Cats, crowds & other considerations in medical imaging

IAEA Fusion & Plasma Science workshop

1 December 2023

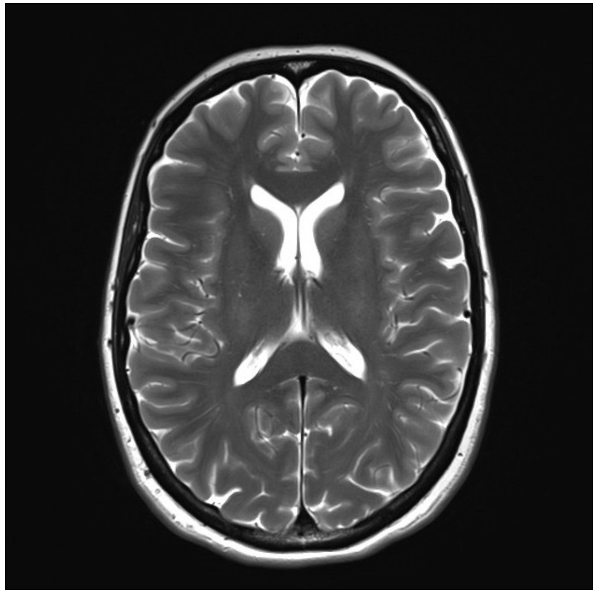
Dr. Veronika Cheplygina
IT University of Copenhagen

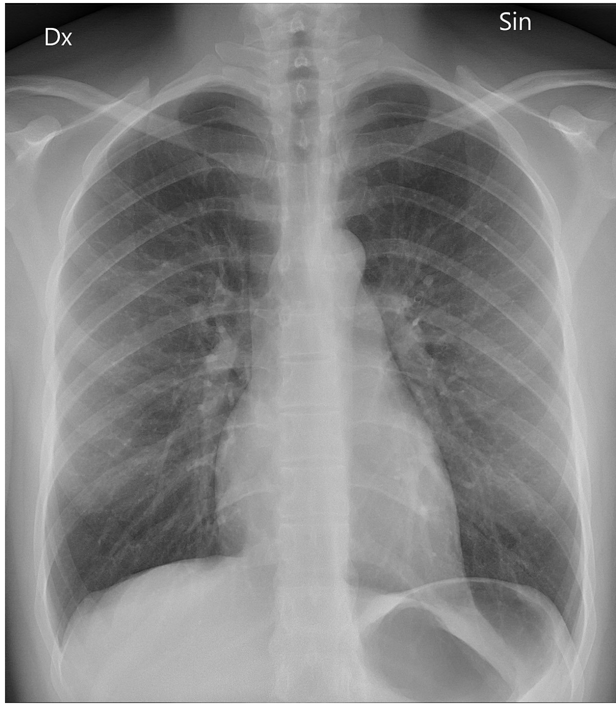
vech@itu.dk
https://www.veronikach.com

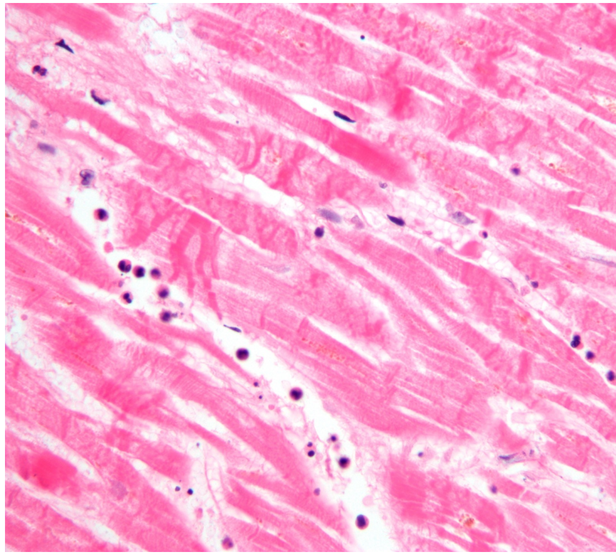
Medical imaging

- Different organs/modalities
- X-ray, CT, MR, ultrasound, histopathology, …
- 2D, 3D, 3D with time, …
- Detection of abnormalities/diagnosis

(Example images from Wikipedia)

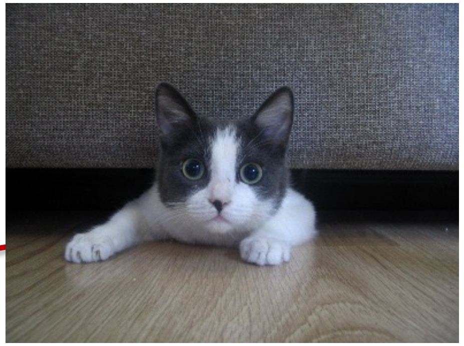


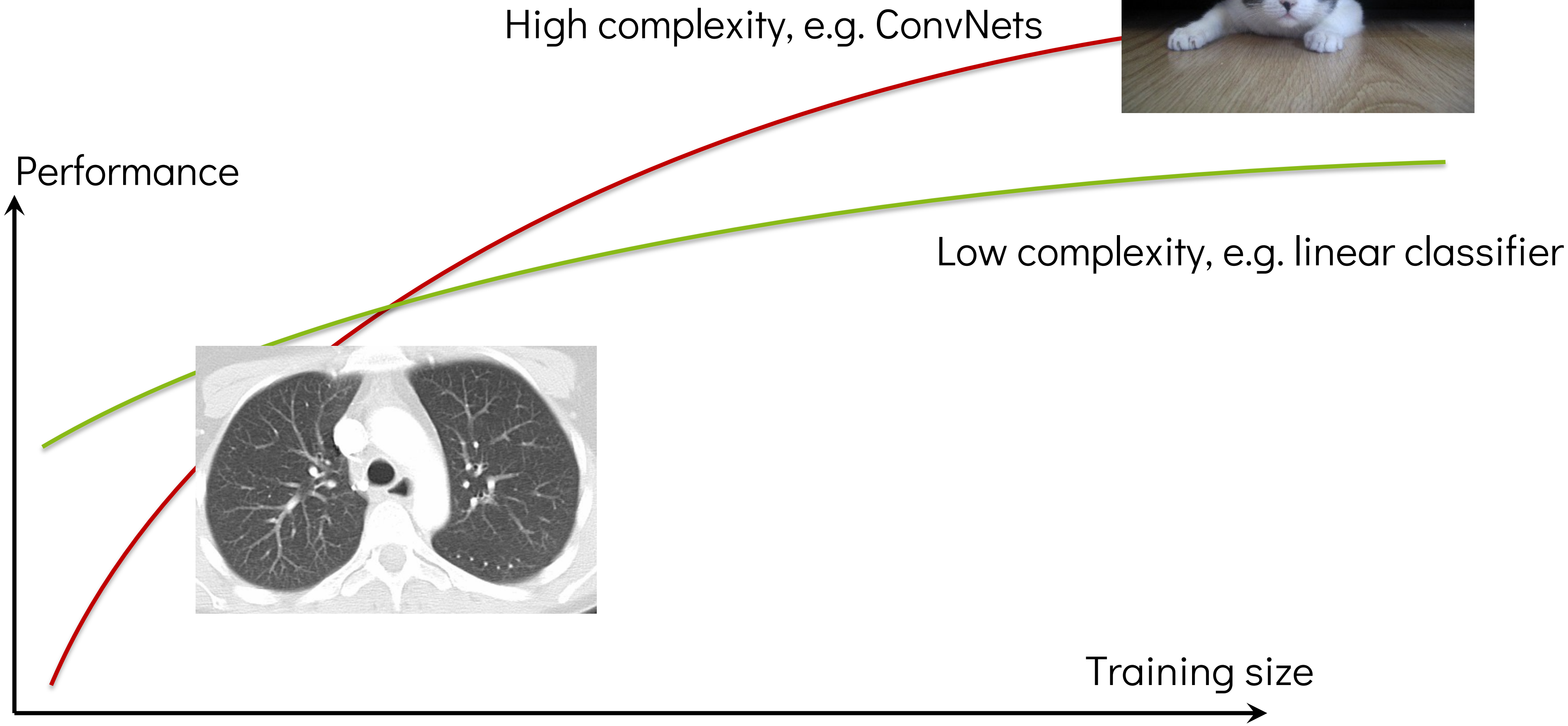




Ideal data size

- Computer vision: millions of images





Medical data size

Dataset size in Alzheimer's disease [Varoquaux, Cheplygina 2022]

