

Medical Image Processing With Neural Networks

Classification of Chest X-rays
and detection in endoscopy

Applications

- Classification
 - Total diagnosis of single images
 - Tumor classes, thorax pathologies
- Segmentation
 - Segment exact contour/volume based on relevance
 - White and grey matter of the brain
- Detection
 - Localization often in real time
 - Polyp detection in endoscopy, diagnosing lung nodules in CT scans

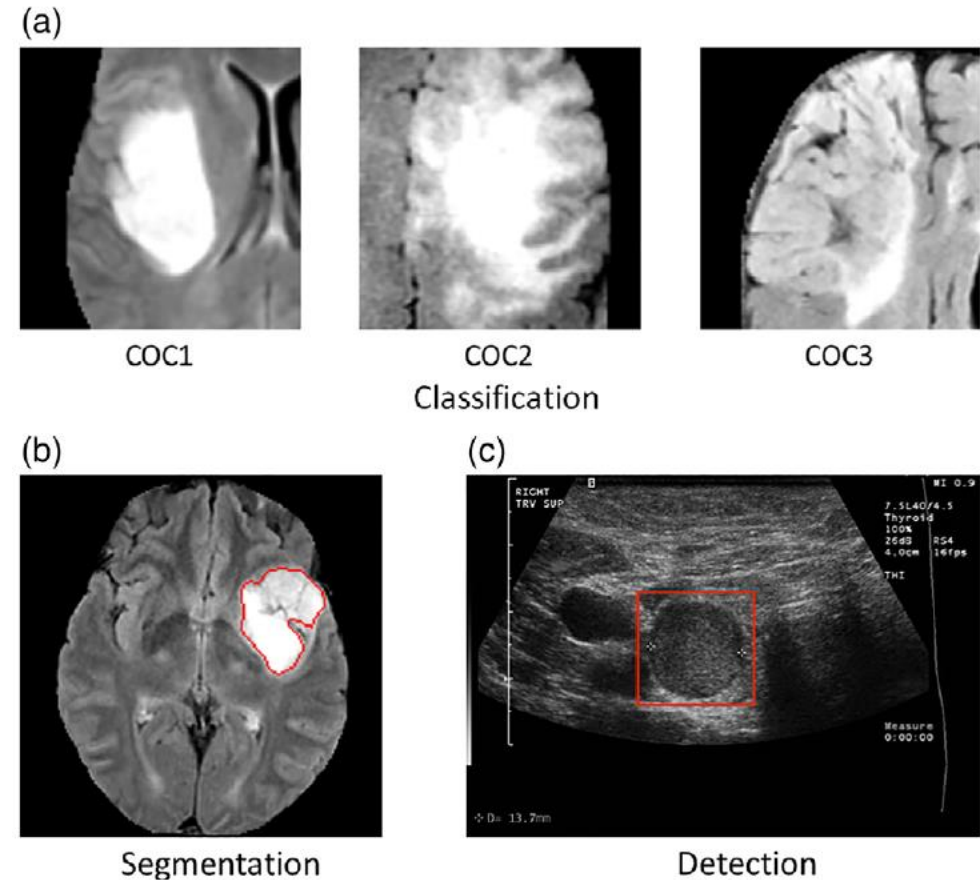


Table of contents

1. Classifying Chest X-ray pathologies
2. Pathology detection in endoscopy
3. Outlook: The role of AI in medicine

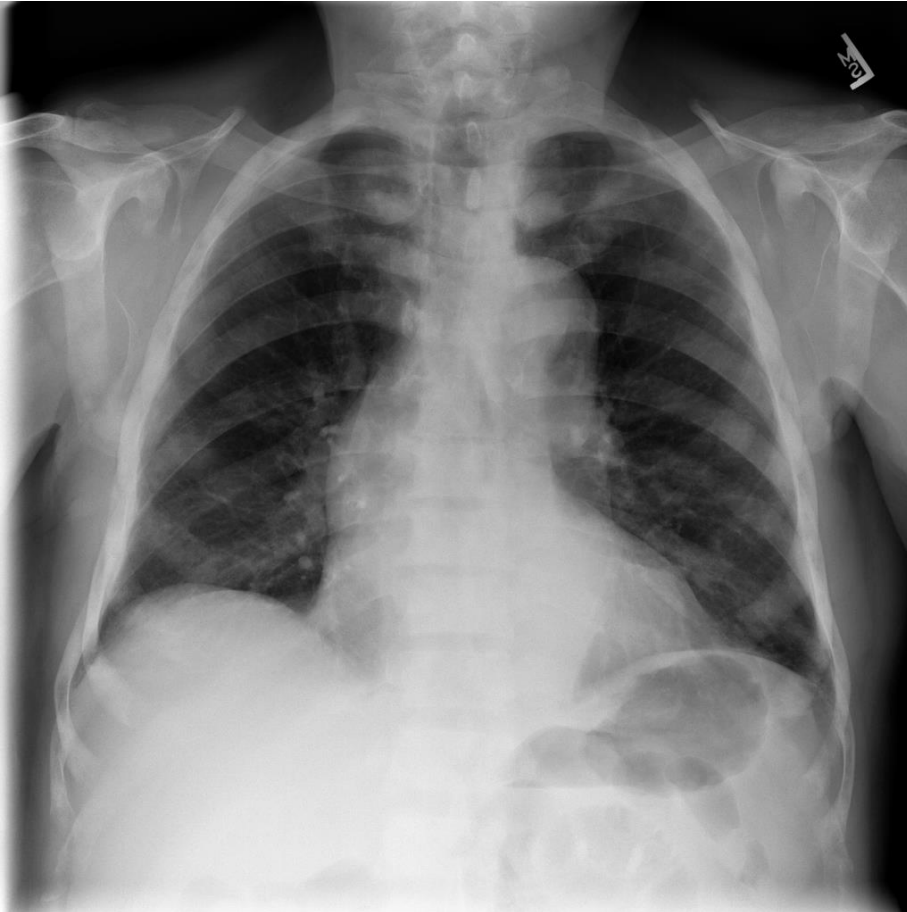


Chest X-ray

Datasets and state of the art



What is a Chest X-ray?



- 2D scan of the chest
- grayscale
- Easy and quick method
- Allows for early diagnosis
- Contains many information
 - Heart size
 - Lung shape
 - Nodules
 - ...

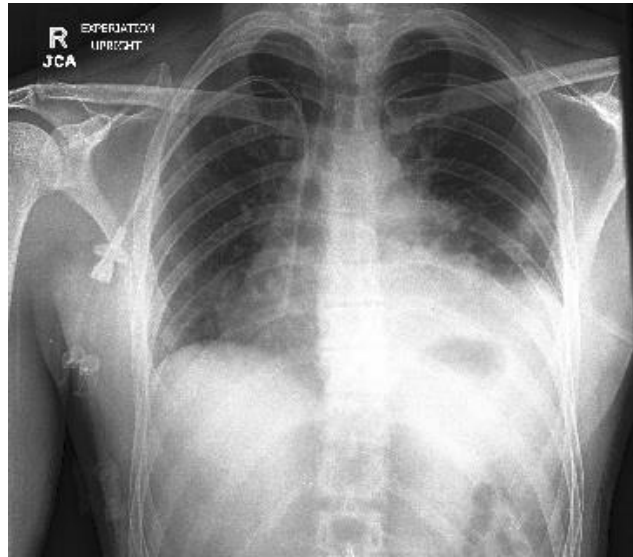


Types of Chest X-rays

- Image properties change based on the view



Postero-Anterior (PA)



Antero-Posterior (AP)



Lateral

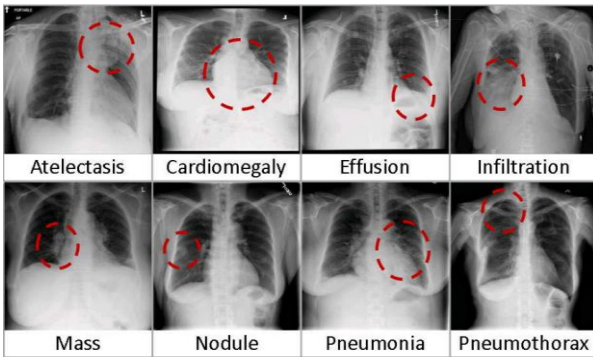
Frontal

Meaning of Chest X-rays

- Allows for fast detection of a number of pathologies:
 - Pneumonia
 - Cancer
 - Injuries or fractures etc.
 - Time for a radiologist to assess a scan: ~ **min** / image
 - Time for a CNN to assess a scan: ~ **ms** / image
- With hundreds of X-rays per day this is an enormous time save



Openly accessible datasets



- ChestXray14
 - > 100.000 Images
 - 14 Classes
- CheXpert
 - > 220.000 Images
 - 14 Classes

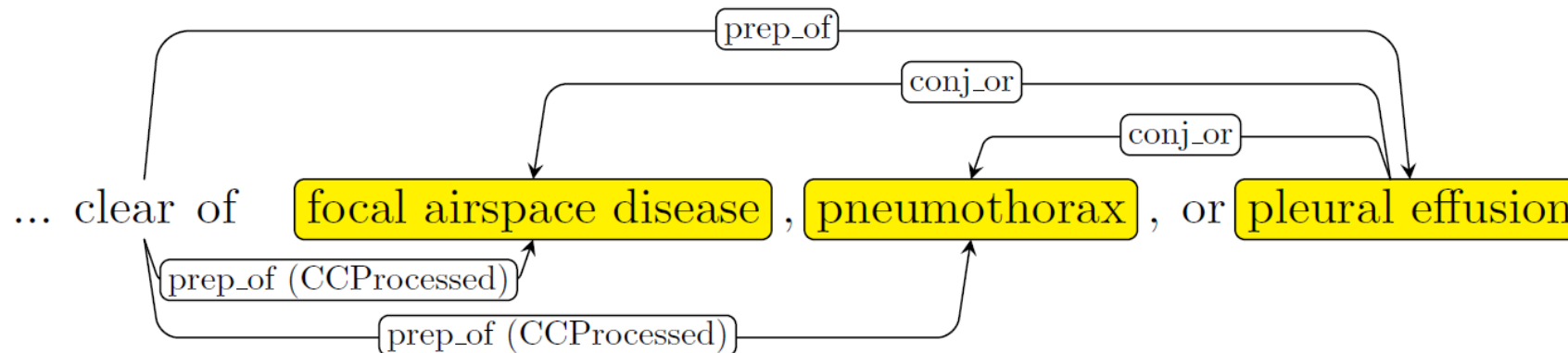


- MIMIC
 - Needs special access
- OpenI
 - Labels need to be extracted from reports first



ChestXray14

- 112.120 Images
- from 30.805 patients
- Uniform resolution of 1024×1024
- 14 different pathologies
- Labels generated from reports
 - DNorm, MetaMap
- Critique: Labels can be false
 - Parsing accuracy



ChestXray14

- Labels cover the 14 most common pathologies from the reports
- No nuance in labels
 - Only present or not present

Important:

- Label: „No Finding“
 - Means none of THESE 14 pathologies are present
 - Does NOT mean healthy

Pathology	Count
No Finding	60361 (53.83%)
Infiltration	19894 (17.74%)
Effusion	13317 (11.88%)
Atelectasis	11559 (10.31%)
Nodule	6331 (5.65%)
Mass	5782 (5.16%)
Pneumothorax	5302 (4.73%)
Consolidation	4667 (4.16%)
Pleural Thickening	3385 (3.02%)
Cardiomegaly	2776 (2.48%)
Emphysema	2516 (2.24%)
Edema	2302 (2.05%)
Fibrosis	1686 (1.50%)
Pneumonia	1431 (1.28%)
Hernia	227 (0.20%)

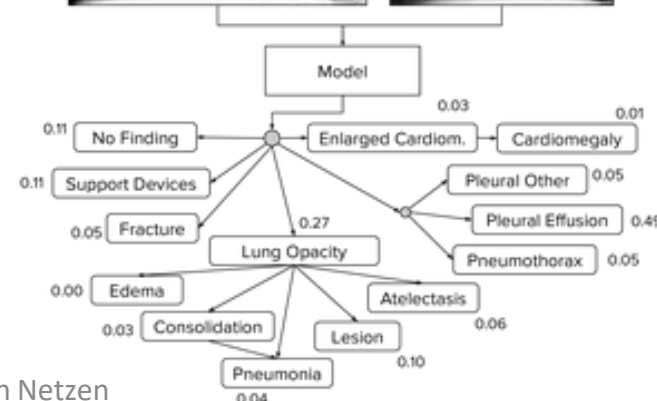
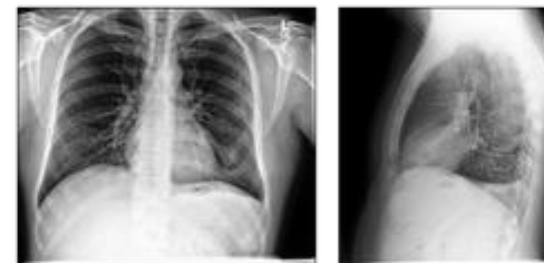


CheXPert

- 224.316 Images from 65.240 patients
 - Variable resolution
 - Contains lateral images
- Label extraction:
 - Mention Extraction,
 - Mention Classification
 - Mention Aggregation
- Labeler Performance (F1)
 - Mention: 0.948
 - Negation: 0.899
 - Uncertain: 0.770
- Labels have node structure from subclasses

Example labeling from the CheXpert paper

	Observation	Labeler Output
1. <i>unremarkable</i> <u>cardiomediastinal silhouette</u>	No Finding	
	Enlarged Cardiom.	0
	Cardiomegaly	
2. diffuse <u>reticular pattern</u> , which can be seen with an atypical <u>infection</u> or chronic fibrotic change. <i>no</i> focal <u>consolidation</u> .	Lung Opacity	1
	Lung Lesion	
	Edema	
	Consolidation	0
	Pneumonia	u
3. <i>no</i> <u>pleural effusion</u> or <u>pneumothorax</u>	Atelectasis	
	Pneumothorax	0
	Pleural Effusion	0
	Pleural Other	
4. mild degenerative changes in the lumbar spine and old right rib <u>fractures</u> .	Fracture	1
	Support Devices	



Hierarchical
label structure



CheXPert

- 14 Labels with nuance
 - Present
 - Uncertain
- During training: How to use uncertain labels?
 - Ignore
 - Set to 0/1
 - random?

