Supplementary Information

Discrimination Transfers but Thresholds Do Not: External Validation of Convolutional and Transformer Pneumonia Classifiers on Adult and Paediatric Chest Radiographs

Contents

Supplementary Methods (reference standards)
Table S1 Paired differences in AUC
Table S2 AUC in subgroups of the RSNA test split
Table S3 Calibration in each cohort
Table S4 Subgroup contrasts in the external cohorts
Table S5 Best epoch versus stochastic weight averaging
Table S6 Errors at the validation threshold
Fig. S1 Training curves
Fig. S2 Best epoch versus stochastic weight averaging, per run
Fig. S3 Examples of errors
CLAIM 2024 checklist

Supplementary Methods: Reference Standards

RSNA. Labels and boxes were produced for the RSNA Pneumonia Detection Challenge by 18 board-certified radiologists from 16 institutions, six representing the Radiological Society of North America and twelve the Society of Thoracic Radiology (Shih et al., Radiol Artif Intell 2019). Most radiographs were read by one radiologist; the pool from which the challenge test set was drawn was read by three, with majority voting and adjudication by an experienced thoracic radiologist. “Lung Opacity” denotes an opacity with at least a medium likelihood of pneumonia in a patient with cough and fever; “No Lung Opacity / Not Normal” covers radiographs with other findings, including opacities not suspicious for pneumonia; “Normal” denotes no findings. No inter-reader agreement statistics were published.

Kermany. The radiographs were screened for quality and graded by two expert physicians, with the evaluation set of the original release checked by a third expert (Kermany et al., Cell 2018). We treated bacterial and viral pneumonia as positive.

Indiana. Findings in the radiologist reports were coded manually with Medical Subject Headings by the creators of the collection, after automatic de-identification with manual verification (Demner-Fushman et al., J Am Med Inform Assoc 2016). Our grouping rule used the major headings of each report: pneumonia-like if any of pneumonia, consolidation, infiltrate or airspace disease was present; normal if “normal” was the only heading; opacity without a pneumonia-like heading as a separate, ambiguous group; and other abnormal otherwise.

Table S1 Paired differences in seed-averaged AUC (first model minus second) with bootstrap 95% confidence intervals and two-sided bootstrap *P* values. Both models were evaluated on the same resamples; Kermany resamples patients

Comparison	RSNA test	Kermany	Indiana
Ensemble: top 5 by validation vs DINOv2-B, lr 1e-5	+0.001 (−0.001 to +0.003); 0.390	−0.002 (−0.005 to +0.001); 0.170	+0.015 (+0.007 to +0.023); <0.001
Ensemble: top 5 by validation vs DenseNet-121	+0.011 (+0.008 to +0.015); <0.001	+0.006 (0.000 to +0.012); 0.056	+0.032 (+0.019 to +0.048); <0.001
DINOv2-B, lr 1e-5 vs DenseNet-121	+0.010 (+0.006 to +0.015); <0.001	+0.008 (+0.001 to +0.015); 0.018	+0.017 (0.000 to +0.038); 0.056
Ensemble: top 5 by validation vs Ensemble: 7 baseline models	+0.014 (+0.011 to +0.018); <0.001	−0.014 (−0.018 to −0.009); <0.001	+0.036 (+0.025 to +0.047); <0.001
DINOv2-B, lr 1e-5 vs ConvNeXt-S	+0.004 (0.000 to +0.008); 0.026	+0.006 (+0.002 to +0.011); 0.006	−0.012 (−0.025 to +0.002); 0.094
DINOv2-B, lr 1e-5 vs Swin-T	+0.001 (−0.003 to +0.005); 0.532	+0.049 (+0.043 to +0.056); <0.001	−0.001 (−0.014 to +0.011); 0.828

Table S2 AUC (bootstrap 95% confidence interval) in subgroups of the RSNA test split. Five patients with recorded ages above 100 years are excluded from the age groups

