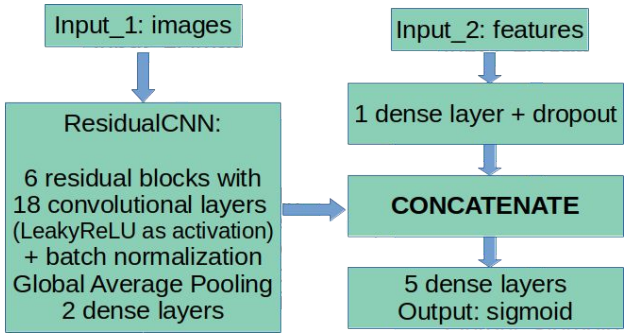
Artificial Intelligence in Medicine

Deep learning in Medical Image Analysis

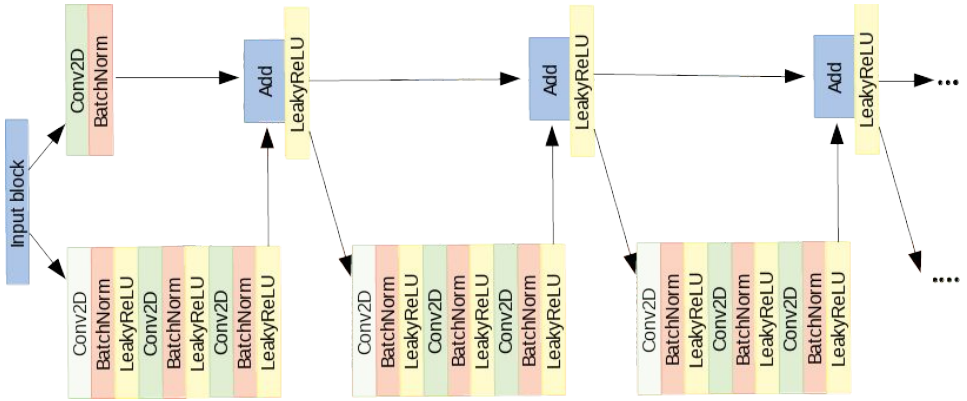
Francesca Lizzi
Scuola Normale Superiore, Pisa
Istituto Nazionale di Fisica Nucleare, Sez. Pisa
Dipartimento di Fisica, Università di Pisa

Deep Learning

Il deep learning è un problema di minimizzazione del rischio empirico



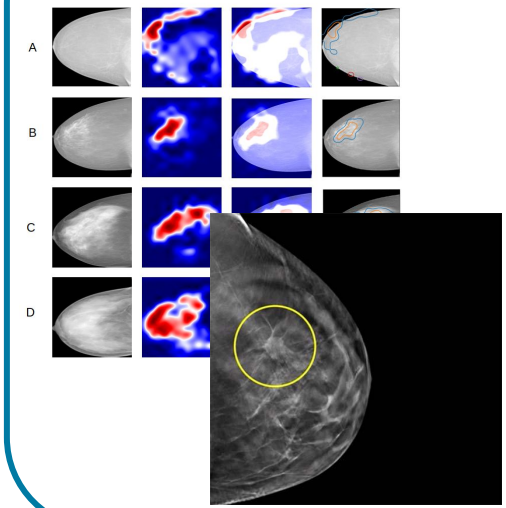
Per quanto riguarda le immagini, gli algoritmi più utilizzati sono Convolutional Neural Netoworks (CNN) e Fully- CNN



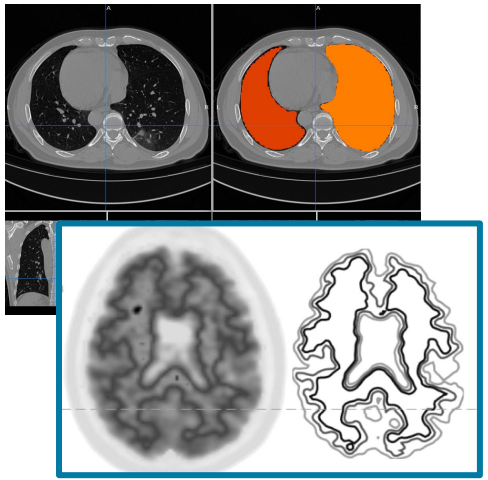
REF: Deep Learning Book, GoodFellow et al, 2016, MIT Press