Pathology	Number (%)	Number uncertain (%)
No Finding	16627 (8.86)	0 (0.0)
Enlarged Cardiom.	9020 (4.81)	10148 (5.41)
Cardiomegaly	23002 (12.26)	6597 (3.52)
Lung Lesion	6856 (3.65)	1071 (0.57)
Lung Opacity	92669 (49.39)	4341 (2.31)
Edema	48905 (26.06)	11571 (6.17)
Consolidation	12730 (6.78)	23976 (12.78)
Pneumonia	4576 (2.44)	15658 (8.34)
Atelectasis	29333 (15.63)	29377 (15.66)
Pneumothorax	17313 (9.23)	2663 (1.42)
Pleural Effusion	75696 (40.34)	9419 (5.02)
Pleural Other	2441 (1.3)	1771 (0.94)
Fracture	7270 (3.87)	484 (0.26)
Support Devices	105831 (56.4)	898 (0.48)



Critique on the available datasets

- ChestXray14 sample from university clinic Wü radiologists

- Suspicious vs. normal
- Labelled 1001 randomly chosen images
- Agreement: 74%
 - 13% wrong for Finding
 - 37% wrong for No Finding

Right: Radiologists Down: ChestXray14	Suspicious	Normal
Finding	401	61
No Finding	198	341

- A radiologist's critique on CXR14:
<https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/>
- Same radiologist on CheXpert:
<https://lukeoakdenrayner.wordpress.com/2019/02/25/half-a-million-x-rays-first-impressions-of-the-stanford-and-mit-chest-x-ray-datasets/>



Critique of a radiologist: Details on CXR14

<https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/>

- Low visual variability (100k images vs. 30k patients)
- Pneumonia, Emphysema, Fibrosis are often diagnosed clinically and not by image
- Labels partially describe visually similar pathologies (z.B. Consolidation, Pneumonia and Infiltration)
- Nodules: Up to 50% are missed in X-Rays → Reports contain errors
- NLP method: Double error source: Extraction Accuracy + report errors

Label	PPV (visual)	PPV (text mining)
Consolidation	35%	97%
Cardiomegaly	80%	97%
Pneumothorax	60%	89%
Pneumonia	35%	89%
Fibrosis	24%	92%
No finding	60%	NA

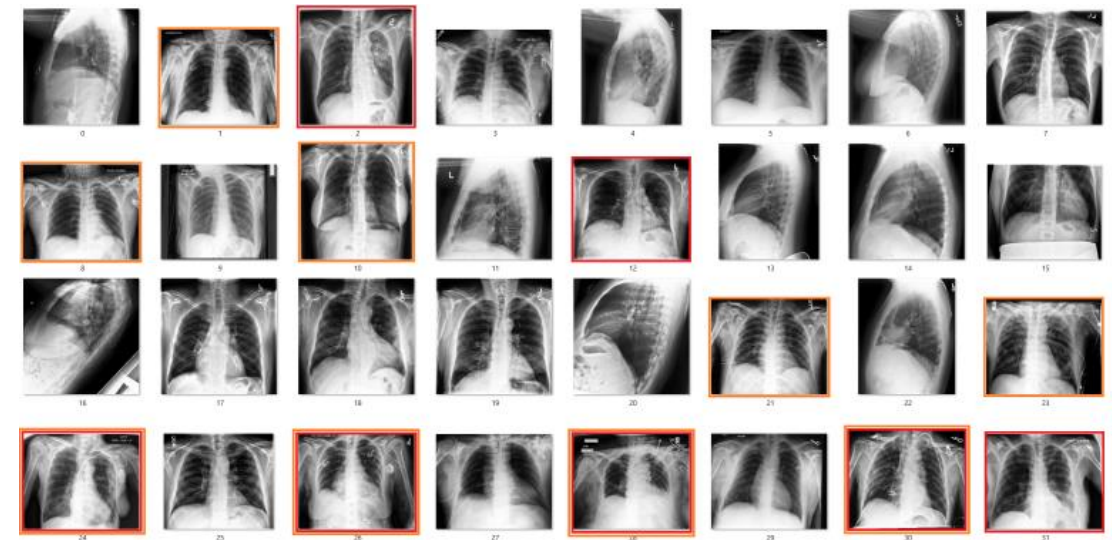
Radiologist's sample („visual“) for CXR14 to some labels with randomly chosen images



Critique of a radiologist: Details on CheXpert

<https://lukeoakdenrayner.wordpress.com/2019/02/25/half-a-million-x-rays-first-impressions-of-the-stanford-and-mit-chest-x-ray-datasets/>

- All in all: Step in the right direction and improves many problems of CXR14
- Label structure positive (see image)
But: „Terminal Nodes“ missing, which could summarize visually similar classes
→ Clinical Outcomes, e.g. treated vs. untreated Pneumothorax
- „No Finding“ sample (no quantitative analysis): Still contains some errors
- General critique on bad documentation of datasets in publications
- Reduced image quality can hide pathologies even for radiologists (but CheXpert offers unscaled data, 439GB)



CheXpert sample to „No Finding“
Red: Pathological, Orange: Support Devices



Data preparation

- Split into **Train/Validation/Test** (e.g.70/10/20)
- Important for X-Rays: Avoid patient overlap between sets!
 - Avoid/minimize correlations
 - Dividing into Train/Val/Test by patients is roughly enough
- Scaling: e.g. $[1024 \times 1024] \rightarrow [256 \times 256]$
- Pixel values
 - Scaling $[0, 255] \rightarrow [0, 1]$
 - Normalization
 - Own dataset (define preprocessing_function for Chest X-rays)
 - Datasets from transfer learning (preprocess_input for Keras packages)



Data preparation

Excerpt from ChestXray14 label file:

Image ID	Finding Labels	Follow-up #	Patient ID	Patient Age	Patient Gender	View Position
0000001_000.png	Cardiomegaly	0	1	58	M	PA
0000001_001.png	Cardiomegaly Emphysema	1	1	58	M	PA
0000002_000.png	No Finding	0	2	81	M	PA
0000003_000.png	Hernia	0	3	81	F	PA

→ Filter relevant information!



Data preparation

Extract labels and encode for the network → One-hot encoding

Image ID	Finding Labels	Follow-up #	Patient ID	Patient Age	Patient Gender	View Position
0000001_000.png	Cardiomegaly	0	1	58	M	PA
0000001_001.png	Cardiomegaly Emphysema	1	1	58	M	PA
0000002_000.png	No Finding	0	2	81	M	PA
0000003_000.png	Hernia	0	3	81	F	PA

Variations possible: filter by age, gender, etc.



Data augmentation

- Flip
 - Horizontal
- Crop
 - 1% - 5% abschneiden
- Rotate
 - $\pm 10^\circ$



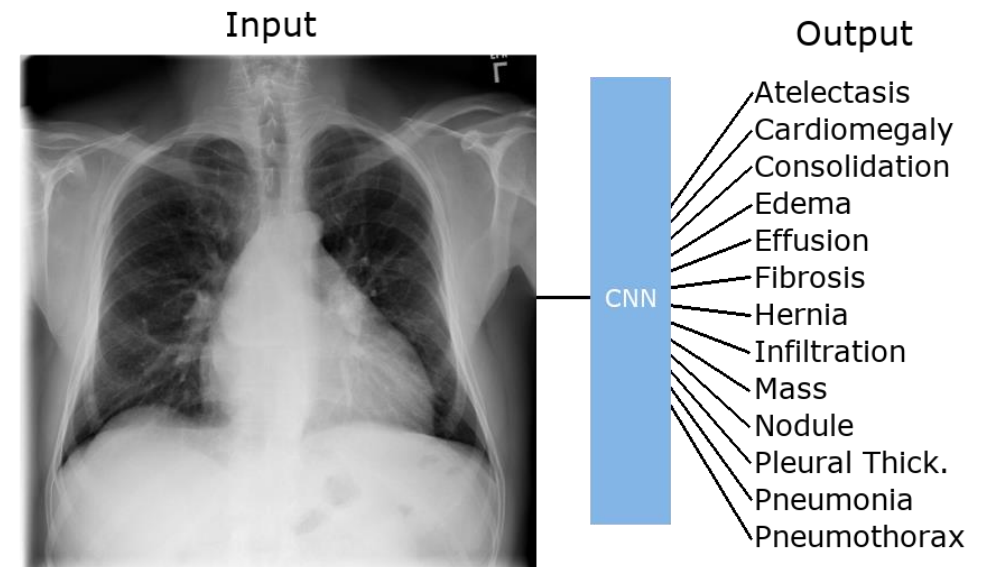
Data augmentation

- Important: **Label/contents** must stay the same!
- Which/how many augmentations are good?
 - Empirical, augmentations vary even between publications
- Which values for rotation, crop etc. are sufficient?
 - Also empirical, but smaller values are usually better
- Always augment in such a way, that a radiologist could diagnose!
 - E.g. no vertical flipping, don't crop on less than half etc.



Loss Function: Multi-Label

- Final layer: NO Softmax, but n Sigmoids
→ n independent binary classific.
- Label output via threshold
- E.g. pos. for $p > 0.5$



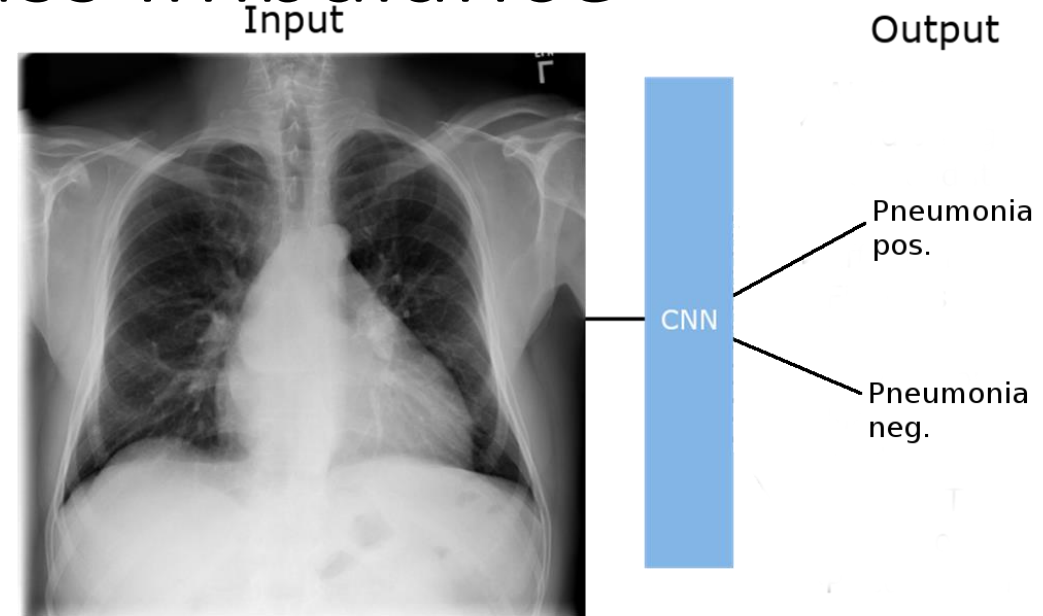
$$L(X, y) = \sum_{c=1}^{14} [y_c \log p(Y_c = 1|X) - (1 - y_c) \log p(Y_c = 0|X)]$$