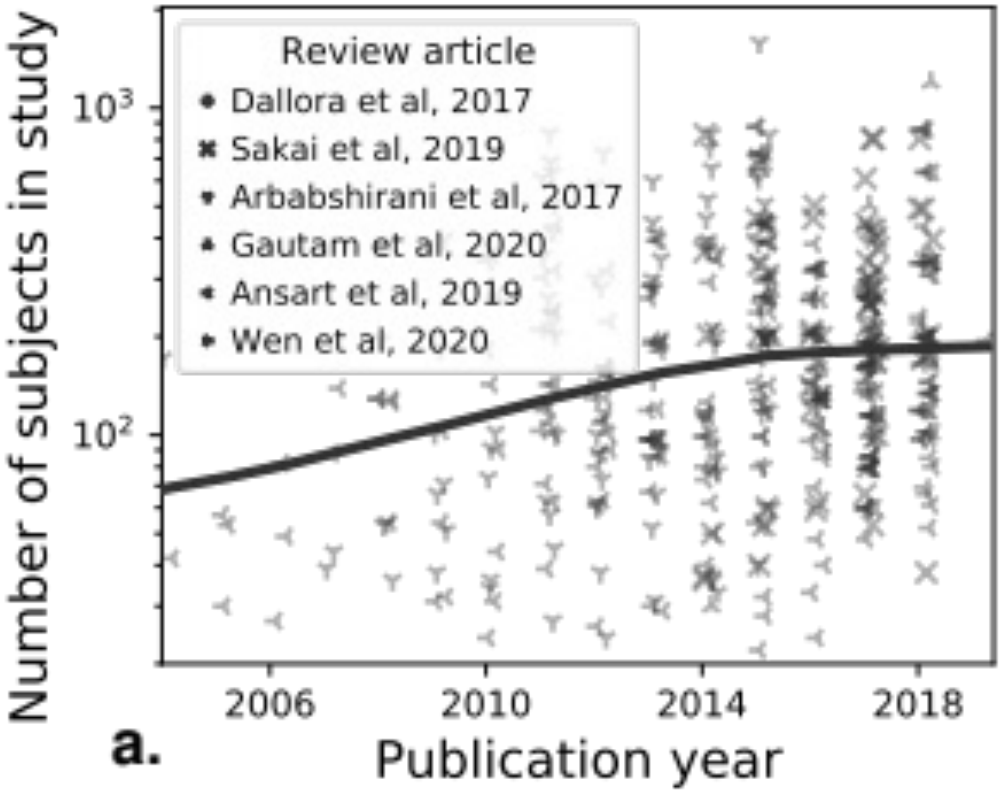


Table 2: Large Open-Source Medical Imaging Data Sets

Data Set Description	Image Types	No. of Patients
American College of Radiology Imaging Network National CT Colonography Trial (ACRIN 6664) (102)	CT	825
Alzheimer's Disease Neuroimaging Initiative (103)	MRI, PET	>1700
Curated Breast Imaging Subset of the Digital Database for Screening Mammography (36)	Mammography	6671
ChestX-ray8, National Institutes of Health chest x-ray database (41)	Radiography	30 805
CheXpert, chest radiographs (79)	Radiography	65 240
Collaborative Informatics and Neuroimaging Suite (104)	MRI	
DeepLesion, body CT (60)	CT	4427
Head and neck PET/CT (105)	PET/CT, CT	298
Lung Image Database Consortium image collection (106)	CT, radiography	1010
MRNet, knee MRI (80)	MRI	1370
Musculoskeletal bone radiographs, or MURA (107)	Radiography	14 863
National Lung Screening Trial (108)	CT, pathology	26 254
PROSTATEx Challenge, SPIE-AAPM-NCI Prostate MR Classification Challenge (109)	MRI	346
Radiological Society of North America Intracranial Hemorrhage Detection (110)	CT	25 000
Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma data collection (111)	CT, MRI	267
Virtual Imaging Clinical Trial for Regulatory Evaluation (112)	Mammography, digital breast tomosynthesis	2994

[Willemink et al 2020]

Outline

- Learning from limited labeled data
  - With cats (transfer learning)
  - With crowdsourcing

- “Other considerations”
  - Evaluation

Idea 1: Transfer learning

Learn from related domains and/or tasks

Domain = input data, e.g. images of different modalities

Task = input → output, e.g. prediction of different diseases

Domain ↓ Task →	**Same**	**Different**
Same	Supervised learning	Multi-task learning
Different	Domain adaptation	Pretraining + Fine-tuning

Learning from any dataset?

Learn a generalized representation (pretraining), then extract features or fine-tune

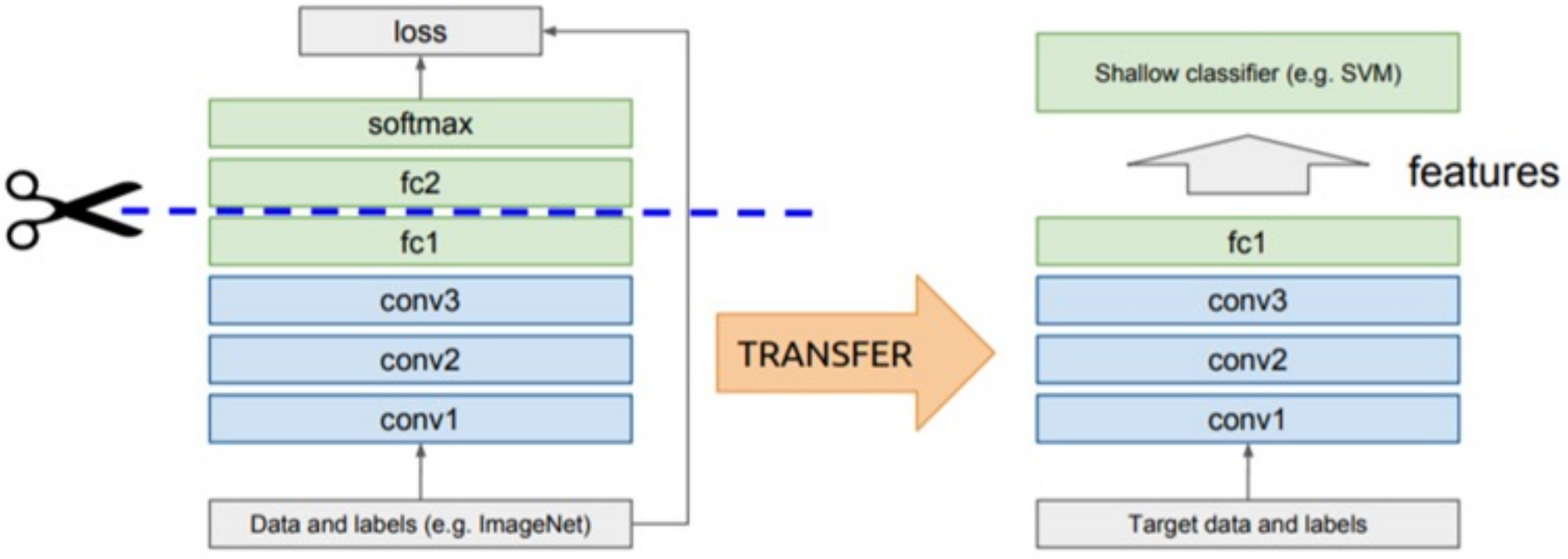
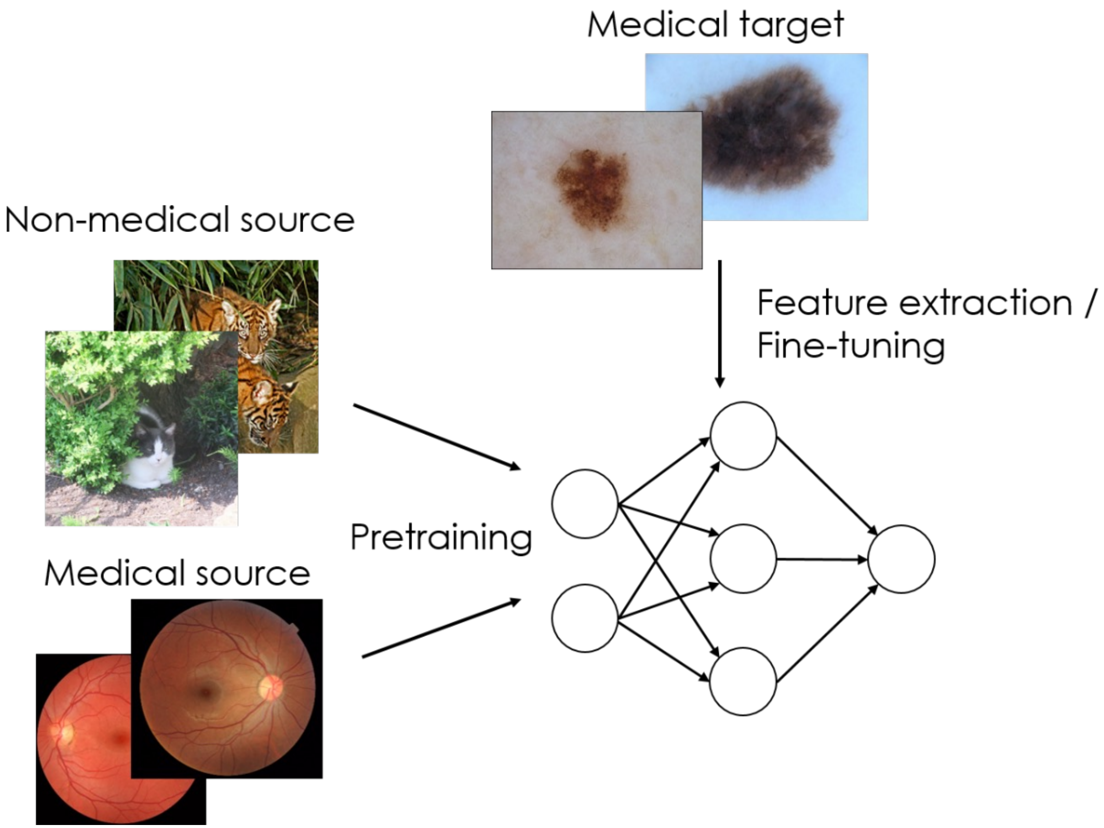


