Individual predictions matter: Assessing the effect of data ordering in training fine-tuned CNNs for medical imaging
eJournal 7th April 2020

David Wallis

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 764458.

Aim of the Paper

HYBRID

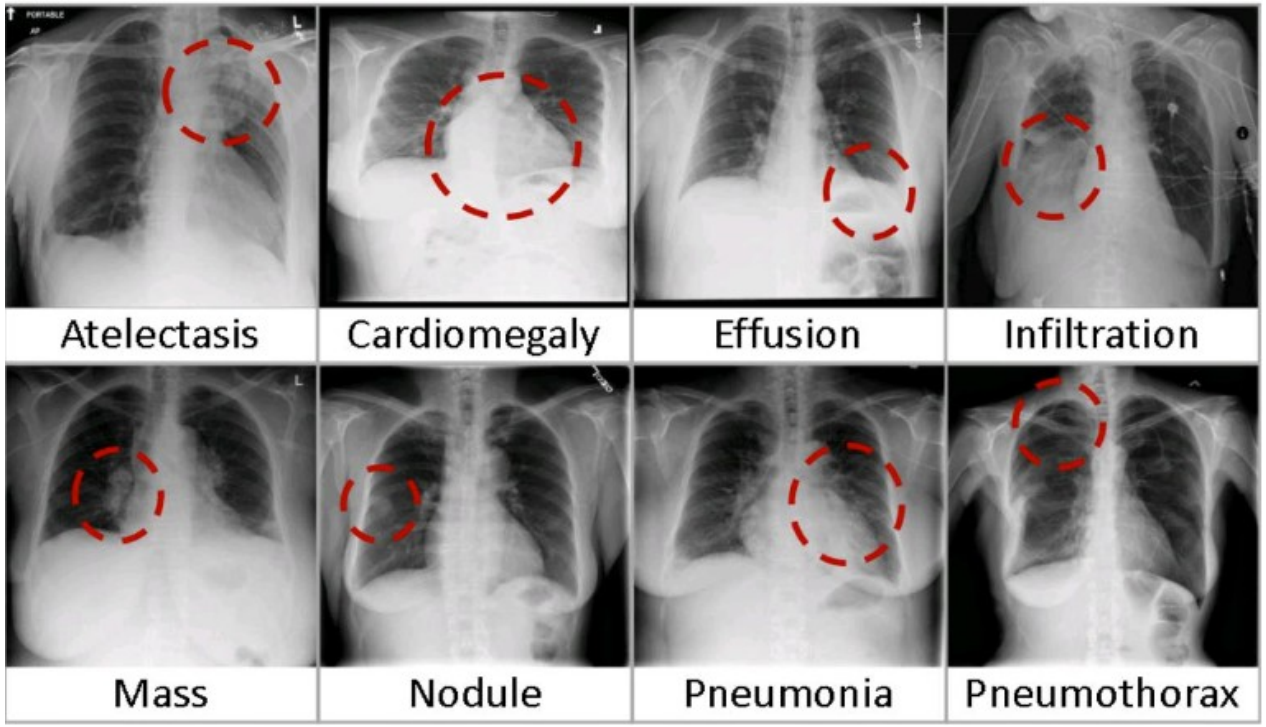
- CNNs are widely used in research, but are they reliable and reproducible enough to be used in a clinical setting?
- If models generate different predictions when retrained, they could make inconsistent predictions for the same patient
- To test this, in this paper they train a CNN multiple times to see how predictions vary

Marie Skłodowska-Curie Grant Agreement No 764458

Dataset

HYBRID

- NIH chest radiography dataset
- 112,129 radiographs used to identify 14 findings
- Train:Validate:Test ratio of 70:10:20 used

