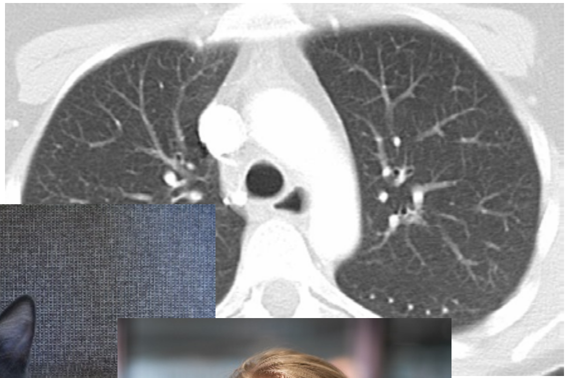
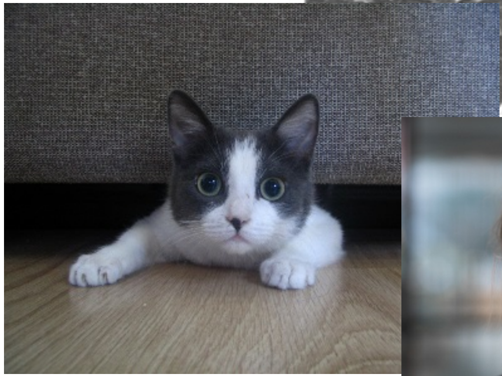
Technische Universiteit Eindhoven
University of Technology

Cats and Crowds: Augmenting Limited Labelled Data in Medical Image Analysis

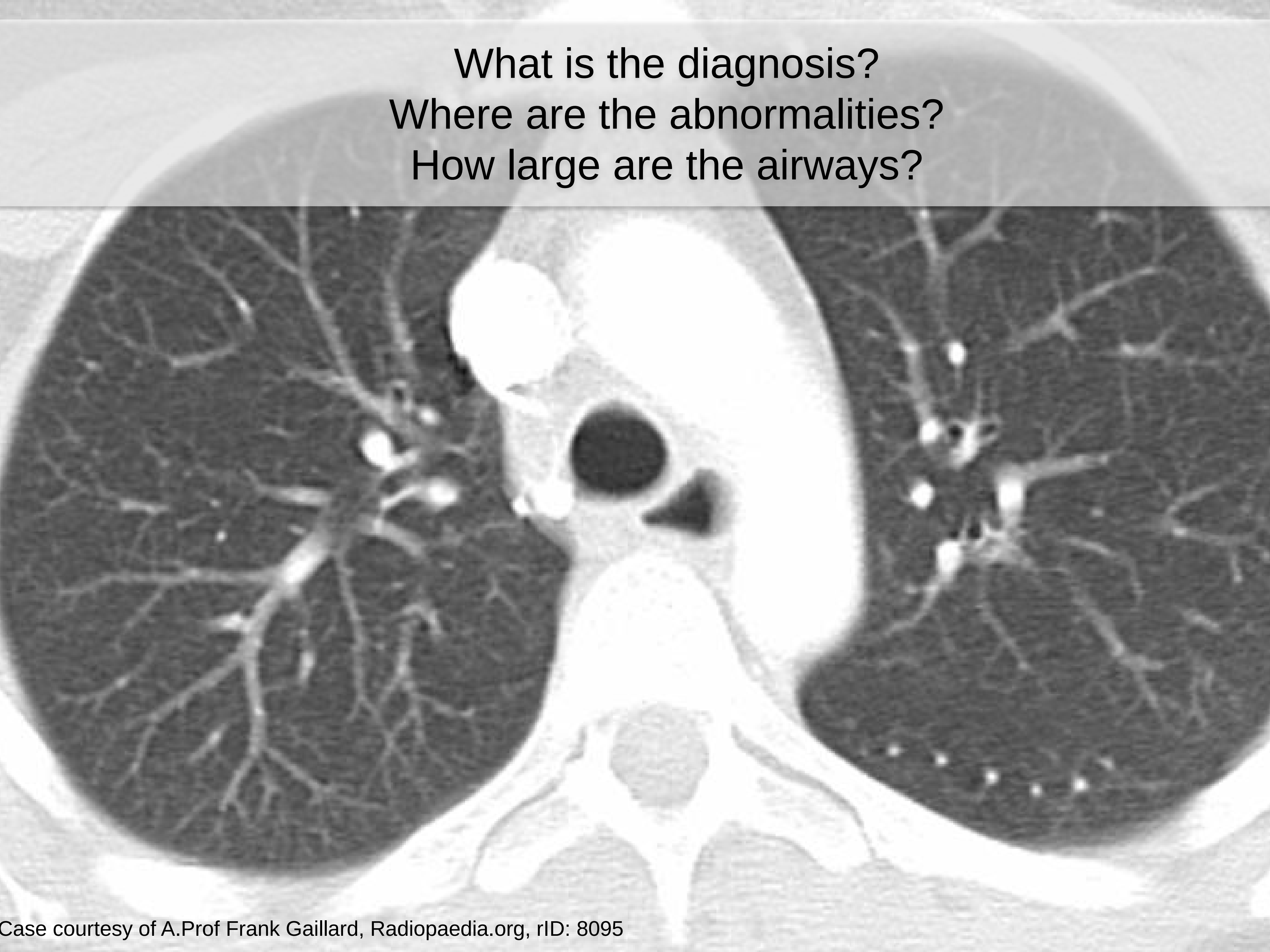
Veronika Cheplygina





@drveronikach

http://www.veronikach.com



What is the diagnosis?
Where are the abnormalities?
How large are the airways?

Case courtesy of A.Prof Frank Gaillard, Radiopaedia.org, rID: 8095


Data?

Representative
& annotated data

Overfitting

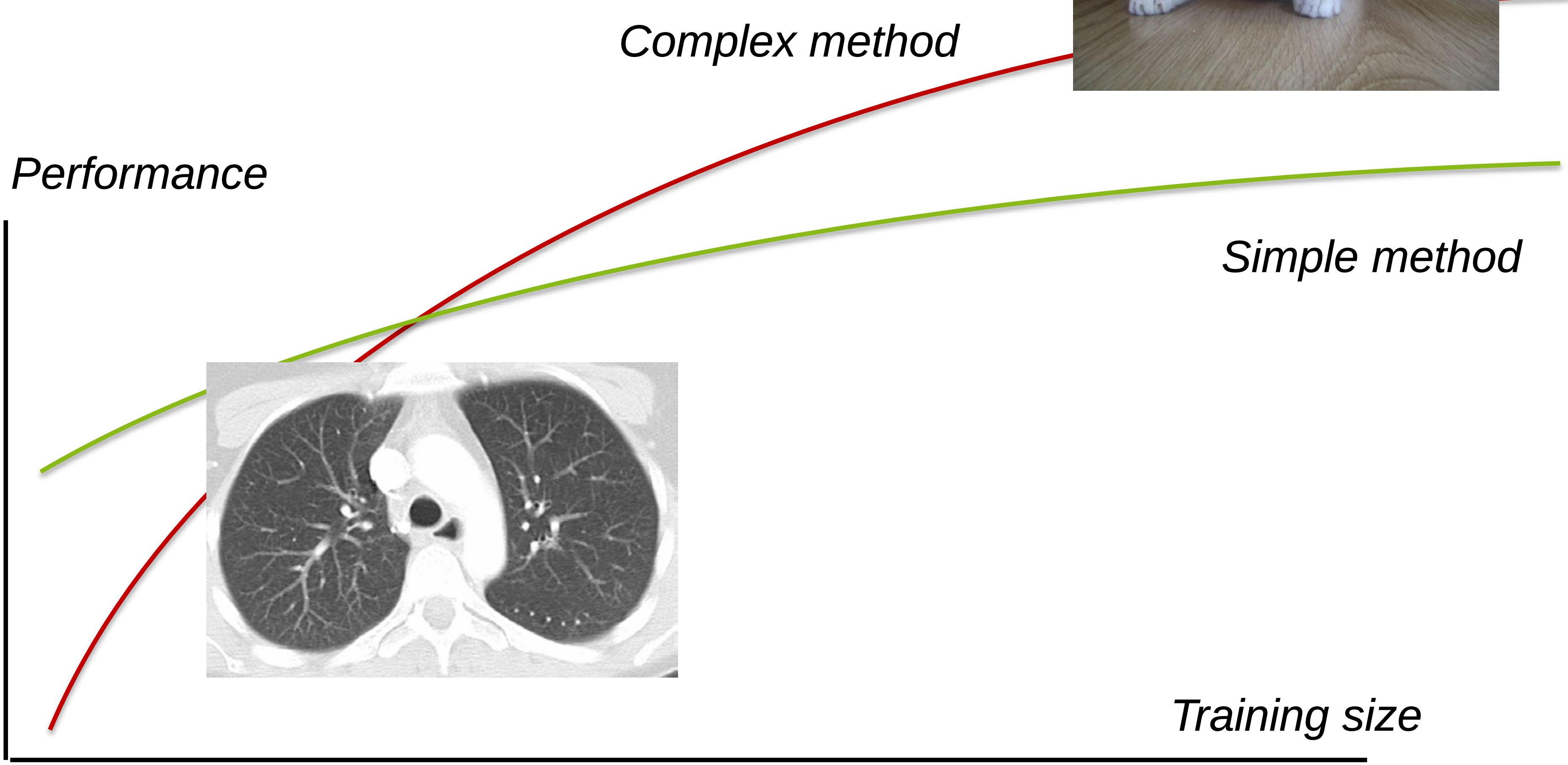
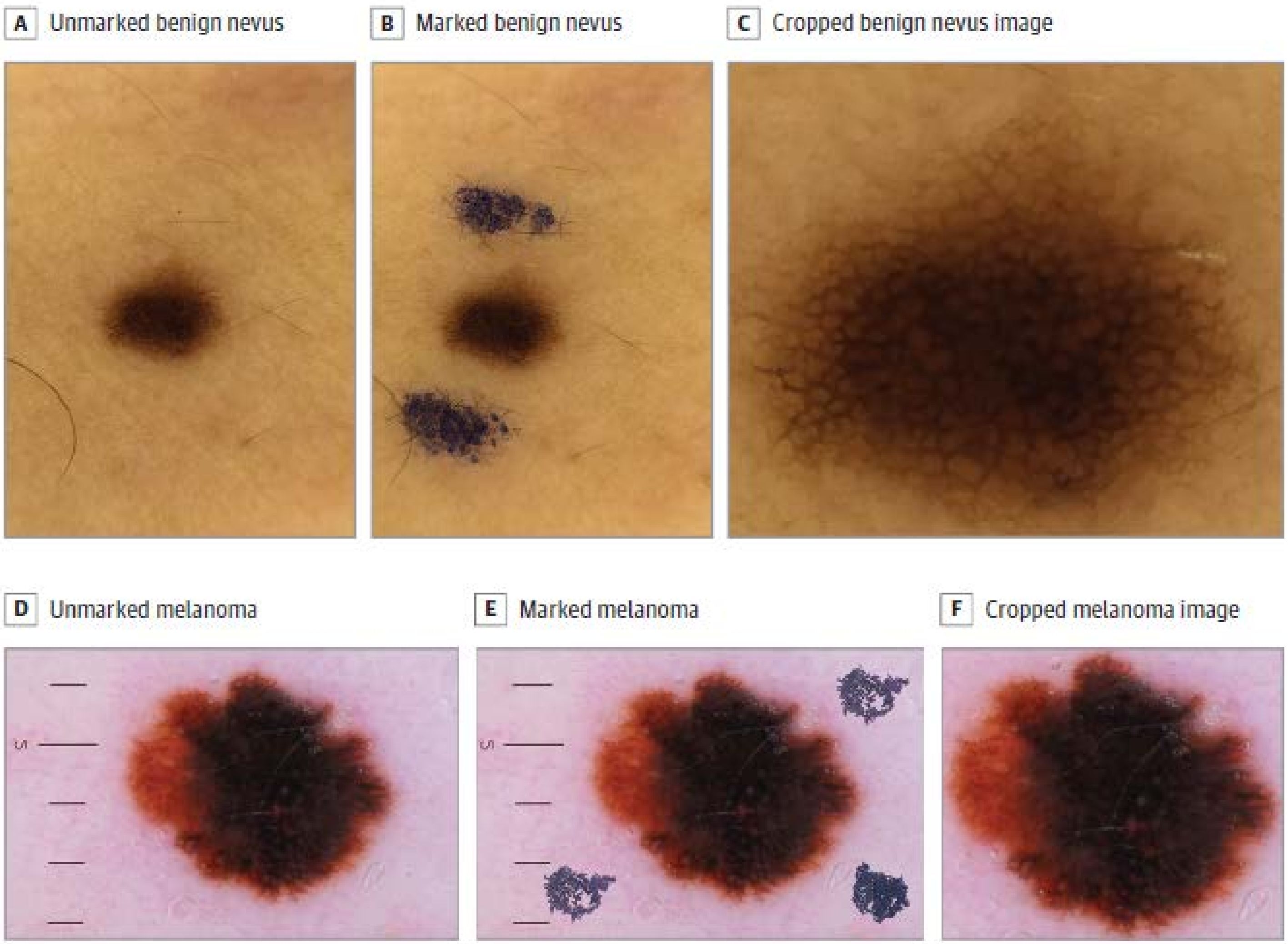