Image: towardsdatascience.com

Medical or non-medical source data?

- 2014-2015 first papers with non-medical sources (often ImageNet)
- May be suboptimal for medical data
- Few comparisons in literature, conflicting results
- Our early comparisons: ImageNet best but is much larger

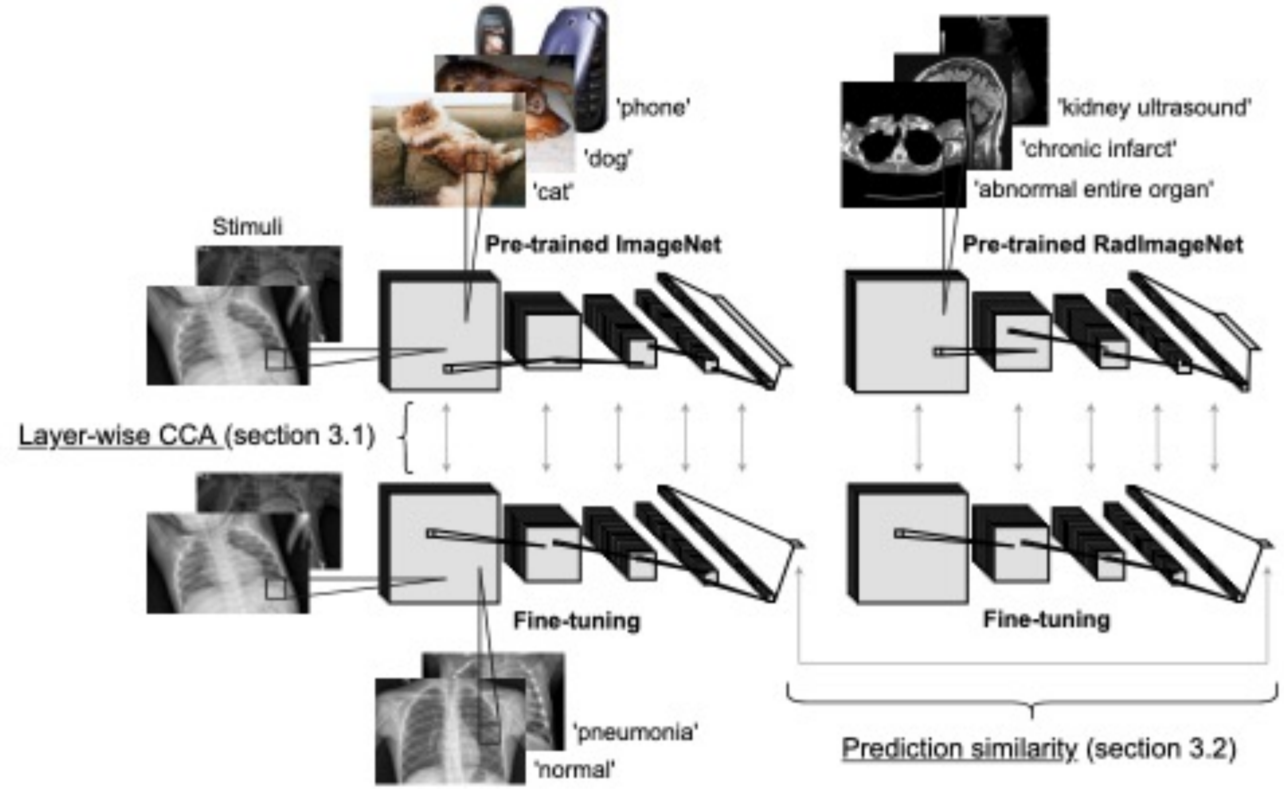


Cheplygina, V. (2019). Cats or CAT scans: transfer learning from natural or medical image source datasets?. *Current Opinion in Biomedical Engineering*. URL

