The Abdus Salam International Centre for Theoretical Physics

IAEA International Atomic Energy Agency

unesco United Nations Educational, Scientific and Cultural Organization

AI Lecture: Advanced applications of AI in Diagnostic Imaging and Imaging in Radiotherapy

College on Medical Physics 2026

Michele Avanzo
Medical Physics
CRO Aviano IRCCS
Trieste, 2026/09/11

24 August - 12 September 2026

Trieste, Italy

Outline

01 Deep Learning Neural Networks for Image Classification

02 Other tasks: Regression, Segmentation, Detection

03 Overfitting, Regularization

04 Other DL architectures

05 AI in Radiology

06 Interpretability

07 The Role of the Medical Physicist in AI

SECTION 01

Deep Learning Neural Networks

What is Artificial Intelligence? AI · ML · DL

AI

Artificial Intelligence

Techniques that allow machines to imitate human cognitive functions: learning, reasoning, problem-solving.

ML

Machine Learning

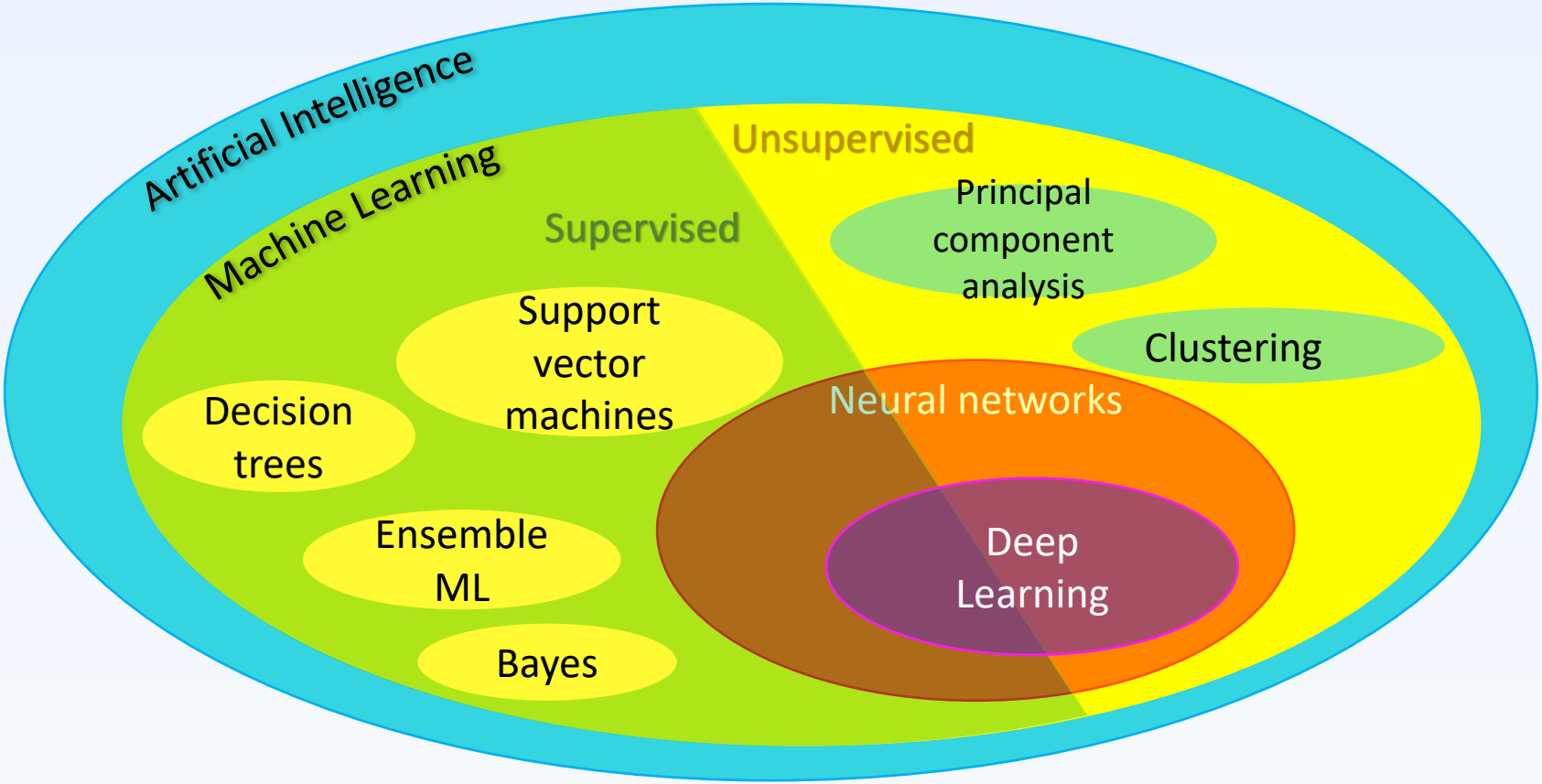
Algorithms that learn cognitive functions from data. Traditional ML processes only numerical data.

DL

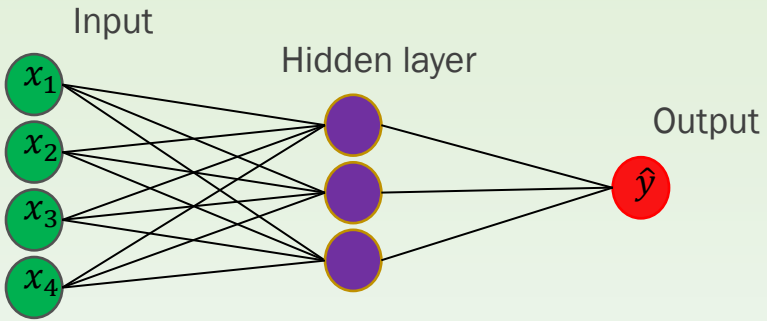
Deep Learning

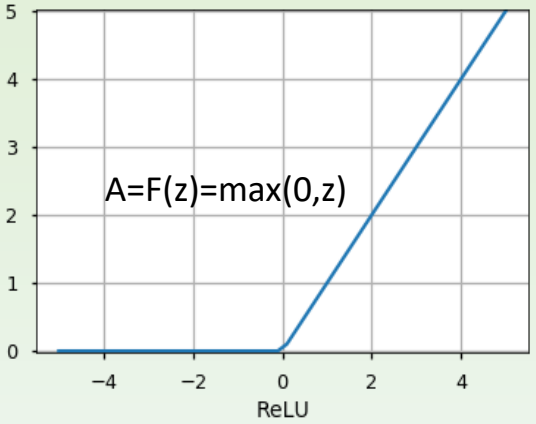
Multi-layer neural networks that learn directly from raw data

ML · DL

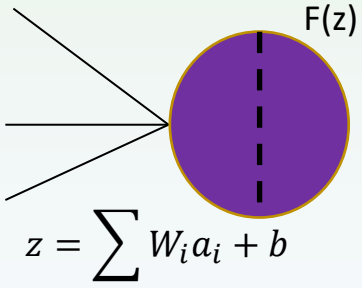

Artificial Intelligence
Machine Learning
Supervised
Unsupervised
Principal component analysis
Clustering
Support vector machines
Decision trees
Neural networks
Ensemble ML
Deep Learning
Bayes

Neural networks



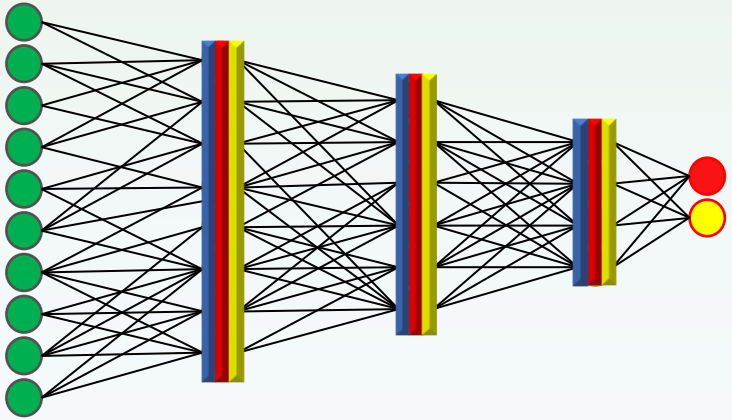


ReLU=rectified linear unit



Each neuron i:

- Computes the pre-activation: $\mathbf{z = \Sigma\, w_i\, a_i + b}$
- Applies the activation function **a = f(z)** and sends it to the next layer



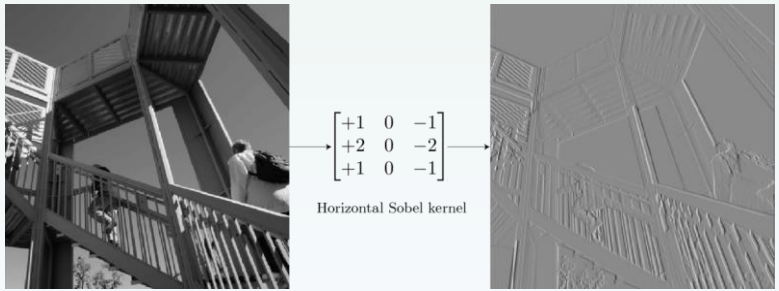
Types of Layers in Deep Neural Networks

Convolutional layer

Applies learned filters (kernels) to the images

3_0	3_1	2_2	1	0
0_2	0_2	1_0	3	1
3_0	1_1	2_2	2	3
2	0	0	2	2
2	0	0	0	1

12.0	12.0	17.0
10.0	17.0	19.0
9.0	6.0	14.0



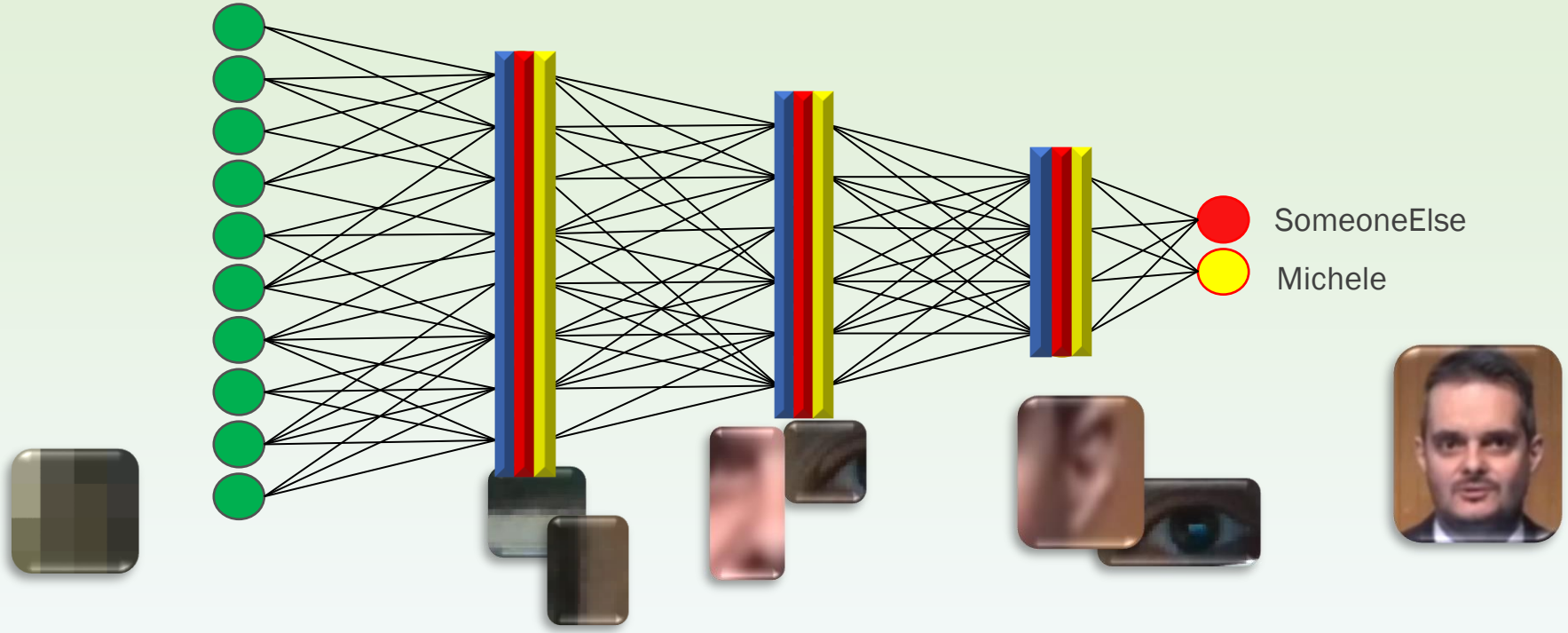
Pooling Layer

Reduces the spatial resolution (downsampling). E.g. Max pooling: takes the maximum in each window

21	60	37	-10	4
28	53	63	20	15
14	22	23	8	19
-8	20	49	1	23
11	55	17	22	1

60			

CNN: Features of Increasing Complexity per Layer



Input
Raw image
(pixel RGB)

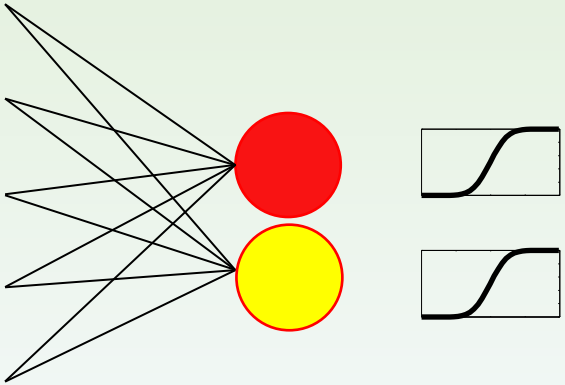
Layer 1
Edges · Contrast
Elementary colors

Layer 2
Texture · Gradients
Local patterns

Layer 3+
Face parts
(eye, lips, ear...)

Output
Class:
Michele /
SomeoneElse

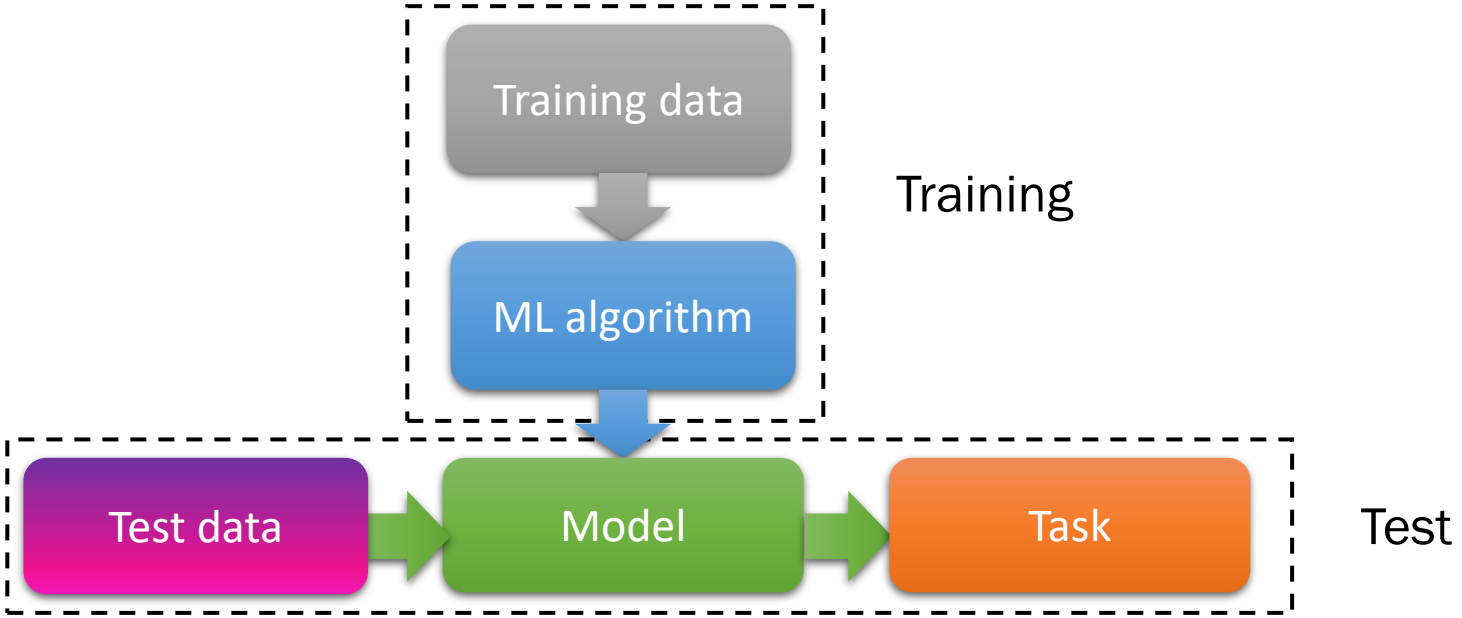
Types of Layers in Deep Neural Networks



Softmax

Calculates probability for each label (e.g. 75% probability of positive)
The probability is converted to a binary label using a threshold

Training and testing

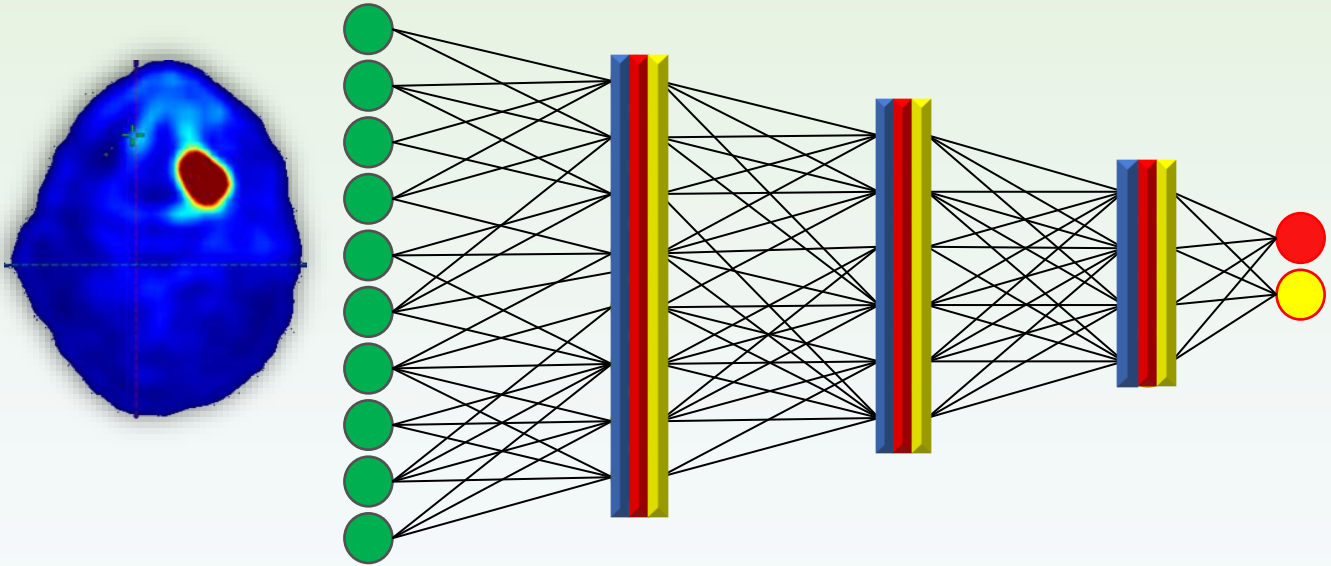


Training: Backpropagation and Optimization

1

Forward Pass

Data → Network → output (predicted label)



Training: Backpropagation and Optimization

1

Forward Pass

Data → Network → output (predicted label)

