Trustworthy Deep Learning in critical applications: a biomedical experience

AI LAB seminar 25/03/2022 - INTESA SANPAOLO INNOVATION CENTER

Prof. Marco Grangetto

Dipartimento di Informatica
Università di Torino
marco.grangetto@unito.it

Outline

- Deep Learning and critical applications: the scenario
- Neural Network representations (background)

- Interpretability by model design
- Representation learning
- Simple is better: prune to generalize

- Conclusions and discussion

I

The scenario

The Deep Learning hype

- Lin, Henry W., Max Tegmark, and David Rolnick. "**Why does deep and cheap learning work so well?."** Journal of Statistical Physics (2017)

“ Deep learning works remarkably well, and has helped dramatically improve the state-of-the-art in areas ranging from speech recognition, translation and **visual object recognition** to drug discovery, genomics and automatic game playing ”

The Deep Learning hype

- Lin, Henry W., Max Tegmark, and David Rolnick. "**Why does deep and cheap learning work so well?.**" Journal of Statistical Physics (2017)

“neural networks are understood only at a heuristic level, where we empirically know that certain training protocols employing large data sets will result in excellent performance… we know that if we train a child according to a certain curriculum, she will learn certain skills but we lack a deep understanding of how her brain accomplishes this”

AI Act, il Parlamento europeo approva la prima legge al mondo sull'intelligenza artificiale

di Francesca Basso

Con 523 voti a favore, si è concluso il lungo iter legislativo per provare a regolamentare (per la prima volta) le applicazioni di intelligenza artificiale. La legge entrerà ufficialmente in vigore tra due anni

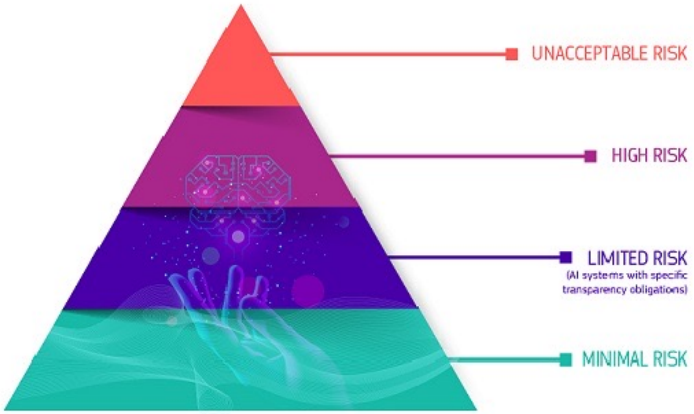
CORRIERE TV

Arriva la prima legge europea che regola l'Intelligenza Artificiale: cosa sono i 4 livelli di rischio

Corriere.it 13 marzo 2024

EU act: AI risk assesment

EU AI act identified as **high-risk AI** technology used in:



- critical infrastructures (e.g. transport), that could put the life and health of citizens at risk;
- educational or vocational training, that may determine the access to education and professional course of someone's life (e.g. scoring of exams);
- safety components of products (e.g. AI application in robot-assisted surgery);
- employment, management of workers and access to self-employment (e.g. CV-sorting software for recruitment procedures);
- essential private and public services (e.g. credit scoring denying citizens opportunity to obtain a loan);
- law enforcement that may interfere with people's fundamental rights (e.g. evaluation of the reliability of evidence);
- migration, asylum and border control management (e.g. verification of authenticity of travel documents);
- administration of justice and democratic processes (e.g. applying the law to a concrete set of facts).

EU act: AI risk assessment

High-risk AI systems will be subject to **strict obligations** before they can be put on the market:

- adequate risk assessment and mitigation systems;
- **high quality of the datasets** feeding the system to minimise risks and discriminatory outcomes;
- **high level of robustness, security and accuracy**
- and others..

Trustworthy

How do we enforce high level of robustness, security and accuracy?

How do we **convey trust**?

Biomedical imaging

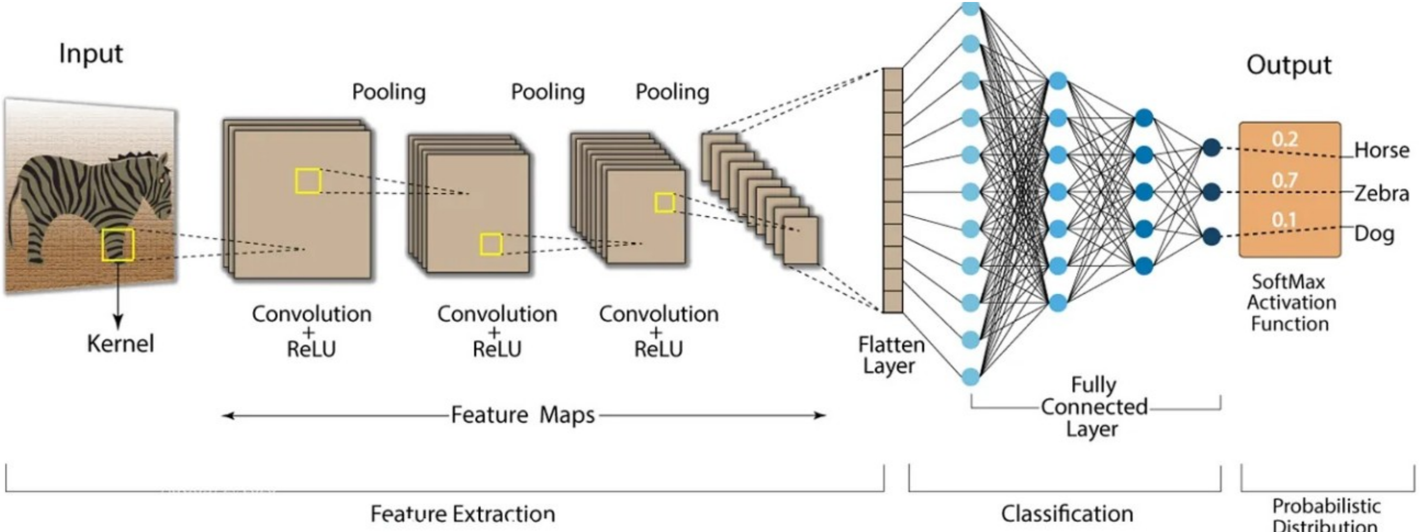
- Computer-aided diagnosis (**CADx**) assists doctors in the interpretation of medical images (X-ray, MRI, Endoscopy, and ultrasound, etc.)
- CAD is an interdisciplinary tech. combining elements of artificial intelligence and computer vision with radiological and pathology image processing

Neural Network representations (background)

NN as information bottleneck

- R. Shwartz-Ziv, N. Tishby "Opening the black box of deep neural networks via information", 2017