ImageNet vs RadImageNet

Project CATS - Choosing A Transfer Source for medical image classification

ImageNet vs RadImageNet [Mei et al 2022] - similar size/properties



Dovile Juodelyte

novo nordisk fonden

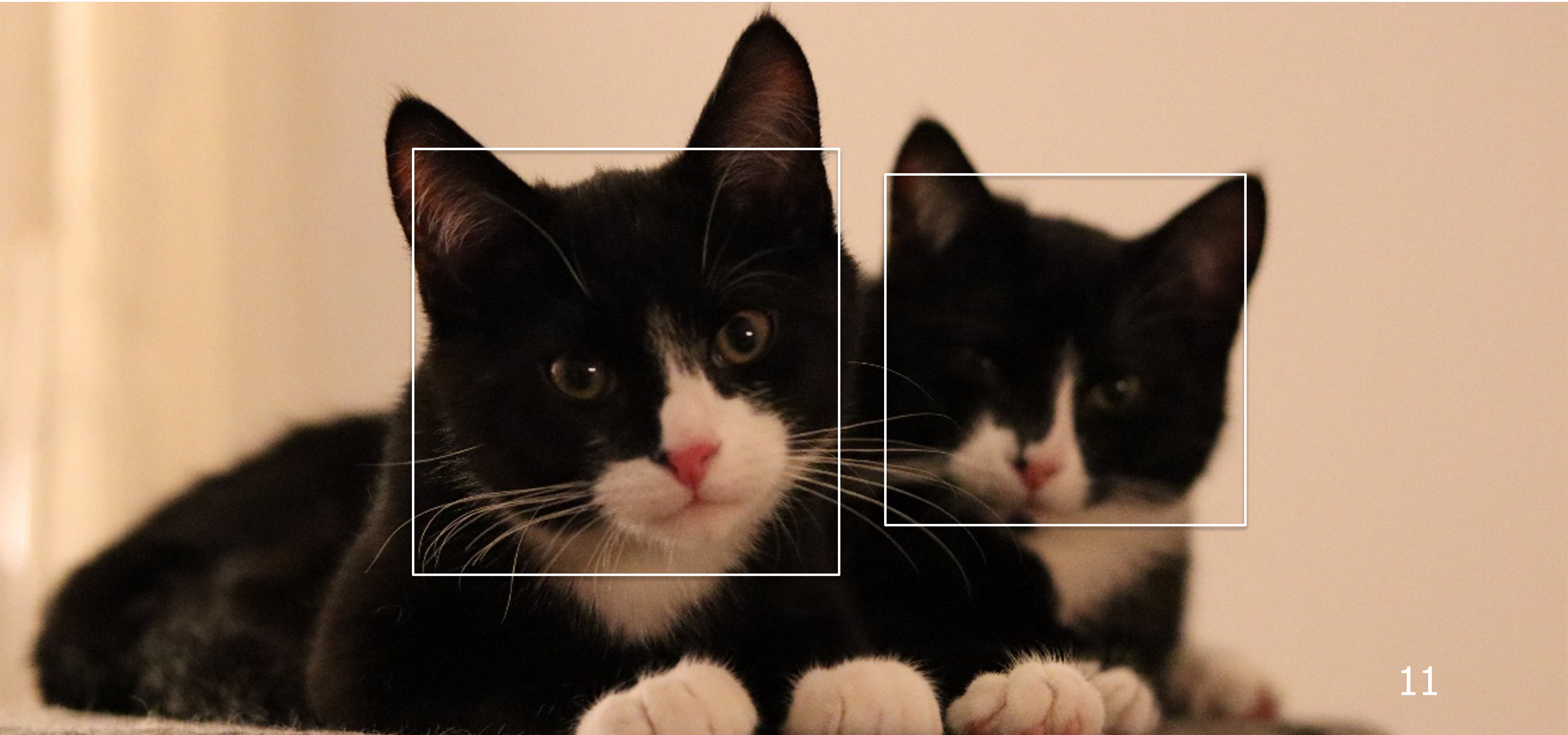
ImageNet vs RadImageNet

- ImageNet tends to outperform RadImageNet Juodelyte et al 2023
- But ImageNet may be more sensitive to label noise, artifacts

Table 2: Mean AUC ± std (both ×100) after fine-tuning on target datasets. Underlined is the highest mean AUC per dataset.

Target dataset	No freeze	Freeze	No freeze	Freeze	Random init
	ImageNet		RadImageNet		
Thyroid	52.3 ± 8.9	<u>58.5 ± 5.1</u>	50.2 ± 4.1	49.4 ± 5.9	52.9 ± 5.1
Breast	<u>90.5 ± 3.8</u>	89.8 ± 3.0	88.2 ± 4.2	72.4 ± 22.1	51.2 ± 10.0
Chest	98.7 ± 0.4	98.6 ± 0.6	<u>98.8 ± 0.2</u>	98.7 ± 0.3	82.5 ± 1.0
Mammograms	63.4 ± 4.3	<u>68.8 ± 2.0</u>	62.0 ± 12.2	66.2 ± 6.3	49.6 ± 3.7
Knee	<u>91.5 ± 1.1</u>	89.0 ± 3.1	89.3 ± 6.3	63.8 ± 5.7	68.3 ± 11.0
ISIC	<u>94.2 ± 1.3</u>	93.1 ± 2.3	90.8 ± 0.8	90.6 ± 0.7	84.0 ± 2.3
Pcam-small	<u>93.8 ± 1.1</u>	93.2 ± 2.5	87.1 ± 2.3	86.0 ± 1.9	73.4 ± 8.2