2

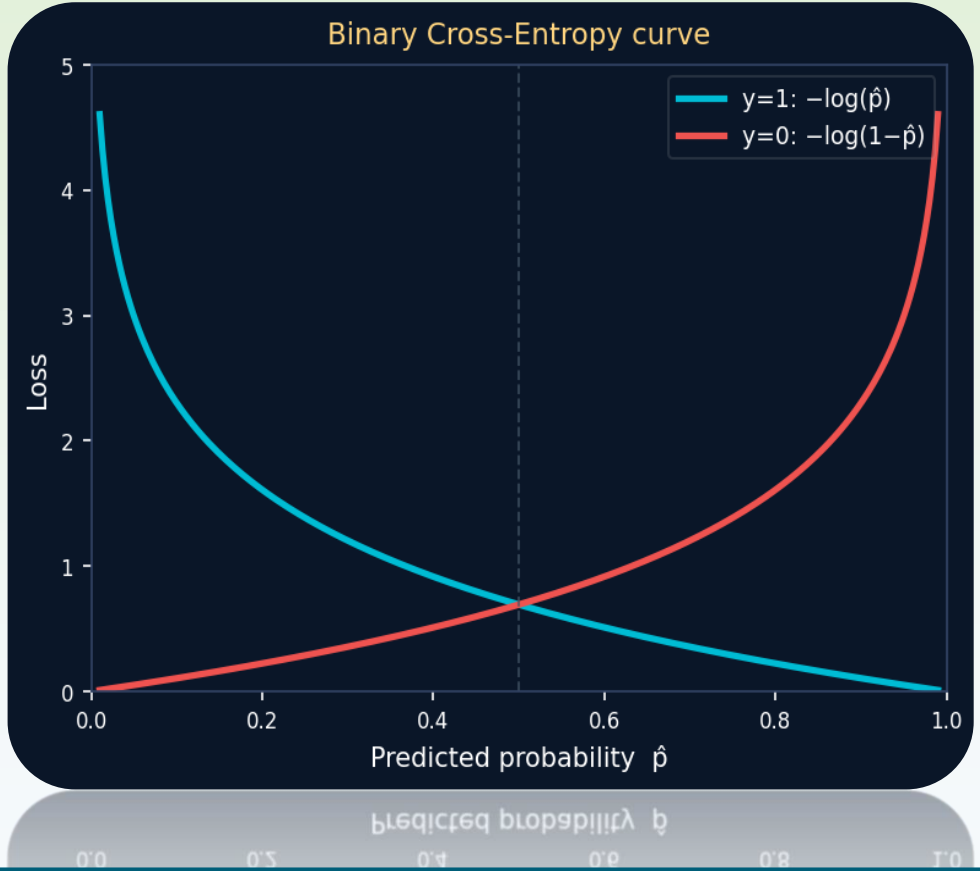
Cost function

Quantity to measure difference between predicted and true label

Cost Function for classification: Binary Cross-Entropy

$$Loss(BCE) = -\frac{1}{N}\sum[y - log\hat{p} + (1-y)\cdot log(1-\hat{p})]$$

Measures the difference between predicted probabilities, ŷ, and true labels, y (“ground truth”)



Training: Backpropagation and Optimization

1

Forward Pass

Data → Network → output (predicted label)

2

Cost function

Quantity to measure difference between predicted and true label

3

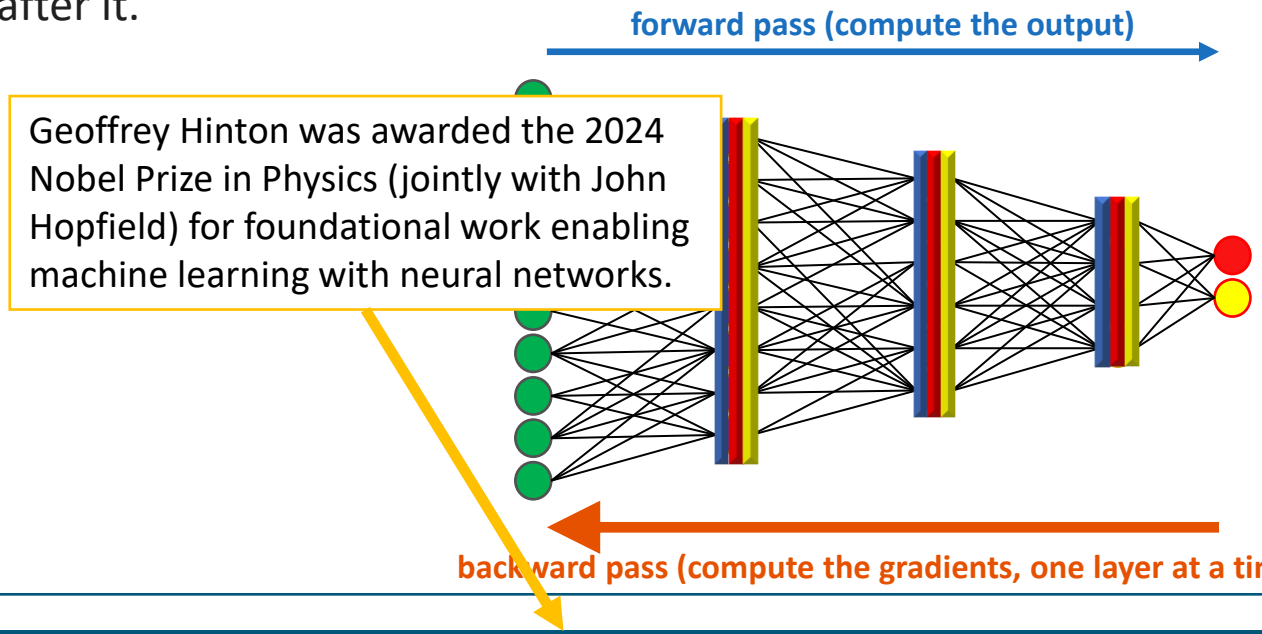
Backpropagation

Computation of the gradient of the cost function versus network weights

Backpropagation: which way do the calculations go?

To learn, the network needs to know how much each weight, W, contributed to the loss function, L: ∂L/∂W

That can only be computed starting from the output and moving backward, via the chain rule of derivatives, because each step needs a result that is only available from the step after it.



Rumelhart, Hinton & Williams, Nature 1986 · backpropagation algorithm

Training: Backpropagation and Optimization

1. **Forward Pass**
   Data → Network → output (predicted label)

2. **Cost function**
   Quantity to measure difference between predicted and true label

3. **Backpropagation**
   Computation of the gradient of the cost function versus network weights

4. **Weight update**
   We update each weight using the computed gradients $W \leftarrow W - \eta \cdot \partial L/\partial W$

Training Terminology: Practical Glossary

Epoch

A complete training over the entire training set.

Batch

Subset of data processed before updating the weights.

Iteration (Step)

A single weight update on one batch.
Iterations/epoch = N_samples / batch_size

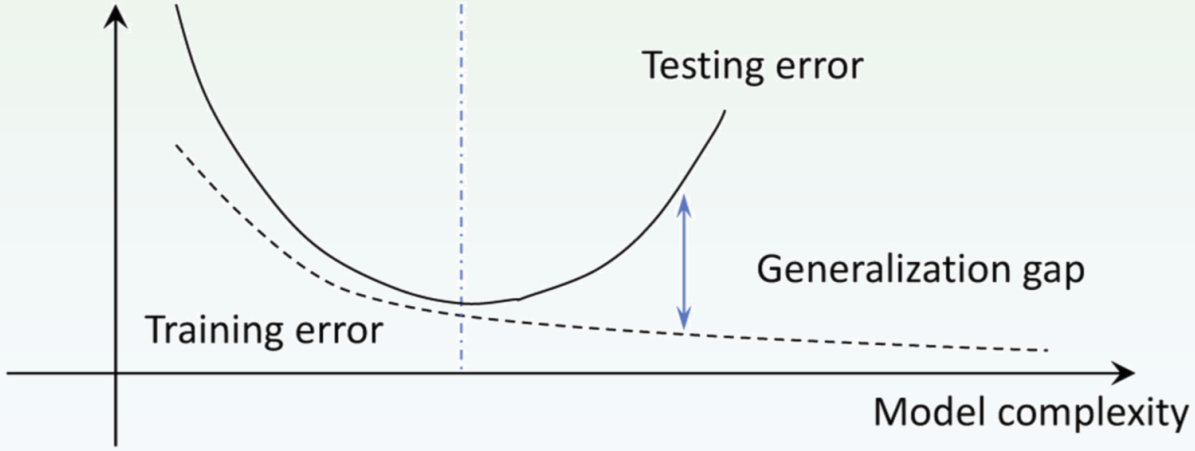
Learning Rate (η)

Size of the weight-update step.
Too high → divergence; too low → slow convergence.

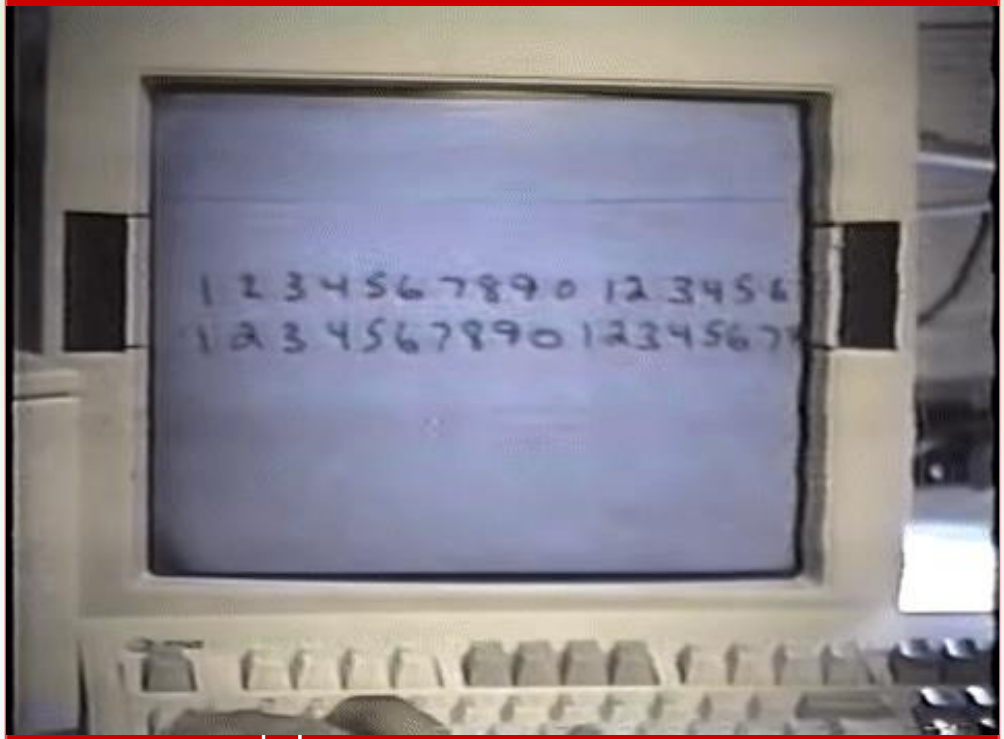
Training Terminology: Practical Glossary

Patience

Number of consecutive epochs without improvement in the validation loss before acting (early stopping or LR reduction). E.g.: patience=10.



The First Deep Learning Classification



youtube.com/watch?v=FwFduRA_L6Q

1989 — LeNet published
LeCun et al., Neural Computation: first CNN with backpropagation to read images (USPS postal digits).

1993 — Real-Time Demo
First public real-time video demonstration at Bell Labs AT&T on a 486 PC

1998 — Industrial Impact
LeCun's CNN processes 10–20% of US bank checks in the late 1990s

Metrics for classification

Confusion Matrix

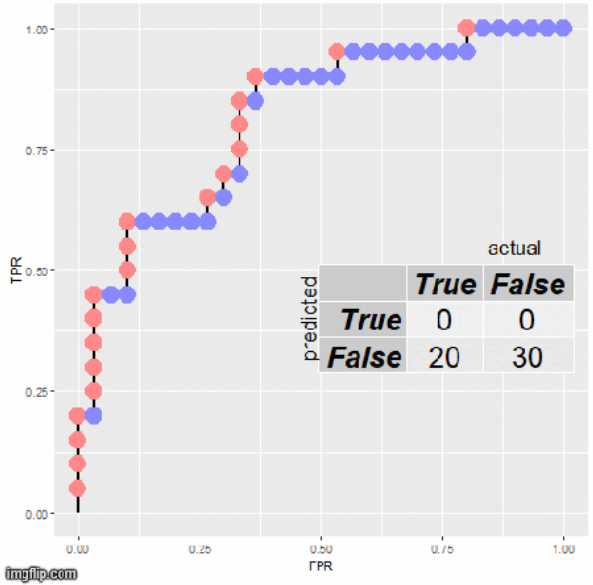
		Predicted	
		N	P
Actual	N	TN	FP
	P	FN	TP

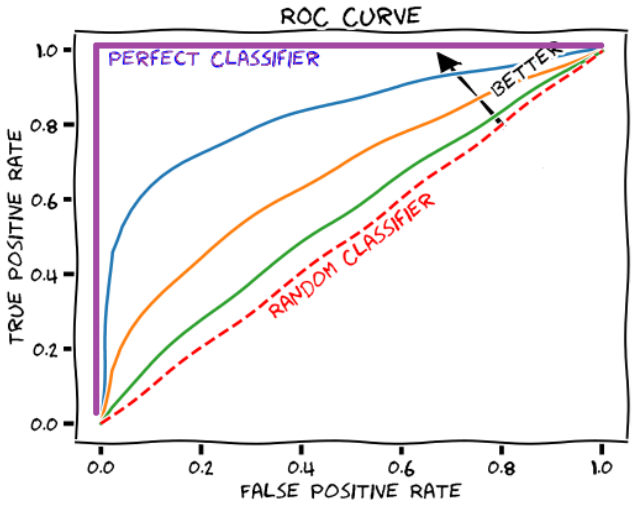
$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}$$

Receiver Operating Characteristic (ROC) curve





SECTION 02

Other tasks: Regression, Segmentation, Detection

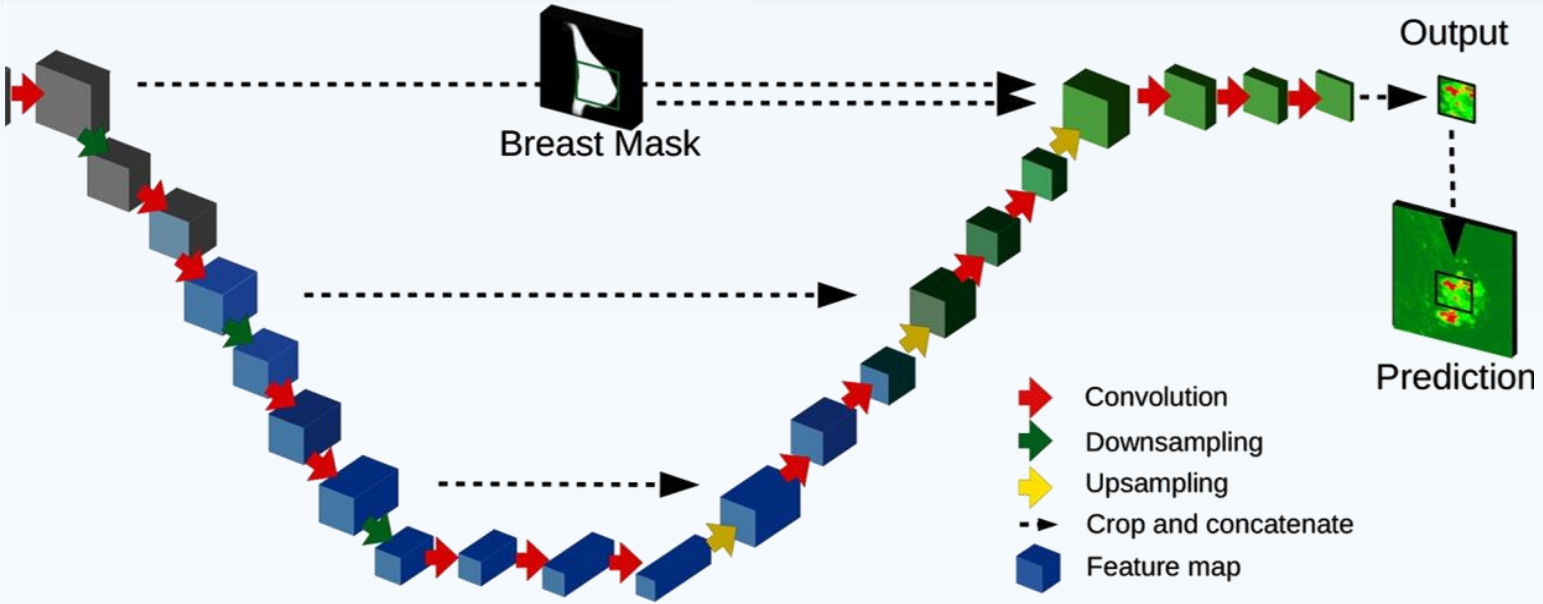
Advanced Layers: Skip Connections, Attention, Upsampling

Transposed Convolution

Inverse operation to convolution: increases spatial resolution

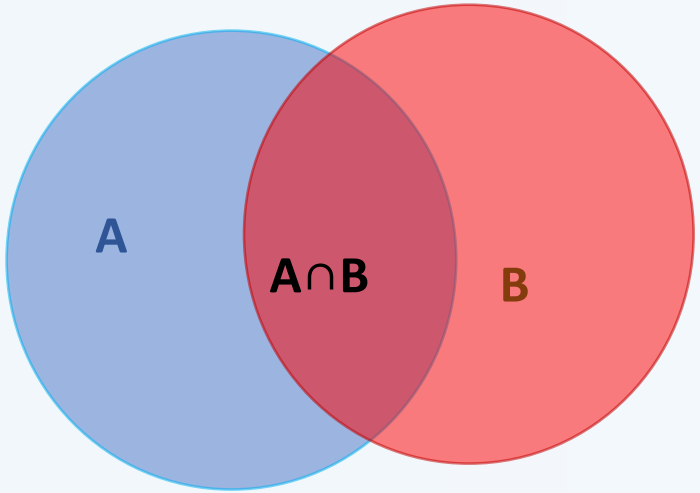
Skip Connection

Connect encoding layers to decoding layers



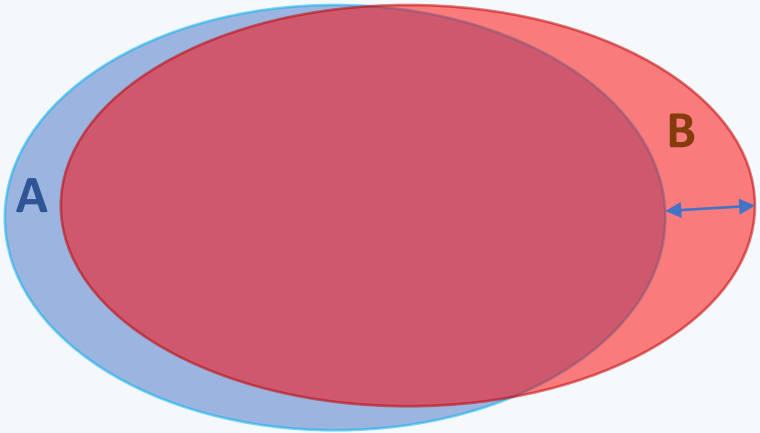
$$\text{Dice} = \frac{2A \cup B}{A+B}$$

Dice loss = 1-Dice

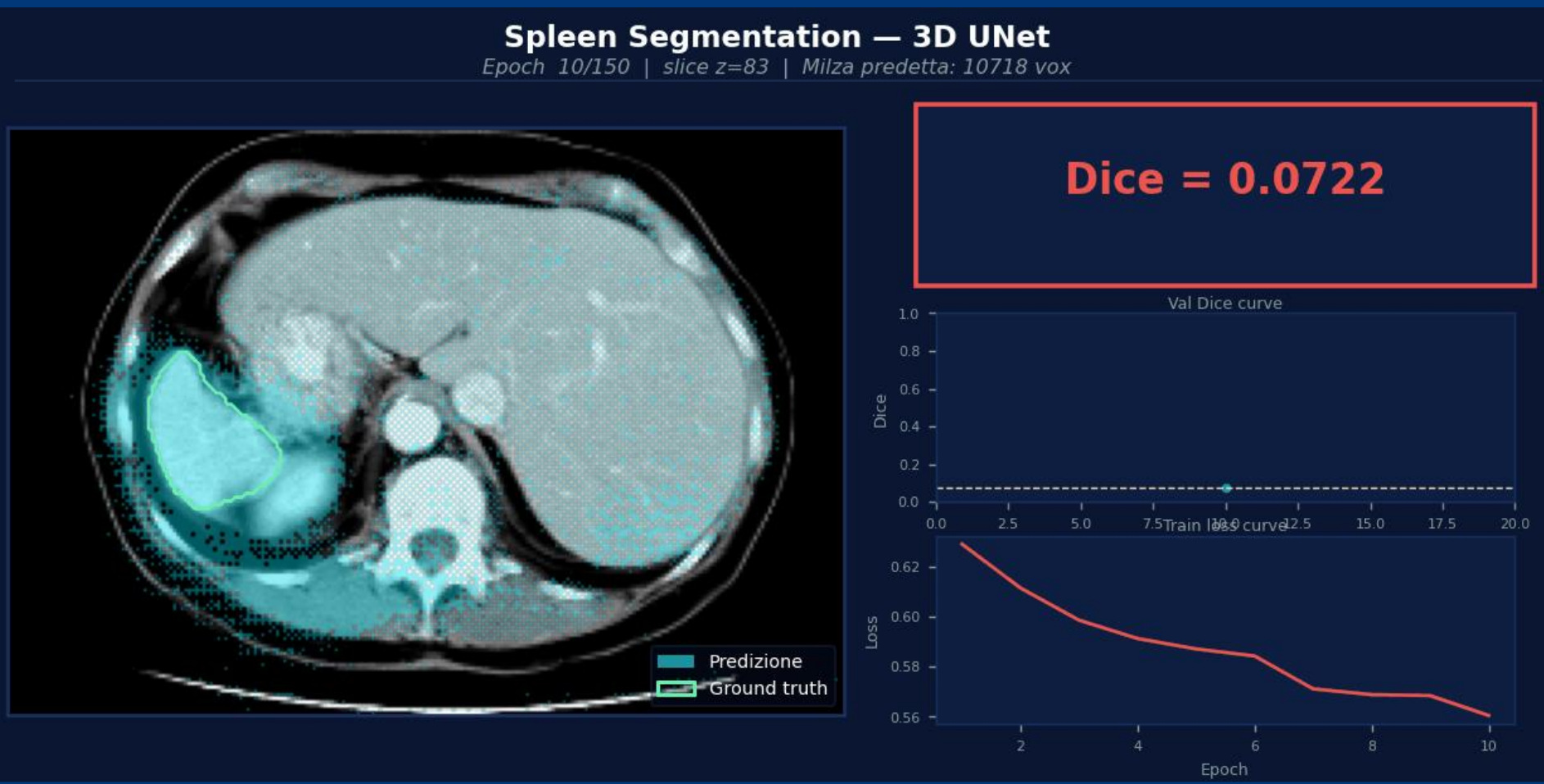


Hausdorff distance H(A,B) = max(max min d(a,b), max min d(b,a))