Deep Learning e immagini mediche

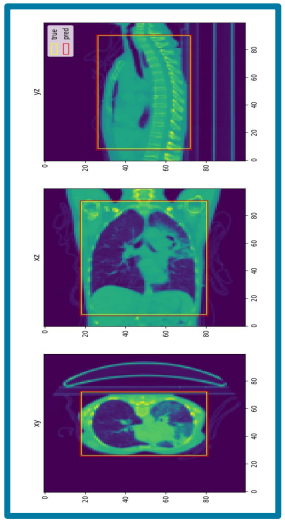
Classificazione:



Segmentazione:



Regressione:

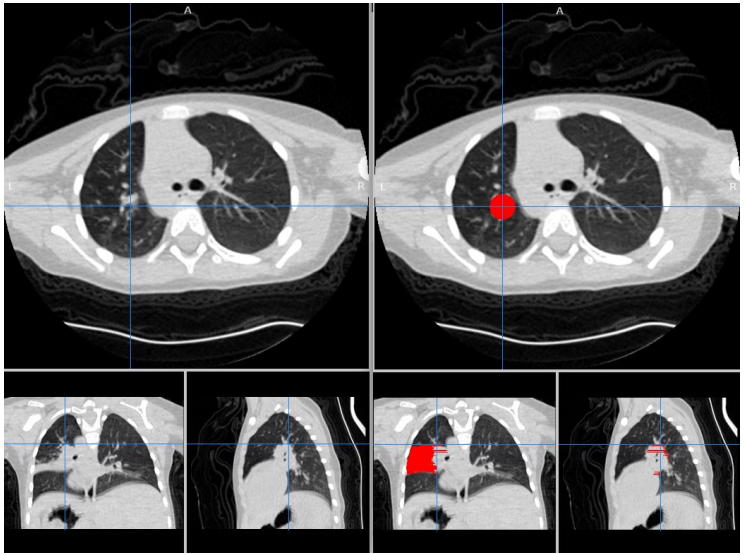


- **Super-risoluzione**: ricostruire immagini ad alta risoluzione da immagini a bassa risoluzione;
- **Generazione di nuovi dati;**

1. Ground Truth in Medical Image Analysis

Quando alleniamo un algoritmo di deep learning, calcoliamo sempre una funzione di costo (**loss function**) che va minimizzata.

- Il GT è dato dall'**opinione di un medico, più medici o un consenso**;
- Può esistere un GT oggettivo : in campo oncologico, per esempio una biopsia ma ottenerla significa incrociare più dati un paziente ed esistono problemi di privacy;
- label noise, polarizzazione dello spettro della malattia;



CT scan of patient volume-covid19-A-0013
COVID-19 Challenge data set

2. Public and private data

The Cancer Imaging Archive

- 147 collezioni;
- la più popolata è il National Lung Screening Trial (NLST) che contiene **26254** Computed Tomography (CT);
- la seconda è il Breast Cancer Screening-DBT che contiene **5060** Digital Breast Tomosynthesis (DBT)

8 dataset contengono più di 1000 soggetti, 51 dataset ne contengono tra 100 e 1000 e il restante contiene meno di 100 soggetti.

Il problema dei formati: **NIfTI e DICOM**

- informazioni conservate
- anonimizzazione e armonizzazione

3. Statistical vs Clinical Validation

Validazione statistica:

- training, validation e test set;
- cross-validation, leave-one-out, ...

Validazione "intermedia": una volta allenato e valutato un algoritmo su training, validation e test set, l'algoritmo va validato su un altro test set esterno, possibilmente multicentrico.

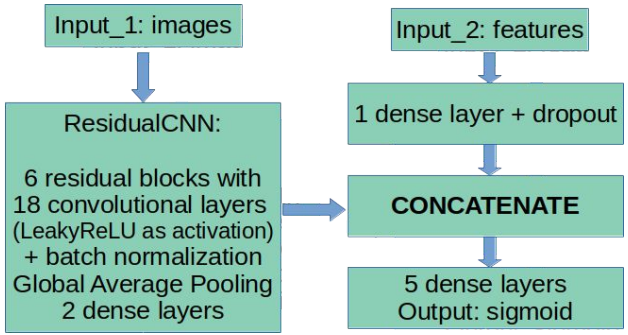
Validazione clinica: trial studies -> prospettici, richiedono tempo, e costosi

È immediatamente vero che migliorare l'accuracy di una diagnosi, ad esempio, facendo confronti controllati tra medici e algoritmo, porti ad un miglioramento nella pratica clinica?