Subgroup	n	Positive, %	Ensemble	DINOv2-B, lr 1e-5	DenseNet-121
Male	2,334	23.2	0.898 (0.884–0.912)	0.898 (0.884–0.910)	0.889 (0.873–0.903)
Female	1,669	21.6	0.911 (0.896–0.926)	0.910 (0.895–0.925)	0.899 (0.882–0.915)
Age <40 years	1,278	25.8	0.922 (0.906–0.936)	0.922 (0.907–0.938)	0.909 (0.892–0.926)
Age 40–59 years	1,781	21.1	0.901 (0.886–0.917)	0.898 (0.882–0.914)	0.893 (0.875–0.909)
Age ≥60 years	943	20.8	0.878 (0.854–0.903)	0.879 (0.853–0.904)	0.864 (0.837–0.889)
Anteroposterior (AP)	1,879	37.6	0.850 (0.832–0.867)	0.847 (0.828–0.865)	0.838 (0.819–0.856)
Posteroanterior (PA)	2,124	9.2	0.911 (0.890–0.930)	0.910 (0.889–0.929)	0.893 (0.870–0.916)

Table S3 Calibration in each cohort: prevalence, mean predicted risk before and after the Platt map fitted on RSNA validation, and expected calibration error (ECE, 15 bins)

Model	Cohort	Prevalence, %	Mean risk, raw, %	Mean risk, Platt, %	ECE, raw	ECE, Platt
Ensemble: top 5 by validation	RSNA test	22.5	39.8	23.0	0.173	0.019
Ensemble: top 5 by validation	Kermany	73.0	49.1	25.2	0.247	0.478
Ensemble: top 5 by validation	Indiana	5.8	16.9	5.0	0.112	0.011
DINOv2-B, lr 1e-5	RSNA test	22.5	32.4	23.1	0.099	0.016
DINOv2-B, lr 1e-5	Kermany	73.0	41.1	26.7	0.319	0.463
DINOv2-B, lr 1e-5	Indiana	5.8	9.3	5.0	0.036	0.018
DenseNet-121	RSNA test	22.5	35.9	23.2	0.134	0.018
DenseNet-121	Kermany	73.0	47.1	26.4	0.258	0.466
DenseNet-121	Indiana	5.8	14.6	7.7	0.088	0.025

Table S4 Subgroup contrasts in the external cohorts. Sensitivities at the Youden threshold fixed on RSNA validation

Model	Kermany				Indiana, AUC		
	AUC bacterial vs normal	AUC viral vs normal	Sens. bacterial	Sens. viral	vs normal	vs other abnormal	Opacity as positive
Ensemble: top 5 by validation	0.936	0.904	0.637	0.488	0.927	0.859	0.807
DINOv2-B, lr 1e-5	0.939	0.903	0.640	0.501	0.914	0.843	0.798
DenseNet-121	0.930	0.898	0.729	0.578	0.896	0.826	0.783
Ensemble: 7 baseline models	0.951	0.915	0.826	0.707	0.897	0.819	0.779

Table S5 Robust recipe: RSNA validation AUC (with test-time flip averaging) of the best epoch and of the stochastic weight average (SWA), mean over seeds, and the number of seeds in which the SWA average was kept

Backbone	Best epoch	SWA average	Difference	SWA kept
ConvNeXtV2-T	0.898	0.841	-0.057	0/3
DenseNet-121	0.889	0.852	-0.037	0/3
DINOv2-S	0.878	0.843	-0.035	0/3
EfficientNet-B0	0.884	0.835	-0.049	0/3
EfficientNetV2-S	0.885	0.828	-0.057	0/3
ResNet-50	0.884	0.854	-0.030	0/3
Swin-T	0.896	0.886	-0.010	0/3
DenseNet-121, CheXpert init	0.882	0.883	+0.001	3/3

Table S6 Errors at the Youden threshold fixed on RSNA validation. TP true positives, FN false negatives, TN true negatives, FP false positives

Cohort	TP	FN	TN	FP	Sens. / spec.	Breakdown
Ensemble: top 5 by validation						
RSNA test	757	145	2,489	612	0.839 / 0.803	FP rate: normal 0.7%, other abnormality 34.0%; AP sens./spec. 0.90/0.58, PA 0.61/0.94
Kermany	2,499	1,774	1,550	33	0.585 / 0.979	FN rate: bacterial 36.3%, viral 51.2%
Indiana	75	119	3,059	111	0.387 / 0.965	FP rate: normal 2.1%, other abnormal 4.5%
DINOv2-B, lr 1e-5						
RSNA test	745	157	2,504	597	0.826 / 0.807	FP rate: normal 0.8%, other abnormality 33.1%; AP sens./spec. 0.88/0.59, PA 0.62/0.94
Kermany	2,527	1,746	1,553	30	0.591 / 0.981	FN rate: bacterial 36.0%, viral 49.9%
Indiana	76	118	3,038	132	0.392 / 0.958	FP rate: normal 2.8%, other abnormal 5.2%

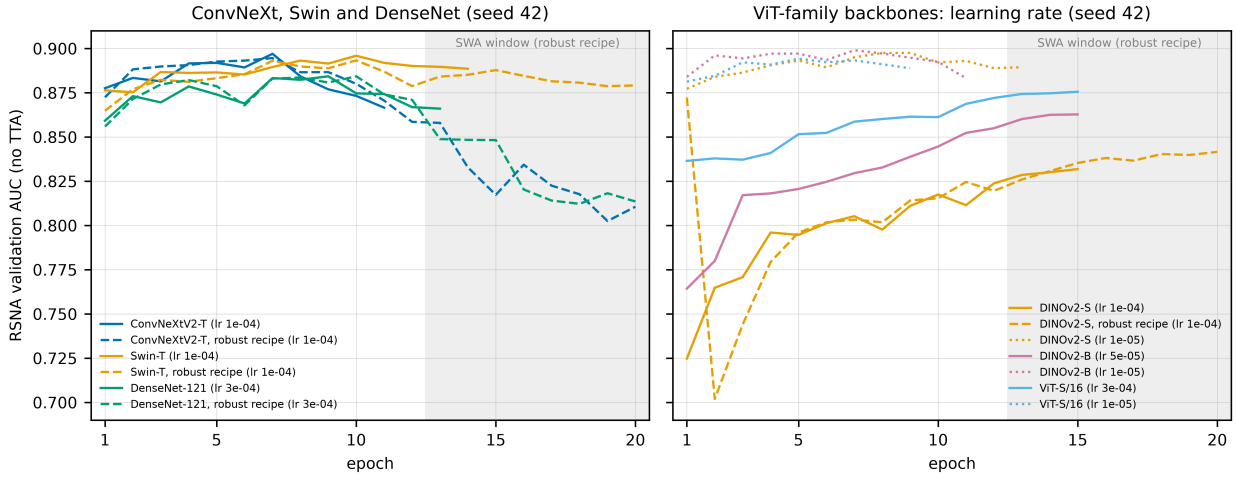


Figure S1 RSNA validation AUC per epoch (without test-time augmentation) for seed 42. Left: convolutional, ConvNeXt and Swin backbones under the standard (solid) and robust (dashed) recipes; validation AUC peaks early and, under the fixed 20-epoch robust schedule, declines during the stochastic weight averaging window. Right: ViT-family backbones at different learning rates. The first epoch of the robust recipe runs at a tenth of the base learning rate