Map to representative representation to generalize



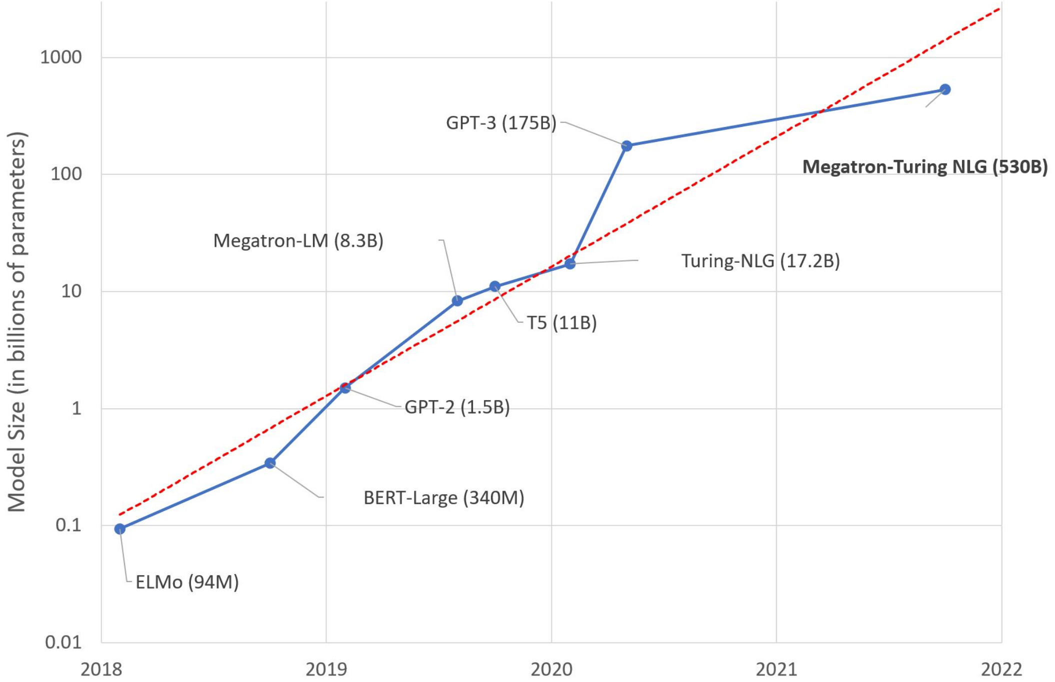
Neural networks learn a lower dimensional representation of data. Credits to Haque 2023.

UNIVERSITÀ DI TORINO

EIDOSLAB

The complexity culprit

- Neural models' size and complexity are increasing
- more data needed, how do we enforce data quality
- The black-box effect is amplified



by Julien Simon, 2021

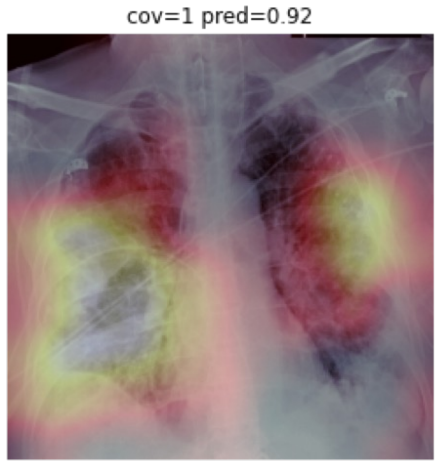
II

Interpretability by model design

Interpretability

- The model can be rather complex and obscure
- Trust can be conveyed by the interpretability of the results or decision mechanics
- Interpretability approaches:
    - post-hoc: methods that analyze the model after training
    - intrinsic: constraints imposed on the model (structure, regularization, etc.)

Post-hoc: GRADCAM



Risks

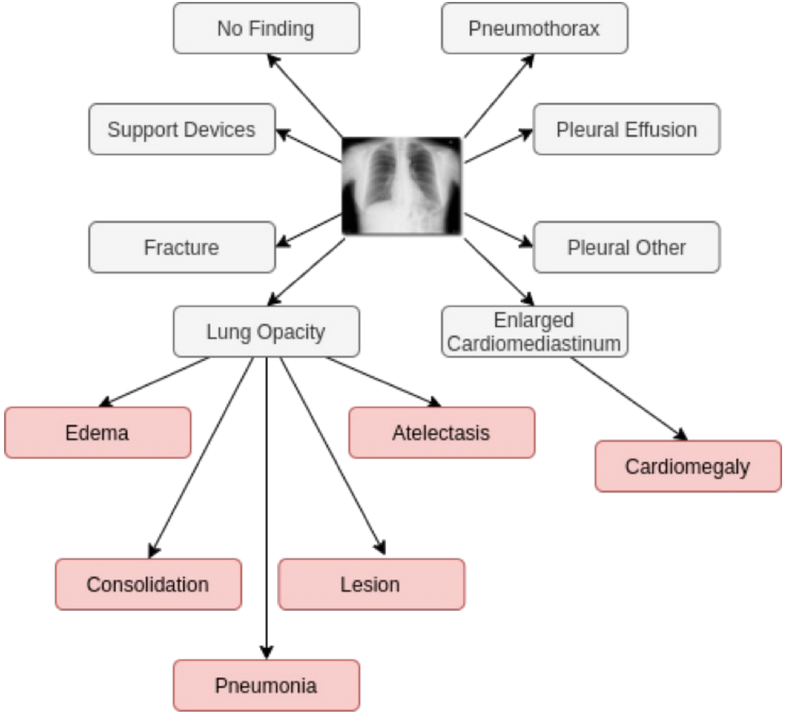
- Low resolution (due to interpolation of the feature maps
- Evidence-based per single image (not a demonstration)
- Amplify artifacts learned by the model instead of actual knowledge from the data

Intrinsic interpretability: the radiology use case

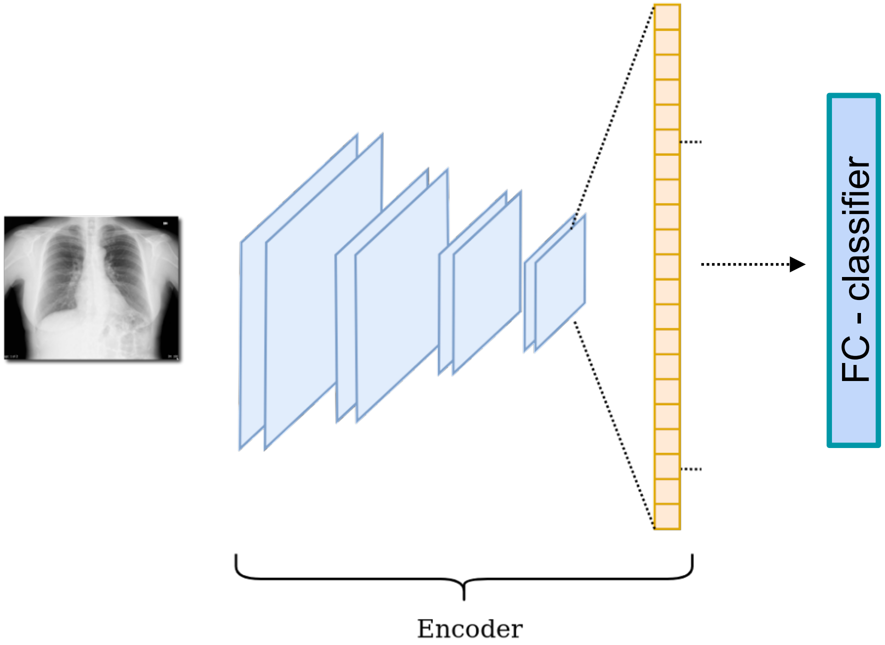
How do we include the additional radiologic expertise?