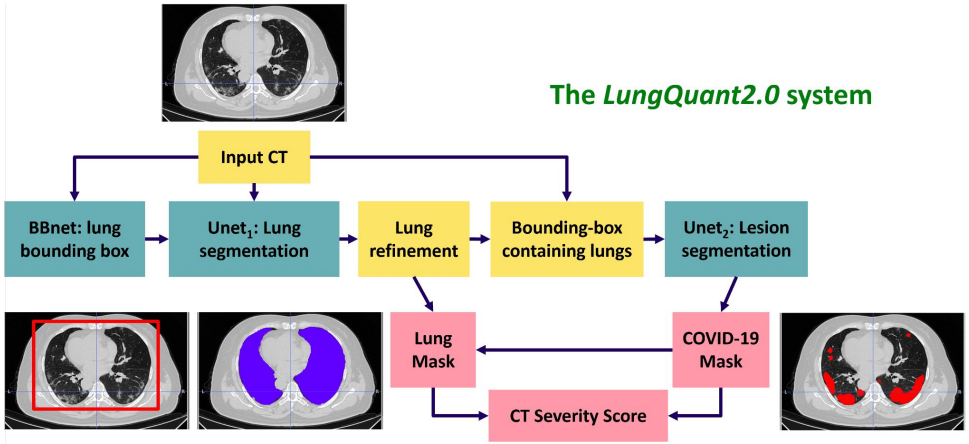
Medical Image Analysis: AIM and next-AIM

Il deep learning nasce nell'ambito della classificazione di immagini naturale ma è stato fin da subito applicato anche all'analisi di immagine medica. Alcuni esempi:

CNN + MLP per CXR COVID-19: multi input CNN per predizione della prognosi in pazienti con COVID-19



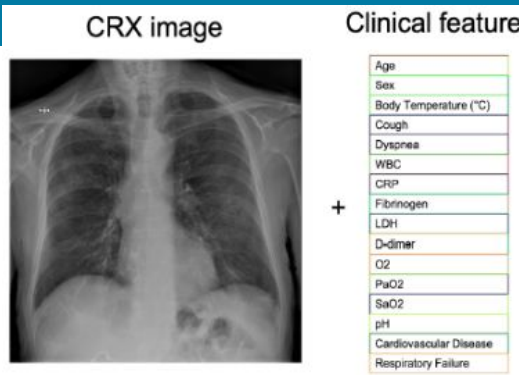
LungQuant: sistema a 3 CNN in cascata per la quantificazione del danno polmonare su CT di pazienti con COVID-19



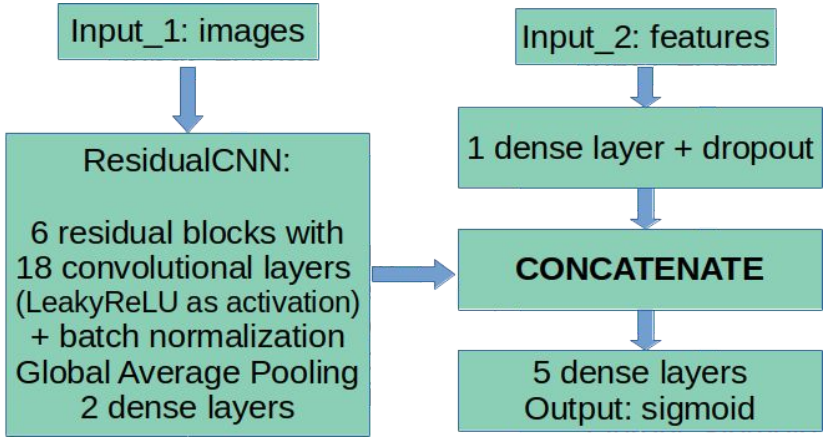
F.Lizzi et al. Quantification of pulmonary involvement in COVID-19 pneumonia by means of a cascade of two U-nets: training and assessment on multiple datasets using different annotation criteria. International Journal of Computer Assisted Radiology and Surgery, 2021

Deep learning per predizione della prognosi su CXR di pazienti con COVID-19

Abbiamo partecipato alla challenge
AI4COVID Hackathon
(https://ai4covid-hackathon.it/)



