Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

**Deepak Kumar**

deepak.kumar@igf.edu.pl

Institute of Geophysics, Polish Academy of Sciences https://orcid.org/0000-0001-7677-860X

**Vandana Pandit**

United Medicity, Prayagraj, India

**Rakshita Ojha**

United Medicity, Prayagraj, India

**Research Article**

**Keywords:** Pneumonia, Chest radiography, Deep learning, External validation, Calibration, Vision transformer

**Posted Date:** November 12th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11686682/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare no competing interests.

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

Deepak Kumar[1*], Vandana Pandit[2] and Rakshita Ojha[2]

[1*]Institute of Geophysics, Polish Academy of Sciences, Księcia Janusza 64, Warsaw, 01-452, Poland.
[2]United Medicity, Prayagraj, Uttar Pradesh, India.

*Corresponding author(s). E-mail(s): deepak.kumar@igf.edu.pl;
Contributing authors: dr.vandanapandit@gmail.com;
rakshitaojha5may@gmail.com;

**Abstract**

Deep-learning classifiers for pneumonia on chest radiographs are usually chosen and reported on data from a single source, and whether the chosen model, its decision threshold and its probability estimates hold elsewhere is rarely tested. We trained 15 convolutional and transformer backbones in 28 configurations, three seeds each, on 18,678 radiographs from the Radiological Society of North America (RSNA) pneumonia dataset, and fixed the best single configuration, an ensemble of the five best configurations, decision thresholds and a calibration map on 4,003 validation radiographs. The frozen models were tested on 4,003 held-out RSNA radiographs and on two external public cohorts: 5,856 paediatric radiographs and 3,364 adult studies from Indiana University. A self-supervised DINOv2 transformer fine-tuned at a low learning rate was the best single configuration. The ensemble reached an area under the receiver operating characteristic curve (AUC) of 0.904 (95% confidence interval 0.893–0.914) internally, 0.925 (0.917–0.932) in children and 0.888 (0.867–0.909) in Indiana adults. Validation rankings of the configurations predicted their adult external ranking (Spearman $\boldsymbol{\rho = 0.92}$) much better than their paediatric one ($\boldsymbol{\rho = 0.49}$). The threshold fixed on validation detected 84% of pneumonia cases internally but only 59% in children and 39% in Indiana adults; re-estimating it on 10% of each external cohort restored sensitivity to 83% and 79%. Calibration fitted on validation held in Indiana but not in the high-prevalence paediatric cohort, and all models were less accurate on anteroposterior than posteroanterior films. External validation of such classifiers should report thresholds and calibration, not discrimination alone.

**Keywords:** Pneumonia, Chest radiography, Deep learning, External validation, Calibration, Vision transformer

# 1 Introduction

Lower respiratory infections caused an estimated 2.38 million deaths worldwide in 2016 [1], and chest radiography is the first imaging examination when pneumonia is suspected [2]. Its interpretation is variable, even among specialists [3], which has made pneumonia a natural target for automated reading. Deep neural networks trained on large public collections such as ChestX-ray14 [4], CheXpert [5], MIMIC-CXR [6] and the RSNA pneumonia dataset [7] detect pneumonia and other findings with high discrimination on held-out data from the same source [8, 9].

A large share of the pneumonia literature, however, trains and tests on a single paediatric collection [10] with random splits and reports accuracies above 95% [11–18]. Such internal figures are a weak guide to performance elsewhere. A pneumonia detector trained at one hospital system performed worse at outside sites and could identify the hospital from the radiograph alone [19]; a paediatric model with an internal AUC of 0.95 fell to 0.54 on adult radiographs [20]; and domain shift between chest radiograph collections is substantial [21, 22]. Only a minority of imaging artificial-intelligence studies validate externally [23], most of those that do report lower performance [24], and reviews of chest radiograph models list selection on the test set, single-source data and missing external evaluation among the commonest flaws [25, 26].

Model choice adds a further layer. Newer backbones such as ConvNeXt [27], ConvNeXt V2 [28], the Swin transformer [29], the vision transformer [30] and self-supervised DINOv2 encoders [31] outperform older convolutional networks on natural images, but ImageNet accuracy predicts chest radiograph performance poorly [32], transformers are sensitive to how they are fine-tuned [33, 34], and the benefit of ImageNet pretraining itself is debated [35]. Comparative studies usually report the configuration that scores best on their own test set; whether a ranking established on internal data carries over to other populations is seldom examined.

Discrimination is also only part of what a deployed classifier needs. A screening or triage tool operates at a fixed threshold, and its sensitivity and specificity at that threshold, like the probabilities it outputs, depend on case mix and prevalence, whereas the AUC does not [36–38]. Modern networks are, moreover, often miscalibrated [39]. Finally, chest radiograph models can exploit cues that correlate with the label without being part of the pathology, such as hospital markers or acquisition settings [19, 40, 41], and their performance can differ markedly between clinical subgroups [42–44]. In the RSNA data, anteroposterior films, typically bedside examinations of patients who cannot stand, carry a much higher pneumonia rate than posteroanterior films, which makes view position a candidate shortcut.

We therefore asked three questions. Which of a broad set of convolutional and transformer backbones and training recipes generalises best when every choice is made on internal validation data? Do discrimination, operating thresholds and calibration of the chosen models transfer to an adult and a paediatric external cohort? And does performance depend on view position, and where do the models look? We trained 15 backbones in 28 configurations with three seeds each, selected models, thresholds and calibration maps on RSNA validation data only, and evaluated

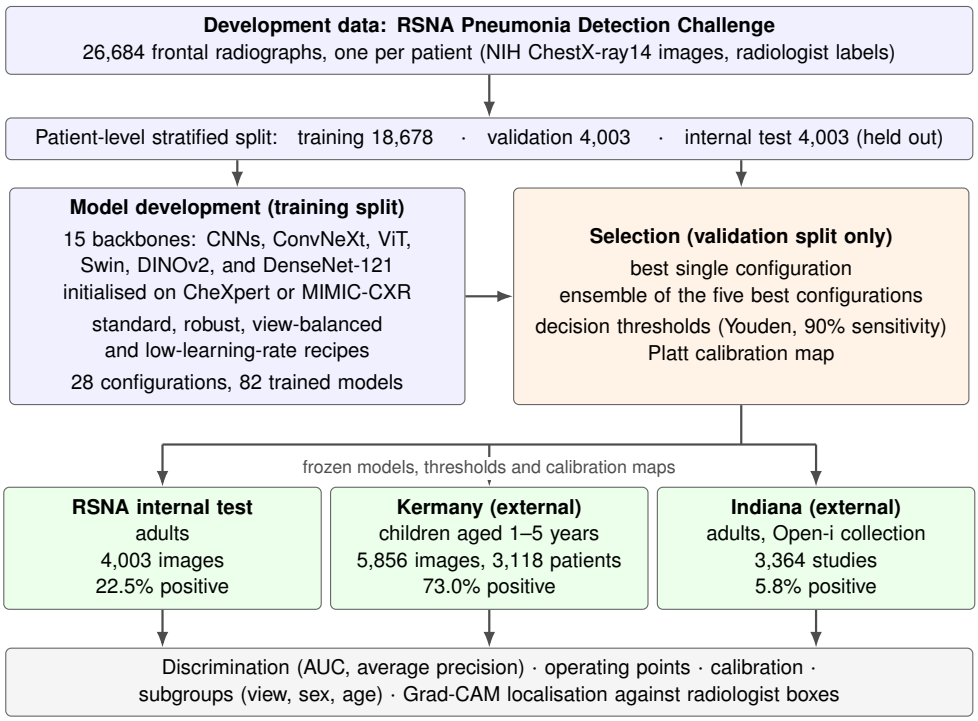



**Fig. 1** Study design. All choices (best single configuration, ensemble members, decision thresholds and calibration maps) were made on the RSNA validation split; the internal test split and both external cohorts were used only to evaluate frozen models

the frozen models on the RSNA test split and on two independent public cohorts. Reporting follows the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) 2024 update [45], whose completed checklist is provided as supplementary material, and the TRIPOD+AI guidance for prediction models [46].

# 2 Materials and Methods

## 2.1 Study Design

This retrospective diagnostic-accuracy study used three publicly available, de-identified datasets (Fig. 1). Models were developed on the RSNA pneumonia dataset and evaluated on its held-out test split and on two external cohorts that played no part in training, model selection, threshold setting or calibration, except in one explicitly labelled analysis of local threshold re-estimation.

## 2.2 Datasets

***Development data.***

The RSNA Pneumonia Detection Challenge dataset [7] contains 26,684 labelled frontal radiographs, one per patient, taken from the NIH ChestX-ray14 collection [4] and annotated by radiologists, who marked lung opacities suspicious for pneumonia with bounding boxes.