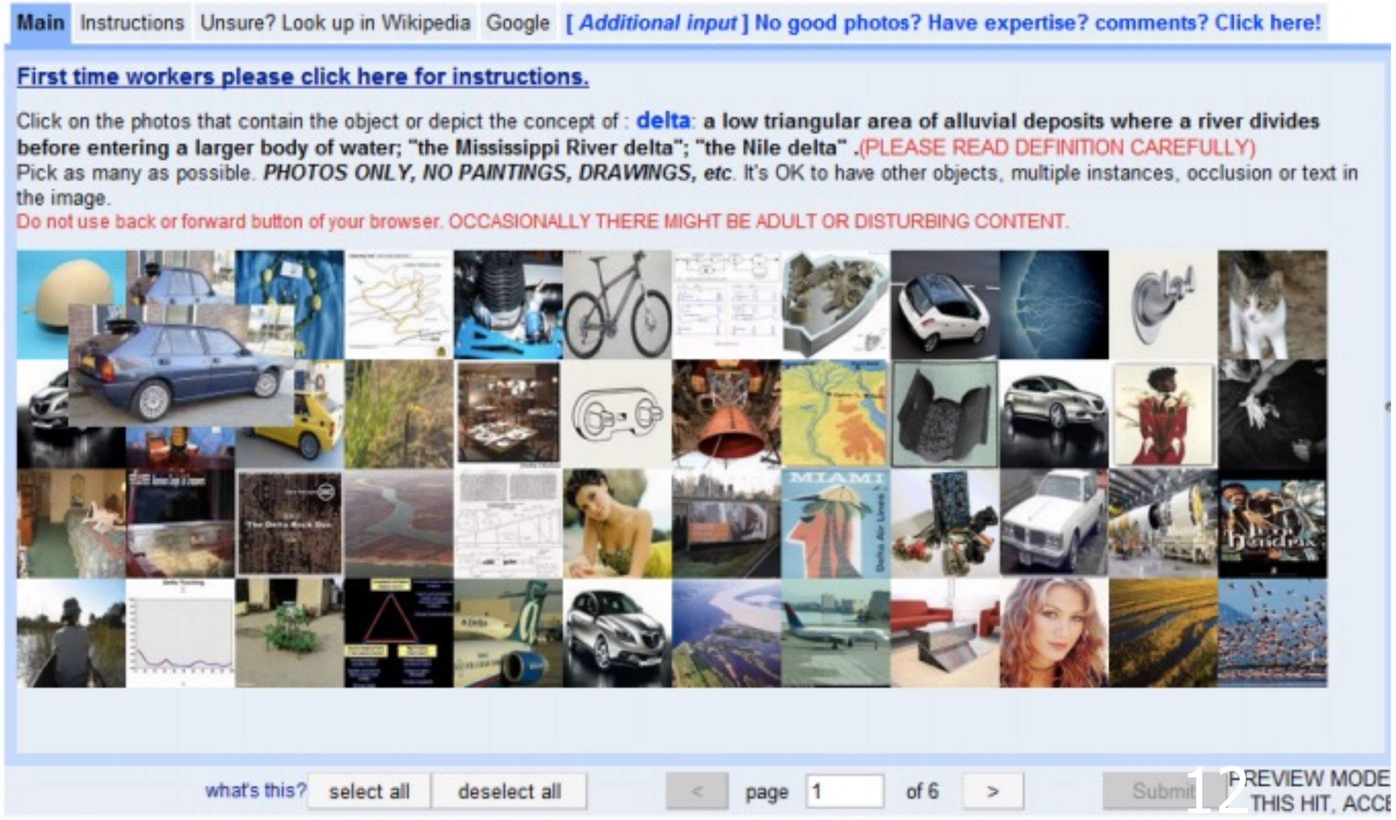
Idea 2: Label more data

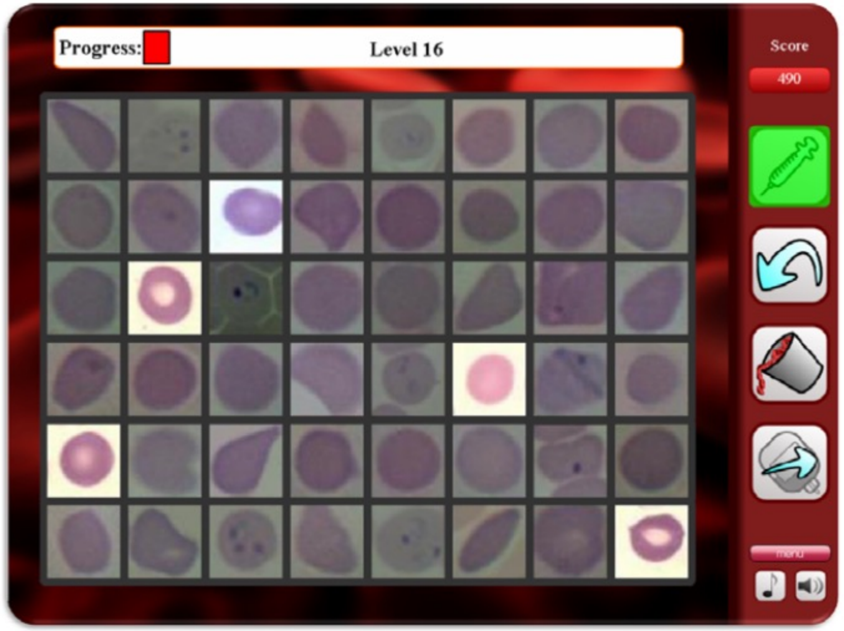


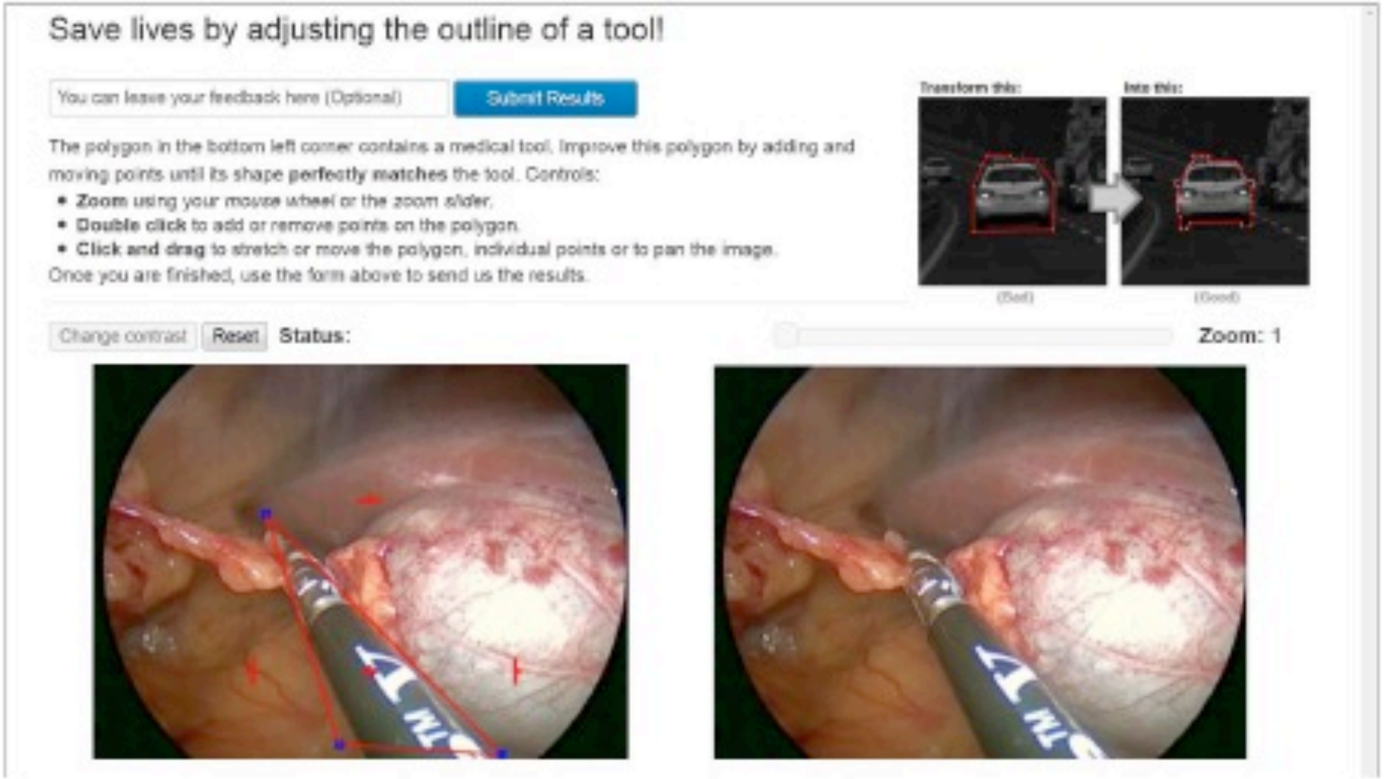
Crowdsourcing

ImageNet [Deng et al 2009] crowdsourced on Amazon Mechanical Turk - workers asked to label cats, bicycles etc.

Counterintuitive for medical?

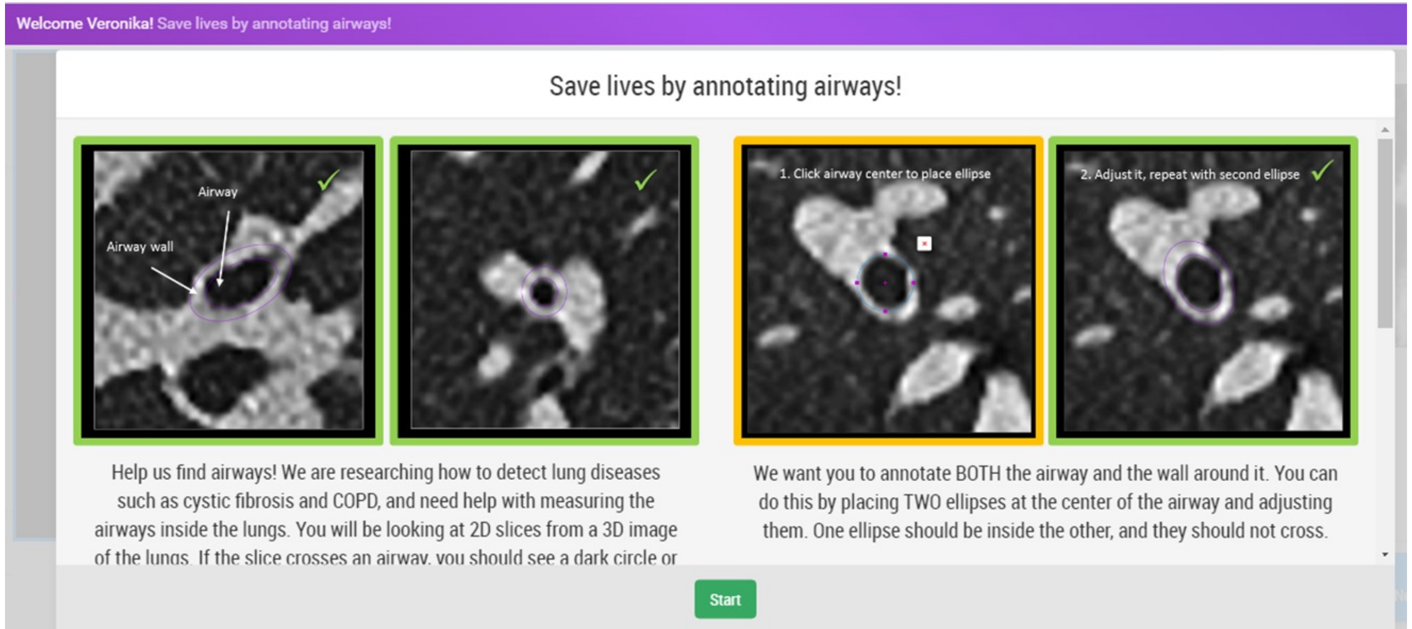






Mavandadi et al 2012
Maier-Hein et al 2014
Cheplygina et al 2016

Survey:
[Ørting et al 2020]



Crowdsourcing annotations

Simplify task, e.g. in skin lesion classification instead of benign/malignant, use more intuitive features (also used by experts):