1103 soggetti nel training set
486 blind test set



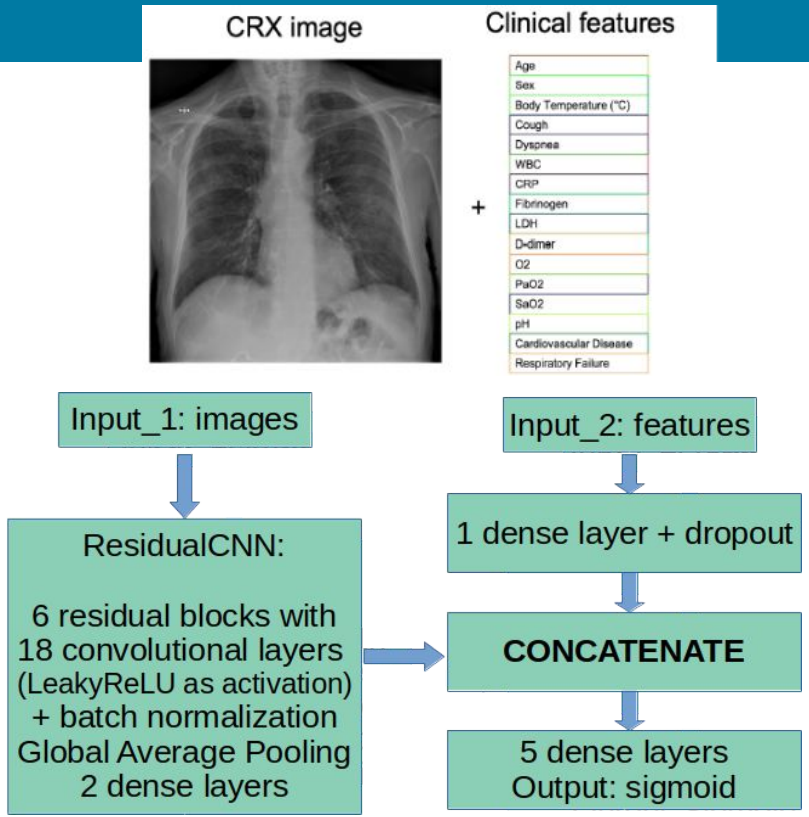
100 epoche, Adam opt, Binary Cross-Entropy, Accuracy come metrica

Deep learning per predizione della prognosi su CXR di pazienti con COVID-19

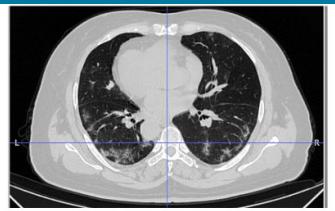
Abbiamo ottenuto una **balanced accuracy sul test set di 73.87%**, classificandoci al quarto posto;

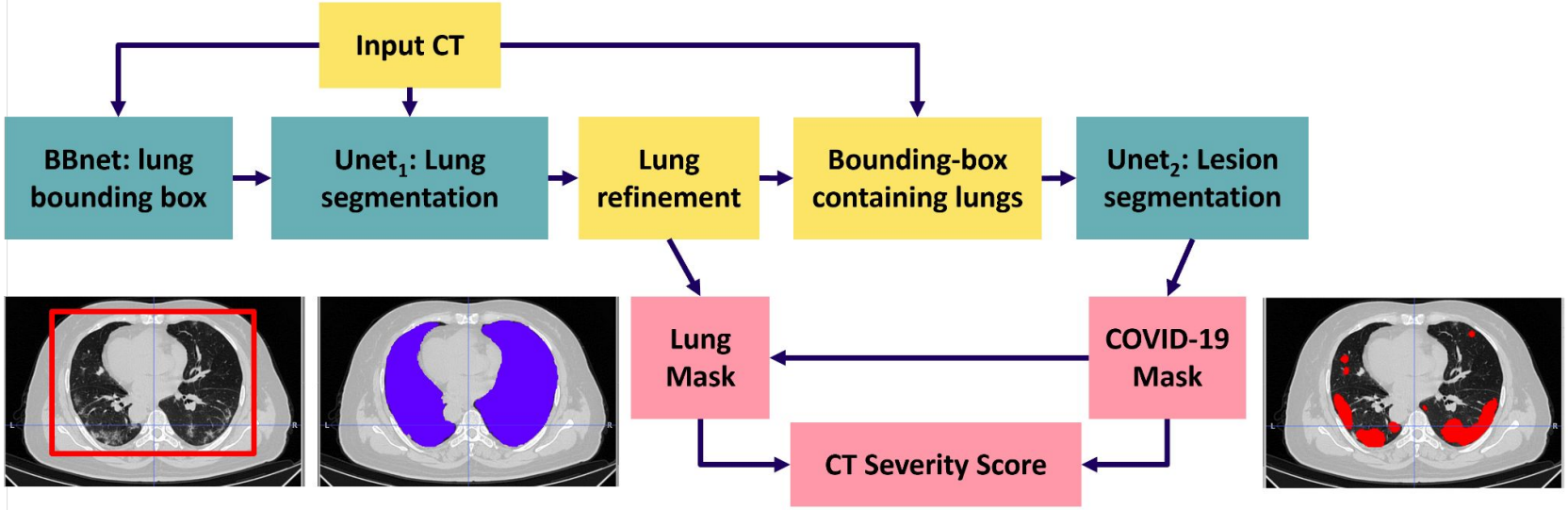
Abbiamo utilizzato l'algoritmo **grad-CAM**, che sottolinea i pixel dell'immagine utilizzati per fare la classificazione, per spiegare la CNN.