**Table 1** Characteristics of the development splits and the external cohorts

| | RSNA (development) | | | Kermany | Indiana |
|---|---|---|---|---|---|
| | Training | Validation | Test | (external) | (external) |
| Images | 18,678 | 4,003 | 4,003 | 5,856 | 3,364 |
| Patients | 18,678 | 4,003 | 4,003 | 3,118 | – |
| Pneumonia-positive, n (%) | 4,208 (22.5) | 902 (22.5) | 902 (22.5) | 4,273 (73.0) | 194 (5.8) |
| Age, median (IQR), years | 49 (34–59) | 49 (35–59) | 49 (34–59) | 1–5[a] | – |
| Female, % | 43.4 | 43.3 | 41.7 | – | – |
| AP view, n (%) | 8,427 (45.1) | 1,867 (46.6) | 1,879 (46.9) | – | 247 (6.7)[b] |
| Positive rate in AP / PA films, % | 38.6 / 9.4 | 37.9 / 9.1 | 37.6 / 9.2 | – | – |

AP anteroposterior, IQR interquartile range, PA posteroanterior. [a] Age range reported for the collection [10]; individual ages are not distributed. [b] Of all 3,685 frontal studies, before the ambiguous opacity group was removed.

Radiographs labelled "Lung Opacity" were positive (6,012, 22.5%); those labelled "Normal" (8,851) or "No Lung Opacity / Not Normal" (11,821) were negative. We split the data at the patient level, stratified by label, into training (18,678), validation (4,003) and test (4,003) sets (Table 1); in CLAIM terms the validation split is the tuning set, and the evaluation on the other cohorts is external testing. Age, sex and view position were read from the DICOM headers; five ages above 100 years were treated as missing.

***External cohort 1: paediatric.***

The Kermany collection [10, 47] contains anteroposterior radiographs of children aged one to five years from Guangzhou Women and Children's Medical Center, graded by experts as normal, bacterial pneumonia or viral pneumonia. We used all 5,856 images of version 2 (both of the original training and test folders) from 3,118 patients, identified from the file names; 1,583 were normal, 2,780 bacterial and 1,493 viral, so 4,273 (73.0%) were positive.

***External cohort 2: adult.***

The Indiana University collection of the National Library of Medicine Open-i service [48] contains chest radiographs with reports whose findings were coded manually as Medical Subject Headings (MeSH). We kept one frontal image per study (posteroanterior if available, otherwise anteroposterior, from the DICOM headers distributed with TorchXRayVision [49], whose image-based view labels filled gaps), giving 3,685 studies. Studies were grouped by their major MeSH headings into pneumonia-like (pneumonia, consolidation, infiltrate or airspace disease; 194), normal (1,345), other abnormal (1,825) and opacity without any pneumonia-like heading (321). The last group is ambiguous for pneumonia and was excluded from the main analysis, leaving 3,364 studies (5.8% positive), and counted as positive in a sensitivity analysis. As in RSNA, negatives therefore include radiographs with other abnormalities.

## 2.3 Image Preprocessing

RSNA DICOM files were decoded with pydicom; MONOCHROME1 images were inverted, pixel values were min–max scaled per image to 8 bits, and the 1,024 × 1,024 images were resized to 512 × 512 pixels. The external images, distributed as 8-bit image files, were converted to greyscale, padded to a square with their minimum value, min–max scaled and resized in the same way. Networks received 448 × 448 crops: random crops during training and the central

**Table 2** Backbones, pretraining and training configurations

| Backbone | Pretraining data | Parameters (M) | Learning rate | Recipes |
|---|---|---|---|---|
| Custom CNN | none (random initialisation) | 1.2 | 3e-4 | S |
| EfficientNet-B0 | ImageNet-1k | 4.0 | 3e-4 | S, R |
| DenseNet-121, CheXpert init | CheXpert radiographs | 6.9 | 3e-4 | S, R, V |
| DenseNet-121, MIMIC init | MIMIC-CXR radiographs | 6.9 | 3e-4 | S |
| DenseNet-121 | ImageNet-1k | 7.0 | 3e-4 | S, R, V |
| DenseNet-201 | ImageNet-1k | 18.1 | 3e-4 | S |
| EfficientNetV2-S | ImageNet-21k → 1k | 20.2 | 3e-4 | S, R |
| ViT-S/16 | ImageNet-21k → 1k | 21.9 | 3e-4, 1e-5 | S, $S_{low}$ |
| DINOv2-S | DINOv2 self-supervised, LVD-142M | 21.9 | 1e-4, 1e-5 | S, $S_{low}$, R |
| ResNet-50 | ImageNet-1k | 23.5 | 3e-4 | S, R |
| Swin-T | ImageNet-22k → 1k | 27.5 | 1e-4 | S, R |
| ConvNeXt-T | ImageNet-12k → 1k | 27.8 | 3e-4 | S |
| ConvNeXtV2-T | FCMAE, ImageNet-22k → 1k | 27.9 | 1e-4 | S, R |
| ConvNeXt-S | ImageNet-22k → 1k | 49.5 | 1e-4 | S |
| DINOv2-B | DINOv2 self-supervised, LVD-142M | 86.3 | 5e-5, 1e-5 | S, $S_{low}$ |

Recipes: S standard, R robust, V view-balanced sampling, $S_{low}$ standard recipe at a learning rate of $10^{-5}$. Parameters include the two-class head. LVD-142M is the curated image set used to train DINOv2.

crop at evaluation. Inputs were replicated to three channels and normalised with ImageNet statistics, except for the TorchXRayVision models, which received one channel scaled to $[-1024, 1024]$ as during their pretraining [49].

## 2.4 Backbones

Fifteen backbones covered five families (Table 2): a compact convolutional network trained from scratch; ImageNet-pretrained convolutional networks (ResNet-50 [50], DenseNet-121 and DenseNet-201 [51], EfficientNet-B0 [52] and EfficientNetV2-S [53]); ConvNeXt-T and ConvNeXt-S [27] and ConvNeXtV2-T [28]; transformers (ViT-S/16 [30], Swin-T [29], and DINOv2-S and DINOv2-B [31]); and DenseNet-121 pretrained on chest radiographs, either on CheXpert or on MIMIC-CXR [49]. Transformers were built for 448-pixel inputs with interpolated position embeddings. Because the RSNA images come from ChestX-ray14, we excluded encoders pretrained on that collection, such as RAD-DINO [54], to avoid leakage of test images into pretraining. ImageNet weights came from the timm library [55] and chest radiograph weights from TorchXRayVision [49].

## 2.5 Training

All networks ended in a two-class softmax and were trained with class-weighted cross-entropy (weights inversely proportional to class frequency), the AdamW optimiser [56] with weight decay $10^{-4}$, batches of 32, cosine learning-rate decay [57] and bfloat16 mixed precision on single NVIDIA A100 40-GB GPUs. Learning rates were set in advance by family, without tuning: $3 \times 10^{-4}$ for the convolutional networks and ViT-S/16, $10^{-4}$ for ConvNeXt, Swin-T and DINOv2-S, and $5 \times 10^{-5}$ for DINOv2-B.

***Standard recipe.***

Fifteen epochs with random crops and horizontal flips, stopping early after four epochs without improvement of the validation AUC and keeping the epoch with the highest validation AUC.

***Robust recipe.***

Applied to eight backbones: stronger augmentation (random crops covering 70–100% of the image with aspect ratios of 0.9–1.1, rotations within ±10°, contrast factors of 0.85–1.15, brightness shifts of ±0.08 and horizontal flips), label smoothing of 0.1 [58, 59], one warm-up epoch at a tenth of the learning rate, 20 epochs without early stopping, stochastic weight averaging (SWA) over the last 40% of the epochs with recomputed batch-normalisation statistics [60], and test-time averaging over the image and its mirror. Each run kept whichever of the best-validation epoch and the SWA average had the higher validation AUC.

***View-balanced sampling.***

To test the view-position shortcut, DenseNet-121 and the CheXpert DenseNet-121 were also trained with sampling weights that equalised the four combinations of view position and label in every epoch.

***Lower learning rate for the ViT family.***

The validation curves of the first two stages showed that ViT-family backbones degraded once the base learning rate took effect (Results), so ViT-S/16, DINOv2-S and DINOv2-B were also trained with the standard recipe at $10^{-5}$. This choice was made from validation curves alone.

Configurations were added in three stages (baseline panel; additional backbones and the robust recipe; lower learning rate for the ViT family), each decided from validation results. Every configuration was trained with seeds 42, 43 and 44, except DINOv2-B with the standard recipe, which was stopped after one seed because its validation AUC was clearly below the leading configurations.