CheXpert: a large dataset comprising about 224k CXRs.

This dataset consists of 14 different observations on the radiographic image: differently from many other datasets which are focused on disease classification based on clinical diagnosis, the main focus here is “chest radiograph interpretation”, where anomalies are detected.

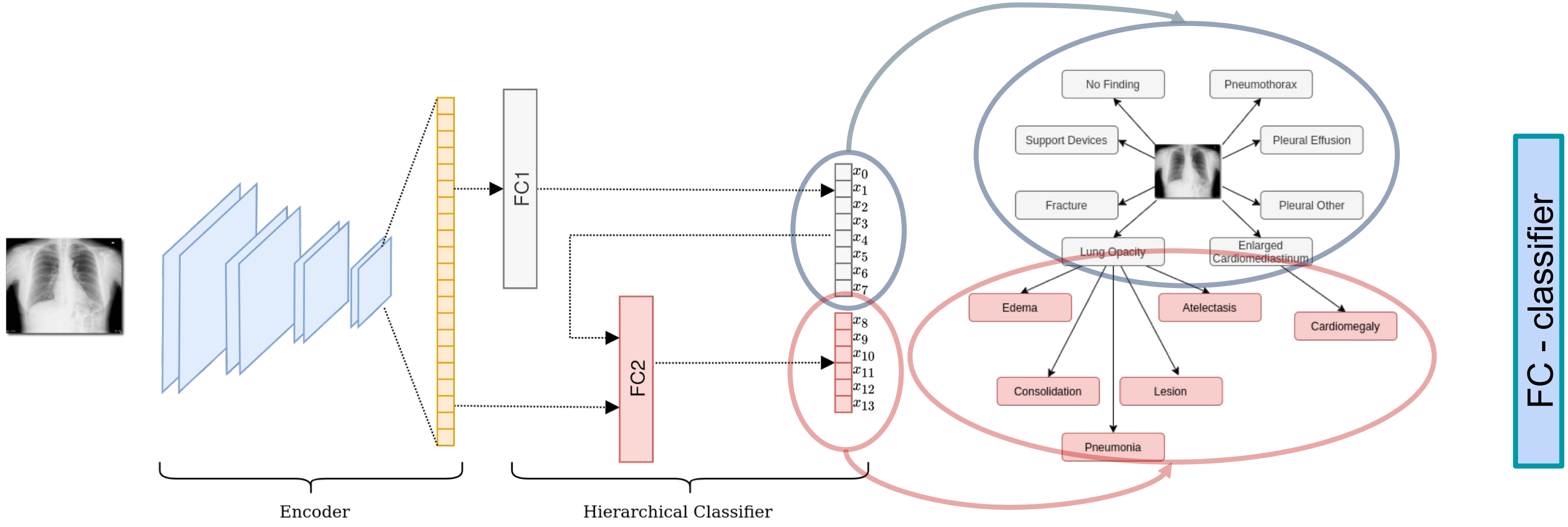


Building a COVID-19 classifier



- A standard approach based on Convolutional Neural Network
  - Low control on the learning process
  - Trust training CXR data

Redesigned for interpretability



Barbano, Carlo Alberto, et al. "A Two-Step Radiologist-Like Approach for Covid-19 Computer-Aided Diagnosis from Chest X-Ray Images

III

Representation learning

Collateral Learning: domain shift

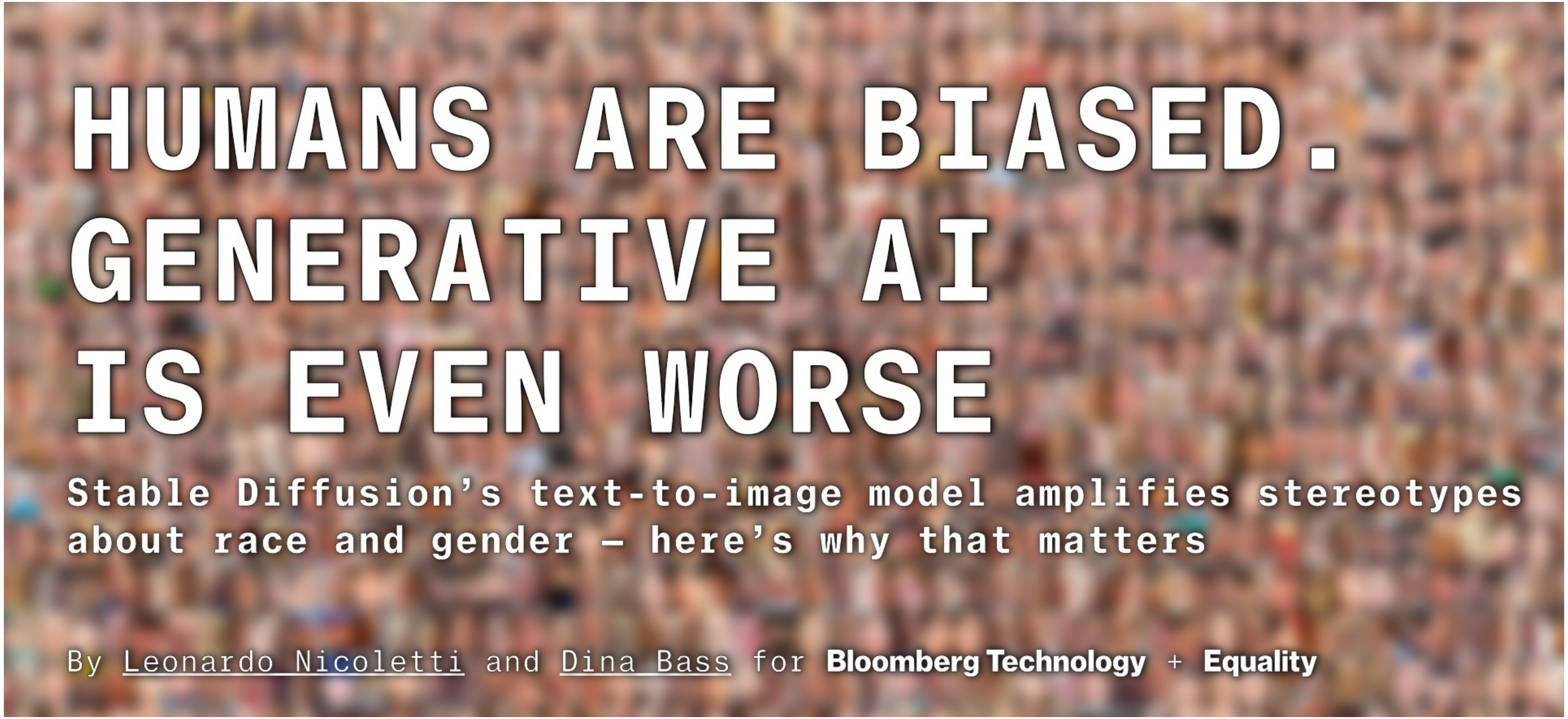
- Learning aims at representations that should be **Robust** to confounding factors and spurious information in the data
- **Collateral Learning** occurs when a model learns more information than intended