Feature importance: WBC, LDH, PaO2, SaO2 e" Difficulty in Breathing"



LungQuant2.0: software per la quantificazione del danno polmonare su CT





F.Lizzi et al. Quantification of pulmonary involvement in COVID-19 pneumonia by means of a cascade of two U-nets: training and assessment on multiple datasets using different annotation criteria. International Journal of Computer Assisted Radiology and Surgery, 2021

LungQuant2.0: dati

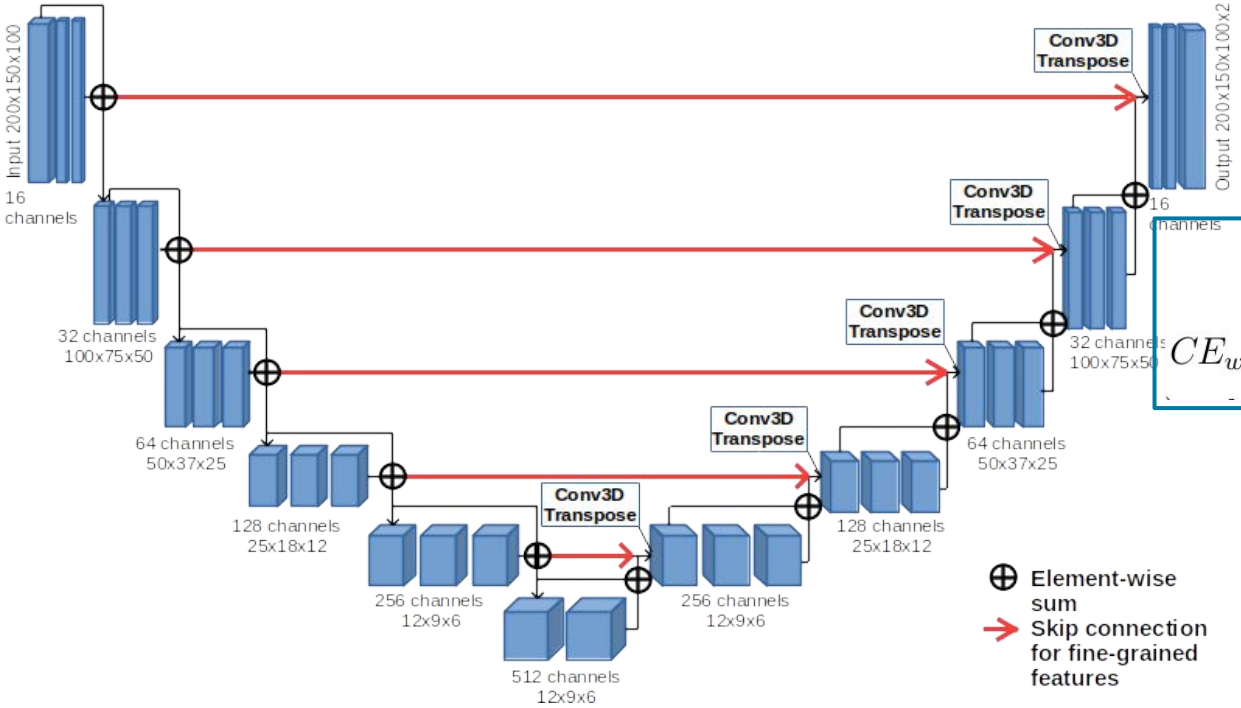
Per ottenere un numero sufficiente di immagini abbiamo aggregato il maggior numero possibile di dati che avessero le maschere del polmone e/o le maschere dell'infezione da COVID-19

Dataset name	Lung mask	GGO mask	CT-SS	N. of cases
Plethora [8]	Yes	No	No	402
Lung CT Segmentation Challenge [15]	Yes	No	No	60
COVID-19 Challenge [1]	No	Yes	No	199
MosMed [12]	No	No	No	1110
MosMed (annotated subsample)	No	Yes	Inferable	50
MosMed (in-house annotated subsample)	Yes	No	No	91
COVID-19-CT-Seg [11]	Yes	Yes	Inferable	10