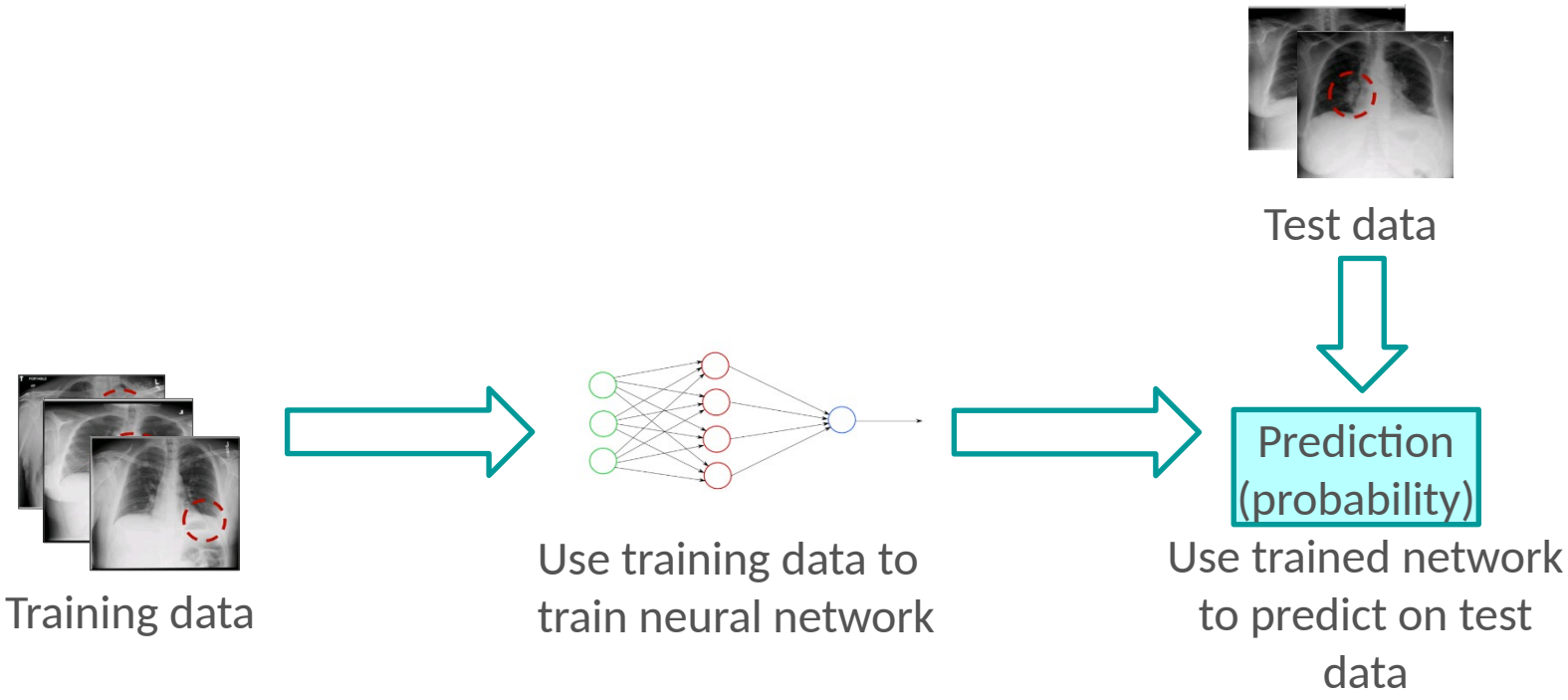


Finding
Atelectasis
Cardiomegaly
Consolidation
Edema
Effusion
Emphysema
Nodule
Pneumonia
Fibrosis
Hernia
Infiltration
Mass
Pleural Thickening
Pneumothorax