- A - Asymmetry
- B - Border
- C - Color

Image source

Multi-task with crowd annotations

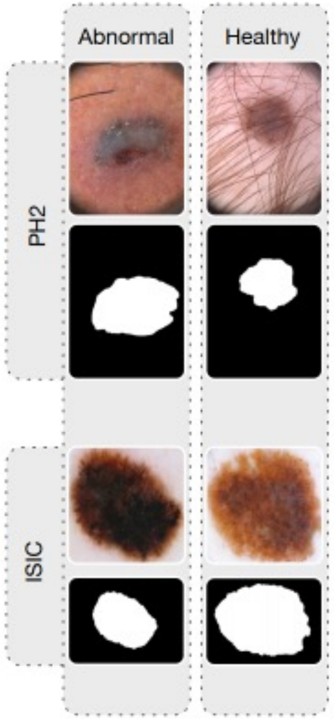
- Noisy annotations of visual features (e.g. asymmetry) by students & crowdsourcing

- Multi-task learning (diagnosis & annotations) outperforms baseline (diagnosis only)

[Raumanns et al 2021]

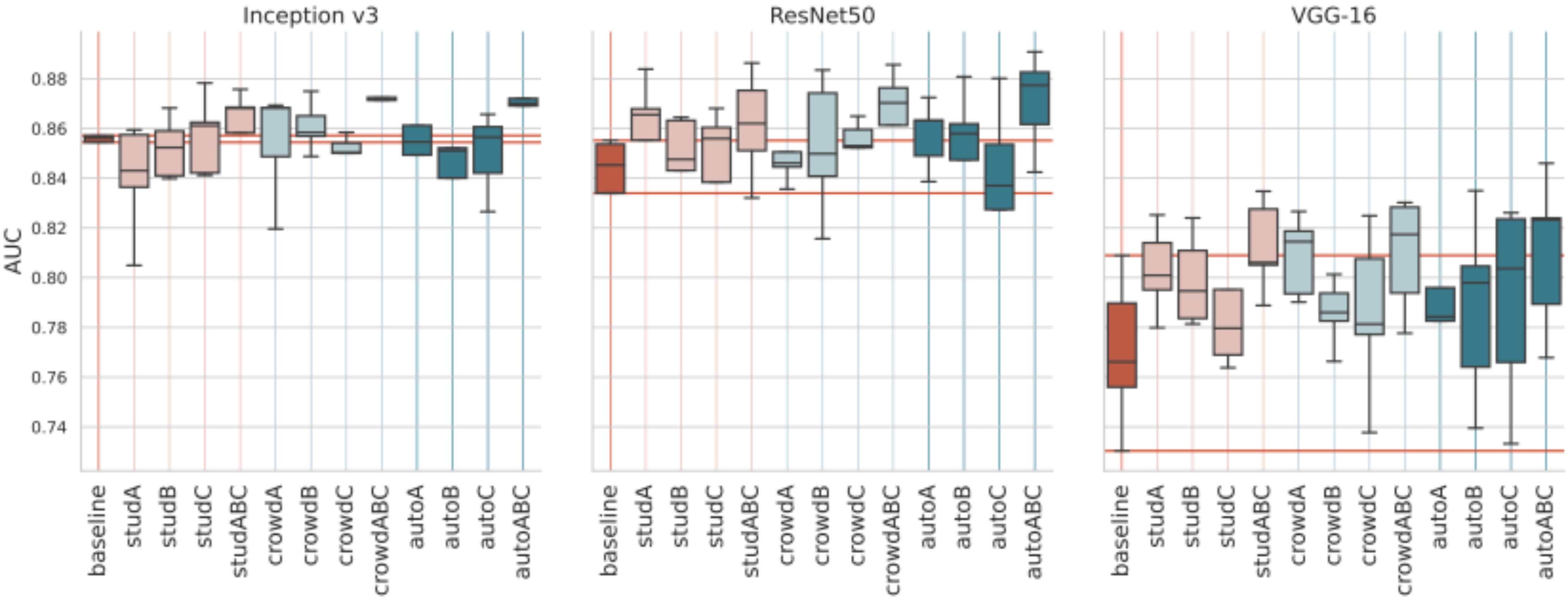
Netherlands Organisation for Scientific Research

Ralf Raumanns



	Asymmetry	Border	Color
Abnormal			
Automated	2	10	5
Crowd	1,2,0	4,8,0	4,4,3
Expert	2	-	2
Healthy			
Automated	2	7	6
Crowd	0,0,0	8,5,7	2,1,1
Expert	0	-	1
Abnormal			
Automated	2	50	5
Crowd	1,1,2	8,2,7	1,1,1
Student	1,1,2	5,3,6	4,4,4
Healthy			
Automated	2	62	6
Crowd	0,0,1	7,6,6	3,3,1
Student	2,2,2	5,3,8	3,3,3

Ensembles with crowd annotations best



Other considerations

What to choose?

Data augmentation
Self-supevised learning
Semi-supervised learning
Active learning
Weakly supervised learning
Transfer learning
Crowdsourcing
Synthetic data
…

https://unsplash.com/photos/Wpg3Qm0zaGk

Practical clinical use

"none of the models identified are of potential clinical use" [Roberts et al 2021]

"[...] narrow use cases [...] limited external validation [...]" [Kelly et al 2022]

"Studies were identified for 26 of the 53 neuroalgorithms [...] exploring the use of algorithms in clinical practice were available for 7 algorithms."

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala & Carola-Bibiane Schönlieb

Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE)

Brendan S. Kelly [1,2,3,4,5,6] · Conor Judge [5,6] · Stephanie M. Bollard [4,5,6] · Simon M. Clifford [1,6] · Gerard M. Healy [1,6] · Awsam Aziz [4,6] · Prateek Mathur [2,6] · Shah Islam [6,7] · Kristen W. Yeom [7,8] · Aonghus Lawlor [2,6] · Ronan P. Killeen [2,4,6]

FDA-approved machine learning algorithms in neuroradiology: A systematic review of the current evidence for approval

Alexander G. Yearley [a b], Caroline M.W. Goedmakers [b c], Armon Panahi [d], Joanne Doucette [b e], Aakanksha Rana [b f], Kavitha Ranganathan [g], Timothy R. Smith [a b]

Why?