$\mathbf{U\text{-}net_1}$	train	val	test
Plethora	319	40	40
MosMed (91 CT-0)	55	18	18
LCTSC	36	12	12
COVID-19-CT-Seg	/	/	10
$\mathbf{U\text{-}net_2^{60\%}}$	train (60%)	val (20%)	test
COVID-19 Challenge	119	40	40
MosMed (50 CT-1)	30	10	10
COVID-19-CT-Seg	/	/	10
$\mathbf{U\text{-}net_2^{90\%}}$	train (90%)	val (10%)	test
COVID-19 Challenge	179	20	/
MosMed (50 CT-1)	45	5	/
COVID-19-CT-Seg	/	/	10

Lizzi F, et al.. "Making data big for a deep-learning analysis: aggregation of public COVID-19 datasets of lung Computed Tomography scans" in Proceedings of the 10th International Conference on Data Science, Technology and Applications (DATA), 2021, pages 316-321, doi: https://doi.org/10.5220/0010584403160321

LungQuant2.0: U-Net



Ottimizzatore: ADAM

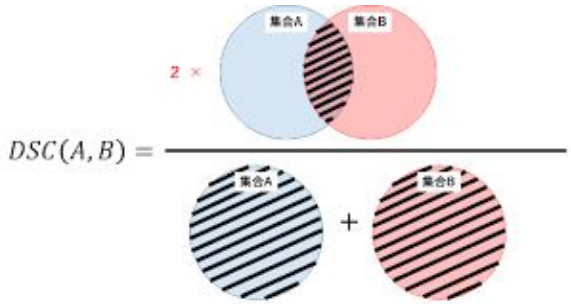
Loss: vDSC (lung)

Metrica: vDSC

Infection Loss:

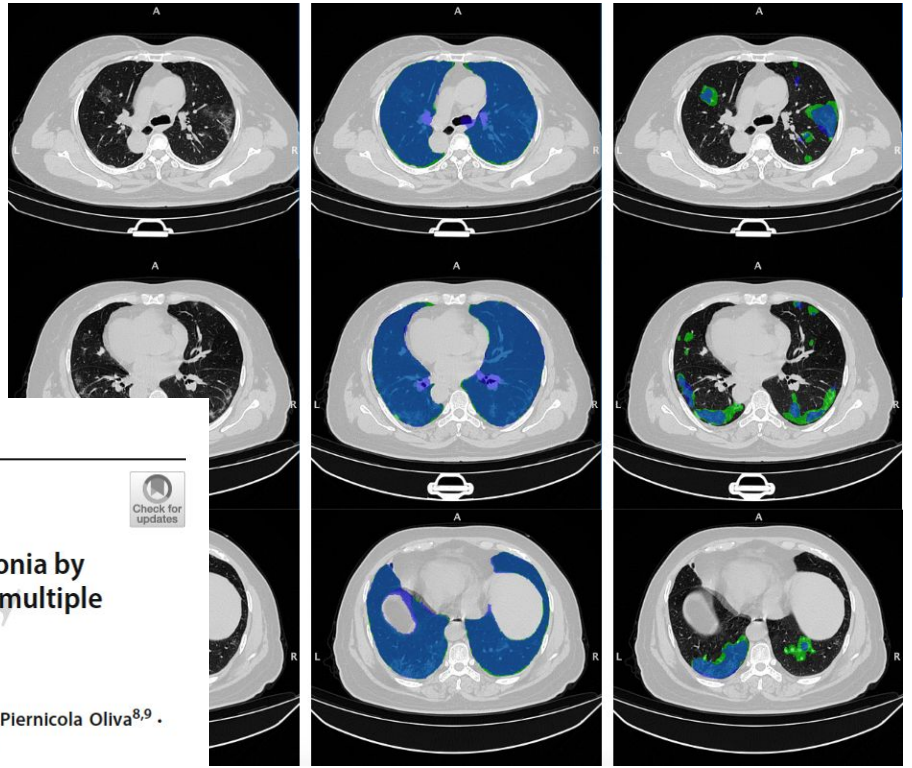
$$L = Dice_{loss} + CE_{weighted}$$

$$CE_{weighted} = w(x) \sum_{x \in \Omega} log(M_{true}(x) \cdot M_{pred}(x))$$



LungQuant2.0: risultati

	sDSC (5mm)	vDSC
Lung	0.95 ± 0.02	0.95 ± 0.01
COVID infection	0.76 ± 0.18	0.66 ± 0.13



International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-021-02501-2

ORIGINAL ARTICLE

Check for updates

Quantification of pulmonary involvement in COVID-19 pneumonia by means of a cascade of two U-nets: training and assessment on multiple datasets using different annotation criteria