Loss Function with Class Imbalance

- Example: Pneumonia on roughly 5% of all images
- Weighting by portion of total data

$$w_+ = \frac{|N|}{|P|+|N|}, w_- = \frac{|P|}{|P|+|N|}$$

$$\rightarrow w_+ = 0.95, w_- = 0.05$$



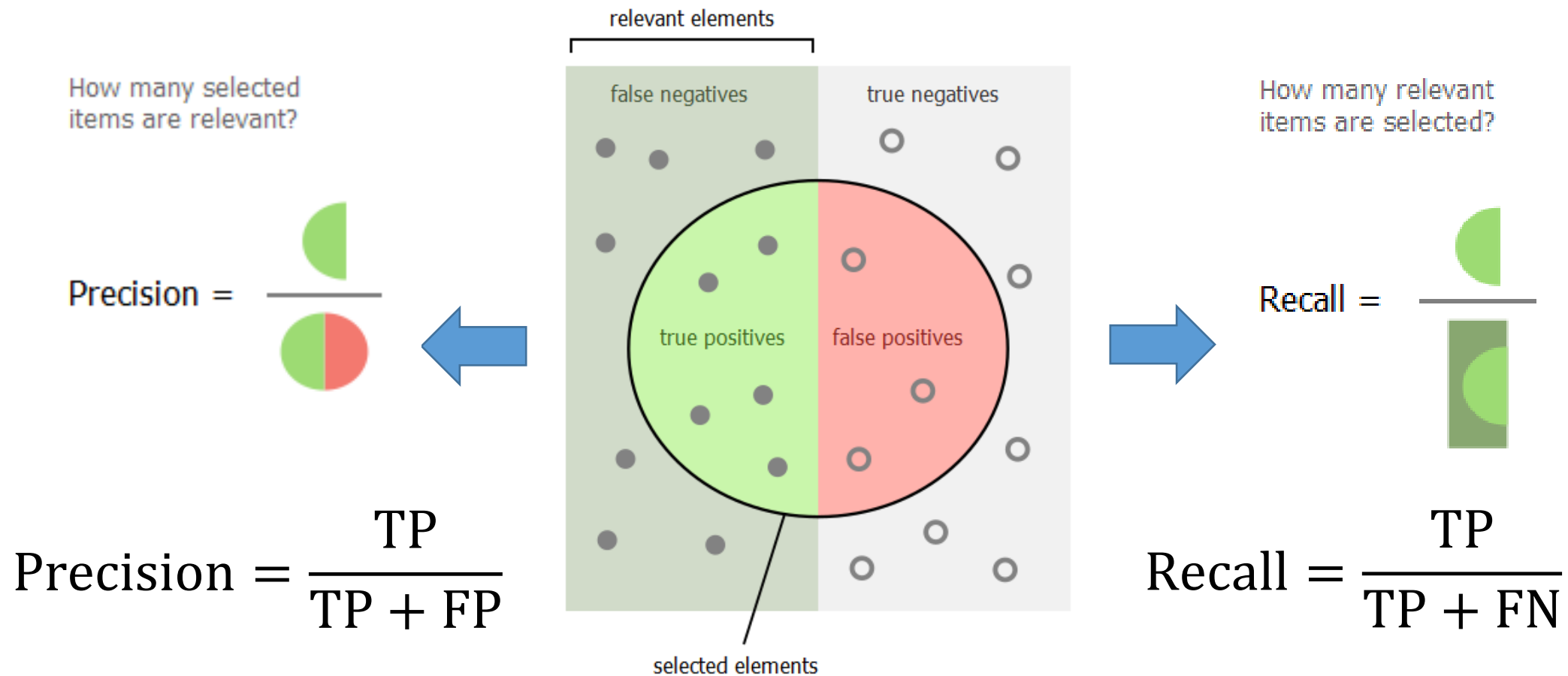
Binary Cross Entropy:

$$L(X, y) = -w_+ y \log p(Y = 1|X) - w_- (1 - y) \log p(Y = 0|X)$$

→ With many classes, analogously one weight for each class/sample



Evaluation (Recap)




Evaluation (Recap)

- A single number is often better to compare models
→ Harmonic mean of precision and recall: F1-Score

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$


$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$



Receiver Operating Characteristic Curve

- TPR/Recall vs. FPR/Fall-out

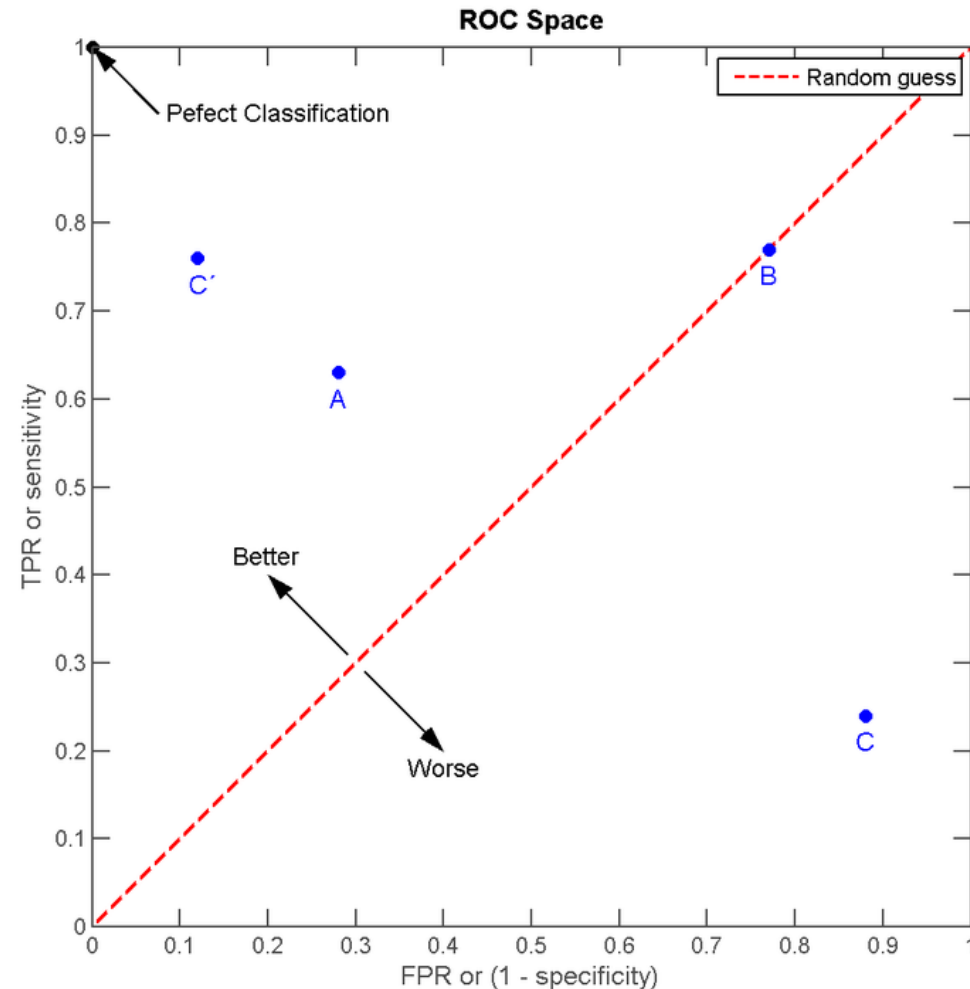
→ ROC space

- $TPR = \frac{TP}{TP+FN}$

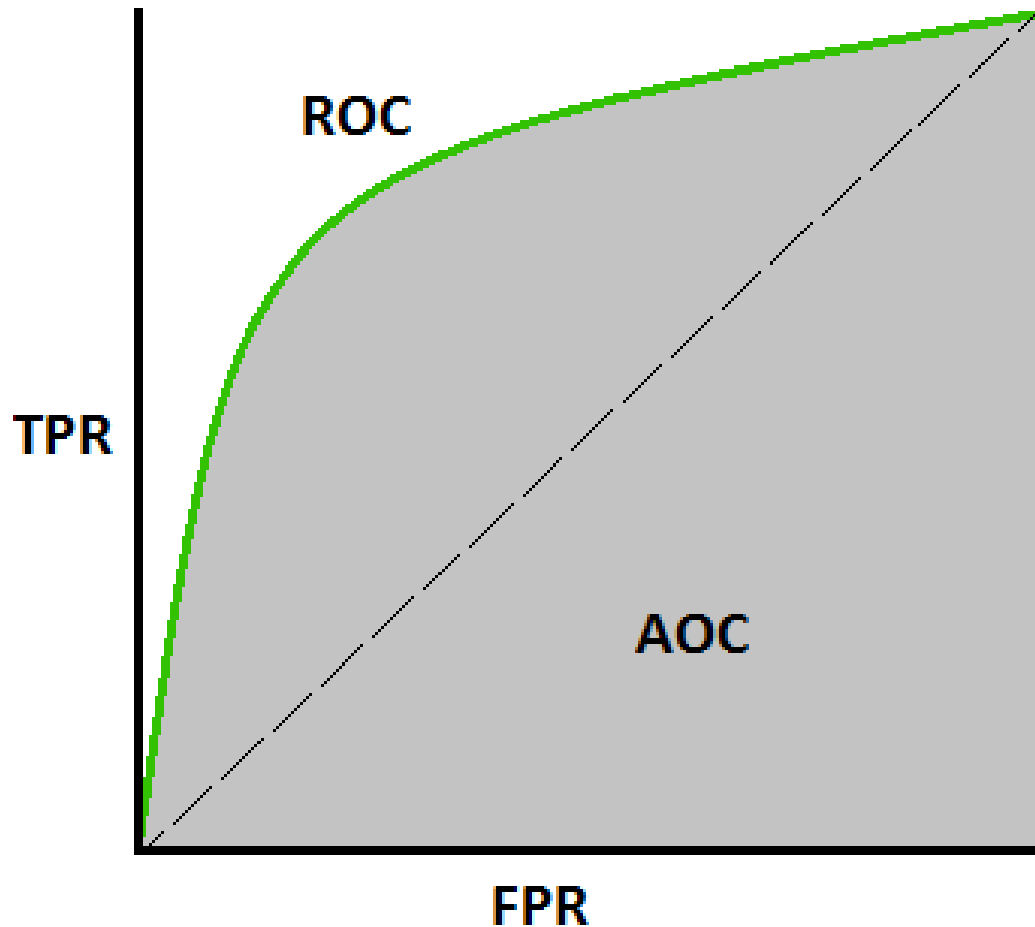
- $FPR = \frac{FP}{FP+TN}$

- For class imbalance:
Also PRC curve!

- Precision vs Recall



Area Under ROC



- Varying the threshold creates the curve
- Area under the curve: AUROC/AUC bzw. AUPRC
- $0 < AUC < 1$
- 0.5 = guessing
- Good metric for comparison
- One AUC per class



State of the Art

Architectures and results



State of the Art: Approaches

There is a steady stream of new publications on X-ray classification with the following commonalities:

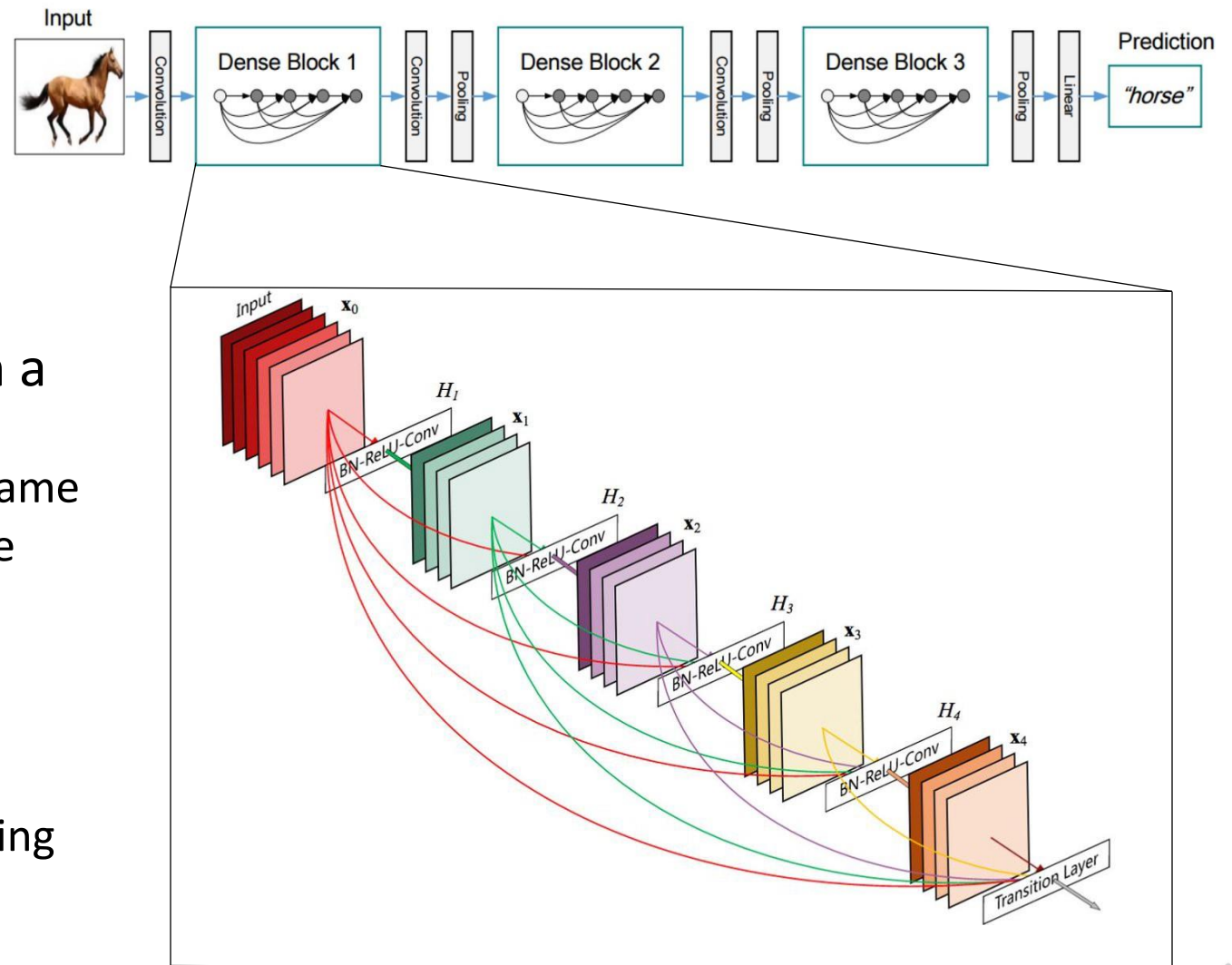
- Using the most recent established (pretrained) architectures
- Using new datasets
- Implementing new approaches in areas of
 - Image Preprocessing
 - Image Utilization (Information)
 - Relationships between labels/pathologies