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

Beery et al., “Recognition in Terra Incognita”, ECCV 2018

Collateral learning: bias & fairness



Collateral Learning: privacy preservation


The Guardian

Sport Culture Lifestyle More

Asia Australia Middle East Africa Inequality Global development

Italy's privacy watchdog bans ChatGPT over data breach concerns

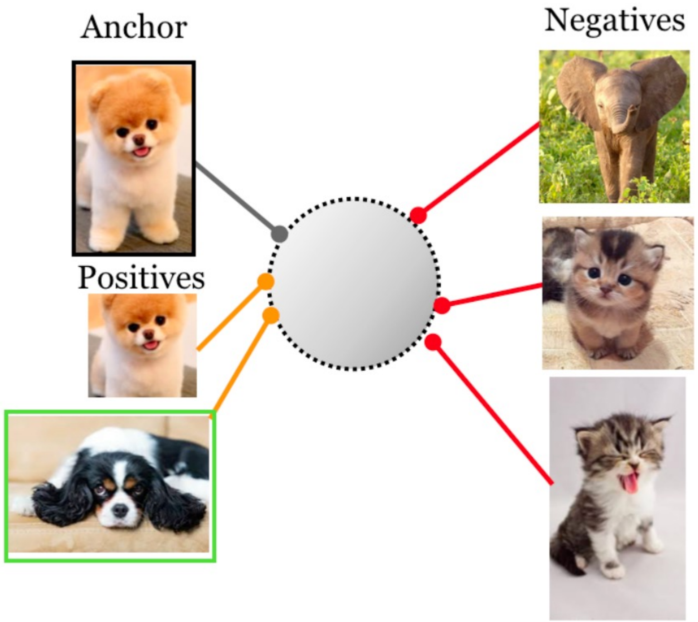
Measure is in place 'until ChatGPT respects privacy', says Italian Data Protection Authority

Fredrikson, Jha, Ristenpart. "**Model inversion attacks** that exploit confidence information and basic countermeasures." ACM SIGSAC 2015

Unbiased representation learning

- Contrastive learning (background)
  - pull together an anchor and a “positive” sample in embedding space
  - push apart the anchor from many “negative” samples

$$\mathcal{L}_{i,j}^{SupCon} = -\frac{1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k \neq i} \exp(sim(z_i, z_k)/\tau)}$$



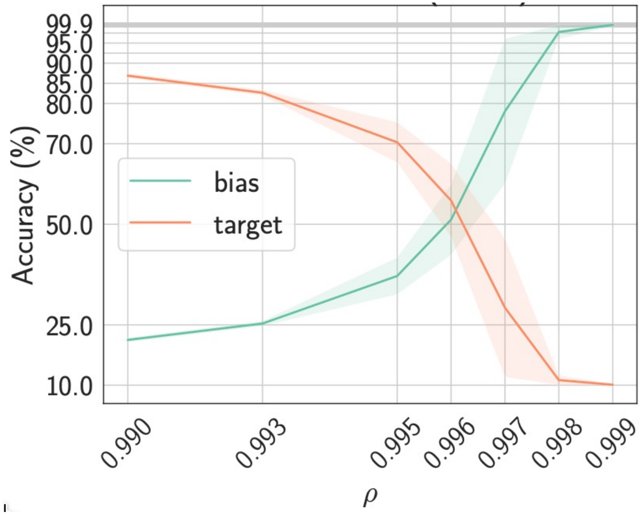
Khosla, Prannay, et al. "Supervised contrastive learning." in Neurips 2020.

Unbiased representation learning

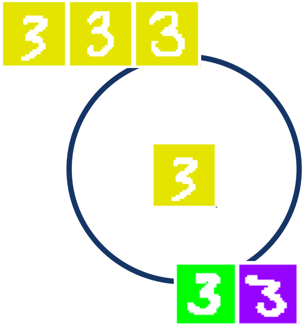
Training on biased MNIST

Testing on unbiased data

- Higher bias / Lower classification accuracy



Unbiased Contrastive learning setup



- Let's constrain the distributions of similarity value of bias aligned $B_{+,b}$ and bias conflicting $B_{+,b'}$ samples:

FairKL loss: $\mathcal{R}^{FairKL} = D_{KL}(B_{+,b} || B_{+,b'})$

- Tartaglione, Barbano, Grangetto. "End: Entangling and disentangling deep representations for bias correction." CVPR 2021.
- Barbano, Dufumier,Tartaglione, Grangetto, Gori "Unbiased supervised contrastive learning", ICLR 2023

Unbiased Contrastive learning setup

- Accuracy obtained on toy data: biased-MNIST

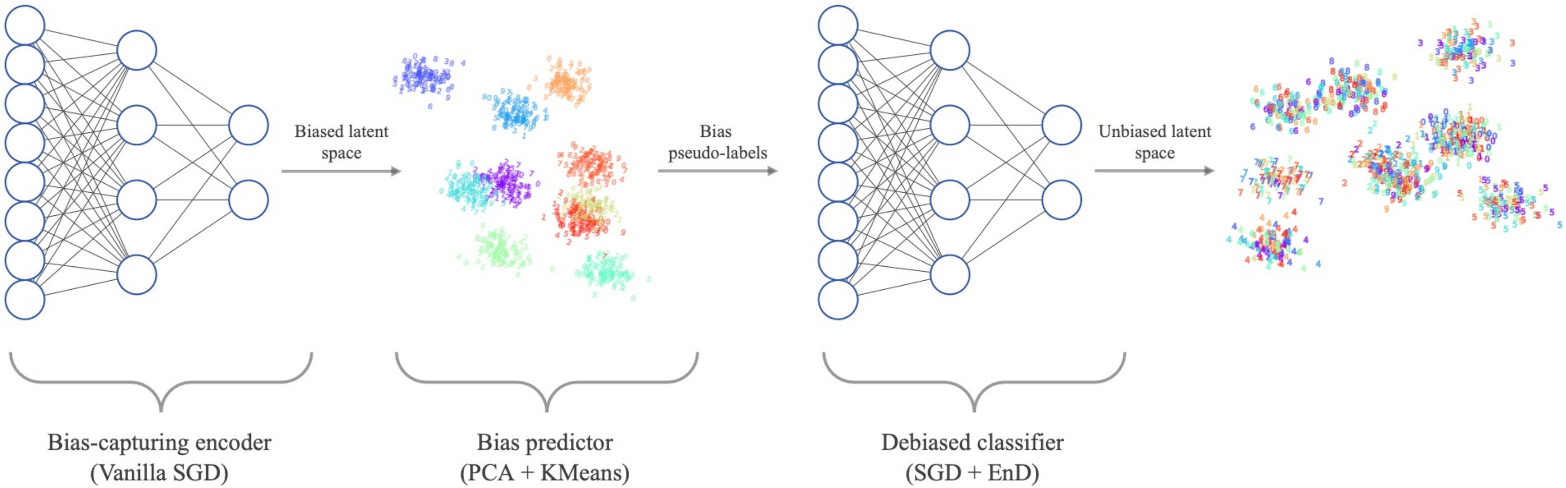
Method	Correlation (%) 99.9	99.7	99.5	99
CE Hong et al. 2021	11.8±0.7	62.5±2.9	79.5±0.1	90.8±0.3
LNL Kim et al. 2019	18.2±1.2	57.2±2.2	72.5±0.9	86.0±0.2
EnD Tartaglione et al. 2021	59.5±2.3	82.70±0.3	94.0±0.6	94.8±0.3
BC+BB* Hong et al. 2021	30.26±11.08	82.83±4.17	88.20±2.27	95.04±0.86
BB Hong et al. 2021	76.8±1.6	91.2±0.2	93.9±0.1	96.3±0.2
BC+CE* Hong et al. 2021	15.06±2.22	90.48±5.26	95.95±0.11	97.67±0.09
FairKL	**90.51**±1.55	**96.19**±0.23	**97.00**±0.06	**97.86**±0.02