Francesca Lizzi[1,2] · Abramo Agosti[6] · Francesca Brero[4,5] · Raffaella Fiamma Cabini[4,6] · Maria Evelina Fantacci[2,3] · Silvia Figini[4,11] · Alessandro Lascialfari[4,5] · Francesco Laruina[1,2] · Piernicola Oliva[8,9] · Stefano Piffer[7,10] · Ian Postuma[4] · Lisa Rinaldi[4,5] · Cinzia Talamonti[7,10] · Alessandra Retico[2]

Conclusioni e discussione

LungQuant è in validazione da parte di 14 medici radiologi di 5 sedi diverse (Pisa, Pavia, Milano, Firenze, Catania).

Accedere a dati di qualità è una delle sfide più importanti e urgenti nel campo del deep learning applicato all'analisi di immagini mediche:

- Quantità di dati;
- Label corrette;
- Distribuzioni corrette dello spettro della malattia;
- Dati che abbiano i parametri di acquisizione, gli algoritmi di ricostruzione, ...

Poter disporre di tutte le informazioni necessarie renderebbe possibile l'armonizzazione e uno studio più preciso dei parametri fisici che influenzano il comportamento delle CNN.

Grazie per l'attenzione.
Domande?

francesca.lizzi@sns.it

3. Caso studio: interpretazione di CRX

In questo studio, si è valutato l'effetto dell'utilizzo di un software basato su **deep learning** per l'interpretazione di radiografie toraciche nel **Pronto Soccorso.**

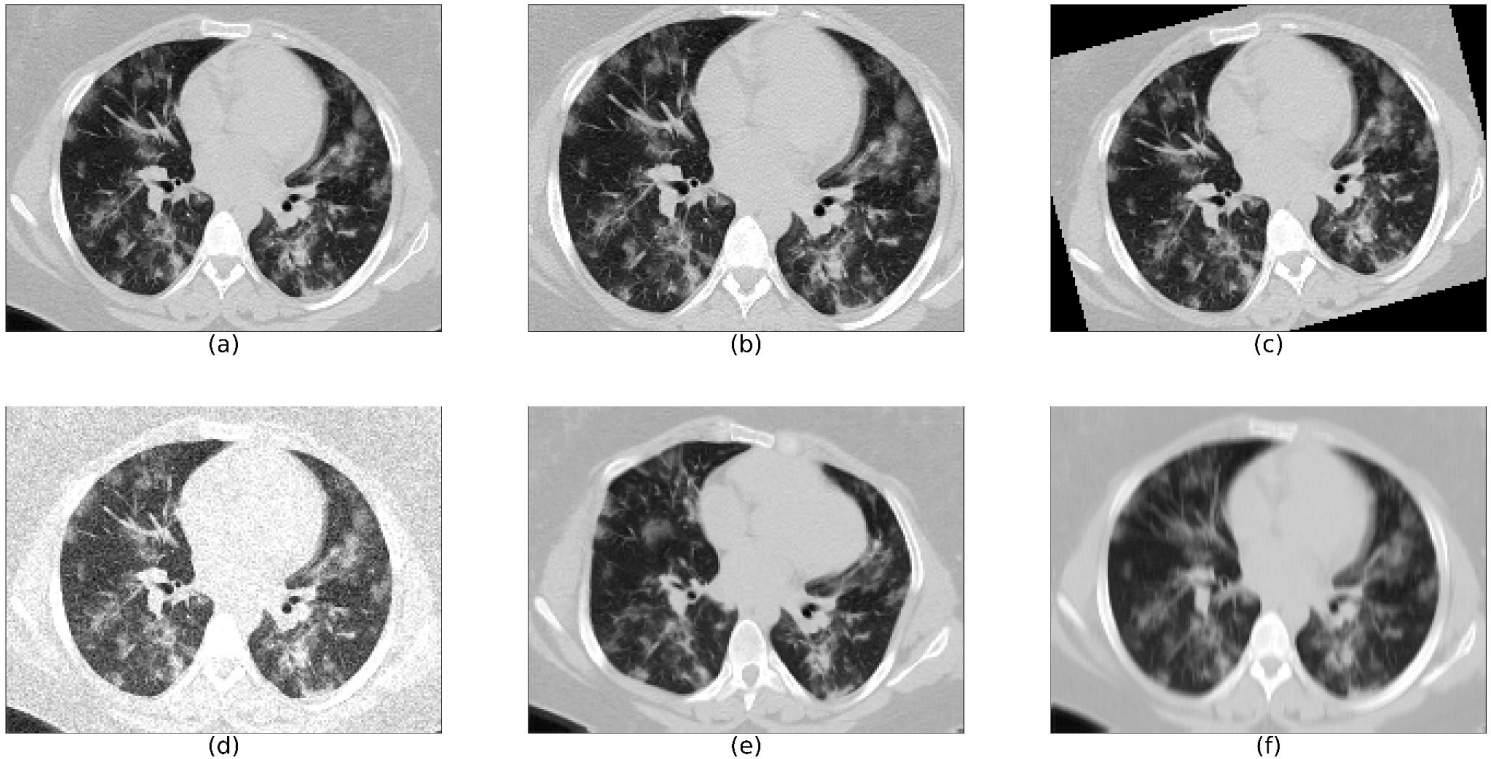
- Interpretare radiografie toraciche per un occhio non specializzato è un compito molto difficile (secondo alcuni studi il rate di errore è superiore al 22%);
- I medici radiologi esperti potrebbero non essere disponibili immediatamente;