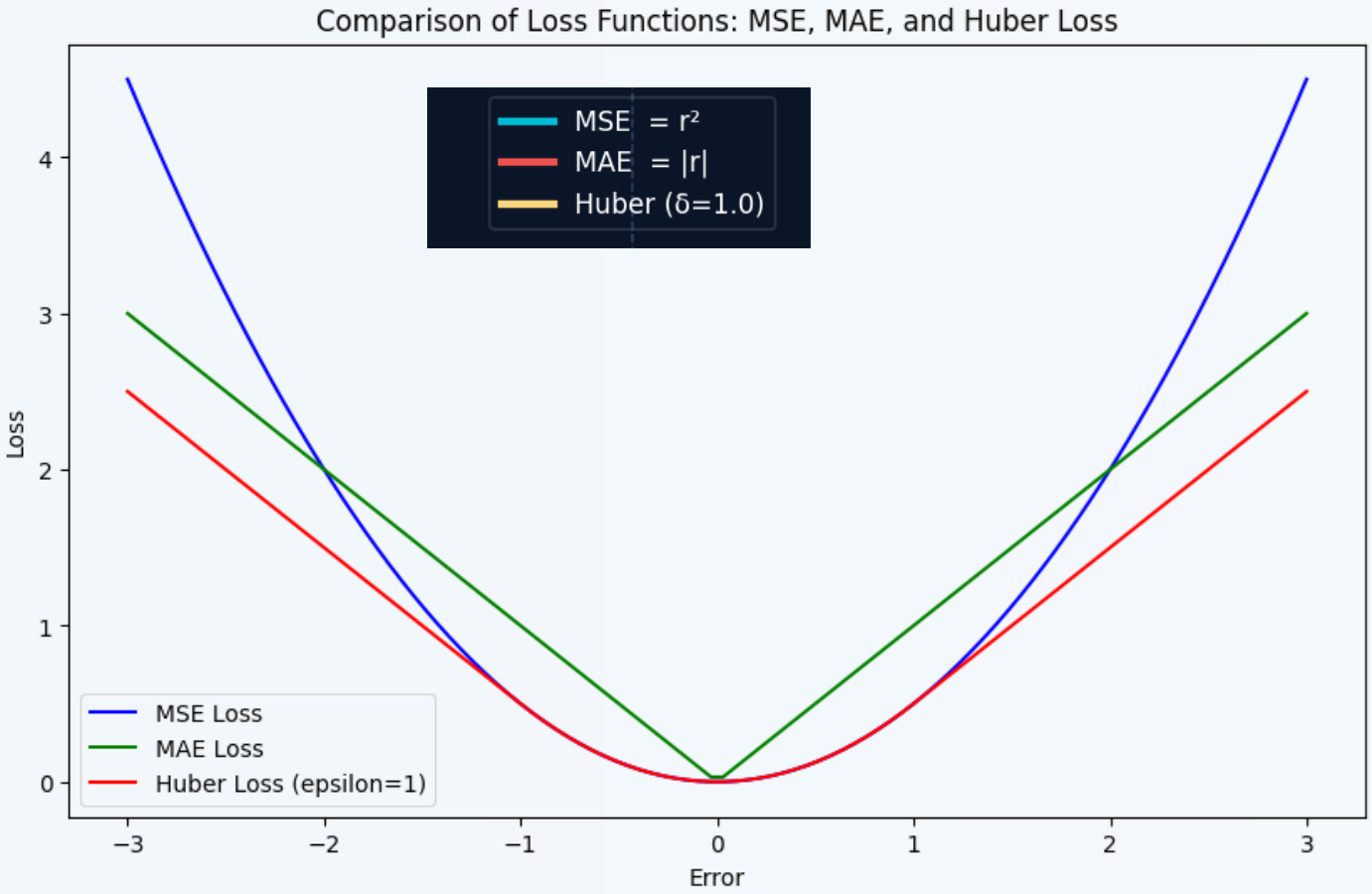
- For each point in A, the nearest point in B is found;
- One takes the maximum of these minimum distances. Not usable as a loss



Dice Loss Function

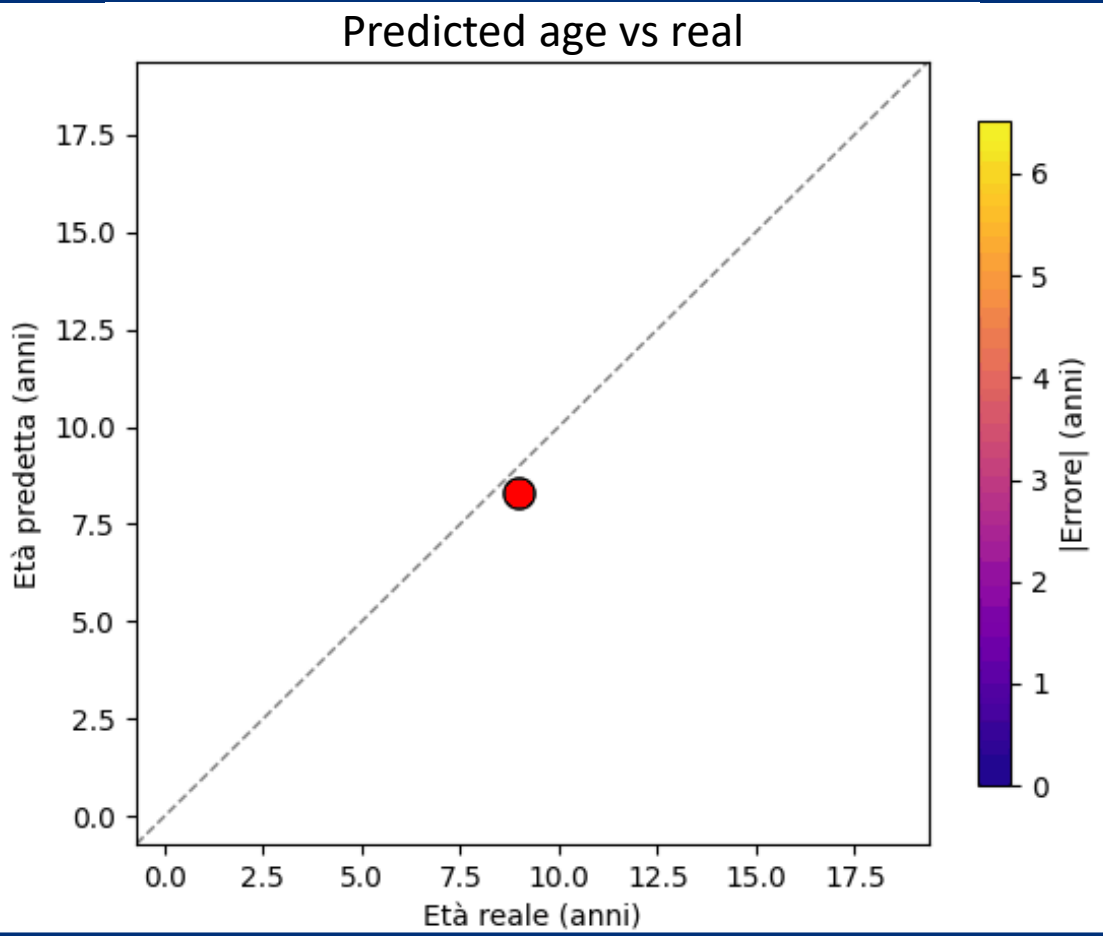


Regression: Mean Squared Error and Mean Absolute Error



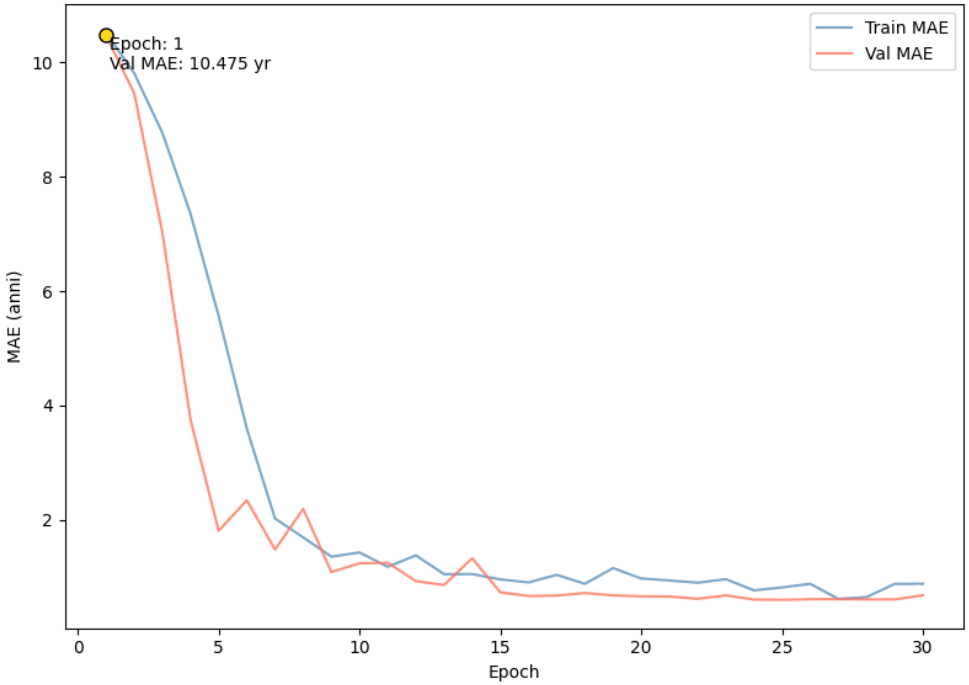
Huber Loss combines the strengths of Mean Squared Error (MSE) and Mean Absolute Error (MAE), switching from one to the other beyond an error threshold.

Regression example: bone age

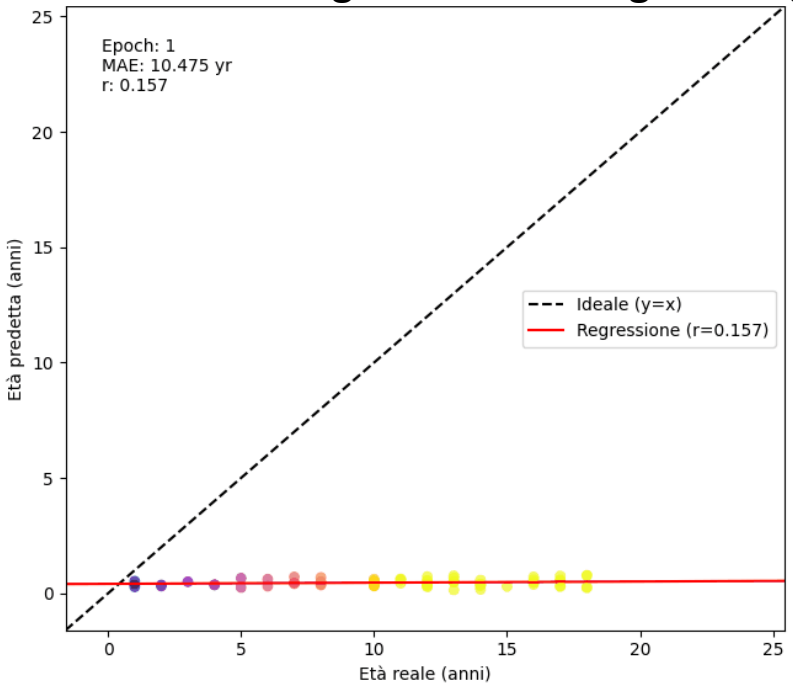


Training a regression network

Learning curve



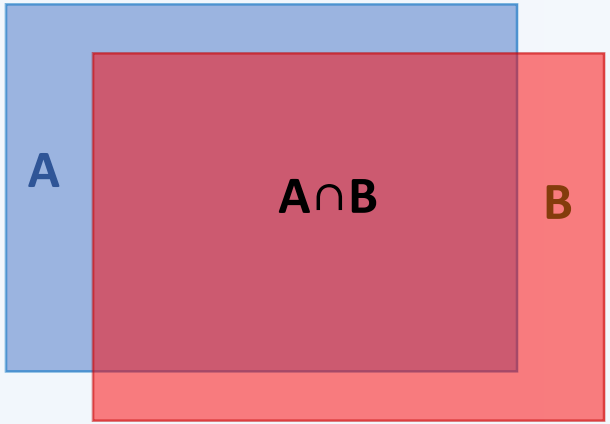
Predicted age vs real during training

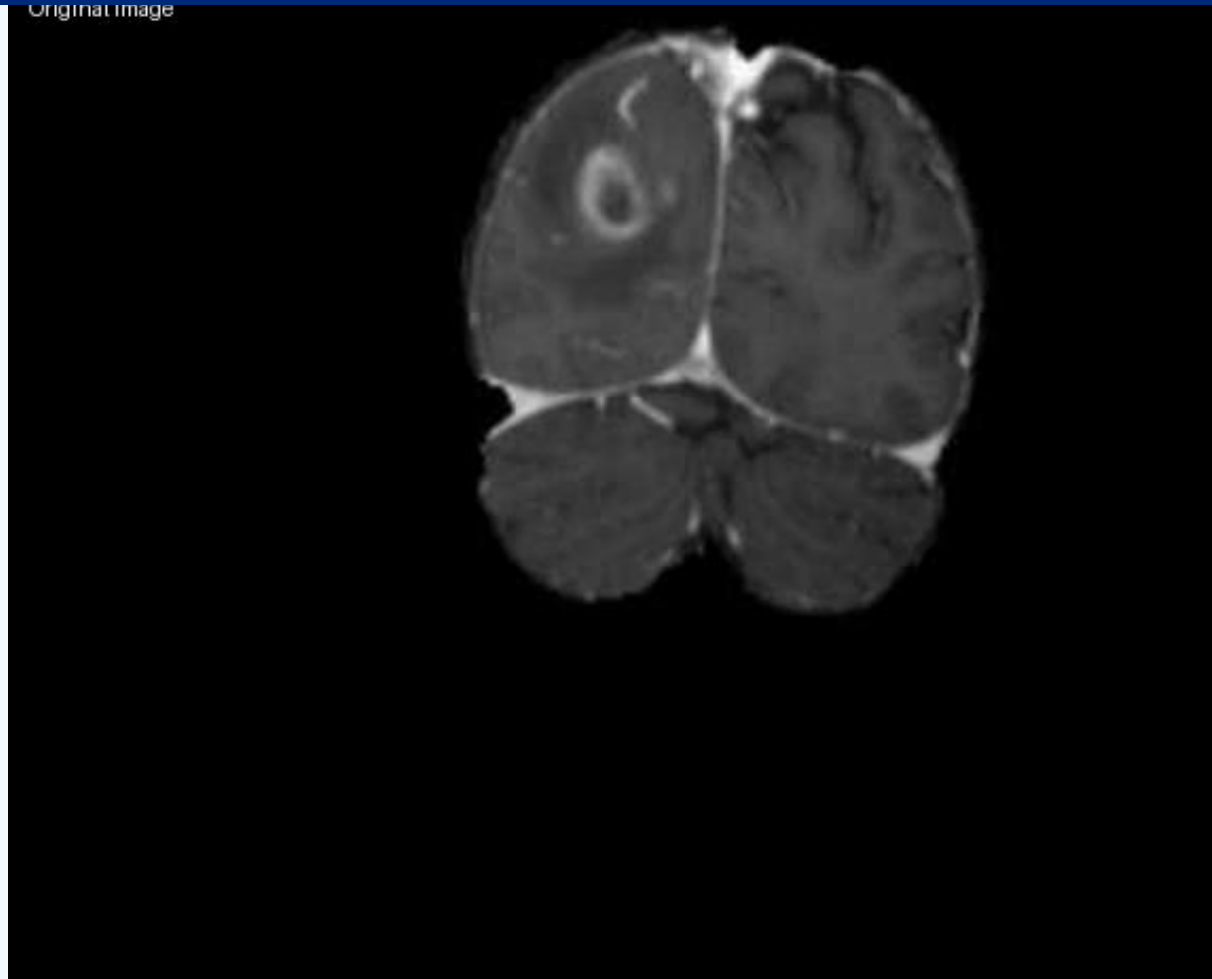


Detection: YOLO (you only look once)

Intersection over Union =

$$\frac{A \cap B}{A \cup B}$$





SECTION 03

Overfitting, Regularization

Overfitting, Regularization and Cross-Validation

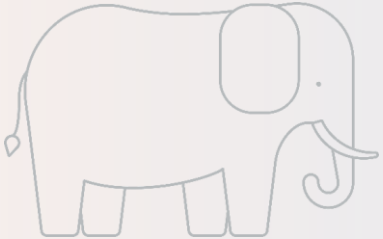
Overfitting

- The model memorizes the training set
- Does not identify the important features in the images
- Training error ↓ but testing error ↑

Underfit

The model fails to capture the features of the examples

Good fit

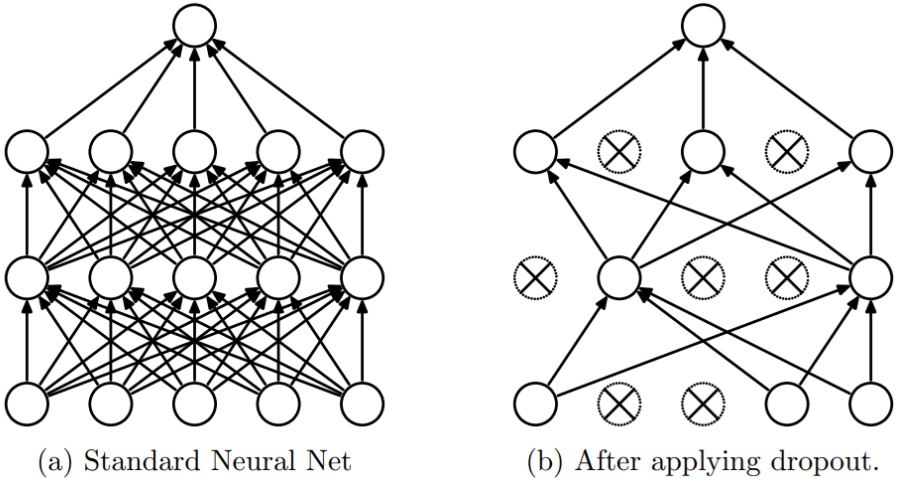


The model captures the features of the sample

Overfitting

The model memorizes the training set without capturing its features

Activation Functions and Regularization



(a) Standard Neural Net

(b) After applying dropout.

