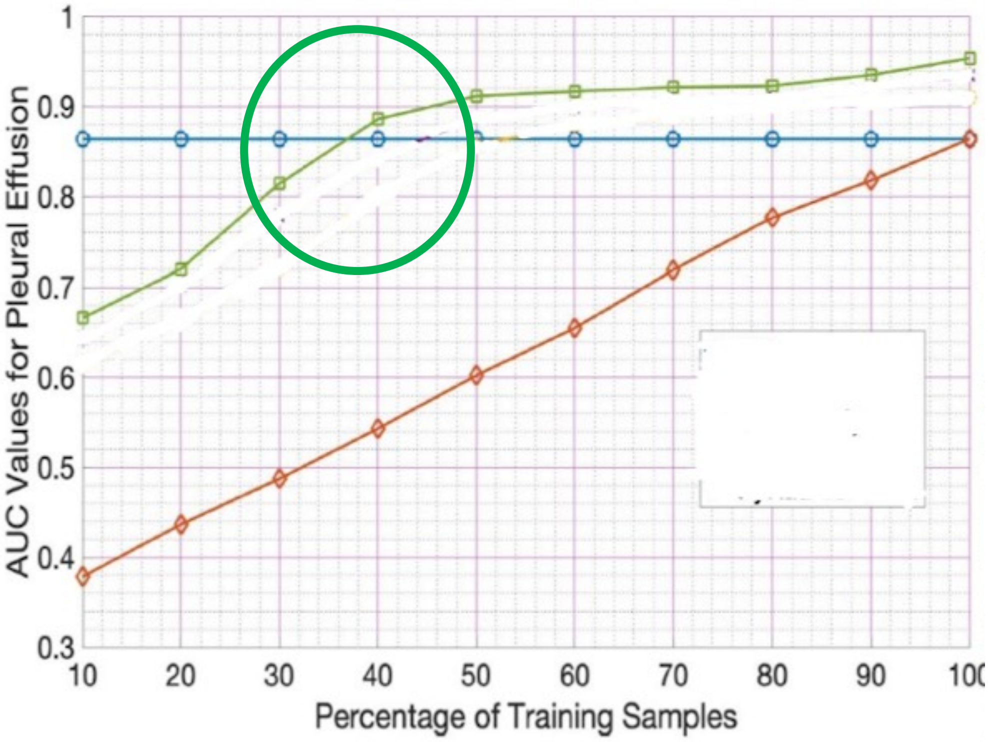


Slide courtesy prof Reyes

Mahapatra et al. IEEE TMI 2022

@basvandervelden

Active learning using XAI



← Active learning

← Fully supervised model

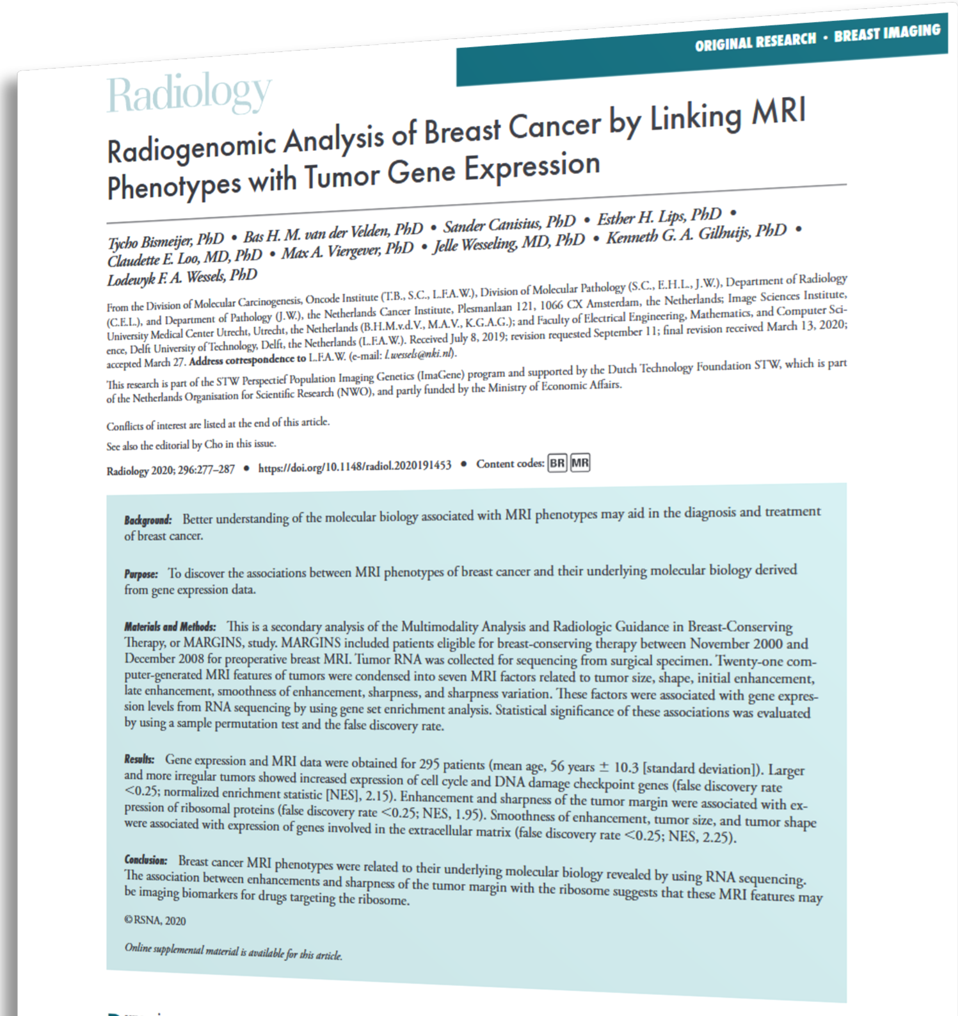
← Random sample selection

Slide courtesy prof Reyes

Mahapatra et al. IEEE TMI 2022

@basvandervelden

Holistic explainable AI



ORIGINAL RESEARCH • BREAST IMAGING

Radiology

Radiogenomic Analysis of Breast Cancer by Linking MRI Phenotypes with Tumor Gene Expression

Tycho Bismeijer, PhD • Bas H. M. van der Velden, PhD • Sander Canisius, PhD • Esther H. Lips, PhD • Claudette E. Loo, MD, PhD • Max A. Viergever, PhD • Jelle Wesseling, MD, PhD • Kenneth G. A. Gilhuijs, PhD • Lodewyk F. A. Wessels, PhD

From the Division of Molecular Carcinogenesis, Oncode Institute (T.B., S.C., L.F.A.W.), Division of Molecular Pathology (S.C., E.H.L., J.W.), Department of Radiology (C.E.L.), and Department of Pathology (J.W.), the Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands; Image Sciences Institute, University Medical Center Utrecht, Utrecht, the Netherlands (B.H.M.v.d.V., M.A.V., K.G.A.G.); and Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Delft, the Netherlands (L.F.A.W.). Received July 8, 2019; revision requested September 11; final revision received March 13, 2020; accepted March 27. **Address correspondence to** L.F.A.W. (e-mail: *l.wessels@nki.nl*).