UNIVERSITÀ DI TORINO

EIDOSLAB

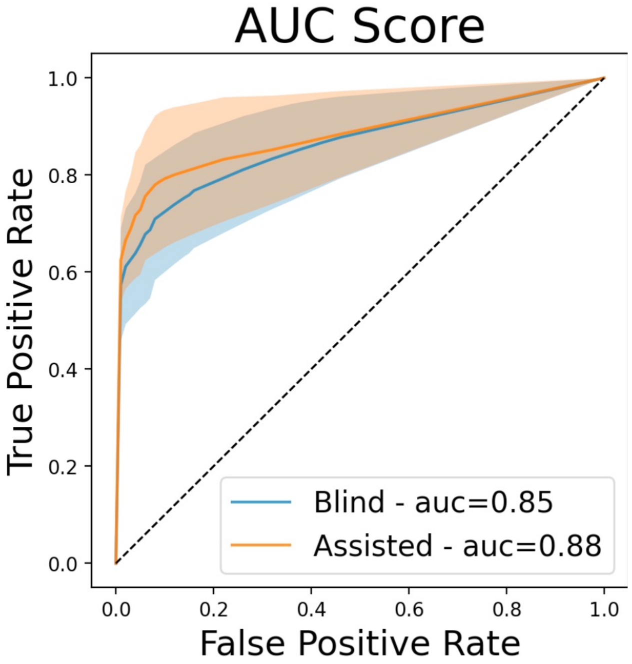
Working with unknown biases

- FairKL assumes to know the bias attribute
- What if we don't know it?



Use case: Co.R.S.A project

- Co.R.S.A. is funded by Regione Piemonte: AI system for COVID identification from CXR
  - Public Dataset, model development, prototype and validation
- State-of-the-art performance
  - using FairKL (multi-site effect mitigation)



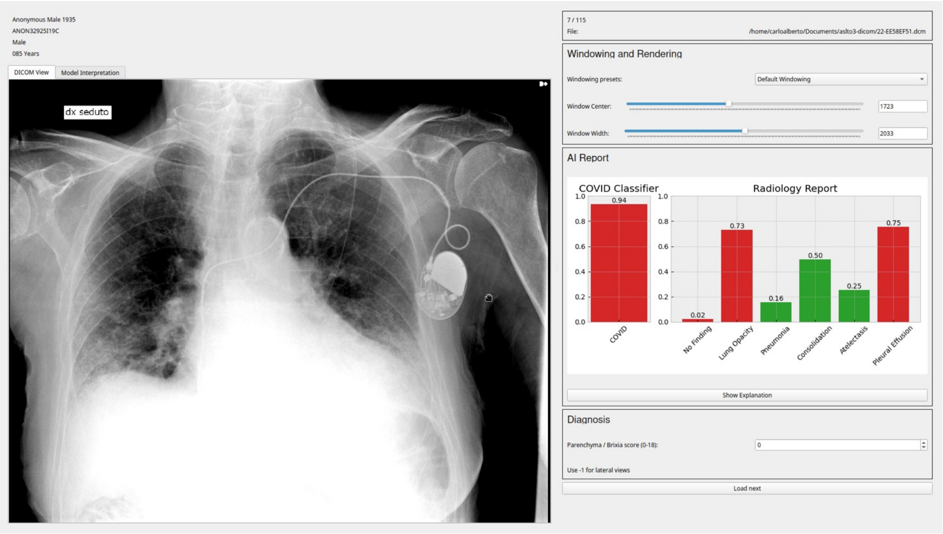
UNIVERSITÀ DI TORINO

EIDOSLAB

Use case: Co.R.S.A project

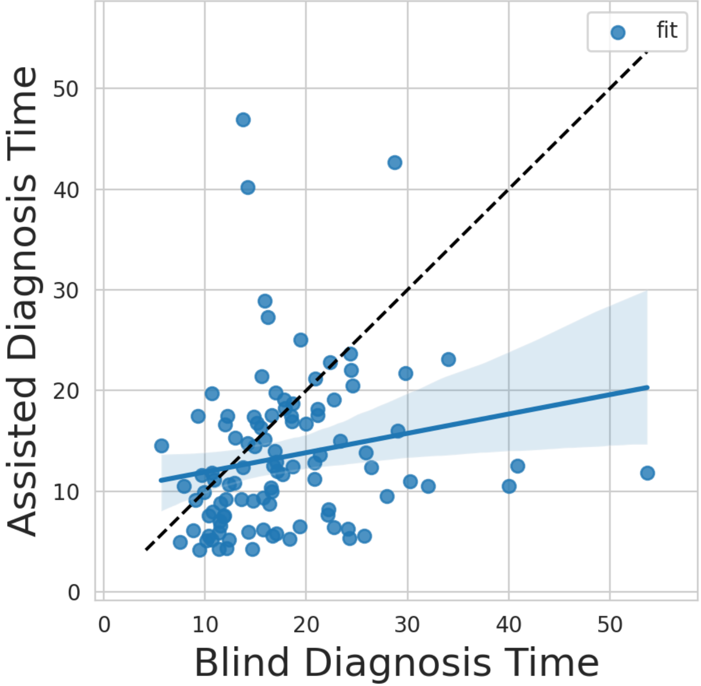
- Co.R.S.A. is funded by Regione Piemonte: AI system for COVID identification from CXR
  - Public Dataset, model development, prototype and validation

- State-of-the-art performance
  - using FairKL (multi-site effect)
  - using interpretability by design