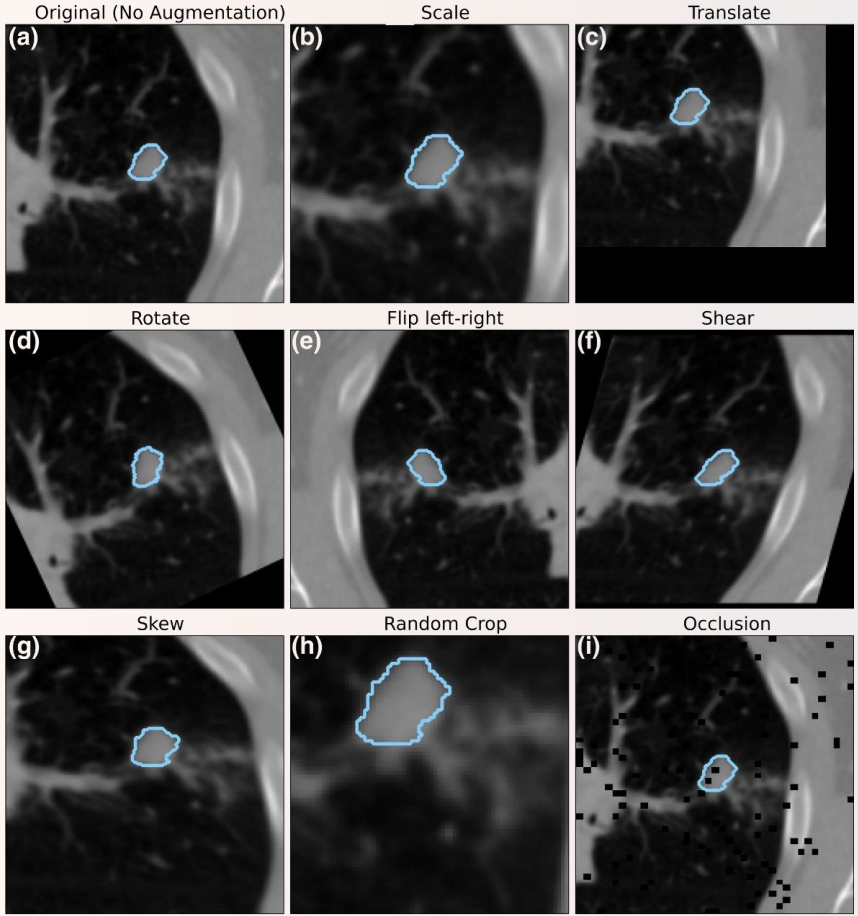
Regularization

Penalty on weights, reduces overfitting by limiting the modulation of the weights.

Dropout

Randomly deactivates neurons (20–50%) during training.

Data augmentation



SECTION 04

Other DL architectures

Transformer · GAN · Foundation Models

Vision Transformers (ViT)

1) treat an image as a sequence of small regions (“patches”) similarly to Natural Language processing. E.g. image divided into 16x16 patches
2) Use an attention mechanism that identifies the weights of the patches

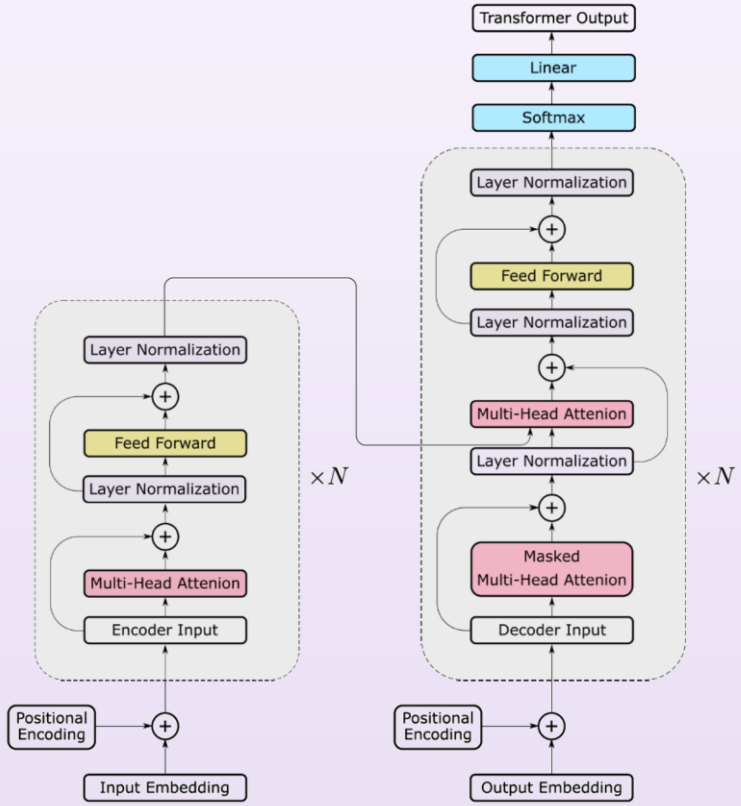


Fig. 3. A brief illustration of a typical Transformer architecture [29].

Vision Transformers (ViT)

Used in various domains.
Classification: performance similar to CNN, require larger datasets

Segmentation: they work well in contexts where modeling spatial dependencies matters, e.g. multi-organ segmentation of the abdomen. Where accurate segmentation of small organs is required, U-Net is better

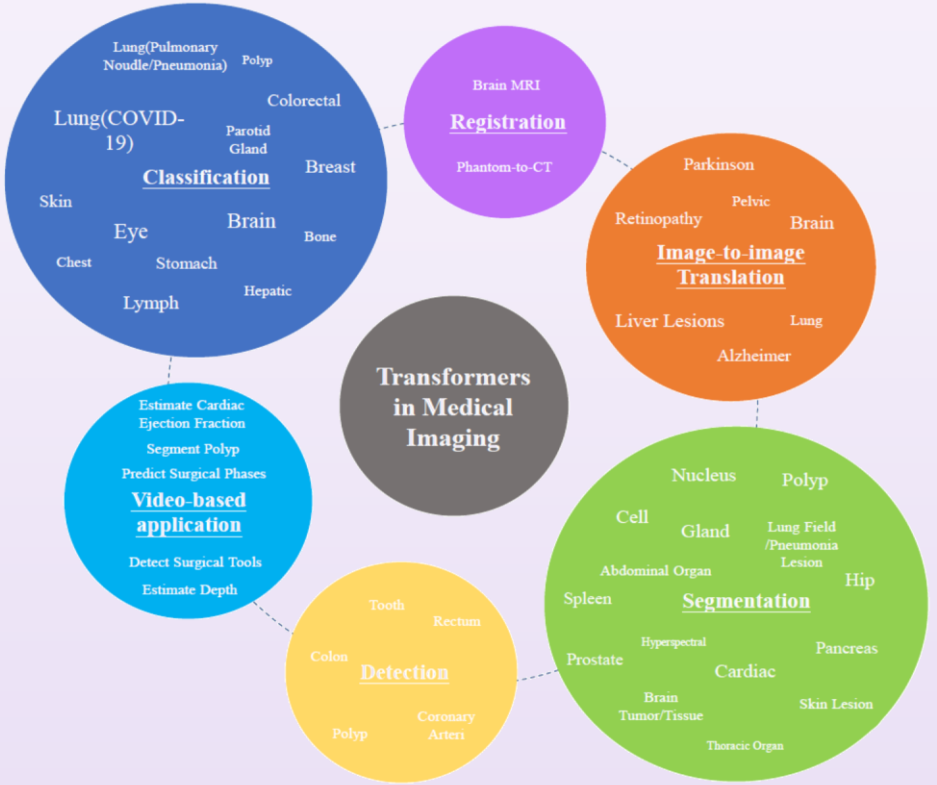


Fig. 5. Applications of Transformers in medical image analysis, as reviewed in this work.

Generative Adversarial Network (GAN)

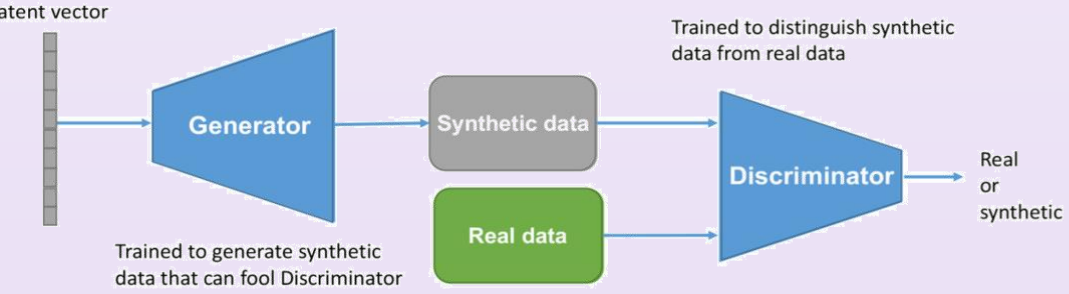
Generator

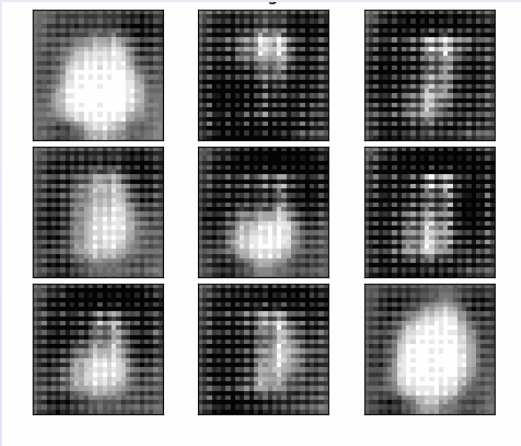
- Receives random vectors as input
- Produces realistic synthetic data
- Goal: fool the discriminator

VS

Discriminator

- It is trained as a binary classifier to distinguish real data from synthetic data
- Feedback to the generator to improve the synthesis





Shen et al., Phys Med Biol 2020

Loss Metrics: Structural Similarity Index Measure

SSIM evaluates the similarity between the real and synthetic images

It is the product of three terms: luminance, contrast, structure

X = patch of the reference image
Y= patch of the synthetic image
C = constant to ensure function stability
μ = mean, σ = variance or covariance

$$I(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_x^2 + C_1}$$

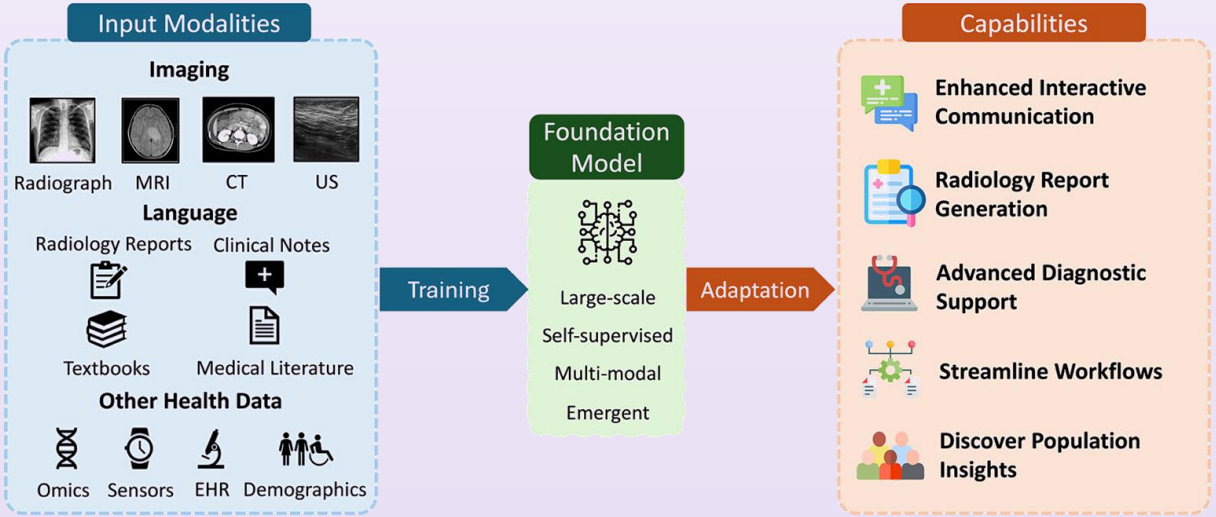
$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

$$SSIM = I^{\alpha}\, c^{\beta}\, s^{\gamma}$$

Foundation models

- Generalist models trained on very large and diverse datasets
- Able to analyze multiple data modalities
- They use self-supervised training techniques to alleviate the need for large expert-supervised databases
- They show emergent capabilities that go beyond the data they were trained on (“zero-shot”)

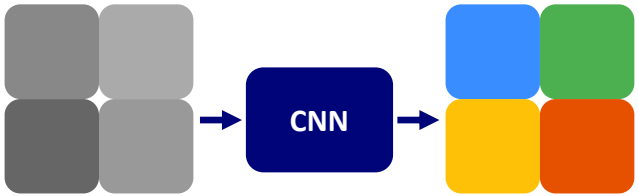


Nguyen E, Med Phys 2025

Self-supervised learning: labels for free

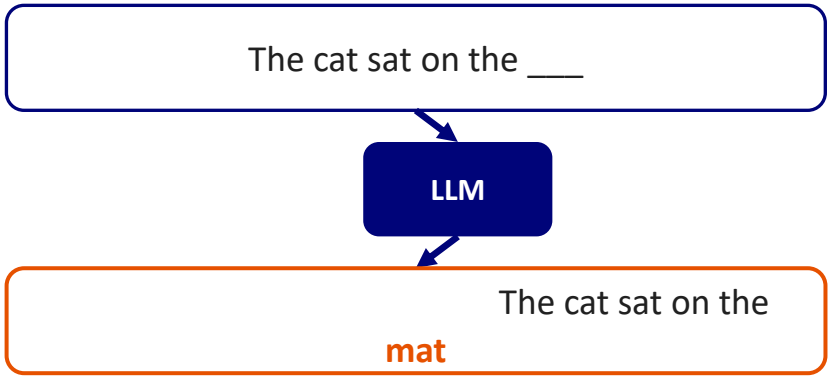
A self-supervised task hides part of the data and asks the model to predict it — the "label" is just the hidden part, already there for free. No human annotator needed.

Vision: colorization



The "label" is just the original color photo — desaturate it, and you already have a free training pair.

Language: next-word prediction



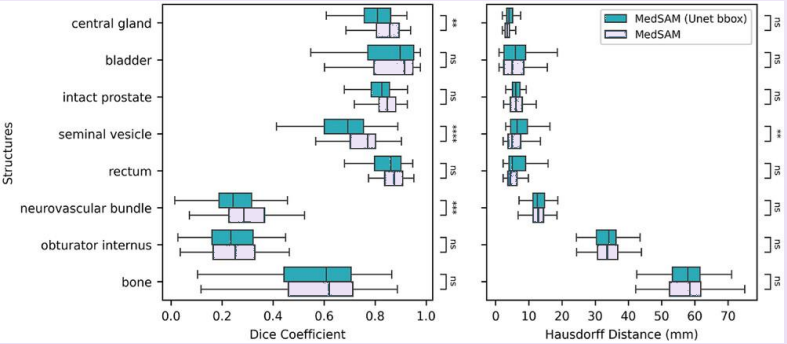
The "label" is just the next word, already sitting in the text — hide it, and you have a free training example.

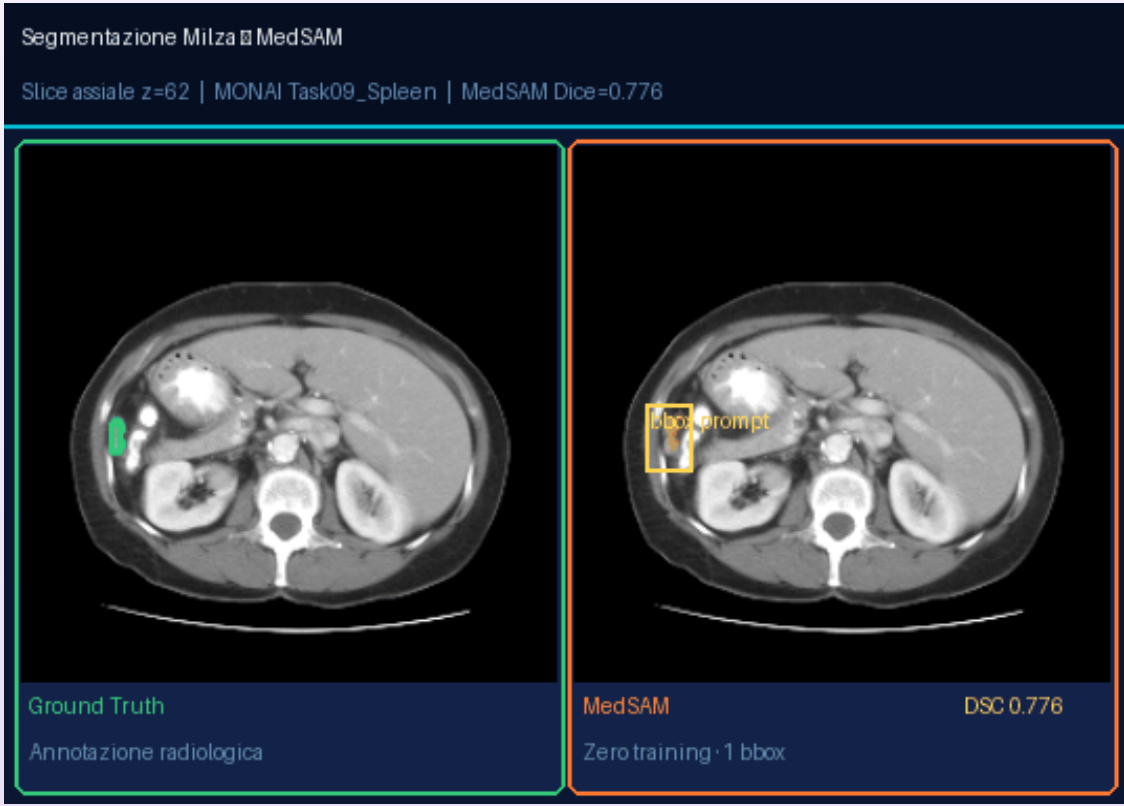
Same idea, two domains: hide a piece of real data, predict it, compare — no manual labeling required.