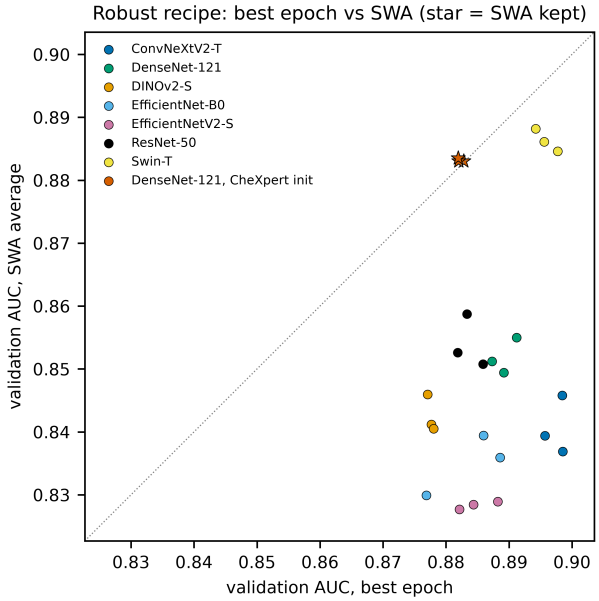


Figure S2 RSNA validation AUC of the best epoch against that of the stochastic weight average for every robust-recipe run. Stars mark runs in which the average was kept

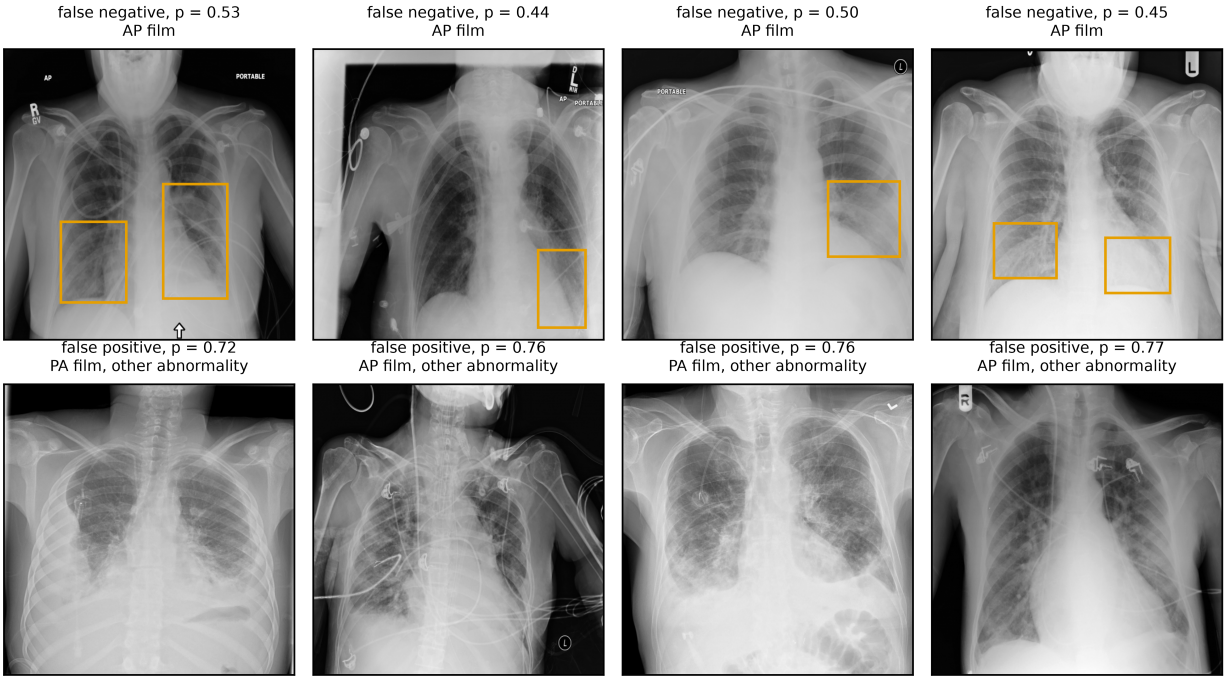


Figure S3 Randomly drawn false negatives (top, radiologist boxes in orange) and false positives (bottom) of the validation-selected ensemble on the RSNA test split at the Youden threshold fixed on validation; p is the ensemble probability. Almost all false positives were radiographs with other abnormalities

CLAIM 2024 Checklist

Item numbers follow the Checklist for Artificial Intelligence in Medical Imaging, 2024 update (Tejani et al., Radiol Artif Intell 2024). Topics are abbreviated; locations refer to sections of the manuscript.