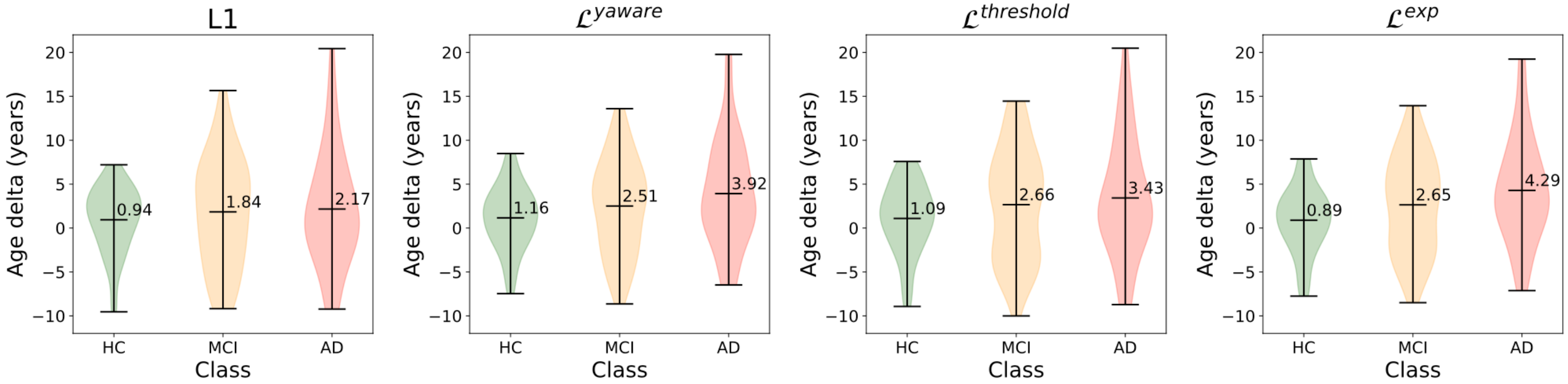
Use case: Co.R.S.A project

- Co.R.S.A. is funded by Regione Piemonte: AI system for COVID identification from CXR
  - Public Dataset, model development, prototype and validation
- State-of-the-art performance
  - using FairKL (multi-site effect)
  - using interpretability by design
  - improved radiologist efficiency



The OpenBHB Challenge

- OpenBHB challenge (Dufumier et al. 2022): brain MRIs from 64 different acquisition sites
- Brain aging involves complex biological processes, (e.g. cortical thinning) → highly heterogeneous, people do not age in the same manner
- Brain age gap (BAG) greater than a certain threshold → unhealthy aging

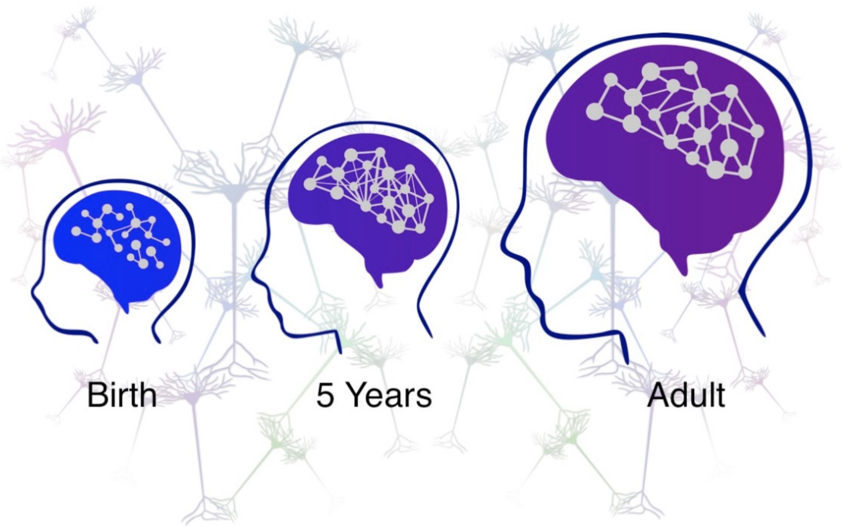


Barbano et al., “Contrastive learning for regression in multi-site brain age prediction”, *IEEE ISBI* 2023

IV

Simple is better: prune to generalize

The brain removes unused connections



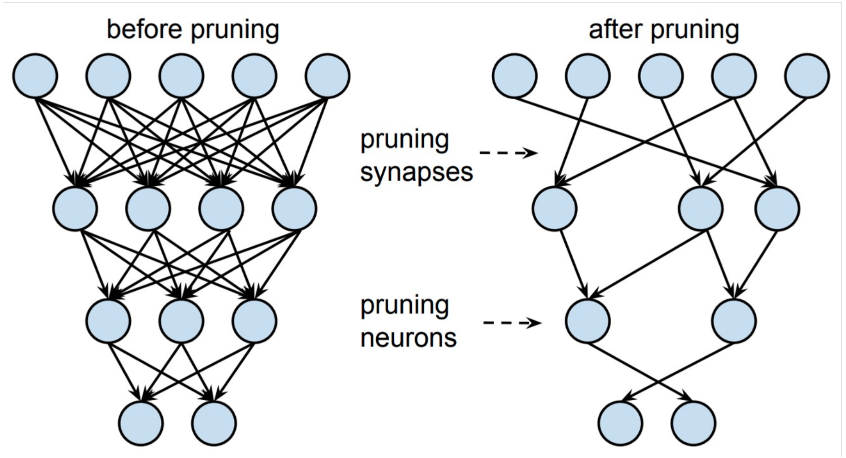
Connecting early learning and brain development, The Institute for Learning & Brain Sciences, University of Washington

Neural Network Pruning

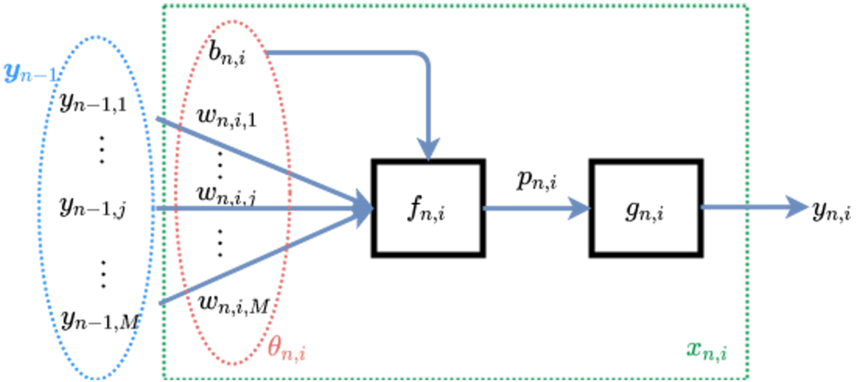
Removes less influential elements while preserving the generalization capabilities.

Reduces the resources required to use the model.

Studied since the late ‘80s has seen a resurgence in 2015.



Neuron sensitivity



How to evaluate the Sensitivity of a neuron?