- Results may appear good, but not generalize, even with larger datasets

Overfitting to spurious patterns / shortcuts

- Pen marks correlated with melanoma
- Network flips diagnosis

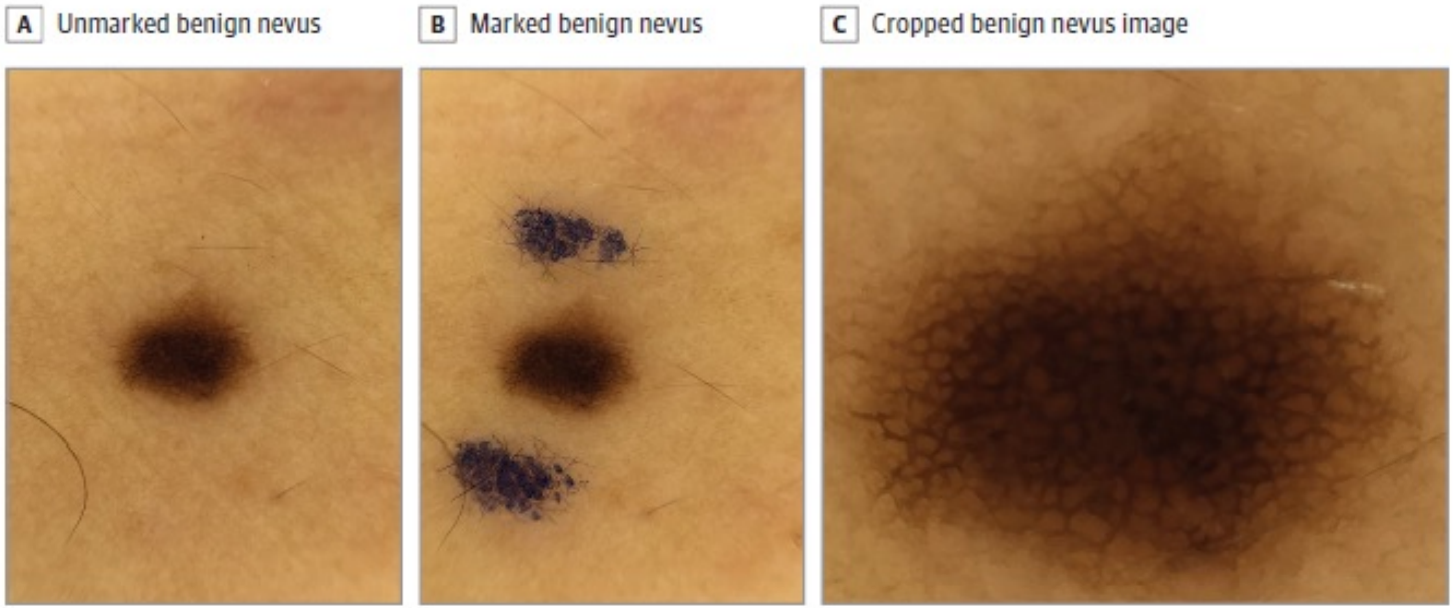


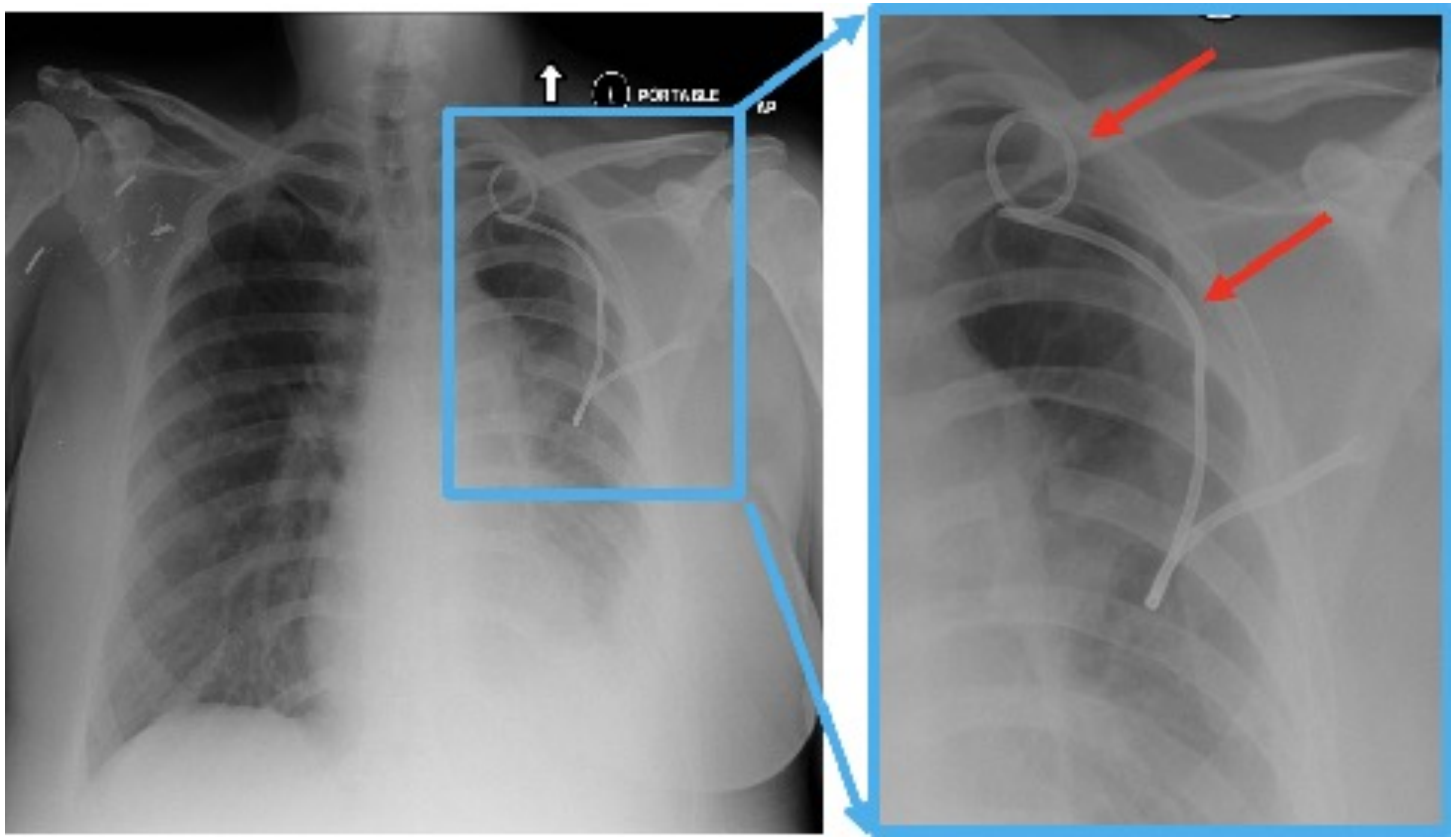
Figure 1. Convolutional Neural Network (CNN) Classification and Melanoma Probability Scores for Dermoscopic Images of Unmarked, Marked, and Cropped Benign Nevus and Melanoma

[Winkler et al]

21

Overfitting

- Chest drain associated with a collapsed lung
- AUC 0.94 vs 0.77



[Oakden-Rayner et al 2019]

[Image from Graf et al 2020]

22

Shortcuts outside the object of interest...

[Bissoto et al 2019]

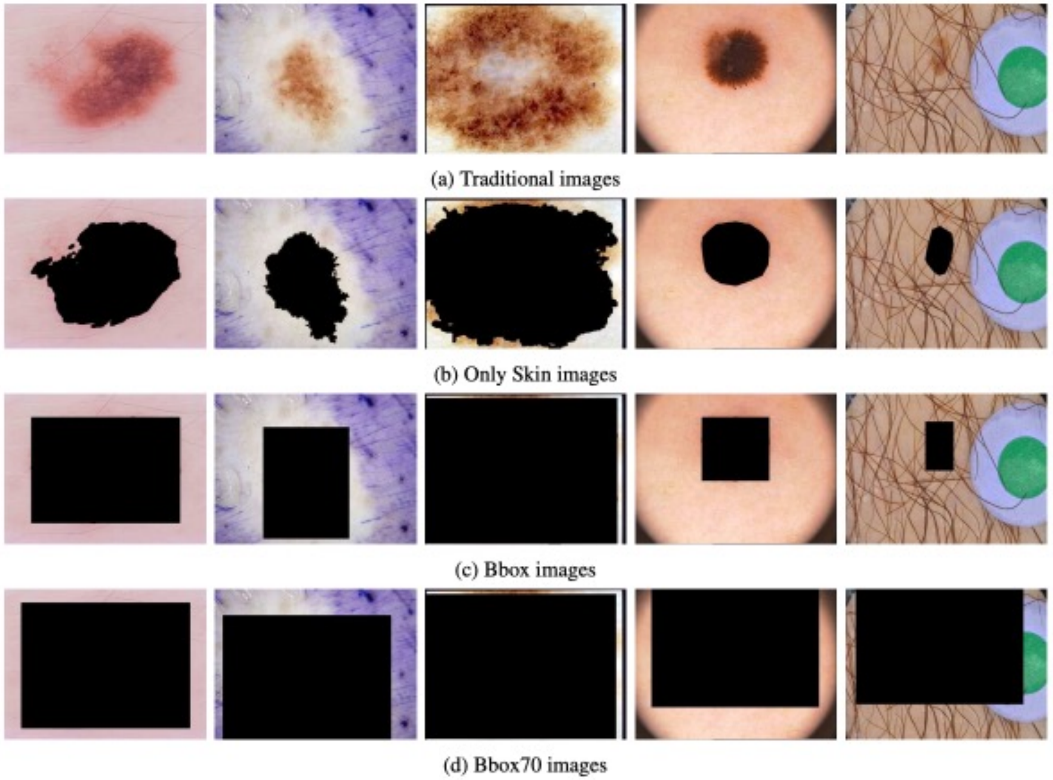
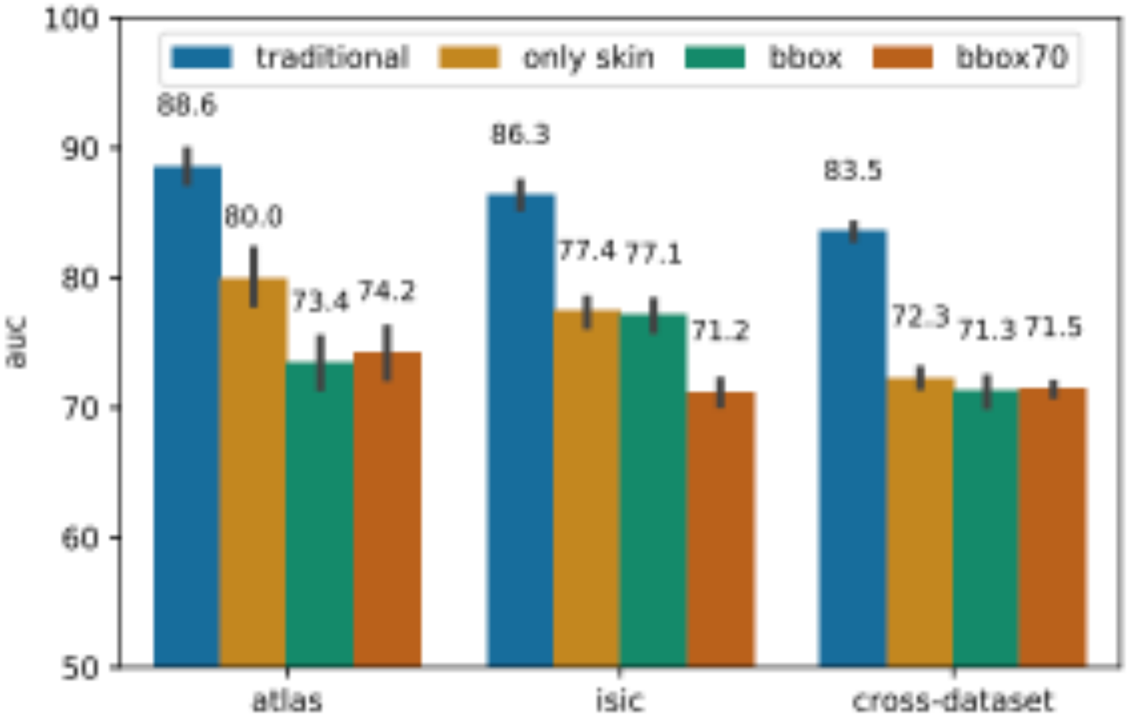


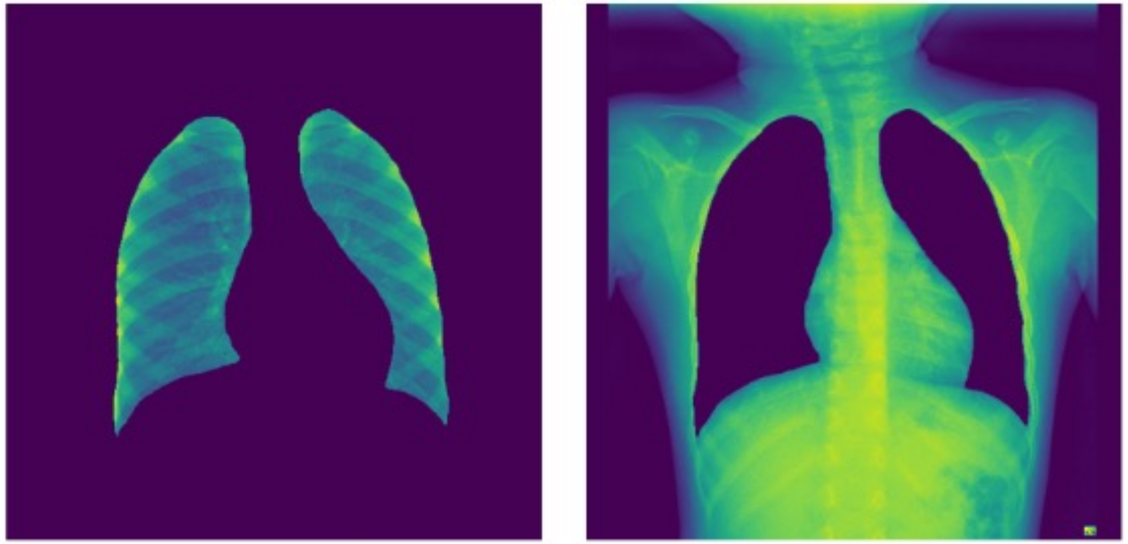
Figure 1: Samples from each of our disrupted datasets. We gradually remove cogent information, until there is no informatio left to apply any aspect of medical score algorithms [4, 12]. Next, we use those sets to evaluate if the network can still lear patterns with the information left to correctly classify skin lesions. Best seen in digital format.

(De)Constructing Bias on Skin Lesion Datasets

Alceu Bissoto[1] Michel Fornaciali[2] Eduardo Valle[2] Sandra Avila[1]
stitute of Computing (IC) [2]School of Electrical and Computing Engineering (FEEC)
RECOD Lab., University of Campinas (UNICAMP), Brazil



Shortcuts



(a) Only lungs (b) Without lungs

Amelia Jiménez-Sánchez

Fine-tuning data	Evaluation data	Effusion	Pneumothorax	Atelectasis	Cardiomegaly	Pneumonia
		Area Under the ROC curve (AUC)				
PadChest	PadChest	94.2 $\pm$ 0.1	81.4 $\pm$ 2.3	86.9 $\pm$ 0.1	89.2 $\pm$ 0.2	**81.0** $\pm$ 0.3
PadChest, no lungs	PadChest, no lungs	**94.4** $\pm$ 0.1	82.2 $\pm$ 1.4	**87.0** $\pm$ 0.6	**90.5** $\pm$ 0.1	79.0 $\pm$ 0.1
PadChest, only lungs	PadChest, only lungs	93.1 $\pm$ 0.0	80.5 $\pm$ 1.2	86.4 $\pm$ 0.2	90.1 $\pm$ 0.1	79.3 $\pm$ 0.2

Webinar: Datasets through the L👀king-Glass

More about datasets & shortcuts

Next webinar: Shortcuts and bias

Date: **4 December 2023 at 4pm CET**

Where: **Zoom** [Registration]

Add to: Google Calendar / Outlook Calendar / Yahoo Calendar

Dr. Jessica Schrouff (Google DeepMind, UK)
Dr. Enzo Ferrante (CONICET, Argentina)
Rhys Compton and **Lily Zhang** (New York University, USA)

https://purrlab.github.io/webinar

Previous talks:

- S01E01 - **Dr. Roxana Daneshjou** (Stanford University School of Medicine, Stanford, CA, USA). 27th Feb 2023. **Challenges with equipoise and fairness in AI/ML datasets in dermatology**. Video.
- S01E02 - **Dr. David Wen** (Oxford University Clinical Academic Graduate School, University of Oxford, Oxford, UK). 27th Feb 2023. **Characteristics of open access skin cancer image datasets: implications for equitable digital health**. Video.
- S01E03 - **Prof. Colin Fleming** (Ninewells Hospital, Dundee, UK). 27th Feb 2023. **Characteristics of skin lesions datasets**. Video.
- S02E01 - **Prof. Amber Simpson** (Queen's University, Canada). 5th June 2023. **The medical segmentation decathlon**. Video.
- S02E02 - **Dr. Esther E. Bron** (Erasmus MC - University Medical Center Rotterdam, the Netherlands). 5th June 2023. **Image analysis and machine learning competitions in dementia**. Video.
- S02E03 - **Dr. Ujjwal Baid** (University of Pennsylvania, USA). 5th June 2023. **Brain tumor segmentation challenge 2023**. Video.
- S03E01 - **Dr. Thijs Kooi** (Lunit, South Korea). 18th September 2023. **Optimizing annotation cost for AI based medical image analysis**. Video.
- S03E02 - **Dr. Andre Pacheco** (Federal University of Espírito Santo, Brazil). 18th September 2023. **PAD-UFES-20: the challenges and opportunities in creating a skin lesion dataset**. Video.

Conclusions

- Lots of methods, we can do many things to improve training, but evaluation is key
- Need more focus on datasets for better generalizability & robustness
- More (systematic) reviews and “real-world” evaluation

Thank you!

@drveronikach@dair-community.social

https://www.veronikach.com