Item	Topic	Reported	Where, or explanation
Title and abstract			
1	AI methodology identified in title or abstract	Yes	Title; Abstract
2	Summary of design, data, partitions, statistics and outcomes	Yes	Abstract (unstructured, as the journal requires)
Introduction			
3	Background, current practice and role of the AI approach	Yes	Introduction, paragraphs 1–5
4	Aims, objectives and hypotheses	Yes	Introduction, last paragraph
Methods: study design			
5	Prospective or retrospective	Yes	Study Design (retrospective)
6	Study goal and intended use	Yes	Introduction; Study Design. Research evaluation of pneumonia classification; not for clinical use
Methods: data			
7	Data sources and match to the target population	Yes	Datasets; Data Availability; Table 1; Discussion
8	Inclusion and exclusion criteria	Yes	Datasets (all labelled RSNA and Kermany images; one frontal image per Indiana study; ambiguous opacity group excluded from the main analysis)
9	Preprocessing	Yes	Image Preprocessing
10	Selection of data subsets	Yes	Datasets (frontal-image selection in Indiana)
11	De-identification	Yes	Public datasets de-identified by their providers; Supplementary Methods; Ethics Approval
12	Missing data	Yes	Datasets (ages above 100 years treated as missing; missing Indiana view labels taken from image-based labels)
13	Image acquisition protocol	No	Acquisition parameters are not distributed with the public datasets; view position is reported (Table 1); Discussion, limitations
Methods: reference standard			
14	Definition of the reference standard	Yes	Datasets; Supplementary Methods
15	Rationale, errors and limitations of the reference standard	Yes	Discussion, limitations
16	Annotators: sources, number, qualifications	Yes	Supplementary Methods
17	Annotation procedure for test data	Yes	Datasets; Supplementary Methods (Indiana labels from manually coded headings by a fixed rule)
18	Inter- and intrarater variability	No	Not reported by the dataset providers and not measured here; Supplementary Methods; Discussion
Methods: data partitions and testing data			
19	Partition into training, tuning and testing sets	Yes	Datasets; Table 1. The validation split serves as the tuning set
20	Level of disjointness	Yes	Datasets (patient level; one radiograph per patient in RSNA)
21	Test sample size	Yes	All available data; no a priori sample-size calculation (Datasets)
Methods: model and training			
22	Architecture	Yes	Backbones; Table 2
23	Software and hardware	Yes	Training; Statistical Analysis
24	Initialisation and transfer learning	Yes	Backbones; Table 2
25	Training details	Yes	Training
26	Selection of the final model	Yes	Model Selection and Ensembling
27	Ensembling	Yes	Model Selection and Ensembling
Methods: evaluation			
28	Metrics and their clinical relevance	Yes	Statistical Analysis
29	Uncertainty, statistical tests and software	Yes	Statistical Analysis
30	Robustness and sensitivity analyses	Yes	Three seeds per configuration; opacity sensitivity analysis; local threshold analysis (Results)
31	Explainability and its validation	Yes	Localisation (Grad-CAM against radiologist boxes)
32	Internal testing	Yes	Internal Discrimination (validation and test AUC agree)
33	External testing	Yes	External Discrimination and Transfer of Rankings; Operating Thresholds; Calibration
34	Trial registration	NA	Retrospective study of public data
Results			
35	Numbers included and excluded, flowchart	Yes	Fig. 1; Datasets; Table 1
36	Characteristics of each partition	Yes	Table 1 (age and sex are not distributed for the external cohorts)
37	Performance across partitions and subgroups	Yes	Tables 3 and 4; Fig. 2; Table S2
38	Estimates with confidence intervals, calibration, class imbalance	Yes	Tables 3 and 4; Fig. 4; Table S3; Training (class weights)
39	Failure analysis	Yes	Table S6; Fig. S3
Discussion and other information			
40	Limitations	Yes	Discussion, final paragraph
41	Implications for practice	Yes	Discussion; Conclusions
42	Full protocol	No	No registered protocol; all methods are described in the manuscript
43	Access to software, models and data	Yes	Data Availability; Code Availability

Item	Topic	Reported	Where, or explanation
44	Funding	Yes	Declarations