Marie Skłodowska-Curie Grant Agreement No 764458

Basic Idea - Network Setup

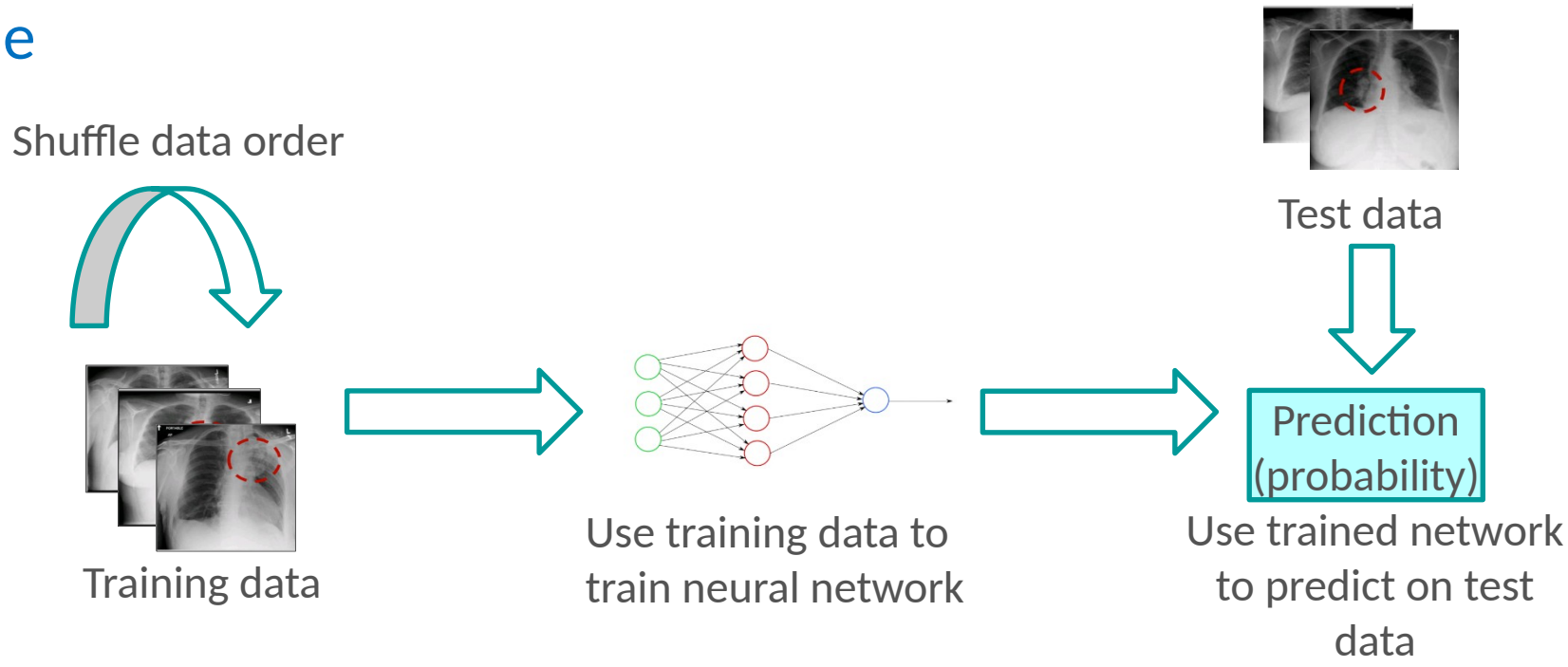
- Normal machine learning setup – use training data to train a neural network, then use the trained network to predict the test data outcome



Marie Skłodowska-Curie Grant Agreement No 764458

Basic Idea – Repeat Experiment Varying Input Order

- Exactly the same setup as previously, but randomly shuffle the order of the training data
- All other parameters (epochs, learning rate, initialisation weights etc) kept the same



Marie Skłodowska-Curie Grant Agreement No 764458

Method – Neural Network Used

- Used a DenseNet-121 CNN pre-trained on ImageNet, then fine-tuned on the chest dataset
- Experiment repeated 50 times, varying the training data input order in each case
- Record the test set results in each case, see how consistent they are