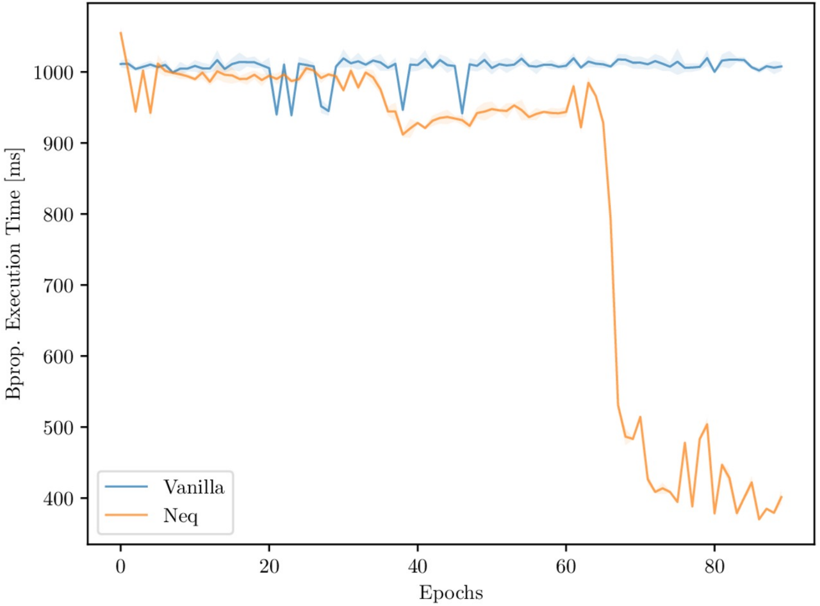
Post-synaptic potential:

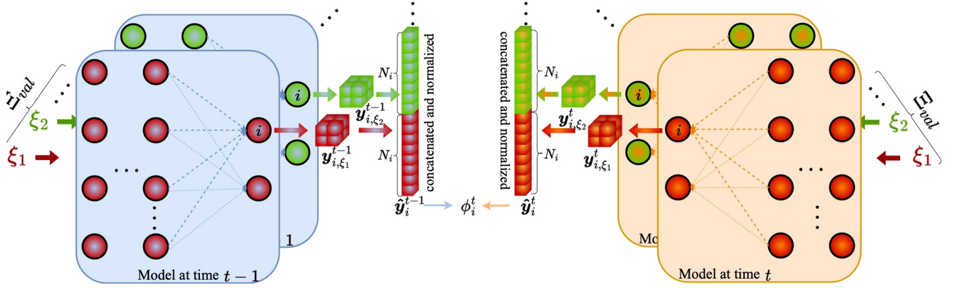
It consider the contribution of all parameters.

It allows us to evaluate the Neuron Sensitivity regardless of the non-linearity.

- Tartaglione, E., Bragagnolo, A., Odierna, F., Fiandrotti, A., & Grangetto, M. (2021). Serene: Sensitivity-based regularization of neurons for structured sparsity in neural networks. IEEE Transactions on Neural Networks and Learning Systems
- Tartaglione, E., Bragagnolo, A., Fiandrotti, A., & Grangetto, M. (2022). Loss-based sensitivity regularization: towards deep sparse neural networks. Neural Networks

Neuron equilibrium



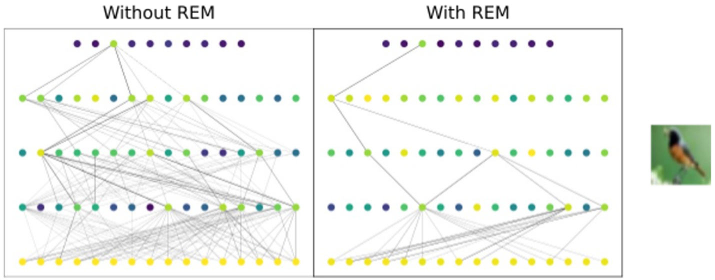


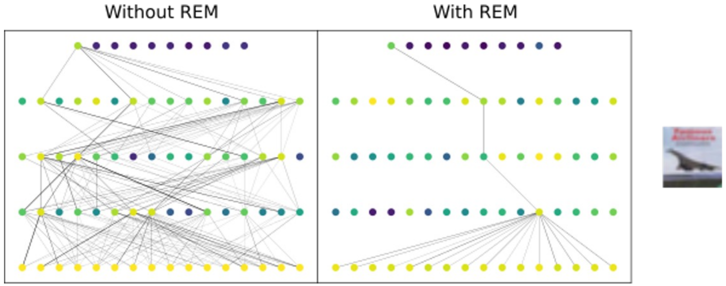
Backpropagation execution time for vanilla and NEq ResNet-18.

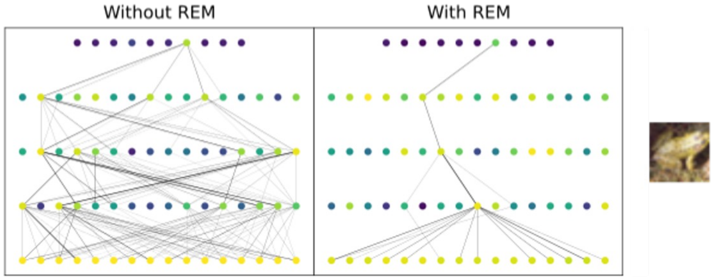
We observe a reduction in the wall-clock time of around -17.52%.

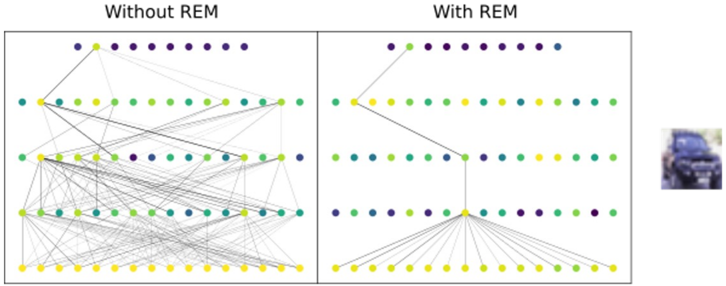
Bragagnolo, A., Tartaglione, E., & Grangetto, M. “To update or not to update? Neurons at equilibrium in deep models”, in Neurips 2022

Pruning & Interpretabiliy









Renzulli, Riccardo, Enzo Tartaglione, and Marco Grangetto. "REM: Routing entropy minimization for capsule networks