In the following some examples



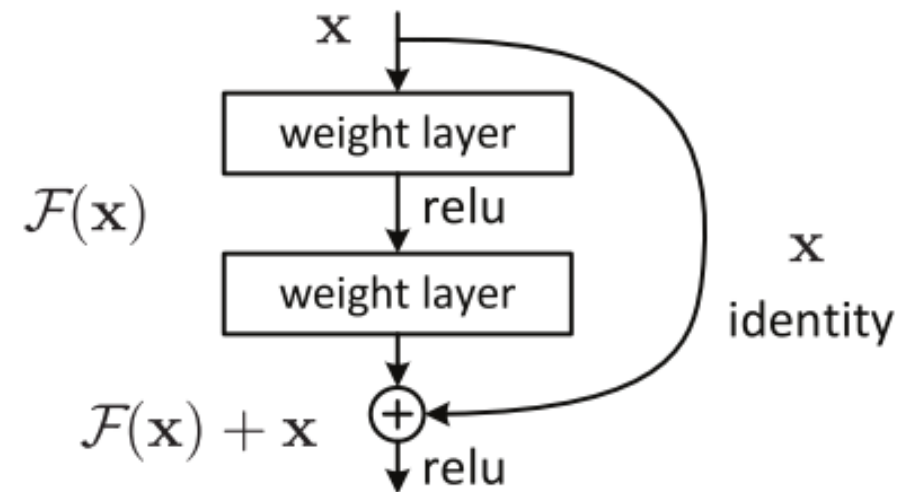
DenseNet-121

- Popular architecture
- Sequence of dense blocks and transition layers
- “Dense Information Flow” within a dense blocks
 - Feature Map resolution stays the same
 - All previous maps inside a block are concatenated to the current layer
→ Dense Connections
- E.g. Rajpurkar et al. 2018
 - Ensemble of 10 networks
 - Network trained → relabeling training data → Train new network



ResNet

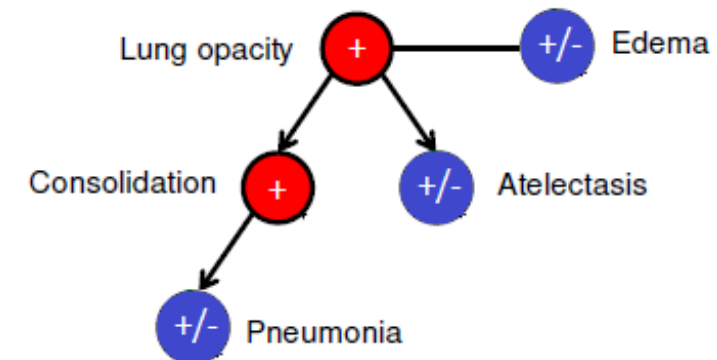
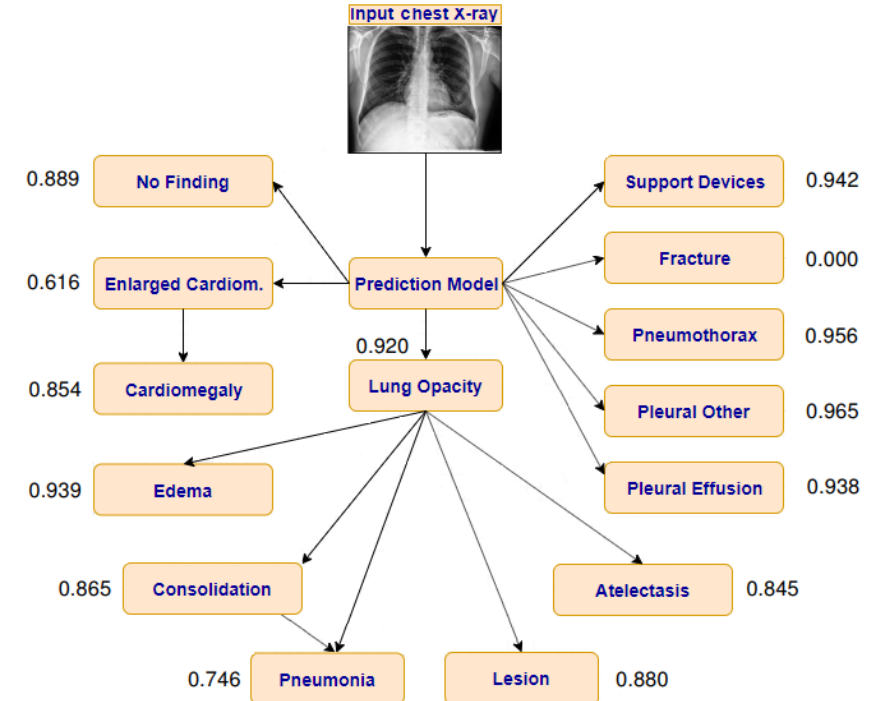
- Rakshit et al. 2019
 - ResNet-18
 - Comparatively small architecture
- Similar principle to DenseNet:
 - Residual instead of Dense Connections
 - Sum instead of concatenation
- ResNet-18 as „light-weight“ variant to many very deep networks with less parameters



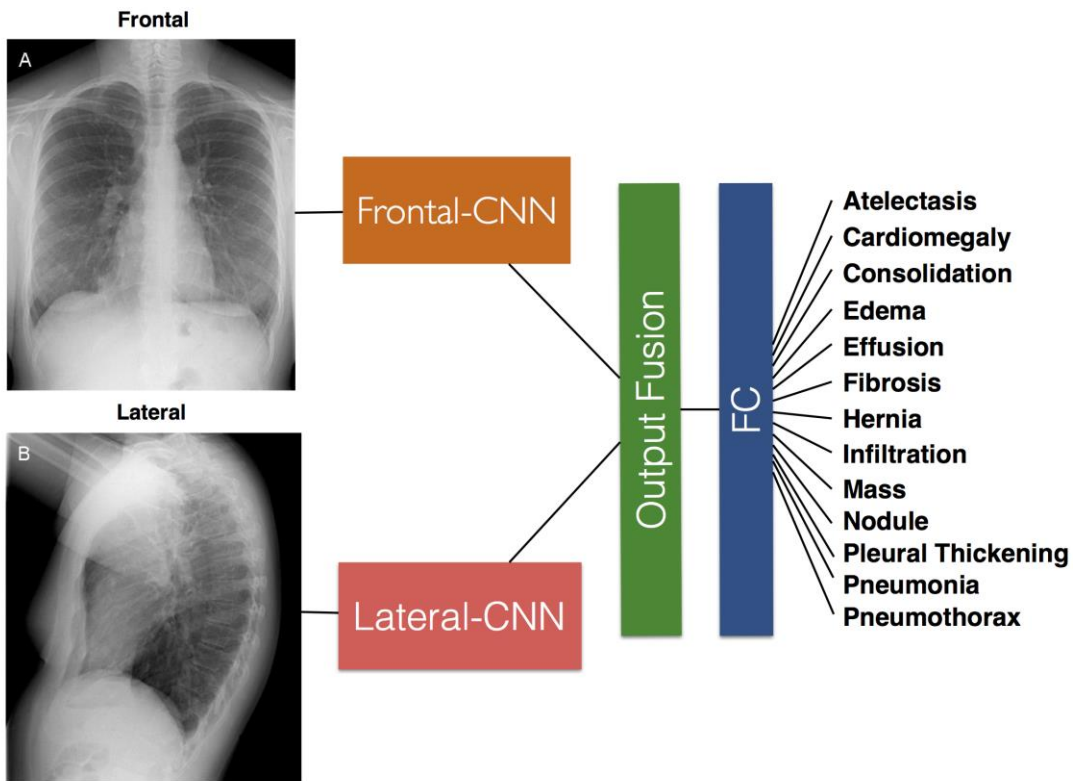
Residual Connections in ResNet

Label Dependencies

- Pham et al. 2019
- Network trained on hierarchy
 - Conditional Training
 - Only trained on images, where the parent is **positive**
 - Real probability is multiplicative (conditional probability)
- Additionally: Random label smoothing, to utilize „uncertain“ labels from CheXpert

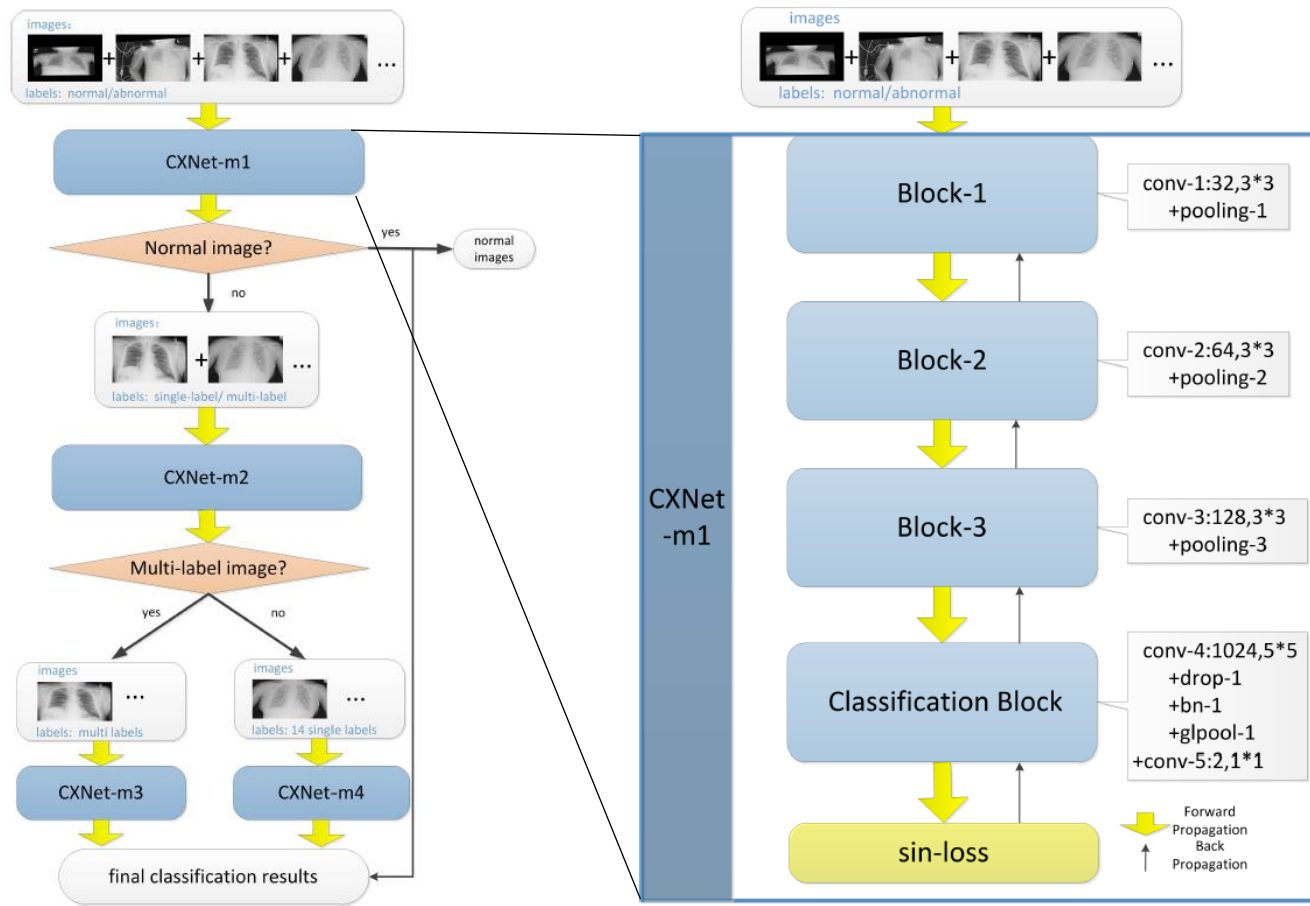


DualNet



- DualNet
- Rubin et al. 2018
- Two networks trained
 - One for frontal scans
 - One for lateral Scans
- 2x Densenet-121
- Pairs of images necessary

CXNet-m1

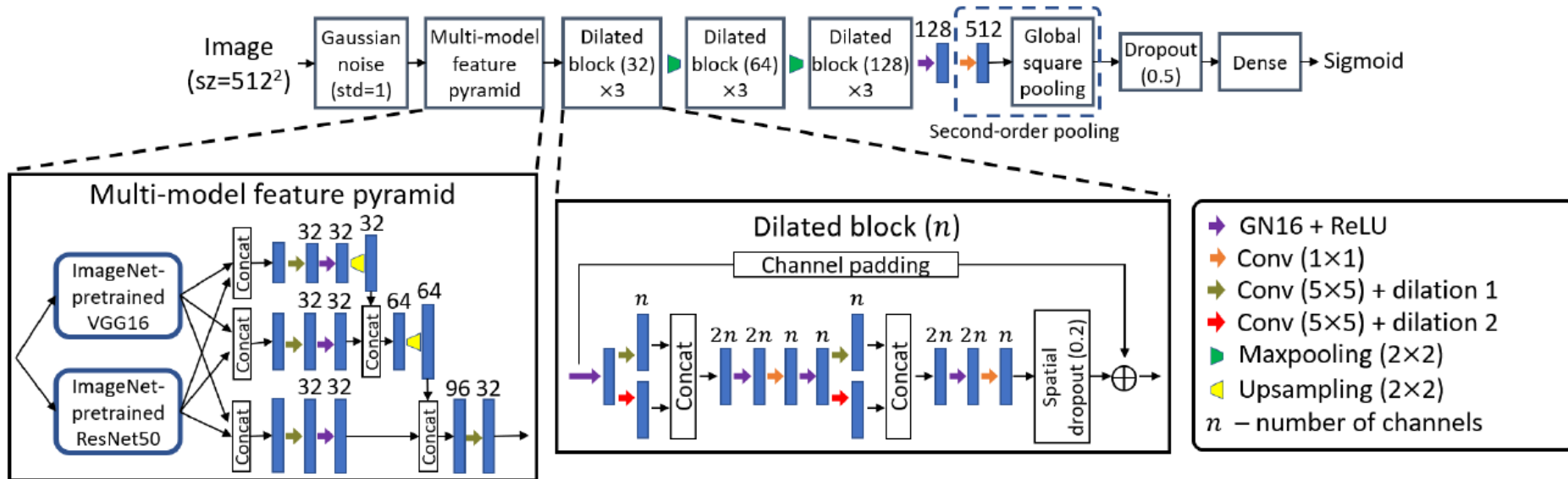


- Xu et al. 2018
- Custom CNN
 - Not pretrained
- Part of a bigger architecture consisting of several networks
- CXNet-m2 as next step