## 2.6 Model Selection and Ensembling

All selection used the RSNA validation split. The best single configuration was the one with the highest mean validation AUC over seeds. The ensemble averaged the softmax outputs of the five configurations with the highest mean validation AUC, excluding view-balanced variants, separately for each seed. As comparators we kept two ensembles fixed before the later stages: the seven baseline configurations (ImageNet ResNet-50, DenseNet-121, EfficientNet-B0, ConvNeXt-T and ViT-S/16, and the two chest radiograph DenseNets) and the five ImageNet baselines among them. Two decision thresholds were fixed on validation, the one maximising Youden's index [61] and the highest threshold giving at least 90% sensitivity, together with a Platt recalibration map [62], all from seed-averaged probabilities.

## 2.7 Statistical Analysis

The primary measure was the AUC; average precision summarised the precision–recall curve. Results are given as the mean ± standard deviation over seeds and, for confidence intervals and tests, from the seed-averaged probabilities. Percentile bootstrap 95% confidence intervals used 1,000 replicates [63], resampling patients rather than images in the Kermany cohort because some children contribute several images. Correlated AUCs on the RSNA test set were compared with the DeLong test [64, 65]; AUC differences on all three test sets were assessed with a paired bootstrap that used the same resamples for both models. Agreement between the rankings of the single configurations in different cohorts was measured with Spearman's

rank correlation of their mean AUCs. Calibration was summarised by the expected calibration error (ECE) with 15 equal-width bins [39, 66], before and after temperature scaling [39] and Platt scaling fitted on validation. To estimate what local recalibration of the threshold would achieve, we drew a random 10% of each external cohort (of patients for Kermany), chose the Youden threshold on that sample, evaluated it on the remaining 90%, and repeated this 500 times. Subgroups were sex, age (<40, 40–59 and ≥60 years) and view position on the RSNA test set; the discriminative value of view position itself was measured as the AUC of the view indicator alone. Secondary contrasts were normal versus pneumonia in RSNA, bacterial and viral pneumonia in Kermany, and pneumonia-like versus normal and versus other abnormal studies in Indiana. *P* values are two-sided and were not adjusted for multiple comparisons; they are descriptive. Analyses used Python 3.11, PyTorch 2.5.1 [67], timm 1.0.31 [55], TorchXRayVision 1.5.7 [49], scikit-learn [68] and SciPy.

### 2.8 Localisation

For the best single configuration (DINOv2-B), ConvNeXtV2-T with the robust recipe, ConvNeXt-S and DenseNet-121 (seed 42), gradient-weighted class activation maps (Grad-CAM) [69] of the pneumonia log-odds were computed for all 902 positive RSNA test radiographs, all of which carry radiologist boxes. For the convolutional networks the maps came from the last convolutional stage; for DINOv2-B they came from the patch tokens entering its last transformer block, arranged on the $32 \times 32$ patch grid. We recorded whether the map's maximum fell inside a box (the pointing game [70]) and the share of map energy inside the boxes; a map with no spatial information would score the mean box-area share, 15.3%. Models were compared with the exact McNemar test for hits and the Wilcoxon signed-rank test for energy shares.

## 3 Results

### 3.1 Cohorts

Pneumonia prevalence was 22.5% in every RSNA split, 73.0% in the paediatric cohort and 5.8% in the Indiana cohort (Table 1). In the RSNA test split 46.9% of films were anteroposterior, and their pneumonia rate was four times that of posteroanterior films (37.6% versus 9.2%). The view indicator alone discriminated pneumonia with an AUC of 0.703.

### 3.2 Internal Discrimination

On the RSNA test split the mean AUC of the single configurations ranged from 0.828 (DINOv2-S, standard recipe) to 0.898 (Table 3, Fig. 2). The five configurations with the highest validation AUC were DINOv2-B and DINOv2-S trained at the lower learning rate, ConvNeXtV2-T and Swin-T with the robust recipe, and ConvNeXt-S. The best single configuration was DINOv2-B at $10^{-5}$ (validation AUC 0.899; test AUC 0.898 ± 0.003). The ensemble of the five reached 0.903 (standard deviation below 0.001), or 0.904 (0.893–0.914) from seed-averaged probabilities, and did not differ significantly from the best single configuration (DeLong $P = 0.40$). The best single configuration was better than DenseNet-121 (0.903 versus 0.893 from seed-averaged probabilities, $P < 0.001$), the backbone of CheXNet and a common baseline in the pneumonia literature [8]. The chest radiograph-pretrained DenseNets (mean test AUC 0.876 and 0.868)

did not improve on the ImageNet DenseNet-121 (0.885), and the network trained from scratch reached 0.857. Discrimination between normal radiographs and pneumonia was near-perfect (ensemble AUC 0.988); most errors involved the "No Lung Opacity / Not Normal" group.

## 3.3 External Discrimination and Transfer of Rankings

Without any adaptation, the ensemble reached an AUC of 0.925 (0.917–0.932) in the paediatric cohort and 0.888 (0.867–0.909) in the Indiana cohort (Fig. 3). It matched the best single configuration in children (paired difference −0.002, $P = 0.17$) and was better in Indiana adults (+0.015, $P < 0.001$; Supplementary Table S1).

Validation rankings carried over to the adult external cohort far better than to the paediatric one. Across the 28 single configurations, the Spearman correlation of mean validation AUC with mean test AUC was 0.96 and with mean Indiana AUC 0.92, but with mean Kermany AUC only 0.49 ($P = 0.008$). The configurations best suited to children ranked in the middle internally: EfficientNetV2-S with the robust recipe (Kermany AUC 0.926 ± 0.023) and the CheXpert DenseNet-121 with the robust recipe (0.916 ± 0.008), and the seven-model baseline ensemble reached 0.934 ± 0.005. The same models were weaker in Indiana adults (0.828, 0.825 and 0.848). Compared with the baseline ensemble, the validation-selected ensemble was 0.014 better internally and 0.036 better in Indiana adults, but 0.014 worse in children (all $P < 0.001$). Seed-to-seed variation was also larger externally: the standard deviation of the Kermany AUC reached 0.078, against at most 0.014 internally.

Within the external cohorts, the ensemble separated bacterial pneumonia from normal radiographs better than viral pneumonia (AUC 0.936 versus 0.904), and pneumonia-like Indiana studies from normal studies (0.927) better than from studies with other abnormalities (0.859; Supplementary Table S4). Counting the ambiguous opacity studies as positive lowered the Indiana AUC to 0.807.

## 3.4 Operating Thresholds

At the Youden threshold fixed on validation, the ensemble had a sensitivity of 0.839 and a specificity of 0.803 on the RSNA test split (Table 4). The same threshold detected only 58.5% of pneumonia cases in children and 38.7% in Indiana adults, while specificity rose to 0.979 and 0.965. The threshold chosen for 90% sensitivity on validation gave a sensitivity of 0.906 internally but 0.778 and 0.572 externally. In children, sensitivity was lower for viral than for bacterial pneumonia (0.488 versus 0.637). With the low prevalence of Indiana, the positive predictive value at the Youden threshold was 0.403. Re-estimating the threshold on a random 10% of each external cohort restored balanced performance: sensitivity 0.834 and specificity 0.865 in children, and 0.791 and 0.826 in Indiana adults, although the Indiana estimates varied widely between samples (sensitivity 0.55–0.94) because a 10% sample holds only about 19 positive studies. The best single configuration behaved in the same way.

Errors at the Youden threshold had a clear structure (Supplementary Table S6 and Fig. S3). On the RSNA test split, 603 of the 612 false positives were radiographs with other abnormalities: the false-positive rate was 34.0% in the "No Lung Opacity / Not Normal" group against 0.7% among normal radiographs. In Indiana, too, false positives were commoner among studies with other abnormalities than among normal studies (4.5% versus 2.1%). The single threshold

also produced different operating points by view: sensitivity 0.90 and specificity 0.58 on anteroposterior films, against 0.61 and 0.94 on posteroanterior films.

## 3.5 Calibration

Raw outputs overestimated risk: the mean predicted probability of the ensemble on the RSNA test split was 0.398 against a prevalence of 0.225, the result of class-weighted training (Fig. 4). The ECE was 0.173 for the ensemble, 0.11–0.18 for configurations trained with the standard recipe and 0.16–0.24 for those trained with the robust recipe, which adds label smoothing; DINOv2-B at the lower learning rate had the lowest raw ECE (0.107). Temperature scaling barely changed the ensemble's ECE (0.165) because it cannot correct a shift in the mean, whereas Platt scaling fitted on validation reduced it to 0.019 on the RSNA test split and to 0.011 in the Indiana cohort. In the paediatric cohort the same map made matters worse (ECE 0.478; mean predicted risk 0.252 against a prevalence of 0.730; Supplementary Table S3).