In questo studio prospettico, è stato visto che **l'utilizzo di software basato su DL potrebbe effettivamente portare un beneficio nei reparti di Pronto Soccorso.**

Se il software fosse stato confrontato con il responso di radiologi esperti, probabilmente la sua performance non sarebbe stata migliore.

Ji Hoon Kim et al. "Effect of deep learning-based assistive technology use on chest radiograph nterpretation by emergency department physicians: a prospective interventional simulation-based study". BMC Medical In-formatics and Decision Making, 21(1):1–9, 2021

LungQuant2.0: data augmentation



Deep learning: quale rivoluzione?

I metodi basati su deep learning riescono ad avere ottime performance e una buona capacità di generalizzazione, lasciando che sia l'algoritmo ad imparare le caratteristiche dei dati importanti per lo specifico problema da risolvere. Ma a che prezzo?

"Correlation is Enough", 2008 Anderson

"Data mining techniques can extract the complete set of patterns and effects, producing scientific conclusions without any further experiment.", 2009 Prensky

Data analysts should not propose or even "bother" themselves with hypothesis anymore, 2013 Steadman

Deep learning is "a kind of magic", 2018, Stuart J Russell

Secondo alcune interpretazioni, il deep learning e in generale la Data Science stanno causando una rivoluzione nell'analisi dei dati e un cambiamento di paradigma scientifico kuhniano.

Inferire i modelli dai dati è un modo nuovo di fare scienza? L'ipotesi scientifica è davvero inutile? Davvero il deep learning non ha bisogno di teoria?

Deep learning: quale paradigma?

Secondo alcune interpretazioni, semplificando, ci sono due paradigmi ai quali possiamo rifarci quando utilizziamo il deep learning:

1) **Approccio puramente empirico, induttivista, senza teoria e senza ipotesi:**
   - I dati rappresentano un intero campo della conoscenza
   - Non c'è necessità di informazioni o modelli a priori
   - Le tecniche di data mining sono **agnostiche**
   - Il significato trascende un campo specifico della conoscenza, **tutti sono in grado di interpretare**.

2) **Approccio costruttivista:**
   - I dati sono espressione della realtà in cui viviamo, sono costruiti e la loro costruzione ne enfatizza delle caratteristiche e ne sopprime delle altre
   - **Non sono risorse "naturali"** ma socialmente determinate
   - Esprimono una verità che deve essere interpretata e per interpretarla è necessario conoscere il background teorico sul quale l'algoritmo è sviluppato.

4. Bias sociali

L'OMS afferma che l'IA porterà beneficio alla medicina se e solo se i diritti umani ne saranno la base. (WHO. Ethics and Governance of Artificial Intelligence for Health 2021)

Oltre alle difficoltà tecnologiche che abbiamo visto, ci sono molte questioni riguardo alla stratificazione della popolazione che richiedono ancora di essere studiate.

Un algoritmo basato su deep learning e allenato su immagini mediche può essere razzista?

In questo studio hanno allenato diverse CNN per la predizione della self-reported race.

9 dataset, pubblici e privato di CXR, Chest CT, Mammography, Digital Radiography, Lateral C-Spine XR

Per ogni modalità di imaging è stato banale inferire la self-reported race, anche facendo correzione note (per densità ossea, per densità mammografica)

Facendo transfer learning e aggiungendo solo un ultimo softmax, le CNN allenate per predire diagnosi (su CXR) risultano condizionate dalla self-reported race.

I Banerjee et al. "Reading Race : AI Recognizes Patient's Racial Identity In Medical Images." arXiv, https://arxiv.org/abs/2107.10356