VGG16 + ResNet50 Pyramid Network

- VGG16 as „classical“ sequential network
- Pyramid Network with ResNet 50
- Wong et al. 2020



State of the Art

Author	Year	Labels	Training-Set	Method	Test-Set*	Metrics	Results
Rajpurkar et al.	2018	Multi	ChestXray14	DenseNet-121	Re-labeled CXR14	AUC	0.8492
Rakshit et al.	2019	Multi	ChestXray14	ResNet18		AUC	0.8494
Allaouzi et al.	2019	Multi	ChestXray14 CheXpert	DenseNet-121 + diff. separate classifiers		AUC	0.877 0.812
Pham et al.	2019	Multi	CheXpert	CNN (unspecified) + DenseNet-121	CheXpert Test-Set (Rajpurkar et al.)	AUC	0.940
Rubin et al.	2018	Multi	MIMIC-CXR	DualNet: DenseNet-121 x2		AUC	0.721
Rajpurkar et al.	2017	Pneumonia	ChestXray14	DenseNet-121	Re-labeled CXR14	F1	43.5%
Stephen et al.	2019	Pneumonia	Privat	Custom CNN	Siehe Training	Accuracy	93.012%
Xu et al.	2019	Normal/ pathological	ChestXray14	CXNet-m1 (Custom CNN)	CXR14 CXR14 + OpenI OpenI	F1	73.7% 89.1% 92.7%
Wong et al.	2020	Normal/ pathological	NLP re-labeled ChestXray14 + MIMIC	VGG16+ResNet50 pyramid network	several Test-Sets: NLP, double/triple consensus	AUC: ROC AUC: PR	0.821, 0.920, 0.961 0.811, 0.924, 0.967

*if different from training



Comparison with radiologists: Re-labeled CXR14

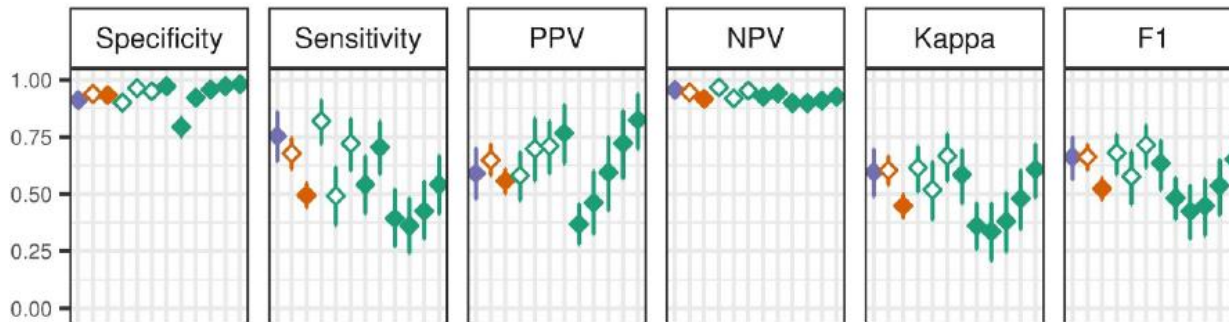
Pathology	Radiologists (95% CI)	Algorithm (95% CI)	Algorithm – Radiologists Difference (99.6% CI) ^a	Advantage
Atelectasis	0.808 (0.777 to 0.838)	0.862 (0.825 to 0.895)	0.053 (0.003 to 0.101)	Algorithm
Cardiomegaly	0.888 (0.863 to 0.910)	0.831 (0.790 to 0.870)	−0.057 (−0.113 to −0.007)	Radiologists
Consolidation	0.841 (0.815 to 0.870)	0.893 (0.859 to 0.924)	0.052 (−0.001 to 0.101)	No difference
Edema	0.910 (0.886 to 0.930)	0.924 (0.886 to 0.955)	0.015 (−0.038 to 0.060)	No difference
Effusion	0.900 (0.876 to 0.921)	0.901 (0.868 to 0.930)	0.000 (−0.042 to 0.040)	No difference
Emphysema	0.911 (0.866 to 0.947)	0.704 (0.567 to 0.833)	−0.208 (−0.508 to −0.003)	Radiologists
Fibrosis	0.897 (0.840 to 0.936)	0.806 (0.719 to 0.884)	−0.091 (−0.198 to 0.016)	No difference
Hernia	0.985 (0.974 to 0.991)	0.851 (0.785 to 0.909)	−0.133 (−0.236 to −0.055)	Radiologists
Infiltration	0.734 (0.688 to 0.779)	0.721 (0.651 to 0.786)	−0.013 (−0.107 to 0.067)	No difference
Mass	0.886 (0.856 to 0.913)	0.909 (0.864 to 0.948)	0.024 (−0.041 to 0.080)	No difference
Nodule	0.899 (0.869 to 0.924)	0.894 (0.853 to 0.930)	−0.005 (−0.058 to 0.044)	No difference
Pleural thickening	0.779 (0.740 to 0.809)	0.798 (0.744 to 0.849)	0.019 (−0.056 to 0.094)	No difference
Pneumonia	0.823 (0.779 to 0.856)	0.851 (0.781 to 0.911)	0.028 (−0.087 to 0.125)	No difference
Pneumothorax	0.940 (0.912 to 0.962)	0.944 (0.915 to 0.969)	0.004 (−0.040 to 0.051)	No difference

CheXNeXt (Rajpurkar et al. 2018) vs. 4 Practicing radiologists on 420 images from ChestXray-14 relabeled by (different) radiologists

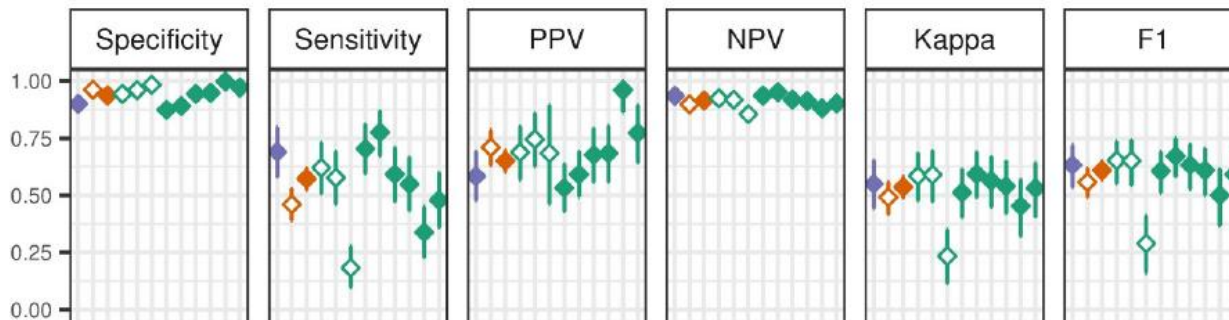


Which metric is best?

a Mass



b Nodule



- Algorithm
- Resident radiologists
- Board-certified radiologists
- Resident1
- Resident2
- Resident3
- BC1
- BC2
- BC3
- BC4
- BC5
- BC6

→ No clear answer



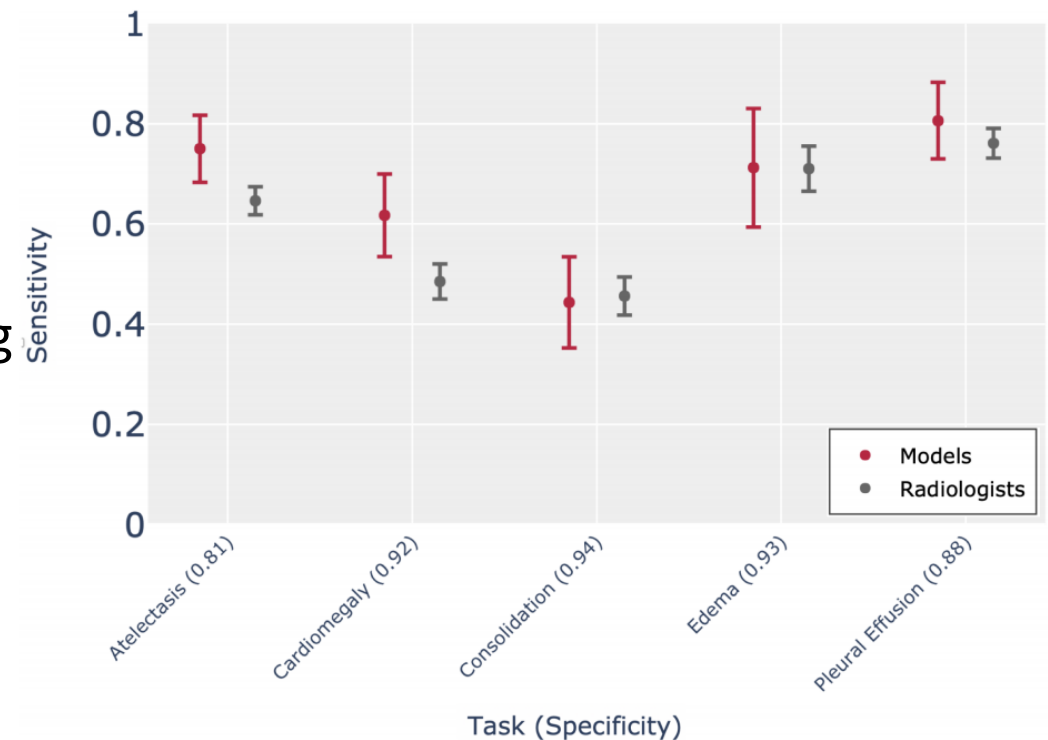
Comments

- Publications only test within their dataset
 - Or with parts of external datasets
- NO „practical test“, so no test on real data
 - Re-labeling of CXR14 from practicing radiologists (CheXNeXt) is the closest approach
- Also no real comparisons to actual practice
 - Diagnosis by independent radiologists (CheXNeXt) is again the closest



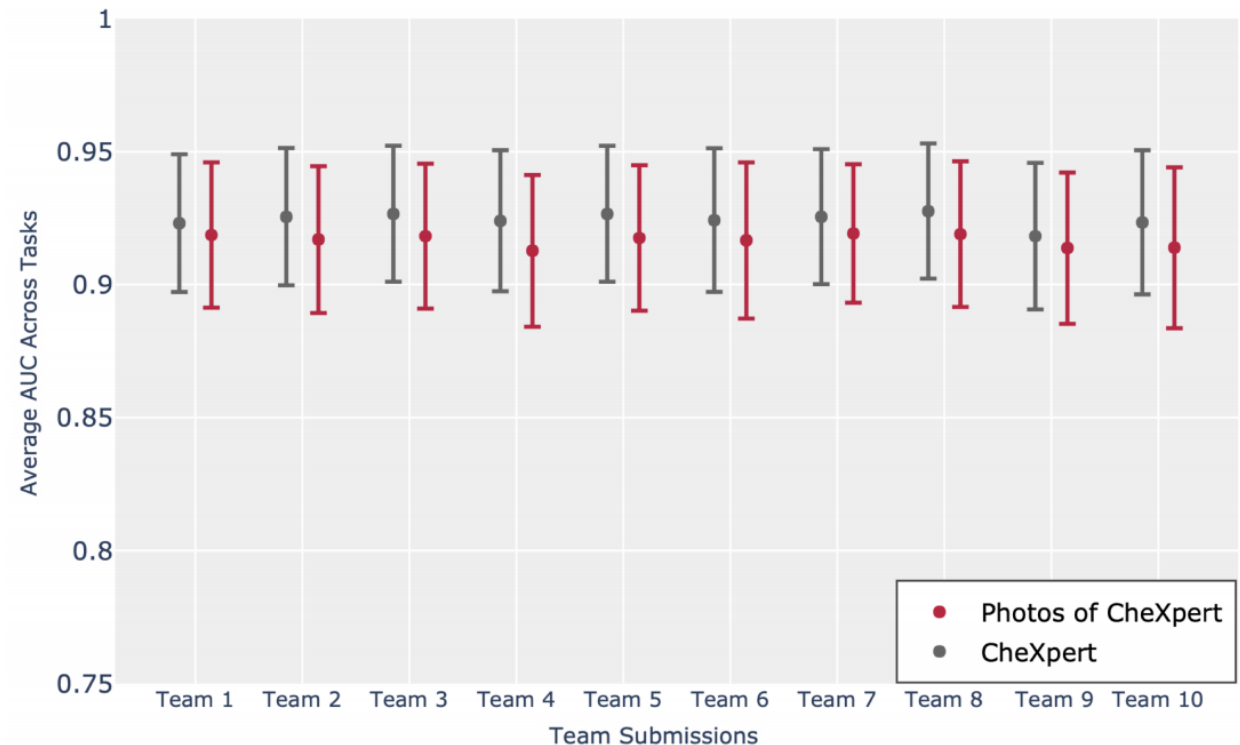
CheXpedition

- Evaluation of the 10 best teams from the CheXPert leaderboard
 - All models are ensembles with 8 to 32 networks, often with DenseNet
- Hidden Test-Set
 - Scores are determined internally after sending in your model
- Analysis of three practical cases
 - Tuberculosis classification
 - Application on Smartphone photos
 - Evaluation on external test set (right)
Same as in Rajpurkar et al. 2018