## 3.6 Subgroups and the View-Position Shortcut

Every configuration discriminated worse on anteroposterior than on posteroanterior films (Fig. 5); for the ensemble the AUC was 0.850 (0.832–0.867) against 0.911 (0.890–0.930). View-balanced sampling did not close this gap. It lowered the internal AUC of DenseNet-121 from 0.885 to 0.861 and of the CheXpert DenseNet-121 from 0.876 to 0.845, and their paediatric AUC from 0.887 to 0.788 and from 0.893 to 0.801. The confidence intervals for women and men overlapped (AUC 0.911 and 0.898), whereas discrimination declined with age, from 0.922 below 40 years to 0.878 at 60 years or more (Supplementary Table S2).

## 3.7 Training Recipes

The robust recipe helped some backbones and not others (Table 3): it raised the test AUC of DINOv2-S from 0.828 to 0.874, of the CheXpert DenseNet-121 from 0.876 to 0.885, of ResNet-50 from 0.878 to 0.885 and of DenseNet-121 from 0.885 to 0.889, but lowered those of EfficientNet-B0 (0.889 to 0.881) and Swin-T (0.897 to 0.893). The SWA average was worse on validation than the best epoch for every backbone except the CheXpert DenseNet-121, by 0.02 to 0.06 (Supplementary Table S5 and Fig. S2). Validation AUC of these backbones peaked between epochs 3 and 9 and declined during the averaging window (Supplementary Fig. S1), so the average combined overfitted weights. For the CheXpert DenseNet-121 the two were equal and the average was kept in all three seeds. The ViT-family backbones behaved differently. DINOv2-S reached a validation AUC of 0.87 after the warm-up epoch at $10^{-5}$ and fell to 0.70 one epoch later at $10^{-4}$, recovering only slowly. Trained at $10^{-5}$ throughout, all three improved markedly (Supplementary Fig. S1): the test AUC of DINOv2-S rose from 0.828 to 0.896, of DINOv2-B from 0.853 (one seed) to 0.898 and of ViT-S/16 from 0.865 to 0.891, with Kermany AUCs of 0.899–0.906, and DINOv2-B became the best single configuration.

## 3.8 Localisation

The Grad-CAM maximum fell inside a radiologist box in 85.0% of positive test radiographs for DINOv2-B, 82.5% for ConvNeXtV2-T, 77.7% for ConvNeXt-S and 69.2% for DenseNet-121, against 15.3% for a map without spatial information (Table 5, Fig. 6). The median share of

map energy inside the boxes was 78.1%, 66.4%, 57.9% and 35.2%. All three newer backbones localised better than DenseNet-121 ($P < 10^{-7}$ for hits, $P < 10^{-100}$ for energy). DINOv2-B and ConvNeXtV2-T did not differ in hits ($P = 0.07$), but DINOv2-B placed more of its map inside the boxes ($P < 10^{-50}$). The maps of DenseNet-121 often extended outside the lungs, including onto burned-in text markers such as the labels of portable anteroposterior films.

# 4 Discussion

This study separates three properties that a pneumonia classifier needs in order to work outside its development data, and finds that they transfer to very different degrees. Discrimination transferred well to an adult external cohort and reasonably to a paediatric one: an ensemble chosen on RSNA validation data reached an AUC near 0.9 in both without adaptation. Rankings of configurations transferred to adults but not to children. Operating thresholds and calibration, which determine how a classifier behaves in practice, did not transfer: the threshold that detected 84% of pneumonia cases internally detected 59% in children and 39% in Indiana adults.

The loss of sensitivity follows from how a fixed threshold interacts with a new population. The AUC depends only on how the scores of positive and negative cases are ordered, whereas a threshold depends on where those scores lie, and their location shifts with case mix, image appearance and prevalence [36, 38]. Paediatric radiographs, with their different anatomy and disease spectrum, and the low prevalence and heterogeneous negatives of a general adult population both moved the score distributions relative to RSNA. A small local sample was enough to restore balanced operation, which supports re-estimating thresholds at each site before use [71]; for the low-prevalence cohort, however, a 10% sample yielded unstable thresholds, so sites with few positive cases will need larger samples or pooled data. The same reasoning applies to calibration. Platt scaling fixed on validation transferred to Indiana, where prevalence was lower, but not to the paediatric cohort, where prevalence was 73%; correcting for a known change in prevalence [37] or recalibrating locally would be required there. Reporting only the AUC of external validation, as is common [24], therefore hides the property most likely to fail.

The divergence of rankings between adults and children is a practical warning for model selection. Validation rankings predicted Indiana rankings closely, and the validation-selected ensemble was also the best model in Indiana adults, but the paediatric cohort favoured different configurations, including EfficientNetV2-S and DenseNets pretrained on chest radiographs, and the older baseline ensemble. Children differ in anatomy, disease spectrum and radiographic technique, and the Kermany labels distinguish bacterial and viral pneumonia rather than lung opacity; validation on adult data is not a sufficient basis for choosing a model for children. Seed-to-seed variation was also much larger externally, so external results from a single training run can mislead: the three seeds of DenseNet-121 with the robust recipe gave Kermany AUCs of 0.759, 0.867 and 0.910 with near-identical internal results. Interestingly, models developed on the heterogeneous adult RSNA data reached an AUC near 0.9 in children, whereas a model developed on the paediatric data has been reported to fall to 0.54 on adult radiographs [20].

Among architectures, self-supervised DINOv2 transformers fine-tuned at a low learning rate and ConvNeXt-family backbones led internally and in adults, and the best single configuration, DINOv2-B, was significantly better than DenseNet-121 internally and in children. Pretraining on chest radiographs did not improve internal discrimination, but the CheXpert-initialised DenseNet-121 transferred to children better than its ImageNet counterpart (Kermany AUC

0.916 versus 0.845 with the robust recipe), in line with reports that ImageNet performance is a poor guide to chest radiograph performance [32]. The ensemble matched the best single configuration internally and in children and was better in Indiana adults, the more heterogeneous external cohort, as expected of ensembles under dataset shift [72, 73], at five times the inference cost. Training details mattered as much as architecture. SWA failed because these networks overfitted RSNA after a few epochs, so averaging the last 40% of a fixed schedule combined overfitted weights; choosing the averaging window on validation data would avoid this. The learning rate decided the fate of the ViT-family backbones: at rates that suited convolutional networks they degraded, consistent with work on feature distortion during fine-tuning [34], whereas at $10^{-5}$ DINOv2-B became the best single configuration. A comparison that used one learning rate for all backbones would have ranked these transformers near the bottom.

View position behaved like a shortcut candidate. On its own it discriminated pneumonia with an AUC of 0.70, every model was less accurate on anteroposterior films, and DenseNet-121 maps often lay on burned-in markers of portable films, echoing earlier reports of models that read acquisition cues rather than pathology [19, 40]. Balancing view and label during training did not remove the gap and harmed external performance, which suggests that anteroposterior films are genuinely harder, with supine positioning, lines and overlapping opacities, rather than that the models relied on view alone. Performance should be reported separately by view, as recommended for other clinically meaningful subgroups [42, 43]. The better localisation of the DINOv2 and ConvNeXt models agrees with their higher discrimination, but saliency maps remain an imperfect account of a model's reasoning [74, 75].

The study has limitations. All models were developed on one source, NIH radiographs annotated for the RSNA challenge, and RSNA "lung opacity" labels, Kermany expert grades and Indiana MeSH codes define pneumonia differently, so part of the external drop reflects label definitions rather than model failure. The external images were 8-bit exports and all images were min–max scaled without DICOM lookup tables, a step that affects generalisation [76]. Hyperparameters were fixed in advance by family rather than tuned for each backbone, and the lower learning rate for the ViT family was chosen after seeing validation curves. We did not compare the models with radiologists, Grad-CAM was computed for one seed of four models, and $P$ values were not adjusted for multiplicity. The study is retrospective, and clinical benefit would require prospective evaluation.

# 5 Conclusions

Pneumonia classifiers selected on internal validation data kept their discrimination on adult and paediatric external cohorts, but their decision thresholds and probability calibration did not transfer, and rankings established on adult data did not predict which models were best for children. Self-supervised DINOv2 transformers fine-tuned at a low learning rate, ConvNeXt-family backbones and a validation-selected ensemble of both gave the best internal and adult external results. External validation of such classifiers should report sensitivity and specificity at prespecified thresholds, calibration and view-stratified performance, and deployment should include local re-estimation of the threshold.