Zhang, Isola & Efros, "Colorful Image Colorization," ECCV 2016

Foundation models

- Segment Anything (SAM) is an FM developed by Meta for prompt-based segmentation (point or box).
- MedSAM is a version of SAM trained on CT, MRI, endoscopy, ultrasound, pathology, dermoscopy...
- MedSAM accepts only box prompts





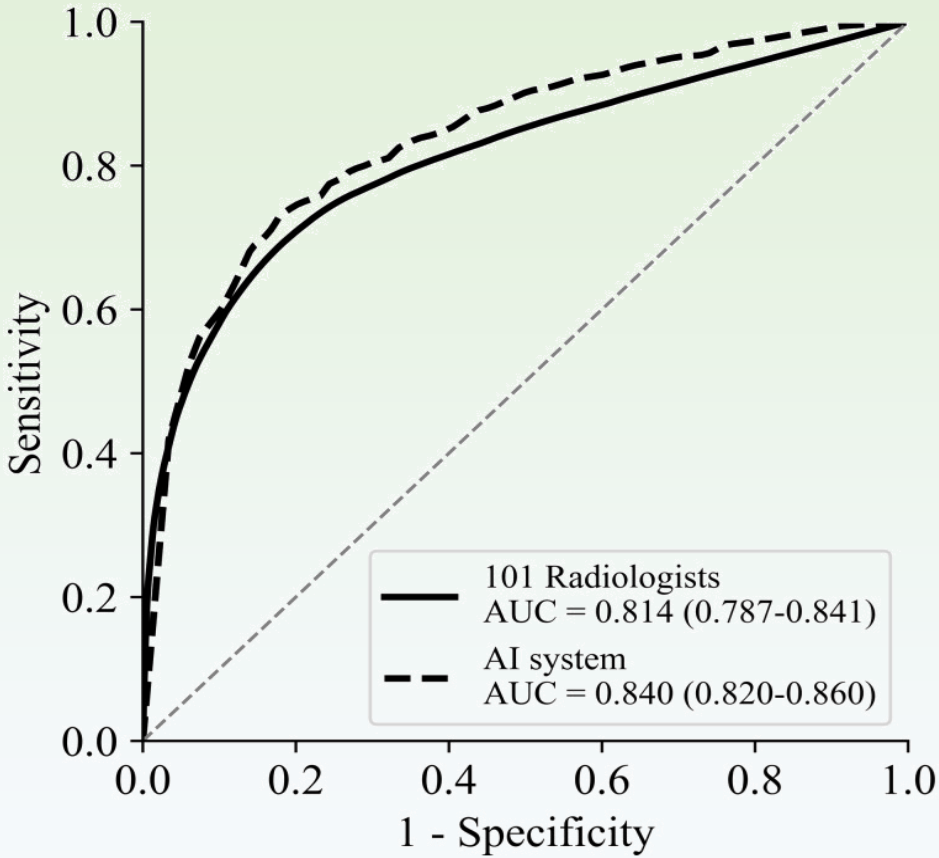
SECTION 05

AI in Radiology

Sinogram → CT · Low-dose CT · MRI Reconstruction · Artifact Reduction

Classification

Breast Cancer Detection in Mammography



Rodriguez-Ruiz, JNCI J Natl Cancer Inst, 2019

Classification

		Predicted	
		N	P
Actual	N	TN	FP
	P	FN	TP

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP+FN}} = \text{recall of positive}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \text{recall of negative}$$

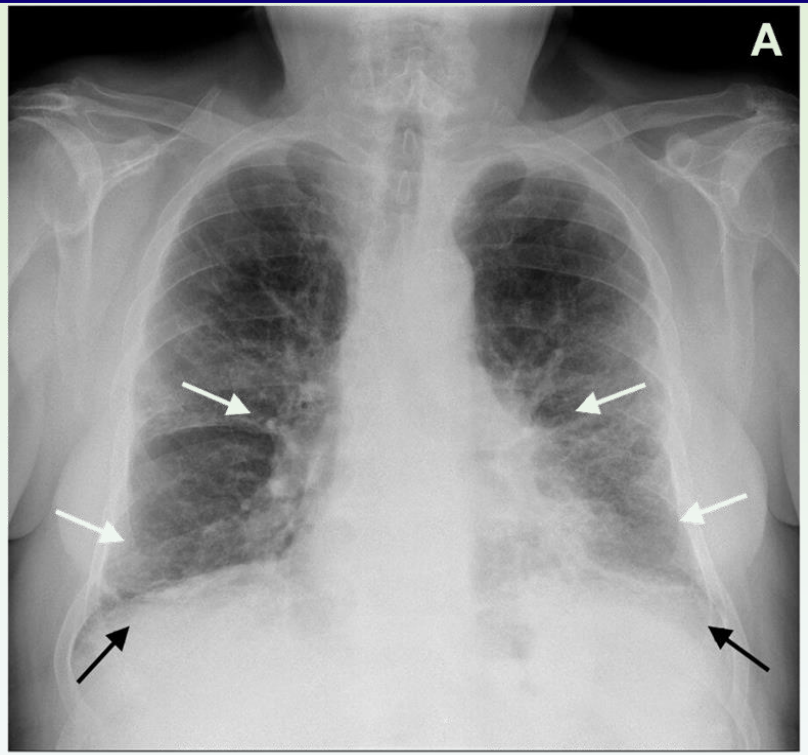
$$\text{accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}$$

Digital medicine

Mammography should include artificial intelligence support

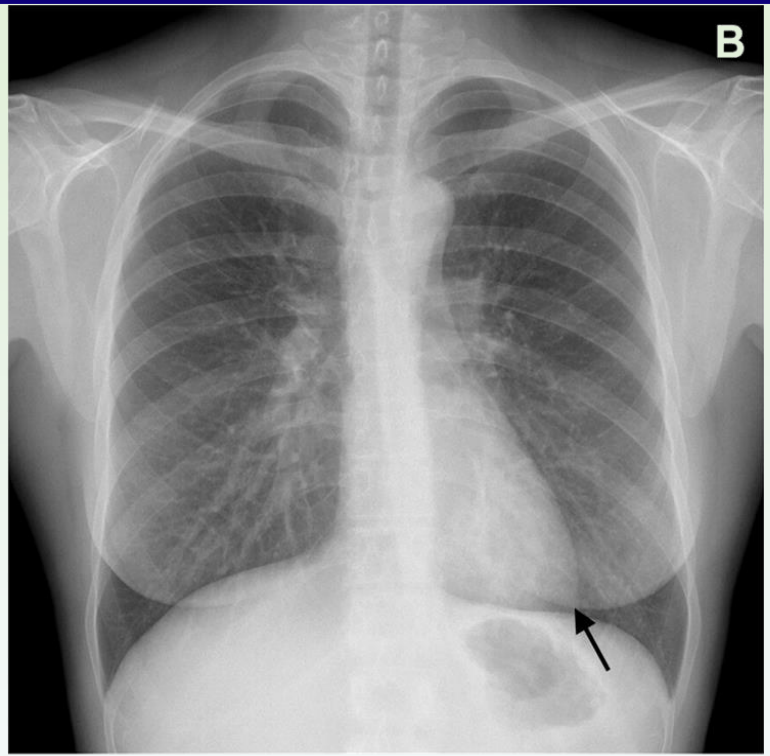
"In the MASAI trial, AI-assisted reading (two radiologists + algorithm) detected 29% more cancers than standard double reading, without increasing false positives"

Report generation from image



AI-generated report

No significant cardiomegaly.
Prominent central pulmonary vasculature.
Increasing interstitial markings bilaterally, more in LML and LLF.
-- r/o pulmonary edema.
Atelectasis or fibrotic changes at lung bases.



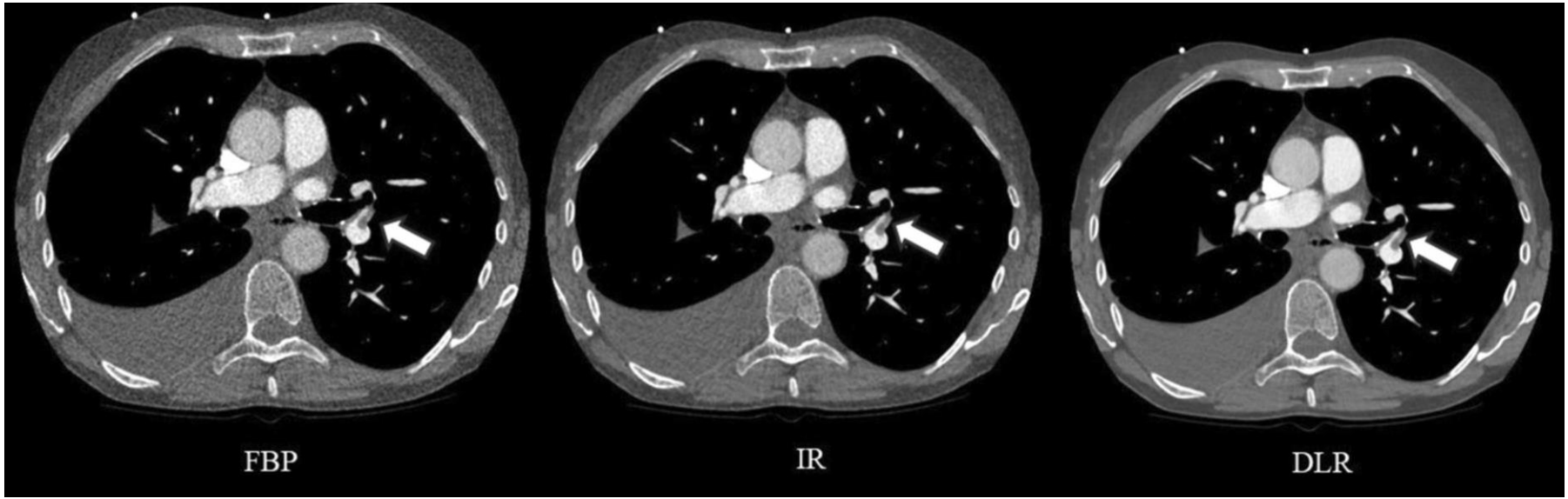
AI-generated report

Slight infiltrate in LLL.
No effusion, nodule, or pneumothorax.
Degenerative changes in the spine.

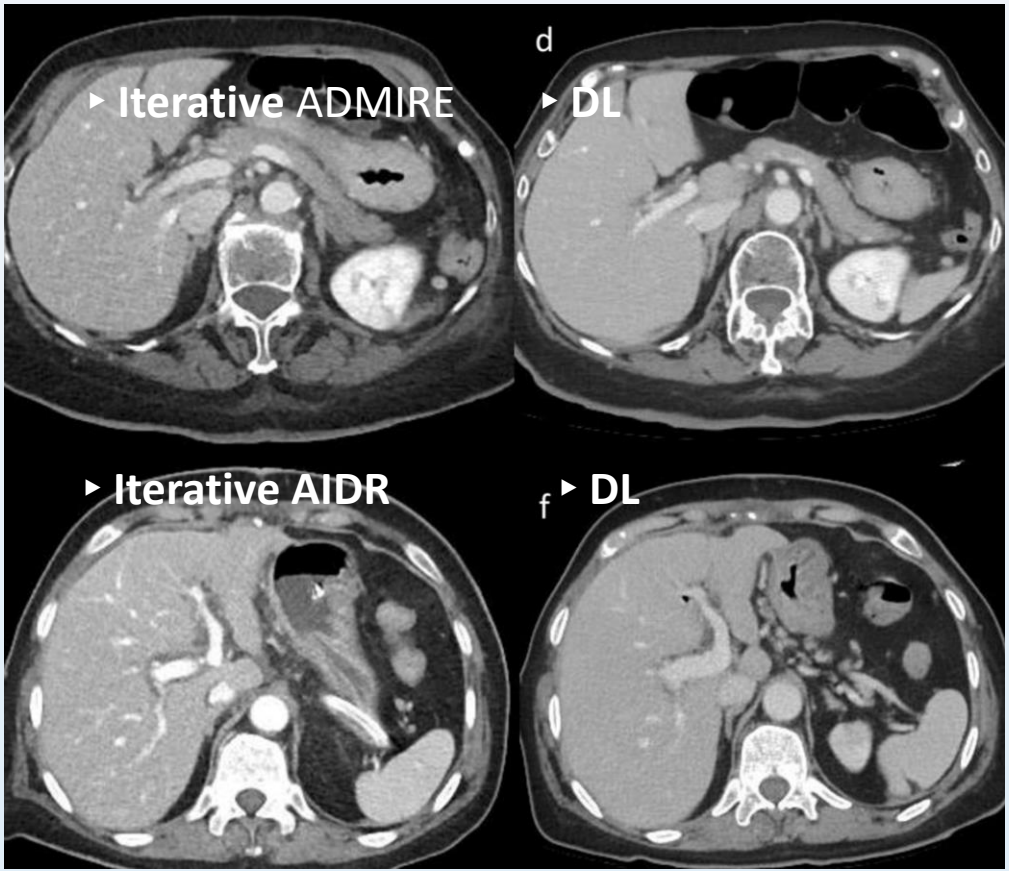
Hong EK, Radiology, 2025

Image reconstruction using deep learning (DRL)

- **Direct DLR** uses neural networks to generate images from raw projection data (sinogram)
- **Projection-space DLR:** Applied to sinogram before conventional reconstruction
- **Image-space DLR:** Applied after reconstruction to reduce noise and enhance clarity

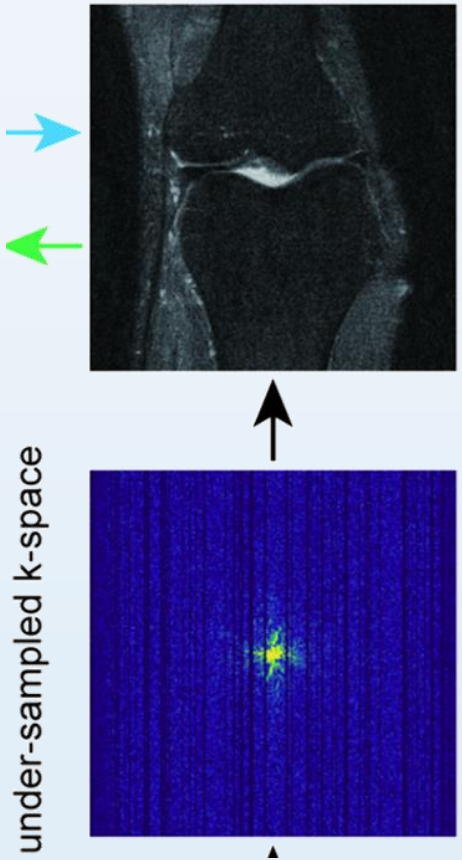


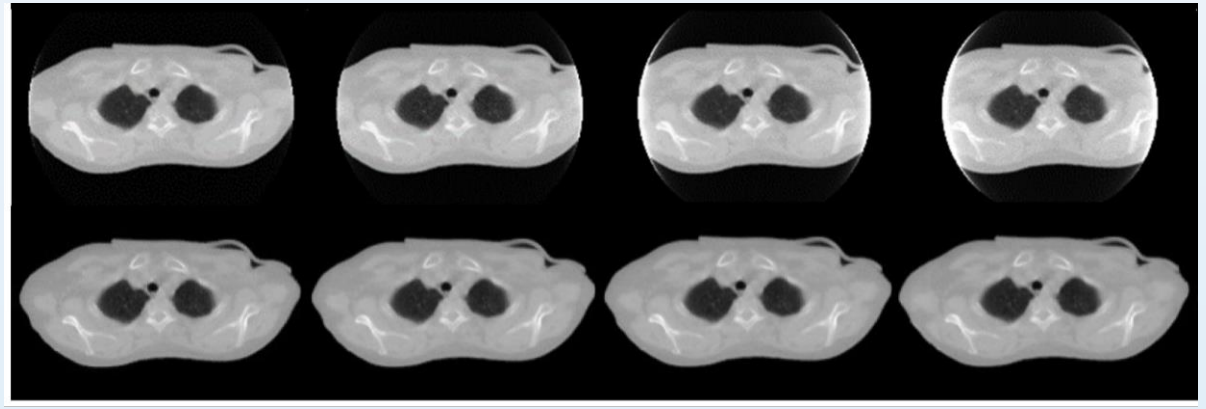
Dose reduction using DLR



- CTDI (mGy) ≈ 29 iterative vs 24 DL

Reconstruction from incomplete data





Shen G, SciRep 2024

Tang, Sensors 2024

AI-based patient dose calculation

- Deep learning models estimate patient-specific organ and effective dose directly from CT localizer/scout images or acquisition parameters, without a full Monte Carlo run
- Replaces or accelerates Monte Carlo dose simulation, enabling near real-time, size-specific dose estimates (SSDE) tailored to the actual patient anatomy

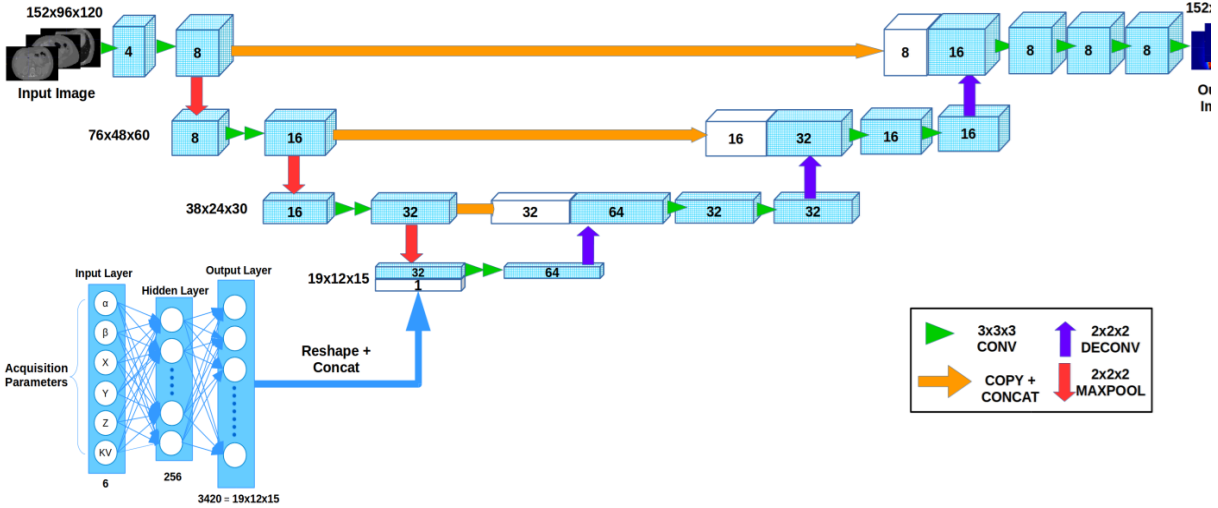


Figure 4: Proposed architecture based on the 3D U-Net with FC layers processing the acquisition parameters.

AI-based patient dose calculation

- Deep learning models estimate patient-specific organ and effective dose directly from CT localizer/scout images or acquisition parameters, without a full Monte Carlo run
- Replaces or accelerates Monte Carlo dose simulation, enabling near real-time, size-specific dose estimates (SSDE) tailored to the actual patient anatomy

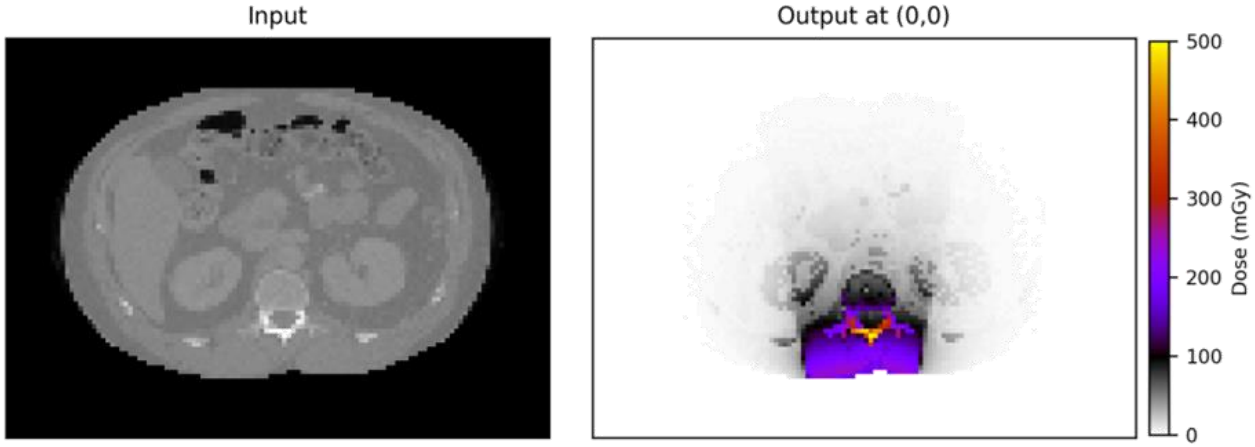


Figure 1: Intermediate slice example of the CT scan image input (left) and 3D dose map output (right) of the proposed DL method

AI-based patient dose calculation

- Deep learning models estimate patient-specific organ and effective dose directly from CT localizer/scout images or acquisition parameters, without a full Monte Carlo run
- Replaces or accelerates Monte Carlo dose simulation, enabling near real-time, size-specific dose estimates (SSDE) tailored to the actual patient anatomy

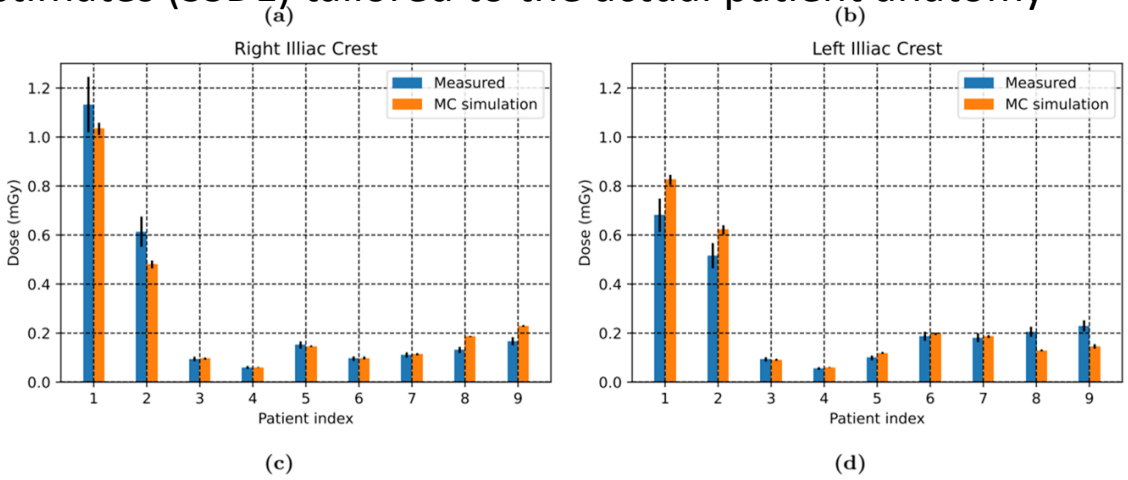


Figure 6: Dose of the 4 anatomical points (Xiphoid Process, Pubic Symphysis, Right Illiac Crest, Left Illiac Crest) for each study patient. We compared the real measurements (blue line) with GGEMS MC simulation (red line).