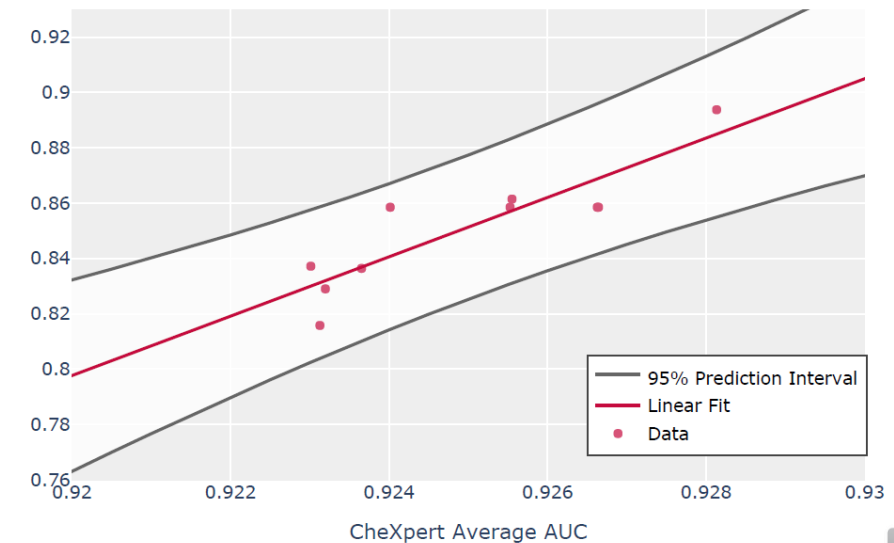
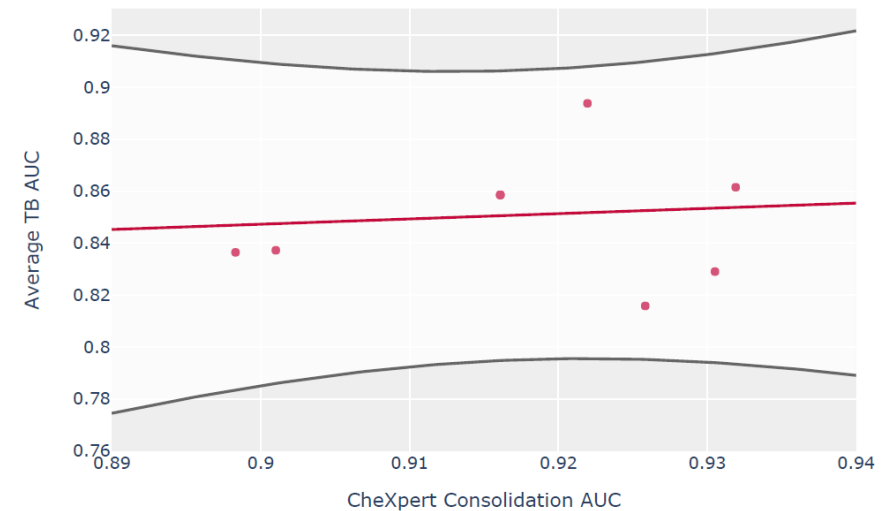
CheXpedition: Smartphones

- Photos taken and labeled in clinical environment
- 500 photos
- Almost no loss in accuracy for all teams
- 0.924 vs. 0.916 mean AUC
- Usage: Access for doctors via messaging services

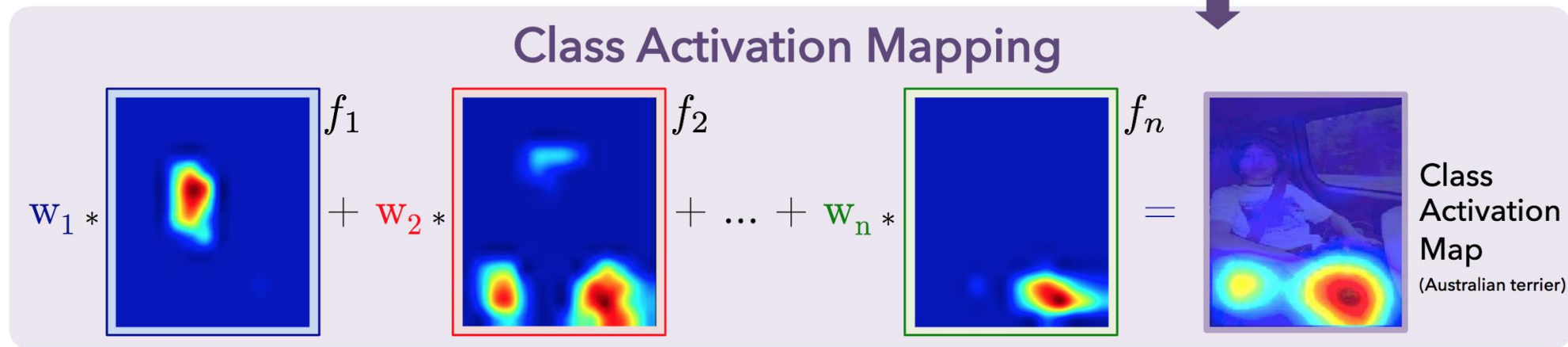
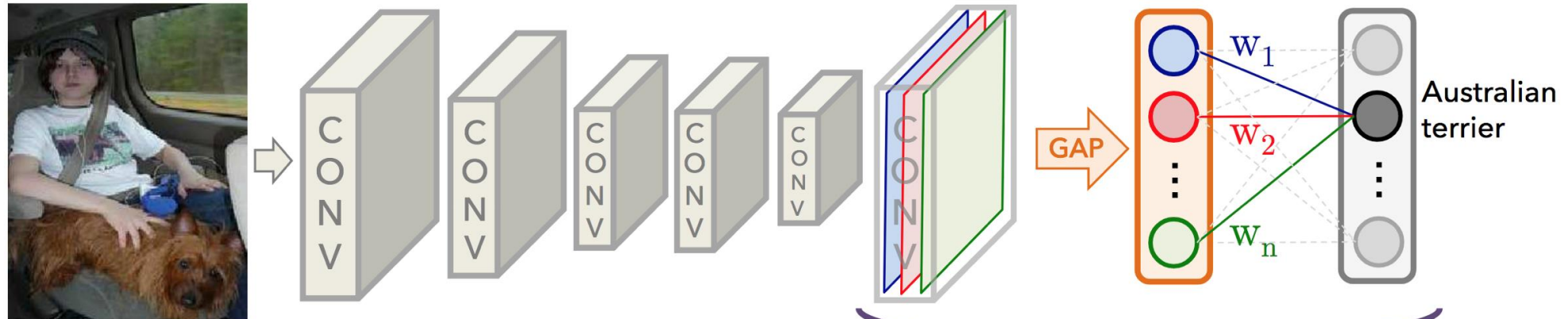


CheXpedition: Tuberkulosis

- Using existing datasets on tuberculosis (Shenzhen, Montgomery)
- Consolidation-% as label for tuberculosis
- TB-AUC depends on multi-task performance rather than consolidation performance



Insert: Class Activation Mapping



Insert : Class Activation Mapping

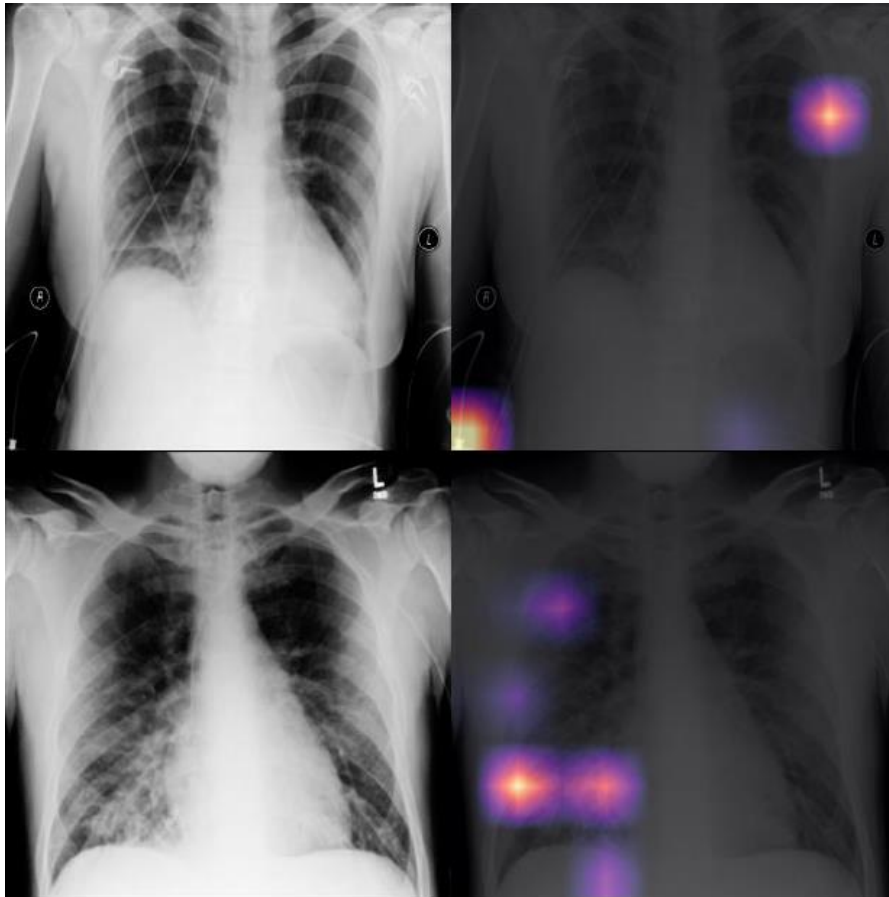
- Only with Global Average Pooling
- Extracting the Feature Maps f_k from the last conv layer
- Summarize the maps with weighting by $w_{c,k}$ from feature map k on class c :

$$M_c = \sum_k w_{c,k} \cdot f_k$$

- Heatmap M_c upscaled and overlain on the input
- Zhou et al. 2016 „Learning Deep Features for Discriminative Localization“



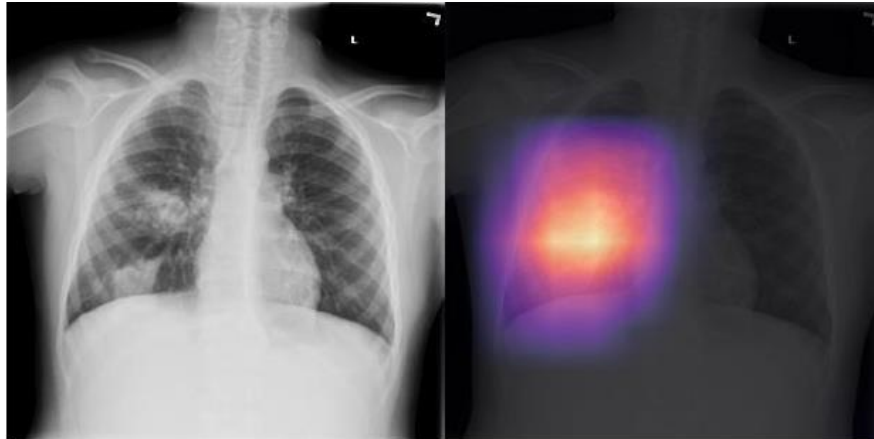
CheXpedition: Errors (Consolidation)



- False Negatives
 - Consolidation not localized
 - Often visually similar feature located
 - 36 errors, 44.44%
- Too uncertain in detection
 - 29 errors, 35.80%



CheXpedition: Errors (Consolidation)



- False Positives
 - „Mimicking“ Feature (visually similar)
 - Often when other similar pathologies like Edema are present
 - 13 errors, 16.05%



- Non-mimicking feature
- Basically completely wrong detection
- 3 errors, 3.70%



State of the Art References

Rajpurkar, Pranav et al. 2018 „Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists.” In: *PLoS medicine* vol. 15,11 e1002686.