This research is part of the STW Perspectief Population Imaging Genetics (ImaGene) program and supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research (NWO), and partly funded by the Ministry of Economic Affairs.

Conflicts of interest are listed at the end of this article.

See also the editorial by Cho in this issue.

Radiology 2020; 296:277–287 • https://doi.org/10.1148/radiol.2020191453 • Content codes: BR MR

Background: Better understanding of the molecular biology associated with MRI phenotypes may aid in the diagnosis and treatment of breast cancer.

Purpose: To discover the associations between MRI phenotypes of breast cancer and their underlying molecular biology derived from gene expression data.

Materials and Methods: This is a secondary analysis of the Multimodality Analysis and Radiologic Guidance in Breast-Conserving Therapy, or MARGINS, study. MARGINS included patients eligible for breast-conserving therapy between November 2000 and December 2008 for preoperative breast MRI. Tumor RNA was collected for sequencing from surgical specimen. Twenty-one computer-generated MRI features of tumors were condensed into seven MRI factors related to tumor size, shape, initial enhancement, late enhancement, smoothness of enhancement, sharpness, and sharpness variation. These factors were associated with gene expression levels from RNA sequencing by using gene set enrichment analysis. Statistical significance of these associations was evaluated by using a sample permutation test and the false discovery rate.

Results: Gene expression and MRI data were obtained for 295 patients (mean age, 56 years ± 10.3 [standard deviation]). Larger and more irregular tumors showed increased expression of cell cycle and DNA damage checkpoint genes (false discovery rate <0.25; normalized enrichment statistic [NES], 2.15). Enhancement and sharpness of the tumor margin were associated with expression of ribosomal proteins (false discovery rate <0.25; NES, 1.95). Smoothness of enhancement, tumor size, and tumor shape were associated with expression of genes involved in the extracellular matrix (false discovery rate <0.25; NES, 2.25).

Conclusion: Breast cancer MRI phenotypes were related to their underlying molecular biology revealed by using RNA sequencing. The association between enhancements and sharpness of the tumor margin with the ribosome suggests that these MRI features may be imaging biomarkers for drugs targeting the ribosome.

Online supplemental material is available for this article.

The Breast 60 (2021) 230–237

Contents lists available at ScienceDirect

The Breast

journal homepage: www.elsevier.com/brst

ELSEVIER

Contralateral parenchymal enhancement on MRI is associated with tumor proteasome pathway gene expression and overall survival of early ER+/HER2-breast cancer patients

Max A.A. Ragusi [a, c, *], Tycho Bismeijer [b], Bas H.M. van der Velden [a], Claudette E. Loo [c], Sander Canisius [b, d], Jelle Wesseling [d, e], Lodewyk F.A. Wessels [b, f], Sjoerd G. Elias [g], Kenneth G.A. Gilhuijs [a]

[a] Department of Radiology / Image Sciences Institute, University Medical Center Utrecht, Utrecht University, Heidelberglaan 100, 3584 CX Utrecht, the Netherlands
[b] Division of Molecular Carcinogenesis – Oncode Institute, The Netherlands Cancer Institute – Antoni van Leeuwenhoek Hospital, Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands
[c] Department of Radiology, The Netherlands Cancer Institute – Antoni van Leeuwenhoek Hospital, Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands
[d] Division of Molecular Pathology, The Netherlands Cancer Institute – Antoni van Leeuwenhoek Hospital, Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands
[e] Department of Pathology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, the Netherlands
[f] Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Mekelweg 5, 2628 CD Delft, the Netherlands
[g] Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Universiteitsweg 100, 3584 CG Utrecht, the Netherlands

ARTICLE INFO

Article history:
Received 23 June 2021
Received in revised form
26 September 2021
Accepted 2 November 2021
Available online 2 November 2021

Keywords:
Breast neoplasm
Magnetic resonance imaging
Gene expression
Proteasome endopeptidase complex
Parenchymal tissue

ABSTRACT

Purpose: To assess whether contralateral parenchymal enhancement (CPE) on MRI is associated with gene expression pathways in ER+/HER2-breast cancer, and if so, whether such pathways are related to survival.

Methods: Preoperative breast MRIs were analyzed of early ER+/HER2-breast cancer patients eligible for breast-conserving surgery included in a prospective observational cohort study (MARGINS). The contralateral parenchyma was segmented and CPE was calculated as the average of the top-10% delayed enhancement. Total tumor RNA sequencing was performed and gene set enrichment analysis was used to reveal gene expression pathways associated with CPE (N = 226) and related to overall survival (OS) and invasive disease-free survival (IDFS) in multivariable survival analysis. The latter was also done for the METABRIC cohort (N = 1355).

Results: CPE was most strongly correlated with proteasome pathways (normalized enrichment statistic = 2.04, false discovery rate = .11). Patients with high CPE showed lower tumor proteasome gene expression. Protea

OS in the

ex n

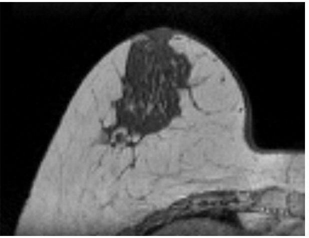
fo

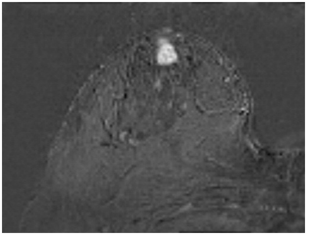
Con PE w

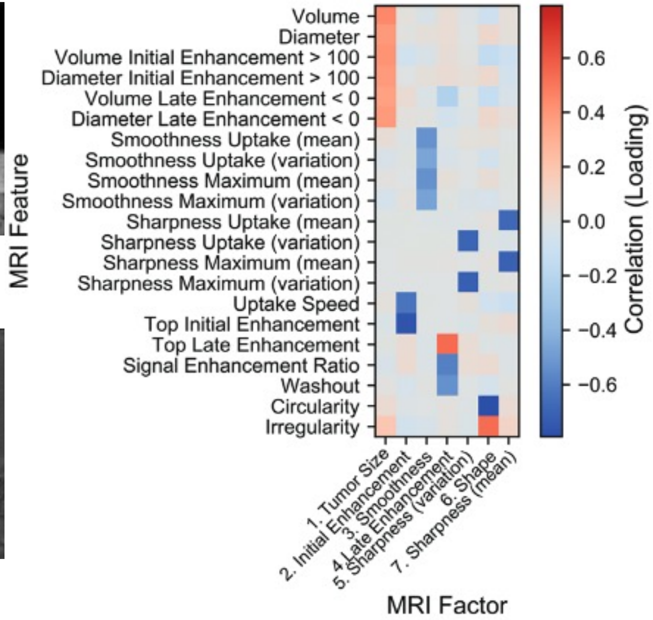
breast cancer patie

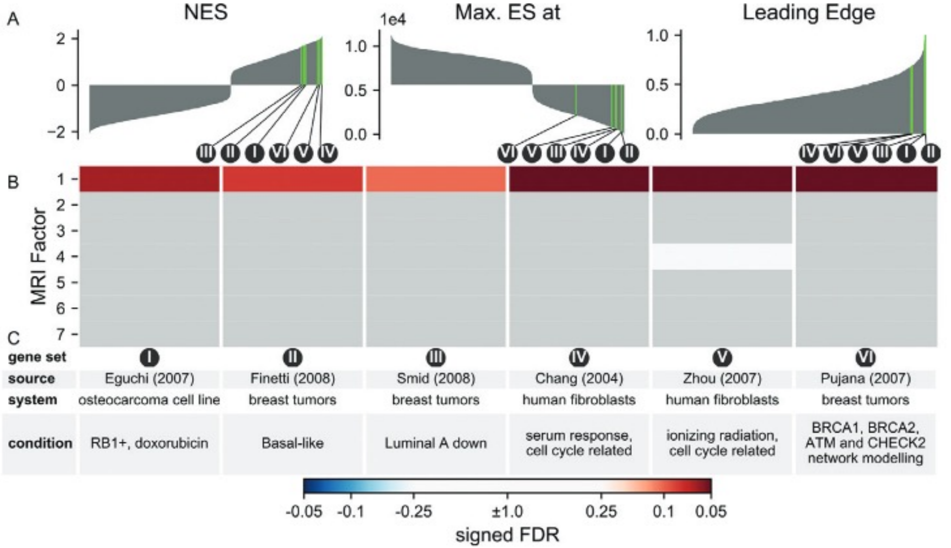
@basvandervelden

Explain imaging using genotype







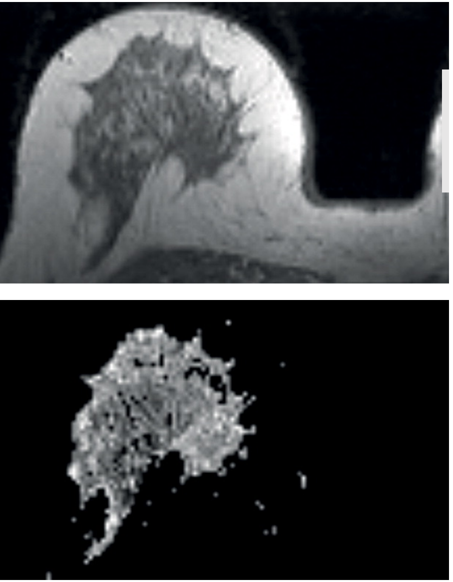


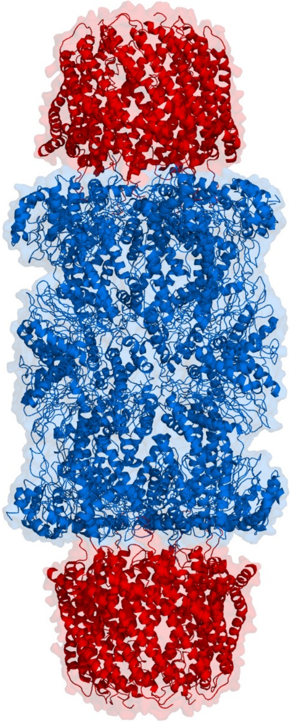
gene set	I	II	III	IV	V	VI
source	Eguchi (2007)	Finetti (2008)	Smid (2008)	Chang (2004)	Zhou (2007)	Pujana (2007)
system	osteocarcoma cell line	breast tumors	breast tumors	human fibroblasts	human fibroblasts	breast tumors