Figure 1. Convolutional Neural Network (CNN) Classification and Melanoma Probability Scores for Dermoscopic Images of Unmarked, Marked, and Cropped Benign Nevus and Melanoma



A gentian violet surgical skin marker was used to highlight the marked examples. A, CNN classification: benign; score, 0.001. B, CNN classification: malignant; score, 0.981. C, CNN classification: benign; score, 0.001. D, CNN classification: malignant; score, 0.999. E, CNN classification: malignant; score, 0.999. F, CNN classification: malignant; score, 0.999.

https://jamanetwork.com/journals/jamadermatology/article-abstract/2740808

Two possibilities

- Transfer learning
- Crowdsourcing

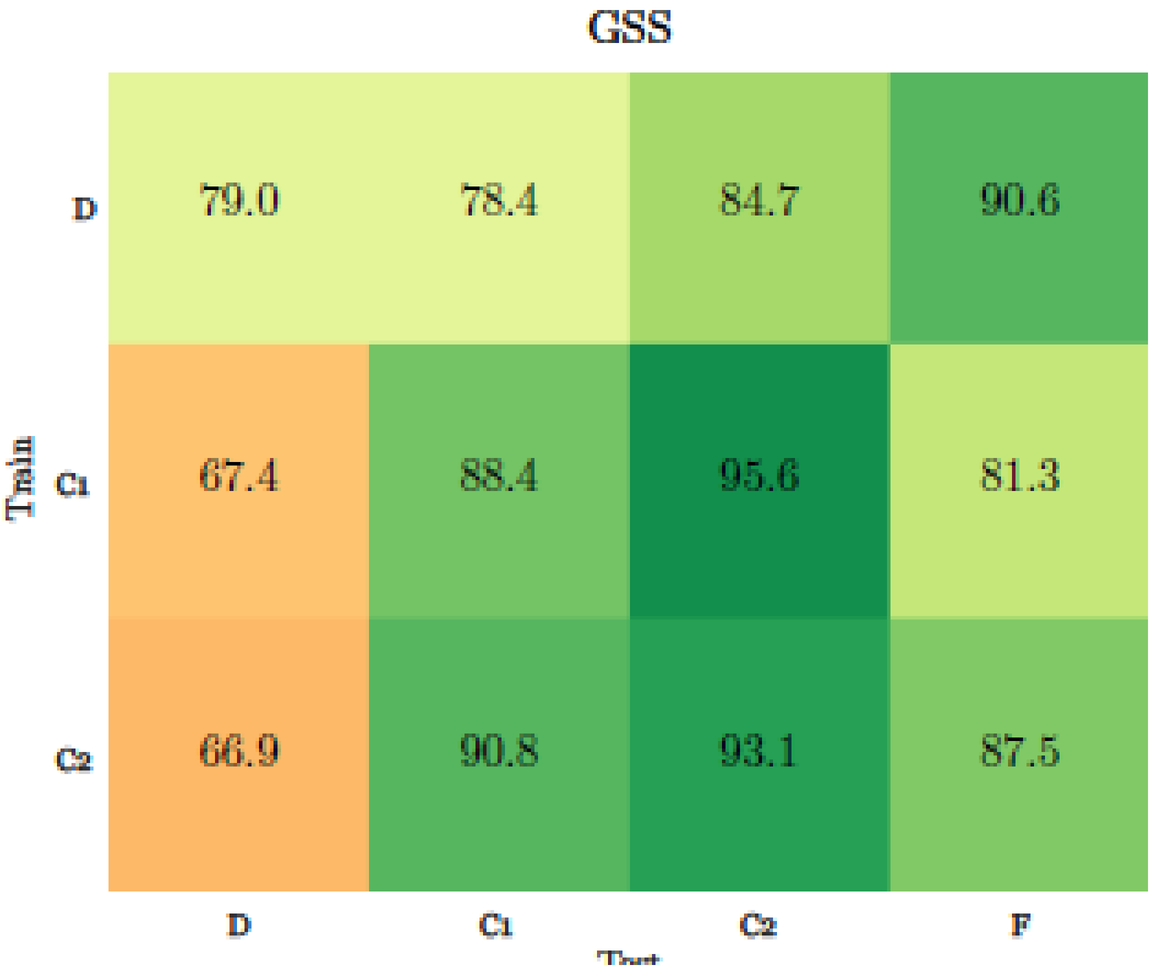
Transfer learning

Not learning “from scratch”

Use other similar datasets

Dataset	Subjects	Age	GOLD (1/2/3/4)	Smoking (c/f/n)	Scanner
DLCST	300 +	59 [50, 71]	69/28/2/0	77/23/0	Philips
	300 -	57 [49, 69]		74/26/0	16 rows Mx 8000
COPDGene1	74 +	64 [45, 80]	21/18/19/16	17/57/0	Siemens
	46 -	59 [45, 78]		23/20/3	Definition
COPDGene2	42 +	65 [45, 78]	9/13/7/13	12/30/0	Siemens
	25 -	60 [47, 78]		9/11/5	Definition AS+
Frederikshavn	8 +	66 [48, 77]	1/3/3/1	1/7/0	Siemens
	8 -	56 [25, 73]		1/2/5	Definition Flash

Performance drops across datasets



Cheplygina, V., Pena, I. P., Pedersen, J. H., Lynch, D. A., Sørensen, L., & de Bruijne, M. (2018). Transfer learning for multicenter classification of chronic obstructive pulmonary disease. *IEEE journal of biomedical and health informatics*, *22*(5), 1486-1496.

Performance drops across datasets

Test set	Training set	Atelectasis	Cardiomegaly	Consolidation
ChestX-ray14	**ChestX-ray14**	**0.8165**	**0.8998**	**0.8181**
	CheXpert	0.7850	0.8646	0.7771
	MIMIC-CXR	0.8024	0.8322	0.7898
CheXpert	ChestX-ray14	0.5137	0.5736	0.6565
	CheXpert	**0.6930**	**0.8687**	**0.7323**
	MIMIC-CXR	0.6576	0.8197	0.7002
MIMIC-CXR	ChestX-ray14	0.5810	0.6798	0.7692
	CheXpert	0.7587	0.7650	0.7936
	MIMIC-CXR	**0.8177**	**0.8126**	**0.8229**

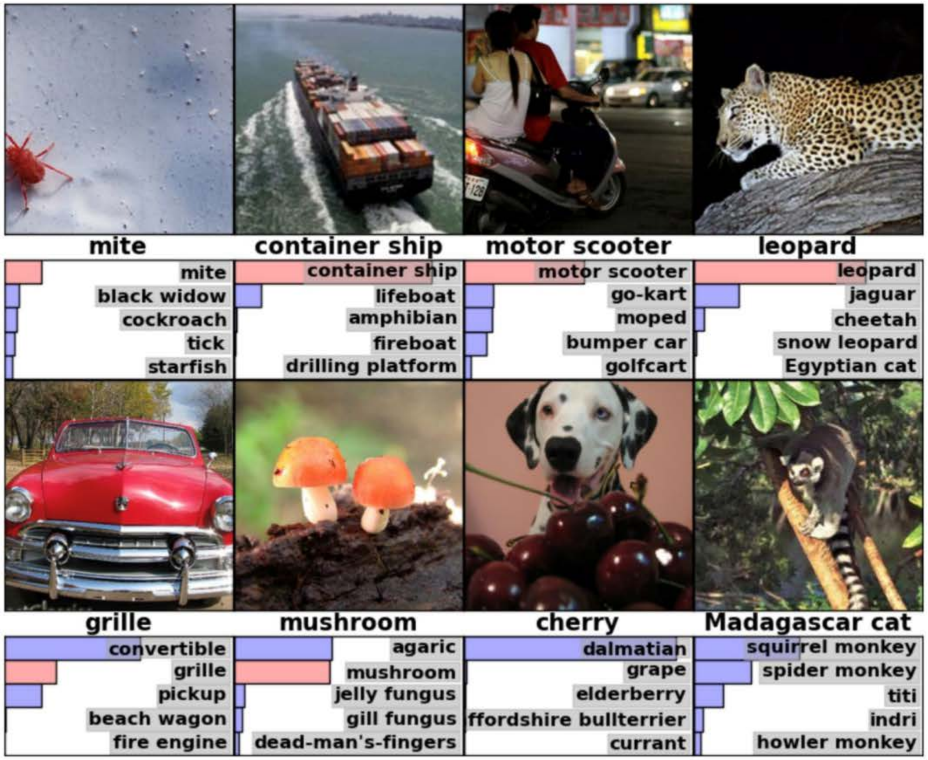
Pooch, E. H., Ballester, P. L., & Barros, R. C. (2019). Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. *arXiv preprint arXiv:1909.01940*.