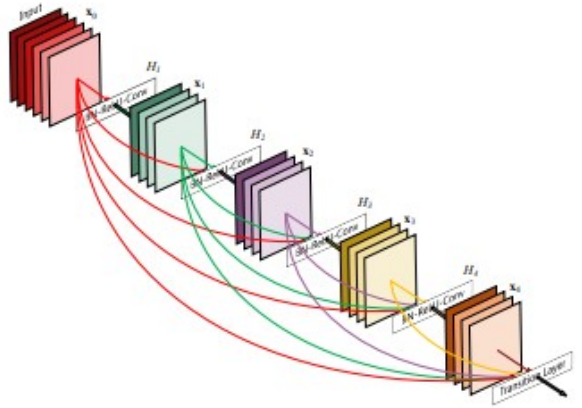
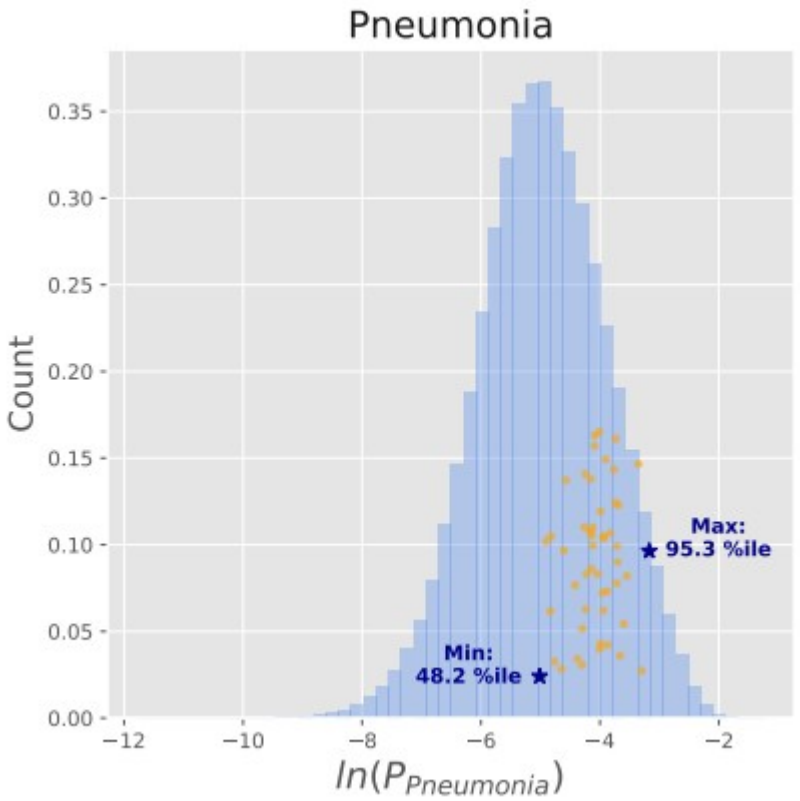


Figure 1: A 5-layer dense block with a growth rate of $k = 4$. Each layer takes all preceding feature-maps as input.

Marie Skłodowska-Curie Grant Agreement No 764458

Results - Pneumonia

- In blue the variability of the full test set across trained models (n=22,433 radiographs × 50 models = 1,121,650)
- In orange the variability in predicted probability of pneumonia for a single test set radiograph across all 50 trained models
- The predicted risk of pneumonia on the single radiograph ranged from the 48.2 to the 95.3 percentile