condition	RB1+, doxorubicin	Basal-like	Luminal A down	serum response, cell cycle related	ionizing radiation, cell cycle related	BRCA1, BRCA2, ATM and CHECK2 network modelling

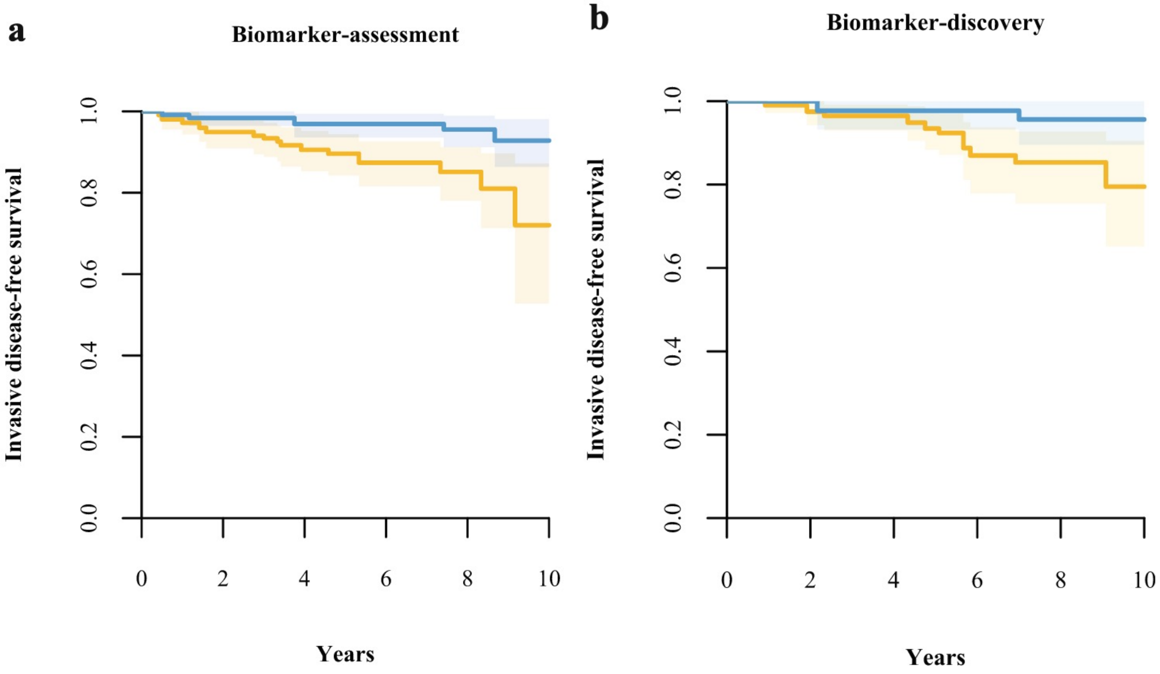
Gilhuijs et al, Med Phys 1998
Bismeijer et al. Radiology 2020

@basvandervelden

Explain imaging using genotype







Van der Velden et al. Radiology 2015, Clin Can Res 2017,
Eur Rad 2018; Ragusi et al. The Breast 2021

@basvandervelden

Take home messages

- XAI adds confidence to decisions
- XAI can improve performance
- Holistic XAI: more than images

@basvandervelden