Rakshit S., Saha I., Wlasnowolski M., Maulik U., Plewczynski D. 2019 „Deep Learning for Detection and Localization of Thoracic Diseases Using Chest X-Ray Imagery” In: Rutkowski L., Scherer R., Korytkowski M., Pedrycz W., Tadeusiewicz R., Zurada J. (eds) *Artificial Intelligence and Soft Computing. ICAISC 2019. Lecture Notes in Computer Science*, vol 11509. Springer, Cham

I. Allaouzi and M. Ben Ahmed 2019 „ A Novel Approach for Multi-Label Chest X-Ray Classification of Common Thorax Diseases” In: *IEEE Access*, vol. 7, pp. 64279-64288

Pham, Le, Tran, Ngo and Nguyen 2019 „Interpreting chest X-rays via CNNs that exploit disease dependencies and uncertainty labels” In: arXiv e-prints

Rubin, Sanghavi et al. 2018 „Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks” In: arXiv e-prints

Rajpurkar et al. 2017 „CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning” In: arXiv e-prints

Stephen et al. 2019 „An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare” In: *Journal of Healthcare Engineering*, vol. 2019, 4180949

S. Xu, H. Wu and R. Bie, 2019 "CXNet-m1: Anomaly Detection on Chest X-Rays With Image-Based Deep Learning," in *IEEE Access*, vol. 7, pp. 4466-4477

K. C. L. Wong *et al.*, "A Robust Network Architecture to Detect Normal Chest X-Ray Radiographs," *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, Iowa City, IA, USA, 2020, pp. 1851-1855

Rajpurkar et al. 2020 „CheXpedition: Investigating Generalization Challenges for Translation of Chest X-Ray Algorithms to the Clinical Setting” In: arXiv e-prints



Applying the predictions

How can you use network predictions effectively in practise?

- Complete AI prediction
 - Scepticism, accuracies are not sufficient, even if radiologists are also prone to make errors
- Partial prediction
 - Example: Perfekte Precision for Normal/Abnormal → Reduced workload since most scanned people are healthy
- Supporting prediction
 - Network outputs suggestion for the radiologists
 - Implement into documentation software for doctors



Application in documentation

- Idea: X-ray uploaded
 → CNN returns predictions
 → (check-)boxes are filled out
 → Radiologist only has to correct
 → Time save



Aufnahme	<input type="radio"/> Thorax p.a.	<input type="radio"/> Thorax 2 Ebenen	<input type="radio"/> Thorax im Liegen	
Voraufnahme	<input type="radio"/> von heute mit VA gestern	<input type="radio"/> von heute mit VA variabel Voraufnahme.....	<input type="radio"/> von heute ohne VA	<input type="radio"/> von variabel mit VA variabel von..... Voraufnahme.....
Instrum.	<input type="checkbox"/> Keine	<input type="checkbox"/> ZVK von (re. / li.) (jugulär / subclaviculär) (korrekte Lage / Fehllage / entfernt)	<input type="checkbox"/> 2. ZVK von (re. / li.) (jugulär / subclaviculär) (korrekte Lage / Fehllage / entfernt)	<input type="checkbox"/> Shaldon-Katheter von (re. / li.) (jugulär / subclaviculär) (korrekte Lage / Fehllage / entfernt)
	<input type="checkbox"/> Pleuradrainage von (re. / li.) (korrekte Lage / Fehllage / entfernt)	<input type="checkbox"/> 2. Pleuradrainage von (re. / li.) (korrekte Lage / Fehllage / entfernt)	<input type="checkbox"/> 3. Pleuradrainage von (re. / li.) (korrekte Lage / Fehllage / entfernt)	<input type="checkbox"/> Mediastinaldrainage (korrekte Lage / Fehllage / entfernt) Drainag
	<input type="checkbox"/> ETT (korrekte Lage / Fehllage / entfernt) Abstand cm	<input type="checkbox"/> Port von (re. / li.) (brachial / pectoral) (korrekte Lage / Fehllage)	<input type="checkbox"/> Schrittmacher/ICD (SM / ICD) (korrekte Lage / Fehllage / Sondenbruch) von (re. / li.) (1 / 2 / 3 / 4) konnektierte Sondenkabel	<input type="checkbox"/> OP-Status mit [Herzklappen-OP / ACB bzw. ACVB]
PE li.	<input type="radio"/> kein PE	<input type="radio"/> gering (Neu / Idem / Progred. / Regred.)	<input type="radio"/> mäßig (Neu / Idem / Progred. / Regred.)	<input type="radio"/> deutlich (Neu / Idem / Progred. / Regred.)
PE re.	<input type="radio"/> gering (Neu / Idem / Progred. / Regred.)	<input type="radio"/> mäßig (Neu / Idem / Progred. / Regred.)	<input type="radio"/> deutlich (Neu / Idem / Progred. / Regred.)	
PE bds.	<input type="radio"/> gering (Neu / Idem / Progred. / Regred.)	<input type="radio"/> mäßig (Neu / Idem / Progred. / Regred.)	<input type="radio"/> deutlich (Neu / Idem / Progred. / Regred.)	
Pleura	<input type="radio"/> kein Pneumothorax	<input type="radio"/> Pneumothorax (re. / li. / bds.) (apikal / lateral / basal / ventral / mantelförmig) mit mm pleuraler Dehiszenz	<input type="radio"/> Seropneumothorax (re. / li. / bds.) mit mm pleuraler Dehiszenz Mediastinalshift (gering / mäßig / deutlich)	<input type="radio"/> Spannungspneumothorax (re. / li.) mit mm pleuraler Dehiszenz Mediastinalshift (gering / mäßig / deutlich)
Mediastinum	<input type="checkbox"/> normal	<input type="checkbox"/> verlagert nach (re. / li.)	<input type="checkbox"/> verbreitert Im (oberen / mittleren / unteren) Anteil	

Example documentation software



Image Processing in Gastroenterology

Endoscopy

- Computer Vision is tested in many medical areas to support gastroenterologists
- The gastroenterologist uses the endoscope as their tool, which generates image and video data, which can then be processed by the computer scientist and their algorithms



Examples of image processing in gastroenterology

- Detection of inflammation:
 - Bossuyt, Peter, et al. "Automatic, computer-aided determination of endoscopic and histological inflammation in patients with mild to moderate ulcerative colitis based on red density." *Gut* (2020).
- 3D modelling of the intestine
 - Tavanapong, Wallapak, et al. "Reconstruction of a 3D Virtual Colon Structure and Camera Motion for Screening Colonoscopy." *Medical Research Archives* 5.6 (2017).
- Recognition, detection and segmentation of pathologies during endoscopy:
 - Wang, Dechun, et al. "AFP-Net: Realtime Anchor-Free Polyp Detection in Colonoscopy." *arXiv preprint arXiv:1909.02477* (2019).
 - Challenges: MICCAI2015(<https://polyp.grand-challenge.org/Home/>), EDD(<https://edd2020.grand-challenge.org/>), EAD(<https://ead2020.grand-challenge.org/>)



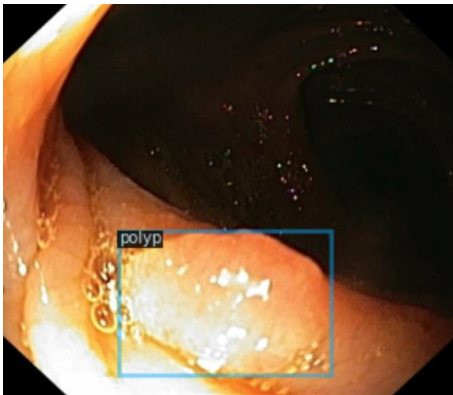
Recognition and localization of pathologies



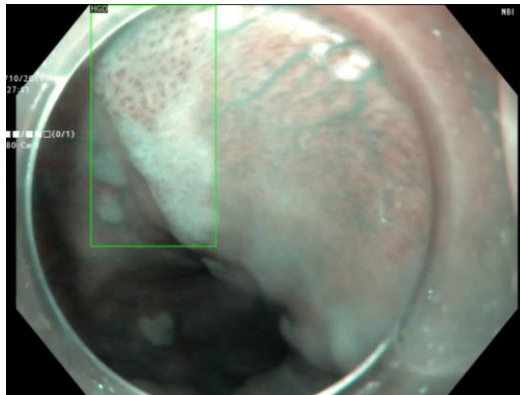
Recognition and localization of pathologies

- Example: 4 different diseases during an endoscopy: Polyps, HGD (high grade dysplasia), Cancer, BE (Barrett Esophagus)

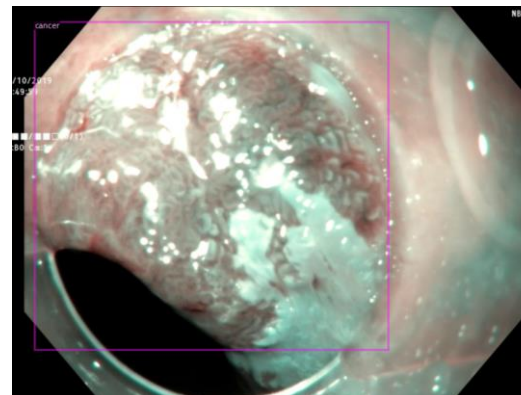
Polyp



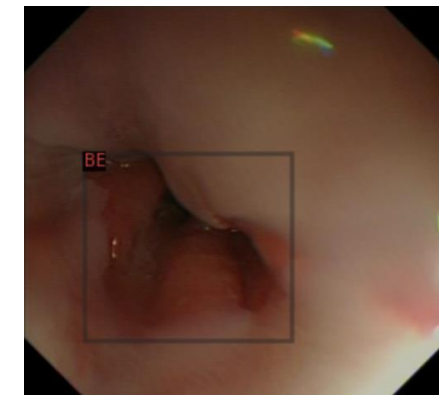
HGD



Cancer



BE



Classes are from the EDD2020 challenge. The class cancer is used like this in the challenge. However, it is not clearly defined as it can appear in several kinds (e.g. HGD/Polyp).

Pathology Detection Pipeline

- Generate a Ground Truth
- Pre Processing
- Choosing the network for training
- Post Processing
- Evaluation

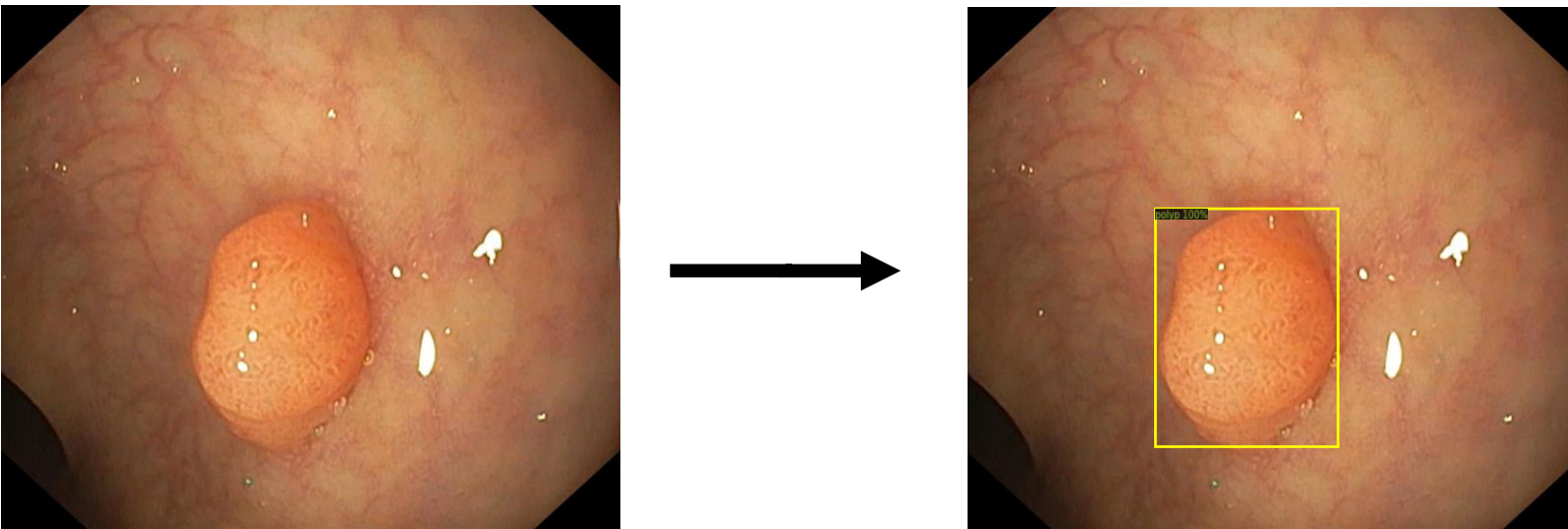
In the following shown on the example of polyp detection

What are polyps?

- Polyps are abnormal proliferations, which often look like small, flat bulges or tiny mushroom-like stems. Due to their abnormal cell growth they can eventually be malignant or cancer-like
- Together with the university clinic of Würzburg we are researching computer-aided polyp detection. Here, the endoscopist is supported during polyp detection and is shown in real time on their video feed, where polyps are located.

Problem Definition: Classification on freeze images

- Using the input image, positions of a certain class of objects can be predicted within an image
- Positions are represented by bounding boxes



Generating the Ground Truth

- Integrating openly accessible datasets:
 - ETIS-Larib Polyp database (196 polyp images)
 - CVC-ClinicDB (612 polyp images)
 - Gastrointestinal Image Analysis (GIANA) challenge (412 polyp images).
 - Kvasir-SEG dataset (1000 polyp images)

Generating the Ground Truth

- Generate and edit using an annotation tool
- Annotated by gastroenterologists:



Pre Processing

Example data augmentation in endoscopy:



Normalization:

- Scale pixel values from $[0, 255]$ to $[0, 1]$
- Normalization on standard values (\pm standard deviation) on
 - own dataset (`datagen.fit(x)`, possible with the small polyp datasets)
 - or: on datasets from pretraining

Post Processing

- Known post processing from object detection
 - Thresholding
 - Non-Maximum-Suppression
- Domain-specific post processing:
 - Utilize knowledge specific to the task
 - Example: BE and polyps only appear mutually in different areas of the intestine
 - If both are predicted, only choose the more likely one

Evaluation

As known from „regular“ object detection:

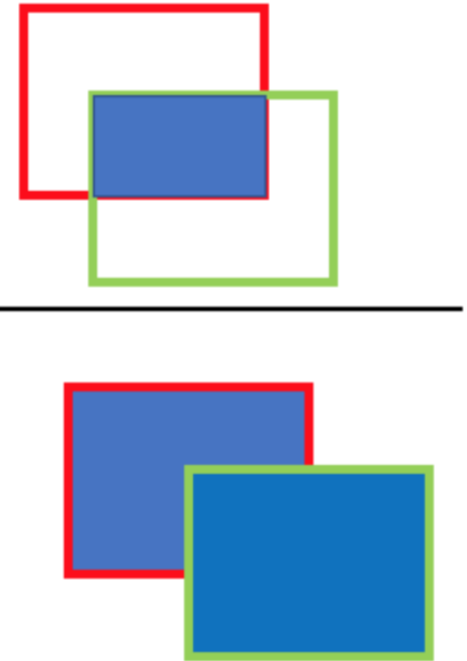
- First Matching of Predictions to a class
- Matching with Intersection over Union (IoU)

$$\text{IoU} \triangleq \frac{A \cap B}{A \cup B}$$

typically, $\text{IoU} \geq 0.5 \rightarrow \text{Match}$

- Then using the usual classification metrics
 - Accuracy
 - F1
 - Precision
 - ...

Intersection over Union=



Green is the Ground Truth Box, red the prediction



Evaluation

- Benchmark datasets are often used for evaluation
- In the case of polyps there are mainly two datasets:
 - ETIS-Larib Polyp database (196 polyp images)
 - CVC-ClinicDB (612 polyp images)
- Introduced in the first international Polyp detection challenge 2015
- Hence, in areas of polyp detection and many other object detection areas, a steady competition is established, where scientists from all over the world try to improve on these datasets

Architectures

The familiar known architectures from object detection

- One-Stage
 - SSD
 - YOLO
 - CenterNet
- Two-Stage
 - Faster R-CNN
 - Mask R-CNN



State of the art polyp detection

Author	Year	Methods	Training Data	Evaluation Metrics	Test Data	Results	Research contribution
(Tajbakhsh et al.)	2016	CNNs	CVC-Clinic VideoDB	F1-Score	EITS-LARIB	77,4 %	- Training of various CNN architectures for polyp detection with a completely new architecture and fine tuning or architectures pre-trained on ImageNet.
(Yuan et al.)	2017	AlexNet (CNN)	CVC-Clinic EITS-LARIB VideoDB split2	Accuracy	CVC-Clinic VideoDB	91.5 %	- Training of a two-step approach with feature extraction first and classification afterwards
(Shin et al.)	2018	Mask RCNN (Segmentation CNN)	CVC-Clinic EITS-LARIB VideoDB split2	F1-Score	CVC-Clinic VideoDB	81,7%	- Detection and segmentation of polyps with a Mask-RCNN that was pretrained on the MS-COCO dataset
(Mo et al.)	2018	Faster RCNN	CVC-Clinic EITS-LARIB VideoDB split2	F1-Score	CVC-Clinic VideoDB	91,7%	- Improving region based CNNs for detecting polyps in images using Faster R-CNN.
(Tian et al.)	2019	Resnet-50	CVC-Clinic EITS-LARIB VideoDB split2 GIANA	Accuracy	CVC-Clinic VideoDB	91,9%	- One-step approach for polyp classification using the ResNet-50 architecture - Extending the CNN to a five class polyp classification
(Zhang et al.)	2019	Single shot detection (SSD)	CVC-Clinic EITS-LARIB VideoDB split2 GIANA	Precision	CVC-Clinic VideoDB	88,5%	- First Single-Shot approach on the CVC-Clinic VideoDB dataset - Application and adaptation of the YOLOv2 real time object detection algorithm
(Liu et al.)	2019	Single shot detection (SSD)	CVC-Clinic VideoDB split2 GIANA KVASIR	F1-Score	EITS-LARIB	83,4%	- Implementation of their own SSD algorithm for polyp detection. - State of the Art results on the EITS-LARIB dataset
(Wang et al.)	2019	SSD Anchor free	CVC-Clinic EITS-LARIB VideoDB split2 GIANA KVASIR	Accuracy	CVC-Clinic VideoDB	96,44%	- First anchor-free real time detection of polyps in colonoscopy. - State of the Art results on the CVC-Clinic VideoDB split1 dataset



State of the art polyp detection

Literatur

- Liu, M., Jiang, J., and Wang, Z. 2019. "Colonic Polyp Detection in Endoscopic Videos with Single Shot Detection Based Deep Convolutional Neural Network," IEEE Access (7), pp. 75058-75066.
- Mo, X., Tao, K., Wang, Q., and Wang, G. 2018. "An Efficient Approach for Polyps Detection in Endoscopic Videos Based on Faster R-Cnn," 2018 24th International Conference on Pattern Recognition (ICPR): IEEE, pp. 3929-3934.
- Shin, Y., Qadir, H. A., Aabakken, L., Bergsland, J., and Balasingham, I. 2018. "Automatic Colon Polyp Detection Using Region Based Deep Cnn and Post Learning Approaches," IEEE Access (6), pp. 40950-40962.
- Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. 2016. "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," IEEE transactions on medical imaging (35:5), pp. 1299-1312.
- Tian, Y., Pu, L. Z., Singh, R., Burt, A. D., and Carneiro, G. 2019. "One-Stage Five-Class Polyp Detection and Classification," 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019): IEEE, pp. 70-73.
- Wang, D., Zhang, N., Sun, X., Zhang, P., Zhang, C., Cao, Y., and Liu, B. 2019. "Afp-Net: Realtime Anchor-Free Polyp Detection in Colonoscopy," arXiv preprint arXiv:1909.02477).
- Yuan, Z., IzadyYazdanabadi, M., Mokkaapati, D., Panvalkar, R., Shin, J. Y., Tajbakhsh, N., Gurudu, S., and Liang, J. 2017. "Automatic Polyp Detection in Colonoscopy Videos," Medical Imaging 2017: Image Processing: International Society for Optics and Photonics, p. 101332K.
- Zhang, X., Chen, F., Yu, T., An, J., Huang, Z., Liu, J., Hu, W., Wang, L., Duan, H., and Si, J. 2019. "Real-Time Gastric Polyp Detection Using Convolutional Neural Networks," PloS one (14:3), p. e0214133.



How many data do you need for a robust detection system?

- Polyp detection through Bounding Boxes
(Test on CVC-Clinic-DB dataset; 612 images):

Number of training images	F1-score
50	49 %
100	58 %
500	70 %
1000	78%
2000	85%

Training settings:

Algorithm: YOLOv3

Batch size: 4

Learning rate: 0.001

Epochs: 50

Optimizer: Adam

Practice: Real time object detection in videos



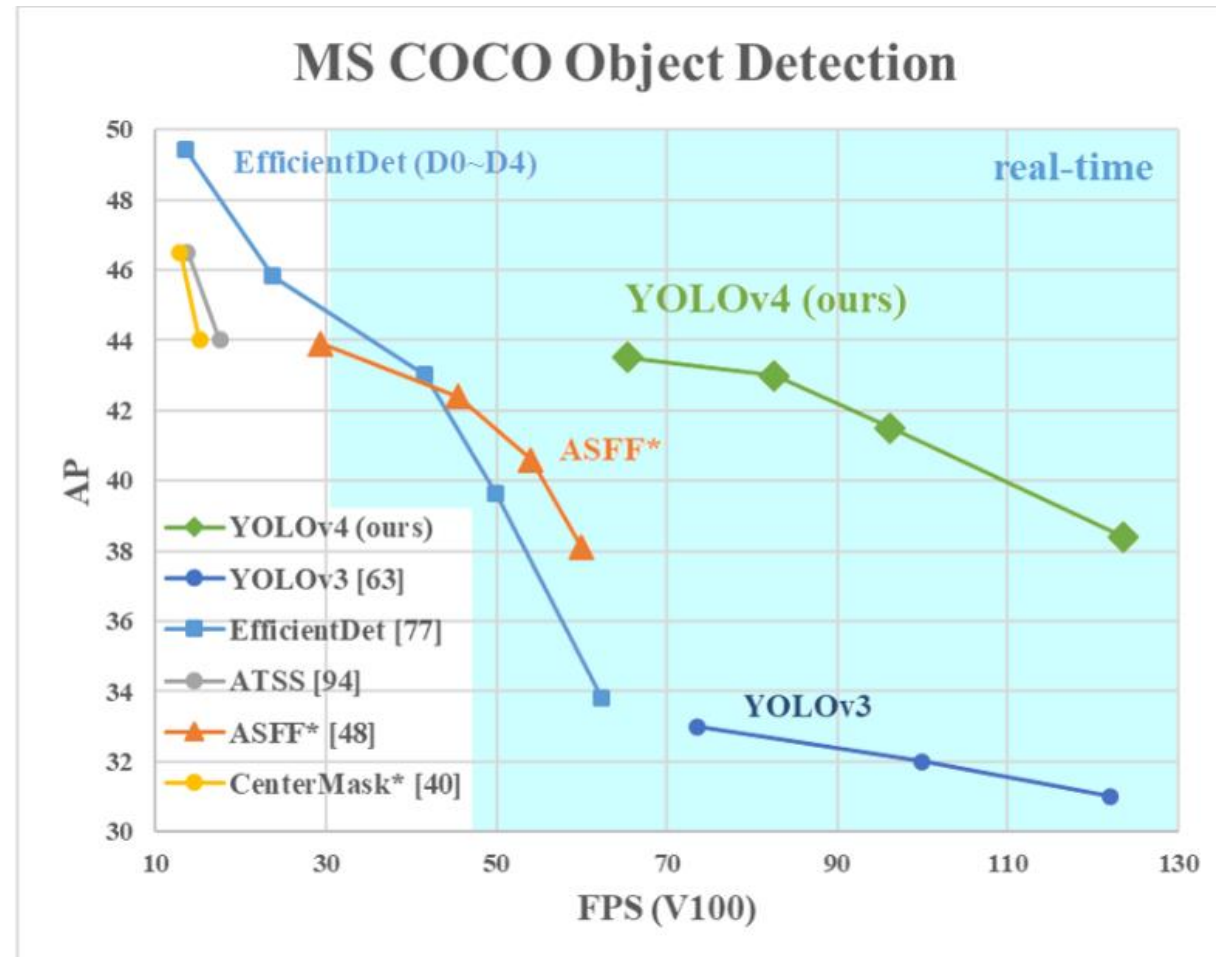
Practice: Real time

- In practice the polyp is removed directly. Hence, real time detection is required
- A system is considered real time if the detection speed of the neural network exceeds 30 fps.
- To reach such a speed often smaller networks need to be utilized. More layers also means less speed.
- However, more layers in most cases means higher accuracy (if there are enough data) → There is a trade-off between accuracy and speed



Trade-off Real Time Detection

- The image on the right shows YOLOv4 [Bochkovskiy et al.] performance.
- There is a clear trade-off between accuracy (AP) and speed (FPS). The results are computed on a Tesla V100 (at the time of YOLOv4 the fastest available GPU)
- All algorithms with higher FPS have smaller AP.



Usage: Polyp Detection

- A practical usage for gastroenterologists can only be achieved with real time polyp detection
- Therefore, the proper trade-off between accuracy and speed must be determined
- Then, it must be researched if the additional support really leads to improved cancer prevention, such as in "Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CAdE-DB trial): a double-blind randomised study."
Wang, Pu, et al.



Outlook

Roles of AI and doctors

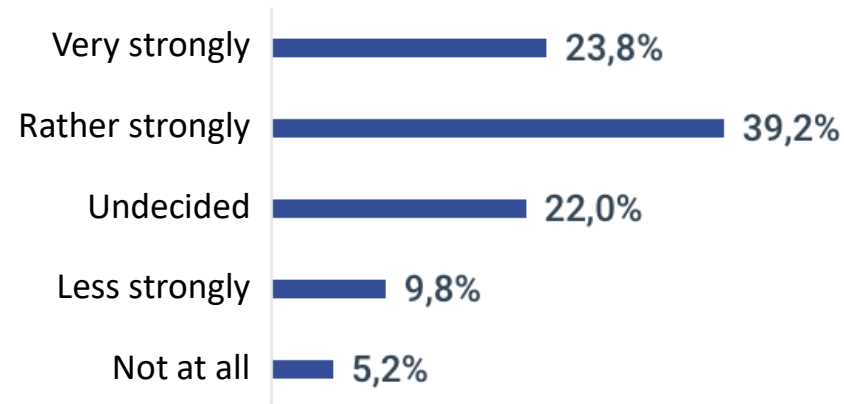


Roles of AI and doctors

In some specific tasks an AI is already better than doctors, for instance in detecting breast cancer in mammograms of a radiologist. Will radiologists as a result have to focus more and more on intervention radiology, i.e. image controlled interventions, where the software does their job first and better than currently?

What do doctors think?

How strongly, in your opinion, can AI contribute to diagnosing diseases earlier and with more certainty?



● 2,5%



Repräsentativ ▾

Quelle: Ärzteblatt



Ethical question: Who is responsible for false AI diagnoses?

Currently, AIs cannot be responsible themselves. There are EU level discussions about introducing a so-called electrical person; however, it is not about classical responsibility, but rather about the adoption of monetary accountability and creating a contact person for the potentially harmed.

Translated quote of Prof. Dr. Susanne Beck
holder of the chair for criminal law, criminal procedural law,
comparative criminal law and legal philosophy at the university of
Hannover