Learn from any dataset

ImageNet Challenge

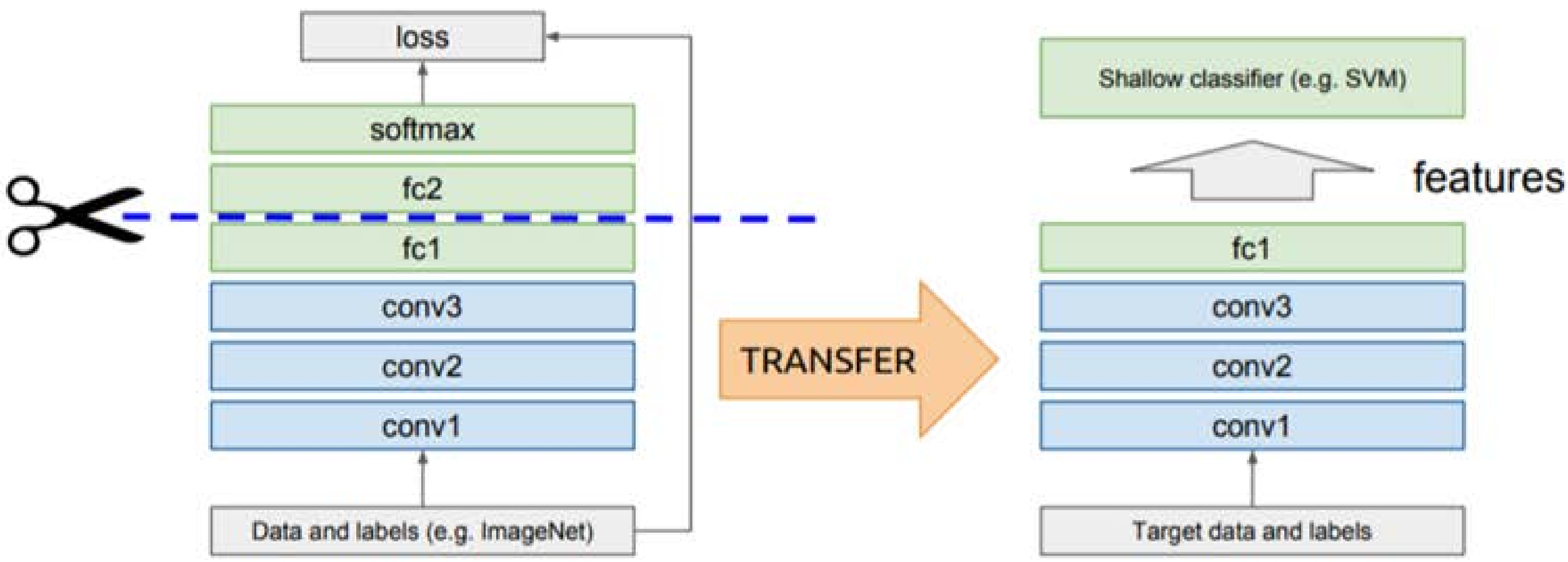
IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.



The ImageNet Large Scale Visual Recognition Challenge. (Source: [Xavier Giro-o-Nieto])

Learn from any dataset

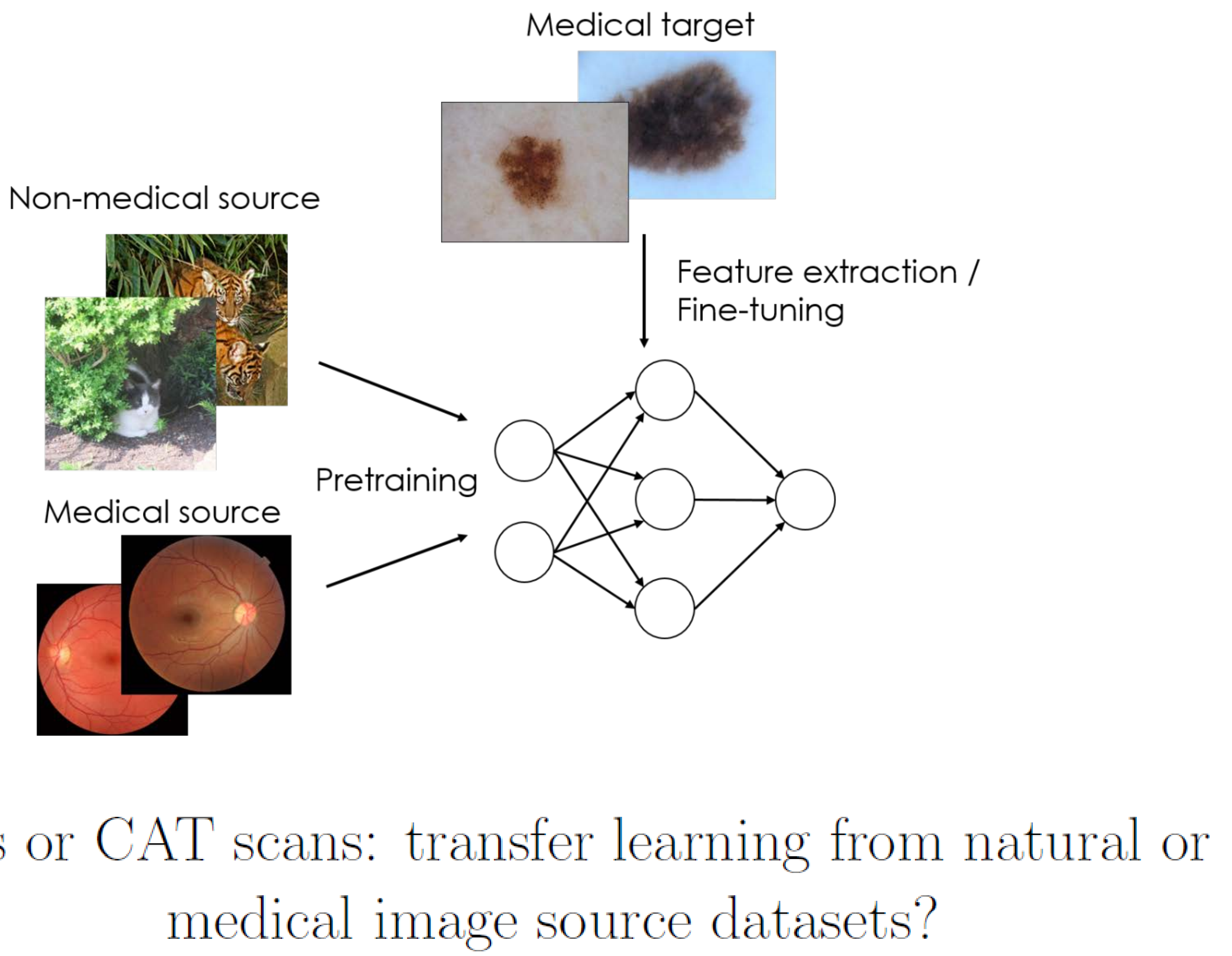



```
import keras
import numpy as np
from keras.applications import vgg16

#Load the VGG model
vgg_model = vgg16.VGG16(weights='imagenet')
```

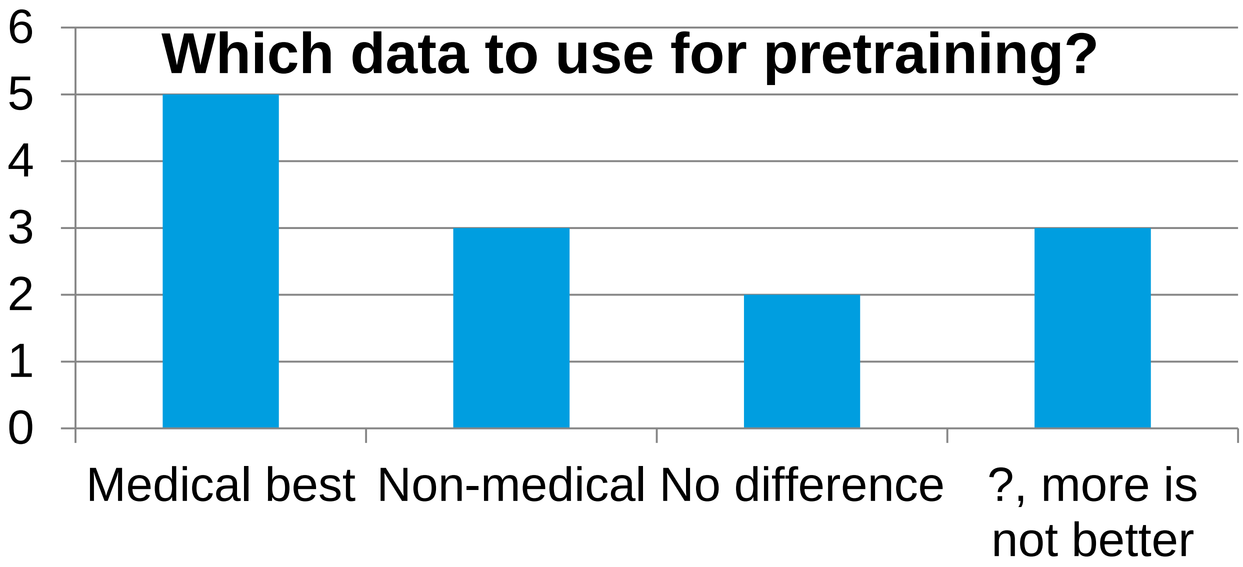
Image: towardsdatascience.com

Learn from any dataset – medical or non-medical?