AI-based acquisition parameter optimization

- Deep learning models recommend or automatically adjust acquisition parameters (kVp, mAs, pitch, collimation) from scout images or patient size, extending classical automatic exposure control

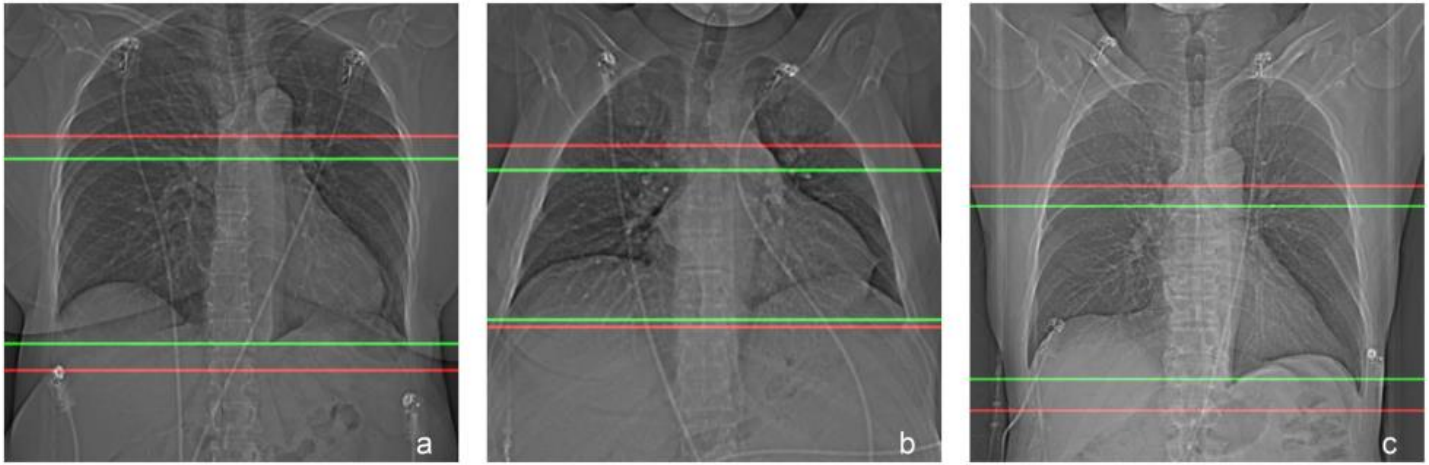
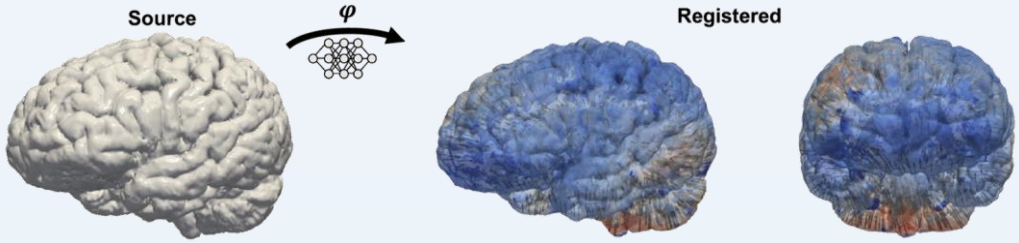
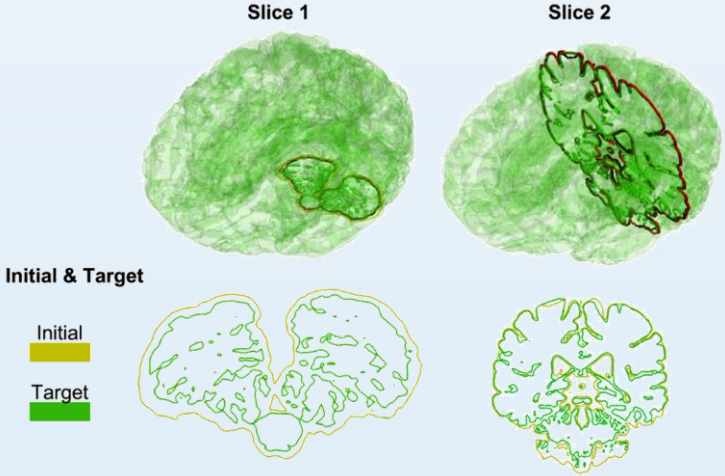


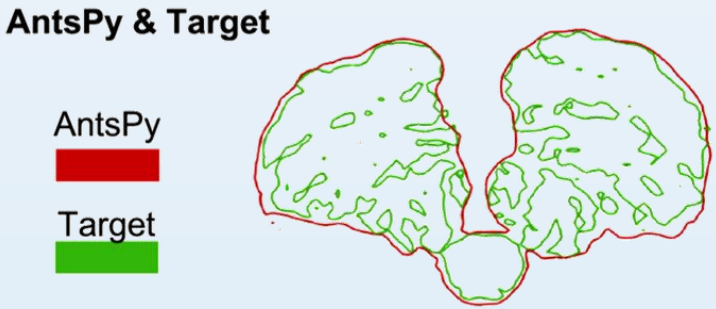
Fig. 2 CT localizers for three patients with a scan range from the radiologists. The radiologists' scan range is depicted in green, and the radiographers in red. **a** The CT localizer of a female patient (55 years). **b** The CT localizer of a male patient (36 years). **c** The CT localizer of a female patient (59 years)

Demircioğlu, Eur. Radiology 2024

Deformable registration







Amiri-Hezaveh, Int. J. of Computer Vision, 2025

Summary

Task	Architecture	Clinical examples
Classification	CNN, ViT	Nodules
Segmentation	U-Net, ViT	OAR, GTV radiotherapy
Detection	YOLO, Faster R-CNN	Lesions
Synthesis (GAN)	CycleGAN, Diffusion	Reconstruction, low dose, denoising
DIR	GAN, Diffusion	Adaptive RT
Reports	FM	Report generation from chest X-ray
Dose calculation	CNN, DL surrogate	Organ/effective dose, SSDE
Acquisition optimization	DL parameter selection	CT, MR

SECTION 06

Interpretability

Model interpretability

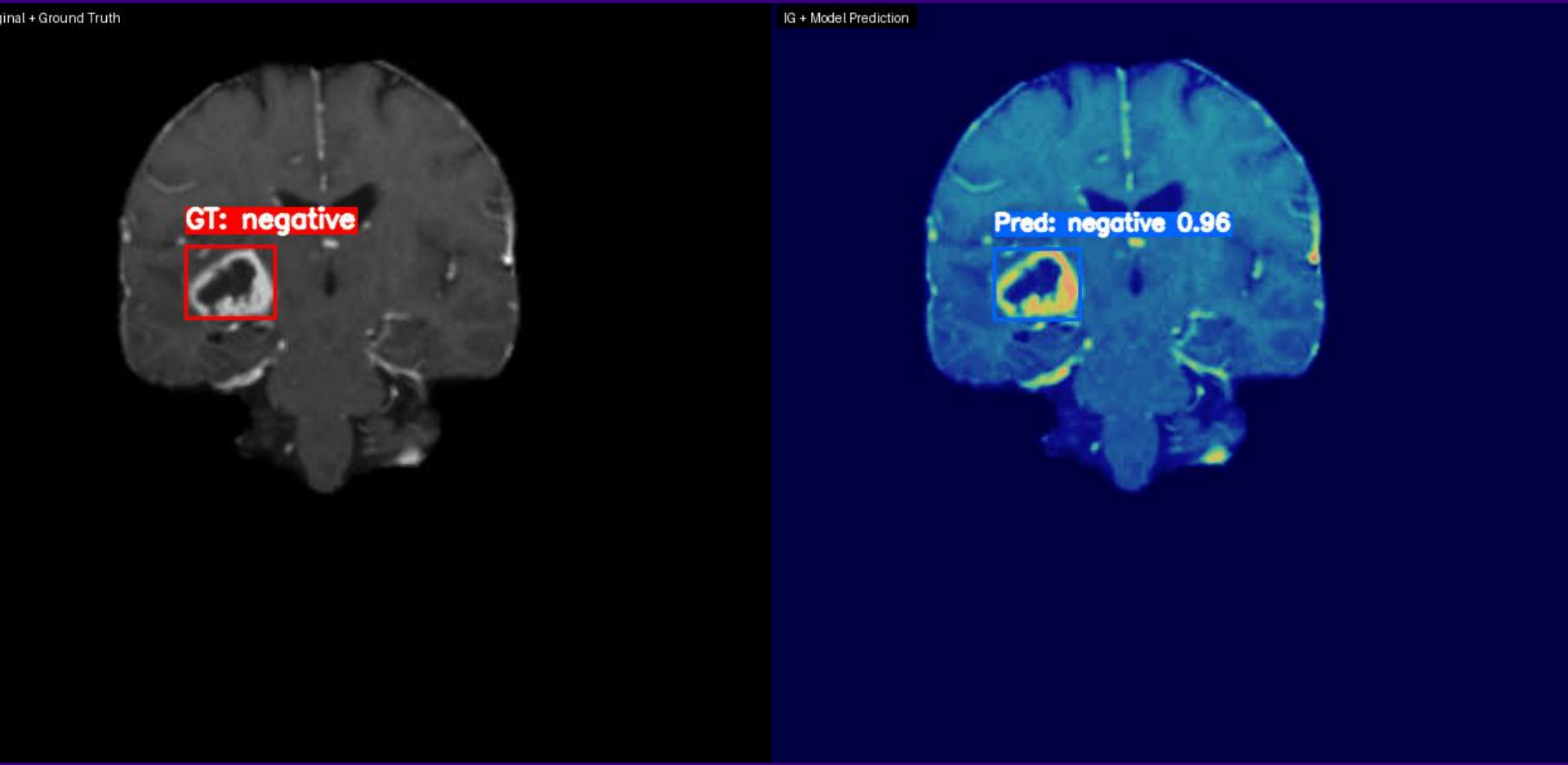
Grad-CAM (CNN)

They visualize the image regions most influential for the model’s decision.

Attention Maps (Transformer)

Indicates which patches/tokens the model considers relevant.

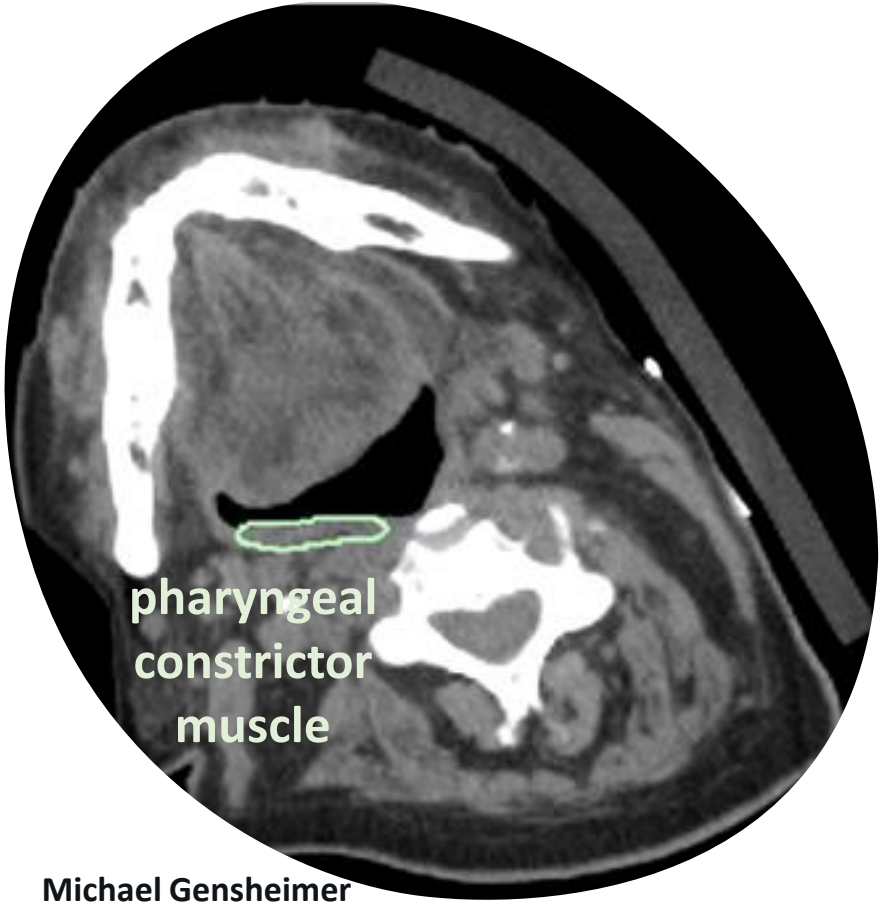
Grad-CAM example in YOLO



SECTION 07

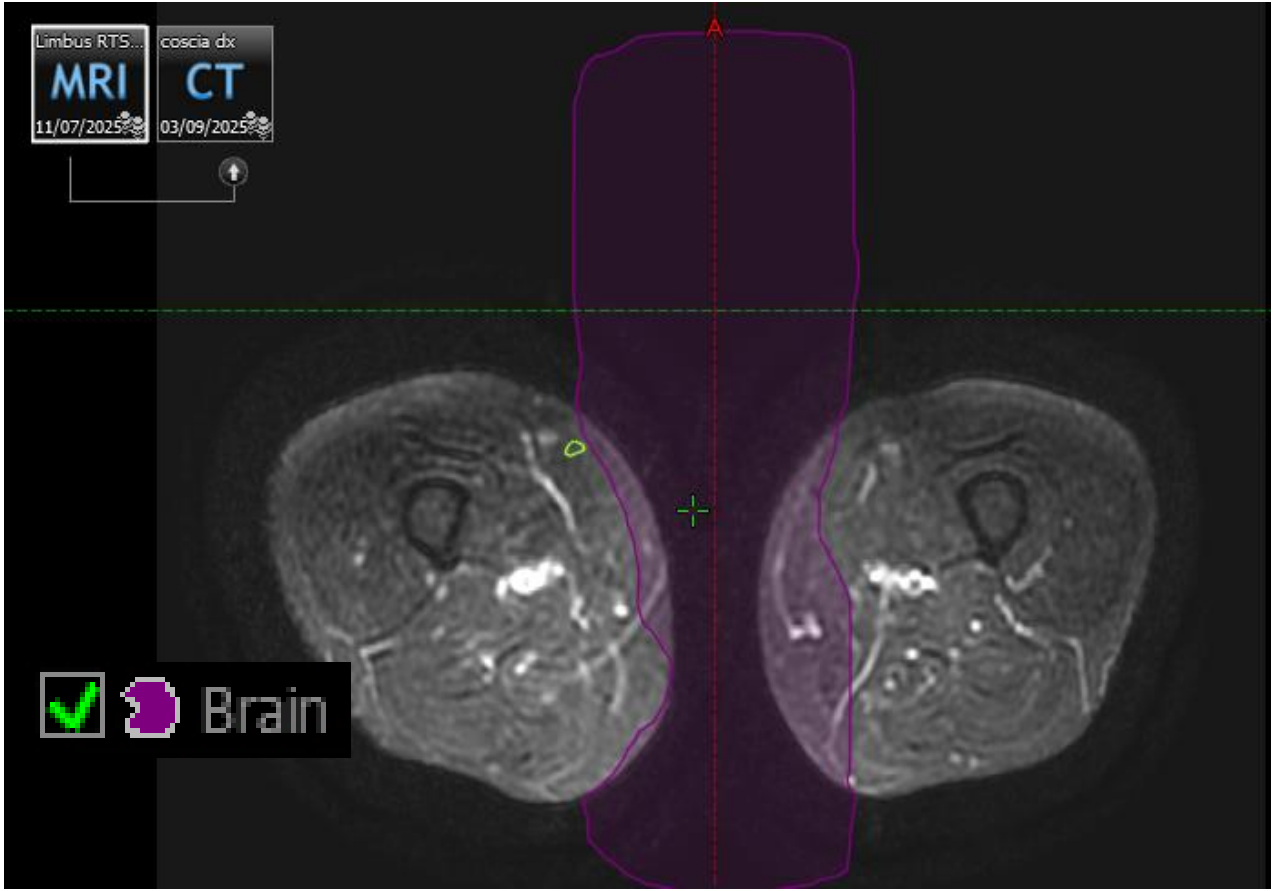
The Role of the Medical Physicist in AI

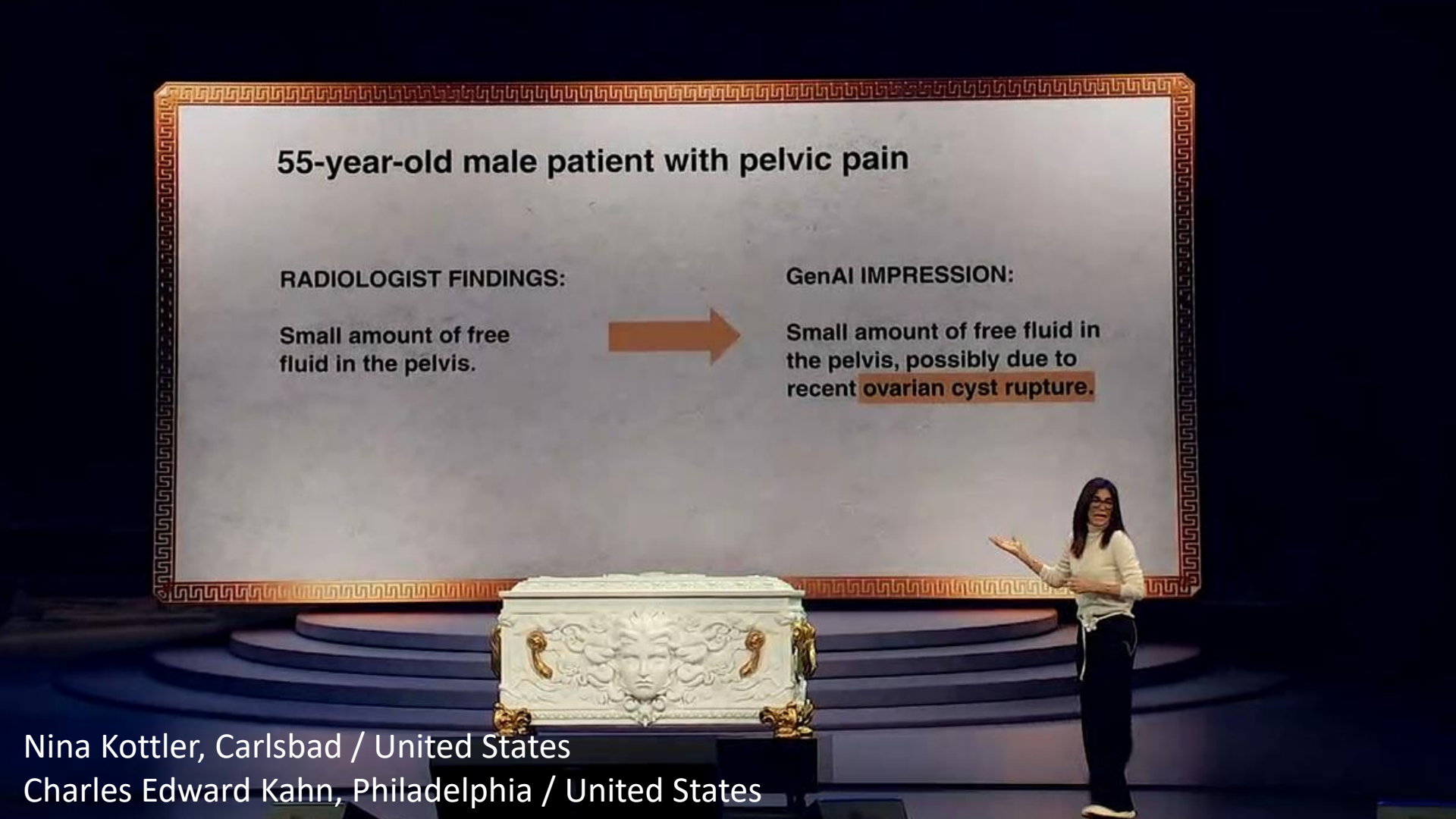
Case examples



Michael Gensheimer
@MFGensheimer

Case examples





Nina Kottler, Carlsbad / United States
Charles Edward Kahn, Philadelphia / United States

Covid diagnosis challenge

- A popular open-source dataset, COVIDx, was used to develop many deep learning tools to diagnose COVID from chest radiographies
- It was found that these models performed poorly on other datasets

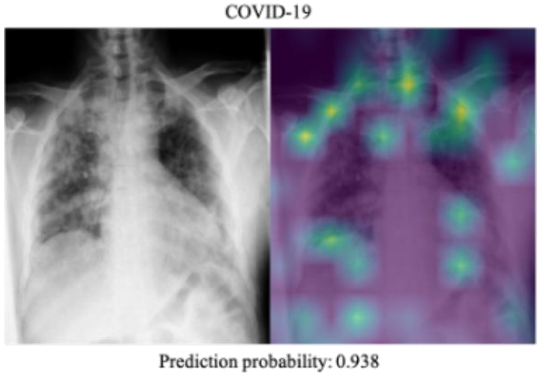


Figure 5: Grad-CAM saliency map generated by DarkCovidNet COVIDx test data predictions. This is a correct prediction of COVID-19, with a prediction probability of 0.938

Table 1– DarkCovidNet test data performance metrics.

Test set	Precision	Recall	F1 score	Accuracy
COVIDx	0.87±0.00	0.80±0.00	0.82±0.00	0.88±0.00
External	0.44±0.00	0.43±0.00	0.41±0.00	0.43±0.00
LTHT	0.47±0.01	0.46±0.00	0.44±0.01	0.45±0.00

Table 2 - CoroNet test data performance metrics.

Test set	Precision	Recall	F1 score	Accuracy
COVIDx	0.81±0.05	0.90±0.01	0.84±0.05	0.88±0.03
External	0.18±0.07	0.34±0.02	0.19±0.03	0.35±0.01
LTHT	0.24±0.01	0.30±0.00	0.15±0.01	0.22±0.00

Table 3 – COVIDNet test data performance metrics.

Test set	Precision	Recall	F1 score	Accuracy
COVIDx	0.86±0.03	0.69±0.05	0.72±0.05	0.86±0.02
External	0.34±0.05	0.36±0.01	0.29±0.02	0.38±0.01
LTHT	0.43±0.01	0.39±0.00	0.37±0.01	0.44±0.03

Covid diagnosis challenge

- the pneumonia class within the COVIDx contains other pathologies, including, pleural effusion, infiltration, consolidation, emphysema and masses.
- The non-COVID data included paediatric chest Xray images which were the only significant source of paediatric images within COVIDx
- the images had different resolutions: 1024x1024 and 299x299. Models resize images (smaller images up-sampled and larger images must be down-sampled). This risks generation of artefacts.
- A dataset had presence of disk-shaped markers in COVID-19 chest X-rays

Bias in the training dataset

- Bias in a dataset may include: confounding factors (e.g. comorbidities) but also dataset imbalance in factors such as gender, sexual orientation, ethnic, social, environmental, or economic factors.
- A biased training dataset produces biased model (“garbage in, garbage out”)

REUTERS

World Business Markets Breakingviews Vi

RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 2 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

SPECIAL REPORT

Check for updates

Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement

SA-CME

J. Raymond Geis, MD[a,b], Adrian P. Brady, MB, FFRRCSI[c], Carol C. Wu, MD[d], Jack Spencer, PhD[e], Erik Ranschaert, MD, PhD[f], Jacob L. Jaremko, MD, PhD[g], Steve G. Langer, PhD[h], Andrea Borondy Kitts, MS, MPH[i], Judy Birch, BEd[j], William F. Shields, JD, LLM[k], Robert van den Hoven van Genderen, PhD, MSc, LLM[l], Elmar Kotter, MSc, MD, MB[m], Judy Wawira Gichoya, MBChB, MS[n,o], Tessa S. Cook, MD, PhD[p], Matthew B. Morgan, MD, MS[q], An Tang, MD, MSc[r], Nabile M. Safdar, MD, MPH[p], Marc Kohli, MD[s]

Can Assoc Radiol J (2019) **70**:329-34 doi: S0846-5371(19)30130-5 [pii].

Other Risks from AI

Table 1 A general framework for considering clinical artificial intelligence (AI) quality and safety issues in medicine

Issue	Summary	Example
Short term		
Distributional shift	A mismatch between the data or environment the system is trained on and that used in operation, due to bias in the training set, change over time, or use of the system in a different population, may result in an erroneous 'out of sample' prediction.	The accuracy of a system which predicts impending acute kidney injury based on other health records data, became less accurate over time as disease patterns changed.[40]
Insensitivity to impact	A system makes predictions that fail to take into account the impact of false positive or false negative predictions within the clinical context of use.	An unsafe diagnostic system is trained to be maximally accurate by correctly diagnosing benign lesions at the expense of occasionally missing malignancy.[6]
Black box decision making	A system's predictions are not open to inspection or interpretation and can only be judged as correct based on the final outcome.	A X-Ray analysis AI system could be inaccurate in certain scenarios because of a problem with training data, but as a black box this is not possible to predict and will only become apparent after prolonged use.[9]
Unsafe failure mode	A system produces a prediction when it has no confidence in the prediction accuracy, or when it has insufficient information to make the prediction.	An unsafe AI decision support system may predict a low risk of a disease when some relevant data is missing. Without any information about the prediction confidence, a clinician may not realise how untrustworthy this prediction is.[46]
Medium term		
Automation complacency	A system's predictions are given more weight than they deserve as the system is seen as infallible or confirming initial assumptions.	The busy clinician ceases to consider alternatives when a usually predictable AI system agrees with their diagnosis.[48]
Reinforcement of outmoded practice	A system is trained on historical data which reinforces existing practice, and cannot adapt to new developments or sudden changes in policy	A drug is withdrawn due to safety concerns but the AI decision support system cannot adapt as it has no historical data on the alternative.
Self-fulfilling prediction	Implementation of a system indirectly reinforces the outcome it is designed to detect.	A system trained on outcome data, predicts that certain cancer patients have a poor prognosis. This results in them having palliative rather than curative treatment, reinforcing the learnt behaviour.

The accuracy of a system which predicts impending acute kidney injury based on other health records data, became less accurate over time as disease patterns changed.[40]

An unsafe diagnostic system is trained to be maximally accurate by correctly diagnosing benign lesions at the expense of occasionally missing malignancy.[6]

A system trained on outcome data, predicts that certain cancer patients have a poor prognosis. This results in them having palliative rather than curative treatment, reinforcing the learnt behaviour.

Challen R, et al. BMJ Qual Saf 2019;**28**:231–237. doi:10.1136/bmjqs-2018-008370

Bias isn't fixed at deployment — it can drift

- A model that was fair at validation can develop performance gaps across subpopulations over time, as the real patient population and clinical practice evolve — this is called fairness drift.
- In a national 11-year study, fairness drift emerged in both the original model and models that were periodically updated; updating helped restore fairness in some periods and worsened it in others.
- Practical implication: algorithmic fairness cannot be assured with a one-time check at development — it needs the same kind of ongoing monitoring as any other quality metric, alongside phantom-based and performance QA.

Davis et al., JAMIA 2025

How will AI change the role of the Medical Physicists?

Received: 1 December 2017 | Revised: 1 December 2017 | Accepted: 4 December 2017
DOI: 10.1002/acm2.12244

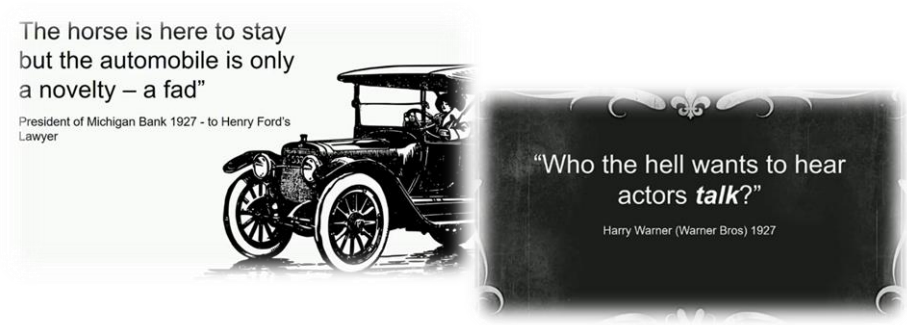
PARALLEL OPPOSED EDITORIAL

WILEY

Artificial intelligence will reduce the need for clinical medical physicists

Xiaoli Tang[1,*] | Brian Wang[2,*] | Yi Rong[3]

[1]Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, West Harrison, NY, USA
[2]Department of Radiation Oncology, University of Louisville, Louisville, KY, USA
[3]Department of Radiation Oncology, University of California Davis Comprehensive Cancer Center, Sacramento, CA, USA



THE MEDICAL PHYSICS CONSULT

MAHADEVAPPA MAHESH, MS, PHD, RICHARD L. MORIN, PHD

Check for updates

Role of the Medical Physicist in the Health Care Artificial Intelligence Revolution

William F. Sensakovic, PhD, Mahadevappa Mahesh, MS, PhD

Artificial intelligence will soon change the landscape of medical physics arch and practice

Lei Xing, Ph.D.
Department of Radiation Oncology, Stanford University, Stanford, CA 94305, USA
(Tel: 650-498-7896; E-mail: lei@stanford.edu)

Elizabeth A. Krupinski, Ph.D.
Department of Radiology & Imaging Sciences, Emory University, Atlanta, GA 30322, USA
(Tel: 404-712-3868; E-mail: ekrupin@emory.edu)

Jing Cai, Ph.D., Moderator

(Received 12 February 2018; revised 19 February 2018; accepted for publication 19 February 2018; published 13 March 2018)

[https://doi.org/10.1002/mp.12831]

"**AI won't replace radiologists**, but radiologists who use AI will replace radiologists who don't,"
Curtis Langlotz, a radiologist at Stanford.

IAEA Human Health Series No. 50 (2026)

- Clinical Implementation of AI Systems in Medical Imaging and Radiotherapy — Guidelines for Medical Physicists
- Aimed at clinically qualified medical physicists (CQMPs): the professionals who bridge complex AI systems and clinical practice
- Covers the whole life cycle — from assessing needs to selection, commissioning, quality management and decommissioning

Life cycle of AI

- Identification of organizational needs
- Market research
- Preselection and demonstration
- Selection and purchasing
- Installation
- Acceptance and commissioning
- Introduction into the clinical setting
- Quality management
- Clinical evaluation and impact assessment
- Decommissioning

The medical physicist's role — selection & commissioning

- Translate clinical needs into technical specifications for procurement and IT integration
- Collaborate in market research, preselection and vendor demonstrations against the intended use
- Lead selection and purchasing with clear, testable acceptance criteria
- Perform acceptance testing and commissioning; establish baseline performance
- Verify data quality, interoperability and integration into the existing workflow

The medical physicist's role — QA and monitoring

- Design and run the quality management programme: acceptance, commissioning, routine performance testing
- Monitor the AI system's impact on users, decision-making and clinical workflow
- Detect performance drift as patient populations, protocols and equipment change
- Lead risk assessment (e.g. FMEA) and incident reporting
- Govern updates and retraining; advise on deviations; decide on safe decommissioning

Steps of QA program

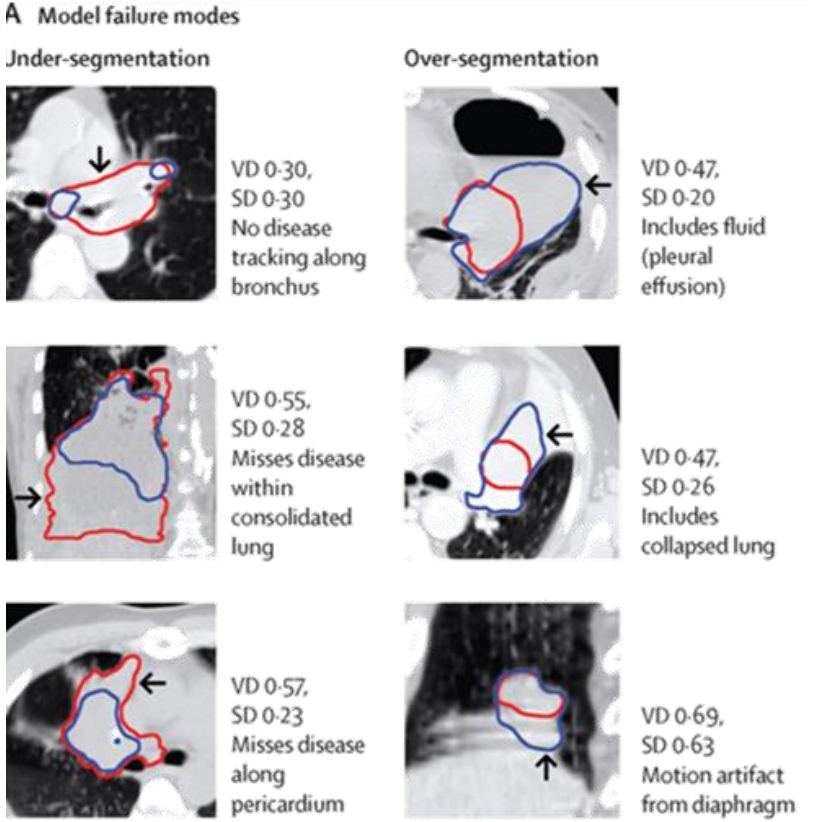
In general, a QA program should include four steps:

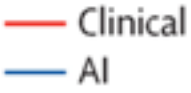
- AT of newly installed tools, which is typically more rigorous than the ongoing routine QC tests.
- Determination of baseline performance.
- Ongoing monitoring of the tools to ensure early detection of any changes in performance.
- Periodic re-validation to verify performance after any changes to the workflow that could impact the AI tool output.

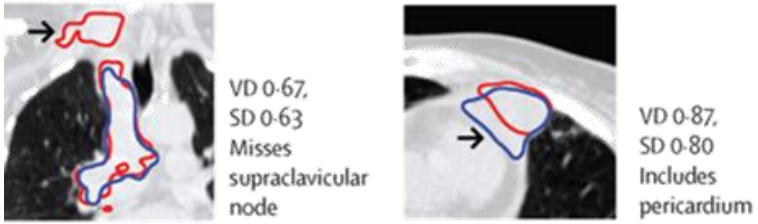
A QC could be performed using:

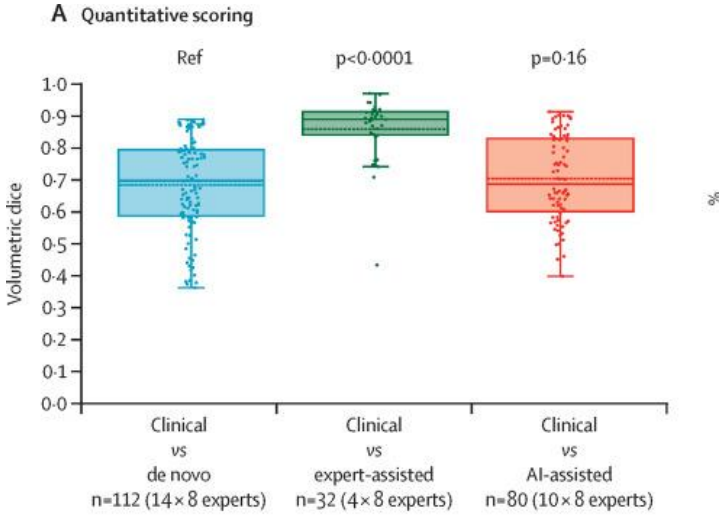
- Physical Phantom
- Digital Phantom
- A retrospective patients dataset

Analysis of retrospective patients / failure modes









Hosny, Lancet Digit Health 2022

QA of AI based registration

- The accuracy of image registration can be validated and verified qualitatively using visualization tools
- Compute metrics such as target registration errors, mean distance to agreement, dice similarity coefficient

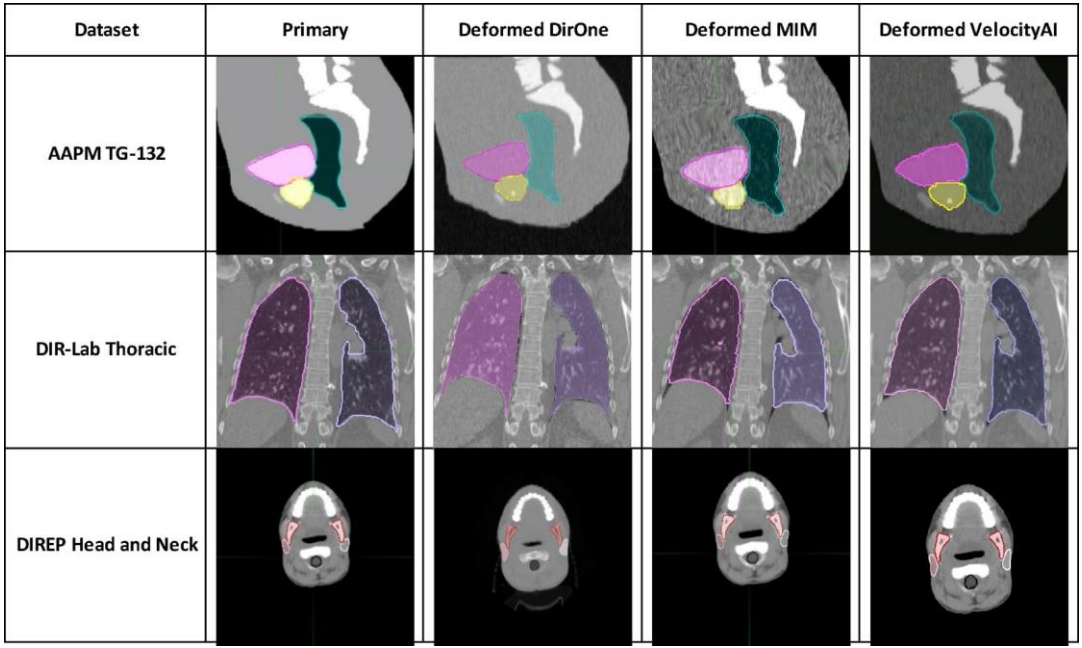
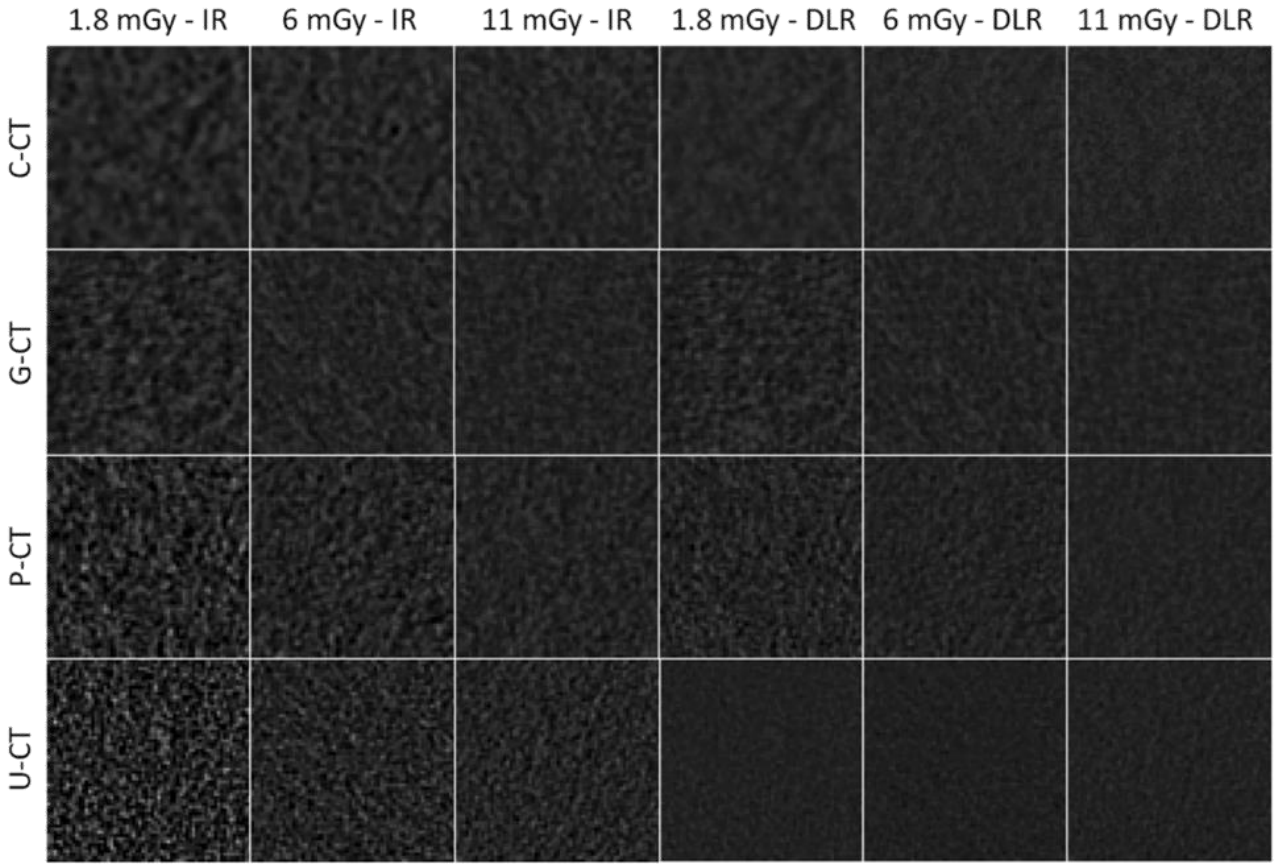
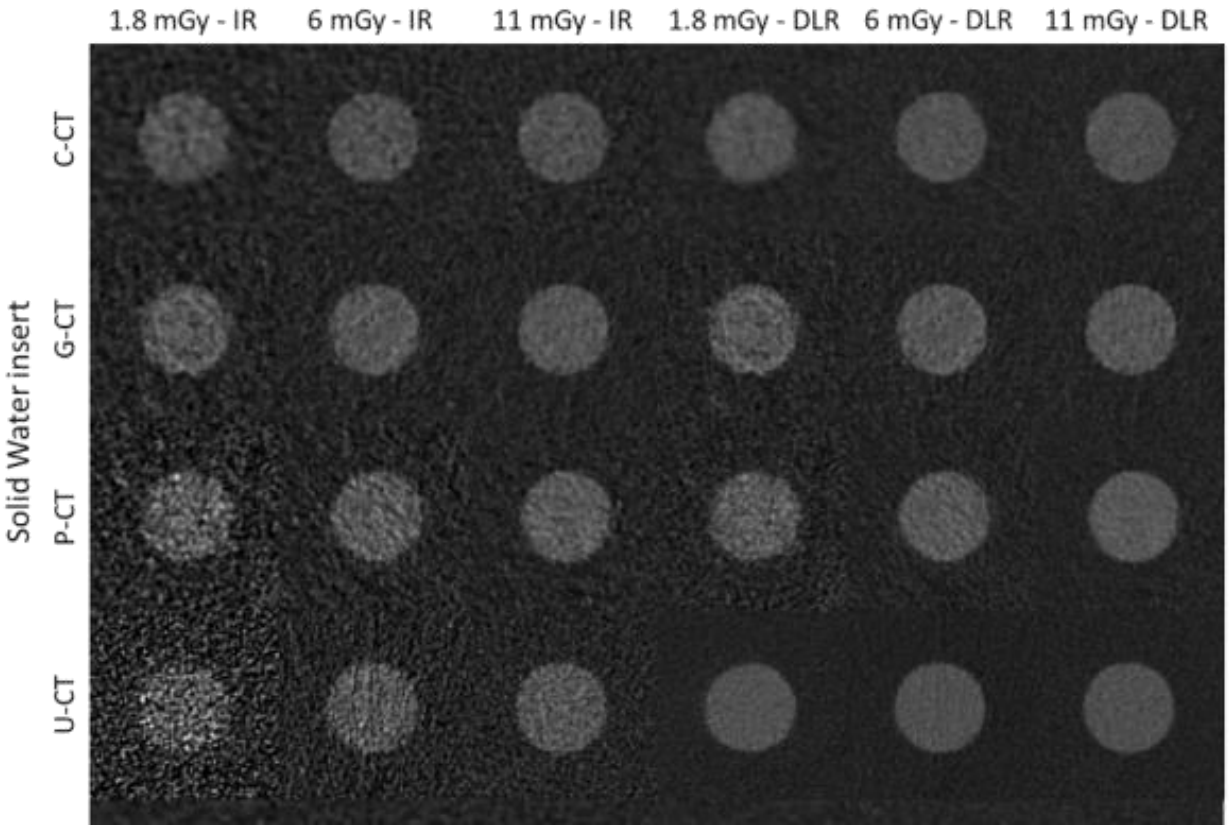