**Supplementary Information.** Supplementary Methods on the reference standards, Supplementary Tables S1–S6 (paired AUC differences, subgroup AUCs, calibration in each cohort, external subgroup contrasts, the SWA comparison and errors at the validation threshold),

Supplementary Figs. S1–S3 (training curves, best epoch versus SWA average, examples of errors) and the completed CLAIM 2024 checklist are available online.

**Acknowledgements.** We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2026/020015.

# Declarations

**Funding.** This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

**Competing Interests.** The authors declare no competing interests.

**Author Contributions.** Deepak Kumar: conceptualisation, methodology, software, formal analysis, visualisation, writing of the original draft. Vandana Pandit: conceptualisation, clinical interpretation, validation, supervision, review and editing. Rakshita Ojha: investigation, validation, review and editing. All authors read and approved the final manuscript.

**Ethics Approval.** Not applicable. This retrospective study analysed publicly available, de-identified datasets; no patients were contacted and no new data were collected.

**Consent to Participate.** Not applicable.

**Consent to Publish.** Not applicable. The radiographs shown are from a public de-identified collection.

**Data Availability.** The RSNA Pneumonia Detection Challenge data are available from https://www.kaggle.com/c/rsna-pneumonia-detection-challenge, the paediatric collection from Mendeley Data (https://doi.org/10.17632/rscbjbr9sj.2, CC BY 4.0) and the Indiana University collection from the National Library of Medicine Open-i service (https://openi.nlm.nih.gov, CC BY-NC-ND 4.0). The patient-level split and per-image predictions are available from the corresponding author on reasonable request.

**Code Availability.** The code for preprocessing, training, evaluation and figures is available from the corresponding author on reasonable request, including to reviewers during peer review, and will be made public on acceptance.

# References

[1] GBD 2016 Lower Respiratory Infections Collaborators (2018) Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016. Lancet Infect Dis 18(11):1191–1210. https://doi.org/10.1016/s1473-3099(18)30310-4

[2] Franquet T (2018) Imaging of community-acquired pneumonia. J Thorac Imaging 33(5):282–294. https://doi.org/10.1097/rti.0000000000000347

[3] Neuman MI, Lee EY, Bixby S, et al (2011) Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children. J Hosp Med 7(4):294–298. https://doi.org/10.1002/jhm.955

[4] Wang X, Peng Y, Lu L, et al (2017) Chestx-ray8: Hospital-scale chest X-Ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 3462–3471, https://doi.org/10.1109/cvpr.2017.369

[5] Irvin J, Rajpurkar P, Ko M, et al (2019) CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proc AAAI Conf Artif Intell 33(01):590–597. https://doi.org/10.1609/aaai.v33i01.3301590

[6] Johnson AEW, Pollard TJ, Berkowitz SJ, et al (2019) MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 6(1):317. https://doi.org/10.1038/s41597-019-0322-0

[7] Shih G, Wu CC, Halabi SS, et al (2019) Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiol Artif Intell 1(1):e180041. https://doi.org/10.1148/ryai.2019180041

[8] Rajpurkar P, Irvin J, Zhu K, et al (2017) CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:171105225 https://doi.org/10.48550/arxiv.1711.05225

[9] Rajpurkar P, Irvin J, Ball RL, et al (2018) Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 15(11):e1002686. https://doi.org/10.1371/journal.pmed.1002686

[10] Kermany DS, Goldbaum M, Cai W, et al (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172(5):1122–1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

[11] Stephen O, Sain M, Maduh UJ, et al (2019) An efficient deep learning approach to pneumonia classification in healthcare. J Healthc Eng 2019:1–7. https://doi.org/10.1155/2019/4180949

[12] Rahman T, Chowdhury MEH, Khandakar A, et al (2020) Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray. Appl Sci 10(9):3233. https://doi.org/10.3390/app10093233

[13] Chouhan V, Singh SK, Khamparia A, et al (2020) A novel transfer learning based approach for pneumonia detection in chest X-ray images. Appl Sci 10(2):559. https://doi.org/10.3390/app10020559

[14] Liang G, Zheng L (2020) A transfer learning method with deep residual network for pediatric pneumonia diagnosis. Comput Methods Programs Biomed 187:104964.

https://doi.org/10.1016/j.cmpb.2019.06.023

[15] Hashmi MF, Katiyar S, Keskar AG, et al (2020) Efficient pneumonia detection in chest Xray images using deep transfer learning. Diagnostics (Basel) 10(6):417. https://doi.org/10.3390/diagnostics10060417

[16] Kundu R, Das R, Geem ZW, et al (2021) Pneumonia detection in chest X-ray images using an ensemble of deep learning models. PLoS One 16(9):e0256630. https://doi.org/10.1371/journal.pone.0256630

[17] Mabrouk A, Díaz Redondo RP, Dahou A, et al (2022) Pneumonia detection on chest X-ray images using ensemble of deep convolutional neural networks. Appl Sci 12(13):6448. https://doi.org/10.3390/app12136448

[18] Szepesi P, Szilágyi L (2022) Detection of pneumonia using convolutional neural networks and deep learning. Biocybern Biomed Eng 42(3):1012–1022. https://doi.org/10.1016/j.bbe.2022.08.001

[19] Zech JR, Badgeley MA, Liu M, et al (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 15(11):e1002683. https://doi.org/10.1371/journal.pmed.1002683

[20] Xin KZ, Li D, Yi PH (2021) Limited generalizability of deep learning algorithm for pediatric pneumonia classification on external data. Emerg Radiol 29(1):107–113. https://doi.org/10.1007/s10140-021-01954-x

[21] Pooch EHP, Ballester P, Barros RC (2020) Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In: Thoracic Image Analysis (TIA 2020), Lecture Notes in Computer Science, vol 12502. Springer, Cham, pp 74–83, https://doi.org/10.1007/978-3-030-62469-9_7

[22] Cohen JP, Hashir M, Brooks R, et al (2020) On the limits of cross-domain generalization in automated X-ray prediction. In: Proceedings of the Third Conference on Medical Imaging with Deep Learning, pp 136–155

[23] Kim DW, Jang HY, Kim KW, et al (2019) Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: Results from recently published papers. Korean J Radiol 20(3):405. https://doi.org/10.3348/kjr.2019.0025

[24] Yu AC, Mohajer B, Eng J (2022) External validation of deep learning algorithms for radiologic diagnosis: A systematic review. Radiol Artif Intell 4(3):e210064. https://doi.org/10.1148/ryai.210064

[25] Roberts M, Driggs D, Thorpe M, et al (2021) Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and ct scans. Nat Mach Intell 3(3):199–217. https://doi.org/10.1038/s42256-021-00307-0

[26] Varoquaux G, Cheplygina V (2022) Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med 5(1):48. https://doi.org/10.1038/s41746-022-00592-y

[27] Liu Z, Mao H, Wu CY, et al (2022) A ConvNet for the 2020s. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 11966–11976, https://doi.org/10.1109/cvpr52688.2022.01167

[28] Woo S, Debnath S, Hu R, et al (2023) ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 16133–16142, https://doi.org/10.1109/cvpr52729.2023.01548

[29] Liu Z, Lin Y, Cao Y, et al (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, pp 9992–10002, https://doi.org/10.1109/iccv48922.2021.00986

[30] Dosovitskiy A, Beyer L, Kolesnikov A, et al (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR), arXiv:2010.11929

[31] Oquab M, Darcet T, Moutakanni T, et al (2024) DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research ArXiv:2304.07193

[32] Ke A, Ellsworth W, Banerjee O, et al (2021) CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-Ray interpretation. In: Proceedings of the Conference on Health, Inference, and Learning. ACM, ACM CHIL '21, pp 116–124, https://doi.org/10.1145/3450439.3451867

[33] Usman M, Zia T, Tariq A (2022) Analyzing transfer learning of vision transformers for interpreting chest radiography. J Digit Imaging 35(6):1445–1462. https://doi.org/10.1007/s10278-022-00666-z

[34] Kumar A, Raghunathan A, Jones R, et al (2022) Fine-tuning can distort pretrained features and underperform out-of-distribution. In: International Conference on Learning Representations (ICLR), arXiv:2202.10054

[35] Raghu M, Zhang C, Kleinberg J, et al (2019) Transfusion: Understanding transfer learning for medical imaging. In: Advances in Neural Information Processing Systems, pp 3342–3352

[36] and Van Calster B, McLernon DJ, van Smeden M, et al (2019) Calibration: the achilles heel of predictive analytics. BMC Med 17(1):230. https://doi.org/10.1186/s12916-019-1466-7

[37] Saerens M, Latinne P, Decaestecker C (2002) Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. Neural Comput 14(1):21–41. https://doi.org/10.1162/089976602753284446