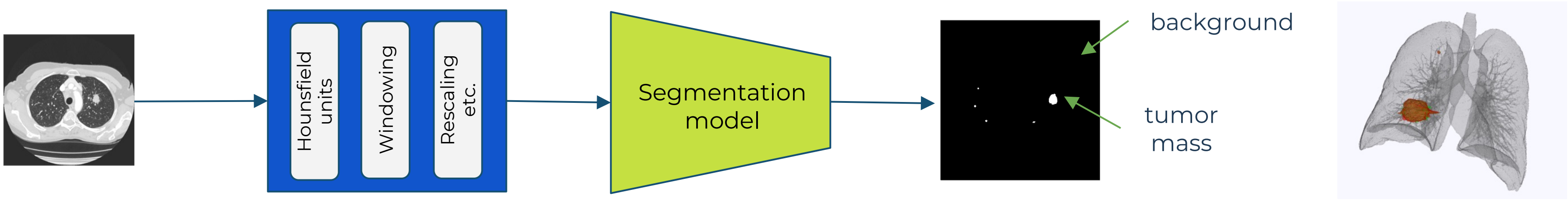
Use case: Lung nodule segmentation

UniToChest

https://zenodo.org/record/5797912#.Y_yoztLMJhE

UniToChest is a collection of anonymized **306440** chest CT scan slices coupled with the proper lung nodule segmentation map, for a total of **10071** nodules from **623** different patients

DEEPHEALTH



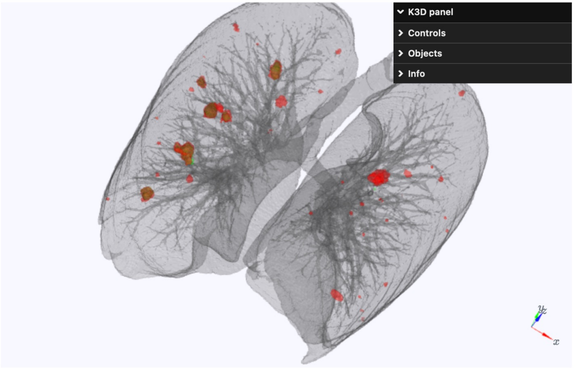
Preprocessing

Segmentation output map

Use case: Lung nodule segmentation

Model	Layers	Training samples (%)	Dice score	Parameters (M)
U-Net		100	0.81	
U-Net	5	50	0.79	31
U-Net		10	0.78	
U-Net		100	0.74	
U-Net	4	50	0.71	7.6
U-Net		10	0.68	
U-Net		100	0.72	
U-Net	3	50	0.69	1.8
U-Net		10	0.66	
SegCaps		100	0.79	
SegCaps	4	50	0.77	1.4
SegCaps		10	0.76	

3D visualization

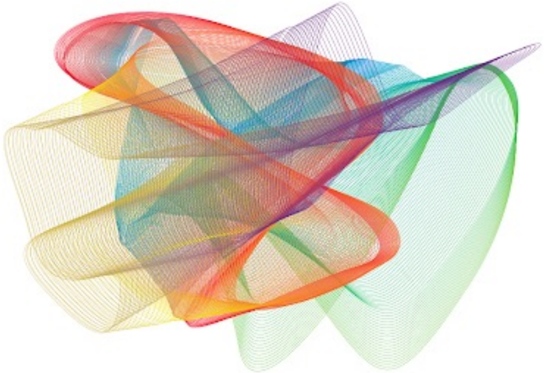


V

Conclusions

Conclusions

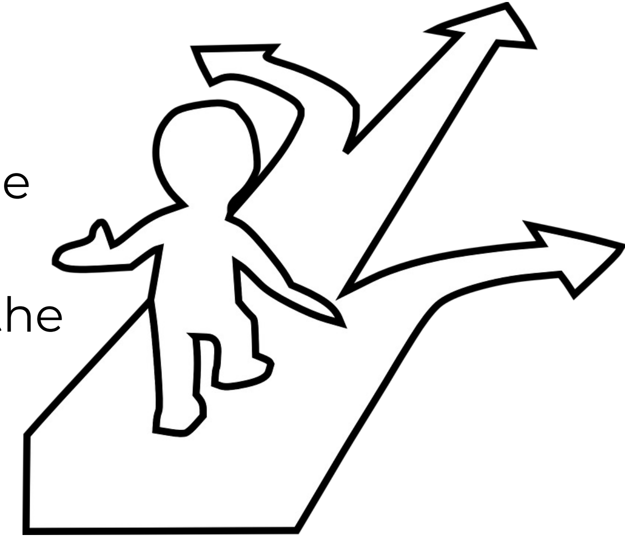
- EU act requirements:
  - high level of **robustness, security and accuracy**
- Trustworthy AI can be tackled from many points: design, interpretability, learning regularization, pruning, human interaction, …
  - Invest in basic & interdisciplinary research



Conclusions

- EU act requirements:
    - **high quality of the datasets** feeding the system to minimise risks and discriminatory outcomes

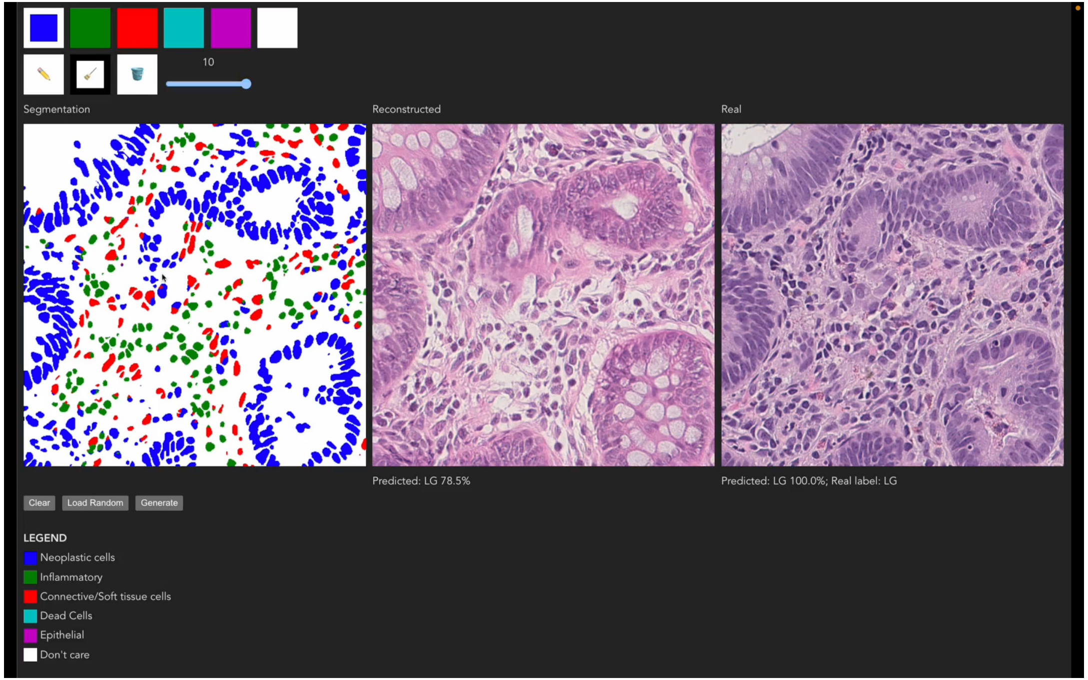
- in many context it is quite difficult to guarantee (or even define) the concept of data quality
- can generative models come into play also in the medical field?



Conclusions

- Generative digital histology

UNITOPATHO (WSI) https://ieee-dataport.org/open-access/unitopatho



All this couldn't be happening **without a group**!

Questions?