Image reconstruction using deep learning (DLR) – in phantom assessment



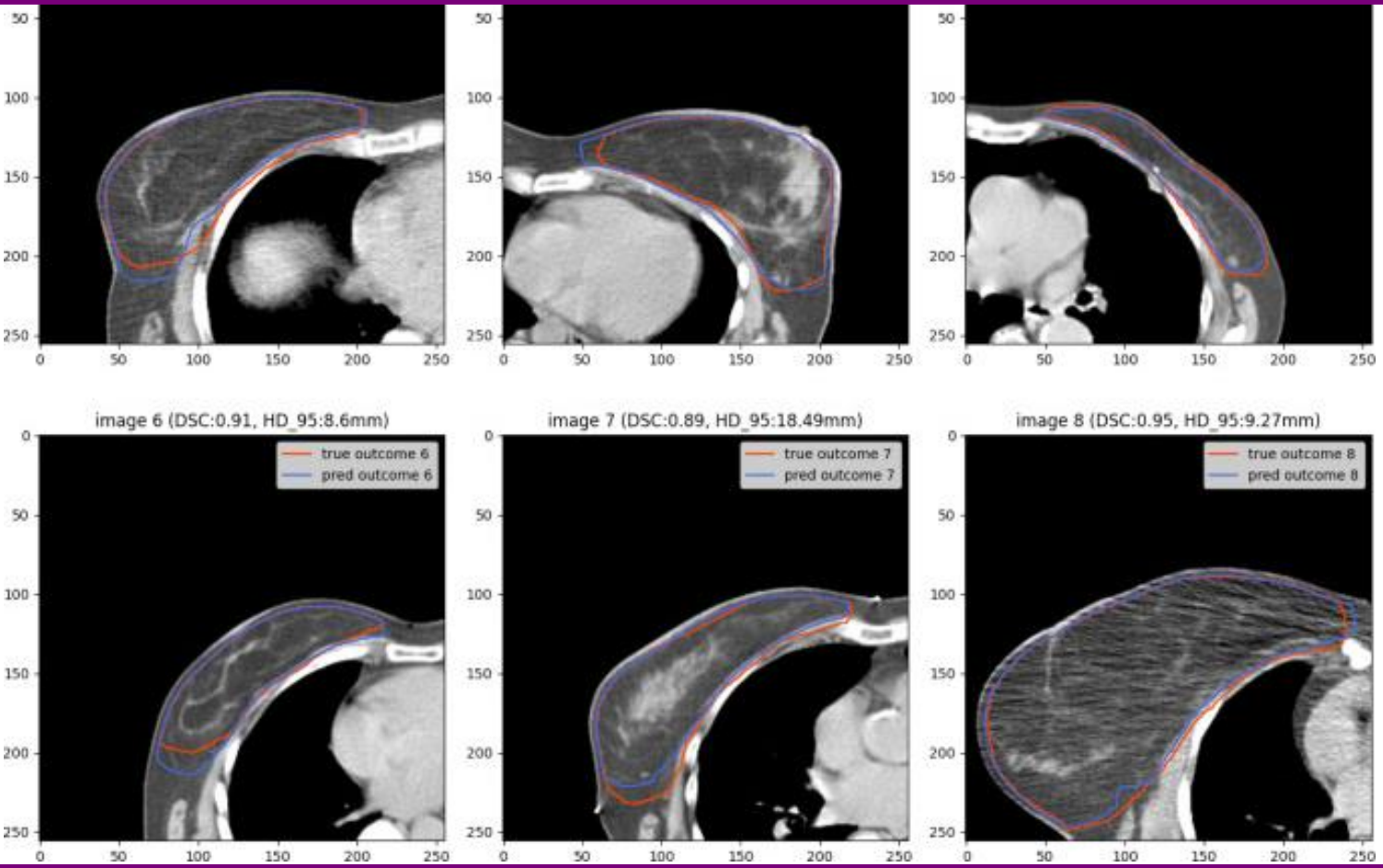
Greffier, Diagnostic and Interventional Imaging 2026

Image reconstruction using deep learning (DLR) – in phantom assessment

- 

Greffier, Diagnostic and Interventional Imaging 2026

QA of automated segmentation



Ubeira Gabellini, Physics and Imaging in Radiation Oncology, 2025

Failure Mode and Effect Analysis

S=severity
D=detectability
O=occurrence
R=risk score=S*D*O

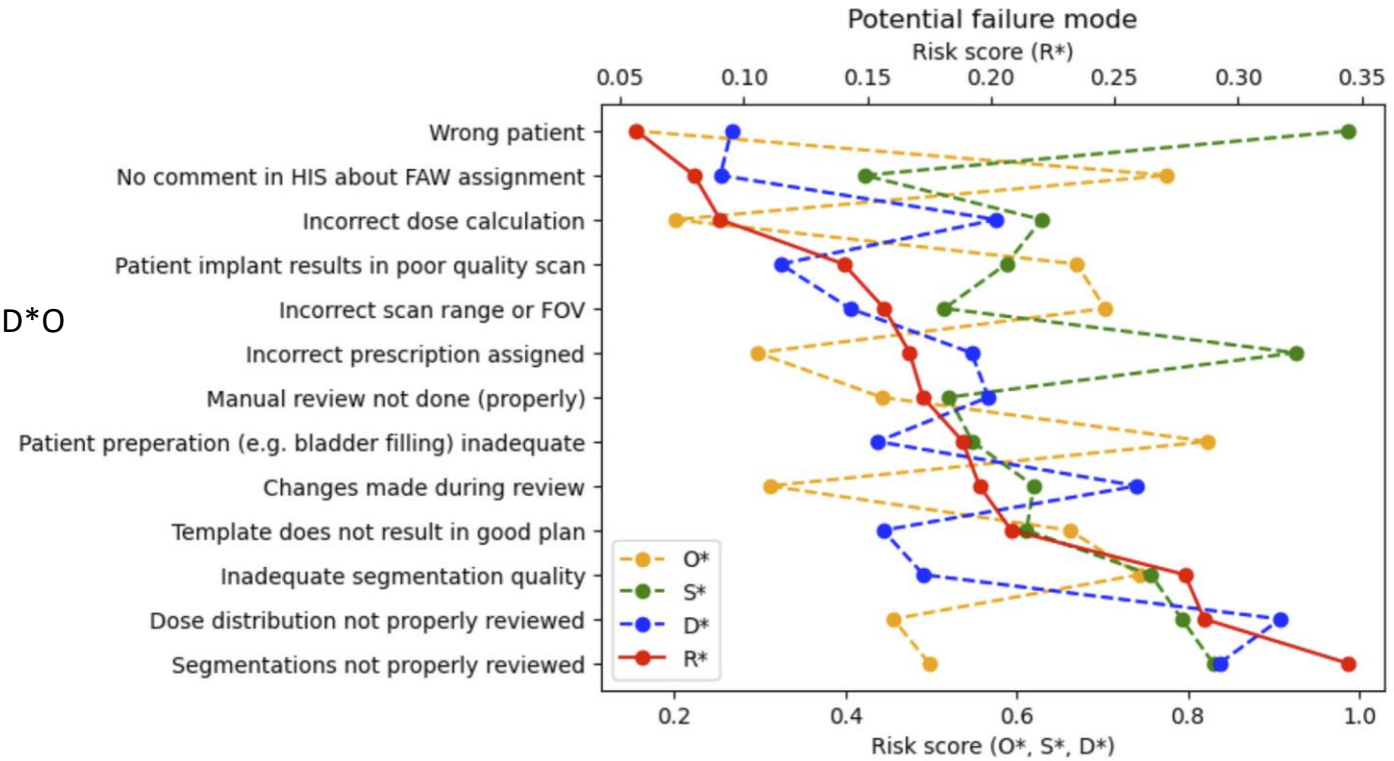
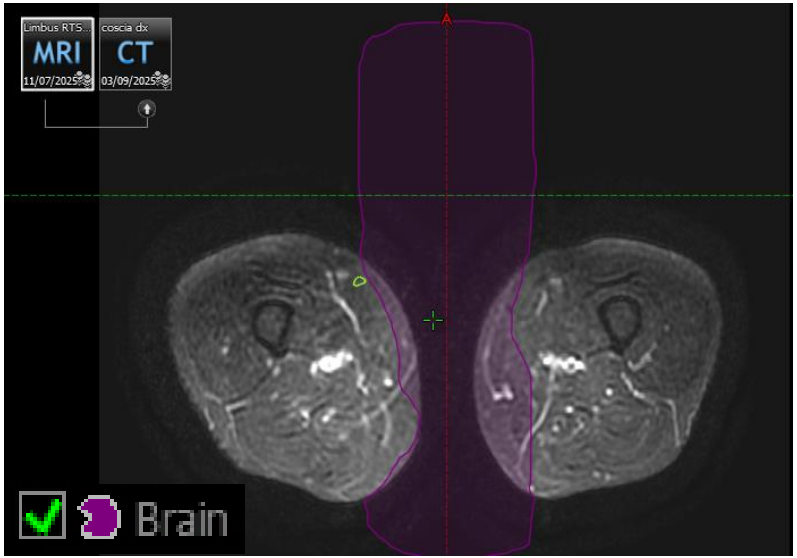


Fig. 3. Top five highest-scoring values for R*, O*, S* and D*and its corresponding failure mode, ordered by increasing R*. Due of overlap, the number of failure modes was limited to thirteen.

Garbage in: AI doesn't know when the input is wrong

- We've seen an auto-contouring AI tool, given a CT scan of the legs, still draw a “brain” contour — floating in mid-air. It had no idea the image was wrong for its job. It just ran.
- Most AI tools don't check whether the input even makes sense for them — that check has to come from us, before we trust the output.



Weng et al., Clinical and Translational Science 2025 (FDA)

AI fails: Whose fault is it?

Accountability: AI cannot be held responsible

- A joint statement by radiology's major professional societies — including AAPM — is direct on this point: “Ultimate responsibility and accountability for AI remains with its human designers and operators.”
- An AI system is not a legal or moral agent — it can produce an error, but it cannot itself be held accountable for one. Responsibility for what happens with its output stays with the people who build, deploy, and use it.
- For clinical practice specifically, the same statement is explicit: “Radiologists will remain ultimately responsible for patient care.”

Geis et al., Radiology 2019 (joint ACR/ESR/RSNA/SIIM/ESMI/CAR/AAPM multisociety statement)

Case-specific QA: checking every single output

- Every output an AI-based application produces — a CAD finding, a triage flag, an auto-generated report — should be checked before it is used clinically: is this particular output trustworthy? This is case-specific QA.
- If the check is satisfactory, the output continues in the normal workflow. If it fails, the output goes to a second verification step before anyone acts on it.
- In practice, the most common form of second verification is human in the loop review of the output combined with quantitative or qualitative checks. Some applications instead benchmark against an independent secondary algorithm, or use automated methods designed to flag divergent behavior.

Framework originally described for radiotherapy AI — the case-specific / second-verification structure applies directly to radiology AI as well.

Routine QA: is the model still working, month after month?

- Routine QA periodically re-tests the model itself, to catch problems that only become visible over time — after a software update, a new scanner, or a shift in the patient population.
- When routine QA results fall short of expectations, the model may need re-commissioning: re-validating it before it is trusted in the clinic again.
- AI models in medical applications may have a “half-life” of data relevance of around 4 months (Chen et al. 2017) — a reminder that routine QA is not an occasional formality, but a recurring necessity

Multidisciplinary team

- A fundamental role of the medical physicist is multidisciplinary team member — working together with physicians, technologists, nurses, therapists, engineers, and even patients in the effective, efficient, and safe delivery of health care»
- another key role of the medical physicist is education and training in AI, not only of junior level and medical physicists in training but also of other health care professionals including residents and fellows

Artificial intelligence will soon change the landscape of medical physics research and practice

Lei Xing, Ph.D.
Department of Radiation Oncology, Stanford University, Stanford, CA 94305, USA
(Tel: 650-498-7896; E-mail: lei@stanford.edu)

Elizabeth A. Krupinski, Ph.D.
Department of Radiology & Imaging Sciences, Emory University, Atlanta, GA 30322, USA
(Tel: 404-712-3868; E-mail: ekrupin@emory.edu)

Jing Cai, Ph.D., Moderator

(Received 12 February 2018; revised 19 February 2018; accepted for publication 19 February 2018; published 13 March 2018)

[https://doi.org/10.1002/mp.12831]

CV of medical physicist

- AI in medical imaging is becoming a core skill for medical physicists — it belongs in the CV, alongside dosimetry and radiation protection
- Medical physicists must also know the regulations governing AI-based medical devices in their own country or region

Expanding the medical physicist curricular and professional programme to include Artificial Intelligence

F. Zanca [a,*,1], I. Hernandez-Giron [b], M. Avanzo [c], G. Guidi [d], W. Crijns [e], O. Diaz [f], G.C. Kagadis [g], O. Rampado [h], P.I. Lønne [i], S. Ken [j], N. Colgan [k], H. Zaidi [l], G.A. Zakaria [m], M. Kortesniemi [n]

ELSEVIER

Regulatory Aspects of the Use of Artificial Intelligence Medical Software

Federica Zanca,* Caterina Brusasco,[†,‡] Filippo Pesapane,[§] Zuzanna Kwade,[‖] Ruth Beckers,[¶] and Michele Avanzo**

IAEA
International Atomic Energy Agency

Artificial Intelligence in Medical Physics

Roles, Responsibilities, Education and Training of Clinically Qualified Medical Physicists

Endorsed by the American Association of Physicists in Medicine

Conclusions and Take-Home Messages

- ✓ AI models can perform a great variety of tasks in Radiology
- ✓ Layers, metrics and architectures are essential to understand the models
- ✓ The medical physicist must acquire these tools to fulfill their role in the AI era.
- ✓ The medical physicist is uniquely placed to make AI adoption safe, effective and sustainable across its whole life cycle

Try it yourself: Lung Cancer Classification

Train your own CNN to tell cancer from normal chest CT images — tune epochs, depth, learning rate, and watch the ROC curve update live.

Open in Colab →

Scan to open

https://colab.research.google.com/github/MedPhysTools/ICTP/blob/main/ICTP_lung_cancer_lab.ipynb

Lab Outline: Lung Cancer Classification

1. Download the dataset and look at real chest CT images — check class balance and image quality before trusting any model built on it.

2. Build your own train / validation / test split, grouped by patient to avoid data leakage.

3. Train a CNN with adjustable settings (epochs, depth, learning rate, batch size) — it starts weak on purpose; your job is to make it better.

4. Read the ROC curve and the error-analysis collage — see exactly which cases the model gets wrong, and which kind of mistake it makes.

5. Log your result and compare it with the rest of the class on the shared dashboard.

Try it yourself: Brain Tumor Detection

Train YOLO26 to detect brain tumors on MRI, and explore the confidence / IoU trade-off behind every detection — including the false positives and negatives.

Open in Colab →

Scan to open

https://colab.research.google.com/github/MedPhysTools/ICTP/blob/main/ICTP_brain_tumor_lab.ipynb

Lab Outline: Brain Tumor Detection (YOLO)

1. Train YOLO26 to detect brain tumors on MRI slices.

2. Explore the confidence and IoU sliders on real images — watch boxes turn from true to false positives as the thresholds change.

3. Find the false negatives — the tumors your current settings are missing entirely.

4. Read the FROC curve, the detection equivalent of ROC: sensitivity vs. false positives per image.

5. Log your point and watch the whole class's FROC curve take shape together.

Try it yourself: Spleen Segmentation

Train a 3D U-Net with MONAI to segment the spleen on abdominal CT, and see the network architecture change live as you adjust depth and width.

Open in Colab →

Scan to open

https://colab.research.google.com/github/MedPhysTools/ICTP/blob/main/ICTP_spleen_segmentation_lab.ipynb

Lab Outline: Spleen Segmentation (3D U-Net)

1. Download the spleen CT dataset and see why every scan needs to be standardized before training.

2. Train your own 3D U-Net — pick depth, neurons, learning rate, CPU or GPU — and watch the network diagram redraw itself live as you change them.

3. Watch the Dice score curve rise — and see why it can start near zero before catching up.

4. Compare a few predictions directly against the ground truth, slice by slice.

5. Log your result and see how the whole class's settings compare on the shared dashboard.