[38] Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, et al (2012) A unifying view on dataset shift in classification. Pattern Recognit 45(1):521–530. https://doi.org/10.1016/j.patcog.2011.06.019

[39] Guo C, Pleiss G, Sun Y, et al (2017) On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, pp 1321–1330

[40] DeGrave AJ, Janizek JD, Lee SI (2021) AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell 3(7):610–619. https://doi.org/10.1038/s42256-021-00338-7

[41] Geirhos R, Jacobsen JH, Michaelis C, et al (2020) Shortcut learning in deep neural networks. Nat Mach Intell 2(11):665–673. https://doi.org/10.1038/s42256-020-00257-z

[42] Ahluwalia M, Abdalla M, Sanayei J, et al (2023) The subgroup imperative: Chest radiograph classifier generalization gaps in patient, setting, and pathology subgroups. Radiol Artif Intell 5(5):e220270. https://doi.org/10.1148/ryai.220270

[43] Oakden-Rayner L, Dunnmon J, Carneiro G, et al (2020) Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM Conference on Health, Inference, and Learning. ACM, ACM CHIL '20, pp 151–159, https://doi.org/10.1145/3368555.3384468

[44] Seyyed-Kalantari L, Zhang H, McDermott MBA, et al (2021) Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 27(12):2176–2182. https://doi.org/10.1038/s41591-021-01595-0

[45] Tejani AS, Klontzas ME, Gatti AA, et al (2024) Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. Radiol Artif Intell 6(4):e240300. https://doi.org/10.1148/ryai.240300

[46] Collins GS, Moons KGM, Dhiman P, et al (2024) TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 385:e078378. https://doi.org/10.1136/bmj-2023-078378

[47] Kermany D, Zhang K, Goldbaum M (2018) Labeled optical coherence tomography (OCT) and chest X-Ray images for classification. Mendeley Data, version 2, https://doi.org/10.17632/rscbjbr9sj.2

[48] Demner-Fushman D, Kohli MD, Rosenman MB, et al (2015) Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc 23(2):304–310. https://doi.org/10.1093/jamia/ocv080

[49] Cohen JP, Viviano JD, Bertin P, et al (2022) TorchXRayVision: A library of chest X-ray datasets and models. In: Proceedings of the 5th International Conference on Medical Imaging with Deep Learning, pp 231–249

[50] He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 770–778, https://doi.org/10.1109/cvpr.2016.90

[51] Huang G, Liu Z, Van Der Maaten L, et al (2017) Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 2261–2269, https://doi.org/10.1109/cvpr.2017.243

[52] Tan M, Le QV (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, pp 6105–6114

[53] Tan M, Le QV (2021) EfficientNetV2: Smaller models and faster training. In: Proceedings of the 38th International Conference on Machine Learning, pp 10096–10106

[54] Pérez-García F, Sharma H, Bond-Taylor S, et al (2025) Exploring scalable medical image encoders beyond text supervision. Nat Mach Intell 7(1):119–130. https://doi.org/10.1038/s42256-024-00965-w

[55] Wightman R (2019) PyTorch image models. GitHub repository, https://github.com/huggingface/pytorch-image-models, https://doi.org/10.5281/zenodo.4414861

[56] Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR), arXiv:1711.05101

[57] Loshchilov I, Hutter F (2017) SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (ICLR), arXiv:1608.03983

[58] Szegedy C, Vanhoucke V, Ioffe S, et al (2016) Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 2818–2826, https://doi.org/10.1109/cvpr.2016.308

[59] Müller R, Kornblith S, Hinton GE (2019) When does label smoothing help? In: Advances in Neural Information Processing Systems

[60] Izmailov P, Podoprikhin D, Garipov T, et al (2018) Averaging weights leads to wider optima and better generalization. In: Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI), arXiv:1803.05407

[61] Youden WJ (1950) Index for rating diagnostic tests. Cancer 3(1):32–35. https://doi.org/10.1002/1097-0142(1950)3:13.0.co;2-3

[62] Platt JC (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola AJ, Bartlett P, Schölkopf B, et al (eds) Advances in Large Margin Classifiers. MIT Press, Cambridge, MA, p 61–74

[63] Efron B, Tibshirani R (1994) An Introduction to the Bootstrap. Chapman and Hall/CRC, https://doi.org/10.1201/9780429246593

[64] DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 44(3):837. https://doi.org/10.2307/2531595

[65] Sun X, Xu W (2014) Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Process Lett 21(11):1389–1393. https://doi.org/10.1109/lsp.2014.2337313

[66] Pakdaman Naeini M, Cooper G, Hauskrecht M (2015) Obtaining well calibrated probabilities using bayesian binning. Proc AAAI Conf Artif Intell 29(1). https://doi.org/10.1609/aaai.v29i1.9602

[67] Paszke A, Gross S, Massa F, et al (2019) PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems

[68] Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830

[69] Selvaraju RR, Cogswell M, Das A, et al (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, pp 618–626, https://doi.org/10.1109/iccv.2017.74

[70] Zhang J, Lin Z, Brandt J, et al (2016) Top-down neural attention by excitation backprop. In: Computer Vision – ECCV 2016. Springer International Publishing, pp 543–559, https://doi.org/10.1007/978-3-319-46493-0_33

[71] Steyerberg EW, Vergouwe Y (2014) Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 35(29):1925–1931. https://doi.org/10.1093/eurheartj/ehu207

[72] Lakshminarayanan B, Pritzel A, Blundell C (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems

[73] Dietterich TG (2000) Ensemble methods in machine learning. In: Multiple Classifier Systems. Springer Berlin Heidelberg, pp 1–15, https://doi.org/10.1007/3-540-45014-9_1

[74] Saporta A, Gui X, Agrawal A, et al (2022) Benchmarking saliency methods for chest X-ray interpretation. Nat Mach Intell 4(10):867–878. https://doi.org/10.1038/s42256-022-00536-x

[75] Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x

[76] Dapamede T, Li F, Khosravi B, et al (2025) DICOM LUT is a key step in medical image preprocessing towards AI generalizability. J Imaging Inform Med 38(5):3040–3048.

603 https://doi.org/10.1007/s10278-025-01418-5

**Table 3** Discrimination of all configurations and ensembles (AUC, mean ± standard deviation over three seeds). Configurations are ordered by mean RSNA validation AUC, the selection criterion

| Model | RSNA validation | RSNA test | Kermany | Indiana |
|---|---|---|---|---|
| Ensemble: top 5 by validation | – | 0.903 ± <0.001 | 0.921 ± 0.008 | 0.885 ± 0.003 |
| Ensemble: 7 baseline models | – | 0.889 ± 0.001 | 0.934 ± 0.005 | 0.848 ± 0.002 |
| Ensemble: 5 ImageNet baselines | – | 0.889 ± 0.002 | 0.918 ± 0.006 | 0.848 ± 0.002 |
| DINOv2-B, lr 1e-5 | 0.899 | 0.898 ± 0.003 | 0.906 ± 0.020 | 0.863 ± 0.010 |
| ConvNeXtV2-T, robust recipe | 0.898 | 0.897 ± 0.002 | 0.885 ± 0.051 | 0.866 ± 0.013 |
| ConvNeXt-S | 0.897 | 0.894 ± 0.005 | 0.908 ± 0.021 | 0.876 ± 0.008 |
| Swin-T, robust recipe | 0.896 | 0.893 ± 0.002 | 0.869 ± 0.023 | 0.871 ± 0.011 |
| DINOv2-S, lr 1e-5 | 0.895 | 0.896 ± 0.001 | 0.899 ± 0.028 | 0.867 ± 0.010 |
| Swin-T | 0.895 | 0.897 ± 0.001 | 0.862 ± 0.041 | 0.865 ± 0.006 |
| ConvNeXtV2-T | 0.895 | 0.896 ± 0.004 | 0.876 ± 0.021 | 0.863 ± 0.004 |
| ViT-S/16, lr 1e-5 | 0.894 | 0.891 ± 0.001 | 0.904 ± 0.021 | 0.848 ± 0.007 |
| DenseNet-121, robust recipe | 0.889 | 0.889 ± 0.005 | 0.845 ± 0.078 | 0.852 ± 0.005 |
| DenseNet-201 | 0.889 | 0.886 ± 0.003 | 0.897 ± 0.005 | 0.838 ± 0.007 |
| EfficientNet-B0 | 0.889 | 0.889 ± 0.001 | 0.831 ± 0.045 | 0.845 ± 0.008 |
| EfficientNetV2-S | 0.885 | 0.879 ± <0.001 | 0.850 ± 0.043 | 0.835 ± 0.018 |
| EfficientNetV2-S, robust recipe | 0.885 | 0.880 ± 0.003 | 0.926 ± 0.023 | 0.828 ± 0.015 |
| EfficientNet-B0, robust recipe | 0.884 | 0.881 ± 0.003 | 0.830 ± 0.038 | 0.820 ± 0.010 |
| ResNet-50, robust recipe | 0.884 | 0.885 ± 0.003 | 0.893 ± 0.040 | 0.844 ± 0.009 |
| DenseNet-121 | 0.883 | 0.885 ± 0.002 | 0.887 ± 0.037 | 0.841 ± 0.005 |
| DenseNet-121, CheXpert init, robust recipe | 0.883 | 0.885 ± 0.001 | 0.916 ± 0.008 | 0.825 ± 0.002 |
| ResNet-50 | 0.879 | 0.878 ± 0.002 | 0.870 ± 0.041 | 0.819 ± 0.010 |
| DINOv2-S, robust recipe | 0.878 | 0.874 ± 0.003 | 0.912 ± 0.005 | 0.844 ± 0.009 |
| DenseNet-121, CheXpert init | 0.875 | 0.876 ± 0.001 | 0.893 ± 0.007 | 0.810 ± 0.005 |
| DenseNet-121, MIMIC init | 0.870 | 0.868 ± 0.001 | 0.879 ± 0.011 | 0.803 ± 0.007 |
| ViT-S/16 | 0.870 | 0.865 ± 0.014 | 0.872 ± 0.018 | 0.804 ± 0.017 |
| DINOv2-B (1 seed) | 0.863 | 0.853 | 0.823 | 0.789 |
| ConvNeXt-T | 0.860 | 0.858 ± 0.009 | 0.851 ± 0.038 | 0.799 ± 0.013 |
| Custom CNN | 0.857 | 0.857 ± 0.004 | 0.690 ± 0.044 | 0.801 ± 0.003 |
| DenseNet-121, view-balanced | 0.857 | 0.861 ± 0.008 | 0.788 ± 0.058 | 0.822 ± 0.018 |
| DenseNet-121, CheXpert init, view-balanced | 0.844 | 0.845 ± 0.001 | 0.801 ± 0.006 | 0.796 ± 0.013 |
| DINOv2-S | 0.834 | 0.828 ± 0.004 | 0.767 ± 0.041 | 0.751 ± 0.008 |