Marie Skłodowska-Curie Grant Agreement No 764458

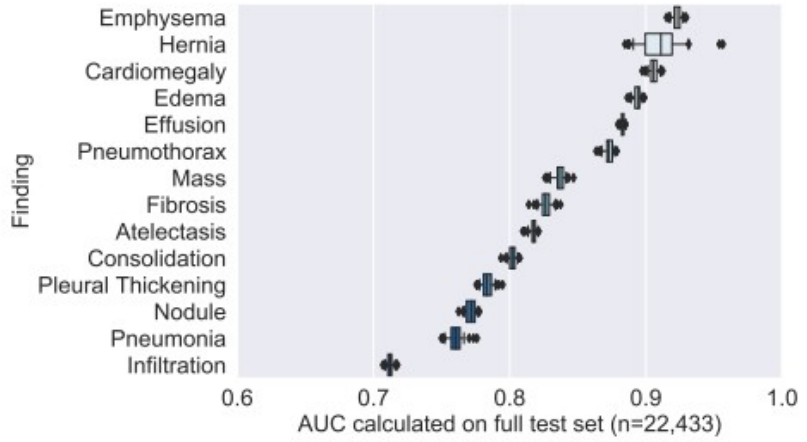
Results

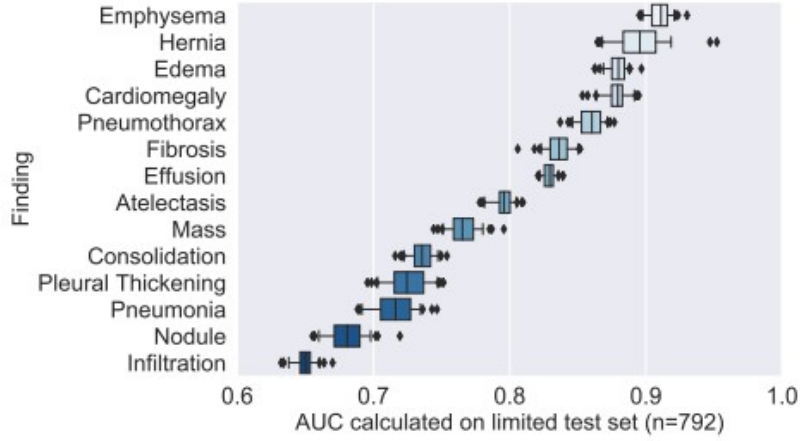
- Table shows the average variation for each finding, across the whole test set
- The mean coefficient of variability mean(σ/μ), was 0.543 for individual models compared to 0.169 for ensembles of 10 runs

	Average across individual models (n=50)					Ensemble of 10 runs (n=5)	
Finding	**Mean (μ)**	**Stdev. (σ)**	σ/μ	$\ln(\frac{P_{max}}{P_{min}})$	**%ile rank range**	σ	σ/μ
Atelectasis	0.107	0.034	0.449	2.085	0.360	0.011	0.142
Cardiomegaly	0.030	0.014	0.686	2.993	0.404	0.004	0.211
Consolidation	0.041	0.014	0.439	2.046	0.368	0.004	0.133
Edema	0.022	0.009	0.654	2.921	0.378	0.003	0.205
Effusion	0.128	0.033	0.523	2.415	0.309	0.010	0.163
Emphysema	0.023	0.010	0.703	3.033	0.479	0.003	0.219
Nodule	0.056	0.021	0.444	2.029	0.493	0.007	0.140
Pneumonia	0.012	0.004	0.403	1.867	0.451	0.001	0.126
Fibrosis	0.016	0.007	0.531	2.435	0.446	0.002	0.171
Hernia	0.002	0.001	0.608	2.784	0.494	0.0004	0.185
Infiltration	0.172	0.042	0.299	1.401	0.425	0.013	0.091
Mass	0.051	0.022	0.624	2.765	0.493	0.007	0.199
Pleural Thickening	0.029	0.012	0.515	2.367	0.457	0.004	0.162
Pneumothorax	0.046	0.022	0.723	3.196	0.465	0.007	0.227

Results – How does this translate to AUC?

- Comparing the AUC and error in AUC for a reduced (n=792) and full (n=22433) dataset
- Shows AUC variability decreases as n increases
- Also note that the AUC is consistently lower for the reduced dataset
- This low variance in AUC masks potential wide variations in predictions on individual radiographs





Marie Skłodowska-Curie Grant Agreement No 764458

Results – How does this translate to AUC?