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

Learn from any dataset – medical or non-medical?



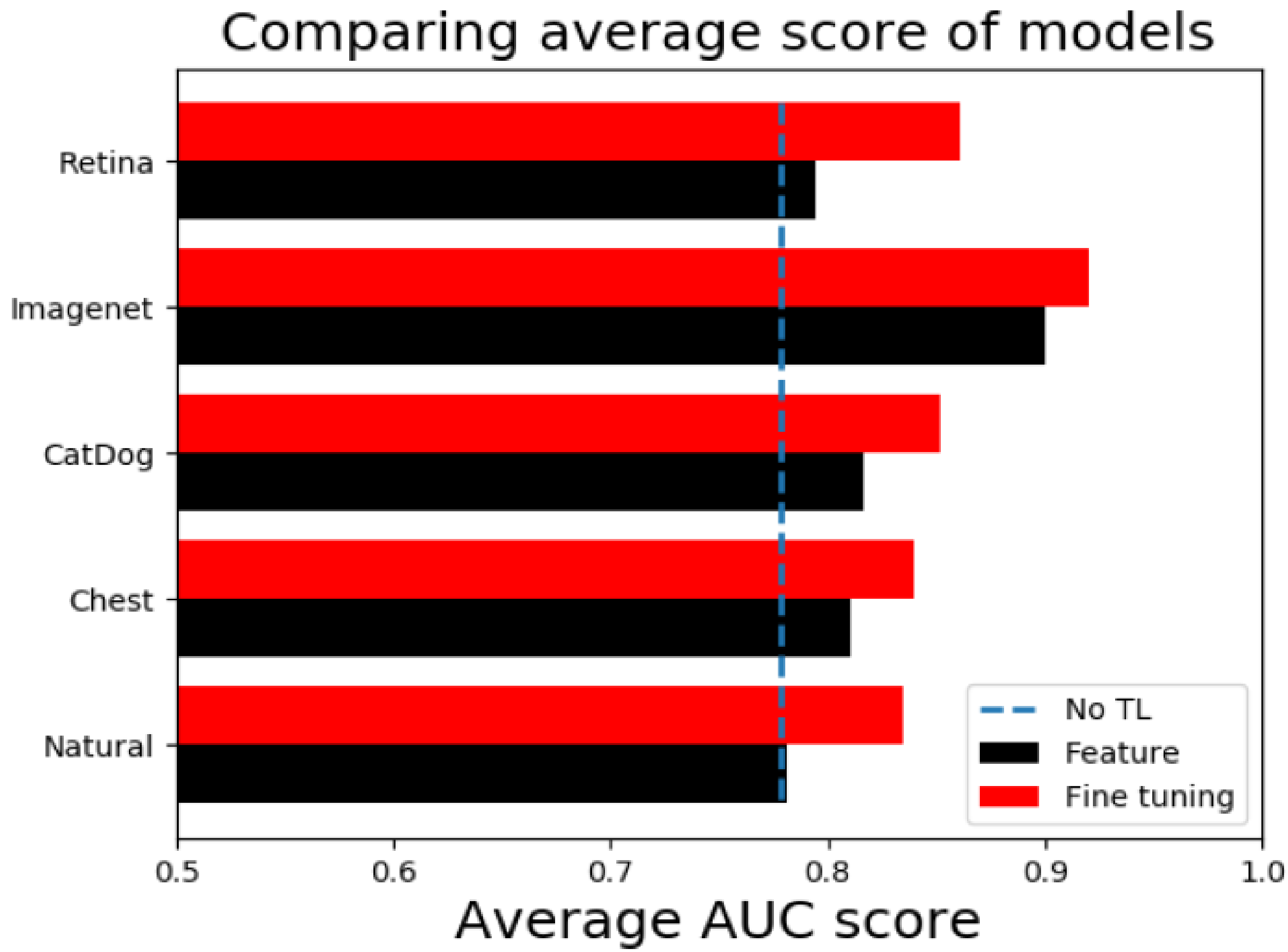
Cheplygina, V. (2019). Cats or CAT scans: transfer learning from natural or medical image source datasets?. *Current Opinion in Biomedical Engineering*. URL

Non-medical vs medical data

ImageNet best as source data

BUT

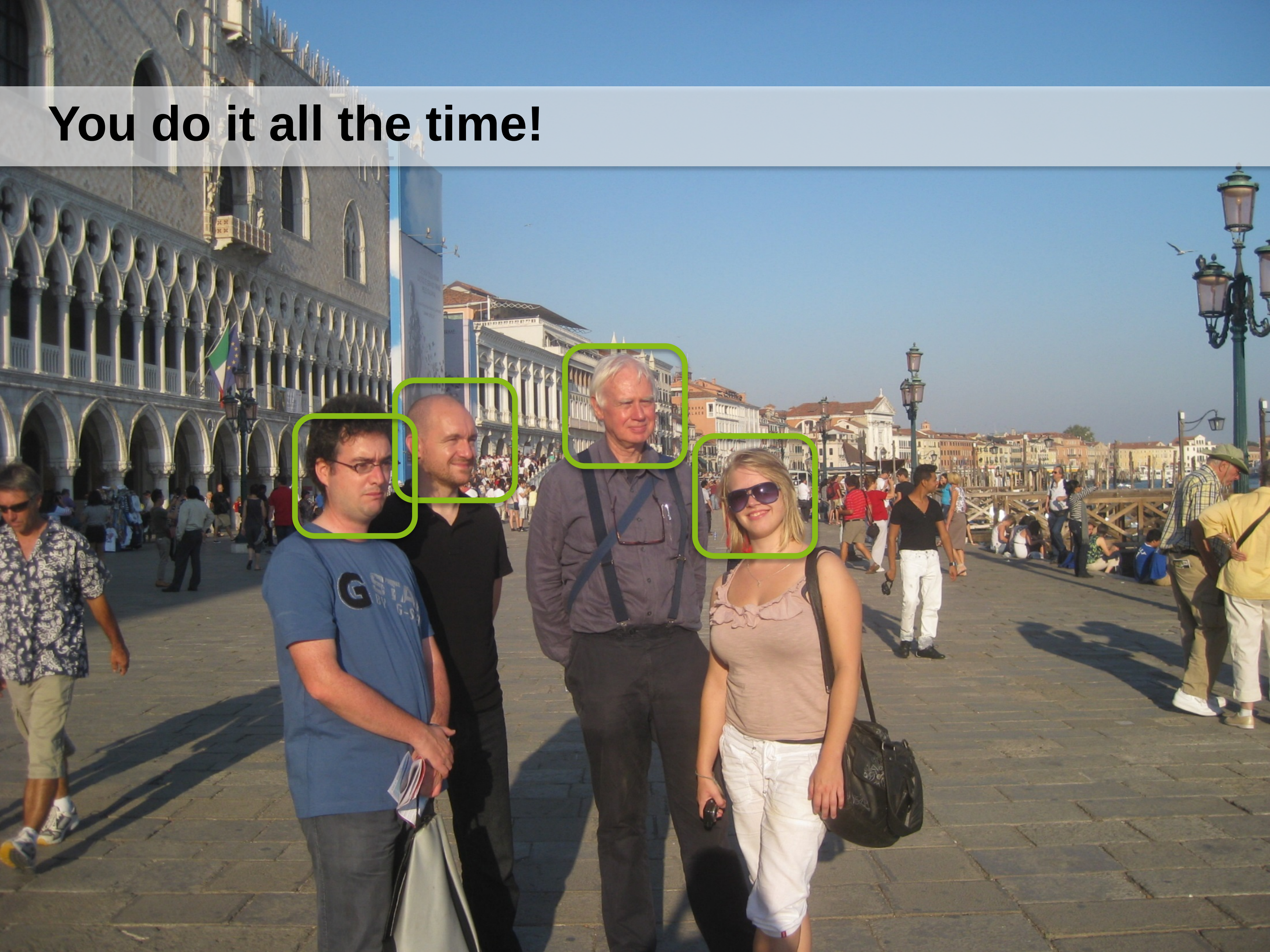
is Imagenet is much larger



Work by Floris Fok

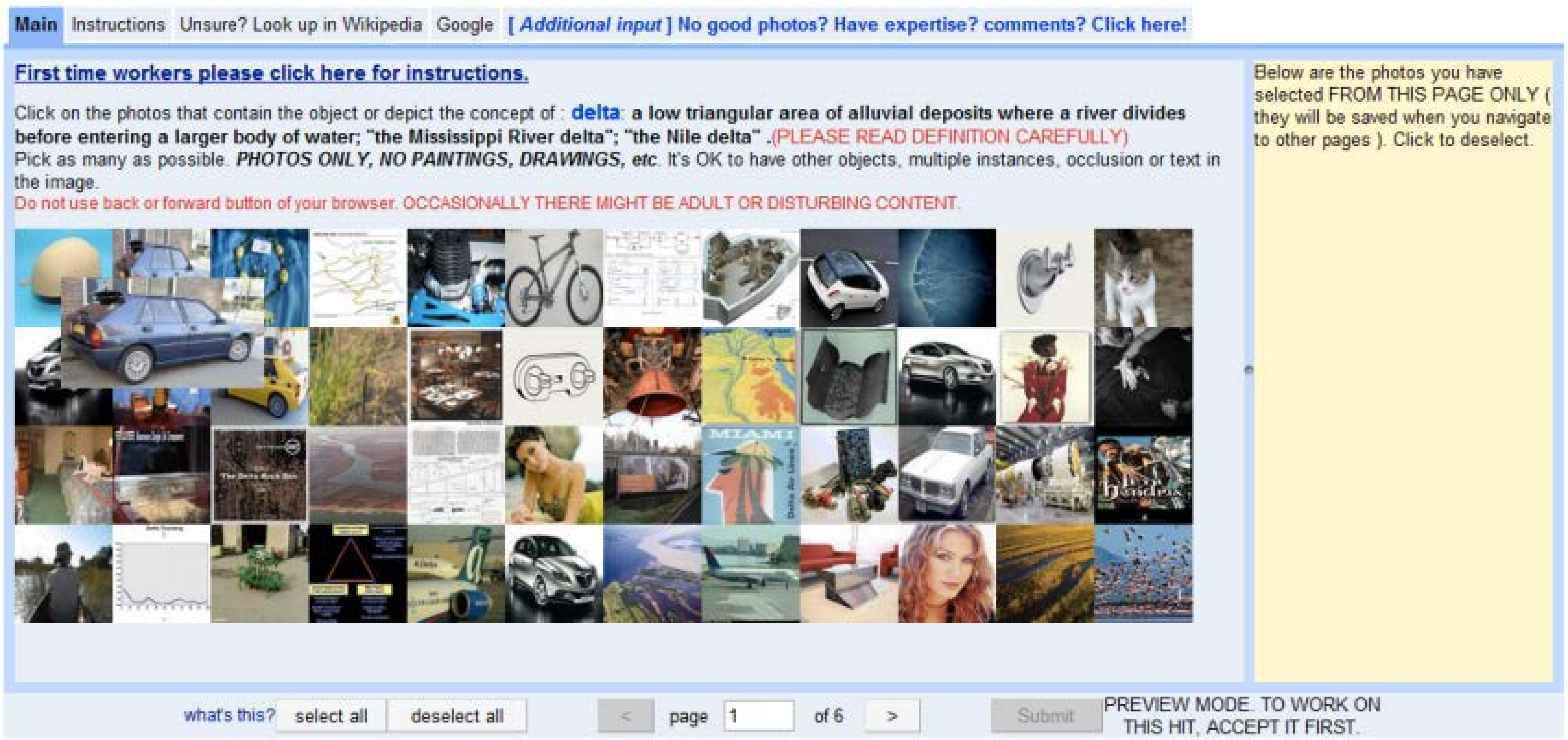
Crowdsourcing

You do it all the time!



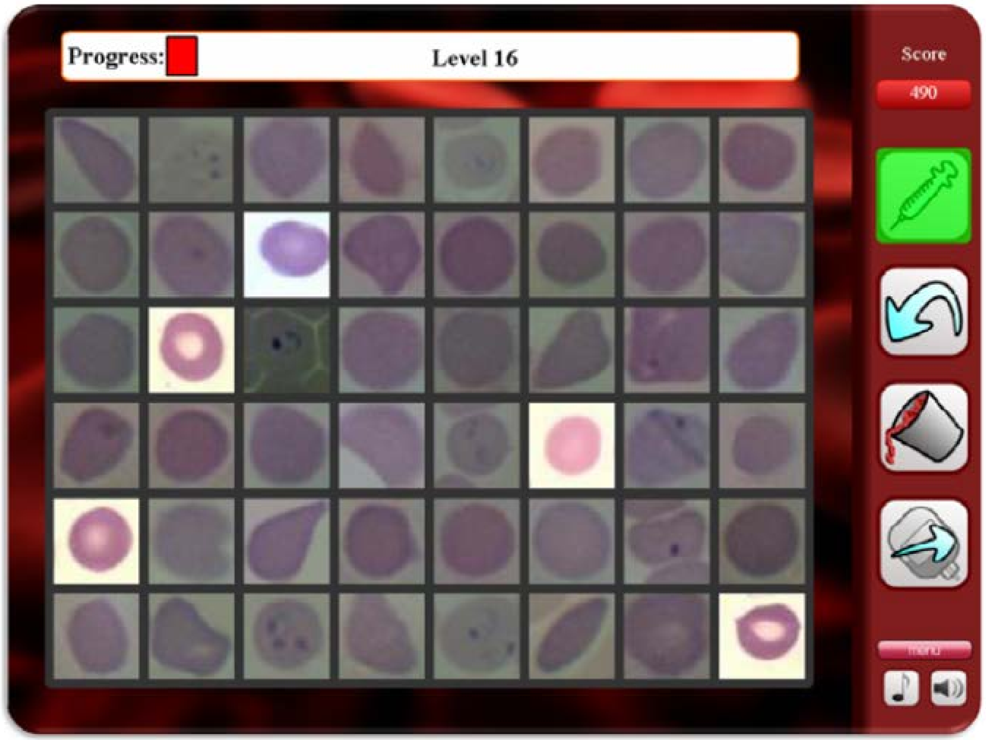
2009: ImageNet

amazon mechanical turk beta



Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 248-255). IEEE.

2012: Malaria diagnosis



Completing this game […] took on average less than one hour for each gamer

accuracy […] is within 1.25% of the diagnostic decisions made by the infectious disease expert.

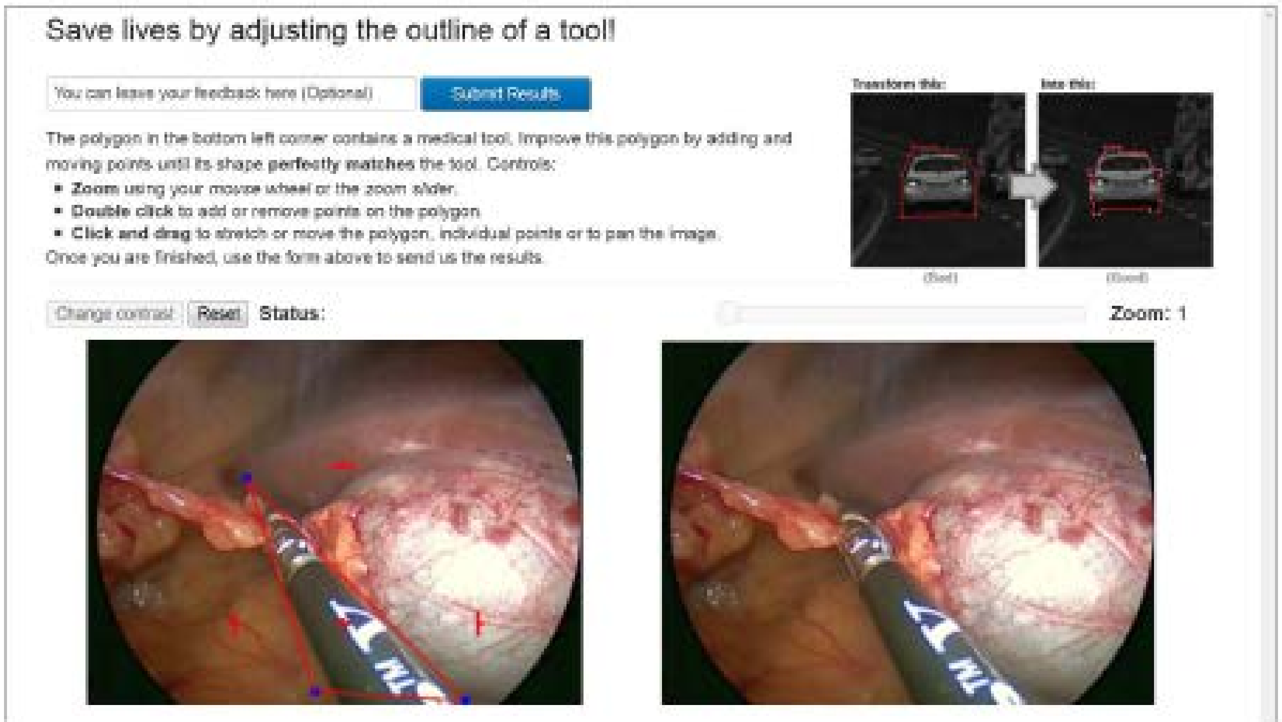
Mavandadi et al. Distributed Medical Image Analysis and Diagnosis through Crowd-Sourced Games: A Malaria Case Study, 2012a

Surgical instrument segmentation

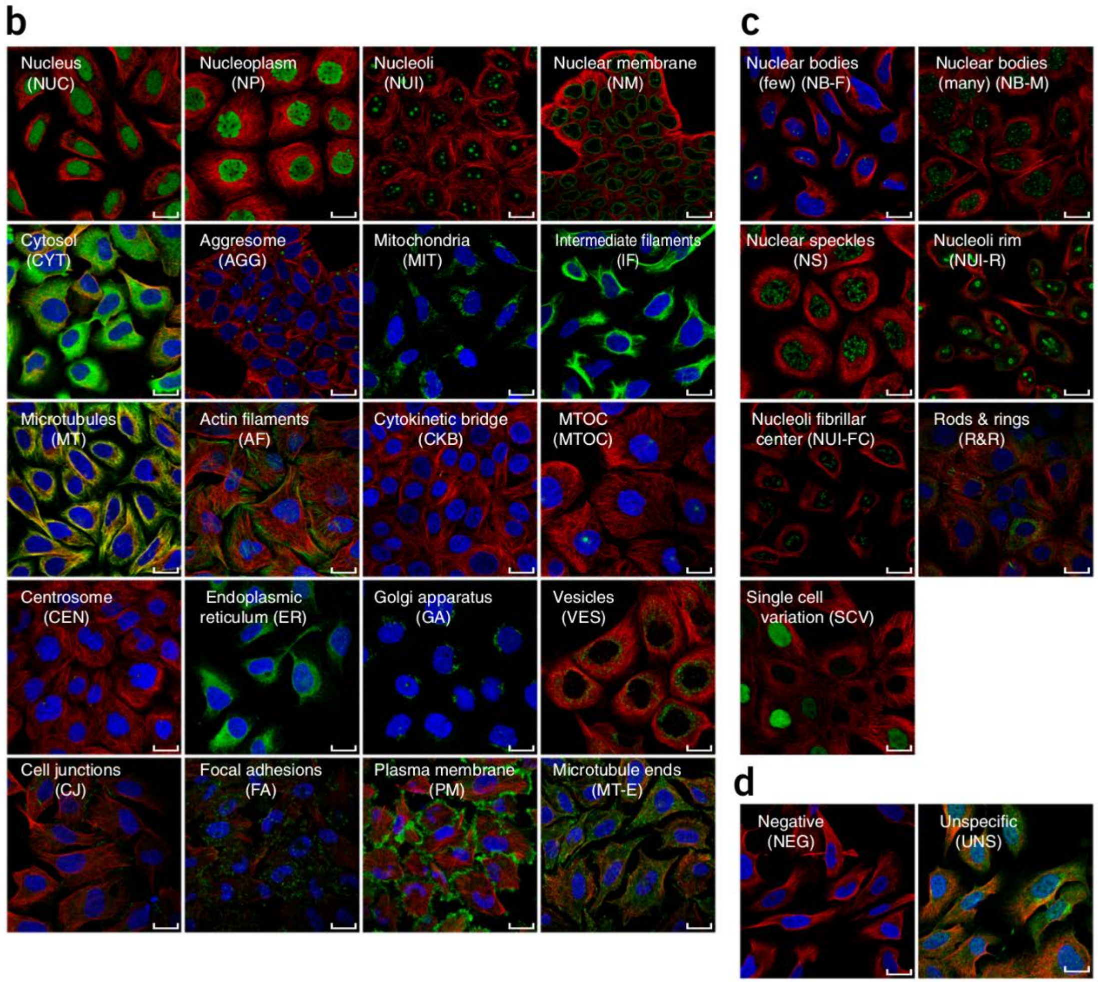
Can Masses of Non-Experts Train Highly Accurate Image Classifiers?

A Crowdsourcing Approach to Instrument Segmentation in Laparoscopic Images

Lena Maier-Hein[1,*,**], Sven Mersmann[1], Daniel Kondermann[2],
Sebastian Bodenstedt[3], Alexandro Sanchez[2], Christian Stock[4],
Hannes Gotz Kenngott[5], Mathias Eisenmann[3], and Stefanie Speidel[3]



Cell pattern classification



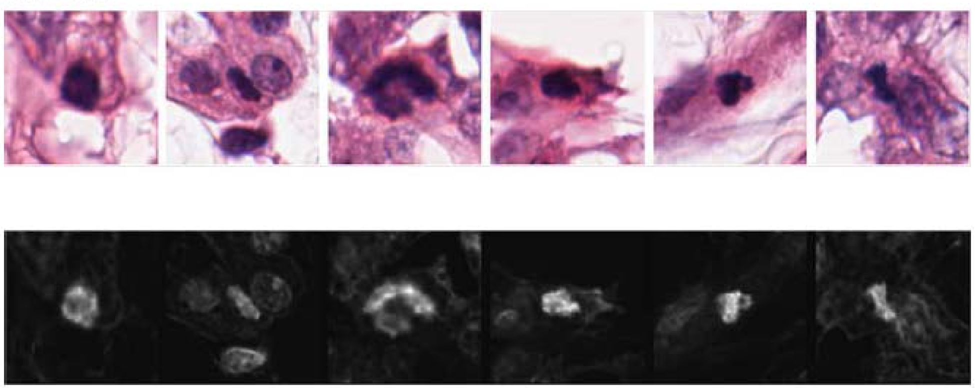
Sullivan et al. Deep learning is combined with massive-scale citizen science to improve large-scale image classification, 2018

Mitosis detection in histopathology

AggNet: Deep Learning From Crowds for Mitosis Detection in Breast Cancer Histology Images

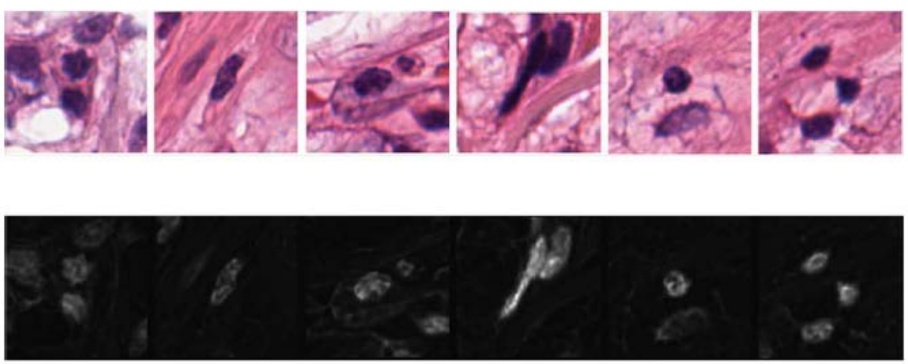
Shadi Albarqouni*, *Student Member, IEEE*, Christoph Baur, Felix Achilles, *Student Member, IEEE*,
Vasileios Belagiannis, *Student Member, IEEE*, Stefanie Demirci, and Nassir Navab, *Member, IEEE*

Mitosis:



The second row shows the corresponding so called "blueRatio" representation of the mitotic figures. Note how they have very bright spots!

Non-Mitosis

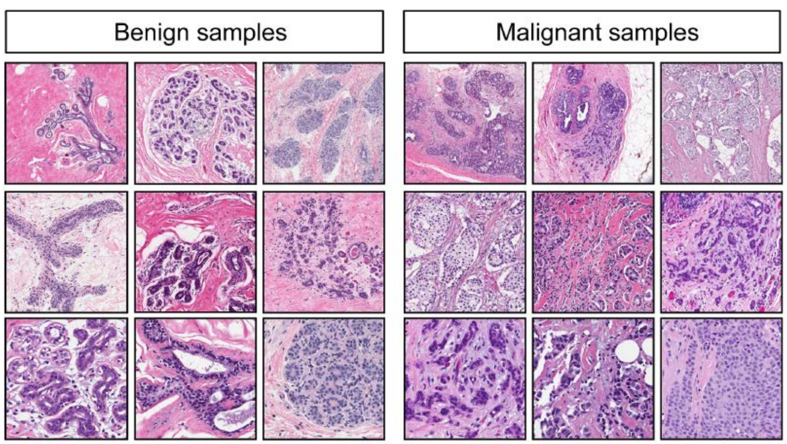


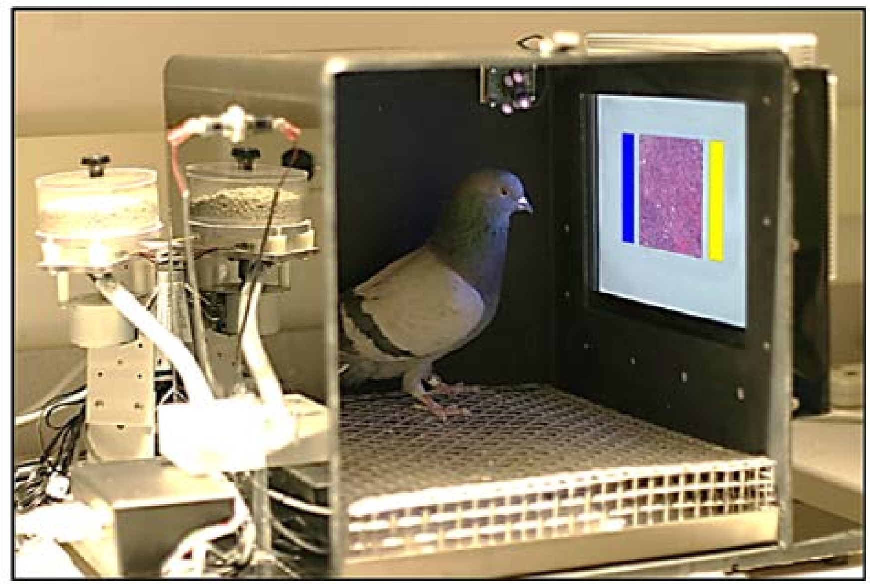
The second row shows the corresponding so called "blueRatio" representation of the non-mitotic figures. Note how they do not have such bright spots as the mitotic blue ratio representations!

RESEARCH ARTICLE

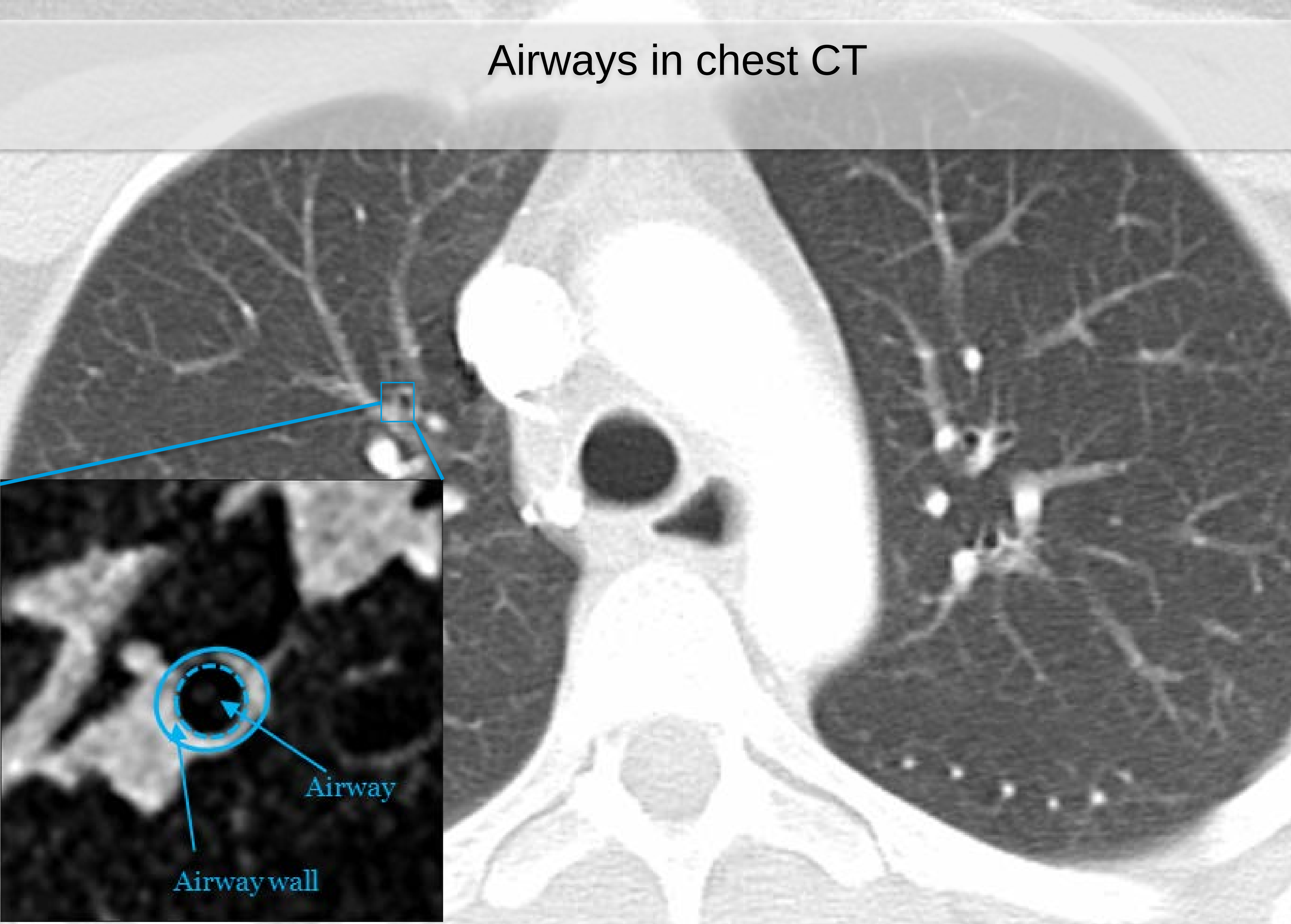
Pigeons (*Columba livia*) as Trainable Observers of Pathology and Radiology Breast Cancer Images

Richard M. Levenson[1]*, Elizabeth A. Krupinski[3], Victor M. Navarro[2], Edward A. Wasserman[2]*

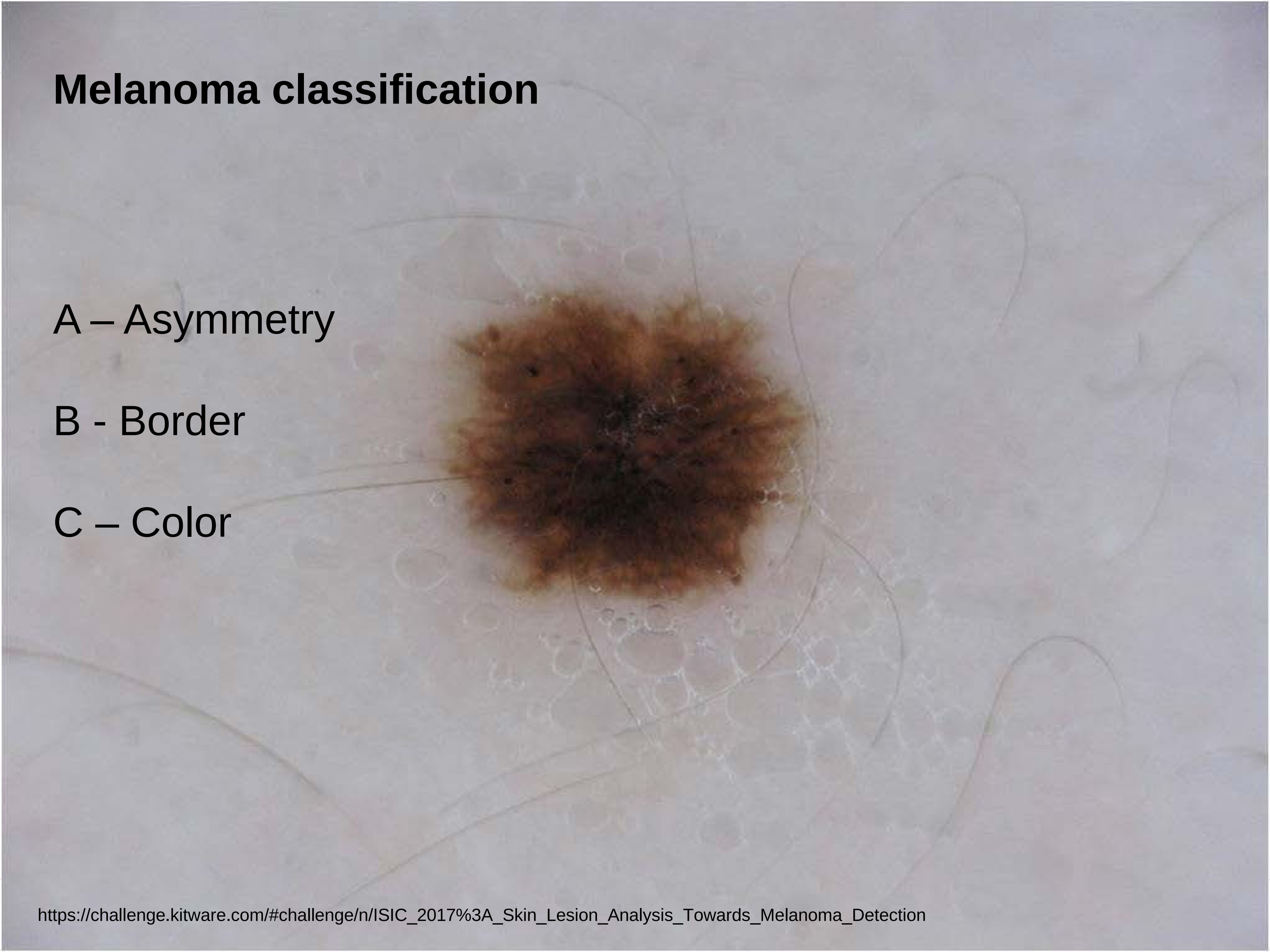




Airways in chest CT



Cheplygina, V et al. (2016). Early Experiences with Crowdsourcing Airway Annotations MICCAI LABELS 2016



Melanoma classification

A – Asymmetry

B - Border

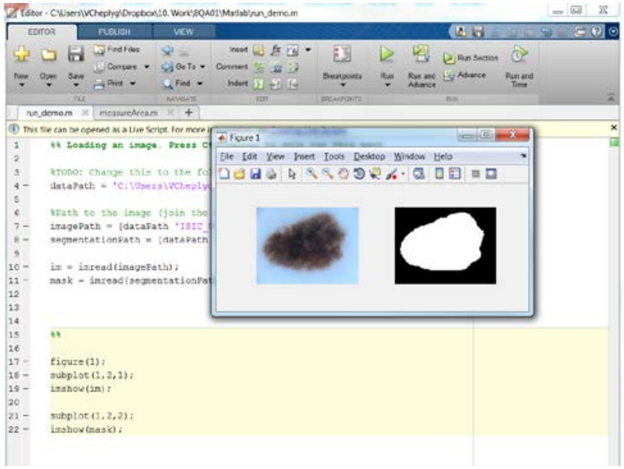
C – Color

https://challenge.kitware.com/#challenge/n/ISIC_2017%3A_Skin_Lesion_Analysis_Towards_Melanoma_Detection

Image analysis project for 1st year students

Technische Universiteit Eindhoven University of Technology

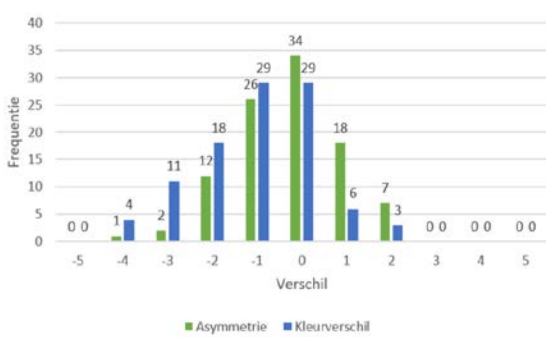
1. Measure features with algorithms



2. Measure features yourself

	A	B	C	D
1	ID	Asymmetry_7_1	Color_7_1	Border_7_1
2	ISIC_0000549	2	4	1
3	ISIC_0000550	1	3	1
4	ISIC_0000551	2	2	1
5	ISIC_0000552	1	4	1
6	ISIC_0000554	2	3	1
7	ISIC_0000555	2	3	1
8	ISIC_0001100	2	5	1
9	ISIC_0001102	2	5	1
10	ISIC_0001103	1	5	1
11	ISIC_0001105	0	2	1
12	ISIC_0001118	2	5	1
13	ISIC_0001119	2	3	1
14	ISIC_0001126	2	2	1
15	ISIC_0001128	1	3	1
16	ISIC_0001131	1	5	1
17	ISIC_0001133	1	5	1
18	ISIC_0001134	2	3	1
19	ISIC_0001140	2	2	1
20	ISIC_0009923	1	2	1
21	ISIC_0009925	2	2	1
22	ISIC_0009929	1	2	1
23	ISIC_0009930	1	2	1
24	ISIC_0009931	1	3	1
25	ISIC_0009932	2	3	1
26	ISIC_0009933	1	2	1
27	ISIC_0009935	1	3	1
28	ISIC_0009936	1	2	0

3. Evaluate

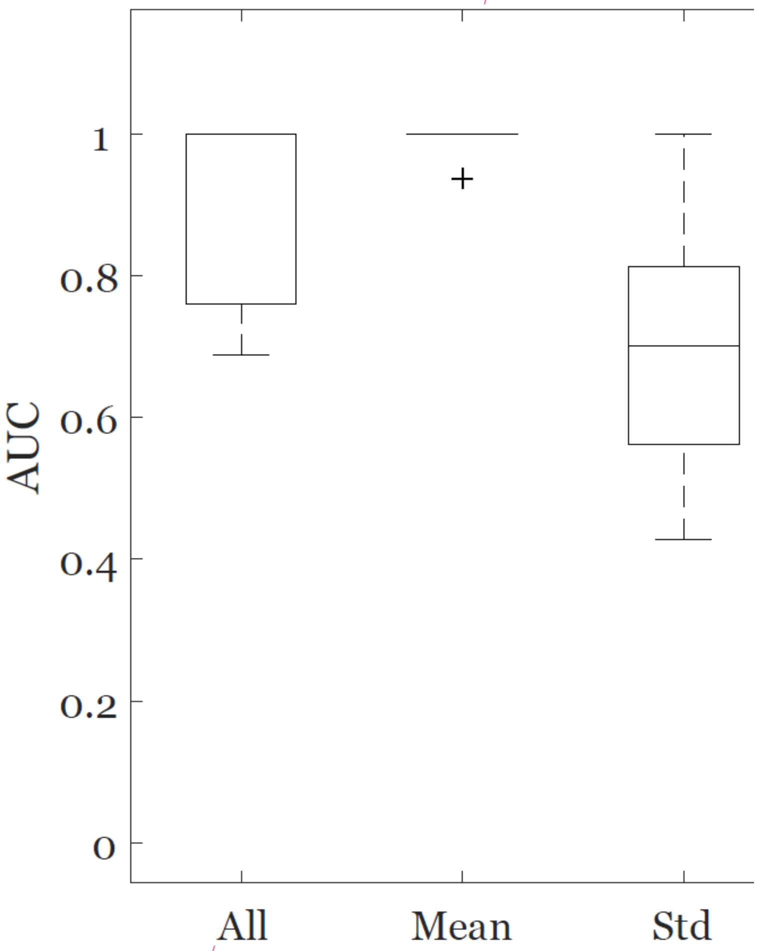


Grafiek 1: De frequenties van de verschilwaardes tussen de metingen in Matlab en de metingen op het oog

Crowdsourcing!

Crowd annotations predict diagnosis

- 100 images, 5 features x 6 people = 30 features
- Averaging annotators best
- Disagreement also informative

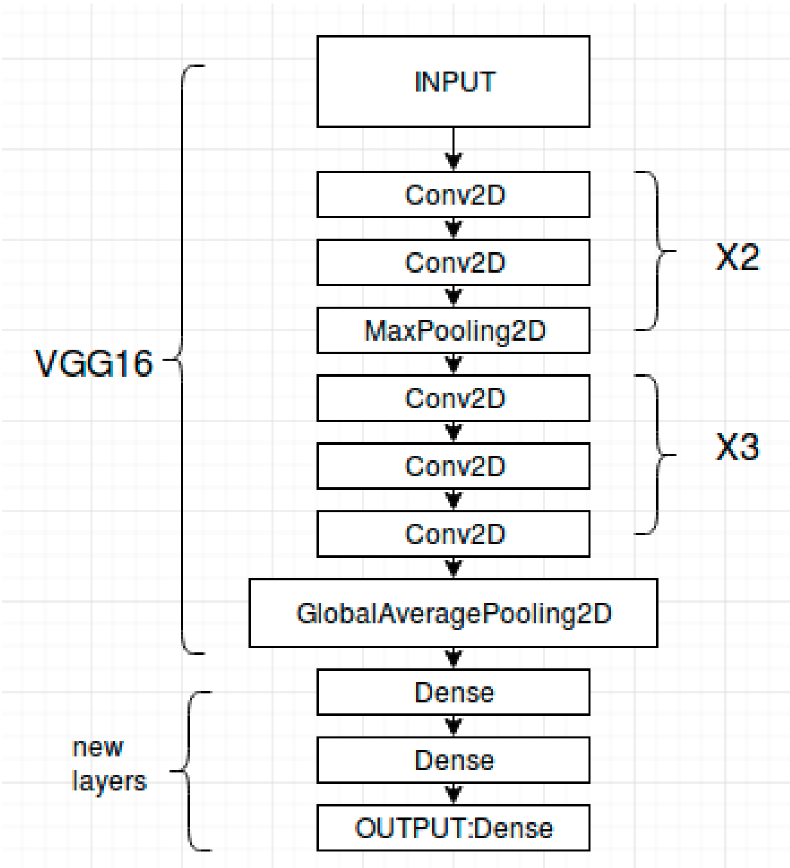


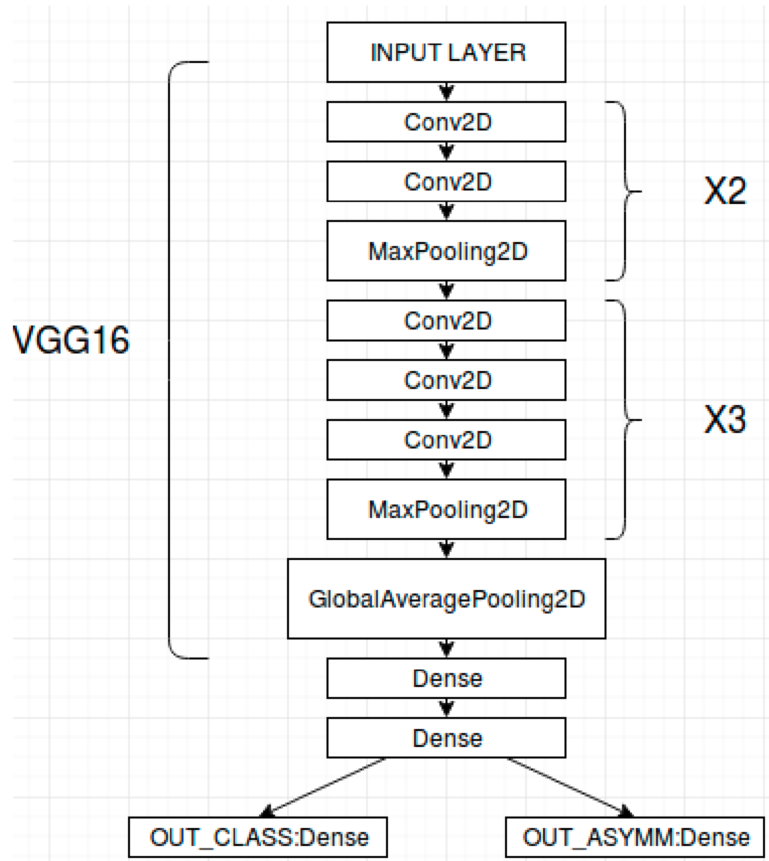
Cheplygina, V., & Pluim, J. P. W. (2018). Crowd disagreement about medical images is informative. [URL]

Work by Elif Kubra Contar

Same network

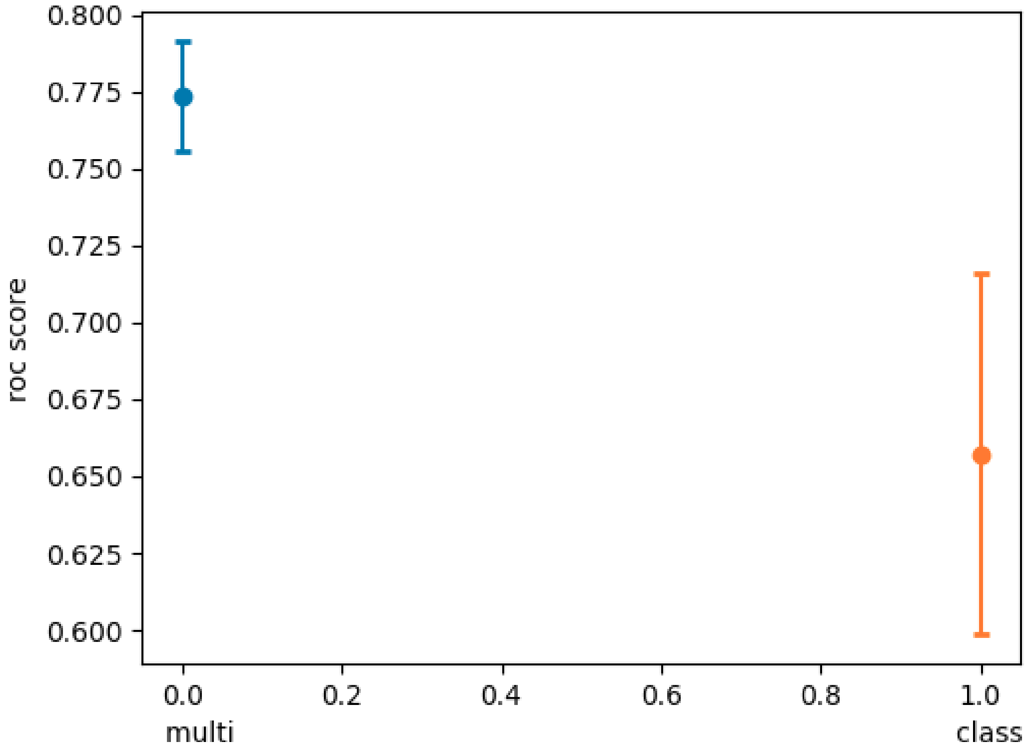
- Single-task with class label
- Multi-task with class label and asymmetry





Work by Elif Kubra Contar

Multi-task network with crowd annotations outperforms single-task network



A Survey of Crowdsourcing in Medical Image Analysis

Silas Ørting[1] ✉, Andrew Doyle[2,*], Matthias Hirth[3,*] Arno van Hilten[4,*], Oana Inel[5,7,*], Christopher R. Madan[6,*], Panagiotis Mavridis[7,*], Helen Spiers[8,9,*], and Veronika Cheplygina[10] ✉

[1] University of Copenhagen, Copenhagen, Denmark
[2] McGill Centre for Integrative Neuroscience, Montreal, Canada
[3] Technische Universität Ilmenau, Ilmenau, Germany
[4] Erasmus Medical Center, Rotterdam, The Netherlands
[5] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[6] University of Nottingham, Nottingham, United Kingdom
[7] Delft University of Technology, Delft, The Netherlands
[8] University of Oxford, Oxford, United Kingdom
[9] Zooniverse, University of Oxford, Oxford
[10] Eindhoven University of Technology, Eindhoven, The Netherlands

Ørting, S., Doyle, A., van Hilten, A., Hirth, M., Inel, O., Madan, C. R., Mavridis, P., ... & Cheplygina, V. (2019). A survey of crowdsourcing in medical image analysis. *arXiv preprint arXiv:1902.09159*

Survey of crowdsourcing – take-aways

- Often 2D images, rating entire image
- Almost all papers report successes

Application	This survey	Cheplygina et al. [2018]	Litjens et al. [2017]
Brain	9%	21%	18%
Eye	15%	4%	5%
Lung	9%	13%	14%
Breast	0%	6%	7%
Heart	2%	4%	7%
Abdomen	22%	14%	9%
Histo/Micro	29%	17%	20%
Multiple	7%	12%	4%
Other	7%	10%	16%

TABLE I
COMPARISON OF THE DISTRIBUTION OF APPLICATIONS IN THIS SURVEY AND TWO OTHER RECENT SURVEYS IN MEDICAL IMAGE ANALYSIS.

Technische Universiteit Eindhoven University of Technology

Survey of crowdsourcing - take-aways

- Setup ad-hoc / details missing
  - Platform, number of annotators, compensation…
- Different use of labels
  - Create labels vs improve/filter labels
  - Compare to experts vs train ML
  - Discover novel patterns
- Discussion of implications?

- Transfer learning
  - Train on similar datasets – performance can drop
  - Transfer weights from any dataset
  - Factors affecting success not 100% clear

- Crowdsourcing
  - Collect labels from the crowd for medical tasks
  - When is it successful?
  - Different ways of using crowd input

@drveronikach

http://**www.veronikach.com**


THANKS
@drveronikach
http://www.veronikach.com