Ensembles average the softmax outputs of their members within each seed. A single value without a standard deviation comes from one seed.

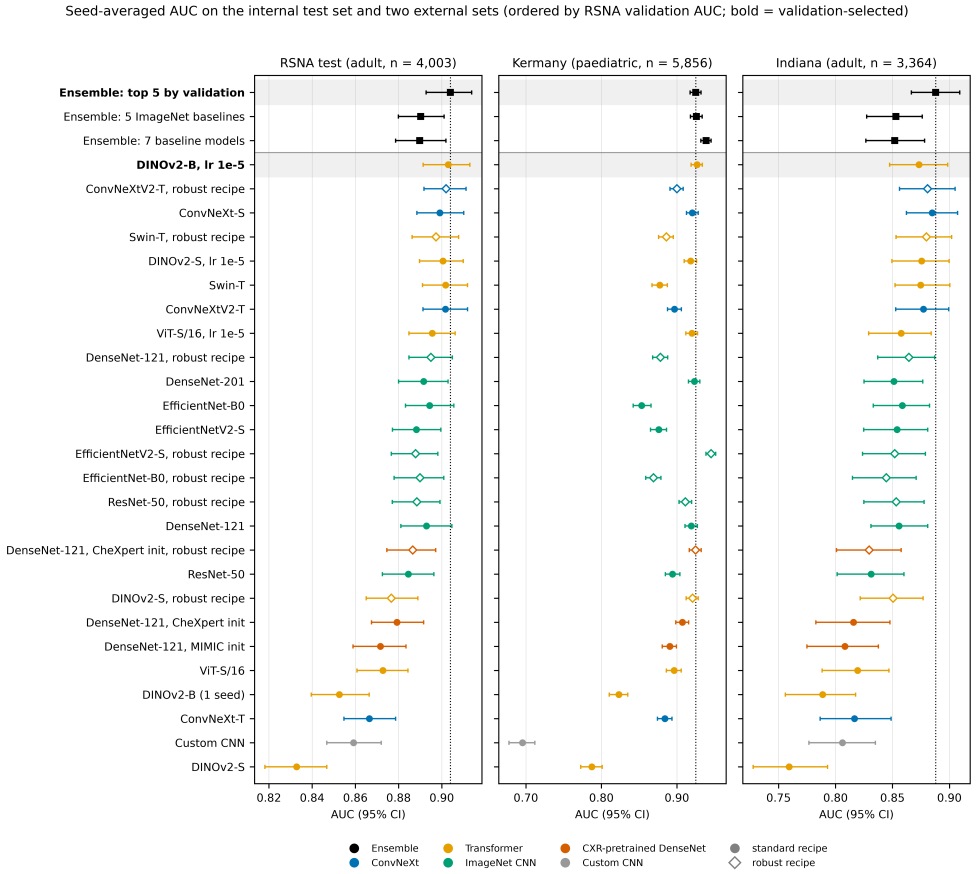



**Fig. 2** Seed-averaged AUC with bootstrap 95% confidence intervals on the RSNA test split and the two external cohorts. Rows are ordered by mean RSNA validation AUC; the validation-selected ensemble and single configuration are shaded, and the dotted line marks the ensemble. Kermany intervals resample patients

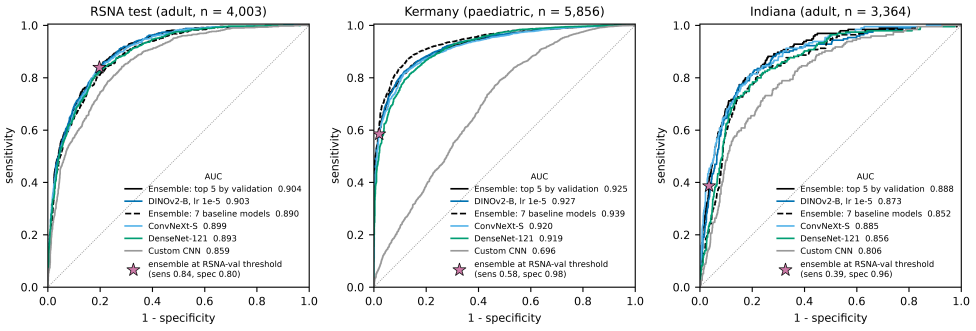



**Fig. 3** Receiver operating characteristic curves of the validation-selected ensemble, the best single configuration and reference models. The star marks the ensemble at the Youden threshold fixed on RSNA validation; the same threshold gives very different operating points in the external cohorts

**Table 4** Operating characteristics at thresholds fixed on RSNA validation and after local re-estimation of the threshold

| Test set | Threshold | Sensitivity | Specificity | PPV | NPV |
|---|---|---|---|---|---|
| *Ensemble: top 5 by validation* | | | | | |
| RSNA test | Youden, 0.54 | 0.839 (0.817–0.862) | 0.803 (0.788–0.815) | 0.553 (0.528–0.580) | 0.945 (0.936–0.954) |
| RSNA test | Sens. 90%, 0.42 | 0.906 (0.887–0.926) | 0.725 (0.709–0.740) | 0.489 (0.466–0.513) | 0.964 (0.956–0.972) |
| Kermany | Youden, 0.54 | 0.585 (0.566–0.604) | 0.979 (0.972–0.986) | 0.987 (0.982–0.991) | 0.466 (0.446–0.489) |
| Kermany | Sens. 90%, 0.42 | 0.778 (0.764–0.794) | 0.907 (0.892–0.921) | 0.958 (0.950–0.964) | 0.602 (0.578–0.627) |
| Kermany | Local 10% | 0.834 (0.737–0.885) | 0.865 (0.788–0.942) | 0.943 (0.919–0.971) | 0.658 (0.571–0.716) |
| Indiana | Youden, 0.54 | 0.387 (0.321–0.453) | 0.965 (0.958–0.971) | 0.403 (0.333–0.472) | 0.963 (0.956–0.969) |
| Indiana | Sens. 90%, 0.42 | 0.572 (0.502–0.644) | 0.933 (0.924–0.941) | 0.344 (0.291–0.393) | 0.973 (0.967–0.979) |
| Indiana | Local 10% | 0.791 (0.545–0.943) | 0.826 (0.585–0.936) | 0.214 (0.124–0.345) | 0.985 (0.971–0.994) |
| *DINOv2-B, lr 1e-5* | | | | | |
| RSNA test | Youden, 0.45 | 0.826 (0.802–0.850) | 0.807 (0.793–0.821) | 0.555 (0.530–0.581) | 0.941 (0.932–0.950) |
| RSNA test | Sens. 90%, 0.31 | 0.915 (0.897–0.933) | 0.712 (0.697–0.728) | 0.480 (0.458–0.504) | 0.966 (0.959–0.974) |
| Kermany | Youden, 0.45 | 0.591 (0.574–0.611) | 0.981 (0.974–0.987) | 0.988 (0.984–0.992) | 0.471 (0.449–0.494) |
| Kermany | Sens. 90%, 0.31 | 0.800 (0.785–0.814) | 0.901 (0.886–0.916) | 0.956 (0.949–0.963) | 0.625 (0.601–0.650) |
| Kermany | Local 10% | 0.823 (0.760–0.885) | 0.882 (0.790–0.925) | 0.949 (0.919–0.965) | 0.648 (0.586–0.717) |
| Indiana | Youden, 0.45 | 0.392 (0.329–0.460) | 0.958 (0.951–0.965) | 0.365 (0.301–0.432) | 0.963 (0.956–0.969) |
| Indiana | Sens. 90%, 0.31 | 0.557 (0.485–0.624) | 0.926 (0.917–0.935) | 0.317 (0.266–0.362) | 0.972 (0.966–0.977) |
| Indiana | Local 10% | 0.788 (0.575–0.903) | 0.812 (0.692–0.918) | 0.207 (0.148–0.302) | 0.984 (0.973–0.991) |