- AUC and empirical (from 50 tests) v theoretical confidence intervals for each finding
- On reduced test set (n=792)
- Empirical intervals don't exceed theoretical intervals

Finding	Mean AUC	Empirical 95% CI width	Average DeLong 95% CI width	Average bootstrap 95% CI width
Atelectasis	0.796	0.029	0.077	0.077
Cardiomegaly	0.878	0.037	0.083	0.082
Consolidation	0.736	0.030	0.097	0.097
Edema	0.879	0.025	0.072	0.071
Effusion	0.829	0.018	0.065	0.065
Emphysema	0.910	0.028	0.067	0.066
Nodule	0.681	0.047	0.111	0.109
Pneumonia	0.715	0.054	0.137	0.136
Fibrosis	0.836	0.033	0.100	0.100
Hernia	0.897	0.082	0.108	0.105
Infiltration	0.650	0.030	0.087	0.087
Mass	0.766	0.040	0.103	0.103
Pleural Thickening	0.725	0.051	0.112	0.111
Pneumothorax	0.860	0.031	0.077	0.077

Marie Skłodowska-Curie Grant Agreement No 764458

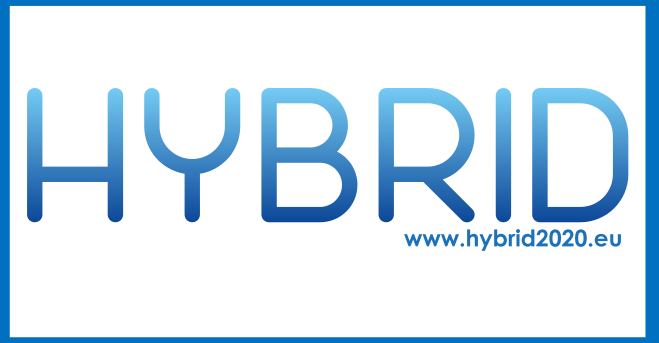
Discussion

HYBRID

- Individual variation high (mean coefficient of variability mean σ/μ was 0.543), but averaging over 10 CNNs reduces this to 0.169
- In a clinical setting this could shift an individual patient from low to high risk (43% percentile range between lowest and highest probability estimation)
- AUC more consistent, but this can mask variations for predictions in single cases
- Studies have shown 30% disagreement between radiologists' interpretations of abdominipelvic CTs and 25% disagreement for the same radiologist at different times

Marie Skłodowska-Curie Grant Agreement No 764458

Thank you for your attention



This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 764458.

Appendix – # of Data in different diseases

A.2 co-occurrence matrix of the fourteen thorax diseases in this chest X-ray dataset

4212	369	3269	3259	727	585	243	772	1222	221	423	220	495	40
369	1094	1060	583	99	108	36	48	169	127	44	51	111	7
3269	1060	3959	3990	1244	909	253	995	1287	592	359	188	848	21
3259	583	3990	9552	1151	1544	571	943	1220	979	447	345	749	33
727	99	1244	1151	2138	894	62	424	602	128	212	115	448	25
585	108	909	1544	894	2706	63	340	428	131	115	166	410	10
243	36	253	571	62	63	307	34	114	330	21	11	45	2
772	48	995	943	424	340	34	2199	222	33	746	80	289	9
1222	169	1287	1220	602	428	114	222	1314	162	103	79	251	4
221	127	592	979	128	131	330	33	162	634	30	9	64	3
423	44	359	447	212	115	21	746	103	30	895	36	151	4
220	51	188	345	115	166	11	80	79	9	36	727	176	8
495	111	848	749	448	410	45	289	251	64	151	176	1127	8
40	7	21	33	25	10	2	9	4	3	4	8	8	110
11535	2772	13307	19871	5746	6323	1353	5298	4667	2303	2516	1686	3385	227

Marie Skłodowska-Curie Grant Agreement No 764458