Values with bootstrap 95% confidence intervals (Kermany: patient resampling). Thresholds on the ensemble's or model's seed-averaged probability: Youden, maximum of Youden's index on RSNA validation; Sens. 90%, highest threshold with at least 90% sensitivity on RSNA validation. For the local threshold, the Youden threshold was re-estimated on a random 10% of the cohort and evaluated on the remaining 90%; values are the median and 2.5th–97.5th percentiles over 500 repetitions. PPV positive predictive value, NPV negative predictive value.

**Table 5** Grad-CAM localisation on the 902 positive RSNA test radiographs (seed 42)

| Model | Positives | Hits, % (all) | Detected | Hits, % (detected) | Energy in boxes, % | Energy/area |
|---|---|---|---|---|---|---|
| DINOv2-B, lr 1e-5 | 902 | 85.0 | 768 | 88.5 | 78.1 (54.9–92.8) | 5.33 |
| ConvNeXtV2-T, robust recipe | 902 | 82.5 | 766 | 86.9 | 66.4 (44.0–83.3) | 4.58 |
| ConvNeXt-S | 902 | 77.7 | 778 | 82.5 | 57.9 (35.6–78.1) | 3.98 |
| DenseNet-121 | 902 | 69.2 | 751 | 76.6 | 35.2 (20.4–52.0) | 2.52 |

A hit is a map maximum inside a radiologist box (pointing game). Detected: probability above the run's validation Youden threshold. Energy in boxes: median (interquartile range) share of map energy inside the boxes; energy/area divides it by the box-area share. A map without spatial information would score 15.3% hits and an energy/area ratio of 1.

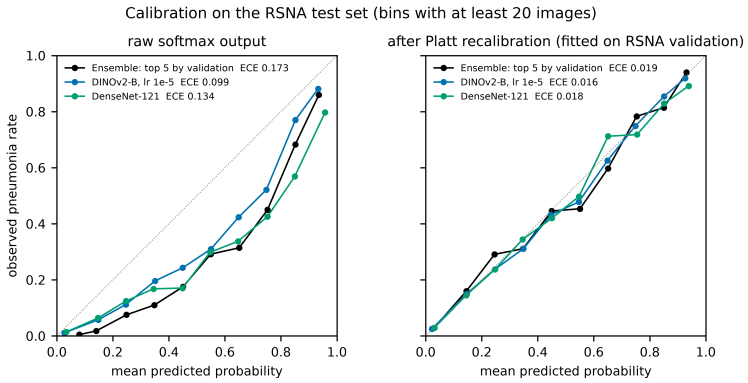



**Fig. 4** Reliability diagrams on the RSNA test split before and after Platt recalibration fitted on RSNA validation (ten equal-width bins with at least 20 radiographs). ECE expected calibration error, computed with 15 bins

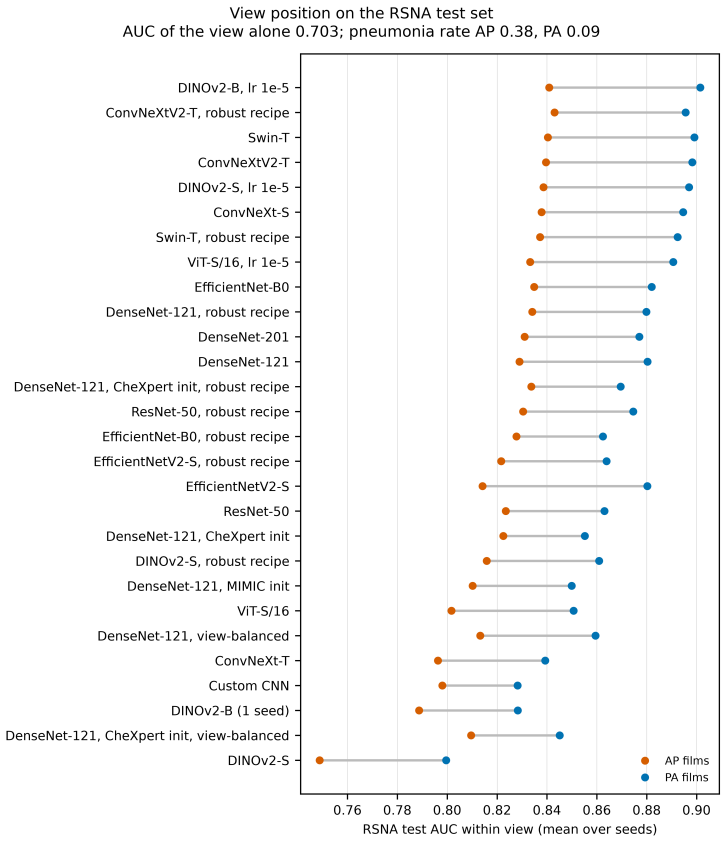



**Fig. 5** AUC within anteroposterior (AP) and within posteroanterior (PA) films of the RSNA test split, mean over seeds, ordered by overall test AUC

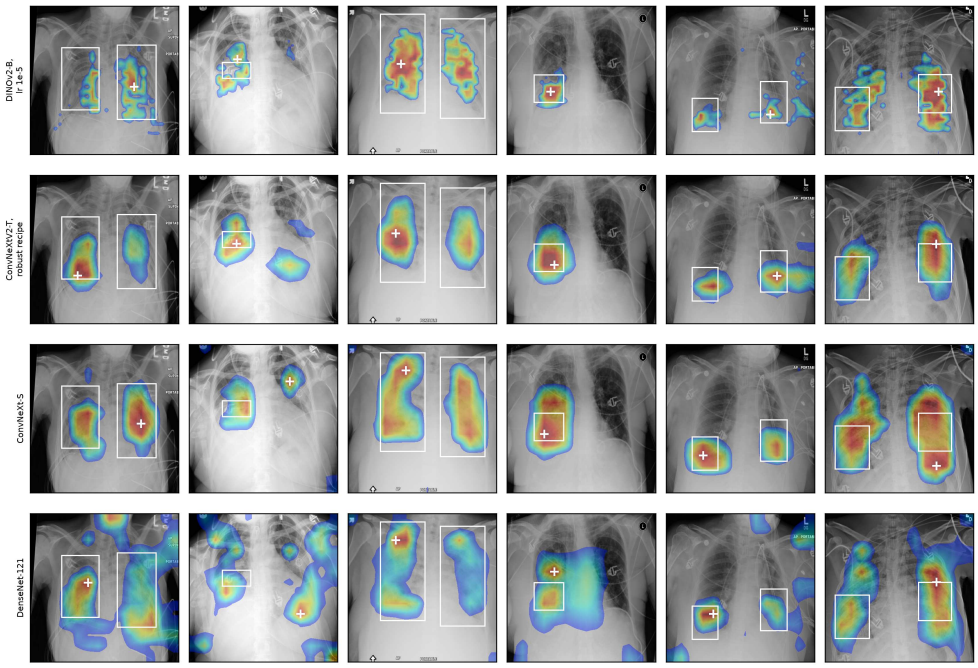



**Fig. 6** Grad-CAM maps for six positive RSNA test radiographs drawn at random from those detected by all four models (rows: DINOv2-B, lr $10^{-5}$; ConvNeXtV2-T, robust recipe; ConvNeXt-S; DenseNet-121). White boxes are radiologist annotations; the cross marks the map maximum

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf